République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
Université Abderrahmane Mira de Béjaia
Faculté des Sciences Exactes
Département de Recherche Opérationnelle

Mémoire Présenté

Pour l'obtention du Diplôme de Master

En Recherche Opérationnelle

Option : Mathématiques Financiéres

Par : BERKANE Zahra
Et : BAROUDI Lidia

Régression non paramétrique : Méthodes d'estimation

Soutenu à l'Université Abderrahmane Mira de Béjaia,
Le 02/07/2023, devant le jury composé de :

M^{me} BERNINE Nassima	Présidente	à l'UAMB - Bejaia.
M^{me} AMROUN Sonia	Encadreur	à l'UAMB - Bejaia.
M^{me} DJERROUD Lamia	Examinatrice	à l'UAMB - Bejaia.
M^{me} HARFOUCHE Lynda	Examinatrice	à l'école - amizour.

Année Universitaire 2022 – 2023

Remerciment

Tout d'abord on aimera d'adresser nos plus sincères remerciements au Dieu le tout puissant et le miséricordieux de nous avoir donné la chance, la patience et le courage pour réaliser ce modeste travail.

On tient à exprimer nos vifs remerciements à madame Sonia AMROUNE pour avoir accepté de nous encadrer lors du présent travail en dépit de son emploi du temps très chargé et de la confiance qu'il nous a témoignée, et les précieux conseils qu'elle a bien voulu prodiguer pour cibler les aspects traités dans ce travail.

On tient à remercier aussi mr ASLI, KABYLE, ABAS, BIBI, et madame TABTI et DJERROUD qui nous donnes des conseils très importants en signe de reconnaissance. On adresse nos sincères remerciements à tous les professeurs, intervenants et toutes les personnes qui par leurs paroles, leurs écrits, leurs conseils et leurs critiques ont guidé nos réflexions et ont accepté à nous rencontrer et répondre à nos questions durant la recherches.

Enfin, on ne voudra pas oublier de remercier tous nos amis qu'ont été d'un soutien que ce soit moral ou matériel, et qui ont contribué de près ou de loin à m'aider.

Beaucoup de charité et Bonne chance à tous.

Merci

Dédicace

Du profond de mon coeur,

je dédie ce travail à tous ceux qui me sont chers.

A mes parents (Larbi, Zahia)

qui nous ont soutenu tout le long de mon parcour d'études.

Mes grands-parents, que Dieu leur fasse miséricorde.

Mon frère : Farid.

Mes soeurs : Fatima, Assia.

Ma petite cousine Thafsouth que Dieu lui

accorde le paradis.

Mes très chers et précieux amis (Salhi, Ark, Yasmine, Nesrine, Mira, ...)

Ma binôme Zahra.

En souvenir des moments heureux passés ensemble, avec mes voeux sincères de réussite,

bonheur, et la santé.

Enfin, à tous les professeurs et les enseignants qui m'ont appris.

Lidia

Dédicace

Je dédie ce modeste travail À Mes chers parents,

Grâce à leurs tendres encouragements et leurs grands sacrifices, ils ont pu créer le climat affectueux et propice à la poursuite de mes études.

À mes deux frére mourad et mouhand.

Ma soeur Rosa.

A mes chers grands-parents (Setti et Jeddi)et mes cousines .

À tous mes chers amis (Romaissa, Cylia, Yamina, Djajiga, Sabrina, Imane, Rahma...).

Ma binome Lidia Baroudi.

A mes collègues de l'Université de Abd mira de Bejaia.

Et à tous ce qui m'ont enseigné au long de

ma vie scolaire.

Zahra

Table des matières

Table des figures

Liste des tableaux

Notations

$f(.)$	Fonction de régression.
$\hat{f}(.)$	Estimateur de fonction de régression.
$m(.)$	Densité.
$M(.)$	Fonction de répartition.
$\hat{m}(.)$	Estmateur de la densité.
$K(.)$	Fonction noyau.
h	Paramètre de lissage ou Fenêtre.
h_{opt}	Fenêtre optimale.
λ	Paramètre de lissage positif.
E	Espérance de probabilité.
Var	Variance d'un estimateur.
$Biais$	Biais d'un estimateur.
ASE	Erreur moyenne qudratique.
$MISE$	Erreur moyenne qudratique integrée.
ISE	Intégrale des erreurs quadratique.
$AMISE$	Erreur moyenne qudratique integrée asymtotique.
CV	Validation croisé.
GCV	Validation croisé généralisée.
$N(p)$	Voisinage de p.
ϵ_i	Erreues aléatoires de moyenne nulle et de variance σ^2.
$\|\|.\|\|$	Norme euclidienne.

Introduction générale

La littérature statistique propose deux types de modèles de régression : la régréssion paramétriques et la régression non paramétriques. La régression non paramétrique est une méthode d'analyse statistique qui permet d'estimer la relation entre une variable dépendante Y et une variable explicative X sans supposer de forme spécifique pour cette relation. Cette méthode est largement utilisée dans divers domaines tels que l'économie, la biostatistique et les sciences de l'environnement. Son développement remonte au 19 ème siècle, et son objectif est d'estimer la dépendance entre les variables sans contraindre sa forme.

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

$$y_i = f(x_i) + \epsilon_i, i = 1, ..., n,$$

ou f est la fonction de régression que l'on cherche à éstimer et les $(\epsilon_i)_{i=1}^n$ sont des erreurs aléatoires supposées que est de loi normal $(0, \sigma^2)$.

Plusieurs méthodes ont été proposées pour estimer la fonction de régression f, la première méthode rencontrée dans la litterature est la méthode du noyau ; également connue sous le nom de régression lissege par l'opérateur du noyau.L'estimateur noyau proposé par [Nadaraya (1964)][25] et de [Watson (1964)][48] repose sur l'utilisation d'une fonction noyau K et d'une fenêtre h, sous la forme suivante qui est un estimateur linéaire par rapport Y :

$$\hat{f} = \frac{\sum_{i=1}^n Y_i K(\frac{X_i - x}{h_n})}{\sum_{i=1}^n K(\frac{X_i - x}{h_n})}.$$

Une autre méthode recontrée dans la littérature est la méthode des fonctions splines, cette méthode est une technique d'analyse numérique utilisée pour l'interpolation. Elle permet de décrire une courbe à travers un ensemble fini de points donnés. Initialement introduite par [Whittaker(1923)][49] et développée par [Schoenberg(1964)][34], elle est largement utilisée dans le calcul scientifique et l'approximation.

Les splines de lissage sont désormais couramment utilisées en infographie pour représenter des courbes et des surfaces. Les travaux de [Wahba][[41]] ont démontré leur utilité pour résoudre des problèmes d'estimation statistique, ce qui les a rendues populaires dans des domaines tels que l'analyse de données de croissance, la médecine et l'économie.
L'estimateur par la méthode des splines de lissage est définit comme étant le fonction $\hat{f}_\alpha$

qui réalise le minimum de $s(\alpha)$, qui s'écrit sous la forme

$$\hat{f}_\alpha = A_\alpha Y;$$

avec A_α une matrice lisse qui ne dépend que des x_i, $i = 1, ..., n$, l'estimateur est linéaire par rapport au vecteur Y.

Ce travail est subdivisé en trois chapitres :

Dans le premier chapitre nous introduirons le modèle de régression non paramétrique ainsi que les méthodes d'estimations de cette fonction (k-plus proche voisin, nayau, spline, et la méthode des séris orthogonal).

Dans le deuxième chapitre, nous présentons les méthodes du noyau et la méthode des splines en détailes ainsi que les propriétés des estimateurs.

Dans le troisième chapitre, nous présentons les résultats obtenus par la simulation effectuée sur les exemples de fonction et un cas réel pour comparer les deux méthodes (noyau et spline), le critère utilisé est celui de ASE.

Ce mémoire se términe par une conclusion générale et quelques perspectives de recherche, suivie d'une bibliographie.

1

Méthodes d'estimation de la régression non paramétrique

Introduction

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

1.1 Régression paramétrique

La régression paramétrtique consiste a supposer que f appartient a une famille de fonctions continues ou discrète qui peuvent être décrites par un certain nombre de paramétres réels. Le but de la régression paramétrtique linéaire simple est d'expliquer une variable Y a l'aide d'une variable X. La variable Y est appelée variable dépendante (a expliquer) et X variable independante (explicative). Elle consiste à estimer des paramètres réelles finis (nombres finis).

1.1.1 Modèle de régression linéaire simple

Dans un modèle de régression paramétrique ; la fonction de régression est :
– (i) De forme explicite.
– (ii) Peut s'écrire en fonction d'un nombre réduit de paramètres.
Considérons un échantillon formé d'une suite de n couple (x_i, y_i) ; $i = 1, ..., n$ à valeurs dans R telque :

$$y_i = a + bx_i + \epsilon_i \ ; \tag{1.1}$$

où les erreurs aléatoires ϵ_i sont non corrélées, de moyenne nulle et de variance σ^2.

Pour estimer y_i on cherchons alors a déterminer les meilleures valeurs de a et b compte tenu d'un critére.

