République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université A. Mira de Béjaïa
Faculté des Sciences Exactes
Département de Recherche Opérationnelle

Mémoire De Magister

En Mathématiques Appliquées

Option

Modélisation Mathématique et Techniques de Décision

Thème

Sur L'Estimation De La Courbe De Régression De La Moyenne

Présenté par :

M^{elle} Sonia AMROUN

Devant le jury composé de :

Président	**M^{r}**	**D. AISSANI**	**Professeur**	**U. de Béjaïa**
Rapporteur	**M^{r}**	**S. ADJABI**	**M. C. A**	**U. de Béjaïa**
Examinateur	**M^{r}**	**M. O. BIBI**	**Professeur**	**U. de Béjaïa**
Examinateur	**M^{r}**	**Z. MOHDEB**	**Professeur**	**U. de Constantine**
Invitée	**M^{me}**	**K. CHERFI-LAGHA**	**M. C. B**	**U. de Béjaïa**

Béjaïa, 2011

Remerciements

En premier lieu, je remercie Dieu, le miséricordieux, sans lui rien de tout cela n'aurait pu être.

Je tiens tout d'abord à exprimer ma profonde gratitude à M^{r} S. ADJABI Maître de Conférences à l'université A. MIRA de Béjaïa, je le remercie pour son soutien et pour la confiance qu'il m'a accordée tout au long de ce travail. Son aide et ses judicieux conseils m'ont été indispensables pendant toute la durée de la progression de ce mémoire.

J'adresse également mes remerciements à M^{r} D. AISSANI Professeur à l'université A. MIRA de Béjaïa qui ma fait l'honneur d'accepter de présider le jury de ma soutenance.

Je remercie l'ensemble des membres du jury M^{r} M. O. BIBI Professeur à l'université A. MIRA de Béjaïa, M^{r} Z. MOHDEB Professeur à l'université de Constantine de me faire l'honneur d'avoir accepté d'examiner mon travail, ainsi que M^{me} K. CHERFI-LAGHA Maître de Conférences à l'université A. MIRA de Béjaïa qui a accepté l'invitation de participer au jury.

Je désire remercier les membres de ma famille qui ont veillé sur moi et qui ont su m'accompagner, mes amis qui m'ont soutenu tout au long de l'accomplissement de ce travail.

J'aimerais remercier toutes les personnes ayant contribué à la réalisation de ce mémoire, que ce soit par leur implication directe ou par leurs encouragements.

Dédicaces

Je dédie ce modeste travail :

- **A** mes chers parents.
- **A** mes frères et sœurs.
- **A** mes amis.

Table des matières

Table des figures

Liste des tableaux

Introduction Générale

La théorie de l'estimation est une des branches les plus basiques de la statistique. Cette théorie est habituellement divisée en deux composantes principales, à savoir, l'estimation paramétrique et l'estimation non paramétrique. Le problème de l'estimation non paramétrique consiste, dans la majeur partie des cas, à estimer, à partir des observations, une fonction inconnue élément d'une certaine classe fonctionnelle. Plus particulièrement, on parle d'estimation non paramétrique lorsque celle-ci ne se ramène pas à l'estimation d'un nombre fini de paramètres réels associés à la loi de l'échantillon.

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 $19^{\text{ème}}$ siecle. 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.

Soit les observations suivantes $(x_1, y_1), \cdots, (x_n, y_n)$. Le modèle de régression étudié dans ce travail est de la forme :

$$y_i = m(x_i) + \epsilon_i, \quad i = 1, \cdots, n,$$

où m est la fonction de régression que l'on cherche à estimer, et les $(\epsilon_i)_{i=1}^n$ sont des variables aléatoires supposées indépendantes identiquement distribuées de moyenne nulle et de variance commune σ^2. Ces erreurs représentent les erreurs de mesure, et plus généralement l'ensemble de la variabilité de Y non expliquée par X.

Les premiers travaux sur ce sujet datent des années 50. La première application est l'estimation de la fonction de densité par des méthodes d'opérateur à noyau (kernel) avec les travaux fondateurs de [Rosenblatt 1956][52] et de [Parzen 1962][46]. Ces premiers travaux ont été étendus à la notion de régression kernel, traduit en français par le terme de régression avec lissage par opérateur à noyau et maintenant connu sous le nom de régression par la méthode du noyau. Dans ce domaine, on identifie deux papiers fondateurs publiés la même année : [Nadaraya (1964)][44] et [Watson (1964)][77]. L'estimateur à noyau proposé par Nadaraya-Watson est construit à partir d'une fonction noyau K et d'une fenêtre h, sous la forme :

$$\hat{m}_h(x) = \frac{\sum_{i=1}^{n} y_i K(\frac{x-x_i}{h})}{\sum_{i=1}^{n} K(\frac{x-x_i}{n})},$$

qui est un estimateur linéaire par rapport à Y. [Collomb (1977) [10], Collomb (1976) [11] et (1977a) Collomb [12]] a étudié les propriétés de convergence asymptotiquement normal, il a donné les résultats sur le biais et la variance de l'estimateur (limite et évaluation asymptotique). On peut également voir : [Collomb (1977b)][12], [Ferraty et Vieu (2002)][22] et [Sarda et Vieu (2000)][59] sur les propriétés : uniforme, presque sûre, presque complète et en moyenne quadratique de l'estimateur $\hat{m}_h$. Stone (1981) [61], Stone(1982) [61] et [62] a traité la vitesse optimale de convergence sous les conditions de régularités de m autour d'un compact.

En ce qui concerne le choix du paramètre de lissage, [Härdle et Marron (1985)][28] ont donné l'optimalité asymptotique de la largeur de fenêtre en termes d'erreur quadratique (voir aussi les revus bibliographiques de [Marron (1988)][41] et [Jones et al. (1996)][34]). Loader (1999) [35] a proposé une étude comparative entre plusieurs méthodes de choix du paramètre de lissage.

Concernant les travaux sur le choix du noyau, [Berlinet (1993)][2] a proposé une méthode automatique du choix du noyau. (voir [Vieu (1999)][66] pour des résultats et références récentes.)

Malgré ses bonnes propriétés asymptotiques, la méthode du noyau n'est pas parfaite

et certains problèmes y sont rattachés. Particulièrement, le paramètre de lissage h lié au problème du manque de variation locale du lissage est aussi lié à la méthode du noyau. En effet, étant donné qu'une seule fenêtre de lissage est utilisée, l'estimateur du noyau ne tient pas compte des différents niveaux de lissage des parties de la fonction. En fait, pour réduire le MSE (Mean Square Error), h devrait augmenter avec $m(x)$, ce qui réduirait la variance, et diminuer avec $|m''(x)|$, ce qui réduirait le biais. D'un point de vue pratique, ce manque d'adaptation se manifeste par du sur-lissage ou par du sous-lissage (Figure 1).

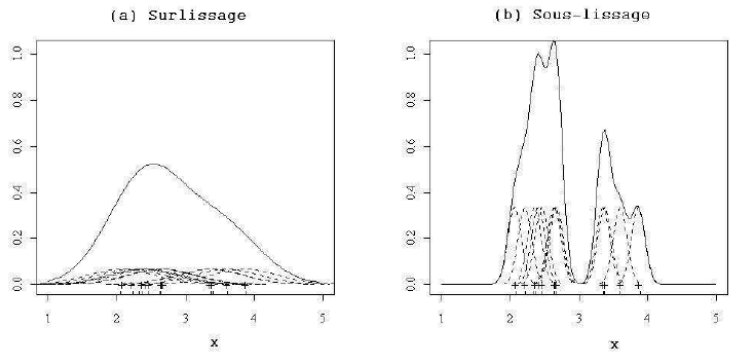

FIGURE 1 – Graphiques de sur-lissage et de sous-lissage.

Une méthode alternative à cette méthode du noyau, est la méthode des fonctions splines. C'est une méthode d'analyse numérique, elle est utilisée dans le but de l'interpolation. Le mot "spline" en anglais signifie (languette élastique), où l'on s'intéresse à la courbe décrite par une languette forcée de passer par un nombre fini de points donnés (x_i, y_i), $i = 1, \cdots, n$. Introduite pour la première fois par [Whittaker (1923)][78] et développée par [Schoenberg (1964)][55] pour servir au calcul scientifique (approximation,...). Actuellement, la méthode des splines de lissage est appliquée pour la représentation de courbes et surfaces en Computer Graphics.

Les fonctions splines sont devenues très populaires, elles trouvent leurs applications dans différents domaines comme l'analyse des données de croissance, en médecine et en économie. Des développement et des applications de la méthode des splines de lissage peuvent être trouvées dans [Wahba (1990)][73] et [Green and Silverman (1994)][25]. Voir aussi [Reinsch (1967)][49], [Silverman (1985)][58], [Eubank (1999)][20] ou [Green and Silverman (1994)][25].

L'estimateur par la méthode des splines de lissage est définit comme étant le fonction

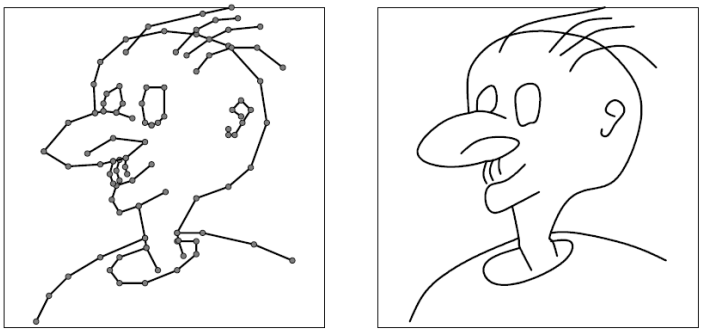

FIGURE 2 – Un dessin en zig-zag (à gauche) et en splines (à droite)

$\hat{m}_\lambda$ qui réalise la solution d'un problème de minimisation, qui s'écrit sous la forme

$$\hat{m}_\lambda = A_\lambda Y,$$

où A_λ est une matrice de lissage qui ne dépend que des $x_i,\ i = 1, \cdots, n$.

Cet estimateur est linéaire par rapport au vecteur réponse Y. [Schultz (1970)][56] a étudié les propriétés de convergence des splines interpolantes et ses dérivées. Wahba (1973) (1974)][68] et [69] a donné les conditions pour obtenir la convergence en moyenne quadratique et [wahba (1976)][71] a traité l'erreur moyenne quadratique. Concernant le degré r de la spline et la selection du paramètre de lissage λ voir [Graven and Wahba (1979)][15] et voir aussi [Wahba et Wold (1975)][75] pour des résultats de ce type. Pour une étude détaillée de cet estimateur voir la monographie de [Eubank (1988)][19].

Dans ce mémoire nous exposerons en détail les deux méthodes d'estimation de la courbe de régression de la moyenne : la méthode du noyau et la méthodes des fonctions splines. Il est 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 patrie du chapitre propose des méthodes d'estimations de la courbe de régression : la méthode des k plus proches voisins, méthode des séries orthogonales et plus particulièrement la méthode du noyau, les propriétés statistiques et asymptotiques de l'estimateur sont données.

- Dans le deuxième chapitre, nous présentons la méthode des fonctions splines dans un cas général et son application à l'estimation de la courbe de régression de la moyenne, nous donnons les propriétés statistiques (biais, variance) de l'estimateur et ses quelques propriétés asymptotiques.
- Le troisième chapitre présente les résultats obtenus par simulation, effectuée sur 3 exemples de fonctions cibles et d'un cas réel, pour comparer les deux méthodes (noyau et spline). Le critère utilisé est celui de MISE (Mean Integrated Square Error).

Ce mémoire se termine par une conclusion générale et quelques perspectives.

1

Régression non paramétrique

1.1 Introduction

L'utilisation des modèles paramétrique est très fréquente lorsque l'on fait appel à la régression afin d'analyser un jeu de données. Or, il y a certaines situations où ces modèles ne sont pas appropriés et où le choix d'un modèle non paramétrique est préférable. Dans ce chapitre, nous étudierons certaines méthodes de régression non paramétrique, à savoir, la méthode des plus proches voisins, la méthode des séries orthogonales et particulièrement la méthode du noyau, qui est une méthode très pratique lorsque l'on s'intéresse à la relation entre une variable réponse Y et une variable explicative X, mais que l'on ne veut supposer aucune forme particulière pour la relation entre ces deux variables, laissant ainsi aux données le choix exclusif de cette forme.

1.2 Principe d'estimation non paramétrique

Lorsque nous souhaitons décrire l'influence d'une variable quantitative sur un événement en faisant le moins d'hypothèses possible sur la forme de la relation, nous distinguons deux approches :

- L'approche de la régression paramétrique,
- L'approche de la régression non paramétrique.

Le but d'un modèle de régression consiste à déterminer la façon dont l'espérance d'une variable dépendante Y dépend d'un ensemble de variables explicatives X. Supposons que $X \in \mathbb{R}$ le problème consiste donc à déterminer pour chaque réalisation x de la variable X ; la valeur de la fonction $m(x)$, dite fonction de lien ou fonction de régression.

Définition 1.1. On appelle fonction de régression, la fonction $m(x)$ qui a pour toute réalisation x de la variable explicative X associe la quantité :

$$\mathbb{E}(Y|X = x) = m(x). \tag{1.1}$$

1) Modèle paramétrique

Pour caractériser cette fonction de régression, la première approche consiste à utiliser un modèle de régression paramétrique. Nous supposons que cette fonction peut s'écrire comme une fonction explicite des valeurs de X. Cette fonction peut être linéaire, logarithmique, non-linéaire etc. Par exemple, dans le cas linéaire on suppose que :

$$\mathbb{E}(Y|X = x) = \alpha + \beta x.$$

Nous cherchons alors à déterminer les meilleures valeurs de α et β compte tenu d'un critère, par exemple celui de la MSE (Mean Square Error).

Définition 1.2. Dans un modèle de régression paramétrique, la fonction de régression est :

- De forme explicite.
- Peut s'écrire en fonction d'un nombre réduit de paramètres.

Exemple

$$\mathbb{E}(Y|X = x) = m(x, \theta),$$

où $m(.)$ est connue avec $\theta \in \mathbb{R}^k$
L'exemple typique est celui d'un modèle linéaire, où l'on suppose que :

$$\mathbb{E}(Y|X = x) = \alpha + \beta x = m(\alpha, \beta, x).$$

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.

Définition 1.3. Dans un modèle de régression non paramétrique, la fonction de régression
- N'a pas de forme explicite.
- Ne peut pas s'écrire en fonction d'un nombre réduit de paramètres.

$$\mathbb{E}(Y|X = x) = m(x).$$

Le principal avantage de cette approche est qu'elle ne nécessite aucune hypothèse a priori sur la forme du lien entre X et Y.

Remarque 1.1. *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 $m(x)$.*

Tout le problème consiste alors à estimer cette fonction de régression $m(x)$, qui est a priori inconnue.
Le modèle qui va nous intéresser est un modèle non paramétrique, en ce sens que la seule condition que nous ferons sur la fonction m est une condition de régularité

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

où

$$m \text{ est } k \text{ fois continûment dérivable,}$$

k étant un entier positif ou nul (le cas $k = 0$ correspondant évidemment à l'hypothèse de continuité de m). Nous allons donner quelques définitions théoriques des notions utilisées

Définition 1.4. On dit qu'un estimateur $\hat{m}$ de m est sans biais si :

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

Définition 1.5. On dit qu'un estimateur $\hat{m}$ de m est asymptotiquement sans biais si :

$$\lim_{n\to\infty} \mathbb{E}(\hat{m}) = m.$$

Définition 1.6. Un estimateur $\hat{m}$ de m est dit asymptotiquement uniformément sans biais si :

$$\lim_{n\to\infty} \sup_x |\mathbb{E}[\hat{m}(x) - m(x)]| = 0.$$

Définition 1.7. L'erreur moyenne quadratique MSE :

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

Définition 1.8. L'erreur moyenne quadratique intégrée MISE :

$$\begin{aligned}
MISE(\hat{m}, m) &= \int MSE(m(x), \hat{m}(x))\ dx \\
&= \int Biais(\hat{m}(x))^2\ dx\ + \int \mathbb{V}ar(\hat{m}(x))\ dx.
\end{aligned}$$

Définition 1.9. On dit qu'un estimateur $\hat{m}$ de m est ponctuellement consistant en moyenne quadratique si :

$$\lim_{n\to\infty} MSE(m(x), \hat{m}(x)) = 0.$$

Définition 1.10. On dit qu'un estimateur $\hat{m}$ de m est uniformément consistant en moyenne quadratique intégrée si :

$$\lim_{n\to\infty} MISE(m(x), \hat{m}(x)) = 0.$$

Définition 1.11. On dit qu'un estimateur $\hat{m}$ de m est asymptotiquement normal si :

$$\hat{m} \longrightarrow \mathcal{N}(\mathbb{E}(\hat{m}), \mathbb{V}ar(\hat{m})), \text{ en loi.}$$

1.3 Lien entre la régression et la minimisation d'une espérance conditionnelle

Soient les observations bivariées $(x_i, y_i),\ i = 1, \cdots, n$, où les x_i représentent les valeurs observées de la variable explicative X et les y_i représentent celles de la variable dépendante Y. La méthode la plus communément utilisée pour étudier la relation entre ces deux variables est la régression linéaire simple, qui suppose un modèle de la forme

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i,\ i = 1, \cdots, n,$$

où les erreurs aléatoires ϵ_i sont non corrélées, de moyenne nulle et de variance σ^2. Cette méthode possède l'avantage d'être facile à interpréter et, lorsque les postulats sur les résidus ϵ_i sont vérifiés, elle permet de faire des tests d'hypothèses statistiques formels sur les paramètres. Par contre, il arrive que la linéarité de la relation ne soit pas toujours respectée. Dans ce cas, il est préférable de choisir un modèle plus flexible qui reflète mieux la relation entre X et Y .

Le modèle de régression non paramétrique suivant peut alors être employé :

$$y_i = m(x_i) + \epsilon_i, \qquad i = 1, \cdots, n,$$

où $m(x_i)$ représente la moyenne conditionnelle de la courbe de régression, c'est-à-dire

$$m(x) = \mathbb{E}(Y|X = x),$$

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 m de régression

$$m :\ \mathbb{R}^2 \longrightarrow \mathbb{R},$$

définie mathématiquement par le fait que $m(X_i)$ est la variable aléatoire fonction de X_i qui approxime le mieux Y_i en moyenne :

$$m = \arg\min_{a(X_i)}\ \mathbb{E}(Y_i - a(X_i))^2.$$

Les résidus ϵ_i représentent la variation de Y autour de $m(x)$. Les postulats sur les termes d'erreur ϵ_i sont les mêmes que ceux du modèle linéaire et, à part certaines

hypothèses de continuité et de lissage [Simonoff][59], il n'y a habituellement aucune contrainte paramétrique ni sur la loi de (X, Y), ni sur la loi conditionnelle, ni sur la forme de $m(x)$, c'est-à-dire pour être claire, nous n'écrivons pas a priori par exemple $m(x) = \beta x$ (régression linéaire).
Il serait intéressant de faire le parallèle avec la définition formelle de la moyenne conditionnelle d'une variable aléatoire, puisque qu'elle fournit une nouvelle expression pour la moyenne conditionnelle de la courbe de régression. A cette fin, la démonstration de l'égalité de ces deux moyennes conditionnelles doit être présentée. Ainsi, soit la définition

$$\begin{aligned} m(x) &= \arg\min_a \mathbb{E}[(Y-a)^2|X=x] && (1.3)\\ &= \mathbb{E}(Y|X=x). && (1.4)\end{aligned}$$

La preuve 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 :

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

$$\Rightarrow a = \mathbb{E}(Y|X=x).$$

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.

Remarque 1.2. *Les estimateurs que nous considérons appartiennent à la vaste classe des estimateurs linéaires (i.e. linéaires en tant que fonction des observations Y_i), cette classe des estimateurs linéaires regroupe la majorité des estimateurs de la régression.*

Définition 1.12. Un estimateur $\hat{m}(x)$ de $m(x)$ est dit **estimateur linéaire** de la régression non-paramétrique si

$$\hat{m}(x) = \sum_i^n y_i w_i(x),$$

où la fonction de poids $w_i(.)$ ne dépend pas des observations y_i.

Nous commencerons par présenter le principe de l'estimation par la méthode du noyau.

1.4 Méthode du noyau

1.4.1 Principe de la méthode

Le problème consiste à estimer la fonction de régression en tous points $x_1, x_2, \cdots, x_n$. Le principe de la méthode du noyau repose en fait sur des méthodes de lissage, elle donne pour estimateur de $\mathbb{E}(Y|X = x)$ 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 $m(x_0)$, doit être fait.

