

RÉPUBLIQUE ALGÉRIENNE DÉMOCRATIQUE ET POPULAIRE  
Ministère de l'enseignement Supérieur et de la Recherche Scientifique  
Université Mohammed Seddik Ben Yahia - Jijel



Faculté des Sciences Exacte et Informatique  
Département de Mathématique

## Mémoire de fin d'études

Présenté pour l'obtention du diplôme de

**Master**

**Spécialité** : Mathématiques

**Option** : Probabilités et statistique

**Thème**

**Analyse de la régression non paramétrique**

**Présenté par :**

Kadi Abir

Zelliche Warda

**Devant le jury :**

Président : *M<sup>me</sup>* Madi Meriem (M.A.A) Université de Jijel

Encadreur : *M<sup>me</sup>* Djeridi Zahra (M.A.A) Université de Jijel

Examinatrice : *M<sup>me</sup>* Abdi Zeyneb (M.A.A) Université de Jijel

Promotion **2018/2019**

---

# RÉSUMÉ

Ce mémoire porte sur l'estimateur non paramétrique de la fonction de régression, pour les données réelles complètes en utilisant la méthode à noyau. Puis nous allons la comparer avec la méthode de LOESS (par polynômes locale et par polynôme locale pondéré) et la méthode de B-spline. Finalement un travail de simulation a permis de vérifier la bonne performance des estimateurs.

---

# REMERCIEMENT

*Tout d'abord, nous tenons à remercier Allah, le tout puissant et le miséricordieux, de nous avoir donné la santé, la volonté et la patience pour mener à terme notre formation de master.*

*C'est avec un grand honneur que nous remercions notre enseignante et promoteur M<sup>me</sup> **Djeridi Zahra** pour nous avoir dirigés pour la réalisation de ce travail, pour ces précieux conseils et ces encouragements.*

*Nous tenons à remercier M<sup>me</sup> **Madi Meriem** d'avoir accepté la présidence du jury de notre travail, qu'il trouve ici toutes nos expressions respectueuses.*

*Nous remercions également M<sup>me</sup> **Abdi Zeyneb** de nous avoir fait l'honneur de faire partie des membres du jury et d'examiner ce travail.*

*Nous tenons à vous remercier.*

*Nos remerciements les plus sincères s'adressent à nos familles pour leur soutien sans faille et pour l'équilibre qu'elles nous ont apporté et pour leurs encouragements.*

*Enfin, nous voulons remercier toutes les personnes qui ont contribué de loin ou de près à l'avancement de ce travail, nos enseignants et collègues à l'université de Jijel.*

---

# TABLE DES MATIÈRES

<b>Notations</b>	<b>7</b>
<b>1 La régression non paramétrique univariée</b>	<b>10</b>
1.1 Présentation générale des fonctions de lissage . . . . .	10
1.2 Estimation par la méthode du noyau . . . . .	12
1.2.1 Principe de la méthode . . . . .	13
1.2.2 Construction de l'estimateur : . . . . .	15
1.3 Biais et variance de l'estimateur . . . . .	19
1.4 Consistance . . . . .	21
1.4.1 Consistance faible . . . . .	21
1.4.2 Consistance forte . . . . .	26
1.5 Convergence asymptotique en loi et intervalles de confiance . . . . .	27
<b>2 Choix du paramètre de lissage</b>	<b>31</b>
2.1 Etude du critère d'erreur quadratique moyenne de $\hat{r}(x)$ . . . . .	31
2.2 Validation croisée . . . . .	43

---

2.2.1	Principe de la méthode . . . . .	43
2.2.2	Plan de travail . . . . .	44
<b>3</b>	<b>Comparaison avec les estimateur Loess et Spline</b>	<b>47</b>
3.1	La régression locale : Loess . . . . .	48
3.2	La régression par polynômes locaux pondérées . . . . .	49
3.3	La régression polynomiale locale (RPL) . . . . .	51
3.4	B-spline : . . . . .	53
3.4.1	Splines linéaires et cubiques : . . . . .	54
3.5	Application : . . . . .	56
<b>A</b>	<b>Les convergences</b>	<b>67</b>
A.1	Convergence en loi . . . . .	67
A.2	Convergence en probabilité : . . . . .	67
A.3	Convergence presque sûre : . . . . .	68
A.4	Convergence presque complète : . . . . .	68
A.5	Liens entre les convergences : . . . . .	69
A.6	Notations $O(h)$ et $o(h)$ . . . . .	69
<b>B</b>	<b>Les théorème utilisès :</b>	<b>70</b>
B.1	Théorème de Bochner . . . . .	70
B.2	Théorème d'Inégalité Cauchy Shwartz . . . . .	71
B.3	Théorème de Fubini . . . . .	71



---

## NOTATIONS

$r(\cdot)$	Fonction de régression.
$\hat{r}(\cdot)$	L'estimateur de la fonction de régression.
$K(\cdot)$	Fonction noyau.
$h$	Paramètre de lissage ou Fenêtre.
$h_{opt}$	Fenêtre optimale.
$\mathbb{E}$	Espérance de probabilité.
$1(\cdot)$	L'indicatrice de l'évènement.
<i>Biais</i>	Biais d'un estimateur.
<i>Var</i>	Variance d'un estimateur.
<i>Cov</i>	la covariance.
$f$	Densité marginale de X.
$\hat{f}$	L'estimateur de la densité marginale de X.
$F$	Densité de répartition.
<i>MSE</i>	L'erreur moyenne quadratique.
<i>MISE</i>	L'erreur moyenne quadratique intégrée.
<i>AMSE</i>	L'erreur moyenne quadratique asymptotique.
<i>AMISE</i>	L'erreur moyenne quadratique intégrée asymptotique.
<i>IC</i>	Intervalle de confiance.
$CV \sim \hat{R}(h)$	Fonction noyau.
$\hat{r}^{(-v(i))}(X_i)$	L'estimateur de r obtenu en enlevant les observations du bloc v(i)
$N(p)$	Voisinage de p.
$V_i(\cdot)$	Base de splines.
$L$	Matrice des poids.
$L_s$	Matrice de lissage.
$X_x$	Matrice des données.
$C(\cdot)$	Critère pénalisé.
$\  \cdot \ $	La norme euclidienne.

---

# INTRODUCTION

La théorie d'estimation est une des branches les plus basiques de la statistique. Cette théorie est habituellement divisée en deux composantes principales : l'estimations paramétrique et l'estimation non-paramétrique.

L'un des modèles le plus fréquemment rencontré en statistique paramétrique ou non paramétrique est le modèle de régression. Le principe de la régression non paramétrique remonte au 19<sup>ème</sup> siècle. La régression non paramétrique est devenue une méthode populaire pour analyser une relation entre une variable dépendante  $Y$  et une variable indépendante  $X$ . Son objet, est d'estimer cette relation de dépendance sans faire d'hypothèses paramétriques sur la forme de cette dépendance. Ces méthodes permettent de construire une forme d'analyse de la régression dans lequel la fonction d'estimation, ne prend pas une forme prédéterminée. Elle est construite selon les informations provenant des données et exige que la taille d'échantillon soit plus importante que celle de la régression basée sur des modèles paramétriques parce que les données doivent fournir la structure du modèle ainsi que l'estimation du modèle. Il n'est pas toujours évident de savoir ce que recouvrent ces méthodes, ce qu'elles peuvent apporter et quelles en sont les limites.

Parmi les méthodes non paramétriques les plus robustes : la méthode à noyau, la méthode Loess (par polynôme locale et par polynôme locale pondéré) sans oublié la méthode spline précisément les B-spline. Qui sont discuté dans le court de ce mémoire.

En ce qui concerne le choix du paramètre de lissage, il existe diverses méthodes pour le choix du paramètre de la fenêtre  $h$ . L'estimation de la fonction de régression par la méthode à noyau et Loess est principalement conditionnée par le paramètre de lissage  $h$ ,



ce paramètre est un facteur important dans l'estimation par ces méthodes. Il représente en quelque sorte une fenêtre qui permet de déterminer le degré de lissage de l'estimation d'une fonction de régression. Un faible paramètre de lissage implique un faible degré de lissage et en résulte une fonction de densité irrégulière. À l'opposé, une large valeur de  $h$  conduit à une estimation lisse.

Plusieurs méthodes ont été introduites afin de choisir le paramètre  $h$ . La fenêtre optimale théorique relative à l'estimateur à noyau minimise l'erreur quadratique moyenne intégrée. Cependant, cette fenêtre est inconnue en pratique, plusieurs travaux se sont intéressés à la procédure du choix d'une fenêtre pouvant approximer celle-ci. Pour la sélection de la fenêtre on utilise la méthode de la validation croisée, ce choix est ainsi asymptotiquement optimal. Notons qu'on considère aussi le choix de la fenêtre qui établit l'équivalence asymptotique entre l' $MSE$  et l' $MISE$ .

Dans ce mémoire nous exposerons, en détail, une méthode d'estimation de la fonction de régression qui est la méthode à noyau, comparée par deux autres méthodes : Loess et B-spline. Le travail composé principalement de trois chapitres :

- ▷ Le premier chapitre est composé de deux parties :
  - ▷ La première partie porte sur des généralités sur la courbe de régression non paramétrique et quelques notions utilisées.
  - ▷ La deuxième partie du chapitre propose des méthodes d'estimations de la courbe de régression qui est l'estimation par noyau de Nadarya Watson, sans oublier les propriétés statistiques et asymptotiques de l'estimateur.
- ▷ Dans le deuxième chapitre, nous présentons deux méthodes de choix du paramètre de lissage par l' $MSE$ , l' $MISE$  et par la méthode de validation croisée avec des exemples de simulations.
- ▷ Finalement, le dernier chapitre composé de trois parties, on donne.
  - ▷ La méthode Loess ( polynômes locale et par polynôme locale pondéré).
  - ▷ La méthode B-spline.
  - ▷ Une comparaison entre les trois estimateurs présenté dans ce mémoire.

On termine le mémoire, par une simulation en utilisant des données réels ramenées de la clinique des 40 Hectares de Jijel pour l'année 2018-2019.

---

---

# CHAPITRE 1

---

## LA RÉGRESSION NON PARAMÉTRIQUE UNIVARIÉE

Ce chapitre est consacré à la présentation de la méthode principale de régression non paramétrique univariée qui est la méthode à noyau. Cette méthode est très utile lorsqu'on veut d'écrire la relation entre une variable à expliquer  $Y$  et une variable explicative  $X$ , sans supposer une forme particulière. On discute la qualité de l'estimateur en donnant la consistance et la convergence en loi pour construire l'intervalle de confiance.

### 1.1 Présentation générale des fonctions de lissage

Soit un échantillon aléatoire (par opposition au cadre déterministe) composé des couples  $(x_i, y_i)$ ;  $i = 1, \dots, n$ , où les  $(x_i)$  représentent les valeurs observées de la variable explicative  $X$  qui est une variable aléatoire réelle de loi donnée par la fonction de répartition  $F$  supposé dérivable, de dérivée  $f$  (qui est donc la densité de  $X$  par rapport à la mesure de Lebesgue sur  $\mathbb{R}$ ), et les  $y_i$  représentent celles de la variable dépendante  $Y$  (dite aussi la variable d'intérêt). Alors, le modèle de régression non paramétrique univariée est donné par :

$$Y_i = r(x_i) + \epsilon_i; \quad i = 1, \dots, n. \quad (1.1)$$

Où  $r(\cdot)$  est une fonction inconnue appelé *fonction de régression* et les  $\epsilon_i$  sont les erreurs aléatoires réelles, appelées "bruit", sous les hypothèses suivantes :

1.  $\mathbb{E}(\epsilon_i)=0, \forall i,$
2.  $X \perp \epsilon,$
3. L'homogénéité des variance, i.e,  $\text{Var}(\epsilon_i, \epsilon_i)=\sigma^2, \forall i,$
4. La non auto-corrélation des erreurs, i.e,  $\text{Cov}(\epsilon_i, \epsilon_j)=0$  , pour  $i \neq j.$

**Définition 1.1.** *Modèle non paramétrique*

Nous pouvons retenir une approche non paramétrique dans laquelle on va estimer la relation entre le niveau moyen de  $Y$  et toutes les valeurs réalisées de  $X$ . Nous ne supposons aucune forme spécifique sur la fonction de régression, tel que,  $r(x)$  est la moyenne conditionnelle de la courbe de régression, c'est-à-dire :

$$\mathbb{E}(Y|X = x) = r(x). \quad (1.2)$$

L'hypothèse de base est que  $(X; Y)$  est un vecteur aléatoire de  $\mathbb{R}^2$ . Le paramètre fonctionnel du modèle, que nous cherchons à estimer, est la fonction  $r$  de régression :

$$r : \mathbb{R}^2 \rightarrow \mathbb{R} .$$

**Remarque 1.1.** *Le principal avantage de cette approche est qu'elle ne nécessite aucune hypothèse a priori sur la forme du lien entre  $X$  et  $Y$ . Avec une approche non paramétrique, on aboutit à :*

1. Une représentation graphique de la relation entre  $X$  et  $Y$ .
2. Il n'existe pas de forme analytique de la fonction de lien  $r(x)$ .

Tout le problème consiste alors à estimer cette fonction de régression  $r(x)$ . C'est donc un problème non paramétrique.

On va donner quelques définitions liées l'estimateur de la fonction de régression  $r$ .

**Définition 1.2.** *(Biais et variance) Le biais d'un estimateur  $\hat{r}$  de  $r$  est l'écart :*

$$\text{Biais}(\hat{r}(x)) = \mathbb{E}[\hat{r}(x) - r(x)].$$

On dit que l'estimateur est sans biais si :

$$\mathbb{E}(\hat{r}) = r.$$

On définit la variance de l'estimateur  $\hat{r}$  par :

$$\text{Var}(\hat{r}(x)) = \mathbb{E}[(\hat{r}(x) - r(x))^2]$$

**Définition 1.3.** On dit qu'un estimateur  $\hat{r}$  de  $r$  est asymptotiquement sans biais si :

$$\lim_{n \rightarrow \infty} \sup \mathbb{E}[\hat{r} - r] = 0$$

**Définition 1.4.** L'erreur moyenne quadratique MSE :

$$\begin{aligned} MSE(r(x), \hat{r}(x)) &= \mathbb{E}(r(x) - \hat{r}(x))^2 \\ &= \mathbb{E}(r(x))^2 + \mathbb{E}(\hat{r}(x))^2 - 2\mathbb{E}(r(x)\hat{r}(x)) \\ &= \mathbb{E}(r(x))^2 + \mathbb{E}(\hat{r}(x))^2 - 2\mathbb{E}(r(x)\hat{r}(x)) + [\mathbb{E}(\hat{r}(x))]^2 - [\mathbb{E}(\hat{r}(x))]^2 \\ &= (r(x))^2 + \mathbb{E}(\hat{r}(x))^2 - 2r(x)\mathbb{E}(\hat{r}(x)) + [\mathbb{E}(\hat{r}(x))]^2 - [\mathbb{E}(\hat{r}(x))]^2 \\ &= [\mathbb{E}(\hat{r}(x) - r(x))]^2 + \mathbb{E}(\hat{r}(x))^2 - [\mathbb{E}(\hat{r}(x))]^2 \\ &= \text{Biais}(\hat{r}(x))^2 + \text{Var}(\hat{r}(x)) \end{aligned}$$

**Définition 1.5.** L'erreur moyenne quadratique intégrée MISE :

$$\begin{aligned} MISE(\hat{r}, r) &= \int MSE(r(x), \hat{r}(x)) dx \\ &= \int \text{Biais}(\hat{r}(x))^2 dx + \int \text{Var}(\hat{r}(x)) dx \end{aligned} \quad (1.3)$$

**Définition 1.6.** On dit qu'un estimateur  $\hat{r}$  de  $r$  est ponctuellement consistant en moyenne quadratique si :

$$\lim_{n \rightarrow \infty} MSE(r(x), \hat{r}(x)) = 0$$

**Définition 1.7.** On dit qu'un estimateur  $\hat{r}$  de  $r$  est uniformément consistante en moyenne quadratique si :

$$\lim_{n \rightarrow \infty} MISE(r(x), \hat{r}(x)) = 0$$

**Définition 1.8.** On dit qu'un estimateur  $\hat{r}$  de  $r$  est asymptotiquement normal si :

$$\hat{r} \longrightarrow \mathcal{N}(\mathbb{E}(\hat{r}), \text{Var}(\hat{r})) \text{ en loi.}$$

## 1.2 Estimation par la méthode du noyau

La classe des estimateurs linéaires regroupe la majorité des estimateurs de la régression, comme l'estimation par fonctions splines ou B-splines, l'estimation par des polynômes par morceaux, et par la méthode du noyau. Dans la section suivante, nous présenterons le célèbre estimateur à noyau de la régression introduit par Nadaraya et Watson et quelques unes de ses propriétés essentielles. Sans oublier l'étude de la consistance et la convergence de cet estimateur.

### 1.2.1 Principe de la méthode

Le problème consiste à estimer la fonction de régression en tous points  $x_1, x_2, \dots, x_n$ . Le principe de la méthode du noyau repose en fait sur des méthodes de lissage, elle donne, pour estimateur de  $\mathbb{E}(Y/X = x) = r(x)$ , une moyenne pondérée des valeurs  $y_i$  pour les  $i$  dont le point  $x_i$  est proche du point d'estimation. Il est évident que le choix d'un point d'estimation  $x_0$ , une valeur de  $x$  pour laquelle on veut estimer  $r(x_0)$ , doit être fait. [1]

#### L'estimateur à noyau

Parmi les estimateurs non-paramétriques, l'estimateur à noyau est, de loin, le plus populaire. Les méthodes à noyaux permettent de trouver des fonctions de décision non linéaires, tout en s'appuyant fondamentalement sur des méthodes linéaires. Grâce à l'utilisation de fonctions noyau, il devient ainsi possible d'avoir le meilleur traitement des problèmes non linéaires.

**Définition 1.9.** *Un estimateur à noyau Parzen et Rosenblatt de la densité  $f$  est une fonction  $\hat{f}$  définie par :*

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

Où  $(h)_{n \geq 1}$  est une suite de réels positifs appelés paramètres de lissage ou largeur de la fenêtre, qui tend vers 0 quand  $n$  tend vers l'infini, et  $K$  est une densité de probabilité appelée noyau.