1.1.2 Estimation des paramètres par la méthode des moindres carrés

Les paramètres sont éstimés avec la méthode des moindres carrées, son principe est de minimiser la somme des carrés des résidus en fonction de a et b.

$$\min_{a,b} Q(a,b) = \min \sum_{i=1}^{n} \epsilon_i^2 = \min \sum_{i=1}^{n} (y_i - a - bx_i)^2 \ ; \tag{1.2}$$

1- Calcul des estimateurs :

La première dérivé par rapport à a et b est :

$$\begin{cases} \frac{\partial Q(a,b)}{\partial a} = -2 \sum_{i=1}^{n} (y_i - a - bx_i) = 0 \ ; & (*) \\ \frac{\partial Q(a,b)}{\partial b} = -2 \sum_{i=1}^{n} x_i (y_i - a - bx_i) = 0 \ . & (**) \end{cases}$$

L'équation (*) donne :

$$\hat{a} = \bar{y} - b\bar{x} \ ;$$

l'équation (**) donne :

$$\hat{b} = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) * (y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \ ;$$

telle que :

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i. \quad \text{et} \quad \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i.$$

1.2 Régression non-paramétrique

C'est un outil statistique pérmettant de d'écrire une relation entre une variable dépendante et une ou plusieurs variables explicatives, le problème de l'estimation non paramétrique consiste, dans la majeure 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. Quand on veut écrire l'effet d'une variable sur un évènement, faire le moin d'hypothèse possible sur la forme de la relation. Le but du modèle de régression est de déterminer la variable dépendante Y dépend d'un ensemble de variable éxplicatives X. Supposer $X \in \mathbb{R}$. Ainsi, le prblème est de détérminer pour chaque réalisation x de la variable X, la valeur de la fonction $f(x)$ appelée fonction de régression.

1.3 Notions sur l'estimateur

1.3.1 Estimateur sans biais

On dit qu'un estimateur $\hat{f}$ de f est sans biais si : $E(\hat{f}) = f$.

1.3.2 Estimateur asymptotiquement sans biais

On dit qu'un estimateur $\hat{f}$ de f est asymptotiquement sans biais si : $\lim_{n\to+\infty}(\hat{f}) = f$.

1.3.3 Estimateur asymptotiquement uniformément sans biais

Un estimateur$\hat{f}$ de f est dit asymptotiquement uniformément sans biais si :

$$\lim_{n\to+\infty} \sup_{x} |E[\hat{f}(x) - f(x)]| = 0 \ ;$$

1.3.4 Espérance mathématique

Soit une variable aléatoire X absolument continue de densité de probabilité $f(x)$ définie sur $[a, b]$.

L'espérance mathématique de X est définie par :

$$E(X) = \int_a^b xf(x)dx \ ;$$

On a la formule suivante pour une variable aléatoire discrète X de loi de probabilité (p_k) :

$$E(X) = \sum_{i=1}^{+\infty} p_i \times x_i \ ;$$

où $p_i = P(X = x_i)$

1.3.5 Variance

La variance d'une variable aléatoire X absolument continue est définie par :

$$V = E(X^2) - [E(X)]^2 \ ;$$

Si X une variable aléatoire discrète de loi (p_i). La variance de X se calcule ainsi :

$$V(X) = \sum_{i=1}^{+\infty} p_i \times x_i^2 - [E(X)]^2 \ ;$$

1.3.6 Erreur moyenne quadratique (MSE)

L'erreur moyenne quadratique MSE :

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

Définition (1.2) :

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

$$\lim_{n \to +\infty} MSE(f(x), \hat{f}(x)) = 0 \ ;$$

1.3.7 Erreur moyenne quadratique intégrée (MISE)

L'erreur moyenne quadratique intégrée MISE :

$$\begin{aligned} MISE(\hat{f}, f) &= \int MSE(f(x), \hat{f}(x))dx \ ; \\ &= \int Biais(\hat{f}(x))^2 dx + \int Var(\hat{f}(x))dx \ ; \end{aligned}$$

Définition (1.3) :

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

$$\lim_{n \to +\infty} MISE(f(x), \hat{f}(x)) = 0 \ ;$$

1.3.8 Intégrale des erreurs quadratiques (ISE)

l'intégrale des erreurs quadratiques (ISE) est défini par :

$$ISE(\hat{f}, f) = \int [\hat{f}(x) - f(x)]^2 dx \ ;$$

1.3.9 Convergence en loi

On dit qu'un estimateur $\hat{f}$ de f est asymptotiquement normal si :

$$\hat{f} \to N(E(\hat{f}),Var(\hat{f})) \ ; \quad \text{en loi.}$$

1.4 Méthodes d'estimation de la régression non paramétrique

La particularité de la statistique non-paramétrique est que la fonction inconnu qu'on cherche à calculer, à estimer ou à classifier n'est pas supposé d'appartenir à une famille indicée par un nombre de paramètres réels. En général, dans la théorie non-paramétrique on suppose que le nombre de paramètres qui décrivent la loi des observations est une fonction croissante du nombre d'observations, ou encore que le nombre de paramètres est infini. Dans cette partie on présente quelques méthodes d'éstimation non paramétrique de la fonction de régression.

1.4.1 Méthode des k-plus proche voisin

C'est une méthode non paramétrique utilisée pour la classification et la régression. Dans les deux cas, il s'agit de classer l'entrée dans la catégorie à laquelle appartient les k-plus proche voisins dans l'espace des caractéristiques identifiées par l'apprentissage. L'estimateur k-plus proche voisins est une moyenne pondérée dans un voisinage variable. Ce voisinage est défini par ces variables X qui sont parmi les k-plus proches voisins, en trouve le X par la distance euclidienne. La suite de ces points à été introduit par [Loftsgaarden and Quesenberry] [22] dans le but de l'estimation de la densité, et [cover and hart] [8] dans le but de classification. L'estimateur des k-plus proches voisins est défini par :

$$\hat{f}_k(x) = n^{-}1 \sum_{i=1}^{n} \omega_{ki}(x) y_i \ ; \tag{1.3}$$

où $\omega_{ki}(x)$ est une suite des poids des k-plus proche voisin, n est la taille d'échantillon.

Soit J_i l'ensemble d'indice ; $J_x = \{\text{i} : x_i \text{ est l'une des k observations les plus proches de x}\}$, telque :

$$\omega_{ki}(x) = \begin{cases} \frac{n}{k}, & \text{si i}\in J_i, \\ 0, & \text{sinon.} \end{cases} \tag{1.4}$$

Exemple

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 $\hat{f}_k(x)$ pour $x = 4$ et $k = 3$. Les observations proches de x sont les 3 derniers points des données. Par conséquent $J_x = J_4 = 3,4,5$ et donc

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

ce qui résulte

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

proposition (1.1) [Lai][21]

soit $k \to \infty$, $\frac{k}{n} \to \infty$, $n \to \infty$. Le biais et la variance de l'estimateur k-plus proches voisins$\hat{f}_k$ sont donnés par :

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

$$Var(\hat{f}(x)) \approx \frac{\sigma^2(x)}{k} \ ;$$

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

1.4.2 Méthode du noyau

C'est la méthode la plus connue et plus utilisée, son succès peut s'expliquer par l'expression théorique de l'estimateurs, et sa convergence dans différent sens. La méthode du noyau est une méthode simple et très pratique, lorsqu'en s'intéresse à la relation entre une variable réponse Y est une variable explicative X. L'estimation de la fonction de régression à été proposée par [Nadaraya (1964)][25] et [watson (1964)][48]. Cette approche non paramétrique par noyau a été aussi développé par [Ferraty and vieu(2002) et (2003)][13]. Cet estimation est basé sur le calcul de la moyenne pondérée des observations pour toutes les valeurs du domaine. Dans le cas multivarié continue, L'estimateur à noyau continu symétrique de la fonction de densité m est défini par [Parzen [1962)][27] :

$$\hat{m}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) \ ; \tag{1.5}$$

ou K est la fonction noyau telle que $K(t) \geq 0$ et $\int_R K(t)dt = 1$ et $h > 0$ est le paramètre de lissage. Dans l'expression de l'estimateur à noyau continu (1,7), la fonction noyau K est

une densité de probabilité sur $R \to R^+$ est symetrique par rapport a zero :

$$K(-u) = K(u) \ ;$$

ce qui implique l'égalité suivante :

$$\int_R uK(u)du = 0 \ ;$$

Construction de l'éstimateur a noyau pour la régréssion

Nous supposon que nous avant des paires d'observation $(x_1, y_1)...(x_n, y_n)$ du couple (X, Y). Nous proposons de construire un estimateur $\hat{m}(x)$ de la fonction de régression à partir des observations. L'estimateur à noyau est défini par une fonction $K(.)$ et une fenêtre h similaire à l'estimation par noyau de la fonction de densité.
La suite $\{h_n : n \geq 1\}$ de nombre positif : $h_n \to 0$ et $n \to +\infty$ La fonction $K : R \to R$ mesurable et satisfait les hypothèse suivante :

1.K est bornée ;

2.$\lim_{|u|\to\infty} |u|K(u) = 0$;

3.$K(.) \in L_1(R)$, i.e, $\int_R |K(u)|du < \infty$;

4.$\int_R K(u)du = 1$;

Nous reprenons la fonction de régréssion

$$f(x) = E(Y|X = x) = \frac{r(x)}{m_X(x)} \ ;$$

Nous avons,

$$\begin{aligned} r(x) &= \int_R y m_{X,Y}(x, y)dy \ ; \\ &= \lim_{h\to 0} \frac{1}{2h} \int_{x-h}^{x+h} \int_R y M_{X,Y}(du, dy) \ ; \\ &= \lim_{h\to 0} \frac{1}{2h} E[Y 1(|X_i - x| \leq h) \ ; \end{aligned}$$

où $M_{X,Y}(.,.)$ 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 $f(x)$ est estimée par

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

cet estimateur se présente sous forme la forme d'une moyenne locale pondérée des valeurs. Mais il présente de maniére discontinue sa généralisation naturelle est définie comme suit :

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

cet estimateur a été introduit par [Nadaray et watson][25]

$$\hat{f}(x) = \frac{\sum_{i=1}^{n} y_i K\left(\frac{x_i - x}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{x_i - x}{h}\right)} \times 1\{\sum_{i=1}^{n} K\left(\frac{x_i - x}{h}\right) \neq 0\} \ ;$$

ou $1\{.\} = 1_{\{.\}}$ désigne la fonction indécatrice

$$\hat{f}(x) = \begin{cases} \frac{\sum_{i=1}^{n} y_i K\left(\frac{x_i - x}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{x_i - x}{h}\right)}, & si\{\sum_{i=1}^{n} K\frac{X_i - x}{h} \neq 0\} \\ \frac{1}{n}\sum_{i=1}^{n} y_i \ , & sinon \end{cases}$$