1.4.2 Construction de l'estimateur

Nous supposons que nous disposons des observations $(x_1, y_1), \cdots, (x_n, y_n)$ du couple (X, Y). On se propose de construire un estimateur $\hat{m}(x)$ de la fonction de régression à partir des couples d'observations. Le premier estimateur rencontré dans la littérature est l'estimateur à noyau de Nadaraye-Watson. Il est construit à partir d'une fonction $K(.)$ et d'une fenêtre h, de manière analogue à l'estimateur à noyau de la fonction de densité $f_X(.)$ introduit par [Parzen][46] et [Rosenblatt][51].
Nous rappelons la définition de l'estimateur Parzen et Rosenblatt :

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K(\frac{x - x_i}{h}) \quad \forall x \in \mathbb{R}.$$

Dans un premier temps, nous désignons par fenêtre une suite $\{h_n : \quad n \geq 1\}$ de nombres strictement positifs vérifiant

$$h_n \longrightarrow 0, \quad lorsque\ n \longrightarrow +\infty.$$

La fenêtre $h = h_n$ dénote une suite indexée par $n = 1, 2, \cdots$, mais la dépendance en n ne sera pas précisée afin d'alléger les notations.

La fonction $K : \mathbb{R} \longrightarrow \mathbb{R}$ sera supposée mesurable et satisfait certaines hypothèses basiques parmi celles énoncées ci-dessous :

1. K est bornée, i.e, $\sup_{u \in \mathbb{R}} |K(u)| < \infty$;
2. $\lim_{|u| \to \infty} |u|\ K(u) = 0$;
3. $K(.) \in L_1(\mathbb{R}),\ i.e,\ \int_{\mathbb{R}} |K(u)|\ du\ < \infty$;
4. $\int_{\mathbb{R}} K(u)\ du = 1$.

Nous reprenons le modèle

$$m(x) = \mathbb{E}(Y|X = x) = \frac{r(x)}{f_X(x)}.$$

Nous avons

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

où $F_{X,Y}(.,.)$ est la fonction de répartition du (X,Y).

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

Donc $m(x)$ est estimée par

$$\hat{m}(x) = \frac{\hat{r}(x)}{\hat{f}_X(x)} = \frac{\sum_{i=1}^n y_i\ 1(|x_i - x| \leq h)}{\sum_{i=1}^n 1(|x_i - x| \leq h)}.$$

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 :

$$\hat{m}(x) = \frac{\sum_{i=1}^n y_i\ K(\frac{x_i - x}{h})}{\sum_{i=1}^n K(\frac{x_i - x}{h})}, \quad \forall x.$$

Cet estimateur a été introduit par [Nadaraya][44] et [Watson][77].

$$\hat{m}(x) = \frac{\sum_{i=1}^n y_i\ K(\frac{x_i - x}{h})}{\sum_{i=1}^n K(\frac{x_i - x}{h})} \ \times\ 1\{\sum_{i=1}^{n} K(\frac{x_i - x}{h}) \neq 0\},$$

où $1\{.\} = 1_{\{.\}}$ désigne la fonction indicatrice.

$$\hat{m}(x) = \begin{cases} \frac{\sum_{i=1}^n Y_i\ K(\frac{X_i - x}{h})}{\sum_{i=1}^n K(\frac{X_i - x}{h})}, & \text{si } \{\sum_{i=1}^n K(\frac{X_i - x}{h}) \neq 0\}\,; \\ \frac{1}{n}\sum_{i=1}^n y_i, & \text{sinon.} \end{cases}$$

Le noyau K détermine la forme du voisinage autour du point x et la fenêtre h contrôle la taille de ce voisinage, i.e, le nombre d'observations prises pour effectuer la moyenne locale.

Remarque 1.3. L'estimateur de Nadaraya-Watson est bien linéaire au sens de la définition (1.12) avec comme fonction de poids

$$W_i = \frac{K(\frac{x_i - x}{h})}{\sum_{i=1}^{n} K(\frac{x_i - x}{h})} \quad 1\{\sum_{i=1}^{n} K(\frac{x_i - x}{h}) \neq 0\}.$$

Définition 1.13. L'estimateur à noyau (kernel estimate) introduit par [Watson, (1964), Nadaraya, (1964)][77] [44], de la fonction de régression évaluée au point x_0, noté $\hat{m}(x_0)$; est défini par :

$$\hat{m}(x_0) = \sum_{i=1}^{n} w_i(x_0) y_i \tag{1.5}$$

avec :

$$w_i(x_0) = \frac{K(\frac{x_i - x_0}{h})}{\sum_{i=1}^{n} K(\frac{x_i - x_0}{h})}. \tag{1.6}$$

où $K(.)$ désigne une fonction kernel, $h > 0$ un paramètre de lissage (bandwidth parameter).

On peut faire ici plusieurs remarques :

Remarque 1.4. *La fonction de régression évaluée au point x_0 est donc définie comme une somme pondérée des observations y_i dont les poids $w_i(x_0)$ dépendent de x_0.*

Remarque 1.5. *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 $|x_0 - x_i|$.*

Ces poids dépendent de la fonction noyau qui correspondent tout simplement à des fonctions de densité de probabilité.
Nous pouvons voir l'estimateur (1.5) comme une solution au problème d'optimisation (1.3) lorsque l'espérance conditionnelle est remplacée par une version empirique, c'est-à-dire par la moyenne pondérée par les $w_i(x_0)$ définie dans 1.6 L'équation qui est finalement obtenue après avoir effectué ce remplacement dans le problème d'optimisation (1.3) est

$$\begin{aligned}
\hat{m}(x_0) &= \arg\min_a \hat{\mathbb{E}}[(Y-a)^2 | X = x_0] \\
&= \arg\min_a \sum_{i=1}^{n} \frac{(y_i - a)^2 K(\frac{x_i - x_0}{h})}{\sum_{i=1}^{n} K(\frac{x_i - x_0}{h})} \\
&= \arg\min_a \sum_{i=1}^{n} w_i(x_0)(y_i - a)^2.
\end{aligned}$$

De manière analogue à la démonstration de l'équation (1.3), il est possible de trouver l'estimateur du noyau. En effet, cet estimateur est obtenu en dérivant la version empirique $\hat{\mathbb{E}}[(Y-a)^2|X=x_0]$ par rapport à a, comme suit :

$$\begin{aligned}
\frac{\partial}{\partial a}\sum_{i=1}^{n}\frac{(y_i-a)^2K(\frac{x_i-x_0}{h})}{\sum_{i=1}^{n}K(\frac{x_i-x_0}{h})} &= -2\sum_{i=1}^{n}\frac{(y_i-a)K(\frac{x_i-x_0}{h})}{\sum_{i=1}^{n}K(\frac{x_i-x_0}{h})} \\
&= -2\sum_{i=1}^{n}\frac{y_iK(\frac{x_i-x_0}{h})}{\sum_{i=1}^{n}K(\frac{x_i-x_0}{h})}+\sum_{i=1}^{n}\frac{aK(\frac{x_i-x_0}{h})}{\sum_{i=1}^{n}K(\frac{x_i-x_0}{h})} \\
&= 0 \\
\Rightarrow a &= \sum_{i=1}^{n}\frac{y_iK(\frac{x_i-x_0}{h})}{\sum_{i=1}^{n}K(\frac{x_i-x_0}{h})}.
\end{aligned}$$

la dérivée seconde de cette même version empirique, qui prend la valeur $2\sum_{i=1}^{n}\frac{K(\frac{x_i-x_0}{h})}{\sum_{i=1}^{n}K(\frac{x_i-x_0}{h})}$ est positive. Il est donc possible d'affirmer avec certitude que cette valeur a est bien une valeur minimale.

Définition 1.14. Une fonction noyau $K(\frac{x_i-x_0}{h})=K(u)$ vérifient les propriétés suivantes :

1. $K(u)\geq 0$.
2. $K(u)$ est normalisée de sorte que
$$\int K(u)du=1.$$
3. $K(u)$ atteint son maximum en 0 lorsque $x_i=x_0$ et décroît avec la distance $|x_0-x_i|$.
4. $K(u)$ est symétrique : le noyau ne dépend que de la distance $|x_0-x_i|$ et non du signe de x_0-x_i.

1- Quelque fonctions noyaux usuelles

– Le noyau uniforme
$$K(u)=\frac{1}{2}\quad u\in[-1,1].$$
– Le noyau triangulaire
$$K(u)=1-|u|\quad u\in[-1,1].$$
– Le noyau quartic ou Bi-Weight
$$K(u)=\frac{15}{16}(1-u^2)^2\quad u\in[-1,1].$$

– Le noyau Epanechnikov

$$K(u) = \frac{3}{4}(1-u^2) \quad u \in [-1,1].$$

– Le noyau triweight

$$K(u) = \frac{35}{32}(1-u^2)^3 \quad u \in [-1,1].$$

– Le noyau normal

$$K(u) = \frac{1}{\sqrt{2\pi}} exp(-\frac{u^2}{2}) \quad u \in]-\infty,+\infty[.$$

Remarque 1.6. *Des travaux ont montré qu'en pratique le choix de la fonction noyau n'influence peu les résultats d'estimation. La seule exception notable étant liée à l'utilisation d'une fonction noyau uniforme qui peut donner des résultats sensiblement différents des autres noyaux, les fonctions noyaux triangulaires ou gaussienes donnent plus de poids au voisinage de zéro.*

Enfin, les poids $w_i(x_0)$ dépendent en outre du paramètre de lissage h qui contrôle l'amplitude des poids.

Remarque 1.7. *Plus le paramètre de lissage h (bandwidth parameter) est élevé, plus on attribue un poids relativement important aux observations x_i éloignées du point de référence x_0 dans la construction de $m(x_0)$.*

1.4.3 Propriétés de l'estimateur à noyau

Lorsque l'on veut comparer plusieurs estimateurs, il faut calculer certaines mesures permettant d'évaluer leurs qualités, telles que le biais et la variance. L'erreur quadratique moyenne (MSE) peut aussi être calculée. Cette dernière est en fait une mesure de la différence quadratique espérée entre l'estimateur et sa valeur théorique. Bien entendu, l'estimateur par la méthode du noyau ne donne pas exactement la même valeur que la valeur théorique. Il serait donc intéressant de voir comment se comportent le biais et la variance pour cet estimateur du noyau $\hat{m}(x)$.

1- La dualité biais-variance

Le compromis entre le lissage et la flexibilité de l'estimateur est identifié comme la dualité biais-variance. Ainsi, en augmentant la flexibilité, il est possible de suivre plus

fidèlement les données, ce qui fait diminuer le biais. La courbe obtenue a donc plus tendance à osciller, ce qui implique que la variance augmente. Par contre, on préfère souvent avoir une courbe qui soit assez lisse, avec moins de variance. Pour ce faire, il faut diminuer la flexibilité de l'estimateur, ce qui implique de suivre moins fidèlement les données, donc d'augmenter le biais. Par conséquent, tout utilisateur d'une méthode de régression non paramétrique doit composer avec cette dualité, lorsque vient le temps de choisir la valeur du paramètre de lissage.

2- Le 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 $m(.)$ et $f_X(.)$ qui détermineront l'ordre du biais asymptotique en fonction du paramètre de lissage h.

Proposition 1.1. [Blondin][4] Supposons que $m(.)$ et $f_X(.)$ sont de classe $C^2(\mathbb{R})$ et que le noyau K est d'ordre 2, i.e. tel que

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

Nous avons alors, lorsque $h \to 0$ et $nh \to \infty$

$$\begin{aligned} Biais(\hat{m}(x)) &= \mathbb{E}[\hat{m}(x) - m(x)] \\ &= \frac{h^2}{2}(m''(x) + 2m'(x)\frac{f'_X(x)}{f_X(x)}) \int u^2\ K(u)\ du\ +\ o(h^2). \end{aligned}$$

3- La variance de l'estimateur

Le noyau K est supposé vérifier les hypothèses suivantes :

1. K est bornée, i.e, $sup_{u\in\mathbb{R}}\ |K(u)| < \infty$;
2. $\lim\limits_{|u|\to\infty}\ |u|\ K(u) = 0$;
3. $K(.) \in L_1(\mathbb{R}),\ i.e,\ \int_{\mathbb{R}} |K(u)|\ du\ < \infty$;
4. $\int_{\mathbb{R}} K(u)\ du = 1$.

On note que les hypothèses (1) et (3) impliquent le fait que $K(.)$ soit de carré intégrable. Nous posons, par convenance

$$\sigma^2(x) = \mathbb{V}ar[Y|X = x],$$

lorsque cette expression est bien définie.

Proposition 1.2. **[Blondin]**[4]
On suppose que $\mathbb{E}[Y^2] < \infty$. A chaque point de continuité des fonctions $m(x)$, $f_X(x)$ et $\sigma^2(x)$, tel que $f_X(x) > 0$,

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

1.4.4 Convergence en probabilité

Proposition 1.3. **[Härdle]**[30]
Supposons que les conditions suivantes sont vérifiées

1. $\int |K(u)|\ du\ < \infty$,
2. $\lim\limits_{|u|\to\infty} uK(u) = 0$,
3. $\mathbb{E}Y^2 < \infty$,
4. $n \to \infty,\ h_n \to 0,\ nh_n \to \infty$.

Alors, à chaque point de continuité de $m(x),\ f_X(x)\ et\ \sigma^2(x)$ avec $f_X(x) > 0$, on a :

$$n^{-1}\sum_{i=1}^{n} w_i(x)y_i \longrightarrow m(x),\ \text{en probabilité.}$$

1.4.5 Convergence presque complète

Cette notion de convergence presque complète entraîne à la fois la convergence presque sûre et la convergence en probabilité. Dans un premier temps nous nous plaçons en un point fixé x. Nous reprenons le modèle (1.2) et supposons que la condition suivante est vérifiée

$$m\ et\ f_X \text{ sont } k \text{ fois continûment dérivables autour de } x.$$

C-à-d,

$$\mathbb{E}(Y|X = x) = m(x), \tag{1.7}$$

où

$$m \text{ est } k \text{ fois contniûment dérivable,}$$

et

$$m \text{ et } f_X \text{ sont } k \text{ fois continûment dérivables autour de } x.$$

Théorème 1.1. Vitesse de convergence presque complète ponctuelle sous condition de dérivabilité [Sarda et Vieu][54]
Considérons le modèle (1.7) avec $k > 0$ et supposons que les conditions suivantes sont réalisées.

1. $f(x) > 0$,
2. $|Y| < M < \infty$,

On a alors

$$\hat{m}(x) - m(x) = O(h^k) + O(\sqrt{\frac{\log n}{nh}}), \text{ presque complétement.}$$

Ce résultat est issu de [Sarda et Vieu][54] mais les idées qui servent de base à cette démonstration reviennent à [Collomb][9] et [13].

Pour établir une version uniforme du résultat précédent, plaçons nous sur un compact S de $\mathbb{R}$, tel qu'il existe $\theta > 0$ pour lequel on a

$$m \text{ et } f_X \text{ sont } k \text{ fois continûment dérivable autour de } S, \tag{1.8}$$

et

$$\inf_{x \in S} f_X(x) > \theta. \tag{1.9}$$

Théorème 1.2. Vitesse de convergence presque complète uniforme sous condition de dérivabilité [Ferraty et Vieu] [22]
Considérons le modèle (1.8) avec $k > 0$ et supposons que les conditions suivantes sont vérifiées :

1. $\lim_{n\to\infty} h = 0$ *et* $\lim_{n\to\infty} \frac{nh}{\log n} = +\infty$;
2. K bornée, intégrable et à support compact ;
3. $\int u^j K(u)du = 0, \ \forall \ j = 1, \cdots, k-1$ *et* $0 < \int |u^k K(u)du| < \infty$;
4. $|Y| < M < \infty$;

5. $\inf_{x\in S} f_X(x) > \theta$;
6. $\exists \beta > 0,\ \exists C < \infty,\ \forall x \in S,\ \forall y \in S, |K(x) - K(y)| \leq C|x - y|^{\beta}$.

On a alors

$$\sup_{x\in S} |\hat{m}(x) - m(x)| = O(h^k) + O(\sqrt{\frac{\log n}{nh}}),\ \text{presque complétement.}$$

Remarque 1.8. – *La condition* (6) *est une restriction de type Lipschitz sur le noyau.*
– *Les résultats précédents sont énoncés sous l'hypothèse de dérivabilité.*

Le résultat que nous allons énoncer est établi sous un modèle de régression plus général que le théorème (4.1) puisque l'hypothèse d'existence de dérivées continues pour les fonctions que l'on estime (i.e pour f_X et m) est remplacée par la condition de continuité.

Théorème 1.3. Convergence presque complète ponctuelle sous condition de continuité [Ferraty et Vieu] [22]
Considérons le modèle (1.7) avec $k = 0$ et supposons que les conditions suivantes sont vérifiées

1. $f_X(x) > 0$;
2. $\lim_{n\to\infty} h = 0\ et\ \lim_{n\to\infty} \frac{nh}{\log n} = +\infty$;
3. K bornée, intégrable et à support compact ;
4. $|Y| < M < \infty$.

On a alors

$$\hat{m}(x) \longrightarrow m(x),\ \text{presque complétement.}$$

Le prochain résultat est une version uniforme du théorème précédent. Il est établi sous un modèle de régression plus général que celui du théorème (**??**) puisque l'hypothèse d'existence de dérivées continues pour les fonctions que l'on estime (i.e pour f_X et m) est remplacée par la condition de continuité.

Théorème 1.4. Convergence uniforme presque complète sous condition de continuité [Ferraty et Vieu][22]
Considérons le modèle de régression (1.8) avec $k = 0$ et supposons que les conditions suivantes sont vérifiées

1. $\lim_{n\to\infty} h = 0\ et\ \lim_{n\to\infty} \frac{nh}{\log n} = +\infty$;
2. K bornée, intégrable et à support compact ;

3. $|Y| < M < \infty$;
4. $\inf_{x \in S} f_X(x) > \theta$;
5. $\exists \beta > 0,\ \exists C < \infty,\ \forall x \in S,\ \forall y \in S, |K(x) - K(y)| \leq C|x - y|^{\beta}$.

On a alors

$$\sup_{x \in S} |\hat{m}(x) - m(x)| \longrightarrow 0, \text{ presque complétement.}$$

Maintenant nous allons énoncer les résultats sous l'hypothèse de type Lipschitz. Les conditions de Lipschitz sont des conditions ponctuelles en x fixé du type

$$\exists \beta > 0,\ \exists C < \infty,\ \exists \epsilon > 0,\ \forall y \in]x - \epsilon, x + \epsilon[,\ |\phi(x) - \phi(y)| \leq C|x - y|^{\beta} \tag{1.10}$$

soit des conditions uniformes sur un compact S du type

$$\exists \beta > 0,\ \exists C < \infty,\ \forall x \in S,\ \forall y \in S,\ |\phi(x) - \phi(y)| \leq C|x - y|^{\beta}, \tag{1.11}$$

où ϕ désigne indifféremment f_X ou m.

Théorème 1.5. Vitesse de convergence presque complète ponctuelle sous condition de Lipschitz [Ferraty et Vieu][22]

Considérons le modèle (1.10) et supposons que les conditions suivantes sont vérifiées

1. $f_X > 0$;
2. $\lim_{n \to \infty} h = 0$ *et* $lim_{n \to \infty} \frac{nh}{\log n} = +\infty$,
3. K bornée, intégrable et à support compact ;
4. $|Y| < M < \infty$;

On a alors

$$\hat{m}(x) - m(x) = O(h^{\beta}) + O(\sqrt{\frac{\log n}{nh}}), \text{ presque complétement.}$$

Théorème 1.6. Vitesse de convergence presque complète uniforme sous condition de Lipschitz [Ferraty et Vieu][22]

Considérons le modèle (1.11) et supposons que les conditions suivantes sont vérifiées

1. $\lim_{n \to \infty} h = 0$ *et* $\lim_{n \to \infty} \frac{nh}{\log n} = +\infty$;
2. K bornée, intégrable et à support compact ;
3. $|Y| < M < \infty$;

4. $\inf_{x\in S} f_X(x) > \theta$;
5. $\exists \beta > 0,\ \exists C < \infty,\ \forall x \in S,\ \forall y \in S, |K(x) - K(y)| \leq C|x - y|^{\beta}$;

On a alors

$$\sup_{x\in S} |\hat{m}(x) - m(x)| = O(h^{\beta}) + O(\sqrt{\frac{\log n}{nh}}),\ \text{presque complétement.}$$

1.4.6 Convergence en moyenne quadratique

1- Erreur moyenne quadratique ponctuelle

Nous allons nous intéresser à des résultats asymptotique en termes de convergence quadratique. Le premier résultat établi la vitesse de convergence en moyenne quadratique de l'estimateur à noyau de la régression sous hypothèse de dérivabilité, en un point x fixé. Supposons que

1. $\lim_{n\to\infty} h = 0$ *et* $\lim_{n\to\infty} nh = \infty$;
2. La fonction
$$\mathbb{E}(Y^2|X = x) \text{ est continue au point } x;$$
3. $K(u) > 0,\ \forall\ u$.

Cette 3ème condition est incompatible avec un noyau d'ordre supérieur à 2. Par conséquent, nous en resterons à un modèle où les fonctions m et f_X sont deux fois continûment dérivable autour de x, et les conditions sur le noyau deviennent alors :

K bornée, intégrable, positive, symétrique et à support compact.

Théorème 1.7. Convergence en moyenne quadratique ponctuelle sous condition de dérivabilité [Ferraty et Vieu][22]
Considérons le modèle (1.7) avec $k = 2$ et supposons les conditions suivantes sont vérifiées :

1. $f_X(x) > 0$;
2. $|Y| < M < \infty$;
3. $\lim_{n\to\infty} h = 0$ *et* $\lim_{n\to\infty} nh = \infty$;
4. $\mathbb{E}(Y^2|X = u)$ est continue au point x ;
5. K bornée, intégrable, positive, symétrique et à support compact.