#### Conditions sur le 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  $K$  (pour plus de détails voir [3]) :

- 1-  $K$  est bornée ; i.e,  $\sup_{u \in \mathbb{R}} |K(u)| \leq M < \infty$  ;
- 2-  $\lim_{|u| \rightarrow \infty} |u| K(u) = 0$  ;
- 3-  $\int |K(u)| du < \infty$  ;
- 4-  $\int K(u) du = 1$  ;
- 5-  $\forall u \in \mathbb{R}, K(u) = K(-u)$  ;

$$6- \int_{-\infty}^{+\infty} u^2 K(u) du < \infty.$$

$$7- \int_{\mathbb{R}} u K(u) du = 0$$

**Exemple 1.1.** On donne quelques exemples des noyaux :

1. Noyau rectangulaire (uniforme) :

$$K_1(u) = \begin{cases} 1/2, & |u| \leq 1 \\ 0, & |u| > 1 \end{cases}$$

2. Noyau triangulaire :

$$K_2(u) = \begin{cases} (1 - |u|), & |u| \leq 1 \\ 0, & |u| > 1 \end{cases}$$

3. Noyau parabolique ou d'Epanechnikov :

$$K_3(u) = \begin{cases} \frac{3}{4}(1 - u^2), & |u| \leq 1 \\ 0, & |u| > 1 \end{cases}$$

4. Noyau "biweight" quadratique :

$$K_4(u) = \begin{cases} \frac{15}{16}(1 - u^2)^2, & |u| \leq 1 \\ 0, & |u| > 1 \end{cases}$$

5. Noyau gaussien

$$K_5(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right); \quad \forall u \in \mathbb{R}.$$

6. Noyau cubique :

$$K_6(u) = \begin{cases} \frac{35}{32}(1 - u^2)^3, & |u| \leq 1 \\ 0, & |u| > 1 \end{cases}$$

Les courbes de certains noyaux sont présentées ci-dessous :

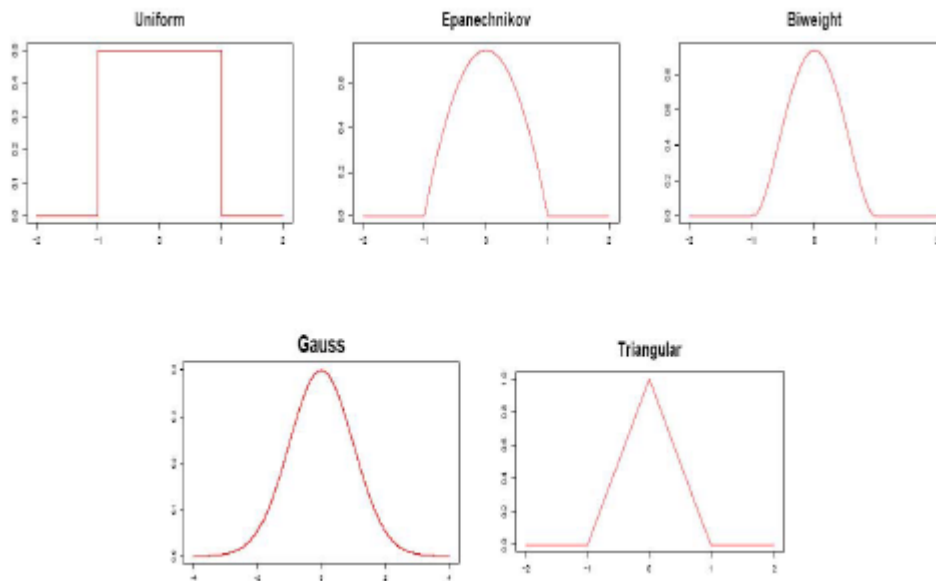


FIGURE 1.1: Les courbes de certains noyaux

### 1.2.2 Construction de l'estimateur :

Dans cette section, on donne deux méthodes différentes de la construction de l'estimateur à noyau, pour plus de détail voir([1],[23])

#### I- Méthode classique :

**Proposition 1.1.** *En supposant que la densité  $f$  soit continue, l'estimateur à noyau de la densité  $f$ , en un point  $x$ , est construit en comptant le nombre d'observation dans l'intervalle autour de  $x$  qui est donné par  $[x-h, x+h]$ . Alors, on écrit :*

$$\begin{aligned}\hat{f}(x) &= \frac{1}{2nh} \sum_{i=1}^n 1_{[x-h, x+h]}(x_i) \\ &= \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - x_i}{h}\right)\end{aligned}$$

**Démonstration.** On sait que :  $x_i \in [x - h, x + h]$  donc :

$$x - h \leq x_i \leq x + h \quad (1.4)$$

$$1 \geq \frac{x - x_i}{h} \geq -1 \quad (1.5)$$

On fait un changement de variable :

$$\begin{aligned}
 \hat{f}(x) &= \frac{1}{2nh} \sum_{i=1}^n 1_{[x-h, x+h]}(x_i) \\
 &= \frac{1}{nh} \sum_{i=1}^n \frac{1}{2} 1_{[x-h, x+h]}(x_i) \\
 &= \frac{1}{nh} \sum_{i=1}^n \frac{1}{2} 1_{[-1, +1]}\left(\frac{x - x_i}{h}\right) \\
 &= \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - x_i}{h}\right).
 \end{aligned}$$

Où  $K(u) = \frac{1}{2} 1_{[-1, +1]}(u)$  ■

Nous reprenons le modèle

$$r(x) = \mathbb{E}(Y/X = x) = \frac{g(x)}{f(x)} = \frac{\int y f_{X,Y}(x, y) dy}{\int f_{X,Y}(x, y) dy} \quad (1.6)$$

où  $f_{X,Y}(\cdot, \cdot)$  est la densité jointe sur  $\mathbb{R}^2$  et nous désignons par  $f(x)$  la densité marginale de  $X$  (par rapport à la mesure de Lebesgue sur  $\mathbb{R}$ ), aussi :

$$\begin{aligned}
 g(x) &= \int_{\mathbb{R}} y f_{X,Y}(x, y) dy \\
 &= \lim_{h \rightarrow 0} \frac{1}{2h} \int_{x-h}^{x+h} \int_{\mathbb{R}} y F_{X,Y}(x, y) (dx, dy) \\
 &= \lim_{h \rightarrow 0} \frac{1}{2h} \mathbb{E}[Y 1_{(|x_i - x| \leq h)}]
 \end{aligned}$$

où  $F_{X,Y}(\cdot, \cdot)$  est la fonction de répartition de  $(X, Y)$

$$\hat{g}(x) = \frac{1}{n} \sum_{i=1}^n y_i \frac{1_{(|x_i - x| \leq h)}}{2h}$$

donc  $r(x)$  est estimé par

$$\hat{r}(x) = \frac{\hat{g}(x)}{\hat{f}(x)} = \frac{\sum_{i=1}^n y_i \frac{1_{(|x_i - x| \leq h)}}{2h}}{\sum_{i=1}^n \frac{1_{(|x_i - x| \leq h)}}{2h}}$$

Cet estimateur se présente sous la forme d'une moyenne locale pondérée des valeurs  $y_i$ , mais il présente le désavantage d'être discontinu. Sa généralisation naturelle est l'estimateur à noyau, défini comme suit :

$$\hat{r}(x) = \frac{\sum_{i=1}^n y_i K\left(\frac{x_i - x}{h}\right)}{\sum_{i=1}^n K\left(\frac{x_i - x}{h}\right)} 1_{\{K\left(\frac{x_i - x}{h}\right) \neq 0\}}, \forall x \quad (1.7)$$



tel que,

$$\hat{g}(x) = \frac{1}{nh} \sum_{i=1}^n y_i K\left(\frac{x_i - x}{h}\right)$$

l'équation (1.7) est l'estimateur de Nadarya Watson [N-W].

où  $1(\cdot)$  désigne la fonction indicatrice. On rappelle que, pour tout événement  $\mathcal{A}$  Borel-mesurable on a,

$$1(\mathcal{A}) := \begin{cases} 1, & \text{si } \mathcal{A} \text{ est vérifié,} \\ 0, & \text{sinon.} \end{cases}$$

De manière similaire, nous pouvons définir l'estimateur Nadarya Watson par,

$$\hat{r}_n^{NW}(x) := \begin{cases} \frac{\sum_{i=1}^n y_i K\left(\frac{x_i - x}{h}\right)}{\sum_{i=1}^n K\left(\frac{x_i - x}{h}\right)}, & \text{lorsque } \sum_{i=1}^n K\left(\frac{x_i - x}{h}\right) \neq 0 \text{ est vérifié,} \\ \frac{1}{n} \sum_{i=1}^n y_i, & \text{sinon.} \end{cases}$$

Le noyau  $K$  détermine la forme du voisinage autour du point  $x$  et la fenêtre  $h$  contrôle la taille de ce voisinage, c'est à dire le nombre d'observations prises pour effectuer la moyenne locale. Intuitivement, il est naturel que la fenêtre  $h$  soit prépondérante pour la consistance de l'estimateur [N-W].

Si on pose

$$W_i(x) = \frac{K\left(\frac{x_i - x}{h}\right)}{\sum_{i=1}^n K\left(\frac{x_i - x}{h}\right)}$$

où  $K(\cdot)$  désigne une fonction noyau,  $h > 0$  un paramètre de lissage (bandwidth parameter), alors, on peut réécrire l'équation (1.7) comme suit :

$$\hat{r}(x) = \sum_i^n y_i W_i(x),$$

dans ce cas l'estimateur  $\hat{r}(x)$  de  $r(x)$  est dit estimateur linéaire de la fonction régression non-paramétrique, et  $W_i(x)$  appelée fonction de poids

## II- Optimisation de l'espérance conditionnelle :

On veut rappeler la relation suivante :

$$\hat{r}(x) = \arg \min_a \mathbb{E}[(Y - a)^2 | X = x] \quad (1.8)$$

le minima de cette égalité est trouvée en différenciant l'espérance :  $\mathbb{E}[(Y - a)^2 | X = x]$  par rapport à  $a$ , en égalant le résultat à 0 et, finalement, en isolant  $a$  :

**Démonstration.**

$$\begin{aligned}
\frac{\partial}{\partial a} \mathbb{E}[(Y - a)^2 | X = x] &= -2\mathbb{E}[(Y - a) | X = x] \\
&= -2\mathbb{E}[Y | X = x] + 2a \\
&= 0 \\
\Rightarrow a &= \mathbb{E}(Y | X = x)
\end{aligned}$$

Le fait que la dérivée seconde, qui se chiffre à 2, soit positive mène à la conclusion que cette valeur  $a$  est bel et bien un minimum, et non un maximum. Ce lien entre la fonction de régression  $r(x)$  et l'optimisation d'une espérance conditionnelle sera exploité à maintes reprises au cours de ce mémoire. ■

L'estimateur est alors, donné par :

$$\begin{aligned}
\hat{r}(x_0) &= \arg \min_a \mathbb{E}[(Y - a)^2 | X = x_0] \\
&= \arg \min_a \frac{\sum_{i=1}^n (Y_i - a)^2 K\left(\frac{x_i - x_0}{h}\right)}{\sum_{i=1}^n K\left(\frac{x_i - x_0}{h}\right)} \\
&= \arg \min_a \sum_{i=1}^n W_i(x_0) (Y - a)^2
\end{aligned}$$

De manière analogue à la démonstration de la proposition (1.1) , il est possible de trouver l'estimateur du noyau. En effet, cet estimateur est obtenu en dérivant la version empirique  $\mathbb{E}[(Y - a)^2 | X = x_0]$  par rapport à  $a$ , comme suit :

**Démonstration.** On pose

$$m = \frac{\sum_{i=1}^n (Y_i - a)^2 K\left(\frac{x_i - x_0}{h}\right)}{\sum_{i=1}^n K\left(\frac{x_i - x_0}{h}\right)}$$

Alors :

$$\begin{aligned}
\frac{\partial m}{\partial a} &= \frac{\partial}{\partial a} \frac{\sum_{i=1}^n (Y_i - a)^2 K\left(\frac{x_i - x_0}{h}\right)}{\sum_{i=1}^n K\left(\frac{x_i - x_0}{h}\right)} \\
&= -2 \frac{\sum_{i=1}^n (Y_i - a) K\left(\frac{x_i - x_0}{h}\right)}{\sum_{i=1}^n K\left(\frac{x_i - x_0}{h}\right)} \\
&= -2 \frac{\sum_{i=1}^n Y_i K\left(\frac{x_i - x_0}{h}\right) - a \sum_{i=1}^n K\left(\frac{x_i - x_0}{h}\right)}{\sum_{i=1}^n K\left(\frac{x_i - x_0}{h}\right)} \\
&= -2 \frac{\sum_{i=1}^n Y_i K\left(\frac{x_i - x_0}{h}\right)}{\sum_{i=1}^n K\left(\frac{x_i - x_0}{h}\right)} + 2a \frac{\sum_{i=1}^n K\left(\frac{x_i - x_0}{h}\right)}{\sum_{i=1}^n K\left(\frac{x_i - x_0}{h}\right)} \\
&= 0
\end{aligned}$$

$$\Rightarrow a = \frac{\sum_{i=1}^n Y_i K\left(\frac{x_i - x_0}{h}\right)}{\sum_{i=1}^n K\left(\frac{x_i - x_0}{h}\right)}$$

La dérivée seconde de cette même version empirique,  $\frac{\partial^2 m}{\partial a^2} = 2 \sum_{i=1}^n \frac{K\left(\frac{x_i - x_0}{h}\right)}{\sum_{i=1}^n K\left(\frac{x_i - x_0}{h}\right)}$  est positive. Certitude que cette valeur  $a$  est bien une valeur minimale. ■

### Remarque 1.2.

1.  $W_i(x)$  représente des poids dépendant de  $x$  et de  $(x_1, \dots, x_n)$ . La valeur de  $W_i(x)$  dépend du type de l'estimateur considéré et ne dépend pas des observations  $y_i$ .
2. Généralement, plus les points  $x_i$  sont proches de  $x_0$ ; plus le poids sera important :  $W(x_0)$  est donc décroissante dans la distance  $|x_0 - x_i|$ .
3. Pour tout  $x$ , les poids  $W(x_i)$  satisfont à la relation suivante :

$$\sum_{i=1}^n W(x_i) = 1.$$

## 1.3 Biais et variance de l'estimateur

### Biais de l'estimateur

Le traitement du biais est purement analytique et repose essentiellement sur le développement de Taylor. Il nous faut supposer certaines conditions de régularités sur les fonctions  $g(\cdot)$  et  $f(\cdot)$  qui détermineront l'ordre du biais asymptotique en fonction du paramètre de lissage  $h$ . [1]

**Propriété 1.1.** *Supposons que  $g(\cdot)$  et  $f(\cdot)$  sont de classe  $C^2(\mathbb{R})$  et que le noyau  $K$  est d'ordre 2, i.e, tel que : les conditions 2,4,6 de noyau  $K$  sont vérifiées. Nous avons alors, lorsque  $h \rightarrow 0$  et  $nh \rightarrow \infty$*

$$\begin{aligned} \text{Biais}(\hat{r}(x)) &= \mathbb{E}[\hat{r}(x) - r(x)] \\ &= \frac{h^2}{2} \left( r''(x) + 2r'(x) \frac{f'(x)}{f(x)} \right) \int u^2 K(u) du + o(h^2) \end{aligned}$$

### Démonstration.

$$\begin{aligned} \text{Biais}(\hat{r}(x)) &= \mathbb{E}(\hat{r}(x)) - r(x) \\ &= [\mathbb{E}(K\left(\frac{x - X}{h}\right))]^{-1} \left[ \int \frac{1}{h} K\left(\frac{x - t}{h}\right) g(t) dt - g(x) + g(x) - r(x) \int \frac{1}{h} K\left(\frac{x - t}{h}\right) f(t) dt \right] \\ &\simeq \frac{1}{h^2} [f(x)]^{-1} [g''(x) - r(x) f''(x)] \int_{\mathbb{R}} u^2 K(u) du \\ &= \frac{1}{h^2} \left[ r''(x) + 2r'(x) \frac{f'(x)}{f(x)} \right] \int_{\mathbb{R}} u^2 K(u) du \end{aligned}$$



### Variance de l'estimateur

Les hypothèses (1),(3) de noyau sont supposés vérifier le fait que  $K(\cdot)$  soit de carré intégrable. Nous posons, par convenance que :

$$\sigma^2(x) = \text{Var}[Y|X = x],$$

lorsque cette expression est bien définie.

**Propriété 1.2.** *On suppose que  $\mathbb{E}[Y^2] < \infty$ . À chaque point de continuité des fonctions  $g(x)$ ,  $f(x)$  et  $\sigma^2(x)$ , tel que  $f(x) > 0$ ,*

$$\begin{aligned} \text{Var}(\hat{r}(x)) &= \mathbb{E}[(\hat{r}(x) - \mathbb{E}[\hat{r}(x)])^2] \\ &= \frac{1}{nh} \frac{\sigma^2(x)}{f(x)} \int K^2(u) du + o\left(\frac{1}{h}\right) \end{aligned}$$

**Démonstration.** En utilisant le lemme de Bochner on obtenons aisément :

$$\begin{aligned} \text{Var}[\hat{f}(x)] &= \frac{1}{nh^2} \{ \mathbb{E}[K^2(\frac{x-X}{h})] - \mathbb{E}[K(\frac{x-X}{h})]^2 \} \\ &= \frac{1}{nh} \{ \int_{\mathbb{R}} K^2(u) f(x-hu) du - h \{ \int_{\mathbb{R}} K(u) f(x-hu) du \}^2 \} \\ &= \frac{1}{nh} f(x) \int_{\mathbb{R}} K^2 du (1 + o(1)) \end{aligned}$$

Lorsque  $h \rightarrow 0$ . Soit la fonction  $s(x) = \int_{\mathbb{R}} y^2 f(x, y) dy$ . nous avons :

$$\begin{aligned} \text{Var}[\hat{g}_n(x)] &= \frac{1}{nh^2} \{ \mathbb{E}[Y^2 K^2(\frac{x-X}{h})] - \mathbb{E}[Y K(\frac{x-X}{h})]^2 \} \\ &= \frac{1}{nh} \{ \int_{\mathbb{R}} K^2(u) s(x-hu) du - h \{ \int_{\mathbb{R}} K(u) r(x-hu) du \}^2 \} \\ &= \frac{1}{nh} s(x) \int_{\mathbb{R}} K^2 du (1 + o(1)) \end{aligned}$$

De même,

$$\mathbb{E}[\{\hat{f}(x) - \mathbb{E}[\hat{f}(x)]\}\{\hat{g}(x) - \mathbb{E}[\hat{g}(x)]\}] = \frac{1}{nh} g(x) \int_{\mathbb{R}} K(u)^2(u) du (1 + o(1))$$

Soit le vecteur

$$A(x) = \begin{pmatrix} \hat{f}(x) \\ \hat{g}(x) \end{pmatrix}$$

et  $\Sigma[A(x)]$  sa matrice de variance covariance. Il s'ensuit :

$$\Sigma[A_n(x)] = \frac{1}{nh} \begin{pmatrix} f(x) & g(x) \\ g(x) & s(x) \end{pmatrix} \int_{\mathbb{R}} K^2(u) du (1 + o(1))$$

En remarquant que,

$$\begin{pmatrix} -\frac{g(x)}{\{f(x)\}^2} & \frac{1}{f(x)} \end{pmatrix} \begin{pmatrix} f(x) & g(x) \\ g(x) & s(x) \end{pmatrix} \begin{pmatrix} -\frac{g(x)}{\{f(x)\}^2} \\ 1 \\ \frac{1}{f(x)} \end{pmatrix} = \frac{s(x)}{\{f(x)\}^2} - \frac{\{g(x)\}^2}{\{f(x)\}^3}$$

On obtient alors,

$$\begin{aligned} \text{Var}(\hat{r}(x)) &= \frac{1}{nh} \left[ \frac{s(x)}{\{f(x)\}^2 - \frac{\{r(x)\}^2}{\{f(x)\}^3}} \right] \int_{\mathbb{R}} K^2(u) du (1 + o(1)) \\ &= \frac{1}{nh} \left[ \frac{\sigma^2(x)}{\{f(x)\}} \int_{\mathbb{R}} K^2(u) du \right] (1 + o(1)) \end{aligned}$$

■

**Remarque 1.3.**  $o(1)$  dans la démonstration précédente vérifiée l'égalité suivante :

$$o(1) = o(h) + o((nh)^{-1})$$

## 1.4 Consistance

La consistance est la propriété la plus importante d'un estimateur. On distingue une consistance forte et aussi une consistance faible. L'estimateur à noyau de la fonction de régression dépend de deux paramètres : la fenêtre  $h$  et le noyau  $K$ . Le noyau  $K$  établit l'aspect du voisinage de  $x$  et  $h$  contrôle la taille de ce voisinage, donc  $h$  est le paramètre prédominant pour avoir de bonnes propriétés asymptotiques, néanmoins le noyau  $K$  ne doit pas être négligé en se basant alors sur l'étude du biais, et de la variance.