1. Le biais de l'estimateur

$$Biais((\hat{f}(x)) = \frac{h^2}{2} f''(x) \int_{-\infty}^{+\infty} t^2 K(t) dt + o(h^2) \ ;$$

2. La variance

La variance de $\hat{f}(x)$ est :

$$Var(\hat{f}(x)) = \frac{1}{nh} f(x) \int_{-\infty}^{+\infty} K^2(t) dt + o(\frac{1}{nh}) \ ;$$

Exemples de noyaux

Voici quelques exemples de noyaux les plus communément utilisés :
1.Rectangulaire (Uniforme) :

$$k_1(u) = \begin{cases} \frac{1}{2}, & |u| \leq 1 \\ 0, & |u| > 1 \end{cases}$$

2.Triangulaire :

$$k_2(u) = \begin{cases} (1 - |u|), & |u| \leq 1 \\ 0, & |u| > 1 \end{cases}$$

3.Epanechnikov :

$$k_3(u) = \begin{cases} \frac{3}{4}(1 - u^2), & |u| \leq 1 \\ 0, & |u| > 1 \end{cases}$$

4.Biweight :

$$k_4(u) = \begin{cases} 1/\sqrt{2\pi} * \exp(u^2/2), & |u| \leq 1 \\ 0, & |u| > 1 \end{cases}$$

5.Gaussien :

$$k_5(u) = \left\{ (1/\sqrt{2\pi}) \exp(-u^2/2) \quad ; \forall u \in R \right.$$

6. Cubique :

$$k_4(u) = \begin{cases} (3/4) * (1 - u^2)^3, & |u| \leq 1 \\ 0, & |u| > 1 \end{cases}$$

Les courbes de certains noyaux sont présentées ci-dessous :

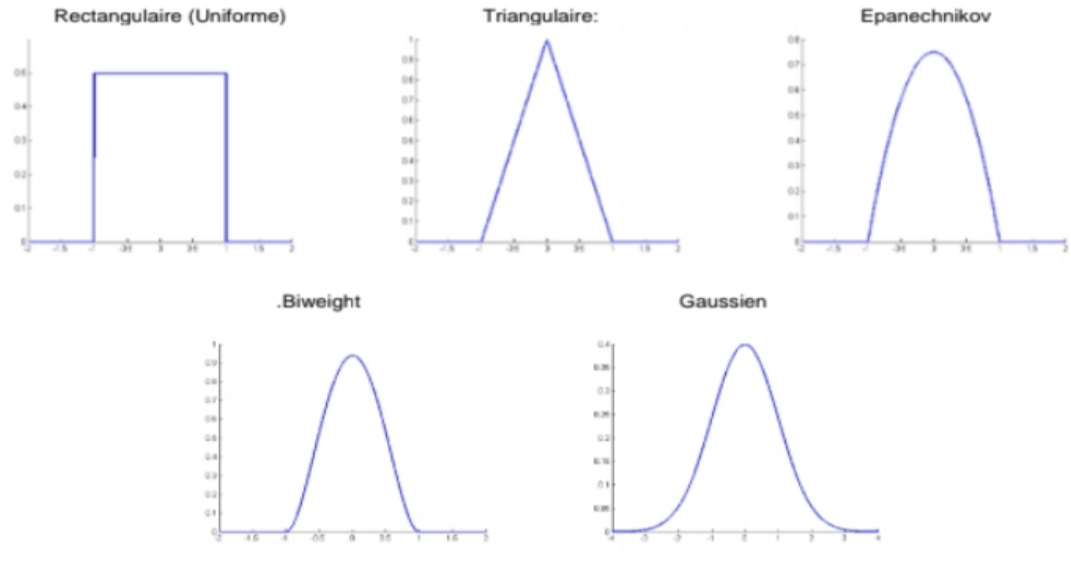

FIGURE 1.1 – Les courbes de certains noyau

1.4.3 Méthode des splines

La méthode des fonctions splines est une méthode d'analyse numérique, elle est utilisée dans le but de l'interpolation ; cette méthode est introduite pour la première fois par [Whittaker (1923)][49] et développée par [Schoenberg (1964)][34] pour servir au calcul scientifique. La méthode des spline est une technique de régression non paramétrique pouvant être vue comme une extension de régression linéaire qui modélisent automatiquement des interactions et la non-linéarités, cette méthode est énormément importante dans des différentes branches des mathématiques, est très souvent préférée à l'interpolation polynomiale, les spline sont également utilisées dans les problème de lissage des données expérimentales ou de statistique et pour représenter numériquement des contours complexes.

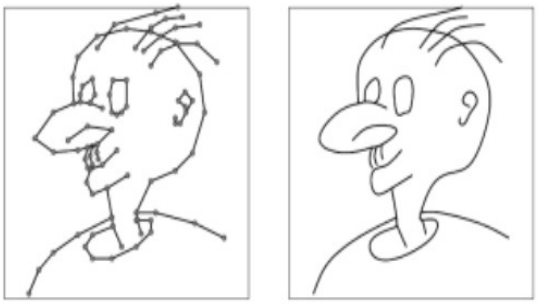

FIGURE 1.2 – Un dessin en zig-zag (à gauche) et en splines (à droite)

Une spline est une fonction définie par des polynômes par morceaux. En statistique, on l'utilise aussi pour lisser un nuage de points. Le principe des splines est de diviser l'intervalle $[a; b]$ où la fonction de spline est définie, en plusieurs sous intervalles $[a, x_1]; [x_1, x_2]; ...; [x_n, b]$. Les points $x_1; ...; x_n$ sont appelés les noeuds.

1- Splines de lissage

Les splines de lissage sont une façon d'utiliser les fonctions splines pour estimer la fonction de régression ; cette méthode permet de déterminer la valeur de l'estimateur en minimisant un critére bien précis.

Voir [Reinch(1967)][29], [Silverman (1985)][37], [Wahba(1990)][45] et [eubank(1999)][12]. 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. Les splines de lissage est une classe de fonctions non-paramétriques définie par [Whittaker (1923)][49].

Un "spline" $f : [a, b] \to R$ est une fonction polynomiale par morceaux, définie par :

$$f(x) = P_i(x), x_{i-1} \leq x \leq x_i \tag{1.6}$$

Avec $i = 1, .., n$ et $a = x_0 < x_1 < ... < x_{n-2} < x_{n-1} = b$ et $P_i(x)$ est un polynôme. La

fonction f minimise la somme des carrés résidus S donnée par :

$$S(f) = \sum_{i-1}^{n}(y_i - f(x_i))^2 + \alpha \int_b^a (f^{(r)}(x))^2 dx \ ; \tag{1.7}$$

c-à-d :

$$\hat{f}(x) = \arg\min \sum_{i-1}^{n}(y_i - f(x_i))^2 + \alpha \int_b^a (f^{(r)}(x))^2 dx \ ; \tag{1.8}$$

f est une fonction différenciable telleque $f \in W_2^r[a,b]$ et α un paramètre de lissage positif.

Le paramètre α permet de réaliser un compromis entre le suivie des données exprimé par $\sum_{i-1}^{n}(y_i - f(x_i))^2$ et les fluctuations locales exprimé par$\int_b^a (f^{(r)}(x)^2 dx$.

Définition de l'espace de Sobolev

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

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

et

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

$W_2^r = \left\{ f : [a,b] \to \mathbb{R} f^{(k)} \right.$ absolument continus pour $k = 0, ..., r-1$, et $\int_a^b (f^{(r)}(x))^2 < \infty$.
i.e :
On dit que f est une fonction absolument continue, si il existe un réel a et une fonction g tels que :

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

2- Estimateur spline de lissage [Cao][4]

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

$$\hat{f}(x_i) = (I + \alpha K)^{-1} Y = A_\alpha Y \ ; \tag{1.9}$$

avec K une matrice et A_α une matrice lisse qui ne dépend que des x_i, $i = 1, ..., n$, l'estimateur est linéaire par rapport au vecteur Y.

3-Propriétés de l'estimateur splines de lissage

3.1 Biais de l'estimateur

$$Biais(\hat{f}, f) = E\hat{f} - f = (A_\alpha - I)f \ ; \tag{1.10}$$

En effet,

$$\begin{aligned} E(\hat{f}) - f &= EA_\alpha Y - f, \\ &= EA_\alpha^i Y - f, i = 1, ..., n, \\ &= A_\alpha^i EY - f, i = 1, ..., n, \\ &= A_\alpha^i f - f, i = 1, ..., n, \\ &= A_\alpha f - f = (A_\alpha - I)f. \end{aligned}$$

3.2 La variance de l'estimateur

$$Var(\hat{f}) = E||\hat{f} - E\hat{f}||^2 = \sigma^2 tr(A_\alpha^t A_\alpha) \ ; \tag{1.11}$$

En effet ;

$$\begin{aligned} Var(\hat{f}) &= E||\hat{f} - E\hat{f}||^2, \\ &= E||A_\alpha Y - A_\alpha f||^2, \\ &= E\sum_{i=1}^{n}(A_{ij}^2)(y_i - f_{xi})^2, j = 1, ..., n, \\ &= \sigma^2 tr(A_\alpha^t A_\alpha). \end{aligned}$$

3.3 Trace d'une matrice

La trace d'une matrice $A = (a_{ij})1 \leq i, j \leq n$ notée par $tr(A)$ est donée par :

$$tr(A) = \sum_{i=1}^{n}(a_{ii}) \ ;$$

1.4.4 Méthode des séries orthogonale

Les séries orthogonale sont des utiles pour la classifications des surfaces quadratiques ; elles ont principalement appliquées dans plusieurs domaines différents (traitement du signal, statistiques fonctionnelles,...), le problème de l'estimation du mode de densité dans des conditions non paramétriques basées sur les séries orthogonales, à été développé par [Cencov (1962)][6] et proposé par [Kronmal et tarter(1968)][20], pour estimer des densité continues et étudiée par plusieurs auteurs [Bosq (2005)][3], [Saadi and Adjabi(2009)][32] , [Schwartz(1967)][36] et [Wahba (1981)][44] ;