On a alors :

$$\begin{aligned}\mathbb{E}[\hat{m}(x) - m(x)]^2 &= [\frac{h}{2}(m''(x) + 2m'(x)\frac{f'_X(x)}{f_X(x)}) \int u^2\ K(u)\ du]^2 \\ &+ \frac{1}{nh}\frac{\sigma^2(x)}{f_X(x)} \int K^2\ du\ +\ o(h^4 + \frac{1}{nh}).\end{aligned}$$

Le résultat suivant est établi sous un modèle de régression plus général que celui du théorème (1.7), puisque l'hypothèse d'existence de dérivées continues pour les fonctions que l'on estime (i.e. pour f_X et m) est remplacée par la simple condition de continuité.

Théorème 1.8. Convergence en moyenne quadratique ponctuelle sous condition de continuité [Ferraty et vieu][22]
Considérons le modèle (1.7) avec $k = 0$ et supposons que les conditions suivantes sont vérifiées :

1. $f_X(x) > 0$;
2. $|Y| < M < \infty$;
3. $\lim\limits_{n\to\infty} h = 0\ et\ \lim\limits_{n\to\infty} nh = \infty$;
4. K bornée, intégrable, positive, symétrique et à support compact.

On a alors

$$\mathbb{E}[\hat{m}(x) - m(x)]^2 \longrightarrow 0.$$

2- Erreur moyenne quadratique intégrée

[Ferraty et Vieu][22] ont donné des versions uniformes sur un compact des deux théorèmes précédents. L'erreur moyenne quadratique intégrée est définie par :

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

La fonction w est une fonction de poids vérifiant :

$$w \text{ est positive, bornée et à support compact } S.$$

Cette fonction est fixée a priori, et sera souvent dans la pratique prise égale à une indicatrice sur un intervalle borné de $\mathbb{R}$, ou égale à un produit d'une telle indicatrice par la densité f_X de X [Ferraty et vieu][22].
Supposons aussi que

$$\mathbb{E}(Y^2|X = u) \text{ est continue autour de} S.$$

Théorème 1.9. Erreur moyenne quadratique intégrée sous condition de dérivabilité [Ferraty et Vieu][22]
Considérons le modèle (1.8) avec $k = 2$ et supposons que les conditions suivantes sont vérifiées

1. $\inf_{x\in S} f_X(x) > \theta,\ \theta > 0$;
2. $\lim_{n\to\infty} h = 0\ et\ \lim_{n\to\infty} nh = \infty$;
3. K bornée, intégrable, positive, symétrique et à support compact ;
4. w est positive, bornée et à support compact S ;
5. $\mathbb{E}(Y^2|X = u)$ est continue autour de S.

On a alors

$$\begin{aligned} MISE(\hat{m}) &= \int [\frac{h^2}{2}(m''(x) + 2m'(x)\frac{f'_X(x)}{f_X(x)}) \int u^2 K(u)du]^2 w(x)dx \\ &\quad + \int [\frac{1}{nh}\frac{\sigma^2(x)}{f_X(x)} \int K^2 du]w(x)dx + o(h^4 + \frac{1}{nh}). \end{aligned}$$

Maintenant, nous allons donner un résultat analogue sous le modèle plus général $k = 0$.

Théorème 1.10. Erreur moyenne quadratique intégrée sous condition de continuité [Ferraty et Vieu][22]
Considérons le modèle (1.8) avec $k = 0$ et supposons que les conditions suivantes sont vérifiées

1. $\inf_{x\in S} f_X(x) > \theta;\ \theta > 0$;
2. $\lim_{n\to\infty} h = 0\ et\ \lim_{n\to\infty} nh = \infty$;
3. K bornée, intégrable, positive, symétrique et à support compact ;
4. w est positive, bornée et à support compact S ;

On a alors

$$MISE(\hat{m}) \longrightarrow 0.$$

1.4.7 Choix du paramètre de lissage h

Comment choisir le paramètre de lissage h dans le cadre d'une régression par noyau ? C'est sans doute le point le plus important de ce type de méthodes. Rappelons que pour certaines fonctions noyaux, les points x_i qui sont distants de plus de h du point de référence

x_0, ne sont pas pris en compte dans le calcul de $m(x_0)$.

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

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

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 $w_i(x_0) = 0$.

Proposition 1.4. [Hurlin][33]

Le choix du paramètre de lissage h correspond à un arbitrage variance / biais :

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

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

Exemple

supposons que l'on choisisse h tel que $h \to \infty$. Alors, on a :

$$\lim_{h\to\infty} K(\frac{x_i - x_0}{h}) = K(0), \quad \forall x_i.$$

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

$$\begin{aligned} \lim_{h\to\infty} w_i(x_0) &= \lim_{h\to\infty} \frac{K(\frac{x_i-x_0}{h})}{\sum_{i=1}^n K(\frac{x_i-x_0}{h})} \\ &= \frac{K(0)}{n\ K(0)} \\ &= \frac{1}{n}. \end{aligned}$$

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

$$\lim_{h\to\infty} \hat{m}(x_0) = \lim_{h\to\infty} \sum_{i=1}^n w_i(x_0) y_i = \frac{1}{n}\sum_{i=1}^n y_i = \overline{y}.$$

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

Exemple

supposons au contraire que l'on choisisse h tel que $h \to 0$. Alors, pour tous les points x_j différents du point x_i de référence :

$$\lim_{h\to 0} K(\frac{x_i - x_j}{h}) = K(\pm\infty) = 0 \quad \forall i \neq j;$$

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

$$K(\frac{x_i - x_i}{h}) = K(0), \quad \forall h.$$

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

$$\lim_{h\to 0} w_j(x_i) = \lim_{h\to 0} \frac{K(\frac{x_j - x_i}{h})}{\sum_{i=1}^{n} K(\frac{x_j - x_i}{h})} = 0 \quad \forall i \neq j$$

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

$$\begin{aligned} \lim_{h\to 0} w_i(x_i) &= \lim_{h\to 0} \frac{K(\frac{x_i - x_i}{h})}{\sum_{i=1}^{n} K(\frac{x_i - x_i}{h})} \\ &= \lim_{h\to 0} \frac{K(0)}{\sum_{j\neq i} K(\frac{x_j - x_i}{h}) + K(\frac{x_i - x_i}{h})} \\ &= \frac{K(0)}{K(0)} = 1. \end{aligned}$$

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

$$\lim_{h\to 0} \hat{m}(x_0) = \lim_{h\to 0} \sum_{j\neq i} w_j(x_i) y_j + \lim_{h\to 0} w_i(x_i) y_i = y_i.$$

Ainsi si le paramètre de lissage tend vers zéro, 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 $\hat{m}(x)$ est aussi grande que la variance de y, mais le biais est faible.

Nous avons constaté que le terme de biais est en terme en h^4, tandis que le terme de variance est en $1/(nh)$. L'un est proportionnel à h tandis que l'autre est inversement proportionnel à h. L'existence d'un biais n'est pas liée à la méthode d'estimation par noyau, mais que c'est une chose inhérente au modèle non paramétrique lui-même (une propriété d'inexistence d'estimateur non biaisé de la régression dans des contextes non paramétrique est donnée par [Collomd (1976)][9] ou [Sarda et Vieu (2000)][54]).

Ainsi, une grande valeur de h se traduit par un estimateur fortement biaisé et alors

qu'une trop petite valeur de h entraîne un estimateur à forte variabilité.

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. Nous adopterons pour la mesure d'erreur une notation qui fait apparaître explicitement ce paramètre :

$$MISE(\hat{m}) = MISE(h).$$

Définition 1.15. La MISE associé à un paramètre de lissage h, correspond à la quantité :

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

Dans l'absolu on cherche la valeur optimale de h telle que :

$$h_{opt} = \arg \min_{h \in \mathbb{R}^{*+}} MISE(h).$$

Regardons le résultat du théorème (1.9)

$$MISE(\hat{m}) = B^2 h^4 + \frac{V}{nh} + o(h^4 + \frac{1}{nh}),$$

où

$$B = \int [\frac{1}{2}(m''(x) + 2m'(x)\frac{f'_X(x)}{f_X(x)}) \int u^2 \; K(u) \; du]^2 \; w(x)dx,$$

et

$$V = \int [\frac{\sigma^2(x)}{f_X(x)} \int K^2 \; du \;] \; w(x) \; dx.$$

La fonction $MISE(\hat{m})$ est une fonction convexe en h, il est facile de minimiser ce développement asymptotique. Le minimum est atteint pour la valeur optimale h_{opt} :

$$h_{opt} = (\frac{V}{4nB^2})^{\frac{1}{5}}.$$

Le même raisonnement peut se faire à partir du résultat (1.8). on arrive aux deux corollaires suivants.

Corollaire 1.1. Convergence optimale en moyenne quadratique ponctuelle [Ferraty et Vieu][22]

Considérons le modèle (1.7) avec $k = 2$ et supposons que les conditions suivantes sont vérifiées

1. $f_X(x) > 0;\ \theta > 0;$
2. $|Y| < M < \infty;$
3. $\mathbb{E}(Y^2|X = u)$ est continue au point x ;
4. K bornée, intégrable, positive, symétrique et à support compact.

Supposons en outre que la fenêtre h soit de la forme :

$$h = Cn^{-\frac{1}{5}},\ 0 < C < \infty.$$

On a alors

$$\mathbb{E}[\hat{m}(x) - m(x)]^2 = O(n^{-\frac{4}{5}}).$$

Corollaire 1.2. Erreur moyenne quadratique intégrée optimale [Ferraty et Vieu][22]

Considérons le modèle (1.8) avec $k = 2$ et supposons que les conditions suivantes sont vérifiées

1. $\inf_{x \in S} f_X(x) > \theta; \theta > 0;$
2. K bornée, intégrable, positive, symétrique et à support compact ;
3. w est positive, bornée et à support compact S ;
4. $\mathbb{E}(Y^2|X = u)$ est continue autour de S;
5. $h = Cn^{-\frac{1}{5}},\ \ 0 < C < \infty.$

On a alors

$$MISE(\hat{m}) = O(n^{-\frac{4}{5}}).$$

Dans les résultats de convergence presque complète donnés dans les théorèmes (1.1) et (1.2). Il s'agit de minimiser en h des expressions du type

$$O(h^k) + O(\sqrt{\frac{\log n}{nh}}).$$

[Ferraty et Vieu][22] donne une nouvelle hypothèse sur h pour atteindre asymptotiquement le minimum est alors

$$h = C(\frac{n}{\log n})^{-\frac{1}{2k+1}}, \quad 0 < C < \infty.$$

Les corollaires suivants sont des conséquences des théorèmes (1.1) et (1.2).

Corollaire 1.3. Convergence presque complète ponctuelle optimale [Ferraty et Vieu][22]

Considérons le modèle (1.7) avec $k > 0$ et supposons que les conditions suivantes sont vérifiées

1. $f_X(x) > 0; \theta > 0;$
2. K bornée, intégrable et à support compact ;
3. $\int u^j K(u)du = 0, \ \forall \ j = 1, \cdots, k-1 \ et \ 0 < \int |u^k K((u)du| < \infty;$
4. $|Y| < M < \infty;$
5. $h = C(\frac{n}{\log n})^{-\frac{1}{2k+1}}, \quad 0 < C < \infty.$

On a alors

$$\hat{m}(x) - m(x) = O((\frac{n}{\log n})^{-\frac{k}{2k+1}}), \text{ presque complétement.}$$

Corollaire 1.4. Convergence presque complète uniforme optimale [Ferrat et Vieu][22]

Considérons le modèle (1.8) avec $k > 0$ et supposons que les conditions suivantes sont vérifiées

1. K bornée, intégrable et à support compact ;
2. $\int u^j K(u)du = 0, \ \forall \ j = 1, \cdots, k-1 \ et \ 0 < \int |u^k K((u)du| < \infty;$
3. $\inf\limits_{x \in S} f_X(x) > \theta; \ \theta > 0;$
4. $\exists \beta > 0, \ \exists C < \infty, \ \forall x \in S, \ \forall y \in S, |K(x) - K(y)| \leq C|x - y|^{\beta};$
5. $h = C(\frac{n}{\log n})^{-\frac{1}{2k+1}}, \quad 0 < C < \infty.$

On a alors

$$\sup_{x \in S} |\hat{m}(x) - m(x)| = O((\frac{n}{\log n})^{-\frac{k}{2k+1}}), \text{ presque complétement.}$$

Ces résultats sont intéressants d'un points de vue théorique, puisque d'après [Stone (1981) et (1982)][61][62] la vitesse optimale de convergence pour le modèle (1.8) est :

$$n^{-\frac{k}{2k+1}}, \quad \text{en norme } L_p, \ p < \infty,$$

et

$$(\frac{n}{\log n})^{-\frac{k}{2k+1}}, \ \text{en norme } L_\infty.$$

1.4.8 La validation croisée

Le problème dans la première méthode c'est que l'on ne connaît pas la quantité $m(x)$ et que l'on ne peut donc directement évaluer cette MISE. On utilise alors une approche qui asymptotiquement nous donne une valeur proche de h_{opt} : l'approche de la fonction validation croisée généralisée (General Cross-Validation GCV).
On peut faire l'analogie avec la méthode simple qui consisterait à déterminer la valeur de h qui minimiserait la variance estimée des résidus.

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

Ce critère nous permettra d'obtenir la valeur de h telle que les données sont parfaitement ajustées. En effet, si l'on cherche :

$$\tilde{h} = \arg\min_{h \in \mathbb{R}^{*+}} \hat{\sigma}^2(h),$$

on va alors aboutir au résultat $\tilde{h} \to 0$. Si le paramètre de lissage tend vers 0, alors nous avons $\hat{m}(x_i; h) = y_i$ et donc $\hat{\sigma}^2(0) = 0$. Ce critère est a priori sans intérêt, mais on peut considérer une légère variation connue sous le nom de la fonction validation croisée (cross validation function).

Définition 1.16. La fonction validation croisée est définie par la quantité :

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

La seule différence avec le critère précédent réside dans l'utilisation de l'indice $\hat{m}^{-i}$. Cet indice signifie que pour chaque $i = 1, \cdots, n$; la valeur de $m(x_i)$ est obtenue en enlevant la $i^{\text{ème}}$ observation x_i. Le modèle est estimé sur toutes les autres observations $x_j, j \neq i$,

puis on estime la valeur de $m(.)$ au point x_i à partir de cette régression. C'est cette valeur estimée qui figure dans la formule $CV(h)$ sous la notation $m^{-i}(x;h)$.

$$\hat{m}^{-i}(x) = \frac{\sum_{j\neq i} y_j K(\frac{x-x_j}{h})}{\sum_{j\neq i} K(\frac{x-x_j}{h})}.$$

Les références principales à ce sujet sont [Hall (1984)][27] , [Härdle et Marron][28], [Härdle et Kelly (1987)][31], concernant l'estimation non paramétrique de la régression. La procédure de validation croisée peut s'interpréter comme étant le meilleur choix de h qui fait de $\hat{m}^{-i}(x_i)$ un estimateur efficace de Y_i.

Proposition 1.5. [Hurlin][33]

Soit h_{CV} la valeur de h telle que :

$$h_{CV} = \arg \min_{\{h\in\mathbb{R}^{*+}\}} CV(h).$$

Alors

$$MISE(h_{CV}) \longrightarrow MISE(h_{opt}), \text{ en probabilité quand } n \to \infty.$$

L'utilisation de la fonction CV permet ainsi d'obtenir un estimateur du paramètre optimal h_{opt}.

[Ferraty et Vieu][22] ont donné une autre forme pour la fonction validation corisée en faisant intervenir la fonction poids définie précédemment

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

Nous posons $H \subset \mathbb{R}$ désigne l'ensemble des valeurs possible de h (i.e la zone de variation de $\hat{h}$)

Théorème 1.11. Validation croisée : optimalité asymptotique [Härdle et Marron][28]
Sous les hypothèses du théorème (1.9) et si H ne contient que des largeurs de fenêtres vérifiant

$$h = Cn^{-\frac{1}{5}}, \quad 0 < C < \infty,$$

alors on a la propriété suivante

$$\frac{\inf\limits_{h\in H} MISE(h)}{MISE(h_{CV})} \longrightarrow 1, \text{ presque sûrement.}$$

1.4.9 Normalité asymptotique

La première démonstration de la normalité asymptotique de l'estimateur Nadaraya-Watson est due à [Schuster][57]. On se réfère également aux théorèmes (1.3) et (1.4) de [Nadaraya][45] et au théorème (4.2.1) de [Härdle][67], qui proposent d'autres méthodes de démonstration. Le noyau K est supposé borné, à support compact et d'ordre 2. La fenêtre h est choisie égale à $cn^{-1/5}$.

Théorème 1.12. [Härdle][67]

Supposons Y bornée ou admettant un moment d'ordre $l > 2$. Les fonctions $f_X(.)$ et $m(x)$ sont supposées deux fois continûment dérivables sur $\mathbb{R}$. A chaque point de continuité de $\sigma^2(x)$, tel que $f_X(x) > 0$,

$$(nh)^{1/2}\{\hat{m}(x) - m(x)\} \longrightarrow \mathcal{N}(B(x), \mathbb{V}(x)), \text{ en loi,}$$

avec

$$\mathbb{V}(x) = \frac{\sigma^2(x)}{f_X(x)} \int K^2(u)\, du, \text{ (la variance asymptotique),}$$

et

$$B(x) = [m''(x) + 2m'(x)\frac{f'_X(x)}{f_X(x)}] \int u^2 K(u)\, du, \text{ (le biais asymptotique).}$$

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

$$(nh)^{1/2}\{\hat{m}(x) - m(x)\} \longrightarrow \mathcal{N}(0, \frac{\sigma^2}{2f_X(x)}), \text{ en loi,}$$

Ce résultat permet en outre de construire des intervalles de confiance pour les valeurs de $m(x_i)$.

1.5 Autres méthode d'estimation

1.5.1 La méthode des k-plus proches voisins

La construction des estimateurs k-plus proches voisins différents de l'estimateur à noyau. L'estimateur à noyau $\hat{m}(x)$ est défini comme une moyenne pondérée des variables réponses dans un voisinage fixé autour de x, déterminé dans la forme par le noyau K et la largeur de bande h. L'estimateur k plus proche voisin est une moyenne pondérée dans un voisinage variable. Ce voisinage est défini par ces variables X qui sont parmi les k-plus proches voisins se x par rapport à la distance euclidienne. La suite des poids des k-plus proches voisins a été introduite par [Loftsgaarden and Quesenberry][39] dans le domaine lié à l'estimation de la densité et utilisé par [Cover and Hart][14] dans le but de classification.

Définition 1.17. Dans le cadre présent de la régression, le lisseur k-plus proche voisin est défini par

$$\hat{m}_k(x) = n^{-1} \sum_{i=1}^{n} w_{ki}(x) y_i,$$

où $\{w_{ki}(x)\}_{i=1}^{n}$ est une suite de poids définie par l'ensemble d'indices

$$J_x = \{i : \ x_i \text{ est l'une des } k\text{-plus proches observations de } x\},$$

avec cet ensemble d'indices des observations voisine, la suite des poids des k-plus proches voisins est construite comme suit

$$w_{ki}(x) = \begin{cases} \frac{n}{k}, & \text{si } i \in J_x\,; \\ 0, & \text{sinon.} \end{cases} \tag{1.12}$$

Exemple : Considérons l'exemple suivant.
Soit $\{(x_i, y_i)\}_{i=1}^{5} = \{(1,5),(7,12),(3,1),(2,0),(5,4)\}$ et calculons les k-plus proches voisins $\hat{m}_k(x)$ pour $x = 4$ et $k = 3$. Les observations proches de x sont les 3 derniers points des données. Par conséquent $J_x = J_4 = \{3,4,5\}$ donc

$$w_{k1}(4) = w_{k2}(4) = 0, \ w_{k3}(4) = w_{k4}(4) = w_{k5}(4) = \frac{5}{3}$$

ce qui résulte

$$\hat{m}_3(4) = \frac{1}{5}(1+0+4)\frac{5}{3} = \frac{5}{3}.$$

Dans une experience où la variable X est choisie dans une grille équidistante, les poids des k-plus proches voisins sont équivalents au poids du noyau. Soit $k = 2nh$ et comparons $\{w_{ki}(x)\}$ avec $\{w_{hi}(x)\}$ pour un noyau uniforme $K(u) = \frac{1}{2}1(|u| \leq 1)$ pour un x pas très proche de la limite.
Ainsi pour $i \in J_x$:

$$w_{ki}(x) = \frac{n}{2nh} = \frac{1}{2}h^{-1} = w_{hi}(x).$$

Le paramètre de lissage k règle le degré de lissage de la courbe estimée. Il joue un rôle similaire à la largeur de bande pour le noyau. Quand on fait varier k l'influence sur les caractéristiques qualitatives de la courbe estimée est similaire à celle observée par l'estimation à noyau avec le noyau uniforme.
Considérons pour n fixé, le cas où k devient plus grand que n. Le lisseur k-plus proche voisin est alors égal à la moyenne des variables réponses. L'autre cas limite est $k = 1$ dans lequel les observations se répètent à x_i, et pour un x entre deux variables adjacentes, une fonction seuil est obtenue avec un saut au milieu des deux observations. Un autre problème de sélection du paramètre de lissage est observé : k doit être choisi comme une fonction de n ou même des données. Comme 1^{er} objectif,c'est de pouvoir réduire le bruit, posons $k = k_n$ tend vers l'infini comme fonction de la taille de l'échantillon. Le second objectif est de garder l'erreur approximative (le biais) petite. Le second objectif est réalisé si le voisinage autour de x se rétréci asymptotiquement vers zéro, ceci peut être fait par définir $k = k_n$ tel que $\frac{k_n}{n} \to 0$. Malheureusement, cette condition contredit le 1^{er} objectif. Pour garder la variance petite autant que possible il faut choisir k grand que possible.