### 1.4.1 Consistance faible

**Théorème 1.1.** *Supposons que le noyau  $K$  vérifie les hypothèses (1,6), que  $\mathbb{E}(Y^2) < \infty$  et que :  $f(x)$  est strictement positive. Si  $h \rightarrow 0$ ,  $nh \rightarrow +\infty$  (quand  $n \rightarrow \infty$ ), alors  $\hat{r}(x)$  est un estimateur consistant de  $r(x)$  [3].*

**Démonstration.** Nous déduisons du théorème de Bochner (Voir Annexe B) que, lorsque  $h \rightarrow 0$

$$\begin{aligned}\mathbb{E}[\hat{f}(x)] &= \mathbb{E}\left[\frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - x_i}{h}\right)\right] \\ &= \frac{1}{h} \mathbb{E}\left[K\left(\frac{x - X}{h}\right)\right] \\ &= \frac{1}{h} \int_{\mathbb{R}} K\left(\frac{x - t}{h}\right) f(t) dt \longrightarrow f(x) \int_{\mathbb{R}} K(t) dt \\ &= f(x).\end{aligned}$$

Nous constatons que le biais de l'estimateur converge vers zéro quand la fenêtre tend vers zéro, de plus, on constate qu'il ne dépend pas du nombre de variables, il dépend surtout du noyau  $K$ . Donc  $\hat{f}$  est un estimateur asymptotiquement sans biais. D'autre part, comme les  $x_i$  sont indépendantes et identiquement distribuées, il vient que :

$$\begin{aligned}\text{Var}(\hat{f}(x)) &= \frac{1}{n} \text{Var}\left[\frac{1}{h} K\left(\frac{x - X}{h}\right)\right] \\ &\leq \frac{1}{n} \mathbb{E}\left[\frac{1}{h} K\left(\frac{x - X}{h}\right)\right]^2 \\ &= \frac{1}{h} \int_{-\infty}^{+\infty} \frac{1}{h} K^2\left(\frac{x - t}{h}\right) f_X(t) dt\end{aligned}$$

D'après le théorème de Bochner :

$$\int_{-\infty}^{+\infty} \frac{1}{h} K^2\left(\frac{x-t}{h}\right) f(t) dt \longrightarrow f(t) \int_{\mathbb{R}} K^2(t) dt < \infty, \text{ quand } n \rightarrow \infty$$

■

donc

$$\text{Var}(\hat{f}(x)) \rightarrow 0 \text{ quand } nh \rightarrow \infty$$

Donc la variance est convergente puisque  $\hat{f}(x)$  est un estimateur consistant de  $f(x)$ .

Puisque  $\hat{f}(x)$  est un estimateur consistant de  $f(x)$ , il suffit donc de montrer que  $\hat{g}(x)$  est un estimateur consistant de

$$g(x) = \int_{-\infty}^{+\infty} y f_{x,Y}(x, y) dy$$

Nous avons

$$\begin{aligned}
\mathbb{E}[\hat{g}(x)] &= \mathbb{E}\left[\frac{1}{nh} \sum_{i=1}^n Y_i K\left(\frac{x - x_i}{h}\right)\right] \\
&= \frac{1}{h} \mathbb{E}\left[Y K\left(\frac{x - X}{h}\right)\right] \\
&= \frac{1}{h} \int_{\mathbb{R}} \mathbb{E}(Y/X = x) K\left(\frac{x - t}{h}\right) f(t) dt \\
&= \frac{1}{h} \int_{\mathbb{R}} r(t) K\left(\frac{x - t}{h}\right) f(t) dt \\
&= \frac{1}{h} \int_{\mathbb{R}} K\left(\frac{x - t}{h}\right) r(t) f(t) dt \\
&= \frac{1}{h} \int_{\mathbb{R}} K\left(\frac{x - t}{h}\right) g(t) dt \longrightarrow g(x)
\end{aligned}$$

par le théorème de Bochner De plus, si  $s(x) = \int y^2 f_{x,y}(x, y) dy$

$$\mathbb{V}ar(\hat{g}(x)) = \mathbb{E}[\hat{g}(x)]^2 - \mathbb{E}^2[\hat{g}(x)] \simeq \frac{1}{h} s(x) \int_{\mathbb{R}} K^2(t) dt \longrightarrow 0 \text{ quand } n \longrightarrow \infty$$

donc

$$\lim_{n \rightarrow \infty} \mathbb{V}ar(\hat{g}(x)) = 0$$

d'où  $\hat{r}(x) = \frac{\hat{g}(x)}{\hat{f}(x)} \longrightarrow \frac{g(x)}{f(x)} = r(x)$  en probabilité

### Absence de biais asymptotique

**Théorème 1.2.** *Sous les conditions (1,6) et si  $f(x)$  est strictement positive, il vient que :*

a- *Lorsque  $Y$  est bornée, p.s,  $h \longrightarrow 0$  et  $nh \longrightarrow \infty$  (quand  $n \longrightarrow \infty$ ) alors*

$$\mathbb{E}(\hat{r}(x)) = \mathbb{E}[\hat{g}(x)]/\mathbb{E}[\hat{f}(x)] + o((nh)^{-1}) \quad (1.9)$$

b- *Lorsque  $\mathbb{E}[Y^2] < \infty$ ,  $h \longrightarrow \infty$  et  $nh^2 \longrightarrow \infty$  (quand  $n \longrightarrow \infty$ ) alors :*

$$\mathbb{E}(\hat{r}(x)) = \mathbb{E}[\hat{g}(x)]/\mathbb{E}[\hat{f}(x)] + o((n^{\frac{1}{2}}h)^{-1}) \quad (1.10)$$

(a), (b) et le théorème de Bochner impliquent que  $\hat{r}(x)$  est un estimateur asymptotiquement sans biais de  $r(x)$

**Démonstration.** En utilisant l'identité suivante

$$\frac{1}{\hat{f}(x)} = \frac{1}{\mathbb{E}[\hat{f}(x)]} - \frac{\hat{f}(x) - \mathbb{E}[\hat{f}(x)]}{\{\mathbb{E}[\hat{f}(x)]\}^2} + \frac{\{\hat{f}(x) - \mathbb{E}[\hat{f}(x)]\}^2}{\hat{f}(x)\{\mathbb{E}[\hat{f}(x)]\}^2}$$

On multiplie par  $\hat{g}(x)$  des deux côtés, puis on passe à l'espérance

$$\begin{aligned} \mathbb{E}[\hat{r}(x)] &= \frac{\mathbb{E}[\hat{g}(x)]}{\mathbb{E}[\hat{f}(x)]} - \mathbb{E}\left[\hat{g}(x) \frac{\hat{f}(x) - \mathbb{E}[\hat{f}(x)]}{\{\mathbb{E}[\hat{f}(x)]\}^2}\right] \\ &\quad + \mathbb{E}\left[\hat{g}(x) \frac{\{\hat{f}(x) - \mathbb{E}[\hat{f}(x)]\}^2}{\hat{f}(x)\{\mathbb{E}[\hat{f}(x)]\}^2}\right] \\ &= \frac{\mathbb{E}[\hat{g}(x)]}{\mathbb{E}[\hat{f}(x)]} - \frac{a_n(x) + b_n(x)}{\{E[\hat{f}(x)]\}^2}, \end{aligned}$$

tel que

$$a_n(x) = [\mathbb{E}(\hat{g}(x)) - \mathbb{E}(\hat{g}(x))] - [\hat{f}(x) - \mathbb{E}(\hat{f}(x))];$$

et

$$b_n(x) = \mathbb{E}[(\hat{f}(x))^{-1} \hat{g}_n(x) (\hat{f}(x) - \mathbb{E}(\hat{f}(x)))^2].$$

Soit  $s(x) = \int_{\mathbb{R}} y^2 f_{x,Y}(x, y) dy$ . Nous calculons la variance asymptotique de  $\hat{g}(x)$  puis  $\hat{f}(x)$  via le théorème de Bochner

$$\begin{aligned} \text{Var}[\hat{g}(x)] &= \frac{1}{nh} \int_{\mathbb{R}} K^2(u) s(x - uh) du - \frac{1}{n} \left\{ \int_{\mathbb{R}} K(u) g(x - uh) du \right\}^2 \\ &\approx \frac{1}{nh} s(x) \int_{\mathbb{R}} K^2 du \end{aligned}$$

$$\begin{aligned} \text{Var}[\hat{f}(x)] &= \frac{1}{nh} \int_{\mathbb{R}} K^2(u) f(x - uh) du - \frac{1}{n} \left\{ \int_{\mathbb{R}} K(u) f(x - uh) du \right\}^2 \\ &\approx \frac{1}{nh} f(x) \int_{\mathbb{R}} K^2(u) du \end{aligned}$$

En utilisant l'inégalité de Cauchy-Schwartz (Voir Annexe B) combinée aux formules ci-dessus, on obtient

$$a_n(x) = o((nh)^{-1}) \tag{1.11}$$

Lorsque la variable Y est bornée, i.e  $|y| \leq M$  pour une certaine constante M fixée, nous remarquons que l'estimateur de [N-W] est lui aussi naturellement borné,

$$\frac{\hat{g}(x)}{\hat{f}(x)} = \frac{\sum_1^n y_i K\left(\frac{x-x_i}{h}\right)}{\sum_1^n K\left(\frac{x-x_i}{h}\right)} \leq \frac{\sum_1^n MK\left(\frac{x-x_i}{h}\right)}{\sum_1^n K\left(\frac{x-x_i}{h}\right)} = M \tag{1.12}$$

Cette dernière inégalité (1.12) permet de borner  $b_n(x)$

$$b_n(x) \leq M \times \mathbb{E}[\{\hat{f}(x) - \mathbb{E}[\hat{f}(x)]\}^2] \approx \frac{M}{Mh} f(x) \int_{\mathbb{R}} K^2(u) du = o((nh)^{-1})$$



Les relation (1.11) et (1.13) entraînent (1.9). Pour démontrer le cas b) ; il suffit de remarquer que la relation (1.11) est toujours valable mais la relation (1.12) devient

$$b_n(x) \leq M \times \mathbb{E}[\hat{f}(x) - E[\hat{f}(x)]]^2 \approx \frac{M}{Mh} f(x) \int_{\mathbb{R}} K^2(u) du = o((nh)^{-1}) \quad (1.13)$$

$$\begin{aligned} |b_n(x)| &\leq \mathbb{E}[\max_{1 \leq i \leq n} |Y_i| \{ \hat{f}(x) - E[\hat{f}(x)] \}^2] \\ &\leq \left\{ \sum_1^n Y_i^2 \right\}^{1/2} \times \{ \mathbb{E}[\{ \hat{f}(x) - E[\hat{f}(x)] \}^4] \}^{1/2} \\ &= \sqrt{n} \{ \mathbb{E}[Y_i^2] \}^{1/2} \times o((nh)^{-1}) \\ &= o((n^{1/2}h)^{-1}) \end{aligned} \quad (1.14)$$

Les relations (1.14) et (1.11) impliquent (1.10), d'où le résultat énoncé ■

#### Remarque 1.4.

- 1- Dans l'estimation de  $r(x)$  par le noyau rectangulaire , le même poids est accordé à toutes les observations comprise entre  $x - h$  et  $x + h$  . Dans les autres noyaux , le poids d'une observation est d'autant plus fort qu'elle est proche de  $x$ .
- 2-  $\hat{r}$  a les mêmes propriétés de continuité et de différentiabilité que  $K$ . Par exemple, si  $K$  est le noyau gaussien  $\hat{f}$  admet des dérivées de tous ordres
- 3- Pour choisir quel noyau prendre et surtout choisir le paramètre de lissage  $h$  , il faut étudier la qualité de l'estimation de  $r$  par  $\hat{r}$

#### Exemple 1.2. [16]

Dans le cas d'une fonction noyau Epanechnikov, on a :

$$K(u) = \begin{cases} \frac{3}{4}(1 - u^2), & |u| \leq 1 \\ 0, & |u| > 1 \end{cases}$$

avec  $u = (x_i - x_0)/h$ . Donc si  $|x_i - x_0| > h$ , alors  $u \in ] - \infty, -1 [ \cup ] 1, +\infty [$ ,  $K(u) = 0$  et par conséquent  $W_i(x_0) = 0$ .

Les vitesses de convergence dépendent de deux paramètres : la fonction de noyau  $K$  dont l'efficacité est peu influente et le paramètre de lissage  $h$ , dont le choix est crucial aussi bien pour l'approche ponctuelle que pour la globale que nous exposons ci après.

1.  $f(\cdot)$  et  $r(\cdot)$  de classe  $C^2(\mathbb{R})$ ,
2.  $\lim_{n \rightarrow \infty} h = 0$ ,

3.  $|Y|$  est bornée,
4.  $K$  satisfait les propriétés (1,4,6,7) suivantes :

### 1.4.2 Consistance forte

Parzen a établi la normalité asymptotique, ainsi que la convergence uniforme en probabilité. Son travail est un outil important et a été largement développé par plusieurs chercheurs ( Devroye et Györfi (1985)[10], Silverman (1986)[25] ). En (1976) Nadaraya [22] a énoncé le théorème de la consistance forte de l'estimateur.

L'étude de la convergence, presque complète des estimateurs à noyau introduite par Ferraty et Vieu (2000) [13], s'inspire énormément du cas réel exposé ci dessous. Cette convergence implique la convergence en probabilité et la convergence presque sûre mais n'implique pas la convergence en moyenne quadratique. (Voir annexe 1)

#### Convergence presque complète

En se basant sur la preuve donnée dans Ferraty et Vieu (2003) [14], nous traitons dans ce paragraphe la convergence presque complète de l'estimateur à noyau de la fonction de régression. Nous gardons quelques conditions précédentes, et nous introduisons des hypothèses de base permettant de donner un théorème général sur la convergence presque complète.

1.  $f(x), r$  sont des fonctions continues au voisinage de  $X$ , un point fixé de  $\mathbb{R}$ .
2. La densité  $f(x)$  et la variable  $Y$  sont telles que

$$f(x) > 0$$

et

$$|Y| < M < \infty$$

où  $M$  est une constante réelle positive.

3. Le paramètre de lissage  $h$  est tel que

$$\lim_{n \rightarrow \infty} h = 0 \text{ et } \lim_{n \rightarrow \infty} \frac{\log n}{nh} = 0,$$

**Théorème 1.3.** [5] *Sous les hypothèses (1),(2),(3) de plus (H.4) et le fait que  $K$  est borné, intégrable et à support compact".*

$$\lim_{n \rightarrow \infty} \hat{r}(x) = r(x). \text{ p.co}$$

## 1.5 Convergence asymptotique en loi et intervalles de confiance

Pour construire un intervalle de confiance, i.e, *IC* sur  $r(x_i)$ , on admet le résultat suivant :

**Théorème 1.4.** [16] *L'estimateur à noyau de Nadaraya-Watson vérifie :*

$$\sqrt{h}\sqrt{n}\{\hat{r}(x_0) - r(x_0) - \frac{1}{2}a_k h^2[r''(x_0) + 2r'(x_0)\frac{f'(x_0)}{f(x_0)}]\} \longrightarrow \mathcal{N}(0, \frac{b_k \sigma^2}{f(x_0)})$$

où  $f(\cdot)$  désigne la densité de  $x$  et

$$a_k = \int u^2 K(u) du \quad b_k = \int k^2(u) du$$

Voir Wand et Jones (1995) pour les valeurs de  $a_K$  et  $b_K$  pour de nombreux noyaux. On admettra en particulier que :

Noyau	$b_k$
Uniforme	1/2
Triangulaire	2/3
Gaussien	1/2 $\pi$
Biweight	5/7
TRWeight	350/429
Epanechnikov	3/5

Des simplifications peuvent être apportées si la valeur de  $h$  décroît avec  $n$  plus rapidement que  $h = n^{-1/5}$ . Dans ce cas, le terme de biais disparaît et donc on obtient le résultat suivant dans le cas d'un noyau uniforme  $b_k = 1/2$

$$\sqrt{h}\sqrt{n}[\hat{r}(x_0) - r(x_0)] \longrightarrow \mathcal{N}(0, \frac{\sigma^2}{2f(x_0)})$$

peut donc en déduire la manière de construire des *IC* sur les valeurs de  $r(x_i)$ .

**Théorème 1.5.** [24] *Sous l'hypothèse que  $h$  vérifie  $hn^{-\frac{1}{5}} \rightarrow 0$ , l'écart type de l'estimateur à noyau de Nadaraya-Watson  $\hat{r}(x_0)$  vérifie :*

$$\sqrt{h}\sqrt{n}[\hat{r}(x_0) - r(x_0)] \longrightarrow \mathcal{N}(0, \frac{b_k \sigma^2}{f(x_0)})$$

$f(\cdot)$  est la densité de  $x$ . dans un intervalle de confiance sur  $\hat{r}(x_0)$  au seuil de  $\alpha\%$  donc on définit par :

$$IC_\alpha = [\hat{r}(x_0) - Z_{1-\alpha/2} s_{\hat{r}(x_0)}, \hat{r}(x_0) + Z_{1+\alpha/2} s_{\hat{r}(x_0)}]$$

où  $Z_{1-\alpha/2}$  désigne le fractile de la loi  $\mathcal{N}(0;1)$  ou :

$$s_{\hat{r}(x_0)} = \sqrt{\frac{b_k \sigma^2}{\hat{f}(x_0)}}$$

avec

$$\hat{f}(x_0) = K\left(\frac{x_i - x_0}{h}\right)$$

La procédure pour obtenir les  $IC$  est donc la suivante :

1. On choisit  $h$  vérifie  $hn^{1/5} \rightarrow 0$  et une fonction noyau  $K(u)$  ; d'où l'on déduit  $b_k$
2. On construit l'estimateur à noyau de Nadaraya-Watson  $\hat{r}(x_0)$ . On recommence pour toutes les valeurs  $x_1, \dots, x_n$ .
3. On calcule l'estimateur de la variance des résidus

$$\sigma^2 = 1/n[y_i - \hat{r}(x_i)]^2$$

4. On estime la valeur de  $f(x_0)$ , densité de  $X$  au point  $x_0$  par :

$$\hat{f}(x_0) = K\left(\frac{x_i - x_0}{h}\right)$$

5. On calcule  $IC$  sur  $\hat{f}(x_0)$  au seuil défini par :

$$IC_\alpha = [\hat{r}(x_0) - Z_{1-\alpha/2} \sqrt{\frac{b_k \sigma^2}{\hat{f}(x_0)}}, \hat{r}(x_0) + Z_{1-\alpha/2} \sqrt{\frac{b_k \sigma^2}{\hat{f}(x_0)}}]$$

Il existe des procédures numériques de choix d'un  $h$  optimal. La première méthode consiste à choisir  $h$  de sorte à minimiser la MISE (Mean Integrated Squared Error). C'est la définition même du paramètre de lissage optimale.