Principe de la méthode

Soit $X_1...X_n$ une suite des variables aléatoires indépendantes. Si on veut estimer f à partir des observations $\{X_i\}_{i=0}^n$, pour cela supposons que nous avons une fonction de régression $f \in L_2(\mathbb{R})$ puisse être représentée comme une série de Fourier.

$$f(x) = \sum_{k=0}^{\infty} a_k e_k(x) \ ; \quad x \in \mathbb{R} \tag{1.12}$$

où $\{e_k\}_{k=0}^{\infty}$ une base orthonormale de $\mathbb{R}$, et $\{a_k\}$ sont des coefficients de Fourier innconu [Szegö][38].
Suposons que le système des fonctions $\{e_k\}, k = 1...n$ forment un système orthonormal dans $[-1, 1]$ et que la variable x est limitée dans cet intervale tel que (on coisier l'interval pour la simplicité de représentation) :

$$\int_{-1}^{1} e_j(x)e_k(x)dx = \sigma_{jk} = \begin{cases} 0, & j \neq k \\ 1, & j = k \end{cases} \tag{1.13}$$

Alors maintenant on peut calculée les coeficients de fourier comme suite :

$$\begin{aligned} a_j &= \sum_{k=0}^{\infty} a_k \sigma_{jk}, \\ &= a_k \int e_j(x)a_k(x)dx, \\ &= \int_{-1}^{1} f(x)e_j(x)dx. \end{aligned}$$

Ce derniére intégrale ne comporte pas uniquement la base connue de fonctions, mais aussi la fonction inconnue $f(x)$. Nous donnons automatiquement une estimation de e_j si nous pouvons l'estimer de manière raisonnable. La formule du coefficient de Fourier peut écrire comme suite :

$$a_j = \sum_{i=1}^{n} \int_{-1}^{1} f(x)e_j(x)dx \approx \sum_{i=1}^{n} f(x_i) \int_{-1}^{1} e_j(x)dx. \tag{1.14}$$

Pour estimer $f(x_i)$, nous proposons de construire un estimateur sans biais de $\hat{f}(x)$. Par la méthode des moments, on peut estimer le coefficient $\{a_k; k \in N\}$ par :

$$\hat{a}_j = \sum_{i=1}^{n} y_i \int_{-1}^{1} e_j(x)dx. \tag{1.15}$$

Si $N(n)$ termes dans la représentation (1.23) sont considérés la fonction de régression est approximée par :

$$\hat{f}_N(x) = \sum_{k=0}^{N} \hat{a}_k e_k(x). \tag{1.16}$$

Cet estimateur est appelé estimateur de f par les séries orthogonales. Il est une moyenne pondérée des variables Y avec le poids

$$\omega_{Ni} = n \sum_{j=1}^{N(n)} \int_{-1}^{1} e_j(u)\ \mathrm{du}\ e_j(x).$$

proposition (1.2) [Härdle][17]

Si pour un s tel que $0 < s < 1$

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

et

$$E|\epsilon_i|^{\frac{s+1}{s}} < \infty,$$

dés lors

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

$$\hat{f}_N(x) \to f(x). \text{ en probabilité}$$

Une preuve détaillée de la consistance de $\hat{m}_N$ peut etre trouver dans [Rutkowski][31], [Szegö][38], montre que :

$$\sup_x |a_j(x)| \sim j^{\rho}, j = 1, 2, 3, ...$$

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 (1.19) apporte alors la condition sur $N(n)$ de la forme de croissance.

$$\frac{N(n)^{2\rho+1}}{n^{1-s}} \leq c < \infty \text{ quand } n \to \infty; \tag{1.20}$$

Il faut que le paramétre de lissage tend vers ∞ pour assurer la consistance, mais pas trop rapidement comme le suggère (1.20). Le taux de convergence de l'estimateur par les séries orthogonales est donné par [Härdle][15].

Conclusion

Dans ce chapitre,nous avons rappeler la notions de l'estimation non paramétrique. Ainsi nous avons donner quellques méthodes d'estimation de la régression non paramétrique, à savoir la méthode des k-plus proches voisins, la méthode du noyau, la methode des splines, et la méthode des séries orthogonales. Enfin nous avons présenté les propriétés statistiques, biais, variance. Nous allons présenter dans le chapitre suivant en détaille les deux méthode noyau et spline et leurs principe.

2

Estimation non paramétrique par la méthode du noyau et par la méthode des splines de lissages

Introduction

Dans ce chapitre, nous allons présenté en détail l'estimateur à noyau et spline ainsi que leurs proprietés. Nous parlons aussi de son intéret particulier sur le choix du paramétre de lissage.

2.1 Estimation non paramétrique par la méthode noyau

On suppose que lon a observé un échantillon $(X_i, Y_i); i = 1...n$ et on veut expliquer la variable aléatoire Y_i par X_i et nous considérons le modèle de régression non paramétrique donné pour $i = 1...n$

$$y_i = f(x_i) + \epsilon_i \ ; \tag{2.1}$$

où ϵ_i est une erreur aléatoire centré et indépendant de X_i et f est une application mesurable réelle.
La fonction de régression $f(.) = E[Y/X]$, fournit des informations sur les dépendances inconnues de Y et X ; un problème important est d'estimer f à partir de n observations. Supposons que la densité de $(X; Y)$ est $m : (x; y) \to m(x; y)$ sur R^2 ;

$m_X(x) \to m_X(x) = \int m(x;y)dy > 0$ (densité de X),

$$\forall x \in R, f(x) = E[Y/X = x] = \frac{\int ym(x,y)\partial y}{m_X(x)} \ ;$$

Les densités m et m_X sont inconnues mais on peut les estimer par :

$$\hat{m}(x,y) = \frac{1}{nh^2}\sum_{i=1}^{n} K(\frac{X_i - x}{h_n})K(\frac{Y_i - y}{h_n}) \ ;$$

$$\hat{m}_X(x) = \frac{1}{nh}\sum_{i=1}^{n} K(\frac{X_i - x}{h_n}) \ ;$$

par conséquent l'estimateur de la régressionest donné par :

$$\forall x \in R, f(x) = E[Y/X = x] = \frac{\int ym(x,y)\partial y}{m_X(x)} I_{(m_X(x)\neq 0)} \ ; \tag{2.2}$$

Définition

Si K est un noyau dordre 1, l'estimateur défini par (2.2) verifié :

$$\begin{aligned} \forall x \in R, \hat{f}(x) &= \frac{\sum_{i=1}^{n} Y_i K(\frac{X_i - x}{h_n})}{K(\frac{X_i - x}{h_n})}; \\ &= \frac{Y_i K_{hn}(X_i - x)}{\sum_{i=1}^{n} K_{hn}(X_i - x)} \ ; \end{aligned}$$

Où $K_{hn}(.) = K(./h_n)$; donc lestimateur noyau de régression est donné

$$\hat{f}(x) = \frac{\sum_{i=1}^{n} Y_i K(\frac{X_i - x}{h_n})}{K(\frac{X_i - x}{h_n})}; \tag{2.3}$$

$$= \frac{\hat{g}_X(x)}{\hat{m}_X(x)} \ ; \tag{2.4}$$

C'est l'estimateur à noyau introduit par [Nadaraya(1964)][25] et [Watson(1964)][48]

2.2 Propriétés asymptotiques de l'estimateur à noyau

Nous étudions dans cette partie deux modes de convergence, la convergence en moyenne quadratique et la convergence presque complète. nous supposons que K est un noyau vérifant les condition suivantes : [khadraoui][19]

1.K est bornée ,cest dire $\sup_{x\in R}|K(x)|\partial x<\infty$.

2.$\lim|x|K(x)\to 0$,quand $|x|\to+\infty$.

3.$\kappa\in L_1(R)$, c'est à dire $\int_R|K(x)|\partial x<+\infty$.

4.$\int_R|K(x)|dx=1$.

5.$\int_R uK(x)\partial u=0$.

6.$\int_R u^2K(u)\partial u<+\infty$.

7.K est bornée, intégrable et a support compact.

1-Etude asymptotique du biais

L'etude asymptotique du biais repose sur la proposition suivante :

Proposition (2.1)

i) Si $|Y|\leq C1<\infty$ et $nh_n\to\infty$, quand n $\to\infty, alors$:$E[\hat{f}(x)]=\frac{E[\hat{g}_X(x)]}{E[\hat{m}_X(X)]}+O(\frac{1}{nh_n})$.
ii)Si $EY^2<\infty$, $nh_n^2\to\infty$; quand $n\to\infty$ alors :

$$E[\hat{f}(x)]=\frac{E[\hat{g}_X(x)]}{E[\hat{m}_X(X)]}+O(\frac{1}{\sqrt{n}h_n}).$$

Proposition (2.2)

Si condition (4),(5) et (6) sont vériffées si $m_X(.)$et $f(.)$ sont le classe $C^2(R)$ et si $|Y|$ est borné. Alors :

$$E[\hat{f}(x)-f(x)]=\frac{h_n^2}{2}\{\{f''(x)+2f'(x)\frac{m'_X(x)}{m_X(x)}\}\int_R u^2K(u)\partial u\}(1+O(1)).$$

Preuve

$E[\hat{f}(x)-f(x)]=[EK(\frac{x-X}{h_n})]^-1\{\int_R\frac{1}{h_n}K(\frac{x-t}{h_n})g(t)\partial t-f(x)\int_R\frac{1}{h_n}K(\frac{x-t}{h_n}m(t)\partial t\}$,

$$=\{(m(x))^{-1}\}\frac{h_n^2}{2}g''(x)-\frac{h_n^2}{2}m(x)f''(x)\}\int_R u^2K(u)\partial u\}(1+O(1))\}.$$