Pour cela, nous envisageons encore un compromis entre "trouver une bonne approximation " pour la fonction de régression et une bonne reduction du bruit observé. Ce problème de compromis peut être formuler par un développement de l'erreur moyenne quadratique de l'estimateur k-plus proches voisins.

Proposition 1.6. [Lai][35]
Soit $k \to \infty$, $\frac{k}{n} \to \infty$, $n \to \infty$. Le biais et la variance de l'estimateur k-plus proches voisins $\hat{m}_k$ sont donnés par :

$$\mathbb{E}\hat{m}(x) - m(x) \approx \frac{1}{24f(x)^3}[(m''f_X + 2m'f'_X)(x)](\frac{k}{n})^2;$$

$$\mathbb{V}ar\ \hat{m}_k(x) \approx \frac{\sigma^2(x)}{k}.$$

Le compromis entre le biais^2 est ainsi réalisé dans un sens asymptotique en posant $k \sim n^{4/5}$. Une conséquence est que l'erreur moyenne quadratique elle même converge vers 0 au taux de $k^{-1} \sim n^{-\frac{4}{5}}$.

Autre suite de poids $w_{ki}(x)$ a été proposée par [Stone (1977)][60]. Ajoutant au poids uniforme défini dans 1.12 il a défini les poids triangulaires et quadratiques du k-plus proches voisins. Généralement, les poids peuvent être pris comme étant générés par des fonctions noyaux K,

$$w_{Ri} = \frac{K(\frac{x-x_i}{R})}{\sum_{i=1}^{n} K(\frac{x-x_i}{R})},$$

où $R = R_n$ est la distance entre x et ses k plus proches voisins. Dans l'exemple donné précédemment avec $x = 4$ et $k = 3$, la distance R sera égale à 2 puisque l'observation $(2, 0)$ c'est la plus éloignée parmi les 3 voisins de $x = 4$.

[Mack][40] a calculé le biais et la variance pour cette paramétrisation de poids $\{w_{Ri}\}_{i=1}^{n}$.

Proposition 1.7. [Mack][40]
Soit $k \to \infty$, $\frac{k}{n} \to 0$, $n \to \infty$ et soit c_K et d_K définis comme suit

$$c_K = \int K^2(u)\, du,$$

et

$$d_K = \int u^2 K(u)\, du.$$

Alors on a

$$\mathbb{E}\hat{m}_R(x) - m(x) \approx \frac{k}{n} \frac{m''f + 2m'f'}{8f(x)^3}\, d_K,$$

$$\mathbb{V}ar(\hat{m}_R(x)) \approx 2\frac{\sigma^2(x)}{k}\, c_K.$$

Une conséquence de cette proposition est qu'une contribution d'équilibre entre le $biais^2$ et la variance de l'erreur moyenne quadratique est pour le poids uniforme du k-plus proche voisin est réalisé en posant $k = k_n$ est proportionnel à $n^{\frac{4}{5}}$. Pensant, que la largeur de bande h des lisseurs noyau peuvent être équivalent à $\frac{1}{2}nk^{-1}$ grossièrement, il est vu que les taux du biais et de la variance de $\hat{m}_R$ sont complètement équivalents à ceux des lisseurs noyau, uniquement les constantes différent. Le biais de $\hat{m}_R$ tend d'être grand

(à la fin de la distribution marginale). L'estimateur à noyau présente un comportement différent : la variance est un multiple de $f(x)^{-1}$ et le biais s'avère ne pas d'être une fonction de $f(x)^{-3}$. Une comparaison des propriétés de l'erreur moyenne quadratique des lisseurs noyau et k-plus proches voisins peuvent être trouvées dans le tableau suivant

	noyau	k-plus proche voisin
Biais	$h^2 \frac{(m''f+2m'f')(x)}{2f(x)} d_K$	$(\frac{k}{2})^2 \frac{(m''f+2m'f')(x)}{8f^3(x)} d_K$
Variance	$\frac{\sigma^2(x)}{nhf(x)} c_K$	$\frac{2\sigma^2(x)}{k} c_K$

TABLE 1.1 – Biais et variance des lisseurs noyau et k-plus roches voisins

Les entrées du tableau montrent une dépendance en f, h et k. L'équivalence essentielle dans la ligne du tableau peut être vue par l'utilisation de la relation

$$k = 2nhf(x).$$

L'utilisation de k conduit (asymptotiquement) à la même erreur moyenne quadratique en x pour les k-plus proches voisins et le noyau. Les taux de convergence pour cet estimateur des k-plus proches voisins se trouvent dans [Devroye][17] et [Györfi][26].

1.5.2 La méthode des séries orthogonales

Supposons que la fonction de régression peut être représentée comme une série de Fourier

$$m(x) = \sum_{j=0}^{\infty} \beta_j \varphi_j(x), \tag{1.13}$$

où $\{\varphi_j\}_{j=0}^{\infty}$ est une base connue de fonctions et $\{\beta_j\}_{j=0}^{\infty}$ sont des coefficients de Fourier inconnus.[Szegö][63]. Les conditions sous lesquelles une représentation de m est possible. Des exemples bien connus des fonctions de base sont des polynômes de Laguerre et Legendre. Une fois que la base de fonction est fixée, le problème de l'estimation de m peut

être abordé par l'estimation des coefficients de Fourier $\{\beta_j\}_{j=0}^{\infty}$. Il y a, bien sûr, la restriction qu'il peut infiniment y avoir beaucoup des β_j non nuls dans la formule (1.13). Alors, étant donné un échantillon fini de taille n, uniquement un sous ensemble de coefficients peut être effectivement estimé. Pour la simplicité de la représentation, supposons que la variable X est limitée dans l'intervalle $[-1, 1]$ et que les observation $\{y_i\}_{i=1}^{n}$ sont prises en des points $\{x_i\}_{i=1}^{n}$ équidistants sur cet intervalle.
Supposons que le système de fonctions $\{\varphi_j\}$ constituent une base orthonormale sur $[-1, 1]$ tel que :

$$\int_{-1}^{1} \varphi_j(x)\ \varphi_k(x)\ dx = \ \delta_{jk} = \begin{cases} 0, & j \neq k\,; \\ 1, & \text{j=k.} \end{cases}$$

Alors, les coefficients de Fourier β_j peuvent être calculés par

$$\begin{aligned} \beta_j &= \sum_{k=0}^{\infty} \beta_k\ \delta_{jk} \\ &= \sum_{k=0}^{\infty} \beta_k\ \int \varphi_k(x)\ \varphi_j(x)\ dx \\ &= \int_{-1}^{1} m(x)\ \varphi_j(x)\ dx. \end{aligned}$$

La dernière intégrale ne comporte pas uniquement la base connue de fonctions, mais aussi la fonction inconnue $m(x)$. Si on peut estimer d'une façon raisonnable, on donne automatiquement une estimation de β_j. Rappelons que les observations sont prises en des points discrets dans l'intervalle $[-1, 1]$. Soit $\{A_i\}_{i=1}^{n}$ un ensemble d'intervalles disjoints tels que :

$$\sum_{i=1}^{n} A_i = \ [-1, 1],$$

et

$$x_i \in A_i, \quad i = 1, \cdots, n.$$

Maintenant, la formule des coefficients de Fourier dans (1.14) peut être écrite comme suit :

$$\beta_j = \sum_{i=1}^{n} \int_{A_i} m(x)\ \varphi_j(x)\ dx \ \approx \ \sum_{i=1}^{n} m(x_i) \int_{A_i} \varphi_j(x)\ dx, \tag{1.14}$$

si les intervalles A_i se concentrent autour de x_i. Par insertion de la variable réponse y_i dans $m(x_i)$ nous obtenons une estimation pour β_j

$$\hat{\beta}_j = \sum_{i=1}^{n} y_i \int_{A_i} \varphi_j(x)\ dx.$$

Puisque seulement un nombre d'observations est disponible, les coefficients de Fourier ne peuvent pas être estimés tous à la fois. Si $N(n)$ termes dans la représentation (1.13) sont considérés, la fonction de régression est approximée par

$$\hat{m}_N(x) = \sum_{j=0}^{N(n)} \hat{\beta}_j \ \varphi_j(x). \tag{1.15}$$

Définition 1.18. L'estimateur de m par la méthode des séries orthogonales est défini par

$$\hat{m}_N(x) = \sum_{j=0}^{N(n)} \hat{\beta}_j \ \varphi_j(x), \tag{1.16}$$

il est une moyenne pondérée des variables Y avec le poids

$$w_{Ni} = \ n \ \sum_{j=1}^{N(n)} \int_{A_i} \varphi_j(u) \ du \ \varphi_j(x).$$

Le paramètre de lissage est $N(n)$, le nombre de coefficients de Fourier qui interviennent dans (1.16).

Les aspects statistiques de l'estimation par la méthode des séries orthogonales ont été principalement examiné dans le domaine de l'estimation de la densité, voir [Cenzov][8], [Wahba][70], [Walter (1977)][76]. Pour quelques applications dans le domaine de la régression non paramétrique nons apportons ces résultats, concernant la consistance et le taux exact de convergence. La consistance de $\hat{m}_N(x)$ découle de la proposition suivante

Proposition 1.8. [Härdl][30]
Si pour un s tel que $0 < s < 1$

$$n^{s-1} \sum_{j=0}^{N(n)} sup_x |\varphi_j(x)|^2 < \infty \tag{1.17}$$

et

$$\mathbb{E}|\varepsilon_i|^{\frac{s+1}{s}} < \infty;$$

dès que

$$N(n) \to \infty;$$

$$\hat{m}_N(x) \longrightarrow m(x) \text{ en probabilité.}$$

Une preuve détaillée de la consistance de $\hat{m}_N$ peut être trouvé dans [Rutkowski][53]. [Szegö][63] montre que

$$\sup_x |\varphi_j(x)| \sim j^\rho, \quad j = 1, 2, 3 \cdots ,$$

avec $\rho = -\frac{1}{4}$ pour les systèmes d'Hermite et Laguerre et $\rho = 0, \frac{1}{2}$ pour les systèmes de Fourier et Legendre, respectivement. L'hypothèse (1.17) apporte alors la condition sur $N(n)$ de la forme de croissance

$$\frac{N(n)^{2\rho+1}}{n^{1-s}} \leq c < \infty \qquad quand\ n \longrightarrow \infty. \tag{1.18}$$

Il faut que le paramètre de lissage tend vers ∞ pour assurer la consistance, mais pas trop rapidement comme le suggère (1.18). Le taux de convergence de l'estimateur par les séries orthogonales est donné par [Härdle][29].

1.6 Conclusion

Nous avons présenté quelques méthodes d'estimation de la courbe de régression de la moyenne, à savoir la méthode des k-plus proches voisins, la méthode des séries orthogonales et plus particulièrement la méthode du noyau qui a été présentée en détail. Nous allons dans le chapitre suivant présenter une méthode alternative à ces méthodes qui est la méthode des fonctions splines. Nous donnerons l'estimateur obtenu par la méthode des fonctions splines et ses principales propriétés statistiques.

2

Régression non paramétrique par la méthode des fonctions splines

2.1 Introduction

Depuis plus de 50 ans [I.J Schoenberg][55] introduit "les fonctions splines" dans la littérature mathématique. Depuis, les fonctions splines se sont énormément développées et sont devenues importantes dans différentes branches des mathématiques telles que : la théorie de l'approximation, l'analyse numérique, et les statistiques. Les fonctions splines sont devenues un outil utile dans différents domaines d'application, comme dans l'animation, dans la topographie, et dans des logiciels de dessins. Aujourd'hui, les splines de lissage sont largement employées dans des domaines de l'ingénierie, du design et en construction automobile. On les utilise par exemple, pour approcher les contours les plus complexes. Leurs simplicité d'implémentation les rend très populaires.
Ce chapitre porte sur la présentation de la notion des fonctions splines et de l'estimation de la courbe de régression de la moyenne par la méthode des fonctions splines de lissage et ses propriétés statistiques.

2.2 Généralités sur les fonctions splines

Un spline polynômial d'ordre r (de degré $r-1$) est une fonction polynomiale par morceau sur chaque intervalle défini par des noeuds consécutifs qui admet r dérivées et $r-1$ dérivées continues à l'intérieur de chacun de ces intervalles. Donc la dérivée d'ordre r est une fonction étagée avec des sauts aux noeuds. D'ou nous avons la définition suivante

Définition 2.1. Fonction spline

Soit $I = [a, b]$ un intervalle de $\mathbb{R}$, soit $a < x_1 < \cdots < x_n < b$ une partition de l'intervalle I.
La fonction $f : I \longrightarrow \mathbb{R}$ est une fonction spline de degré $r-1$ (ou d'ordre r) ayant pour noeuds $x_1, \cdots, x_n$ (respectivement à cette partition) si :

1. $f \in \mathcal{C}^{r-2}(I)$ (f est continûment dérivable jusqu'à l'ordre $r-2$),
2. $f \in \mathcal{P}_{r-1}$ sur chaque sous intervalle $[x_i, x_{i+1}]$ pour tout $i = 1, \cdots, n-1$

Où $\mathcal{P}_{r-1} = \{p : \mathbb{R} \longrightarrow \mathbb{R}, P \text{ polynôme de degré } \leq (r-1),\ r \in \mathbb{N}\}$.
Cet ensemble de fonctions sera noté $\mathcal{S}_r(x_1, \cdots, x_n)$. Il contient l'ensemble des polynômes de degré inférieur ou égal à $r-1$.

Théorème 2.1. [Thomas-Agnan][64]
$\mathcal{S}_r(x_1, \cdots, x_n)$ est un sous espace vectoriel de l'espace des fonctions dérivables jusqu'à l'ordre $r-2$ dont une base est donnée par les fonctions $1, x, \cdots, x^{r-1}$ et les fonctions $(x-x_1)_+^{r-1}, \cdots, (x-x_n)_+^{r-1}$, avec

$$dim\mathcal{S}_r(x_1, \cdots, x_n) = r + n$$

et

$$u_+^{r-1} = \begin{cases} u^{r-1}, & \text{si } u > 0\,; \\ 0, & u \leq 0. \end{cases}$$

Définition 2.2. Fonction spline naturelle
Soit $a < x_1 < \cdots < x_n < b$ une partition de l'intervalle I. Une fonction spline d'ordre $2r$ ou de degré impair $(2r-1)$ ayant pour noeuds $x_1, \cdots, x_n$ est dite naturelle, si elle coincide avec un polynôme de degré inférieur ou égale à $(r-1)$ en dehors de l'intervalle $[x_1, x_n]$.
On note par $\mathcal{S}_{2r}(x_1, \cdots, x_n)$ l'espace des fonctions splines naturelles d'ordre $2r$.
C-à-d $f \in \mathcal{S}_{2r}(x_1, \cdots, x_n)$. Nous avons alors :

1. f est une spline de degré $2r-1$,
2. $f \in \mathcal{P}_{r-1}$, pour $x \in [a, x_1] \cup [x_n, b]$.

$dim\mathcal{S}_{2r}(x_1, \cdots, x_n) = n$.

Définition 2.3. Spline cubique naturelle

Soient $x_1 < x_2 < \cdots < x_n$ n points d'un intervalle $[a, b]$, une fonction f définie sur $[a, b]$ est une spline cubique si les deux conditions suivantes sont satisfaites :

1. Sur chaque intervalle $(a, x_1),\ (x_1, x_2), \cdots, (x_n, b)$, f est un polynôme cubique.
2. La fonction f est deux fois continûment differentiable sur $[a, b]$, et donc f et ses dérivées d'ordre 1 et 2 sont continues aux points x_i.

2.3 Représentation des valeurs de la dérivée seconde

Pour faciliter les calculs, nous allons représenter une spline cubique naturelle en utilisant ses valeurs et ses dérivées secondes aux points $x_i,\ i = 1, \cdots, n$. Cette méthode est appelée la représentation des valeurs des dérivées secondes.
Supposons que f est une spline cubique naturelle de noeuds $x_1, \cdots, x_n$. Par définition

$$f(x) = a_i(x - x_i)^3 + b_i(x - x_i)^2 + c_i(x - x_i) + d_i,\ x_i \leq x < x_{i+1},$$

c-à-d f est un polynôme cubique sur chaque sous intervalle $[x_i, x_{i+1}[$, notons par P_i ces polynômes, alors

$P_i :\ [x_i, x_{i+1}[\longrightarrow \mathbb{R},\ P_i(x) = a_i(x - x_i)^3 + b_i(x - x_i)^2 + c_i(x - x_i) + d_i,$
f peut alors s'écrire sous la forme :

$$f(x) = \sum_{i=1}^{n} P_i(x)\ 1_{[x_i, x_{i+1}[}(x).$$

La dérivée première de f est :

$$f'(x) = 3a_i(x - x_i)^2 + 2b_i(x - x_i) + c_i,\ x_i \leq x < x_{i+1},$$

ou bien

$$f'(x) = \sum_{i=1}^{n} P_i'(x)\ 1_{[x_i, x_{i+1}[}(x),$$

où

$$P_i'(x) = 3a_i(x - x_i)^2 + 2b_i(x - x_i) + c_i.$$

La dérivée seconde de f est

$$f''(x) = 6a_i(x - x_i) + 2b_i,\ x_i \leq x < x_{i+1},$$

ou bien

$$f''(x) = \sum_{i=1}^{n} P_i''(x)\ 1_{[x_i, x_{i+1}[}(x),$$

où

$$P_i''(x) = 6a_i(x - x_i) + 2b_i.$$

On pose $h_{i-1} = x_i - x_{i-1}$ et $y_i = f(x_i)$.

- (C_1) ***: La condition que les polynômes adjacents doivent être égaux au point*** (x_i, y_i) ***(i.e*** f ***est continue) est exprimée par :***

$$P_{i-1}(x_i) = P_i(x_i) \Leftrightarrow a_{i-1}h_{i-1}^3 + b_{i-1}h_{i-1}^2 + c_{i-1}h_{i-1} + d_{i-1} = d_i = y_i.$$

  Ce qui donne

$$d_i = y_i.$$

- (C_2) ***: La condition que les dérivées premières doivent être égales au point*** (x_i, y_i) ***est exprimée par :***

$$P_{i-1}'(x_i) = P_i'(x_i) \Leftrightarrow 3a_{i-1}h_{i-1}^2 + 2b_{i-1}h_{i-1} + c_{i-1} = c_i.$$

- (C_3) ***: La condition que les dérivées secondes doivent être égales au point*** (x_i, y_i) ***est exprimée par :***

$$P_{i-1}''(x_i) = P_i''(x_i) \Leftrightarrow 6a_{i-1}h_{i-1} + 2b_{i-1} = 2b_i.$$

- (C_4) ***: La condition que les dérivées seconde et troisième doivent être égales à zéro aux bords est exprimée par :***

$$P_1''(x_1) = P_n''(x_n) = 0 \Leftrightarrow d_1 = d_n = 0,$$

  et

$$P_1'''(x_1) = P_n'''(x_n) = 0 \Leftrightarrow a_1 = a_n = 0.$$

La continuité de f et ses deux dérivées implique différentes relations entre les coefficients.