**Exemple 1.3.** En choisissant  $h = n^{-1/5}$  et  $K(u) = 1_{[-1/2 < (u) < 1/2]}$ , nous étudions, à travers quelques simulations les propriétés des intervalles de confiance de la fonction de densité et de la fonction de régression pour les petits échantillons. Nous générons pour chacune des applications que nous proposons des échantillons de taille  $n=10$ ,  $n=50$ ,  $n=100$ ,  $n=800$ , respectivement, et nous considérons les modèles suivants.

a/-  $Y = X^2 + Z$  où  $X \sim \mathcal{N}(0,1)$  et  $Z \sim \mathcal{N}(0,2)$

b/-  $Y = \sin(\frac{3}{2}X) + Z$  où  $X \sim \mathcal{N}(0,1)$  et  $Z \sim \mathcal{N}(0,2)$

c/-  $Y = \exp(X - 0.2) + Z$  où  $X \sim \mathcal{N}(0, 1)$  et  $Z \sim \mathcal{N}(0, 2)$

Les figures (1.2)\_(1.3)\_(1.4) montrent que la probabilité de trouver les fonctions à estimer dans l'intervalle de confiance  $\sim 95\%$ .

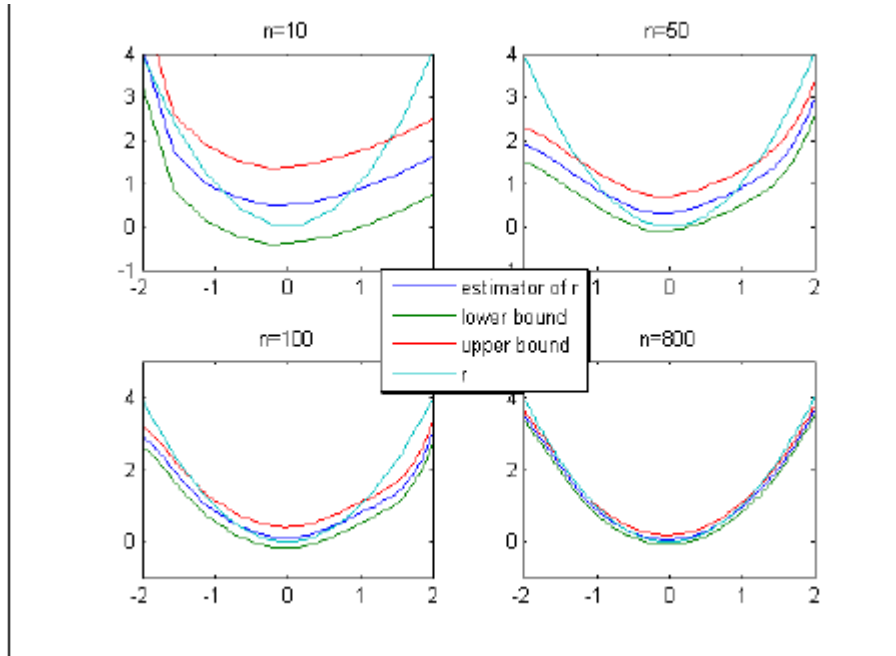


FIGURE 1.2: Les intervalles de confiance de la fonction de régression dans le cas  $a$

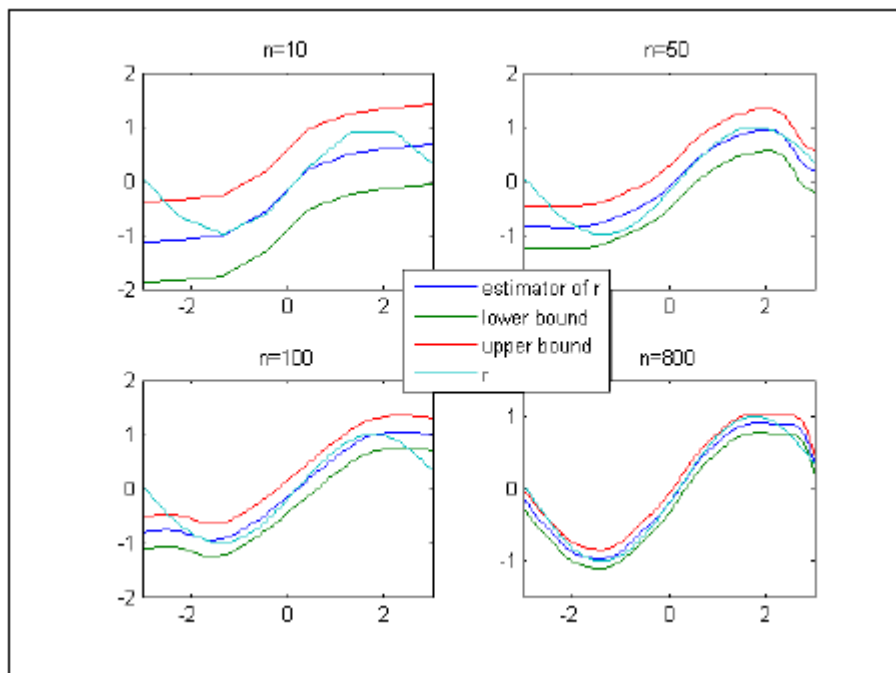


FIGURE 1.3: Les intervalles de confiance de la fonction de régression dans le cas  $b$

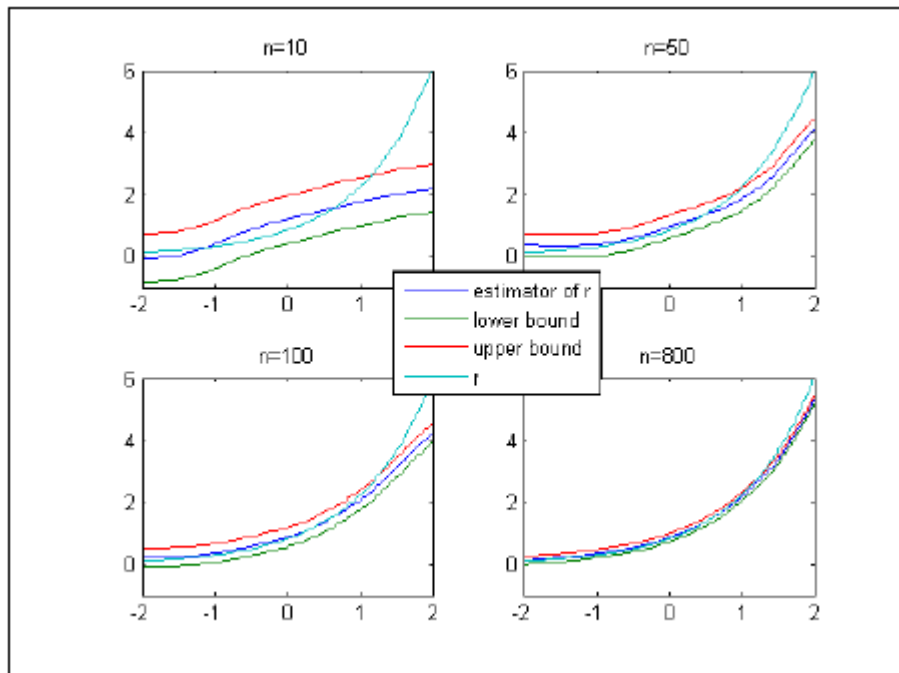


FIGURE 1.4: Les intervalles de confiance de la fonction de régression dans le cas c

*D'après les figures (1.2), (1.3) et (1.4) on remarque que l'estimateur de Nadarya Watson de la fonction de régression  $r$  est toujours convergent lorsque  $n$  est grand et l'intervalle de confiance s'étroie sur la vraie fonction  $r(x)$ .*

---

---

# CHAPITRE 2

---

## CHOIX DU PARAMÈTRE DE LISSAGE

Toute la question est comment choisir une valeur optimale du paramètre de lissage permettant d'arbitrer au mieux entre variance et biais. Il existe des procédures numériques de choix d'un  $h$  optimal. La première méthode consiste à choisir  $h$  de sorte à minimiser la *MISE*, c'est la définition même du paramètre de lissage optimal. Alors, on donnera l'estimation de  $h$  par le critère de la MSE et MISE, puis, on termine par une méthode numérique qui est choisit la validation croisé.

### 2.1 Etude du critère d'erreur quadratique moyenne de $\hat{r}(x)$

L'erreur quadratique moyenne *MSE* (mean square error) est une mesure permettant d'évaluer la similarité de  $\hat{r}(x)$  par rapport à la fonction de régression inconnue  $r(x)$ , au point  $x$  donné de  $\mathbb{R}$  [5]. Notre but est de minimiser

$$MSE(\hat{r}(x)) = \mathbb{E}(\hat{r}(x) - r(x))^2$$

Le développement de cette expression faite précédemment, nous donne

$$MSE(\hat{r}(x)) = \text{Biais}(\hat{r}(x))^2 + \text{Var}(\hat{r}(x))$$

Nous constatons d'une part que les expressions du biais de  $\hat{r}(x)$  et de la variance de  $\hat{r}(x)$  données dans le chapitre (1), permettent de conclure qu'une grande valeur de  $h$  donne une augmentation du biais et une diminution de la variance (estimation fortement biaisée) et qu'un faible paramètre  $h$ , donne une diminution du biais et une augmentation de la variance (phénomène de sous lissage). Dans ce sens nous allons donner l'exemple suivant :

**Exemple 2.1.** *L'importance du choix  $h$*  Le paramètre de lissage  $h$ , est un réel positif dont le choix est prépondérant sur celui du noyau continu symétrique  $K$ . Le choix d'une valeur de  $h$  trop grande conduit à une courbe trop lisse. La courbe estimée ne traduit pas suffisamment les variations de la vraie distribution (voir Figure 2.1).

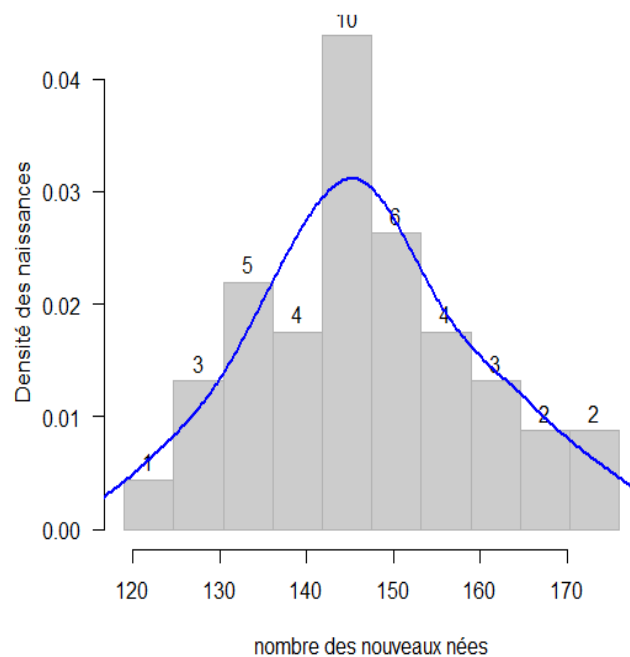


FIGURE 2.1: Illustration d'un phénomène de sous-lissage lors de l'estimation d'une densité où le paramètre  $h$  est très grand.

*Par contre, en choisissant un paramètre de lissage très petit que celui adopté précédemment, l'allure de la distribution change. Il s'agit d'une distribution surestimé (voir Figure 2.2).*



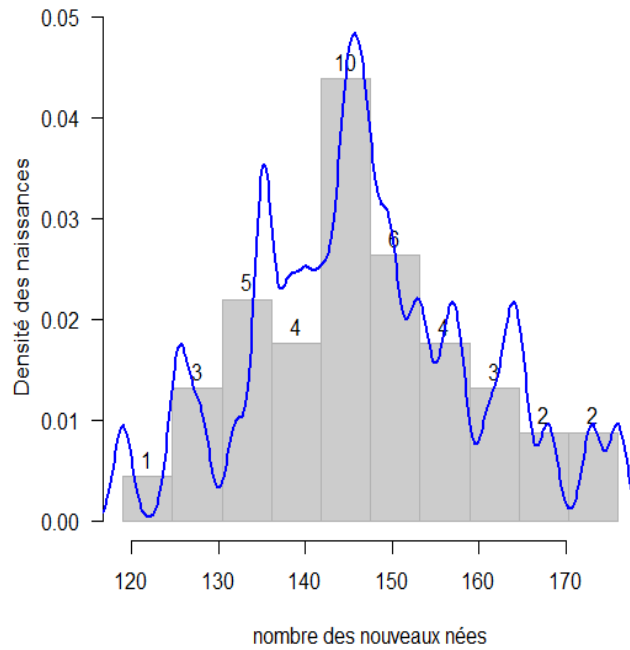


FIGURE 2.2: Illustration d'un phénomène de sur-lissage lors de l'estimation d'une densité où le paramètre  $h$  est très petit.

*L'estimation de densité nécessite également le choix adéquat de la fenêtre  $h$ , et pour cette valeur idéale du paramètre de lissage, nous obtenons une allure qui suit parfaitement la distribution de départ (Voir figure 2.3).*

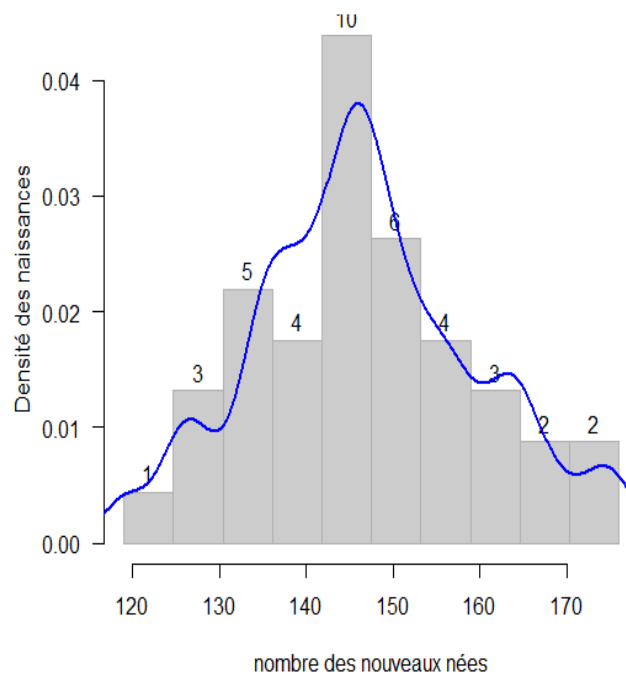


FIGURE 2.3: Illustration d'une estimation idéale  $h$  optimale.

Les courbes obtenues illustrent à quel point les formes estimées sont différentes en fonction de l'ordre de grandeur du paramètre de lissage.

D'autre part, sous les hypothèses de ces mêmes propositions, nous obtenons :

$$MSE(\hat{r}(x)) = \frac{h^4}{4} \left\{ [r''(x) + 2r'(x) \frac{f'(x)}{f(x)}] [u^2 K(u)] + o(1) \right\}^2 + \frac{1}{nh} \left\{ \frac{\sigma^2(x)}{f(x)} [K^2(u)] \right\} (1 + o(1)),$$

où  $[u^p K^q(u)] = \int u^p K^q(u) du$ .

Pour un compromis entre le biais et la variance nous minimisons par rapport à  $h$  l'expression de l'erreur quadratique moyenne asymptotique  $AMSE$  (asymptotique mean squared error) donnée par :

$$AMSE(\hat{r}(x)) = \frac{h^4}{4} \left\{ r''(x) + 2r'(x) \frac{f'(x)}{f(x)} \right\}^2 [u^2 K(u)]^2 + \frac{1}{nh} \frac{\sigma^2(x)}{f(x)} [K^2(u)]$$

. Comme  $AMSE$  est une fonction convexe. La fenêtre  $h_{opt}(\hat{r}(x)) = \arg \min_h [AMSE(\hat{r}(x))]$  est une solution de l'équation suivante

$$\frac{\partial}{\partial h} \left[ \frac{h^4}{4} \left\{ r''(x) + 2r'(x) \frac{f'(x)}{f(x)} \right\}^2 [u^2 K(u)]^2 + \frac{1}{nh} \frac{\sigma^2(x)}{f(x)} [K^2(u)] \right] = 0$$

Lorsque  $[r''(x) + 2r'(x) \frac{f'(x)}{f(x)}]^2 [u^2 K(u)]^2 \neq 0$ , on a :

$$h_{opt} = n^{-\frac{1}{5}} \left( \frac{\frac{\sigma^2(x)}{f(x)} [K^2(u)]}{\left\{ r''(x) + 2r'(x) \frac{f'(x)}{f(x)} \right\}^2 [u^2 K(u)]^2} \right)^{\frac{1}{5}}$$

**Démonstration.** On pose :

$$B^2 = [r''(x) + 2r'(x) \frac{f'(x)}{f(x)}]^2 [u^2 K(u)]^2, \text{ et } V = \frac{\sigma^2(x)}{f(x)} [K^2(u)].$$

Donc :

$$AMSE = \frac{1}{4} B^2 h^4 + \frac{V}{nh}.$$

On calcule la dérivée de l' $AMSE$  :

$$\frac{\partial}{\partial h} \left[ \frac{1}{4} B^2 h^4 + \frac{V}{nh} \right] = B^2 h^3 - \frac{V}{nh^2}.$$

Il est clair que la deuxième dérivée de l' $AMSE$  est positive, alors on a une certitude qu'elle admet un minimum :

$$\begin{aligned} B^2 h^3 - \frac{V}{nh^2} &= B^2 h^5 - \frac{V}{n} \\ &= 0, \end{aligned}$$

on trouve :

$$\begin{aligned} h_{opt} &= \left( \frac{V}{nB^2} \right)^{\frac{1}{5}} \\ &= \left( \frac{\frac{\sigma^2(x)}{f(x)} [K^2(u)]}{n \{ r''(x) + 2r'(x) \frac{f'(x)}{f(x)} \}^2 [u^2 K(u)]^2} \right)^{\frac{1}{5}} \end{aligned}$$

■

On s'intéresse maintenant à l'approche globale pour la sélection du paramètre  $h$ , pour cela on introduit le critère d'erreur quadratique moyenne intégrée ou *MISE* (mean integrated squared error) de  $\hat{r}(x)$ .

$$MISE(\hat{r}(x)) = \mathbb{E} \int_{\mathbb{R}} [(\hat{r}(x) - r(x))^2 dx]$$

En appliquant le Théorème de Fubini (Voir Annexe.B), on a :

$$MISE(\hat{r}(x)) = \int_{\mathbb{R}} \mathbb{E}[(\hat{r}(x) - r(x))^2 dx]$$

**Définition 2.1.** Pour choisir le paramètre de lissage  $h$ , on choisit la méthode de *AMISE* (Approximate Mean Integrated Squared Error), l'*AMISE* associé à un paramètre de lissage  $h$ , correspond à la quantité :

$$AMISE(h) = \frac{1}{4} h^4 [u^2 K(u)]^2 \left[ \int r''(x)^2 dx \right] + \frac{1}{nh} [K^2(u)]$$

ou de façon équivalent avec nos notations ;

$$AMISE(h) = \frac{1}{4} h^4 a_K^2 \int_x f''(x)^2 dx + \frac{b_K}{nh}$$

ou par l'équation suivante :[5]

$$AMISE(\hat{r}(x)) = \frac{h^4}{4} \int \{ r''(x) + 2r'(x) \frac{f'(x)}{f(x)} \}^2 dx [u^2 k(u)]^2 + \frac{1}{nh} \int \frac{\sigma^2(x)}{f(x)} dx [K^2(u)]$$

La fenêtre  $h_{opt(\hat{r}(x))}^{MISE}$  minimisant l'*AMISE* du critère global est :

$$h_{opt(\hat{r}(x))}^{MISE} = n^{-\frac{1}{5}} \left[ \frac{\int \frac{\sigma^2(x)}{f(x)} dx [K^2(u)]}{\int \{ r''(x) + 2r'(x) \frac{f'(x)}{f(x)} \}^2 dx [u^2 K^2(u)]} \right]^{\frac{1}{5}}$$

**Exemple 2.2** (Choix optimal du paramètre  $K$  et  $h$ ). Dans ce que suit, on va estimer deux fonctions de régression, en utilisant l'estimateur de Nadarya Watson :

1- **Cas non-linéaire** : On définit le modèle non linéaire suivant :

$$y = \sin(\exp(x)) + \epsilon$$

avec :  $\epsilon \sim \mathcal{N}(0, 1)$  et  $x \sim \mathcal{N}(0, 0.6)$ . Dans les résultats graphiques, on obtient :

1. La courbe de la fonction de régression  $r(x)$  qu'on cherche à estimer (la droite noire).
2. La courbe de l'estimateur de la fonction de régression  $\hat{r}(x)$  (la droite rouge).