Comme $g(x)=f(x)m(x)$, l'équation précédente peut secrite :

$$E[\hat{f}-f(x)]=\{\frac{h_n^2}{2}\{f''(x)+2f'(x)\frac{m'_X(x)}{m_X(x)}\int_R u^2K(u)\partial u\}(1+O(1)).$$

D'où

$$\lim_{n\to\infty} E[\hat{f}(x)] = f(x).$$

2-Etude asymptotique de la variance

Proposition (2.3)

Sous $E(Y^2) < \infty$, alors en chaque point de continuité des fonctions $m(x)$, $f_X(x)$ et $\sigma^2(x) = var(Y \backslash X = x)$,
on a :

$$var[\hat{f}(x)] = \frac{1}{nh_n}\{\frac{\sigma^2(x)}{m_X(x)}\int K^2(u)du\}(O(1)+1).$$

Preuve

soit la fonction $z(x) = \int y^2 f(x,y)\partial y$, en se basant sur le lemme de Bochner

$$var[\hat{g}, X(x)] = \frac{1}{nh_n}\{E[y^2\kappa^2(\frac{x-X}{h_n})] - [EYK(\frac{x-X}{h_n})]^2\},$$

$$= \frac{1}{nh_n}\{\int_R K^2(u)z(x-h_nu)\partial u - h_n(\int_R K(u)m(x-uh_n)^2\},$$

$$= \frac{1}{nh_n}\{\int_R K^2(u)\partial u(1+O(1))\}.$$

$$E[\{\hat{m}_X(x) - E(\hat{m}_X(x))\}\{\hat{g}_X(x) - E(\hat{g}_X(x))\}] = \frac{1}{nh_n}g(x)\int_R K^2(u)\partial u(1+O(1)),$$

$$var[\hat{f}_X(x)] = \frac{1}{nh_n}f_X\int_R K^2(u)\partial u(1+O(1)).$$

2.2.1 Choix du paramatre de lissage

L'estimateur $\hat{f}$ dépend : le noyau K et la largeur de la fenêtre h.

1-Etude de critére d'erreur quadratique moyenne de $\hat{f}(x)$

L'erreur quadratique moyenne (en anglais : mean squared error MSE) est une mesure qui permet de réduiser la similarité de $\hat{f}$ avec la fonction de régression inconnue f au point x.

Notre but est de minimiser

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

Le développement de cette expression faite précédemment , nous donne

$$MSE(\hat{f}(x)) = var[\hat{f}(x)] + [biais(\hat{f}(x))]^2.$$

Nous constatons d'une part que les expressions du biais de $\hat{f}(x)$et de la variance de $\hat{f}(x)$

$$MSE(\hat{f}(x)) = \frac{h_n^4}{4}[(f''(x)+2f'(x)\frac{m'_X(x)}{m_X(x)}(u^2K(u))+O(1)]^2+\frac{1}{nh_n}(\frac{\sigma^2(x)}{m_X(x)}[K^2(u)]+(1+O(1))$$

où

$$[u^pK^q(u)] = \int t^pK^q(t)dt.$$

2-Fenêtre optimal

Pour trouver un compromis entre biais et variance, nous minimisons l'expression de l'erreur quadratique moyenne asymptotique AMSE (asymptotic mean square error) par rapport à h_n, donnée par :

$$MSE(\hat{f}(x)) = \{\frac{h_n^4}{4}[(f''(x) + 2f'(x)\frac{m'_X(x)}{m_X(x)}\}^2[u^2K(u)]^2 + \frac{1}{nh_n}(\frac{\sigma^2(x)}{m_X(x)}[K^2(u)].$$

Comme AMSE est fonction convexe. La fenêtre $h^{MSE}_{opt(\hat{f}(x))}(\hat{f}(x)) = (AMSE\hat{f}(x))]$ est solution de léquation suivante :

$$\frac{\partial}{\partial h_n}\{\frac{h_n^4}{4}[(f''(x) + 2f'(x)\frac{m'_X(x)}{m_X(x)}\}^2[u^2K(u)]^2 + \frac{1}{nh_n}(\frac{\sigma^2(x)}{m_X(x)}[K^2(u)] = 0.$$

Lorsque :

$$\frac{h_n^4}{4}[(f''(x) + 2f'(x)\frac{m'_X(x)}{m_X(x)}\}^2[u^2K(u)]^2 \neq 0;$$

d'où

$$h^{MSE}_{opt(\hat{f}(x))} = n^{-1/5}\{\frac{\frac{\sigma^2(x)}{m_X(x)}[K^2(u)]}{\frac{m'_X(x)}{m_X(x)}\}^2[u^2K(u)]^2}\}^{1/5}.$$

2-MISE (Mean Integrated Squared Error)

$$MISE[f_n(x)] = E[\int R(F_n(x) - f(x))^2 dx].$$

En appliquant le théorème de Fubini, on a :

$$MISE[\hat{f}(x)] = E[\int R(\hat{F}(x) - f(x))^2 dx].$$

Sous les mêmes hypothèses que les propositions (biais) et (variance),on a

$$MSE(\hat{f}(x)) = \{\frac{h_n^4}{4}[(f''(x) + 2f'(x)\frac{m'_X(x)}{m_X(x)}\}^2[u^2K(u)]^2 + \frac{1}{nh_n}(\frac{\sigma^2(x)}{m_X(x)}[K^2(u)].$$

3-AMISE sous condition de continuité

Supposons que $\exists \beta > 0$ telle que $\inf_{x \in c} m(x) > 0$;

$$AMISE[\hat{f}(x)] \to 0$$

la fenêtre $hMISE^{MISE}_{opt(\hat{f}(x))}$ minimisant l'AMISE du critére global, et $h \to 0$, $nh \to \infty$ et K est borné, intégrable, positif, symétrique et a support compact On a :

$$h^{MSE}_{opt(\hat{f}(x))} = n^{-1/5}\{\frac{\frac{\sigma^2(x)}{m_X(x)}[K^2(u)]}{\frac{m'_X(x)}{m_X(x)}\}^2[u^2K(u)]^2}\}^{1/5}.$$

Un travail similaire se fait pour le choix optimum du paramètre de lissage dans le cas de l'estimateur de [Parzen][27] et [Rosemblatt][30], nous obtenons :

$$h^{\hat{f}(x))}_{MSE} = n^{-1/5}\{\frac{m_X(x)[K^2]}{m''_X(x))^2 dx[t^2K]^2\}^{1/5}}$$

$$h^{\hat{f}(x))}_{MSE} = n^{-1/5}\{\frac{[K^2]}{\int_R (m''(x))^2 dx[t^2K]^2}\}^{1/5}.$$

Nous notons que l'expression de h_n optimal, minimisant asymtotiquement les quatre critères derreurs la forme

$$h_opt = C_n^{-1/5}.$$

Où la constante C est en fonction de la distribution et de termes aléatoires inconnues.

3-Validtaion croisé

La cross validation ou validation croisée est une méthode statistique qui permet d'évaluer les pérformances des modèles d'apprentissage automatique. Lorsqu'on entraîne un modèle sur des données étiquetées, l'hypothèse qu'il doit également fonctionner sur de nouvelles données.

La fonction validation croisée est définie par la quantité :

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

2.2.2 Normalisation asymtotique

La première démonstration de la normalité asymptotique de l'estimateur [Nadaraya][25] et [Watson][48] est due à [Schuster][35]. On se réfère également aux théorèmes

$$f(x) = \arg\min_a E[(Y-a)^2[X=x]];$$

$$= E(Y|X=x).$$

de [Nadaraya][25] et de [Härdle] [16] 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 $cn^{-\frac{1}{5}}$.

2.3 Estimation non paramétrique par la méthode des fonctions splines

2.3.1 Fonction Spline

Soient $I = [a, b]$ un intervalle de $\mathbb{R}$, et soit $a < x_1 < ... < x_n < b$ une partition de l'intervalle I et soit $f : I \to \mathbb{R}$ une fonction.
La fonction f est dit spline polynomiale de degré $(r-1)$ (où d'ordre r) ayant pour noeuds $x_1, ..., x_n$ si elle vérifié les propriétés suivantes :

— f est continûment dérivable jusqu'à l'ordre r.
— Dans chaque sous-intervalle la fonction f est un polynôme de degré $\leq r, r \in \mathbb{N}$.

2.3.2 Fonction Spline Naturelle

soit $a < x_1 < ... < x_n < b$ une partition de l'intervalle I et soit f une fonction spline de degré $\leq r$.
La fonction f est dite naturelle, si elle coincide avec un polynome de degré inférieur ou égale à $(r-1)$ en dehors de l'intervalle $[x_1, x_n]$.

On note par $S_{2r}(x_1, ..., x_n)$ l'espace des fonctions splines naturelles d'ordre $2r$, telque $\dim S_2(x_1, ..., x_n)_r = n$
Alors nous avons

— f est une spline de degré $2r-1$,
— $f \in P_{r-1}$, pour $x \in [x_1, a] \cup [x_n, b]$. P_{r-1}
où $P_{r-1} = p : \mathbb{R} \to \mathbb{R}$, P polynôme de degré $\leq (r-1), r \in \mathbb{N}$.

Théorème (2.1) :[Thomas-Agnan][40]

$S_r(x_1, ..., x_n)$ 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 $1, x, ..., x^{r-1}$ et les fonctions $(x - x_1)_+^{r-1}, ..., (x - x_n)_+^{r-1}$, avec

$$\dim S_r(x_1, ..., x_n) = r + n$$

tel que

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

2.3.3 Spline Cubique Naturelle

Soient $[a, x_1], [x_1, x_2], ..., [x_n, b]$ sous-intervalles dans $[a, b]$, et soit f une fonction définie sur I.
f est appelée une spline cubique si les conditions suivantes sont satisfaites :

— f est un polynome cubique sur chaque sous-intervalle dans $[a, b]$,
— La fonction f est deux fois continument differentiable sur $[a, b]$, et donc f et ses dérivées d'ordre 1 et 2 sont continues aux points x_i.