– La condition (C_1) donne :

$$c_{i-1} = \frac{y_i - y_{i-1}}{h_{i-1}} - a_{i-1}h_{i-1}^2 - b_{i-1}h_{i-1} \tag{2.1}$$

– La condition (C_3) donne :

$$a_{i-1} = \frac{b_i - b_{i-1}}{3h_{i-1}} \tag{2.2}$$

Remplaçons (2.2) dans (2.3), on obtient :

$$c_{i-1} = \frac{y_i - y_{i-1}}{h_{i-1}} - \frac{1}{3}(b_i + 2b_{i-1}h_{i-1}) \tag{2.3}$$

Remplaçons (2) et (3) dans la condition (C_2) :

$$\frac{1}{3}[b_{i-1}h_{i-1} + 2b_i(h_{i-1} + h_i) + b_{i+1}h_i] = y_{i-1}\frac{1}{h_{i-1}} - y_i(\frac{1}{h_i} + \frac{1}{h_{i-1}}) + y_{i+1}\frac{1}{h_i}.$$

Nous avons

$$P_i(x_i) = y_i, \quad et \quad P_i''(x_i) = 2b_i,$$

Notons par $F = (F_1, \cdots, F_n)^t$ et $\Gamma = (\gamma_2, \cdots, \gamma_{n-1})$, tels que

$$F_i = f(x_i) \quad et \quad \gamma_i = f''(x_i), \ avec \ \gamma_1 = \gamma_n = 0$$

La spline cubique peut être construite complètement avec F et Γ.
En effet, soit les matrices Q et R définies comme suit :

– Q est une matrice de taille $n \times (n-2)$ de composantes q_{ij} :

$$\begin{aligned} q_{j-1,j} &= \frac{1}{h_{j-1}}, \\ q_{jj} &= -(\frac{1}{h_{j-1}} + \frac{1}{h_j}), \\ q_{j+1,j} &= \frac{1}{h_j} \\ q_{ij} &= 0, \quad si \ |i-j| \geq 2. \end{aligned}$$

– R est une matrice symétrique de taille $(n-2) \times (n-2)$ de composantes r_{ij} :

$$\begin{aligned} r_{ii} &= \frac{1}{3}(h_{i-1} + h_i), \quad i = 2, \cdots, n-1, \\ r_{i,i+1} &= r_{i+1,i} = \frac{1}{6}h_i, \quad i = 2, \cdots, n-2, \\ r_{ij} &= 0, \quad si \ |i-j| \geq 2. \end{aligned}$$

R est une matrice à diagonale strictement dominante, et définie positive.

Nous pouvons donc définir la matrice K par :

$$K = QR^{-1}Q^t.$$

Théorème 2.2. [Green and Silverman][25]
Les vecteurs F et Γ définissent une spline cubique naturelle si et seulement si ils satisfont :

$$Q^tF = R\Gamma.$$

Si cette relation est vérifiée, alors :

$$\int_a^b f''(x)^2dx = \Gamma^tR\Gamma = F^tKF.$$

2.4 Interpolation

Le problème posé est de reconstituer une fonction à partir de la seule connaissance de ses valeurs en un nombre limité de points. L'interpolation consiste à trouver une fonction passant exactement par les points à notre disposition.
Soit f une fonction inconnue :

$$f : \mathbb{R} \longrightarrow \mathbb{R}$$

$$x \longrightarrow f(x) = y.$$

Nous connaissons les valeurs de f en certains points $x_1, \cdots, x_n$:

$$y_1 = f(x_1), \cdots, y_n = f(x_n).$$

Cherchons la fonction g telle que :

$$\begin{cases} y_1 = f(x_1) = g(x_1) \\ y_2 = f(x_2) = g(x_2) \\ \\ y_n = f(x_n) = g(x_n) \end{cases}$$

Commençons par le problème le plus simple. Comment interpoler deux points de coordonnées $(x_1; y_1), (x_2; y_2)$?
Voici la forme d'une fonction affine. Il nous reste à trouver les coefficients a et b pour deux points donnés.

$$g(x) = ax + b,$$

en résolvant le système,

$$\begin{cases} y_1 = ax_1 + b \\ y_2 = ax_2 + b \end{cases}$$

nous obtenons

$$a = \frac{y_2 - y_1}{x_2 - x_1},\ et\ b = y_1 - ax_1.$$

Maintenant si nous cherchons à interpoler trois points de coordonnées (x_1, y_1), (x_2, y_2), (x_3, y_3). S'ils sont non-alignés, il n'est pas possible d'utiliser une droite. Il faut complexifier notre fonction par un degré polynomial plus élevé.

$$g(x) = ax^2 + bx + c,$$

il faut résoudre le système à 3 équations et 3 inconnues (a, b, c) suivant :

$$\begin{cases} y_1 = ax_1^2 + bx_1 + c \\ y_2 = ax_2^2 + bx_2 + c, \\ y_3 = ax_3^2 + bx_3 + c, \end{cases}$$

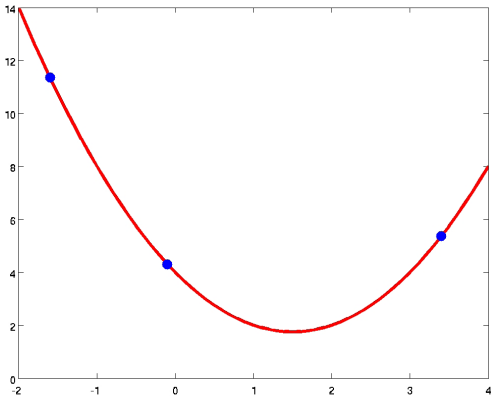

FIGURE 2.1 – Interpolation de 3 points par une parabole

Si nous avons n points à interpoler ?

2.4.1 Polynômes d'interpolation de Lagrange

Le polynôme de Lagrange associé au point (x_i, y_i) est :

$$l_i(x) = \prod_{j=1, j\neq i}^{n} \frac{x - x_j}{x_i - x_j},$$

$l_i(x)$ est de degré $n-1$.
On peut vérifier que

$$l_i(x_i) = 1 \; et \; l_i(x_j) = 0.$$

Avec ces polynômes, nous pouvons interpoler tout ensemble de n points par un polynôme de degré $n-1$.

Théorème 2.3. (Lagrange)[**Barbillon**][**1**]
Le polynôme

$$L(x) = \sum_{i=1}^{n} l_i(x) y_i,$$

est l'unique polynôme de degré $n-1$ vérifiant

$$L(x_i) = y_i, \quad i = 1, \cdots, n.$$

Remarque 2.1. Dans la méthode d'interpolation par les polynômes de Lagrange, où le degré du polynôme est en fonction du nombre de point à interpoler. En effet, quand fait l'interpolation de n point $(x_i, y_i), i = 1, \cdots, n$, le degré du polynôme de Lagrange est de $n-1$, ainsi si l'on dispose d'un nombre grand de points à interpoler le polynôme sera de degré grand et le nombres d'opérations est grand.
Les auteurs ont développé une méthode d'interpolation, dite méthode des fonctions splines interpolantes, cette méthode consiste à interpoler des points sans prendre en consideration le nombre de points, puisque elle se base sur la minimisation d'un critère qui est en fonction de la spline. D'où, la solution est un polynôme de degré indépendant du nombre de points à interpoler.

Définition 2.4. Espace de Sobolev
Une fonction f est dite absolument continue, s'il existe un réel a et une fonction g intégreble tels que :

$$f(x) = \int_a^x g(t)dt.$$

L'espace de sobolev $W_2^r[a,b]$ est l'ensemble des fonctions f définies sur $[a,b]$ telles que :

$$f^{(k)} \text{ est absolument continue, pour } k = 0, \cdots, r-1,$$

et

$$f^{(r)} \text{ est de carré intégrable.}$$

C-à-d

$$W_2^r = \{f : [a,b] \to \mathbb{R}; f^{(k)} \text{ abs. con pour } k = 0, \cdots, r-1, \; et \; \int_a^b (f^{(r)}(x))^2 < \infty\}.$$

2.4.2 Interpolation par spline cubique naturelle

Dans cette section, nous présentons le problème d'interpolation par fonction lisse dans $W_2^2[a,b]$. Résoudre un problème d'interpolation pour un ensemble de points (x_i, y_i), c'est chercher une fonction "régulière" f qui satisfait :

$$f(x_i) = y_i, \quad i = 1, \cdots, n.$$

Les polynômes d'interpolation de Lagrange sont une solution particulière à ce problème. Les splines d'interpolation en sont une autre.

2.4.3 Existence et unicité des splines d'interpolation

Théorème 2.4. [Thomas-Agnan][64]
Étant donnés n points (x_i, y_i) d'abscisses distinctes dans l'intervalle $[a,b]$ et $n \geq r$. Il existe une fonction et une seule f de l'espace de Sobolev $W_2^r[a,b]$ telle que :

1. f satisfait les conditions d'interpolation

$$f(x_i) = y_i, \quad i = 1, \cdots, n.$$

2. f minimise la quantité $\int_a^b f^{(r)}(x)^2 dx$ dans l'ensemble des fonctions $W_2^r[a,b]$ qui satisfait les conditions d'interpolation.

De plus, cette fonction est une spline polynomiale naturelle d'ordre $(2r)$ ayant pour noeuds les points $x_1, \cdots, x_n$.
Le théorème suivant montre que l'interpolation par spline cubique naturelle est l'unique qui minimise $\int_a^b f''(x)^2 dx$ par rapport à toutes les fonctions dans $W_2^2[a,b]$.

Théorème 2.5. [Cao][7]
Supposons que $n \geq 2$ et $\hat{m}_\lambda$ est la spline cubique naturelle pour les valeurs $y_1, \cdots, y_n$ aux points $x_1, \cdots, x_n$, où $a < x_1 < \cdots < x_n < b$. Soit $\tilde{m}$ une fonction dans $W_2^2[a,b]$ telle que $\tilde{m}(x_i) = y_i,\ i = 1, \cdots, n$. Alors

$$\int_a^b \tilde{m}''(x)^2 dx \geq \int_a^b \hat{m}''(x)^2 dx.$$

On a l'égalité si et seulement si $\tilde{m}$ et $\hat{m}$ sont identiques.

2.5 Les splines de lissage pour la courbe de régression non paramétrique

Rappelons que l'interpolation consiste à trouver une fonction passant exactement par les points que l'on souhaite interpoler tandis que la régression ajuste au mieux une fonction simple à ces points. Nous ne souhaitons plus ici passer exactement par ces points, mais en être "proche" et avoir une fonction g simple. Cette problématique est sensée notamment parce qu'il arrive souvent, en pratique, que les observations des évaluations soient bruitées. C'est à dire que nous observons l'évaluation $f(x_i)$ plus une petite erreur. Celle-ci peut provenir d'une imprécision de mesure, par exemple.

Beaucoup d'auteurs proposent d'ajuster les données avec les polynômes splines [De Boor][4]. L'origine des splines de lissage est dûe à [Whittaker (1923)][78] et ont été développées par [Schoenberg (1964)][55] et [Reinsh][53] ; voir aussi les monographies de [Eubank][20] et [Wahba (1990)][74].
Supposons que $a < x_1 < \cdots < x_n < b$, et considérons le modèle non paramétrique suivant :

$$y_i = m(x_i) + \epsilon_i; \tag{2.4}$$

où m est une fonction lisse inconnue dans $W_2^2[a,b]$, $y_i,\ i = 1, \cdots, n$ sont des valeurs observées de la variable réponse Y, $x_i,\ i = 1, \cdots, n$ sont des valeurs observées de la variable X, et $\epsilon_i,\ i = 1, \cdots, n$ sont des erreurs normalement distribuées, de moyenne nulle et de variance σ^2. L'idée générale est d'estimer la fonction lisse m par balancement entre un bon "fit" (ajustement aux données) et une estimation lisse. La caractéristique par rapport à d'autres méthodes de lissage est l'adaptation au changement rapide de la courbure de la fonction de régression,c-à-d de faire une interpolation sous hypothèse de minimiser les ondulations de la fonction. On est ainsi conduit à rajouter une pénalisation qui contrôle les oscillations.
Une mesure courante de la fidélité des données pour une courbe m est la somme résiduelle au carré,

$$\sum_{i=1}^{n} (y_i - m(x_i))^2. \tag{2.5}$$

La minimisation de (2.5) par rapport à la fonction m apportera aux données une courbe interpolante avec des variations trop rapides (trop ondulée). L'approche par splines de

lissage fait intervenir une mesure de lissage fonction de m est définie par :

$$\int_a^b m^{(r)}(x)^2 \, dx. \tag{2.6}$$

L'estimateur de m est donné par la minimisation de la somme convexe :

$$(1-q) \sum_{i=1}^{n}(y_i - m(x_i))^2 \; + \; q \int_a^b m^{(r)}(x)^2 \, dx; \tag{2.7}$$

pour $0 < q < 1$. Une estimation optimale peut être obtenue en minimisant cette fonctionnelle sur $W_2^r[a,b]$. Posons $\lambda = \frac{q}{1-q}$, il est equivalent d'estimer m par la fonction $\hat{m}_\lambda$ qui minimise la fonctionnelle suivante :

$$\sum_{i=1}^{n}(y_i - m(x_i))^2 \; + \; \lambda \int_a^b m^{(r)}(x)^2 \, dx; \quad \lambda > 0 \tag{2.8}$$

par rapport à toutes les fonctions m dans $W_2^r[a,b]$.
La solution est l'estimateur spline de lissage $\hat{m}_\lambda$ pour la courbe de régression qui sera étudié dans ce chapitre.
Le paramètre λ prenant ses valeur dans $[0, +\infty)$ contrôle le compromis entre le lissage et le bon ajustement. Pour cette raison, il est souvent référé au paramètre de lissage. La valeur de λ permet de déterminer la flexibilité de l'estimateur. Quand la valeur de λ est proche de 0 (i.e $q \to 0$) l'estimateur est flexible, car on diminue l'apport de la quantité de lissage donnant un estimateur interpolant les données. Inversement, quand la valeur de λ augmente (i.e $q \to 1$) on donne plus d'importance à la deuxième partie, ce qui oblige l'intégrale à être plus petite et donc l'estimateur à être plus lisse.

2.5.1 Spline de lissage

Les splines de lissage déterminent la valeur de l'estimateur en minimisant un critère bien précis. Celui-ci combine la mesure classique de la qualité de l'ajustement, la somme des résidus au carré, et une mesure de la quantité de lissage.

Supposons que $x_1 < x_2 < \cdots < x_n$, l'estimation de la fonction m par les splines de lissage résulte comme solution du problème de minimisation suivant :
Trouver la fonction $\hat{m}_\lambda \in \; W_2^r[a, \; b]$ qui minimise la somme des carrés des résidus pénalisée (l'erreur quadratique pénalisée)

$$S(m) = \; \sum_{i=1}^{n}(y_i - m(x_i))^2 \; + \; \lambda \int_a^b (m^{(r)}(x))^2 dx, \tag{2.9}$$

c-à-d

$$\hat{m}_\lambda(x) = \arg\min \ \sum_{i=1}^{n}(y_i - m(x_i))^2 \ + \ \lambda \int_a^b (m^{(r)}(x))^2 dx, \tag{2.10}$$

où $m^{(r)}$ est la dérivée d'ordre r de la fonction m.
Le premier terme est la somme des carrés des résidus, mesure la fidélité des données.
Le second terme est pris comme mesure de lissage de m.

Le théorème suivant montre que la solution du problème (2.9) est une fonction spline d'ordre $(2r-1)$.

Théorème 2.6. [Schoenberg][55]
Le minimum du problème (2.9) admet une solution unique $\hat{m}_\lambda(x)$ qui est une fonction spline dans l'ensemble $\mathcal{S}_{2r}(x_1, \cdots, x_n)$.

Nous allons traiter le cas où $r = 2$, déjà vue précédemment, ce cas est souvent utilisé car il donne un algorithme très simple pour la construction de la fonction spline. En plus, les splines cubiques sont facilement évaluées et donnent "en général" des résultats satisfaisants.
Poser $r = 2$ revient à résoudre le problème de minimisation suivant : Trouver la fonction m qui minimise

$$S(m) = \ \sum_{i=1}^{n}(y_i - m(x_i))^2 \ + \ \lambda \ \int_a^b (m''(x))^2 \ dx, \tag{2.11}$$

par rapport à toutes les fonctions m dans $W_2^2[a, b]$.

[Reinsh][49] montre que la solution du problème (2.11) est une spline de lissage cubique naturelle aux noeuds $x_i, \ i = 1, \cdots, n$.

2.5.2 Existence et unicité de la spline de lissage minimisante

Théorème 2.7. [Thomas-Agnan][64]
Étant donnés n points (x_i, y_i), d'abscisses distinctes dans l'intervalle $[a, b]$ et un réel $\lambda > 0$, il existe une fonction et une seule $\hat{m}_\lambda$ de l'espace de Sobolev $W_2^2[a, b]$ qui minimise la quantité $\mathcal{S}(m)$ dans l'ensemble des fonctions $W_2^2[a, b]$. De plus, cette fonction est une spline polynomiale d'ordre 4 ayant pour noeuds les points $x_1, \cdots, x_n$.

Preuve

Soit $Y = (y_1, \cdots, y_n)^t,\ M = (M_1, \cdots, M_n)^t$ tel que $M_i = \hat{m}_\lambda(x_i)$. Donc on a :

$$\sum_{i=1}^{n}(y_i - \hat{m}_\lambda(x_i))^2 = (Y - M)^t(Y - M)$$

D'après le théorème (2.2) on a :

$$\int_a^b \hat{m}''_\lambda(x)^2 dx = M^t K M.$$

D'où

$$\begin{aligned} S(m) &= (Y - M)^t(Y - M) + \lambda M^t K M \\ &= M^t(I + \lambda K)M - 2Y^tY - 2Y^tM. \end{aligned}$$

Dérivons $S(m)$ par rapport à M :

$$\begin{aligned} \frac{dS(m)}{dM} &= 2(I + \lambda K)M - 2Y \\ \frac{dS(m)}{dM} &= 0 \Leftrightarrow M = (I + \lambda K)^{-1}Y. \end{aligned}$$

La dérivée seconde de S donne

$$\frac{d^2S(m)}{dM^2} = 2(I + \lambda K),$$

Comme λK est non négative, alors $(I + \lambda K)$ est strictement définie positive. D'où

$$\hat{M} = (I + \lambda K)^{-1}Y$$

définit bien un minimum.

Théorème 2.8. [Cao][7]
Supposons que $n \geq 0$ et que le paramètre de lissage λ est positif, alors $\hat{m}_\lambda$ est une spline cubique naturelle telle que :

$$M = (I + \lambda K)^{-1}Y,$$

et pour toute m dans $W_2^2[a, b]$, on a

$$S(\hat{m}_\lambda) \leq S(m).$$

Remarque 2.2. L'estimateur spline de lissage est linéaire au sens de la définition (1.12) du chapitre 1.

2.5.3 Propriétés de l'estimateur splines de lissage

Soit $\hat{m}_\lambda$ l'estimateur spline, alors $\hat{m}_\lambda$ est défini comme suit :

$$\hat{m}_\lambda = (I + \lambda K)^{-1} Y = A_\lambda Y.$$

1- Biais de l'estimateur

Le biais est défini par :

$$Biais\ (\hat{m}_\lambda, m)\ = \mathbb{E}\hat{m}_\lambda - m = (A_\lambda - I)m.$$

En effet,

$$\begin{aligned}
\mathbb{E}\hat{m}_\lambda - m &= \mathbb{E}A_\lambda Y - m \\
&= \mathbb{E}A_\lambda^i Y - m,\ i = 1, \cdots, n \\
&= A_\lambda^i \mathbb{E}Y - m,\ i = 1, \cdots, n \\
&= A_\lambda^i m - m,\ i = 1, \cdots, n \\
&= A_\lambda m - m = (A_\lambda - I)m.
\end{aligned}$$

2- La variance de l'estimateur

La variance de $\hat{m}_\lambda$ est définie comme suit :

$$\mathbb{V}ar\ (\hat{m}_\lambda)\ = \mathbb{E}\|\hat{m}_\lambda - \mathbb{E}\hat{m}\|^2 = \sigma^2 tr(A_\lambda^t A_\lambda).$$

En effet,

$$\begin{aligned}
\mathbb{V}ar(\hat{m}) &= \mathbb{E}\|\hat{m}_\lambda - \mathbb{E}\hat{m}_\lambda\|^2 \\
&= \mathbb{E}\|A_\lambda Y - A_\lambda m\|^2 \\
&= \mathbb{E}\sum_{i=1}^{n}(A_\lambda)_{ij}^2(y_i - m_{x_i})^2, j = 1, \cdots, n \\
&= \mathbb{E}\sum_{i=1}^{n}(A_\lambda)_{ij}^2(y_i^2 + m_{x_i}^2 - 2y_i m(x_i))^2, j = 1, \cdots, n \\
&= \sum_{i=1}^{n}(A_\lambda)_{ij}^2\mathbb{E}(y_i^2 + m_{x_i}^2 - 2y_i m(x_i)), j = 1, \cdots, n \\
&= \sum_{i=1}^{n}(A_\lambda)_{ij}^2[\mathbb{E}y_i^2 + \mathbb{E}m_{x_i}^2 - 2\mathbb{E}(y_i m(x_i))], j = 1, \cdots, n \\
&= \sum_{i=1}^{n}(A_\lambda)_{ij}^2[\mathbb{E}y_i^2 + m_{x_i}^2 - 2m(x_i)\mathbb{E}y_i], j = 1, \cdots, n \\
&= \sum_{i=1}^{n}(A_\lambda)_{ij}^2[\mathbb{E}y_i^2 + m_{x_i}^2 - 2m(x_i)m(x_i)], j = 1, \cdots, n \\
&= \sum_{i=1}^{n}(A_\lambda)_{ij}^2[\mathbb{E}y_i^2 + m_{x_i}^2 - 2m(x_i)^2], j = 1, \cdots, n \\
&= \sum_{i=1}^{n}(A_\lambda)_{ij}^2[\mathbb{E}y_i^2 - m_{x_i}^2], j = 1, \cdots, n \\
&= \sum_{i=1}^{n}(A_\lambda)_{ij}^2[\mathbb{E}y_i^2 - \mathbb{E}(y_i)^2], j = 1, \cdots, n \\
&= \sum_{i=1}^{n}(A_\lambda)_{ij}^2\sigma^2, j = 1, \cdots, n \\
&= \sigma^2\sum_{i=1}^{n}(A_\lambda)_{ij}^2, j = 1, \cdots, n \\
&= \sigma^2 tr(A_\lambda^t A_\lambda).
\end{aligned}$$

2.5.4 Propriétés asymptotiques de l'estimateur

1- Convergence en moyenne quadratique

Théorème 2.9. [Wahba][69]

Supposons que $\sigma^2 \neq 0$.