On va chercher le choix optimale des paramètres  $K$  et  $h$  selon l'estimateur de [N-W]

i- Choix du paramètre  $K$  : Le graphe ci dessous représente la fonction  $r(x)$  et l'estimateur  $\hat{r}(x)$ . On choisit le paramètre de lissage  $h = n^{1/5}$  (fixé),  $n$  varié ( $n = 40, 200, 400$ ), et  $K$  noyau gaussien (normale)  $K_5(t) = (\frac{1}{\sqrt{2\pi}}) \exp(-\frac{t^2}{2})$ . En utilisant le code R suivant :

```
rm(list=ls(all=TRUE))
n=40
X=rnorm(n)
E=rnorm(n)
Y=sin(exp(X))+E
# Noyau Normale K(t)
dev.new()
K=function(t){(1/sqrt(2*pi))*exp(-0.5*t^2)}
h=n^(-0.2)
# Initiation
s=100 # taille de l'intervalle [a,b]
a=min(X) #borne inf
b=max(X) # borne sup
x=seq(a,b,length=s) # Intervalle [a,b]
V=numeric(n)
fn=numeric(s)
for(j in 1 :s){
  for(i in 1 :n){ V[i]=K((x[j]-X[i])/h) }
  fn[j]=sum(V)/(n*h)}
# Fonction Hn(.)
W=numeric(n)
Hn=numeric(s)
for(j in 1 :s){ for(i in 1 :n){ W[i]=K((x[j]-X[i])/h)*Y[i] } Hn[j]=sum(W)/(n*h)}
Rn =Hn/fn
# Graphes
x11() # nouvelle fenetre graphique
dev.new()
```

```
op=par(mfrow=c(1,3))
```

```
plot(x,Rn,xlab="x", ylab="Rn(x)", main="n=40",col="red")
```

```
lines(x,sin(exp(x)),lwd= 2)
```

Puis, on refait le programme pour  $n=200$  et  $n=400$ , on obtient le graphe suivant :

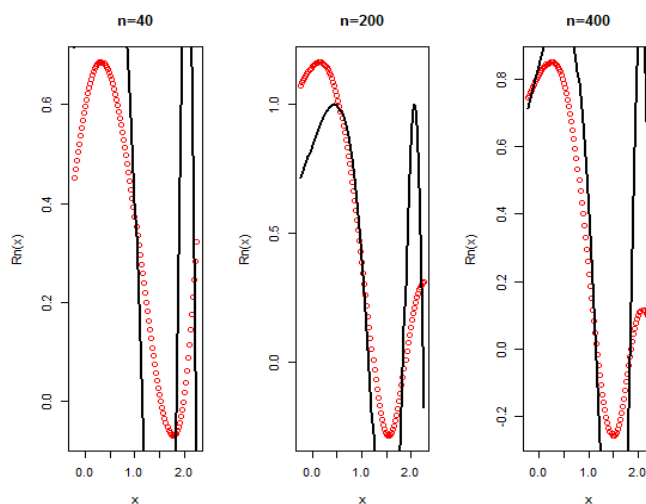


FIGURE 2.4: Cas du noyau gaussien

La courbe de l'estimateur  $\hat{r}(x)$  s'approche vers la courbe de la fonction  $r(x)$ , alors le noyau gaussien peut être le noyau optimale dans ce cas.

En choisissant le noyau triangulaire :  $K_2(t) = (1 - |t|)$ . On obtient le graphe suivant :

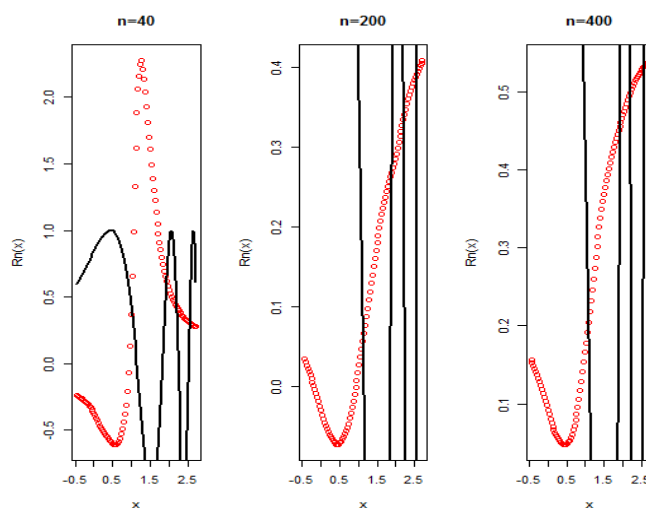


FIGURE 2.5: Cas du noyau triangulaire.

On remarque que le noyau triangulaire est un mauvais choix pour estimateur  $r(x)$ .

i- Choix de paramètre  $h$  : Le graphe ci dessous représente le graphe de la fonction  $r(x)$  et le graphe de l'estimateur  $\hat{r}(x)$ . avec le noyau  $K_5$  gaussien et  $h$  varié de 0.1 jusqu'à 0.9.

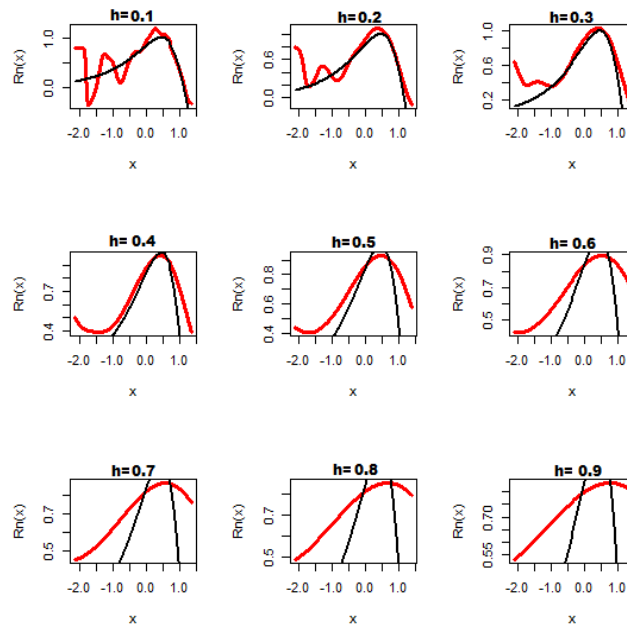


FIGURE 2.6: Représentation de  $\hat{r}(x)$  selon différentes valeurs de  $h$ .

Le code R utilisé :

```
n=350
X=rnorm(n,0,0.6)
E=rnorm(n)
Y=sin(exp(X))+E
# Noyau Normale K(t)
K=function(t){(1/sqrt(2*pi))*exp(-0.5*t^2)}
h=seq(.1,.9,length=9)
K
h
# Initiation
s=100 # taille de l'intervalle [a,b]
a=min(X) #borne inf
b=max(X) # borne sup
x=seq(a,b,length=s) # Intervalle [a,b]
V=array(dim=c(n,s,9))
```

```

fn=array(dim=c(s,9))
W=array(dim=c(n,s,9))
Hn=array(dim=c(s,9))
for(k in 1 :9){
for(j in 1 :s){
for(i in 1 :n){ V[i,j,k]=K((x[j]-X[i])/h[k]) }
fn[j,k]=sum(V[,j,k])/(n*h[k])}}
# fonction Hn(x)
for(k in 1 :9){
for(j in 1 :s){
for(i in 1 :n){ W[i,j,k]=K((x[j]-X[i])/h[k])*Y[i] }
Hn[j,k]=sum(W[,j,k])/(n*h[k])}}
Rn=array(dim=c(s,9))
for(k in 1 :9){ Rn[k]=Hn[,k]/fn[,k]}
# Graphes
x11() # nouvelle fenêtre graphique
op=par(mfrow=c(3,3))
for(k in 1 :9){
plot(x,Rn[k],xlab="x", ylab="Rn(x)",col="red",lwd=3,type="l")
lines(x,sin(exp(x)),lwd= 2)
}
par(op)

```

Il est clair que le  $h$  optimale est représenté dans le 4<sup>ème</sup> graphe, pour  $h = 0.4$ .

**Cas linéaire :** On définit la fonction de régression linéaire suivante :

$$r(x) = 2 + 0.8x + \epsilon$$

avec  $\epsilon \sim \mathcal{N}(0, 1)$  et  $x \sim \exp(1)$ . On veut montrer que le choix des paramètres  $K$  et  $h$  joue un rôle très important pour l'estimateur de  $[N-W]$  dans deux choix pour  $K$  (gaussien et triangulaire) et pour  $h$  varié entre 0.1 et 0.9.

i- Choix des paramètres  $K$  : Le graphe si dessous représente celui de la fonction  $r(x)$ .

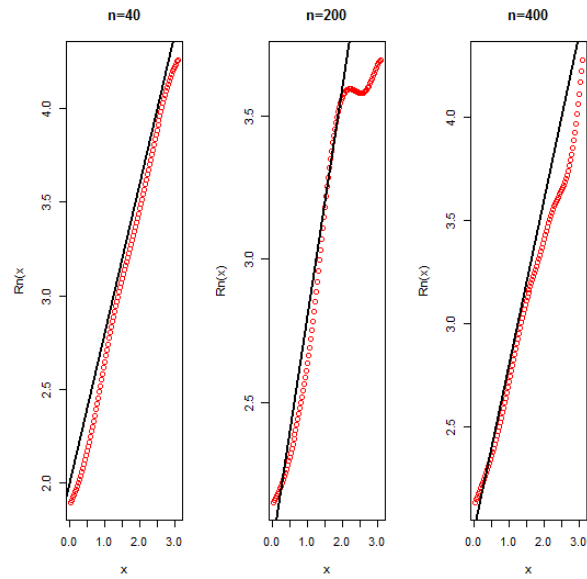


FIGURE 2.7: Cas du noyau gaussien.

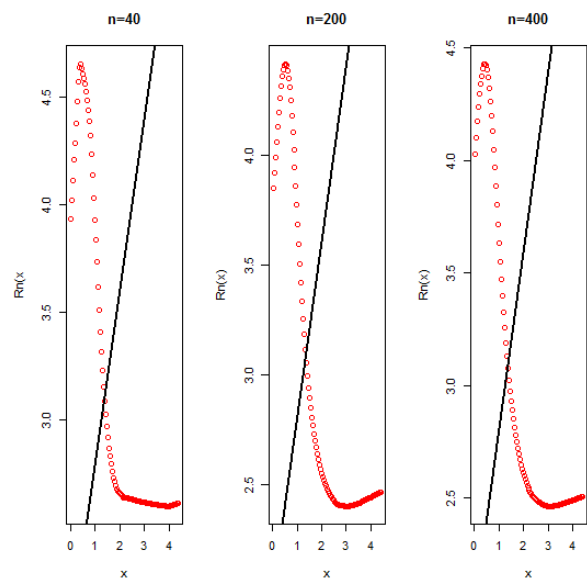
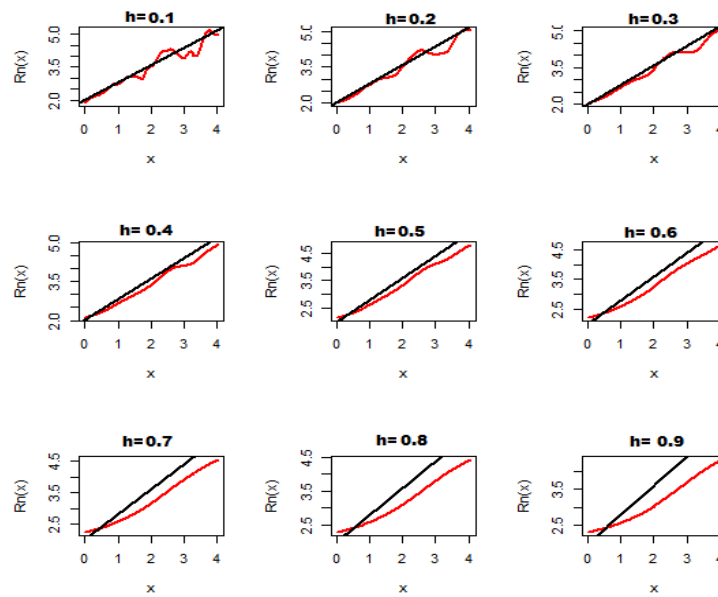


FIGURE 2.8: Cas de noyau triangulaire.

*Tandis-que le noyau triangulaire n'est pas un bon choix dans ce cas.*

ii- *Choix de paramètre  $h$  : Le graphe si dessous représente le graphe de la fonction  $r(x)$ .*



FIGURE 2.9: Représentation de  $\hat{r}(x)$  selon différentes valeur de  $h$ 

On obtient le bon estimateur  $\hat{r}(x)$  de la fonction  $r(x)$  lorsque  $h = 0.5$  qui est la valeur optimale.

**Remarque 2.1.** Le choix du paramètre de lissage  $h$  correspond à un arbitrage variance/biais :

1. Plus  $h$  est élevé, plus la courbe  $\hat{r}(x)$  sera lisse. La variance de l'estimation est limitée, mais l'estimateur  $\hat{r}(x)$  peut être fortement biaisé.
2. Plus  $h$  est faible, plus la courbe  $\hat{r}(x)$  est irrégulière. Les biais d'estimation de  $r(x)$  sont faibles, mais la variance de  $\hat{r}(x)$  est très importante.

Le choix de  $h$  résulte donc d'un arbitrage biais/variance, mais aussi d'un arbitrage lissage/non lissage de  $r(x)$  :

**Exemple 2.3.** Supposons que l'on choisisse  $h$  tel que  $h \rightarrow \infty$ . Alors, on a :

$$\lim_{h \rightarrow \infty} K\left(\frac{x_i - x_0}{h}\right) = K(0), \quad \forall x_i$$

Ceci implique que les poids de tous les indices  $i$  dans le calcul de  $\hat{r}(x_0)$  sont strictement identiques et égaux à :

$$\begin{aligned} \lim_{h \rightarrow \infty} W_i(x_0) &= \lim_{h \rightarrow \infty} \frac{K\left(\frac{x_i - x_0}{h}\right)}{\sum_{i=1}^n K\left(\frac{x_i - x_0}{h}\right)} \\ &= \frac{K(0)}{nK(0)} \\ &= \frac{1}{n} \end{aligned}$$

Ainsi, l'estimateur de  $W_i(x_0)$  est définie par :

$$\lim_{h \rightarrow \infty} W_i(x_0) = \lim_{h \rightarrow \infty} \sum_{i=1}^n W_i(x_0) y_i = \frac{1}{n} \sum y_i = \bar{y}.$$

Donc si le paramètre de lissage tend vers l'infinie, pour tous les points de l'échantillon, l'estimateur à noyau correspond à la moyenne empirique  $\bar{y}$ . La fonction de régression estimée correspond à une droite horizontale : la variance de  $\hat{r}(x)$  est nulle, mais le biais est sans doute fort.

**Exemple 2.4.** Supposons au contraire que l'on choisisse  $h$  tel que  $h \rightarrow 0$ . Alors, pour tous les points  $x_j$  différents du point  $x_i$  de référence :

$$\lim_{h \rightarrow 0} K\left(\frac{x_j - x_i}{h}\right) = K(\pm\infty) = 0, \quad \forall i \neq j$$

Par contre, pour le point de référence  $x_i$ , on a :

$$K\left(\frac{x_i - x_i}{h}\right) = K(0)$$

Donc, pour tous les indices autres que l'indice de référence dans le calcul de  $\hat{r}(x)$ , les poids  $W_j(x_i)$  sont nuls :

$$\lim_{h \rightarrow 0} W_j(x_i) = \lim_{h \rightarrow 0} \frac{K\left(\frac{x_j - x_i}{h}\right)}{\sum_{i=1}^n K\left(\frac{x_j - x_i}{h}\right)} = 0, \quad \forall i \neq j$$

En revanche, le poids de l'individu de référence  $x_i$  vérifie :

$$\begin{aligned} \lim_{h \rightarrow 0} W_j(x_i) &= \lim_{h \rightarrow 0} \frac{K\left(\frac{x_j - x_i}{h}\right)}{\sum_{i=1}^n K\left(\frac{x_j - x_i}{h}\right)} \\ &= \lim_{h \rightarrow 0} \frac{K(0)}{\sum_{j \neq i} K\left(\frac{x_j - x_i}{h}\right) + K\left(\frac{x_i - x_i}{h}\right)} \\ &= \frac{K(0)}{K(0)} = 1 \end{aligned}$$

Ainsi, l'estimateur de  $\hat{r}(x_0)$  est définie par :

$$\lim_{h \rightarrow 0} \hat{r}(x_0) = \lim_{h \rightarrow 0} \sum_{j \neq i} W_j(x_i) y_i + \lim_{h \rightarrow 0} W_i(x_i) y_i = y_i$$

Alors si le paramètre de lissage  $h \rightarrow 0$ , pour tous les points de l'échantillon, l'estimateur à noyau correspond exactement à l'observation  $y_i$ . La fonction de régression estimée passe exactement par tous les points de l'échantillon : la variance de  $\hat{r}(x_0)$  est aussi grande que la variance de  $y$ , mais le biais est faible.

Les méthodes de la minimisation de l'estimation par l'Erreur Moyenne Quadratique ( $MSE$ ) ou par l'Erreur Intégrée Moyenne Quadratique ( $MISE$ ) restent incomplètes ou autrement dit, ces méthodes ont toujours des inconvénients, soit au sens de la qualité de l'estimateur par rapport à une norme d'erreur bien déterminée, soit par l'allure graphique de la courbe (lisse ou non). Parmi ces méthodes, Pour parer à cette difficulté, ils existent trois méthodes célèbres . La famille des méthodes de validation croisées (cross validation) et la famille des méthodes Plug in et aussi la famille Bootstrap. Nous nous limitons à la méthode de la validation croisée pour la régression.

## 2.2 Validation croisée

Une des méthodes les plus utilisées qui fournit une fenêtre proche de l'optimale, la méthode de validation croisée. La définition précise de cette méthode sera donnée plus tard dans cette section.