2.3.4 Interpolation par les fonctions splines

Rappellons que l'interpolation consiste à trouver une fonction passant exactement par les point à notre disposition.
soient $x_0, ..., x_n$ les points d'interpolation. On cherche un polynôme P sur chaque segment $[x_i, x_{i-1}]$ tel que :

$$P_i : [x_i, x_{i+1}[\longrightarrow R, P_i(x) = a_i(x - x_i)^3 + b_i(x - x_i)^2 + c_i(x - x_i) + d_i,$$

Soit f une fonction spline cubique peût s'écrire sous la forme polynomiale suivante :

$$f(x) = \sum_{i=1}^{n} P_i(x) 1_{[x_i, x_{i+1}[}(x). \tag{2.5}$$

— **La 1^{er} dérivée de f est :**

$$\begin{aligned} f'(x) &= \sum_{i=1}^{n} P_i'(x) 1_{[x_i, x_{i+1}[}(x), \\ &= \sum_{i=1}^{n} 3a_i(x - x_i)^2 + 2b_i(x - x_i) + c_i, \quad x_i \leq x < x_{i+1}. \end{aligned}$$

— **La 2^{me} dérivée de f est :**

$$\begin{aligned} f''(x) &= \sum_{i=1}^{n} P_i''(x) 1_{[x_i, x_{i+1}[}(x), \\ &= \sum_{i=1}^{n} 6a_i(x - x_i) + 2b_i, \quad x_i \leq x < x_{i+1}. \end{aligned}$$

les coeifficiants a_i, b_i, c_i et d_i sont déterminés à partir des conditions d'interpolation ($P_i(x_i) = f(x_i)$ pour $i = 1, ..., n$) et les conditions de continuité pour la première et la deuxième dérivées à chaque tous noeuds et sur chaque sous-intervalle.
Notons $h_i = x_{i+1} - x_i$, $i = 1, ..., n$;

1- La condition (1) :

l'égalité des polynômes adjacents au point (x_i, y_i) (i.e f est continue)

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

alors

$$d_i = y_i$$

2- La condition (4) :

l'égalité des dérivées premiéres au point (x_i, y_i)

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

3- La condition (3) :

l'égalité des dérivées seconde au point (x_i, y_i)

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

4- La condition (4) :

les dérivées seconde et troisième doivent être égales à zéro aux bords :

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

et

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

D'aprés La continuité de f et ses deux dérivées on obtient les différentes relations entre les coefficients.
Dans la première condition on obtient :

$$c_{i-1} = \frac{y_i - y_{i-1}}{h_{i-1}} - a_{i-1}h_{i-1}^2 + 2b_{i-1}h_{i-1}. \qquad (1)$$

Le seconde condition donne :

$$a_{i-1} = \frac{b_i - b_{i-1}}{3h_{i-1}}. \qquad (2)$$

Remplaçons (2) dans (1), on trouve :

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

On remplace (2) et (3) dans le seconde condition, on obtient :

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

Nous avons

$$P_i(x_i) = y_i, \text{ et } P_i''(x_i) = 2b_i.$$

Supposons que f est une fonction spline cubique naturelle (SCN) de nœuds $X_1, ..., X_n$ et soient $F = (F_1, ..., F_n)^t$ et $\Gamma = (\gamma_2, ..., \gamma_{n-1})$ telque

$$F_i = f(x_i),\ \gamma_i = f''(x_i),\ i = 1, ..., n, \text{ et } \gamma_1 = \gamma_n = 0.$$

Théorème (2.2) : [Green et Silverman(1994)][14]

Les vecteurs F et Γ définissent une spline cubique naturelle si et seulement s'il satisfait :

$$Q^tF = R\Gamma$$

Si cette relation est vérifiée, alors :

$$\int_a^b f''(x)^2dx = \Gamma^2R\Gamma = F^tKF. \qquad (2.6)$$

Où Q, R, et K des matrice définit comme suit :

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

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

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

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

— K est définie par :

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

Interpolation polynomiale : Polynômes de Lagrange

Soit $f : R \to R$ et $x_1, ..., x_n$ n réels donnés distincts.
Interpoler la fonction f par un polynôme de degré $n-1$ aux points $x_1, ..., x_n$ consiste à trouver un polynôme de degré $\leq (n-1)$ tel que :

$$p(x_i) = f(x_i), 1 \leq i \leq n.$$

Si un polynôme existe, il s'écrit avec une forme unique suivante

$$p(x_i) = \sum_{i=1}^{n} \alpha_i l_i(x). \tag{2.7}$$

Avec $l_i(x)$ sont des polynômes de Lagrange associé aux points (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 $n-1$ et vérifié

$$l_i(x_i) = 1, l_i(x_j) = 0. \tag{2.8}$$

En utilisant (2.8) et en prenant $j = 1, ..., n$, on obtient

$$\alpha_j = p(x_j) = f(x_j)$$

Exemple 1

Pour $n = 1$ le polynôme de Langrang s'écrit

$$P_1(x) = \alpha_0 \frac{x - x_1}{x_0 - x_1} + \alpha_1 \frac{x - x_0}{x_1 - x_0}.$$

C'est l'équation de la droite qui passe par les point (x_0, α_0) et (x_1, α_1).

Exemple 2

Pour $n = 2$ le polynôme de Langrang s'écrit

$$P_2(x) = \alpha_0 \frac{(x - x_1)(x - x_2)}{(x_0 - x_1)(x_0 - x_2)} + \alpha_1 \frac{(x - x_0)(x - x_2)}{(x_1 - x_0)(x_1 - x_2)} + \alpha_2 \frac{(x - x_0)(x - x_1)}{(x_2 - x_0)(x_2 - x_1)}.$$

C'est l'équation de parabole qui passe par les point (x_0, α_0), (x_1, α_1) et (x_2, α_2).

Théorème (2.3) [Barbillon][2]

Le polynôme

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

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

$$L(x_i) = y_i.$$

2.3.5 Existence et unicité des spline d'interpolation

Théorème (2.4) [Thomas-Agnan] [40]

Etant 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 $W_r^2[a, b]$ tel que :

1. f satisfait les conditions d'interpolation

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

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

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

Théorème (2.5) [Cao and Golubev(2006)] [5]

Supposons que $n \geq 2$ et f_λ est la spline cubique naturelle pour les valeurs $y_1, ..., y_n$ en points $x_1, ..., x_n$, ou $a < x_1 < ... < x_n < b$. Soit $\tilde{f}$ fonction dans $W_2^2[a, b]$ telle que :

$$\tilde{f}(x_i) = yi, i = 1, ..., n.$$

Alors

$$\int_a^b \tilde{f}''(x)^2 dx \geq \int_a^b \hat{f}''(x)^2 dx. \tag{2.9}$$

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

2.3.6 Splines de lissage

Supposons que $a < x_1 < ... < x_n < b$, et considérons le modèle non paramétrique suivant :

$$y_i = f(x_i) + \epsilon_i; \tag{2.10}$$

où
y_i, $i = 1, ..., n$ sont des valeurs observées de la variable réponse Y,
f est une fonction lisse inconnue dans $W_2^2[a, b]$,
x_i, $i = 1, ..., n$ sont des valeurs observées de la variable X,
ϵ_i, $i = 1, ..., n$ sont des erreurs normalement distribuées, de moyenne nulle et de variance σ^2.
L'estimateur de f est donné par la minimisation de la somme convexe pour $q \in]0, 1[$:

$$(1 - q) \sum_{i=1}^{n} (y_i - f(x_i))^2 + q \int_a^b f^{(r)}(x)^2 dx, \alpha > 0. \tag{2.11}$$

On pose $\alpha = \frac{q}{1-q}$. La formule précidents devient :

$$\sum_{i=1}^{n} (y_i - f(x_i))^2 + \alpha \int_a^b f^{(r)}(x)^2 dx, \alpha > 0,$$

Il est equivalent d'estimer f par la fonction $\hat{f}$ qui minimise la fonctionnelle suivante par rapport à toute fonctions f dans $W_2^r[a, b]$.

Théorème (2.6) : [Schoenberg][34]

Le minimum du problème (2.11) admet une solution unique $\hat{f}(x)$ qui est une fonction spline dans l'ensemble $S_{2r}(x_1, ..., x_n)$. Nous allons traiter le cas où $r = 2$, déjà vue précédemment, ce cas est souvent utilisé car il donne un algorithme trés simple pour la construction de la fonction spline. En plus, les spline cubique sont facilement évaluées et

donnent "en général" des résultats satisfaisants. Poser $r = 2$ revient à résoudre le problème de minimisation suivant : Trouver la fonction f qui minimise

$$R(\hat{f}) = \sum_{i=1}^{n}(y_i - f(x_i))^2 + \alpha \int_a^b (f''(x))^2 dx, \tag{2.12}$$

par rapport à toute les fonctions f dans $W_2^2[a, b]$.
la solution du problème (2.12) est une spline de lissage cubique naturelle aux noeuds x_i, $i = 1, ..., n$ (voir [Reinsh][29]).

2.3.7 Existence et unicité de la spline de lissage minimisante

Théorème (2.7) : [Thomas-Agnan][40]

Etant donnés n points (x_i, y_i), d'abscisses distinctes dans l'intervalle $[a, b]$ et un réel $\alpha > 0$, il existe une fonction et une seule $\hat{f}$ de l'espace de Sobolev $W_2^2[a, b]$ qui minimise la quantité $R(\hat{f})$ 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 $x_1, ..., x_n$.