Pour avoir la convergence il faut que λ décroît avec n. Si cette condition est vérifiée, on

a alors :

$$\hat{m}_\lambda(x) \longrightarrow m(x), \text{ en moyenne quadratique.}$$

Dans le modèle (2.4), [Craven et Wahba][15] ont majoré l'erreur moyenne quadratique de l'estimateur $\hat{m}_\lambda$ aux points d'observations $x_1, \cdots, x_n$. Avec les notations $m = (m(x_1), \cdots, m(x_n))^t$ et $\hat{m}_\lambda = (\hat{m}_\lambda(x_1), \cdots, \hat{m}_\lambda(x_n))^t$ cette erreur est appelée MASE (pour Mean Average Squared Error), s'écrit

$$n^{-1}\mathbb{E}\|m - \hat{m}_\lambda\|^2.$$

Théorème 2.10. [Craven et Wahba][15]
Dans le modèle (2.4), on suppose que

- Les points d'observations sont tel que

$$\int_0^{x_j} w(x)dx = \frac{j}{n} \; pour \; 1 \leq j \leq n$$

  où w est une fonction strictement positive et continue sur $[0,\ 1]$;
- La fonction m est dans l'espace de Sobolev $W_2^r[0,\ 1]$ pour un $r \geq 1$ donné ;
- Les erreurs $\epsilon_1, \cdots, \epsilon_n$ sont décorelées et de variance commune $\sigma^2 > 0$.

En supposant que m n'est pas un polynôme de degré $< r$, l'estimateur spline de lissage $\hat{m}_\lambda$ d'ordre $2r$ est consistant si et seulement si $\lambda = \lambda(n) \to 0$ et $\lambda n^{2r} \to \infty$ lorsque $n \to \infty$. D'autre part, la MASE (pour Mean Average Squared Error) de cet estimateur, vérifie

$$\frac{1}{n}\mathbb{E}\|m - \hat{m}_\lambda\|^2 = O(\lambda) + O(\frac{1}{\lambda^{\frac{1}{2r}} n}).$$

Considérons maintenant l'erreur moyenne quadratique intégrée, notée MISE. Dans le cas du modèle (2.4), sous l'hypothèse du bruit blanc, [Ragozin(1983)][48] a majoré la MISE des dérivées de la fonction par les les dérivées correspondantes de la spline de lissage $\hat{m}_\lambda$.

Théorème 2.11. [Ragozin][48]
Dans le modèle (2.4), on suppose que

- Les points d'observations sont équidistants : $x_j = \frac{j}{n},\ 1 \leq i \leq n$;
- La fonction m est dans l'espace de Sobolev $W_2^k[0,\ 1]$ avec $0 < k \leq r$;
- Les erreurs ϵ_i sont centrées, décorrelées et de variance commune σ^2.

Soit $\hat{m}_\lambda$ l'estimateur spline de lissage d'ordre $2r$ de m. Si le paramètre de lissage $\lambda = \lambda(n)$ est choisi de sorte que $(n\lambda^{\frac{1}{2r}})^{-1} \leq c$ pour une constante c et pour tout $n \geq r$, alors la MISE de la j-ème dérivée $(j < k)$ de $\hat{m}_\lambda$ vérifie

$$\mathbb{E}|m - \hat{m}_\lambda|_j^2 = \mathbb{E}\int_0^1 (m^j(x) - \hat{m}_\lambda^j(x))^2 dx \leq P(\lambda + n^{-2r})^{\frac{k-j}{r}}|m|_k^2 + \frac{Q\sigma^2}{n^{\frac{\lambda(2j+1)}{2r}}}$$

Pour des constantes, P et Q ne dépendent que de r, k et C. En particulier si $\lambda(n) \sim n^{-2r/(2k+1)}$ lorsque $n \to \infty$, alors

$$\mathbb{E}|m - \hat{m}_\lambda|_j^2 = O(n^{-2^{(k-j)/(2k+1)}}).$$

2.5.5 Choix du paramètre de lissage

Le paramètre de lissage λ varie entre 0 et $+\infty$. La solution du problème (2.8) varie entre interpolation et modèle linéaire.
Lorsque $\lambda \to +\infty$: La pénalité de régularité domine et l'estimateur est forcé d'être constant.
Lorsque $\lambda \to 0$: La pénalité de régularité disparaît et l'estimateur interpole les données.

Cependant, obtenir un choix approprié de λ est un problème important.

On présente ici quelque méthode de sélection du paramètre de lissage λ.
Soit les notations, $Y = (y_1, \cdots, y_n)^t$ et $\hat{m}_\lambda = (\hat{m}_\lambda(x_1), \cdots, \hat{m}_\lambda(x_n))^t$.

Rappelons que les valeurs de la spline de lissage optimale dépendent linéairement des observations,

$$\hat{m}_\lambda = A_\lambda Y$$

avec

$$A_\lambda = (I + \lambda \mathrm{K})^{-1}$$

La matrice A_λ est appelée matrice chapeau ou matrice de lissage. Elle permet de relier le vecteur des observations y_i au vecteur des predicteurs $\hat{y}_i = \hat{m}_\lambda(x_i)$, le risque $R(m, \hat{m}_\lambda)$ est définit par

$$R(m, \hat{m}_\lambda) = \frac{1}{n}\mathbb{E}\|m - \hat{m}_\lambda\|^2,$$

où $\|V\| = \sqrt{V^t V}$ est la norme usuelle de L_2 pour tout vecteur V, la trace de la matrice A est notée $tr(A)$.

1- Méthode de la validation croisée (CV)

Soit $(A_\lambda)_{ii}$ le $i^{\text{ème}}$ élément de la diagonale associée à la matrice de lissage A_λ.

Théorème 2.12. **[Moulines]**[43] Le score de la validation croisée vérifie :

$$CV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}(\frac{y_i - \hat{m}_\lambda(x_i)}{1-(A_\lambda)_{ii}})^2. \tag{2.12}$$

λ est choisi de telle manière qu'il minimise $CV(\lambda)$.

La preuve du théorème découle du lemme suivant :

Lemme 1. **[Moulines]**[43] Soit λ et $i \in \{1, \cdots, n\}$ donnés. Notons m^{-i} le vecteur de composants $m_k^{-i} = \hat{m}_\lambda^{-i}(x_k)$. Définissons Y^* le vecteur

$$\begin{aligned} y_j^* &= y_j \text{ pour } j \neq i, \text{ et} \\ y_i^* &= \hat{m}_\lambda^{-i}(x_i). \end{aligned}$$

Alors

$$m^{-i} = A_\lambda Y^*. \tag{2.13}$$

2- Validation croisée généralisée (GCV)

L'idée de base de la validation croisée généralisée est de remplacer les dénominateurs $1-(A_\lambda)_{ii}$ dans la validation croisée $CV(\lambda)$ par leurs moyenne $1-\frac{1}{n}tr(A_\lambda)$ donnant ainsi le score de la validation croisée généralisée

$$GCV(\lambda) = \frac{1}{n}\frac{\sum_{i=1}^{n}(y_i - \hat{m}_\lambda(x_i))^2}{(1-\frac{1}{n}tr(A_\lambda))^2}. \tag{2.14}$$

Comme pour la validation croisée, λ est choisi de telle manière qu'il minimise $GCV(\lambda)$.

3- Estimation du risque sans biais

Un calcul direct du risque conduit à une décomposition biais-variance pour le risque $R(m, \hat{m}_\lambda)$.

$$R(m, \hat{m}_\lambda) = \frac{1}{n}\mathbb{E}\|m - A_\lambda Y\|^2 = \frac{1}{n}\|(I - A_\lambda)m\|^2 + \frac{\sigma^2}{n}tr(A_\lambda A_\lambda^t).$$

L'idée principale de l'estimation du risque sans biais est de minimiser le membre droit de l'équation précédente en se basant sur les observations. Notons que le risque dépend du

biais $\|(I - A_\lambda)m\|^2$, L'idée principale est de remplacer $\|(I - A_\lambda)m\|^2$ par son estimateur sans biais, qui peut être calculé par

$$\mathbb{E}\|(I - A_\lambda)Y\|^2 = \|(I - A_\lambda)m\|^2 + \sigma^2 n - 2\sigma^2 tr(A_\lambda) + \sigma^2 tr(A_\lambda A_\lambda^t).$$

Par conséquent

$$\|(I - A_\lambda)m\|^2 = \mathbb{E}\|(I - A_\lambda)Y\|^2 - \sigma^2 n + 2\sigma^2 tr(A_\lambda) - \sigma^2 tr(A_\lambda A_\lambda^t).$$

Le risque devient

$$R(m, \hat{m}_\lambda) = \frac{1}{n}\mathbb{E}\|(I - A_\lambda)Y\|^2 - \sigma^2 + 2\frac{\sigma^2}{n}tr(A_\lambda).$$

Puisque l'espérance n'est pas estimable à partir des observations, la seule chose qu'on puisse faire est de prendre $\|(I - A_\lambda)Y\|^2 - \sigma^2 + 2\frac{\sigma^2}{n}tr(A_\lambda)$ comme estimateur du biais. Ainsi, le choix de λ basé sur l'estimation du risque sans biais est donné par la formule suivante :

$$\hat{\lambda} = \arg\min_{\lambda>0}\{\|Y - A_\lambda Y\|^2 + \frac{2}{n}\sigma^2 tr(A_\lambda)\}. \tag{2.15}$$

Dans la pratique, cette méthode a un petit inconvénient, car $\hat{\lambda}$ depend de σ^2 qui est difficilement connu en pratique, par consequent σ^2 sera estimé par

$$\hat{\sigma}^2 = \|Y - A_\lambda Y\|^2/n.$$

On insère cet estimateur dans la formule de $\hat{\lambda}$ et on obtient

$$\hat{\lambda} = \arg\min_{\lambda>0}\{\|Y - A_\lambda Y\|^2\ [1 + \frac{2}{n}tr(A_\lambda)]\}. \tag{2.16}$$

Remarque 2.3. Soit $\hat{\lambda}$ défini précédemment (2.15)

Dans la pratique une modification légère du critère est typiquement utilisée. Ainsi, la méthode des splines fournit un lissage des données visible si $tr(A_\lambda) << n$. Dans ce cas l'ensemble d'admissibilité de λ peut être réduit à $\frac{tr(A_\lambda)}{n} << 1$ et par conséquent, d'après le développement de Taylor on a

$$[1 + \frac{2}{n}tr(A_\lambda)] \approx [1 - \frac{2}{n}tr(A_\lambda)]^{-1} \approx [1 - \frac{1}{n}tr(A_\lambda)]^{-2}.$$

Ainsi, on obtient la formule suivante :

$$\hat{\lambda} = \arg\min_{\lambda>0} \frac{\|Y - A_\lambda Y\|^2}{[1 - \frac{tr(A_\lambda)}{n}]^2}.$$

Nous obtenons la méthode validation croisée généralisée. et le paramètre de lissage λ est choisi tel qu'il minimise le score de la validation croisée généralisée défini par

$$\frac{1}{n}\frac{\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2}{(1 - \frac{1}{n}tr(A))^2}.$$

2.6 Conclusion

Nous avons présenté dans ce chapitre, la méthode des fonctions splines. Nous l'avons ensuite appliquée à l'estimation de la courbe de régression de la moyenne. Nous avons donné les propriétés statistiques de l'estimateur (Biais, Variance et Erreur moyenne quadratique intégrée). On a exposé les différentes méthodes : validation croisée (CV), validation croisée généralisée (GCV) et la méthode du risque sans biais pour le choix du paramètre de lissage qui intervient dans l'estimation de la courbe de régression de la moyenne.

Dans le chapitre suivant nous allons comparer par simulation et sur des données réelles les deux méthodes d'estimation de la courbe de régression de la moyenne à savoir la méthode du noyau et la méthode des splines. Le critère de comparaison sera le MISE (Erreur Moyenne Quadratique Intégrée).

3

Résultats numériques

3.1 Introduction

Nous présentons dans ce chapitre, différentes simulations et un cas réel pour comparer les deux méthodes d'estimation de la courbe de régression de la moyenne $m(x)$: La méthode du noyau et la méthode des splines. La comparaison est basée sur le critère du MISE (erreur moyenne quadratique intégrée)

3.2 Algorithme de simulation

Les étapes de l'algorithme, pour comparer les deux méthodes d'estimation de la courbe de régression de la moyenne sont comme suit :

- Générer un échantillon de couple aléatoire (x_i, y_i) de taille n.
- Estimer la courbe de régression $y = m(x)+z$ par la méthode du noyau et la méthode des fonctions splines :

1. **Pour la méthode du noyau :** L'estimateur de la courbe de régression de la moyenne $y = m(x) + \epsilon$ par la méthode du noyau est :

$$\hat{m}_h(x) = \frac{\sum_{i=1}^{n} y_i K(\frac{x-x_i}{h})}{\sum_{i=1}^{n} k(\frac{x-x_i}{h})},$$

le noyau K utilisé est le noyau normal.

Le paramètre de lissage est choisi par la méthode validation croisée :

$$\hat{h} = \arg\min \sum_{i=1}^{n} (y_i - \hat{m}^{-i}(x_i))^2 w(x_i),$$

avec

$$\hat{m}^{-i}(x) = \frac{\sum_{j=1, j\neq i}^{n} y_j K(\frac{x-x_j}{h})}{\sum_{j=1, j\neq i}^{n} K(\frac{x-x_j}{h})}.$$

2. **Pour la méthode des splines :**

La courbe de régression de la moyenne $y = m(x)+\epsilon$ est estimée par les fonctions splines de lissage cubique (les polynômes qui composent la fonction splines de lissage sont de degré 3),

$$\hat{m}_\lambda(x) = A_\lambda Y,$$

où A_λ est la matrice de lissage définie par

$$A_\lambda = (I + \lambda K)^{-1}.$$

Le paramètre de lissage est choisi par la méthode validation croisée généralisée :

$$\hat{\lambda} = \arg\min \frac{1}{n} \frac{\sum_{i=1}^{n} (y_i - \hat{m}_\lambda(x_i))^2}{(1 - n^{-1} tr(A_\lambda))^2}.$$

– Calculer les erreurs (MISE) associées à chaque méthode.

– Tracer la courbe de régression et les graphes des estimateurs noyau et splines.

Les programmes utilisés dans ce travail sont calculés avec le logiciel **R**. Les simulations sont effectuées pour différentes tailles d'échantillon de plus en plus grandes

$$n \in \{50, 100, 200, 500, 700, 1000, 1500, 2000, 2500, 3000\}$$

3.3 Simulation de modèles de fonctions

considérons le modèle de régression :

$$y = m(x) + z \tag{3.1}$$

Les modèles considérés sont les suivants :

1. $m_1(x) = x + 0.5\exp(-50(x-0.5)^2)$; [Eubank][20].
2. $m_2(x) = 4.26(\exp(-3.25x) - 4\exp(-6.5x) + 3\exp(-9.75x))$; [Eubank][20].
3. $m_3(x) = \sqrt{x(1-x)}\sin(\frac{2.1\pi}{x+0.5})$, fonction de **Doppler**.

3.3.1 Modèle $m_1(x)$:

1. x est uniforme sur $[0, 1]$;
2. z sont les résidus qui sont normalement distribués de moyenne 0 et de variance $\sigma^2 = 0.0225$;
3. La courbe de régression considérée dans le modèle (3.1) est donnée par $m(x) = m_1(x)$;

L'erreur moyenne quadratique associée au modèle m_1

n	MISE noyau	MISE spline
50	0.00639922	0.00061377
100	0.00516139	0.00059959
200	0.00376109	0.00055375
500	0.00311738	0.00053171
700	0.00126655	0.00052573
1000	0.00098935	0.00051186
1500	0.00088708	0.00050892
2000	0.00074196	0.00050615
2500	0.00066193	0.00049508
3000	0.00050997	0.00049338

TABLE 3.1 – Erreur moyenne quadratique donnée par les deux méthodes associée au modèle $m_1(x)$ en fonction de la taille de l'échantillon n.

Pour $n = 50$

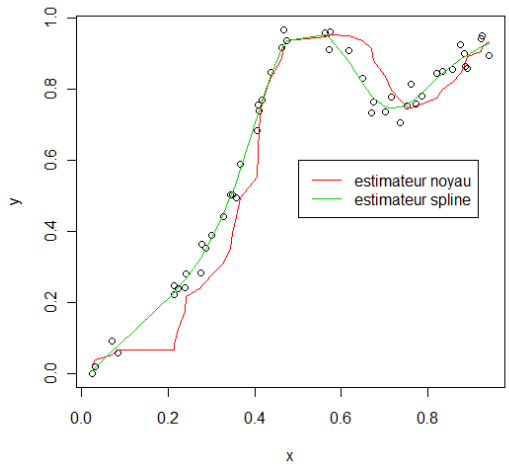

FIGURE 3.1 – Estimation de m_1, $n = 50$

Pour $n = 100$

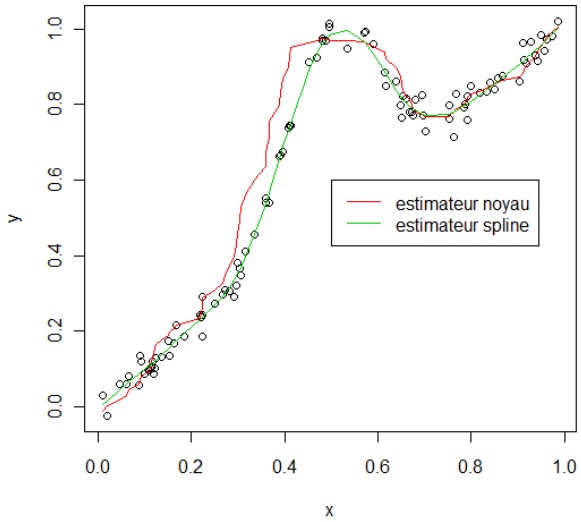

FIGURE 3.2 – Estimation de m_1, $n = 100$

Pour $n = 200$

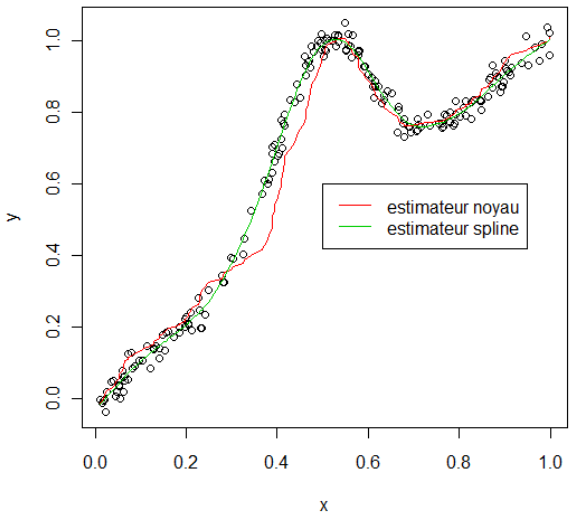

FIGURE 3.3 – Estimation de m_1, $n = 200$

Pour $n = 500$

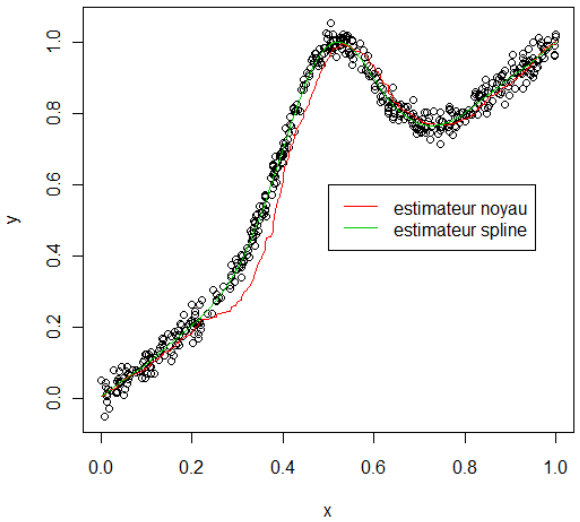

FIGURE 3.4 – Estimation de m_1, $n = 500$

Pour $n = 1000$

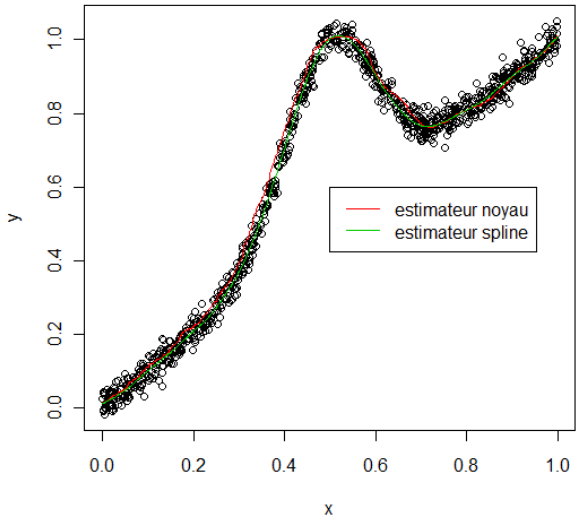

FIGURE 3.5 – Estimation de m_1, $n = 1000$

Pour $n = 2000$

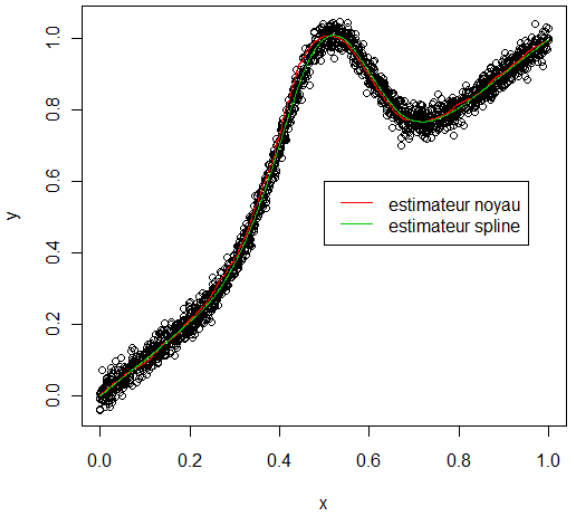

FIGURE 3.6 – Estimation de m_1, $n = 2000$

Pour $n = 3000$

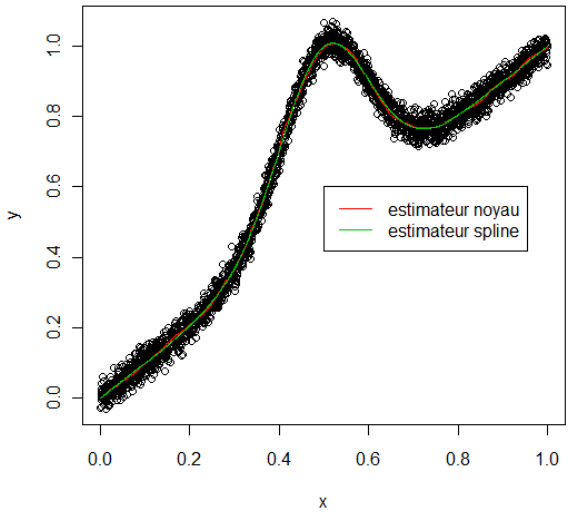

FIGURE 3.7 – Estimation de m_1, $n = 3000$

Interprétation des résultats du modèle m_1 :