### 2.2.1 Principe de la méthode

Dans cette section, nous supposons le noyau  $K$  fixé, et on ne s'intéresse qu'au choix de la fenêtre  $h$ . Nous avons observé dans les précédents paragraphes que l'efficacité de l'estimateur [N-W] est liée au paramètre de lissage, la fenêtre  $h$ . Il faut choisir la fenêtre afin d'équilibrer un terme stochastique (la variance) et un terme déterministe (le biais), si possible indépendamment des propriétés de régularité de la courbe de régression. Dans la section précédente, la fenêtre optimale qui minimise le risque quadratique intégré ( $MISE$ ) est obtenue sous des hypothèses de régularité spécifiques et dépend alors de quantités fonctionnelles inconnues, de la distribution du couple  $(X, Y)$ . Afin de construire un estimateur non oracle qui minimise l'erreur quadratique, il faut utiliser d'autres méthodes dont la plus commune est appelée la procédure de validation croisée. L'idée principale de la validation croisée consiste à minimiser, par rapport à  $h$ , l'estimé d'une mesure de la  $MISE$ . La fenêtre  $h$  n'est alors plus déterministe, elle dépend des observations.[6]

### 2.2.2 Plan de travail

Dans le cas des estimateurs à noyaux (chapitre 1), et pour les estimateurs par polynômes locaux (chapitre 3), on doit choisir la fenêtre  $h$ , pour les estimateurs constants par morceaux. Dans cette section, nous allons décrire la méthode de validation croisée, dont le but est de choisir ce paramètre, ce qui correspond à sélectionner un estimateur dans une collection d'estimateurs.

Notons  $h$  le paramètre à choisir. Soit  $\hat{r}$  l'estimateur de la fonction de régression  $r$ , associé à ce paramètre  $h$ . On considère l'erreur quadratique moyenne :

$$MSE(h) = \mathbb{E}\left(\frac{1}{n}(\hat{r}(x_i) - r(x_i))^2\right) = R(h)$$

Idéalement, on souhaiterait choisir  $h$  de manière à minimiser  $R(h)$ , mais cette quantité dépend de la fonction inconnue  $r$ . Une première idée est d'estimer  $R(h)$  par l'erreur d'estimation :

$$\frac{1}{n} \sum_{i=1}^n (y_i - \hat{r}(x_i))^2$$

mais cette quantité sous-estime  $R(h)$  et conduit à un sur-ajustement. Ceci est dû au fait que l'on utilise les mêmes données pour construire l'estimateur  $\hat{r}$  (qui est construit pour bien s'ajuster à l'échantillon d'estimation) et pour estimer l'erreur commise par cet estimateur. Pour avoir une meilleure estimation du risque, on doit construire l'estimateur du risque avec des observations qui n'ont pas été utilisées pour construire l'estimateur  $\hat{r}$ . Idéalement, si on avait assez d'observations, on pourrait les séparer en un échantillon d'apprentissage et un échantillon test. Ce n'est généralement pas le cas, et on souhaite utiliser l'ensemble des données d'estimation pour la construction de l'estimateur. Alors, recours à la validation croisée. On partitionne l'échantillon d'estimation en  $V$  blocs, notés  $B_1, \dots, B_V$ , de tailles à peu près identiques. Pour tout  $v$  de 1 à  $V$ , on note  $\hat{r}^{-v(i)}$  l'estimateur obtenu en supprimant de l'échantillon d'estimation les données appartenant au bloc  $B_v$ .

**Définition 2.2.** On définit le score de validation croisée  $v$ -fold par :

$$CV = \hat{R}(h) = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{r}^{-v(i)}(x_i))^2$$

où  $\hat{r}^{-v(i)}(x_i)$  est l'estimateur de  $r$  obtenu en enlevant les observations du bloc qui contient l'observation  $i$ . Le principe de la validation croisée est de choisir une valeur  $\hat{h}$  de  $h$  qui minimise la quantité  $\hat{R}(h)$ . Un cas particulier correspond à la validation croisée leave-one-out, obtenue quand on considère  $n$  blocs, chacun est réduit à une observation.

**Définition 2.3.** *Le score de validation croisée leave-one-out est défini par :*

$$CV = \hat{R}(h) = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{r}^{(-i)}(x_i))^2$$

où  $\hat{r}^{(-i)}(x_i)$  est l'estimateur de  $r$  obtenu en enlevant les observations  $(x_i, y_i)$ . L'idée de la validation croisée leave-one-out vient du calcul suivant :

$$\begin{aligned} \mathbb{E}((y_i - \hat{r}^{(-i)}(x_i))^2) &= \mathbb{E}((y_i - r(x_i) + r(x_i) - \hat{r}^{(-i)}(x_i))^2) \\ &= \sigma^2 + \mathbb{E}((r(x_i) - \hat{r}^{(-i)}(x_i))^2) \\ &\simeq \sigma^2 + \mathbb{E}((r(x_i) - \hat{r}(x_i))^2) \end{aligned}$$

On obtient donc  $\mathbb{E}(\hat{R}(h)) \simeq \sigma^2 + R(h)$ . Le calcul de  $\hat{R}(h)$  peut s'avérer long, mais dans certains cas, il n'est pas nécessaire de recalculer  $n$  fois un estimateur de la fonction de régression. Dans cette méthodes l'estimateur correspond à un algorithme de moyennes locales, c'est-à-dire est de la forme

$$\hat{r}(x) = \sum_{j=1}^n y_j W_j(x),$$

avec  $\sum_{j=1}^n W_j(x) = 1$ , d'autre part

$$\hat{r}^{(-i)}(x) = \sum_{j=1}^n y_j W_j^{(-i)}(x)$$

avec

$$W_j^{(-i)}(x) = \begin{cases} 0 & \text{si } j = i \\ \frac{W_j(x)}{\sum_{k \neq i} W_k(x)} & \text{si } j \neq i \end{cases}$$

**Théorème 2.1.** [27] *Sous les hypothèses ci-dessus concernant l'estimateur, le score de validation croisée leave-one-out est égal à :*

$$CV = \hat{R}(h) = \frac{1}{n} \sum_{i=1}^n \left( \frac{y_i - \hat{r}(x_i)}{1 - W_i(x_i)} \right)^2$$

On trouve également dans les logiciels une définition légèrement différente :

**Définition 2.4.** *On appelle score de validation croisée généralisée la quantité :*

$$GCV(h) = \frac{1}{n} \sum_{i=1}^n \left( \frac{y_i - \hat{r}(x_i)}{1 - v/n} \right)^2$$

où  $\frac{v}{n} = \sum_{i=1}^n \frac{W_i(x_i)}{n}$ . Dans cette définition,  $W_i(x_i)$  est remplacé par la moyenne des  $W_i(x_i)$ . En pratique, les deux méthodes donnent généralement des résultats assez proches. En utilisant l'approximation  $(1 - x)^{-2} \approx 1 + 2x$  pour  $x$  proche de 0, on obtient :

$$GCV(h) \approx \frac{1}{n} \sum_1^n (y_i - \hat{r}(x))^2 + \frac{2v\sigma^2}{n},$$

où  $\sigma^2 = \frac{1}{n} \sum_1^n (y_i - \hat{r}(x))^2$

**Remarque 2.2.** *Les méthodes les plus connues et les plus utilisées, pour choisir le paramètre de lissage en plus de la méthode de validation croisée, on a :*

1. *Les méthodes plug-in (ré-injection). Le principe de ces méthodes repose sur l'estimation d'une quantité qui dépend de la dérivée seconde de la densité de probabilité inconnue  $r$  qu'on désire estimer. La méthode donne de meilleurs résultats théoriques et pratiques[4].*
2. *La méthode de bootstrap : L'application de la méthode bootstrap permettent d'estimer une région de confiance pour les courbes des fonctions  $r$ . cette méthode est plus coûteuse en calcul et s'applique aussi bien dans le cas discret que dans le cas continu [20].*

*On donnera les deux méthodes numériques pour pouvoir les comparer avec l'estimateur à noyau.*

---

---

## CHAPITRE 3

---

# COMPARAISON AVEC LES ESTIMATEUR LOESS ET SPLINE

Un des problèmes essentiels pour réaliser la régression à noyau ou l'estimation de la fonction de régression par noyau réside dans le manque de robustesse de ces estimateurs pour les valeurs extrêmes de  $X$  : Une solution alternative, plus robuste pour les valeurs extrêmes consiste à utiliser la régression locale. Elle combine la simplicité de régression linéaire par les moindres carrés avec la flexibilité de la régression non linéaire. Le plus grand avantage de la régression locale qu'elle n'exige que de préciser la valeur du paramètre de lissage et le degré du polynôme local. La méthode de régression locale est aussi très souple en application, et relativement simple à implémenter [16], on propose aussi la méthode de splines de régression tel que, les splines de lissage et la B-spline. Ces méthodes permettent toutes de contrôler la flexibilité de l'estimateur et elle est plus stable par exemple la méthode B-spline permette de changer l'ordre de la courbe sans changer le nombre de sommets du polygone. Cela est avantageux car une courbe d'ordre élevé est plus difficile à calculer avec précision. De plus elles permettent de créer des courbes avec des points anguleux. Au plus le degré de la spline diminue, tant-que la courbe se rapproche du polygone, inversement c'est le degré de la courbe augmente, la courbe est "tendue" entre les points extrême. Chaque sommet affecte la forme de la courbe dans un intervalle donné[26].

### 3.1 La régression locale : Loess

Cette méthode est, probablement, la plus utilisée pour faire de la régression non paramétrique. Ceci s'explique principalement par le fait qu'elle est très simple, ce qui lui donne l'avantage d'être habituellement plus rapide d'exécution que ses concurrentes. C'est aussi un très bon outil lorsque l'on désire connaître l'allure générale de la courbe dans le but de faire une régression paramétrique.[26] On cherche toujours à estimer le modèle (1,1) :

$$Y_i = r(x_i) + \epsilon_i, i = 1, \dots, n.$$

Où  $r(x_i)$  est toujours une fonction inconnue.

L'idée de la régression linéaire locale consiste à utiliser un modèle de régression défini uniquement dans un voisinage du point  $x_0$  d'intérêt [16]. Notons  $N(x_0)$  ce voisinage.

**Définition 3.1.** *Le principe général d'une régression locale est de postuler que la fonction de lien  $r(x_0)$ , évaluée au point  $x_0$ , peut être approximée par la valeur d'une fonction paramétrique évaluée localement au voisinage  $N(x_0)$  du point de référence  $x_0$ . Par exemple, on peut penser à approximer  $r(x_0)$  par son estimateur :*

$$\hat{r}(x_0) = \hat{\beta}_0(x_0) + \hat{\beta}_1(x_0)x_0;$$

Où les estimateurs des paramètres  $\hat{\beta}_0(x_0)$  et  $\hat{\beta}_1(x_0)$  sont déterminés par la formule suivante :

$$\{\hat{\beta}_0(x_0), \hat{\beta}_1(x_0)\} = \arg \min \sum_{x_i \in N(x_0)} [y_i - \hat{\beta}_0(x_0) - \hat{\beta}_1(x_0)x_0]^2.$$

On peut distinguer deux types d'estimateur de la régression local :

1. La régression locale ou LOESS (Local Regression) de Cleveland (1979)[8].
2. La régression locale pondérée ou LOWESS (Locally Weighted Scatterplot Smoothing) de Cleveland et Devlin (1988) [7].

Considérons le cas de la régression LOESS. La principale différence avec la régression polynomial local (introduite dans la section suivante) c'est que la valeur de  $r(x_0)$  estimée n'est pas une moyenne mais une valeur prévue par une droite de régression. Par contre c'est une méthode qui requiert plus de temps de calcul : pour  $n$  observations on doit faire  $n$  régressions.[16]

**Définition 3.2.** *Dans le contexte de la régression locale, souvent on caractérise le voisinage de la variable  $x_0$ , noté  $N(x_0)$ , par un rapport constant, appelé smoothing parameter,*



quel que soit le point considéré :

$$h = \frac{\dim[N(x_0)]}{n} \in ]0, 1], \forall x_0$$

Si  $h$  est trop faible, l'estimateur des paramètres  $\beta_0(x_0)$  et  $\beta_1(x_0)$  manque de précision car le voisinage est trop petit, si au contraire le voisinage couvre l'ensemble des observations ( $h = 1$ ), alors, on retrouve la droite d'ajustement linéaire (modèle de régression simple).

**Remarque 3.1.** Les polynômes de degré un ou deux sont ordinairement favorisés, car ils permettent d'obtenir des résultats satisfaisants en un temps minimal.

Maintenant on va introduire deux autres méthodes d'estimation non paramétrique local, la régression polynomiale locale pondérée et la régression polynomiale locale (RPL).

## 3.2 La régression par polynômes locaux pondérées

La régression par polynômes locaux est très similaire à l'estimation du noyau. Par contre, les valeurs obtenues sont produites par régression polynomiale pondérée par la distance au lieu de la moyenne pondérée par la distance. En fait, ce qui est différent de la méthode du noyau est que l'estimation de  $r(x)$  est obtenue par régression polynomiale de  $y$  sur  $x$  [23]. Il faut maintenant minimiser la somme des carrés résiduels pondérée suivante :

$$\sum_{i=1}^n W_i(x) [Y_i - \beta_0 + \beta_1 x + \dots + \beta_s x^s]^2$$

où  $W_i(x) = \frac{K(\frac{x_i-x}{h})}{\sum_{i=1}^n K(\frac{x_i-x}{h})}$  Voici maintenant la solution à ce problème de minimisation Simonoff, (1996), [20]. Soit la matrice :

$$X_x = \begin{pmatrix} 1 & x - x_1 & \dots & (x - x_1)^s \\ 1 & \vdots & \vdots & \ddots & \vdots \\ 1 & x - x_n & \dots & (x - x_n)^s \end{pmatrix}$$

Et soit la matrice

$$L = \frac{1}{h} \text{diag}[K(\frac{x - x_1}{h}), \dots, K(\frac{x - x_n}{h})]$$

qui s'appelle la matrice des poids. Alors si  $X'_x L X_x$  est, inversible, on peut écrire

$$\hat{\beta} = (X'_x L X_x)^{-1} X'_x L Y,$$

où

$$Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \quad \text{et} \quad \hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \vdots \\ \hat{\beta}_s \end{pmatrix}$$

Ensuite, on pose  $\hat{r}_0(x) = \hat{\beta}_0$ .

Il est important de souligner le fait que la méthode du noyau est en réalité un cas particulier de la méthode par polynômes locaux. En effet, pour un choix de  $s = 0$ , l'équation à minimiser devient exactement celle qui donnait la méthode du noyau :

$$\sum_{i=1}^n W_i(x)[Y_i - \beta_0]^2$$

Comme dans les méthodes précédentes, il faut définir le biais et la variance de l'estimateur. **Biais et variance**

Dans le cas de l'estimateur de polynôme locale pondérée, on peut définir deux type de formule pour le biais et aussi pour la variance de l'estimateur  $\hat{r}_1(x)$  à l'intérieur des bornes et à l'extrémités de données[23]. On commence par le biais et la variance à l'intérieurs des bornes.

$$\mathbb{Biais}(\hat{r}(x)) = \frac{h^2 r''(x) \int u^2 K(u) du}{2} + o_s(h^2)$$

La variance vaut pour sa part à :

$$\mathbb{Var}(\hat{r}(x)) = \frac{r'(x)f'(x)}{f(x)} \int u^2 K(u) du,$$

on remarque que, comme dans le cas de polynôme locale, la variance est la même que l'estimateur à noyau. Ainsi, contrairement au biais trouvé par la méthode du noyau, le biais obtenu par la méthode localement linéaire ne dépend ni de la fonction de densité des valeurs exogènes  $x_i$ , ni de la dérivée de  $f(x)$ . Et comme d'habitude il faut obtenir la formule *AMISE* de l'estimateur.

$$AMISE(\hat{r}(x)) = \left[ \frac{h^2 \int u^2 K(u) du}{2} \right]^2 \int r''(u)^2 f(x) du + \frac{\int K(u^2) du \sigma^2}{nh} + o[h^4 + (nh)^{-1}]$$

la fenêtre de lissage optimale  $h$  :

$$h_{opt} = \left[ \frac{\int K(u^2) \sigma^2 du}{n \int u^2 K(u)^2 du \int r''(x)^2 f(u) du} \right]^{\frac{1}{5}}$$

Pour la démonstration (Voire [23])

On s'adresse, maintenant, aux biais et la variance près des bornes : on définit le terme  $t_{n,c}$  :

$$t_{n,c} = \int_{-\infty}^c u^n K(u) du, \quad n = 0, 1, 2, 3,$$

de nouvelles formules s'appliquant aux estimateurs évalués à des points  $x$  se trouvant aux extrémités des données peuvent être obtenues. Ainsi, le biais est corrigé et il devient alors

$$\mathbb{Biais}(\hat{r}(x)) = \frac{a_K(c)r'(x)h^2}{2} + o_s(h^2)$$

Où

$$a_K(c) = \frac{t_{2,c}^2 - t_{3,c}t_{1,c}}{t_{2,c}t_{0,c} - t_{1,c}^2}$$

Cette formule permet de constater qu'une fois de plus, pour les estimateurs évalués à des points  $x$  situés aux bornes, l'estimateur par la méthode du noyau est plus biaisé que l'estimateur par la méthode localement linéaire. En effet, le biais est d'ordre  $o(h)$  dans le cas de la méthode du noyau, alors qu'il est d'ordre  $o(h^2)$  pour la méthode localement linéaire.

Évidemment, la correction pour le biais implique nécessairement un prix à payer, c'est-à-dire, que la variance devient plus élevée. Cette variance asymptotique conditionnelle de  $\hat{r}$  près des bornes est :

$$\mathbb{Var}(\hat{r}(x)) = \frac{\beta_K(c)\sigma(x)}{nhf(x)} + o_s((nh)^{-1})$$

Où

$$\beta_K(c) = \frac{\int_{-\infty}^c (t_{2,c} - ut_{1,c})K^2(u)du}{(t_{2,c}t_{0,c} - t_{1,c}^2)^2}$$

Aux extrémités des données, si le noyau gaussien est utilisé et que la même fenêtre de lissage est choisie, la variance asymptotique conditionnelle de  $\hat{r}$  est environ 3.17 fois celle de  $\hat{r}^{NW}$  [23],

### 3.3 La régression polynomiale locale (RPL)

Cette méthode est une généralisation de la méthode Local Weighted Regression ou LOWESS de Cleveland (1979) (Voir [8]) proposée par Lejeune (1983) [19], dans le cadre de l'estimation par noyau. Elle consiste, pour estimer  $r(x)$ , à ajuster une fonction polynomiale de degré  $s$  choisi, sur les couples de points  $(x_i, y_i)$  dont les  $x_i$  sont situés dans le voisinage

(fenêtre)  $[x - h, x + h]$  de  $x$ . L'ajustement s'entend au sens classique des moindres carrés des résidus ( $\epsilon_i = y_i - \hat{y}_i$ ) et est donc un cas particulier de la régression linéaire multiple, la fonction polynomiale étant linéaire par rapport aux paramètres inconnus. Il se résout matriciellement. Alors  $r(x)$  est estimé par la valeur ajustée au point  $x$ , que nous noterons  $\hat{r}(x)$ . [20]

Soit  $P(x) = \beta_0 + \beta_1 x + \dots + \beta_s x^s$  un polynôme de degré  $s$ . Cet ajustement s'opère avec les valeurs de  $\beta_0, \beta_1, \dots, \beta_s$  telles que l'expression :

$$\sum_{i=1}^n [y_i - \beta_0 + \beta_1 x + \dots + \beta_s x^s]^2 K\left(\frac{x - x_i}{h}\right)$$

soit minimale, où  $K$  est la fonction indicatrice de l'appartenance de la valeur  $x_i$  à la fenêtre ( $K(u) = 1$  si  $|u| \leq 1$ , 0 sinon). Soit  $a_0, \hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_s$  les valeurs permettant d'atteindre le minimum,  $r(x)$  est alors estimé par  $\hat{r}(x) = \hat{\beta}_0 + \hat{\beta}_1 x + \dots + \hat{\beta}_s x^s$ .