Théorème (2.8) : [Cao (2008)][4]

Supposons que $n \geq 0$ et que le paramètre de lissage α est positif, alors $\hat{f}$ est une spline cubique naturelle telle que :

$$F = (I + \alpha K)^{-1} Y,$$

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

$$R(\hat{f}) \leq R(f). \tag{2.13}$$

2.3.8 Propriétés de l'estimateur splines de lissage

1- Biais de l'estimateur

$$Biais(\hat{f}, f) = E\hat{f} - f = (A_\alpha I)f. \tag{2.14}$$

2-Variance de l'estimateur

$$Var(\hat{f}) = E||\hat{f} - E\hat{f}||^2 = \sigma^2 tr(A_\alpha^t A_\alpha). \tag{2.15}$$

2.3.9 Propriétés asymptotiques de l'estimateur

1- Convergence en moyenne quadratique

Théorème (2.9) : [Wahba][41]

Supposons que $\sigma^2 \neq 0$.
Pour avoir la convergence il faut que α décroît avec n. Si cette condition est vérifiée, on a alors ;

— $\hat{f}_\alpha(x) \longrightarrow f(x)$, en moyenne quadratique.

Dans le modèle (2.4), [Craven and Wahba][9] ont majoré l'erreur moyenne quadratique de l'estimateur $\hat{f}_\alpha$ aux points d'observations $x_1, ..., x_n$. Avec les notations

$$f = (f(x_1), ..., f(x_n))^t \text{ et } \hat{f}_\alpha = (\hat{f}_\alpha(x_1), ..., \hat{f}_\alpha(x_n))^t.$$

cette erreur est appelée MASE s'écrit

$$n^{-1}E||f - \hat{f}_\alpha||^2.$$

Théorème (2.10) : [Craven and Wahba][9]

Dans le modèle (2.4), on suppose que
— Les points d'observations sont tel que $\int_0^{x_j} \omega(x)dx = \frac{j}{n}$ pour $1 \leq j \leq n$ où ω est une fonction strictement positive et continue sur $[0, 1]$;
— La fonction f est dans l'espace de Sobolev $W_2^r[0, 1]$ pour un $r \geq 1$ donné ;
— Les erreurs $\epsilon_1, ..., \epsilon_n$ 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 $\hat{f}_\alpha$ d'ordre $2r$ est consistant si et seulement si

$$\alpha = \alpha(n) \to 0 \text{ et } \alpha n^{2r} \to \infty \text{ lorsque } n \to \infty.$$

D'autre part, la MASE de cet estimateur vérifie

$$\frac{1}{n}E||f - \hat{f}_\alpha||^2 = O(\alpha) + O(\frac{1}{\alpha^{\frac{1}{2r}}n}) \tag{2.16}$$

Considérons maintenant l'erreur moyenne quadratique intégrée, notée MISE. Dans le cas du modèle (2.4), sous l'hypothèse du bruit blanc, [Ragozin(1983)][28] a majoré la MISE des dérivées de la fonction par les les dérivées correspondantes de la spline de lissage $\hat{f}_\alpha$.

Théorème (2.11) : [Ragozin][28]

Dans le modèle (2.4), on suppose que
— Les points d'observations sont équidistants : $x_j = j_n, 1 \leq i \leq n$;
— La fonction f est dans l'espace de Sobolev $W_2^k[0, 1]$ avec $0 < kr$;
— Les erreurs ϵi sont centrées, décorrelées et de variance commune σ^2

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

$$E|f - \hat{f}_\alpha|_j^2 = E\int_0^1 (f^j(x) - \hat{f}_\alpha^j(x))^2 dx \leq P(\alpha + n^{-2r})^{\frac{k-j}{r}} |f|_k^2 + \frac{Q\sigma^2}{n^{\frac{\alpha(2j+1)}{2r}}} \tag{2.17}$$

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

$$E|f - \hat{f}_\alpha|_j^2 = O(n^{\frac{-2r}{2k+1}}).$$

2.3.10 Choix du paramètre de lissage

1-Méthode de la validation croisée(CV)

soit $(A)_{ii}$ le i^{me} élément de la diagonale associée à la matrice de lissage A_α.

Théorème(2.12) [Moulines][24]

Le score de la validation croisée vérifie :

$$CV(\alpha) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{f}(x_i)}{1 - (A_{\alpha ii})}\right)^2, \tag{2.18}$$

α est choisi de telle manière qu'il minimise $CV(\alpha)$.

2-Méthode de la validation croisée généralisée (GCV)

Il suffit de remplacer les dénominateurs $1-(A_{\alpha ii})$ dans la validation croisée $CV(\alpha)$ par leurs moyenne $1 - \frac{1}{n}tr(A_\alpha)$ ainsi le score de la validation croisée généralisée :

$$GCV(\alpha) = \frac{1}{n}\frac{\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2}{(1 - \frac{1}{n}tr(A_\alpha))^2} \tag{2.19}$$

α est choisi de telle manière qu'il minimise $GCV(\alpha)$.

$$\alpha GCV = \arg\min GCV(\alpha)$$

3-Estimation du risque sans biais

Le risque associé à l'estimateur $\hat{f}$ de f est donnée par :

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

Le principe 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 $||(I - A_\alpha)f||^2$. Le principe c'est le remplacer $||(I - A_\alpha)f||^2$ par son estimateur sans biais, qui peut être calculé par

$$E||(I - A_\alpha)Y||^2 = ||(I - A_\alpha)f||^2 + \sigma^2 n - 2\sigma^2 tr(A_\alpha) + \sigma^2 tr(A_\alpha A_\alpha^t).$$

Par consequent

$$||(I - A_\alpha)f||^2 = E||(I - A_\alpha)Y||^2 - \sigma^2 n + 2\sigma^2 tr(A_\alpha) - \sigma^2 tr(A_\alpha A_\alpha^t).$$

Le risque devient

$$R(f, \hat{f}) = \frac{1}{n} E||(I - A_\alpha)Y||^2 - \sigma^2 + 2\frac{\sigma^2}{n} tr(A_\alpha)$$

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

$$\hat{\alpha} = \arg\min_{\alpha>0} ||Y - A_\alpha Y||^2 + 2tr(A_\alpha) \tag{2.20}$$

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

$$\hat{\sigma^2} = \frac{||Y - A_\alpha Y||^2}{n}$$

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

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

Conclusion

Dans ce chapitre nous avons présenter l'estimation non-paramétrique de la fonction de régression par la méthode de noyau et spline. Ainsi nous avons donné leurs propriétés (l'espérance, la variance). Ensuit, nous avons utilisé la méthode de validation croisée pour le choix du paramètre de lissage pour chaques méthodes.

Dans le chapitre suivant nous allons comparer les deux méthodes d'estimation de la courbe de régression de la moyenne à savoir la méthode du noyau et la méthode des splines par la simulation et un cas reél.

3
Application

Introduction

Dans ce chapitre, nous allons comparer deux méthodes d'estimation de la courbe de régression $f(x)$ à travers les différents simulations et un cas réels : Méthodes noyaux et méthodes splines. La comparaison est basée sur les critères suivants ASE (erreur moyenne quadratique)

3.1 Application sur R

Dans cette partie, nous allons utilisée le logiciel R, c'est un langage de programmation, un logiciel libre destiné aux statistiques et à la science des données soutenu par la R Foundation for Statistical Computing. Il fait partie de la liste des paquets GNU et est écrit en C.

On a utilisé le R pour calculer et représenter graphiquement la fonctions de régression et leurs estimateurs afin de pouvoir les comparer dans des situations simulées.

Le ASE est calculer par la formul suivante :

$$ASE(\hat{f}, f) = \frac{1}{n}\sum((\hat{f}(x) - f(x))^2)$$

3.2 Etude comparatif

3.2.1 Méthode de noyau

Rappelons qu'on suppose que lon a observé un échantillon $(X_i, Y_i); i = 1, ..., n$ et on veut expliquer la variable aléatoire Y_i par X_i. De plus, on suppose que le modèle est donné par l'expression :

$$y_i = f(x_i) + \epsilon_i \ ;$$

où ϵ_i est l'aléatoire centré et indépendente de X_i. Aussi la fonction de régression :

$$f(x) = E[Y/X = x] = \frac{\int ym(x, y)\partial y}{m_X(x)}$$

où $m_X(x)$ est la densité de la variable X, et $g_X(x) = \int ym(x, y)\partial y$.

Nous avons vu que $f(x)$ est estimé par la quantité

$$\hat{f}(x) = \frac{\sum Y_i K(\frac{X_i - x}{h_n})}{\sum K(\frac{X_i - x}{h_n})} = \frac{\hat{g}_X(x)}{\hat{m}_X(x)};$$

Il dépend de la taille de l'échantillon n, et aussi du noyau K et de la fenêtre h_n qu'il faut choisir pour calculer $\hat{f}(x)$: avec $\hat{g}_X(x)$ est l'estimateur naturel de $g_X(x)$:

$$\hat{g}_X(x) = \frac{1}{nh_n} \sum Y_i K_{hn}(X_i - x);$$

et $\hat{m}_X(x)$ l'estimateur du noyau de la densité de la fonction m

$$\hat{m}_X(x) = \frac{1}{nh_n} \sum K_{h_n}(X_i - x).$$

Dans la suite de ce chapitre, nous supposons que notre modèle est de la forme

$$y = f(x) + \epsilon, \quad o\grave{u} \quad \epsilon \to N(0; \sigma^2)$$

3.2.2 Méthode des splines

L'estimateur de la fonction de régression spline :

$$\hat{f}_\alpha(x_i) = (I + \alpha K)^{-1} Y = A_\alpha Y.$$

avec A_α une matrice lisse qui ne dépend que des x_i, $i = 1, ..., n$, l'estimateur est linéaire par rapport au vecteur Y.