Les résultats du tableau (3.1) montrent que quelque soit la taille n de l'échantillon, les erreurs associées à la méthode des fonctions splines sont meilleures que celles du noyau (MISE spline $<$ MISE noyau). Cependant, pour n grand ($n \geq 1000$), les MISE des deux méthodes sont proches. Ce qui signifie que la méthode des fonctions splines peut s'appliquer quelque soit la taille de l'échantillon, et la méthode du noyau ne s'applique que si la taille de l'échantillon est grande. Cette constatation est vérifiée graphiquement, en effet, on constate que la méthode des fonctions splines ajuste les données pour toutes les valeurs de n, par contre, la méthode du noyau ajuste les données quand $n \geq 1000$.

3.3.2 Modèle $m_2(x)$:

1. x est uniforme sur $[0, 1]$;
2. z sont les résidus qui sont normalement distribués de moyenne 0 et de variance $\sigma^2 = 0.0025$;
3. La courbe de régression considérée dans le modèle (3.1) est donnée par $m(x) = m_2(x)$;

L'erreur moyenne quadratique associée au modèle m_2

n	MISE noyau	MISE spline
50	0.02555829	0.00000679
100	0.01329819	0.00000643
200	0.00139980	0.00000640
500	0.00072893	0.00000606
700	0.00058020	0.00000600
1000	0.00019613	0.00000595
1500	0.00019445	0.00000580
2000	0.00019248	0.00000575
2500	0.00016873	0.00000575
3000	0.00008621	0.00000564

TABLE 3.2 – Erreur moyenne quadratique donnée par les deux méthodes associée au modèle $m_2(x)$ en fonction de la taille de l'échantillon n.

Pour $n = 50$

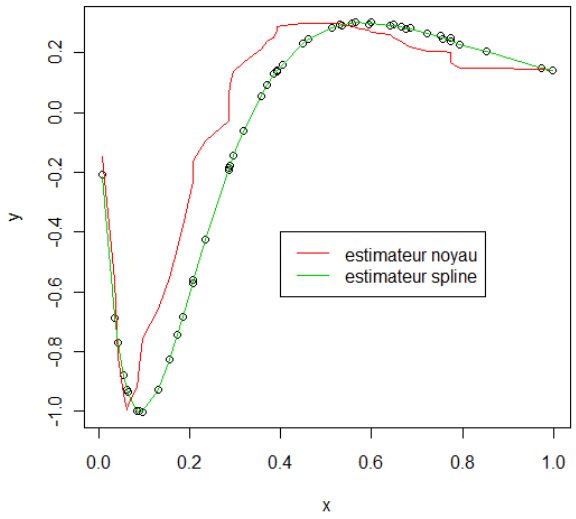

FIGURE 3.8 – Estimation de m_2, $n = 50$

Pour $n = 100$

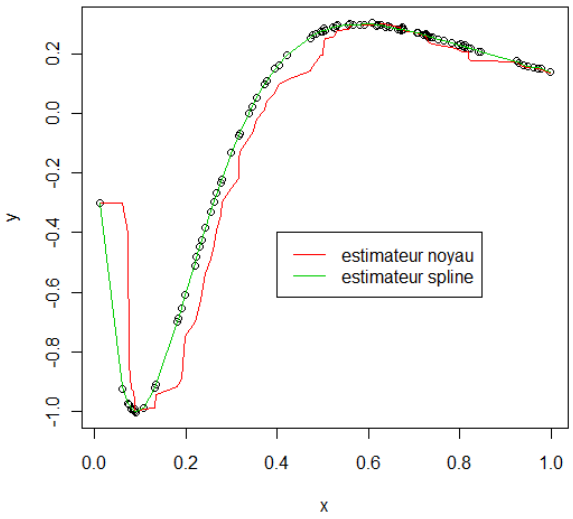

FIGURE 3.9 – Estimation de m_2, $n = 100$

Pour $n = 200$

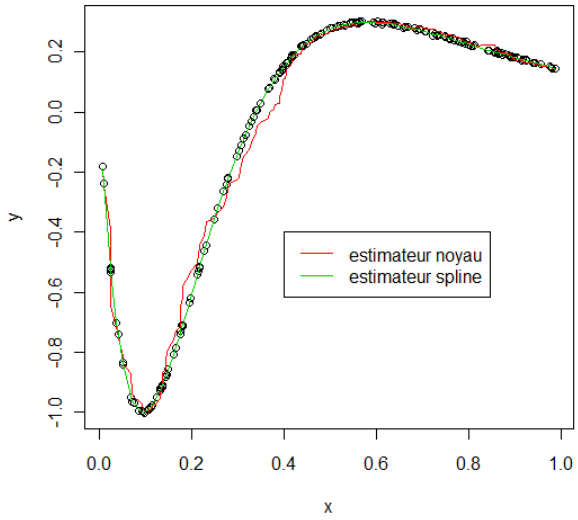

FIGURE 3.10 – Estimation de m_2, $n = 200$

Pour $n = 500$

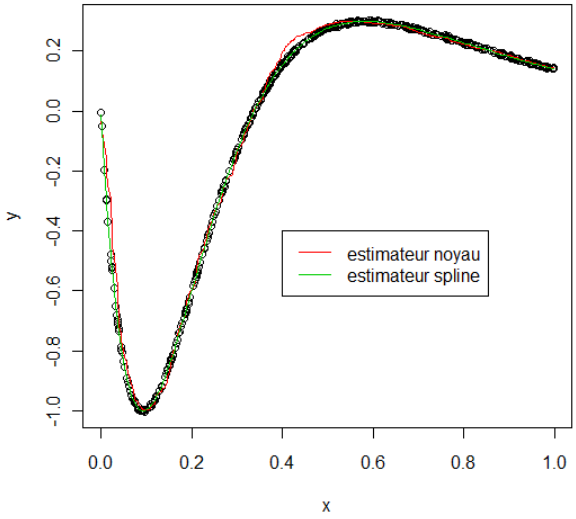

FIGURE 3.11 – Estimation de m_2, $n = 500$

Pour $n = 1000$

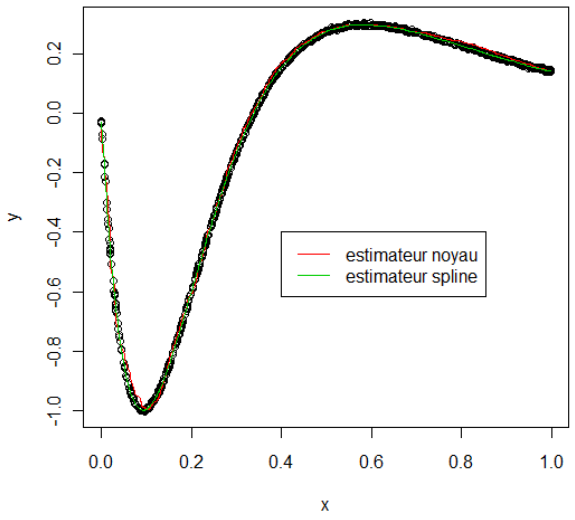

FIGURE 3.12 – Estimation de m_2, $n = 1000$

Pour $n = 2000$

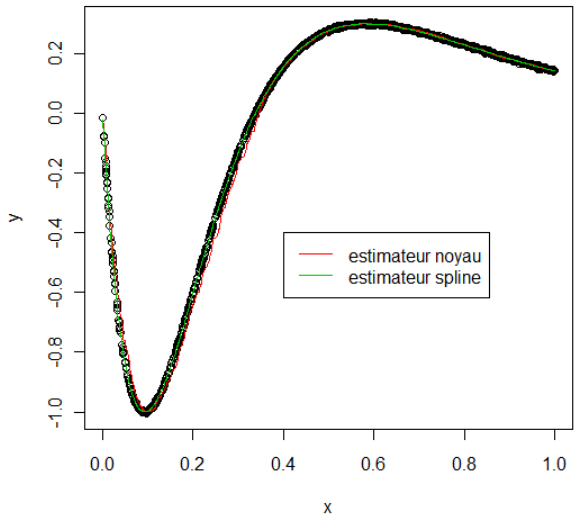

FIGURE 3.13 – Estimation de m_2, $n = 2000$

Pour $n = 3000$

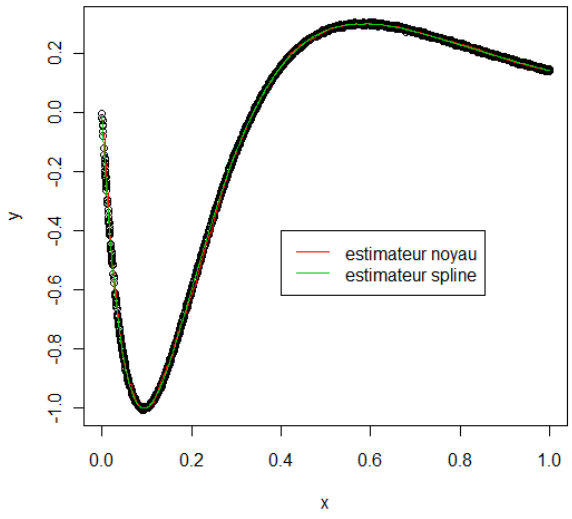

FIGURE 3.14 – Estimation de m_2, $n = 3000$

Interprétation des résultats du modèle m_2 :

Les résultats du tableau (3.2) montrent que quelque soit la taille de l'échantillon n, l'erreur associée à la méthode des fonctions splines est très petite devant la l'erreur associée à la méthode du noyau. Pour $n \in \{50, 100\}$ l'erreur MISE associée à la méthode du noyau est d'ordre 10^{-1}. Quand n augmente le MISE noyau décroît jusqu'à l'ordre 10^{-3}. Concernant l'erreur associée à la méthode des fonctions splines (MISE spline), elle est d'ordre 10^{-5} quelle que soit la taille de l'échantillon. Par conséquent, la méthode des fonctions splines donne une meilleure estimation des données que celle donnée par la méthode du noyau. cette constatation, est vérifiée aussi graphiquement. En effet, si on observe les graphes associés aux deux méthodes d'estimation, nous constatons que la méthode des fonctions splines donne de meilleurs ajustements des données que celles du noyau.

3.3.3 Modèle $m_3(x)$

1. x est uniforme sur $[0, 1]$;
2. z sont les résidus qui sont normalement distribués de moyenne 0 et de variance $\sigma^2 = 0.01$;
3. La courbe de régression considérée dans le modèle (3.1) est donnée par $m(x) = m_3(x)$;

L'erreur moyenne quadratique associée au modèle m_3

n	MISE noyau	MISE spline
50	0.02790672	0.00010490
100	0.01021975	0.00010196
200	0.00597055	0.00010116
500	0.00491645	0.00010037
700	0.00439731	0.00009964
1000	0.00058604	0.00009971
1500	0.00056320	0.00009884
2000	0.00040439	0.00009847
2500	0.00036532	0.00009741
3000	0.00024054	0.00009636

TABLE 3.3 – Erreur moyenne quadratique donnée par les deux méthodes associées au modèle $m_3(x)$ en fonction de la taille de l'échantillon n.

Pour $n = 50$

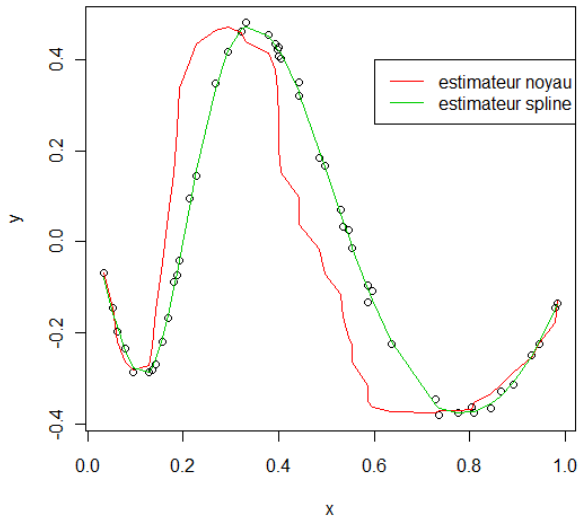

FIGURE 3.15 – Estimation de m_3, $n = 50$

Pour $n = 100$

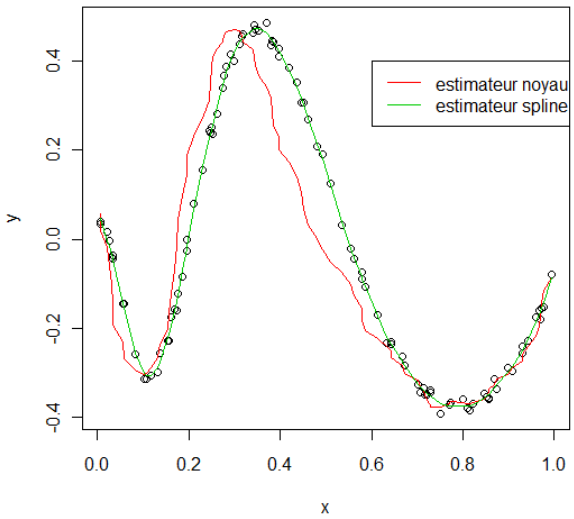

FIGURE 3.16 – Estimation de m_3, $n = 100$

Pour $n = 200$

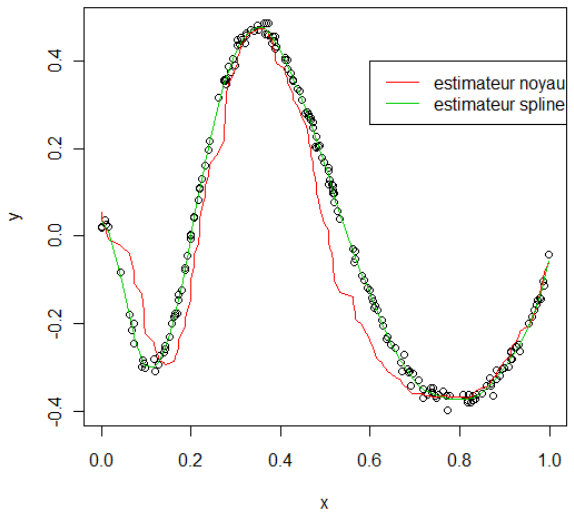

FIGURE 3.17 – Estimation de m_3, $n = 200$

Pour $n = 500$

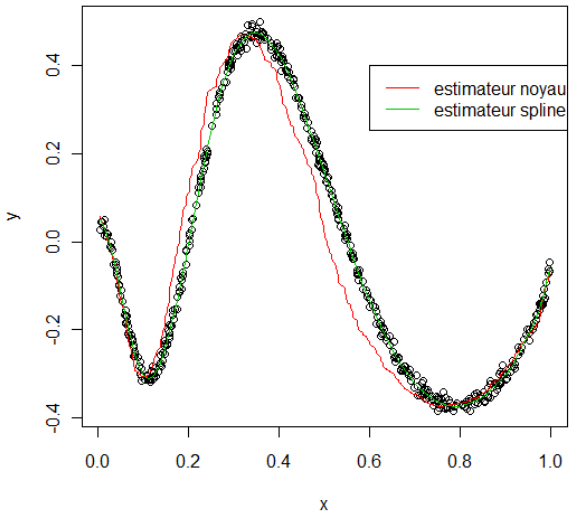

FIGURE 3.18 – Estimation de m_3, $n = 500$

Pour $n = 1000$

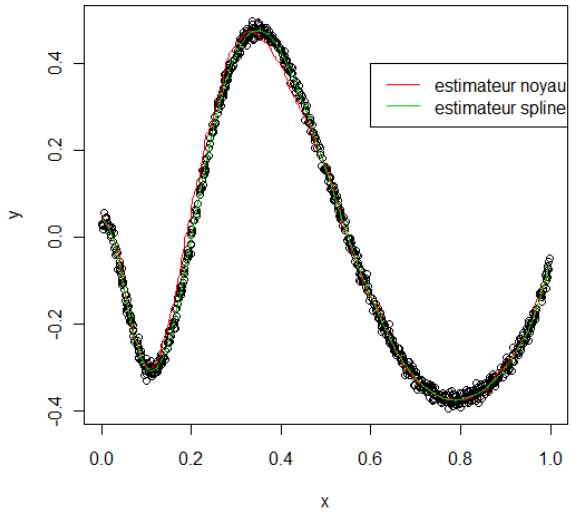

FIGURE 3.19 – Estimation de m_3, $n = 1000$

Pour $n = 2000$

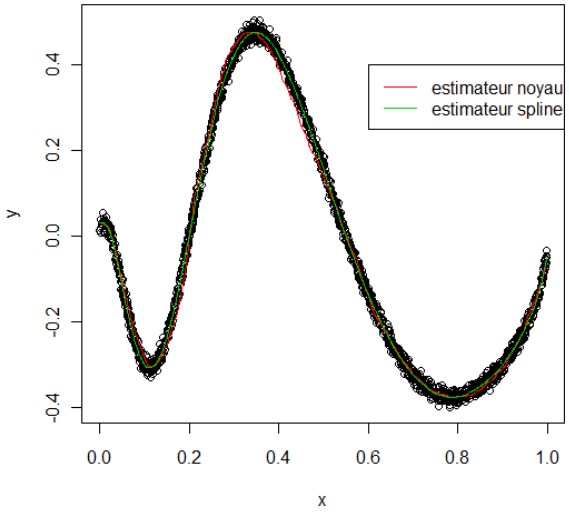

FIGURE 3.20 – Estimation de m_3, $n = 2000$

Pour $n = 3000$

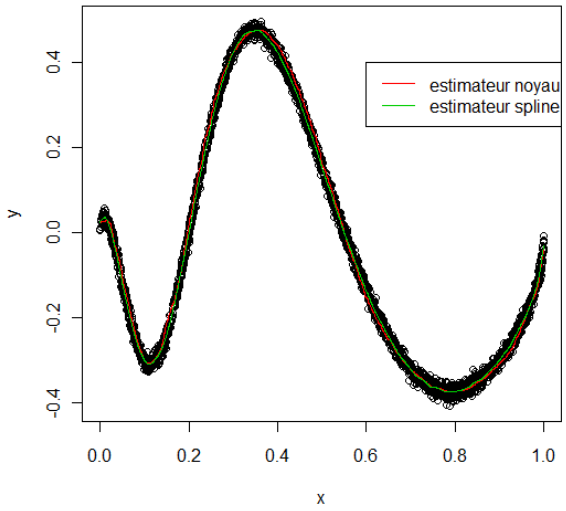

FIGURE 3.21 – Estimation de m_3, $n = 3000$

Interprétation des résultats du modèle m_3 :

Les résultats du tableau (3.3) montrent que l' erreur associée à la méthode du noyau MISE noyau est d'ordre 10^{-1}, elle décroît jusqu'à l'ordre 10^{-3} à partir de $n = 1000$ et reste stationnaire. Par contre l'erreur associée à la méthode des fonctions splines elle est d'ordre 10^{-3} pour $n \leq 500$, puis devient d'ordre 10^{-4} pour $n \geq 700$. Graphiquement, la méthode des fonctions splines représente bien les données pour toutes les tailles de l'échantillon, contrairement à la méthode du noyau qui donne une bonne représentation qu'à partir de $n = 1000$. La méthode des fonctions splines fournie par conséquent de meilleurs résultats et peut s'appliquer quelle que soit la taille n de l'échantillon par rapport à celle du noyau qui ne s'applique que pour des échantillons de grandes tailles.