Comme pour l'estimateur à noyau les propriétés de la fonction  $K$  se transfèrent à la fonction  $\hat{r}(x)$  et l'on aura avantage à substituer à la fonction indicatrice une fonction de pondération dérivable partout, ce qui conduit à une solution des moindres carrés pondérés. On montre aisément que la solution matricielle pour le vecteur des paramètres devient  $(X_x^t L X_x)^{-1} X_x^t L Y$  où  $L$  est la  $n \times n$ -matrice diagonale des poids telle que  $K(\frac{x-x_i}{h})$  affectés aux  $n$  observations, et

$$X_x = \begin{pmatrix} 1 & x - x_1 & \dots & (x - x_1)^s \\ 1 & \vdots & \vdots & \ddots & \vdots \\ 1 & x - x_n & \dots & (x - x_n)^s \end{pmatrix}$$

**Remarque 3.2.** On pourra choisir pour  $K$  le noyau le plus simple possédant les qualités requises, à savoir le biweight de Tukey. Notons d'ailleurs que l'estimation par noyau correspond au cas particulier  $s = 0$ , car la valeur de  $a_0$  qui minimise  $\sum_{i=1}^n [y_i - \beta_0]^2 K(\frac{x-x_i}{h})$  est la moyenne pondérée des  $y_i$  avec les poids  $K(\frac{x-x_i}{h})$ .

**Remarque 3.3.** Pour  $s = 2$ , il est possible d'obtenir un meilleur ajustement pour les zones à forte courbure, notamment en ce qui concerne le problème de l'écrêtement des extrêmes. Nous allons vérifier cela sur les propriétés de la RPL.

### Biais et la variance

Le biais et la variance asymptotiques ont été établis par Fan (1993) [12]. En fait la

variance ne dépend pas du degré de la RPL et reste égale à celle indiquée plus haut pour l'estimateur à noyau correspondant au cas  $s = 0$ . Pour le biais, on obtient pour  $s \geq 1$  :

$$\text{Biais}(\hat{r}(x)) = \frac{h^{s+1}}{(s+1)!} r^{(s+1)}(x) \int_{-1}^1 x^{s+1} K(x) dx + o(h^{s+1}).$$

Où  $r^{(s+1)}(x)$  la dérivée d'ordre  $s + 1$  de la fonction  $r$ . En pratique le choix  $s = 1$ , proposé à l'origine par Cleveland et repris par divers auteurs, n'est cependant pas satisfaisant car il ne traite pas le problème de l'écrêtement des extremas (ou, plus généralement, du biais dans les zones à forte courbure). Pour cela, on peut considérer qu'un ajustement parabolique ( $s = 2$ ) suffira, d'autant qu'à  $n$  fini la variance augmente avec l'ordre du biais. Notons que la RPL n'évite pas le problème du choix de la largeur de fenêtre. Néanmoins, l'estimation de  $r(x)$  est moins sensible à ce paramètre qu'avec un estimateur à noyau. [20]

Pour approfondir la régression non paramétrique, on pourra consulter les ouvrages de Härdle (1990) (Voir [15]) et de Simonoff (1996) (Voir [24]).

**Remarque 3.4.** *A l'origine la méthode a été également appelée «régression polynomiale mobile» par référence à la moyenne mobile analogue dans son esprit à l'estimateur à noyau.*

Une autre alternative pour estimer la fonction de régression est les splines de régression, plus particulièrement les B-splines, qui sont bien connues en analyse numérique et sont très utilisées en modélisation graphique. En statistique, on les préfère aux autres méthodes non paramétriques pour leurs qualités mathématiques qui combinent la simplicité des calculs à une expression explicite et à la possibilité d'effectuer des tests statistiques comme dans la régression linéaire usuelle. Le lecteur intéressé à étudier, plus en détails, les splines de régression univariées peut consulter Eubank (voir [11]).

### 3.4 B-spline :

**Définition 3.3.** *En mathématique B-spline est une combinaison linéaire de spline positive à support compacte minimale de plus c'est une méthode locale. La théorie des fonctions B-splines utilisée pour l'estimation des fonctions de régression a l'intérêt d'être adaptée aux différents types d'espaces de représentation.*

Nous supposons, ici, que les  $x_i$  appartiennent à  $\mathbb{R}$ . Les estimateurs de la section précédente ne sont pas continus, pour obtenir des estimateurs qui sont des polynômes par

morceaux et qui ont des propriétés de régularité, on utilise les bases de splines.

Il y'a plusieurs types de splines comme :

1. Spline uniforme, Spline linéaire, Spline quadratique et Spline cubique.

Ce qui suit, on s'intéresse aux splines cubiques (les polynômes qui composent la fonction splines de lissage sont de degré 3) :

### 3.4.1 Splines linéaires et cubiques :

**Définition 3.4.** Soit  $0 < a < b < c \dots$ , les points qui déterminent les intervalles de la partition (appelés les nœuds), et  $0 < a < b < c \dots$  inclus dans l'intervalle :  $[t_1, t_n]$ .

Une fonction  $r$  est une spline cubique si les deux conditions suivantes sont satisfaites :

- i. Sur chaque intervalle  $[t_1, a], [a, b], [b, c] \dots$ ,  $r$  est un polynôme cubique
- ii. La fonction  $g$  est deux fois continûment différentiables sur  $[t_1, t_n]$  (et donc  $r$  et ses dérivées d'ordre 1 et 2 sont continues).

Il y a de nombreuses façons essentiellement équivalentes de définir des splines cubiques.

La façon la plus naturelle est d'exprimer :

$$r(x) = \beta_0 + \beta_1 x + \beta_2(x - a)_+ + \beta_3(x - b)_+ + \beta_4(x - c)_+ + \dots + .$$

Telle que  $\beta_i \in \mathbb{R}$ , et  $u_+ = \max(0, u)$

$$r(x) = \begin{cases} \beta_0 + \beta_1 x & , x \leq a \\ \beta_0 + \beta_1 x + \beta_2(x - a)_+ & , a \leq x \leq b \\ \beta_0 + \beta_1 x + \beta_2(x - a)_+ + \beta_3(x - b)_+ & , b \leq x \leq c \end{cases}$$

La fonction  $f$  est continue, si on veut imposer plus de régularité (par exemple  $f$  de classe  $\mathcal{C}^2$ ), on utilise des splines cubiques.

$$r(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4(x - a)_+^3 + \beta_5(x - b)_+^3 + \beta_6(x - c)_+^3 + \dots +$$

La fonction  $(x - a)^3$  s'annule ainsi que ses dérivées d'ordre 1 et 2 en  $a$  donc  $f$  de classe  $\mathcal{C}^2$ .

Pour éviter les problèmes de bords, on impose souvent des contraintes supplémentaires

aux splines cubiques, notamment la linéarité de la fonction sur les deux intervalles correspondant aux extrémités. On se place sur  $[0, 1]$ .  $\xi_0 = 0 < \xi_1 < \dots < \xi_K < 1$

$$r(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \sum_{k=1}^K \theta_k (x - \xi_k)_+^3$$

On impose  $f''(0) = f^{(3)}(0) = 0$ ,  $f''(\xi_K) = f^{(3)}(\xi_K) = 0$ . On en déduit :

$$\beta_2 = \beta_3 = 0, \sum_{k=1}^K \theta_k (\xi_K - \xi_k) = 0, \sum_{k=1}^K \theta_k = 0$$

$$\begin{aligned} r(x) &= \beta_0 + \beta_1 x + \sum_{k=1}^K \theta_k [(x - \xi_k)_+^3 - (x - \xi_K)_+^3] \\ &= \beta_0 + \beta_1 x + \sum_{k=1}^K \theta_k (\xi_K - \xi_k) \left[ \frac{(x - \xi_k)_+^3 - (x - \xi_K)_+^3}{(\xi_K - \xi_k)} \right] \end{aligned}$$

On pose  $\gamma_k = \theta_k (x - \xi_k)$  et  $d_k(x) = \frac{(x - \xi_k)_+^3 - (x - \xi_K)_+^3}{(\xi_K - \xi_k)}$ ,  $\sum_{k=1}^K \gamma_k = 0$ , on obtiens :

$$r(x) = \beta_0 + \beta_1 x + \sum_{k=1}^K \gamma_k [d_k(x) - d_{K-1}(x)].$$

On obtient la base de splines naturels :

$V_1(x) = 1, V_2(x) = x, \forall 1 \leq k \leq K - 2, V_{k+2}(x) = d_k(x) - d_{K-1}(x)$ . On doit choisir la position et le nombre de nœuds.

### Méthodes de régularisation :

On se place toujours dans le modèle de régression :  $y_i = r(x_i) + \epsilon_i$ ,  $1 \leq i \leq n$ . On minimise parmi les fonctions  $r$  splines naturels de nœuds en  $x_i$  ( $r(x) = \sum_{k=1}^n \theta_k V_k(x)$ ) le critère pénalisé :

$$C(f, \lambda) = \sum_{i=1}^n (y_i - r(x_i))^2 + \lambda \int_0^1 (r''(t))^2 dt.$$

Où  $\lambda > 0$ . En notant  $\Omega_{l,k} = \int_0^1 V_k''(x) V_l''(x) dx$  et  $V_{i,j} = V_j(x_i)$ , le critère à minimiser est

$$C(\theta, \lambda) = \|y - V\theta\|^2 + \lambda \theta^t \gamma \theta,$$

la solution est :

$$\hat{\theta} = (V^t V + \lambda \gamma)^{-1} V^t Y.$$

Et

$$\hat{r}(x) = \sum_{k=1}^n \hat{\theta}_k V_k(x) \tag{3.1}$$

**Théorème 3.1.** [27] On note :

$$\mathcal{F} = \{r, C^2([0, 1]), \int_0^1 r''^2(t)dt < +\infty\}$$

On se donne  $n \geq 2, 0 < x_1 < \dots < x_n < 1$  et  $(y_1, \dots, y_n)$ . Pour  $r \in \mathcal{F}$ , et  $\lambda > 0$ , on note

$$C(\theta, \lambda) = \|y - V\theta\|^2 + \lambda\theta^*\gamma\theta$$

Pour tout  $\lambda > 0$ , il existe un unique minimiseur dans  $\mathcal{F}$  de  $C(\theta, \lambda)$ , qui est la fonction définie en (3.1)

**Théorème 3.2.** L'estimateur de B-spline  $\hat{r}(x)$  est un estimateur linéaire lisse, ce qui'il existe une fonction pondérée  $W_i(x)$  comme  $\hat{r}(x) = \sum_{i=1}^n y_i W_i(x)$ . En particulier la matrice de lissage  $L_s$  est définie par :

$$L_s = V(V^tV + \lambda\Omega)^{-1}V^t$$

On définit le degré de liberté par :  $df = tr(L_s)$  et puis on choisit le paramètre  $\lambda$  qui minimise le critère de validation croisée (où la validation croisée généralisé)

## 3.5 Application :

Nous sommes intéressées par l'application des méthodes d'estimation dans le domaine médicale précisément la maladie de diabète. Pour cela, on va donner quelques notions biologiques.

L'hémoglobine glyquée (fraction HbA1c) est la forme glyquée de la molécule d'hémoglobine. Sa valeur biologique permet de déterminer la concentration de glucose dans le sang, la glycémie, sur trois mois. Elle est particulièrement utile et constitue le paramètre de référence dans la surveillance de l'équilibre glycémique des patients diabétiques.[28]

Son dosage régulier, par un prélèvement sanguin veineux, permet de surveiller l'équilibre glycémique des patients diabétiques et, ainsi, d'évaluer et d'adapter leurs traitements anti-diabétiques. Sa valeur augmente lorsque les périodes d'hyperglycémie ont été fréquentes dans les 120 jours précédant le dosage et diminue lorsque la glycémie a été correctement équilibrée. Ainsi meilleur est le contrôle glycémique, plus basse est l'HbA1c et moindre est le risque de développer une des complications du diabète (microangiopathie, macroangiopathie...). L'équilibre chronique d'un diabète sucré se juge sur l'HbA1c et pas sur la glycémie, sauf au moment de l'adaptation du traitement ou lors de situations aiguës.



Sa valeur normale se situe entre 4 et 6% [28] de l'hémoglobine totale, et sa valeur de risque se situe entre 9 et 11% de l'hémoglobine totale. une estimation de la fonction de régression Notre étude concerne de trouver la relation non paramétrique entre (HbA1c) et l'age des patients (age) selon les méthodes à noyau, Loess et B-spline. Pour cela nous avons utilisés des données réels de 51 malades diabétiques d'après la clinique 40 Hectares-Jijel.

**premier trimestre :**

age	HbA1c(%)	age	HbA1c(%)	age	HbA1c(%)	age	HbA1c(%)
66	7.7	70	7.6	60	10.9	46	8.7
63	9	75	7.3	48	9.6	72	7.3
66	9	56	9.5	65	7.1	62	6.2
62	9.1	55	10.4	46	7.4	52	8.5
67	8.3	62	9.9	56	8.7	58	7.6
43	9.5	60	6.6	64	6.9	62	7.3
72	7.2	50	9.3	73	6.6	79	9
65	11.5	51	7.9	65	8.1	63	7.6
62	7.9	49	7.6	59	7.7	67	7.9
72	11	64	10	66	6.3	54	8.5
64	7.6	67	6.7	76	10.1	62	6
38	6.7	48	6.4	79	6.9	53	9.4
63	8.2	53	7.3	62	8.4	/	/

**deuxième trimestre :** les données de deuxième trimestre pour les mêmes malades sont comme suit :

age	HbA1c(%)	age	HbA1c(%)	age	HbA1c(%)	age	HbA1c(%)
66	7.4	70	8	60	10	46	9
63	8.9	75	8.1	48	10	72	8
66	11.5	56	9	65	7.3	62	6.2
62	6.6	55	10.4	46	7.9	52	9.1
67	8	62	9	56	8.3	58	7.9
43	8.1	60	12.5	64	7	62	7.1
72	8	50	9.6	73	7	79	10
65	7	51	7	65	8	63	7.8
62	6.8	49	7	59	6.5	67	9
72	9.4	64	10.4	66	7	54	8.7
64	7.7	67	7.9	76	8.1	62	6.7
38	8.5	48	6	79	6.3	53	8.5
63	8	53	7	62	7.3	/	/

l'échantillon est constitué des malades diabétiques de tranche d'âge entre 40 ans et 80 ans.

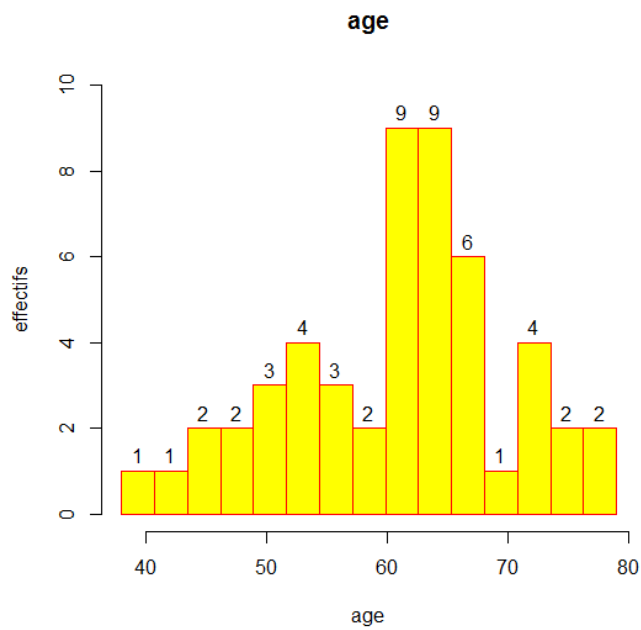


FIGURE 3.1: histogramme selon l'âge.

La distribution de dosage HbAc1 des deux trimestres :

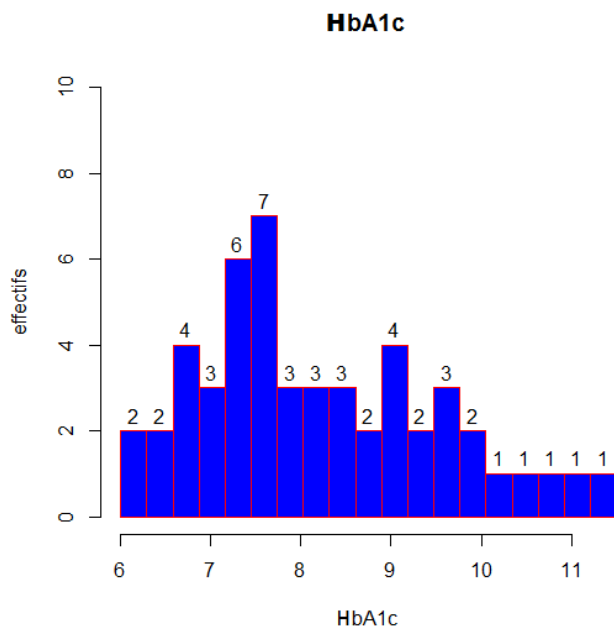


FIGURE 3.2: histogramme selon l'HbAc1 de premier trimestre.

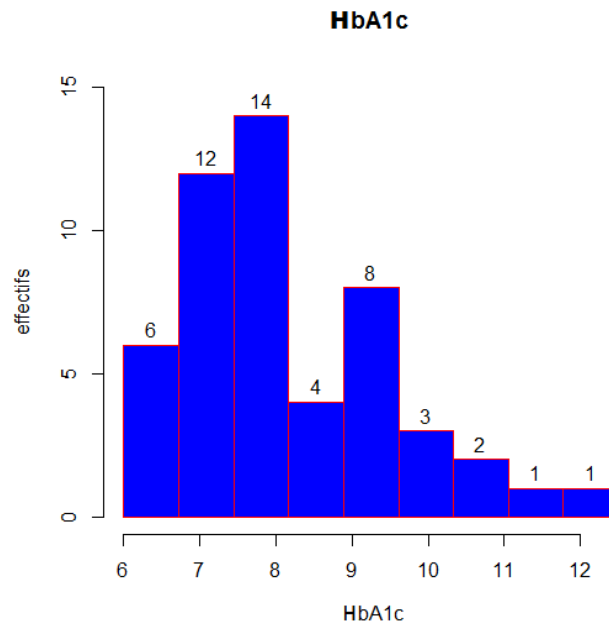


FIGURE 3.3: histogramme selon l'HbAc1 de deuxième trimestre.

On remarque qu'il y a un changement remarquable de la distribution de (HbA1c) dans les deux périodes du temps.

Pour avoir une idée sur le comportement des patients envers leurs maladie, on envisage de comparer les distributions dans les deux périodes de temps (trimestre 1 et trimestre 2)

**Remarque 3.5.** Pour utiliser la méthode B-spline, nous avons besoin de trouver  $df$  par la méthode de validation croisée suivante :

```
Call:
smooth.spline(x = age, y = AbA1c, cv = T)

Smoothing Parameter spar= 1.499804 lambda= 2275.085 (29 iterations)
Equivalent Degrees of Freedom (Df): 2.000042
Penalized Criterion (RSS): 45.89254
PRESS(1.o.o. CV): 1.949115
```

FIGURE 3.4: validation croisée B-spline pour premier trimestre.