3.4 Cas réel

Le tableau (3.3) représente des données de croissance des garçons. Les mesures exposées y sont des tailles établies en millimètres en fonction d'âge en année x (source : D^r Luciano Molinari, université de Zürich [Eubank][20]).

x	y	x	y	x	y
0.083	525	0.25	608	0.5	665
0.75	717	1	745	1.5	803
2	859	3	940	4	1007
5	1065	6	1121	7	1183
8	1238	9	1298	10	1348
10.5	1369	11	1391	11.5	1422
12	1470	12.5	1525	13	1578
13.5	1638	14	1664	14.5	1692
15	1708	15.5	1723	16	1727
16.5	1727	17	1727	18	1729
19	1738	20	1738		

TABLE 3.4 – Données numériques de croissance

L'erreur moyenne quadratique associée au cas réel

MISE noyau	MISE spline	MISE régression linéaire
4427.31	31.87724	5759.474

TABLE 3.5 – Erreur moyenne quadratique associée

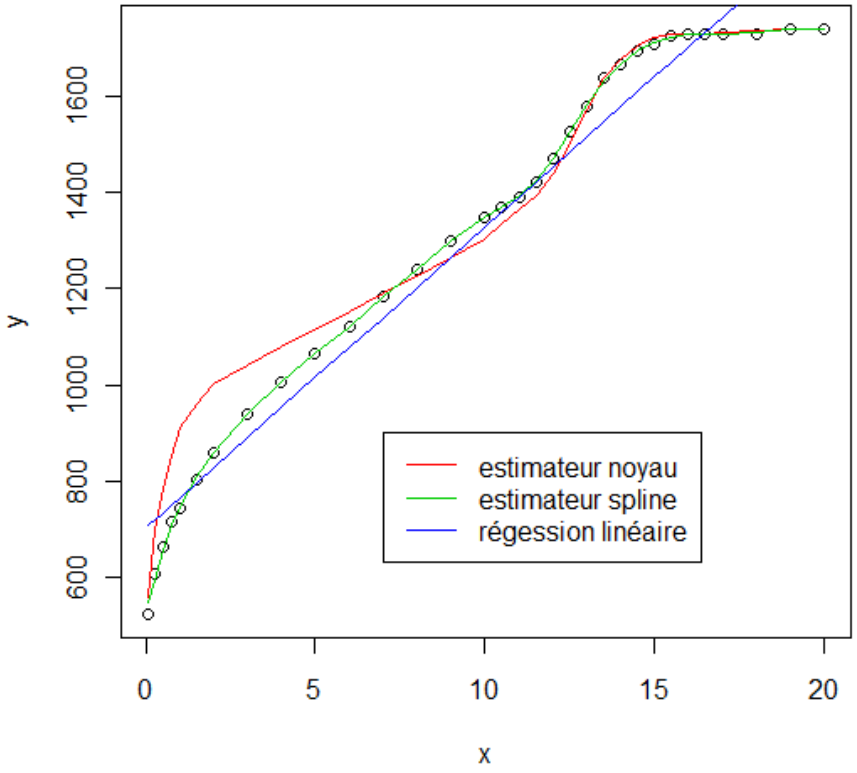

FIGURE 3.22 – Estimation de la courbe de croissance pour la méthode du noyau, la méthode des fonctions splines et par la régression linéaire.

Interprétation des résultats du cas réel :

A partir des résultats du tableau (3.5), nous constatons que l'erreur associée à la méthode des fonctions splines est très petite par rapport à celle donnée par la méthode du noyau (rapport de 138). On constate, ainsi, que l'erreur associée au modèle linéaire de régression est très grande. Ce qui signifie bien que le lien entre croissance et l'âge n'est pas linéaire. Ces constatations sont confirmées graphiquement, on l'on observe que la méthode des fonctions splines est celle qui ajuste le mieux les données réelles.

3.5 Conclusion

Dans ce chapitre, nous avons discuté 4 modèles de régression non paramétrique basés sur la méthode des fonctions splines et la méthode du noyau. Les résultats obtenus avec la méthode des fonctions splines ont été comparés à ceux obtenus par la méthode du noyau :

- Selon le critère du MISE, la méthode des splines est plus performante.
- Les estimateurs obtenus par la méthode des fonctions splines donnent aussi de meilleurs résultats graphiquement.

Ces résultats montrent que les estimateurs basés sur la technique des fonctions splines sont meilleurs que ceux du noyau. La méthode des splines est applicable quelle que soit la taille de l'échantillon et donne une bon ajustement graphique des données, contrairement à celle du noyau qui ne s'applique que pour des échantillons de grande taille.

Conclusion Générale

Dans ce travail nous avons présenté deux méthodes d'estimation non paramétrique de la courbe de régression de la moyenne. Les méthodes étudiées sont la méthode du noyau et la méthode des fonctions des splines de lissage cubique. Le but principal est de comparer ces deux méthodes en s'appuyant sur des exemples de fonctions de régressions simulés et sur un cas réel. Ce travail peut être résumer en deux parties principales :

La première partie est théorique, elle comporte les deux premiers chapitres. Dans cette partie nous avons brièvement défini les méthodes d'estimation non paramétrique de la courbe de régression de la moyenne. Nous nous sommes intéressé à la méthode classique du noyau, on a défini l'estimateur et donné ses différentes propriétés statistiques et asymptotiques. C'est l'objet du premier chapitre.

Par ailleurs, on a présenté une autre méthode d'estimation non paramétrique de la courbe de régression de la moyenne : la méthode des fonctions splines de lissage. C'est une méthode d'analyse numérique initialement, développée afin de résoudre des problèmes d'estimation. Nous avons présenté les étapes de la méthode, les différentes propriétés statistiques et asymptotiques.

La deuxième partie de ce travail, est la partie simulation qui nous a permis de comparer par simulation, sur des fonctions cibles et sur des données réelles, les deux méthodes d'estimation de la courbe de régression de la moyenne.

Les résultats obtenus montrent que pour toutes les fonctions cibles et même dans le cas réel, l'erreur quadratique moyenne intégrée (MISE) associée à la méthode des splines de lissage est toujours inférieure à celle de la méthode du noyau et ceci quelque soit la

taille de l'échantillon. Graphiquement, on constate que la méthode des splines ajuste mieux les données que la méthode du noyau. On constate par ailleurs qu'aussi bien du point de vue du MISE que graphiquement, les méthodes sont équivalentes quand la taille de l'échantillon n est très grande ($n > 1000$).

Nous donnons quelques perspectives de travail :

1. Étudier d'autre propriétés de l'estimateur par les fonctions splines de lissage cubique, par exemple la convergence en probabilité, presque sûre et uniforme.
2. Notre étude s'est focalisée sur l'estimation de la régression non paramétrique univariée. Il serait intéressant d'étendre les propriétés de l'estimateur spline de lissage au cas bidimensionnel.
3. Comparer à l'aide de simulations, la méthode des splines de lissage et la méthode du noyau dans le cas de régression bidimensionnel.
4. Puisque les splines de lissage cubique donnent de bons résultats dans l'estimation de la courbe de régression, on peut l'appliquer à l'estimation de la densité dans le cas uni et multidimensionnel.

Bibliographie

[1] P. Barbillon. Interpolation et régression. 2008.

[2] A. Berlinet. Hierarchies of higher order kernels. *Probability Theory and Related Fields, 94*, 489-504, 1993.

[3] S. Bianconcini. A reproducing kernel perspective of smoothing spline estimators. *Department of Statistics, University of Bologna, Via Belle Arti, 41 - 40126 Bologna, Italy*, 2008.

[4] D. Blondin. Lois limites uniformes et estimation non-paramètrique de la régression. *Thése Doctorat, Université Paris 6*, 1 - 26., 2004.

[5] N. Breaz and M. Aldea. On the smoothing spline regression models. *Acta Universitatis Apulensis. No 15*, 2008.

[6] R. Cao, M. A. Delgado, and W. Gonzalez-Manteiga. Non parametric curve estimation : An overview. *Investigaciones Economicas, vol. XXI(2)*, 209-252, 1997.

[7] Y. Cao. Inégalités d'oracle pour l'estimation de la régression. *Thése Doctorat, Université de Provence*, 2008.

[8] N. N. Cenzov. Evaluation of an unknown distribution density from observations. *Soviet Math. Dokl. 3*, 1962.

[9] G. Collomb. Estimation non-paramétrique de la régression par la méthode du noyau. *Thése 3ème cycle, Toulouse 3*, 1976.

[10] G. Collomb. Estimationnon paramétrique de la régression par la méthode du noyau : Propriétés de convergence asymptotiquement normale indépendante. *Annales csientifiques de l'Université de Clermont-Ferrand 2, tome 65, série mathématiques, n° 15*, p. 24–46, 1977.

[11] G. Collomb. Quelques propriétés de la méthode du noyau pour l'estimation non paramétrique de la régression en un point fixé. *C.R. Acad. Sc. Paris*, 295-314, 1977a.

[12] G. Collomb. Estimation de la régression pour la méthode du noyau : Quelques propriétés de convergence uniforme. *C.R. Acad. Sc. Paris*, 1977b.

[13] G. Collomb. Propriétés de convergence prèsque complète du prédicteur à nouyau. *Z. Wahrscheinlichkeitstheorie verw, Gebiete, 66*, 441-460., 1984.

[14] T. M. Cover and P. E. Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory, 13*, 21-7., 1967.

[15] P. Craven and G. Wahba. Smoothing noisy data with spline functions. estimating the correct degree of smoothing by the method of generalized cross-validation. *Numer. Math., 31(4)*, 377-403, 1979.

[16] C. De Boor. A practical guide to splines. *Springer, New York. Revised Edition*, (2001.

[17] L. P. Devroye. The uniform convergence of nearest neighbor regression function estimators and their application in optimization. *IEEE Transactions on Information Theory 24*, 142-51., 1978a.

[18] A. Dursun. A comparison of the nonparametric regression models using smoothing spline and kernel regression. *World Academy of Science, Engineering and Technology 36*, 2007.

[19] R. Eubank. Spline smoothing and nonparametreic regression. *Dekker, New York*, 1988.

[20] R. Eubank. Nonparametric regression and spline smoothing. *(2nd ed.) New York, Dekker*, 1999.

[21] R. L. Eubank. The hat matrix for smoothing splines. *Technical Report No. SMU/ DS/ TR- 179. Departement of Statistics ONR Contract*, 1983.

[22] F. Ferraty and P. Vieu. Statistique fonctionnelle : Modèle non-paraamétrique de régression. *Notes de cours de DEA*, 2002/2003.

[23] J. Fox. Non parametric regression. 2003.

[24] R. Germán. Smoothing and non-parametric regression. *Spring*, 2001.

[25] P. J. Green and B. W. Silverman. Nonparametric regression and generalized linear models : A roughness penalty approach. *volume 58 of Monographs on Statistics and Applied Probability. Chapman et Hall, London*, 1994.

[26] L. Györfi. The rate of convergence of k_n-nn regression estimation and classification. *IEEE Transactions of Information Theory 27*, 1981.

[27] P. Hall. Asymptotic properties of integrated square error and cross validation for kernel estimation of a regression function. *Z. Wahrsch. Verw. Gebiete, 67*, 1984.

[28] W. Härdl and J.S. Marron. Optimal bandwidth selection in nonparametric regression function. *Ann. Statist.13*, 1465-1481., 1985.

[29] W. Härdle. Robust regression function estimation. *Journal of Multivariate Analysis 14*, 169-80, 1984.

[30] W. Härdle. *Applied Nonparametric Regression.* Humboldt-Universität zu Berlin, 1994.

[31] W. Härdle and G. Kelly. Nonparametric kernel regression estimation -optimal choice of bandwidth. *Statistics, 18.1*, 21-35., 1987.

[32] C. Huang. Smoothing splines with boundary correction. *Communications in Statistics - Theory and Methods, 35 : 6,*, 1101 — 1107, 2006.

[33] C. Hurlin. Régressions non paramétriques univariées. *Master Econométrie et Statistique Appliquée (ESA), Université d'Orléans*, 2007-2008.

[34] M. S. Jones, L. Davies, and S. J. Sheather. A brief survey bandwith selection for density estimation. *J. Amer. Statist. Assoc., 91*, 401-407, 1996.

[35] S. L. Lai. Large sample properties of k-nearest neighbor procedures. *Ph.d. dissertation, Dept. Mathematics, UCLA, Los Angeles*, 1977.

[36] T. C. M. Lee. Smoothing parameter selection for smoothing splines : A simulation study. *Conputational Statistics and Data Analysis 42*, 139-148, 2003.

[37] K.S. Lii and M. Rosenblatt. Asymptotic behavior of a spline estimate of a density function. *Comput. Math. Appl. 1*, 223-235, 1975.

[38] C. Loader. bandwidth selection : Classical or plug-in. *Ann. of Statist., 27*, 1999.

[39] D. O. Loftsgaarden and G. P. Quesenberry. A nonparametric estimate of a multivariate density function. *Annals of Mathematical Statistics, 36*, 1049-51., 2005.

[40] Y. P. Mack. Local properties of k-nn regression estimates. *SIAM, J. Alg. Disc. Meth. 2*, 311-23, 1981.

[41] J. Marron. Automatic smoothing parameter : a survey empirical economics. *J. Multiv. Anal., 13*, 187-208, 1988.

[42] G. Micula. A variational approach to spline fuctions theory. *Rend. Sem. Mat. Univ. Pol. Torino Vol. 61, 3*, 2003.

[43] E. Moulines. Etude de cas de régression non paramétrique.

[44] E. A. Nadaraya. On estimating regression. *Theor. Prob. Appl., 9*, 141-142, 1964.

[45] E. A. Nadaraya. Nonparametric estimation of probability densities and regression curves. *Kluwer, Dordrecht*, 1989.

[46] E. Parzen. On estimation of a probability density function and mode. *Ann. Math. Statist, 33*, 1065-1076., 1962.

[47] G.S.G. Pollock. Smoothing wirt cubic splines. *Queen Mary and Westfield College, The University of London*, 365–397, 2004.

[48] D. L. Ragozin. Error bounds for derivative estimates based on spline smoothing of exact or noisy data. *J. Approx. Theory, 37(4)*, 335-355, 1983.

[49] C. Reinsch. Smoothing by spline functions. *Numererisch Mathematik 10*, 177–183, (1967.

[50] J. Rice and M. Rosenblatt. Smoothing splines : Regression, derivatives and deconvolution. *The Annals of Statistics. Vol. 11, No 1*, 141-156, 1983.

[51] M. Rosenblatt. Remarks on some nonparametric estimates of a density function. *Annals of Mathematical Statistics, 27*, 832-837., 1956.

[52] V. Roy. Régression non paramétrique des percentiles pour données censurées. *Mémoire Maître ès sciences (M.Sc.), UNIVERSITE LAVAL QUEBEC*, 2-18., 2007.

[53] L. Rutkowski. Sequential estimates of a regression function by orthogonal series with applications in discrimination, in revesz, schmetterer and zolotarev (eds). *The First Pannonian Symposium on Mathematical Statistics, Springer-Verlag*, pp. 263-44., 1981.

[54] P. Sarda and P. Vieu. Kernel regression. smoothing and regression : Approches, comutation, and application. *Ed. M.G. Schimek, 43-70, Wiley series in Probability and statistics*, 2000.

[55] I. J. Schoenberg. Spline functions and the problem of graduation. *Mathematics, 52*, 974-50, 1964.

[56] M. Schultz. Error bounds for polynomial spline interpolation. *Math. Computation 24 (111)*, 507-15, 1970.

[57] E. F. Schuster. Joint asymptotic distribution of the estimated regression function at a finite number of points. *Annals of Mathematical Statistics, 43.1*, 84-88., 1972.

[58] B. W. Silverman. Some aspects of the spline smoothing approach to non- parametric regression curve fitting (with discussion). *Journal of the Royal Statistical Society Series B, 47*, 1-52, 1985.

[59] J. S. Simonoff. Smoothing methods in statistics. *Springer Series in Statistics. Springer-Verlag, New York*, 1996.

[60] C. J. Stone. Consistent nonparametric regression (with discussion). *Annals of Statistics 5*, 595-645., 1977.

[61] C. J. Stone. Optimal rates of convergence for nonparametric regression. *Ann. Statist., 9*, 1348-1360, 1981.

[62] C. J. Stone. Optimal global rates of convergence for nonparametric regression. *Ann. Statist., 10*, 1040-1053., 1982.

[63] G. Szegö. Orthogonal polynomials. *Amer. Math. Soc. Coll. Publ*, 1959.

[64] C. Thomas-Agnan. Estimateurs splines. 2006.

[65] B. A. Turlach. Shape constrained smoothing using smoothing splines. *Computational Statistics,School of Mathematics and Statistics (M019), The University of Western Australia, 35 Stirling Highway, Crawley WA 6009, Australia*, 81-103, 2005.

[66] P. Vieu. Multiple kernel procedure : an asymptotic support. *Scand. J. of Statist., 26*, 61-72, 1999.

[67] Härdle W. Applied nonparametric regression. *Cambridge University Press, Cambridge*, 1990.

[68] G. Wahba. Convergence properties of the method of regularization for noisy linear operator equation. *TSR No. 1132, Math. Res. Center, Univ. of Wisconsin-Madison*, 1973.

[69] G. Wahba. Smoothing noisy data by spline functions. *II.Tech. Report No. 380, Dept. of Statist. Univ. of Wisconsin- Madison*, 1974.

[70] G. Wahba. Optimal convergence properties of variable knot, kernel, and orthogonal series methods for density estimation. *Annals of Statistics 3*, 15-29., 1975.

[71] G. Wahba. Histosplines with knots which are order statistics. *Journal of the Royal Statistical society, Series B 38*, 140-151, 1976.

[72] G. Wahba. Improper priors, spline smoothing and the problem of guarding against model errors regression. *Journal of the Royal Statistical society, Series B (Methodological), vol. 40. No, 3*, 364-372., 1978.

[73] G. Wahba. Spline models for observational data. *S.I.A.M., Philadelphia*, 1990.

[74] G. Wahba. Spline models for observational data. *CBMS-NSF series. SIAM, Philadelphia*, 1990.

[75] G. Wahba and S. Wold. A completely automatic french curve : Fitting spline fuctions by cross validation. *Simulation and Computation, 4 : 1*, 1-17, 1975.

[76] G. Walter. Properties of hermite series estimation of probability density. *Annals of Statistics 5*, 1258-64., 1977.

[77] G. S. Watson. Smooth regression analysis. *Sankhyà Ser. A, 26*, 359-372., 1964.

[78] E.T. Whittaker. On a new method of graduation. *Proceedings of the Edinburgh Mathematical Society, vol. 41*, pp. 63–75., 1923.

Résumé

La régression non paramétrique est un outil statistique permettant de décrire une relation entre une variable dépendante et une variable explicative, sans spécifier la forme de cette relation. L'objectif de ce travail est de comparer deux méthodes non paramétrique, la méthode du noyau et la méthode des fonctions splines, pour estimer la courbe de régression de la moyenne. Nous présentons dans ce mémoire en détail les estimateurs obtenus par les deux méthodes ainsi que leurs propriétés statistiques (Biais, variance, Erreur quadratique moyenne intégrée) et leurs propriétés asymptotiques. En utilisant le critère de l'erreur quadratique moyenne intégrée, on compare ces deux méthodes, par simulation sur trois modèles cibles de régression et sur un jeu de données réels qui concerne la croissance d'individus en fonction de l'âge. Les résultats numériques et graphiques montrent que la méthode des splines est meilleure que la méthode du noyau. Cependant quand la taille de l'échantillon observé est suffisamment grande les deux méthodes sont équivalentes.

Mots clés : Estimation, courbe de régression de la moyenne, noyau, fonction spline, paramètre de lissage, matrice de lissage.

Abstract

The non parametric regression is a statistic tool which allows to describe the relation between a dependent variable and explanatory variable, without specifying that relation's forme. The aim of this work is to compare two non parametrical methods which are the kernel method and the splines functions method, in order to estimate the mean regression curve. In this dissertation, we present in detail the obtained estimators by the two methods as well as their statistical properties (bias, variance, mean square integrated error) and their asymptotical properties. With the use of the mean square integrated error criterion, we compare these two methods, by simulation over three regression models and a reel case which concerns the growth data as function of age. The numerical and graphical results show that the splines functions method is better then the kernel method. Nevertheless, when the sample's size is sufficiently large the two methods are equivalent.

Key words : Estimation, mean regression curve, kernel, spline function, smoothing parameter, smoothing matrix.