Les graphes des trois estimateurs pour le première trimestre sont présentés dans la figure (3,5) :

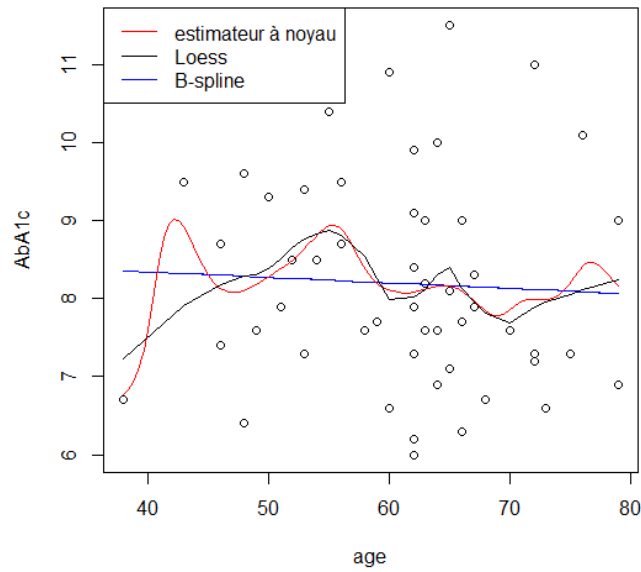


FIGURE 3.5: Les estimateurs à noyau, Loess et B-spline de premier trimestre.

On remarque que les trois méthodes donnent presque le même estimateur. Pour comparer ces trois méthodes, on calcul leurs erreurs  $MSE$  :

$MSE$ Noyau	$MSE$ Loess	$MSE$ B-spline
1.303716	1.277445	1.322461

On remarque que l'erreur associée à la méthode Loess est la plus petite, donc, on conclue que cette dernière est considérée comme optimale par contre la méthode spline est un mauvais choix dans ce cas.

Le degré de liberté de la matrice  $L_s$  peut trouver par la méthode de validation croisée. Le résultat de la méthode de validation croisée pour la méthode B-splines du deuxième trimestre est donné comme suit :

```

Call:
smooth.spline(x = age, y = AbA1c, cv = T)

Smoothing Parameter spar= 1.499804 lambda= 2275.085 (29 iterations)
Equivalent Degrees of Freedom (Df): 2.000042
Penalized Criterion (RSS): 45.89254
PRESS(1.o.o. CV): 1.949115
la validation croisée avec des valeurs de 'x' non uniques paraît douteuse
> fit2

```

FIGURE 3.6: validation croisée

Les graphes des trois estimateurs pour le deuxième trimestre sont présentés dans la figure (3,6) :

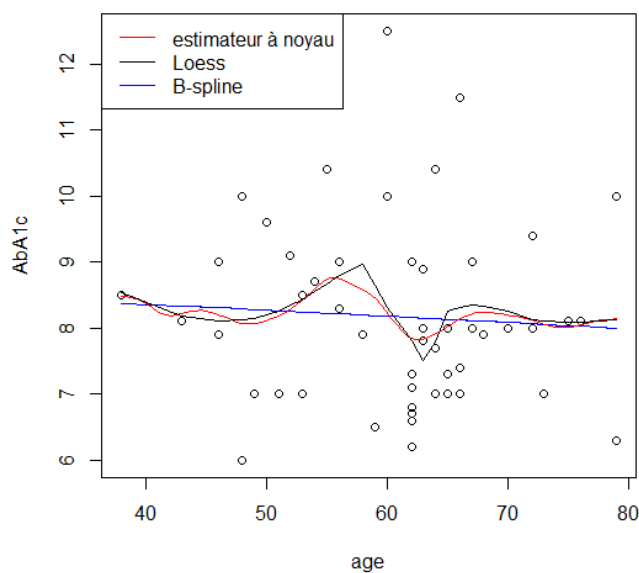


FIGURE 3.7: Les estimateurs à noyau, Loess et B-spline de deuxième trimestre.

Les erreurs de ces trois méthodes sont données comme suit :

$MSE$ Noyau	$MSE$ Loess	$MSE$ B-spline
1.320931	1.318609	1.323155

On trouve les mêmes résultats que le premier trimestre. Donc la méthode optimale pour nos données est toujours la méthode Loess.

Et pour qu'on bien observé les résultats de l'estimateur optimale Loess selon les deux trimestres on va présenté le graphe suivant :

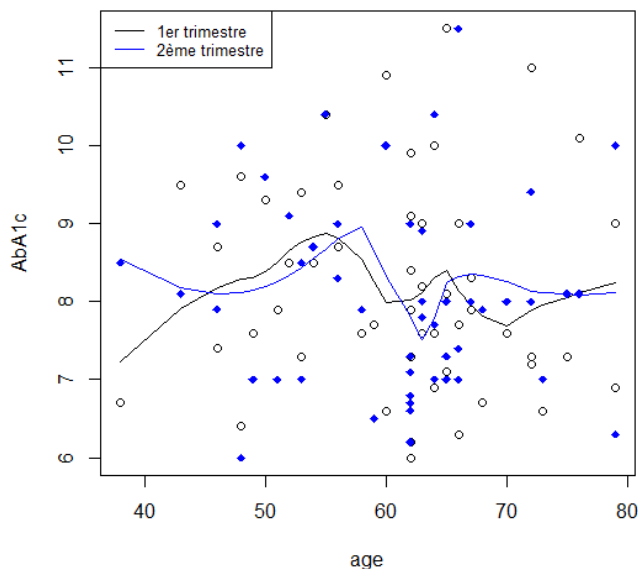


FIGURE 3.8: L'estimateur Loess pour le premier trimestre et le deuxième trimestre.

On remarque qu'il n'y a pas un grand changement selon les deux trimestres. **Discussion des résultat :** À partir de cette étude comparatif, on conclue que le meilleur estimateur dans le problème étudié est l'estimateur Loess mais on ne peut pas généralisé l'optimalité de cette estimateur pour d'autres données.

#### Le code R utilisé :

```
plot(age,HbA1c,type="p")
Ozone.spl <- smooth.spline(age,HbA1c)
Ozone.spl
lines(Ozone.spl, col = "blue")
lines(smooth.spline(age,HbA1c, df=2), lty=2,col = "blue")
lines(Ozone.ker,col="red")
Ozone.ker<-ksmooth(age,HbA1c,kernel="normal",bandwidth=5)
Ozone.ker

lines(Ozone.ker,col="red")
```

```
polyloc=loess(V age,span=0.5,family="gaussian")
pred_app=predict(polyloc,data.frame(age=sort(unique(age))))
lines(sort(unique(age)),pred_app,col="black")
#erreur spline
pred.app.spl=predict(Ozone.spl,age)
sqrt(mean((pred.app.spl$y-age)**2))

#erreur noyau
noyau=function(z,X,Y,h)
#z est la valeur en laquelle on calcule
#l'estimateur, h la fenetre,
#(X,Y) l'échantillon d'apprentissage.
Vect=exp(-((X-z)**2)/(2*h**2)) noyau=(sum(Y*Vect))/(sum(Vect))

m=length(age)
m
pred.ker.app=c(rep(0,m))
for (j in 1 :m)
pred.ker.app[j]=noyau(age[j],age,HbA1c,5)
sqrt(mean((pred.ker.app-HbA1c)**2))
#erreur loess

pred.app.loess=predict(polyloc,age)
sqrt(mean((pred.app.loess-HbA1c)**2))
a<-unique(age)
fit2<-smooth.spline(x=age,y=HbA1c,cv=T)
fit2
```

---

## CONCLUSION

Ce mémoire porte sur l'étude des estimateurs de la fonction de régression non paramétrique. On a présenté trois méthodes permettant d'effectuer de la régression non paramétrique : à noyau, Loess (par polynômes locales et par polynômes locales pondérées) et B-spline.

L'estimateur à noyau de la régression non paramétrique dépend de deux paramètres le noyau  $K$  et le paramètre de lissage  $h$ . À travers les résultats obtenus, nous concluons que : le noyau  $K$  est peu influent sur l'estimateur, par contre le paramètre  $h$  joue un rôle très important et son choix est crucial. On trouve plusieurs méthodes pour le choix de  $h$  optimale et nous avons choisi celles qui sont plus utilisées : celle minimisant les erreurs  $MSE$  et  $MISE$  ainsi que la méthode de validation croisée.

On conclue que ces méthodes numérique sont développé de tel sortes, qu'on peut toujours trouver le meilleur estimateur selon les données à main. De plus la méthode Spline nous donne la chance de faire l'inférence statistique. Et La méthode Loess (par polynômes locales et par des polynômes locaux pondérées) est très simple à implémenter et d'appliquer.



---

# BIBLIOGRAPHIE

- [1] Amroun, S, 2011. *Sur L'Estimation De La Courbe De Régression De La Moyenne*, Mémoire De Magister en Mathématiques Appliquées université de Bêjaïa.
- [2] Arnak, S. *Statistique avancée : méthodes non paramétriques*. École centrale paris de Paris.
- [3] Beghriche, H,. *L'estimation de la Fonction de Régression*, Mémoire de magistère en mathématique université de Mentouri-Constantine.
- [4] Benchoulak, H. 2012. *Bandes de confiance pour les fonctions de densité et de régression*, Université Mentouri Constantine.
- [5] Ben Khalifa, I, 2007. *Estimation non-paramétrique par noyaux associés et données de panel en marketing*, université du 7 Novembre à Carthage. Tunisie.
- [6] Blondin, D, 2004 *Lois limites uniformes et estimation non-paramétrique de la régression*. Thèse Doctorat, Université Paris 6, 1-26.
- [7] Cleveland, W. S. et Devlin, S. J, 1988. *Locally weighted regression : an approach to regression analysis by local fitting*. *Journal of the American statistical association*, vol.83, no 403, p. 596-610.
- [8] Cleveland, W. S. (1979), *Robust locally weighted regression and smoothing scatterplots*. *Journal of the American statistical association*, 74(368), 829-836.
- [9] Devroye, L. Györfi, L. Krzyzak, A. et Lugosi, G. 1994. *On the strong universal consistency of nearest neighbor regression function estimates*. *The Annals of Statistics*, 22(3), 1371-1385.
- [10] Devroye, L. and Györfi, L, (1985). *Nonparametric Density Estimation*. New York.

- [11] Eubank, R. L, 1999 *Nonparametric regression and spline smoothing*. CRC press.
- [12] Fan. J. *Local linear regression smoother and their minimax*.
- [13] Ferraty, F. et Vieu. P, (2000). Dimension fractale et estimation de la régression dans des espaces vectoriels semi-normés. *C. R . Acad. Sci., Paris. 330, No. 2, 139-142*.
- [14] Ferraty, F et Vieu, P. 2002, Modèles Non paramétrique de Régression.
- [15] W. Härdle, Applied Nonparametric Regression. *université de Berlin, 1990*.
- [16] HURLIN, C. Régressions non paramétriques univariées. *Master Econométrie et Statistique Appliquée (ESA), Université d'Orléans, 2007, vol. 2008*.
- [17] Kebabi, K, 2014. Estimation non-paramétrique de la fonction de régression : cas d'un modèle de censure mixte. *Université Constantine 1*.
- [18] Kiwitt, S, N. Neumeier. Estimating the Conditional Error Distribution in Non-parametric Regression. *Scandinavian Journal of Statistics, 2012, vol. 39, no 2, p. 259-281*.
- [19] Lejeune. M, 1984 Optimization in non-parametric regression. *In Compstat 1984 (pp. 421-426). Physica, Heidelberg. efficiencies. Annals of statistics 21, 196-216, 1993*.
- [20] Lejeune. M, (2011). *Statistique. La théorie et ses applications : La théorie et ses applications*. Springer Science & Business Media.
- [21] Mallows. C. L. 1973 *Some Comments on  $C_p$* , *Technometrics*, 15 :4, 661-675.
- [22] Nadaraya, E. A. 1976. *On the nonparametric estimator of Bayesian risk in the classification problem*. Proc. AN. Georg SSR, 82(2), 277-280 (in Russian).
- [23] Roy. V, 2007. *Régression non paramétrique des percentiles pour données censurées*, Mémoire de Maitre des sciences présenté à l'Université Laval.
- [24] Simonoff. J. S, 1996. *Smoothing Methods in Statistics*. Springer Series in Statistics. Springer-Verlag, New York.
- [25] Silverman, B. W. 1986. *Density Estimation for Statistics and Data Analysis*. Chapman Hall, London.
- [26] Vandal, N. 2005, *La régression non paramétrique multidimensionnelle Théorie et application à une étude portant sur la densité mammaire*, Mémoire de Maitre des sciences, l'Université Laval.
- [27] Cours apprentissage non paramétrique en régression. <https://www.math.univ-toulouse.fr/besse/Wikistat/pdf/st-m-app-non-param.pdf>
- [28] [https://fr.wikipedia.org/wiki/Hémoglobine\\_glyquée](https://fr.wikipedia.org/wiki/Hémoglobine_glyquée).

---

---

# ANNEXE A

---

## LES CONVERGENCES

### A.1 Convergence en loi

**Définition A.1.** Soit  $(X_n)_n$  une suite de variables aléatoires réelles définies sur un même espace de probabilité  $(\Omega, \mathcal{A}, \mathcal{P})$ . On dit que la suite  $(X_n)_n$  de fonction de répartition  $F_n$ , converge en loi vers une v.a  $X$  de fonction de répartition  $F$ , si la suite  $F_n(x)$  converge vers  $F(x)$  en tout point  $x$  où  $F$  est continue :

$$X_n \xrightarrow{l} X, \text{ quand } n \rightarrow \infty$$

### A.2 Convergence en probabilité :

**Définition A.2.** On dit que  $X_n$  converge vers  $X$  en probabilité

$$\text{si, } \forall \epsilon > 0 \lim_{n \rightarrow \infty} (\mathbb{P}|X_n - X| \geq \epsilon) = 0$$

On note aussi :  $X_n \xrightarrow{P} X$

### A.3 Convergence presque sûre :

**Définition A.3.** On dit que  $X_n$  converge presque sûrement vers  $X$  si :

$\mathbb{P}(\lim_{n \rightarrow \infty} (X_n = X)) = 1$ . Ou de manière équivalente, s'il existe un sous-ensemble  $\mathbb{P}$ -négligeable  $N$  dans  $\Omega$  tel que :  $\forall \omega \in \Omega \setminus N$ , on a  $(X_n)(\omega) \xrightarrow{P.s.} X(\omega)$  On parle aussi de convergence presque partout ou avec probabilité 1 ou forte, et on écrit :  $X_n \xrightarrow{P.s.} X$

### A.4 Convergence presque complète :

**Définition A.4.** La convergence presque complète implique la convergence presque sûre et se prête bien aux calculs faisant intervenir des sommes de variables aléatoires. Malgré cela, elle ne commence à devenir populaire dans la communauté statistique. Elle est utilisée surtout en statistique non-paramétrique. On dit que la suite de variables aléatoires réelles  $(X_n)_{n \in \mathbb{N}}$  converge presque complètement vers une variable aléatoire  $X$  lorsque  $n \rightarrow \infty$  si :

$$\forall \epsilon > 0, \sum_{n \in \mathbb{N}} \mathbb{P}[|X_n - X| > \epsilon] < \infty$$

et on dit que la vitesse de convergence presque complète de la suite de variables aléatoires réelles  $(X_n)_{n \in \mathbb{N}}$  vers  $X$  est d'ordre  $(u_n)$  si :

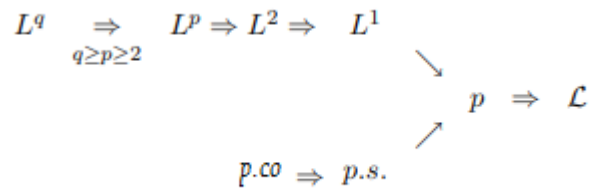
$$\forall \epsilon_0 > 0, \sum_{n \in \mathbb{N}} \mathbb{P}[|X_n - X| > \epsilon_0 u_n] < \infty$$

Cette définition du taux a été introduite par Ferraty et Vieu (2006). Elle a l'avantage théorique d'impliquer les deux vitesses de convergence classiques en probabilité et presque sûre, et l'avantage pratique d'être souvent plus facile à démontrer. Dans les quinze dernières années, ce mode de convergence a été très utilisé dans des travaux concernant la statistique non-paramétrique des données fonctionnelles.

**Remarque A.1.** La convergence presque complète est plus forte que la convergence presque sûre et la convergence en probabilité.

## A.5 Liens entre les convergences :

*On donne la relation entre les différentes modes de convergence :*



## A.6 Notations $O(h)$ et $o(h)$

La définition des notations standards pour la convergence des séries se lit de la façon qui suit :

1- Pour les variables non aléatoires :

Soit  $x$  et  $h$  deux séries de nombres réels. Alors, lorsque  $n \rightarrow \infty$ ,

(a)  $x_n = O(h) \Leftrightarrow \limsup_{n \rightarrow \infty} |x_n/h| < \infty$

(b)  $x_n = o(h) \Leftrightarrow \limsup_{n \rightarrow \infty} |x_n/h| = 0$

2- Pour les variables aléatoires : Soit  $x$  et  $h$  deux séries de nombres réels. Alors, lorsque  $n \rightarrow \infty$ ,

(a)  $x_n = O_s(h) \Leftrightarrow \forall \epsilon > 0, \exists \delta$  et  $N$  tels que  $\mathbb{P}(|x_n/h| > \sigma) < \epsilon, \forall n > N$

(b)  $x_n = o_s(h) \Leftrightarrow \forall \epsilon > 0, \lim \mathbb{P}(|x_n/h| > \sigma) = 0,$

---

---

# ANNEXE B

---

## LES THÉORÈME UTILISÉS :

### B.1 Théorème de Bochner

**Théorème B.1.** *Bochner [3] soit  $K : (\mathbb{R}^m, \mathcal{B}^m) \rightarrow (\mathbb{R}, \mathcal{B})$  une fonction mesurable, où  $\mathcal{B}^p$  est la tribu borélienne de  $\mathbb{R}^p$ , vérifiant :  $\exists M$  (constante) telle-que,*

$$\forall z \in \mathbb{R}^m, |K(z)| \leq M$$
$$\int_{\mathbb{R}^m} |K(z)| dz < \infty$$

*Et*

$$\|z\|^m |K(z)| \rightarrow 0 : \text{quand } \|z\| \rightarrow \infty$$

*Par ailleurs, soit  $g : (\mathbb{R}^m, \beta) \rightarrow (\mathbb{R}, \beta)$  une fonction tq*

$$\int_{\mathbb{R}^m} |g(z)| dz < \infty$$

*Si  $g$  est continue, et si  $0 < h \rightarrow 0$ , quand  $n \rightarrow \infty$  alors :*

$$\lim_{n \rightarrow \infty} \frac{1}{h^m} \int_{\mathbb{R}^m} K\left(\frac{z}{h}\right) g(x - z) dz = g(x) \int_{\mathbb{R}^m} K(z) dz$$

*Si  $g$  est uniformément continue alors la convergence ci dessus est uniforme.*

## B.2 Théorème d'Inégalité Cauchy Shwartz

**Théorème B.2.** *Soit  $(H, \langle \cdot, \cdot \rangle)$  une espace préhilbertien. Alors*

$$|\langle \mu, \theta \rangle|^2 \leq \langle \mu, \mu \rangle \langle \theta, \theta \rangle \quad \forall \mu, \theta \in H$$

*ou bien*

$$|\langle \mu, \theta \rangle| \leq \sqrt{\langle \mu, \mu \rangle} \cdot \sqrt{\langle \theta, \theta \rangle}$$

## B.3 Théorème de Fubini

**Théorème B.3.** *Si  $f(x, y)$  est une fonction Bochner-intégrable sur  $X \times Y$ , Alors la fonction  $\int f dy$  (de  $x$ ) est déterminé presque partout sur  $X$  est Bochner-intégrable sur  $X$ . De même, la fonction  $\int f dx$  (de  $y$ ) est déterminé presque partout sur  $Y$ . En outre,*

$$\int \int f dx dy = \int dx \int f dy = \int dy \int f dx.$$