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

$$\hat{\alpha} = \arg\min \frac{1}{n} \frac{\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2}{(1 - \frac{1}{n}tr(A_\alpha))^2} \; ;$$

Les simulations sont effectuées pour différentes tailles d'échantillon de plus en plus grandes :

$$n \in \{50, 100, 200, 500, 1000, 2000\}$$

3.3 Modèles de fonction

Les modèles considérés sont :
1.$f_1(x) = 4.26(exp(-3.25x_i) - 4exp(-6.5x_i) + 3exp(-9.75x_i)) + e_i$,[Eubank][12]
2.$f_2(x) = x_i + 0.5exp(-50(x_i - 0.5)^2) + e_i$.[Eubank][12]

3.3.1 Modèle $f_1(x)$

1. x est uniforme sur $[0, 5]$;
2. e sont les résidus qui sont normalement distribués de moyenne 0 et de variance $\sigma^2 = 0.05$;

ASE associé à f_1

n	ASE noyau	ASE spline
50	0.1480599	0.0898976
100	0.0975050	0.0654594
200	0.0750013	0.0497245
500	0.0655905	0.0374850
1000	0.0674577	0.0216823
2000	0.0414707	0.0160180

TABLE 3.1 – ASE donnée par les deux méthodes associée au modèle $f_1(x)$ en fonction de la taille de l'échantillon n.

Pour $n = 50$

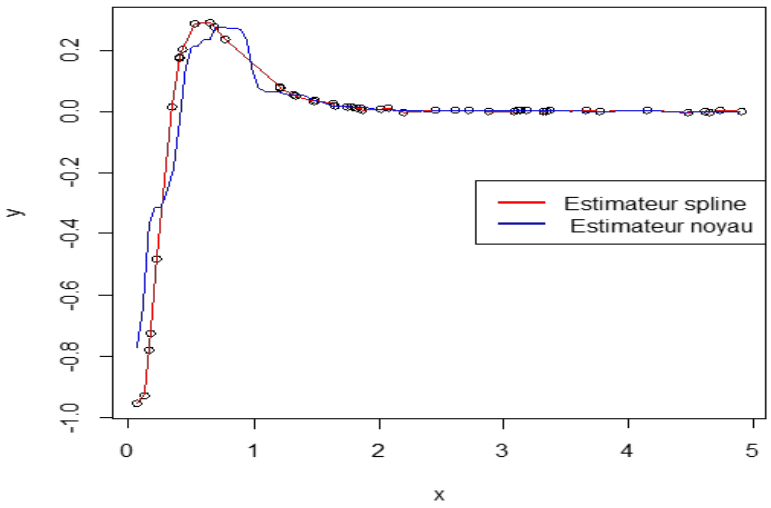

FIGURE 3.1 – Estimation de f_1, $n = 50$.

Pour $n = 100$

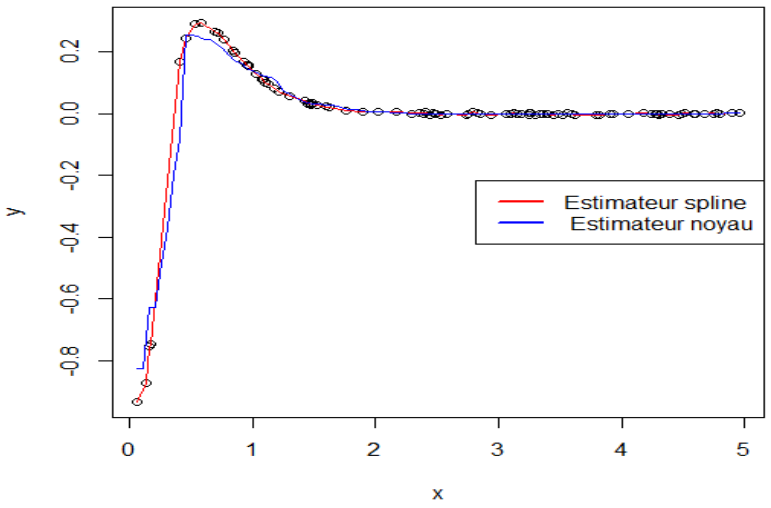

FIGURE 3.2 – Estimation de f_1, $n = 100$.

Pour $n = 200$

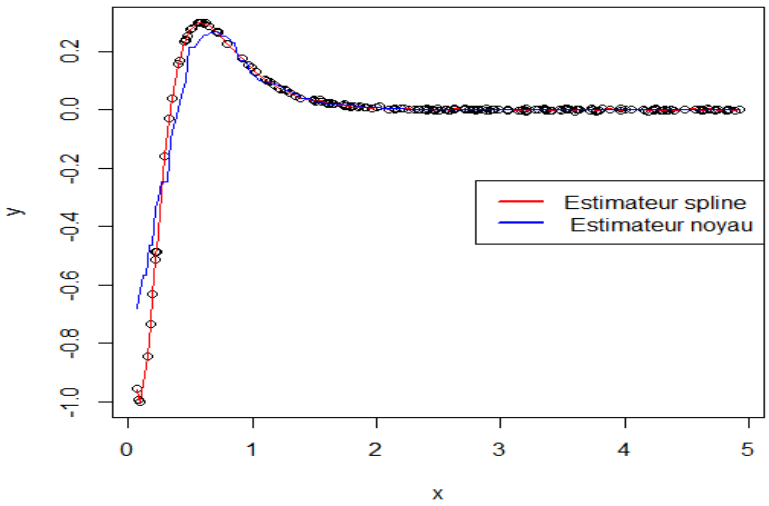

FIGURE 3.3 – Estimation de f_1, $n = 200$.

Pour $n = 500$

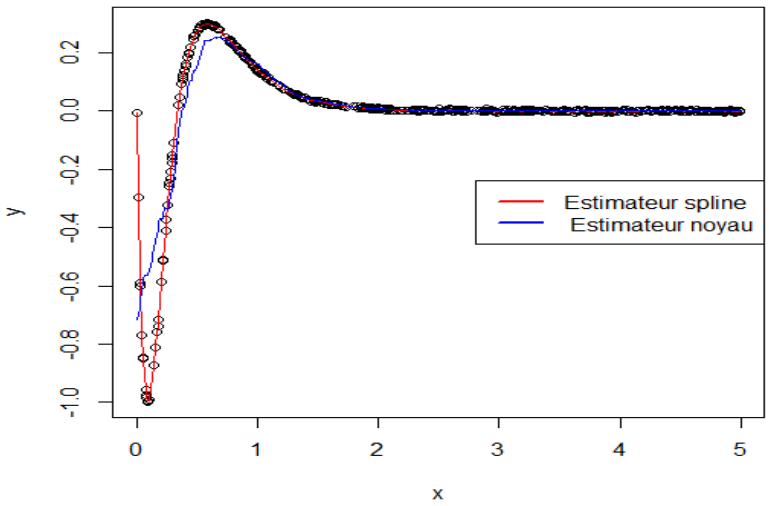

FIGURE 3.4 – Estimation de f_1, $n = 500$.

Pour $n = 1000$

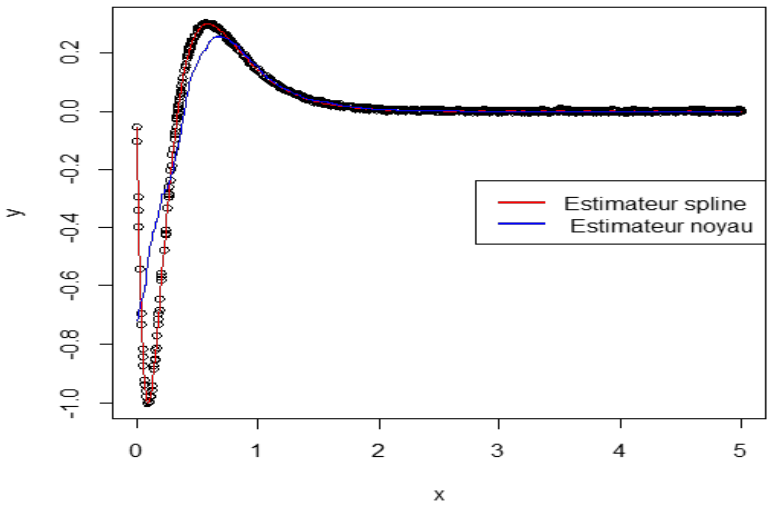

FIGURE 3.5 – Estimation de f_1, $n = 1000$.

Pour $n = 2000$

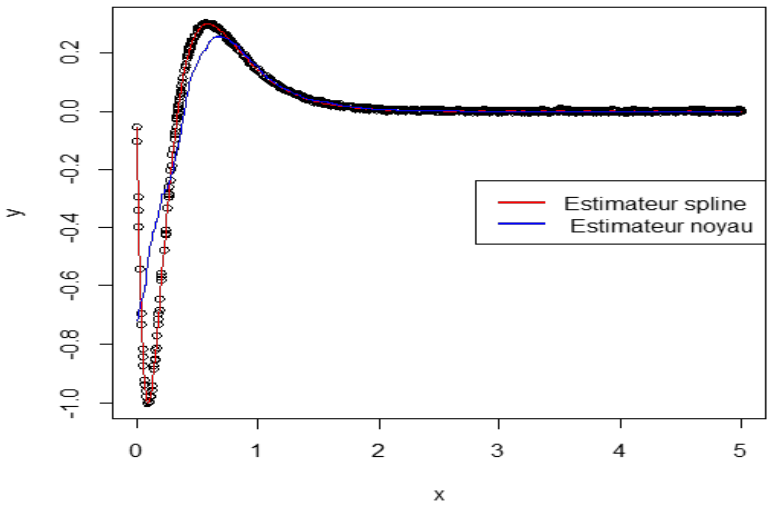

FIGURE 3.6 – Estimation de f_1, $n = 2000$.

Discution des résultats du modèle f_1 :

D'aprés les résultats du tableau(3.1) on peut observée que :

• La valeur de ASE diminue lorsque la taille d'échantillon n augmente et ce résultat est vrai pour les deus méthode.

On remarque que quelque soit la taille n de l'échantillon l'erreur associée à la méthode des fonction spline est la plus petite par rapport a la méthode du noyaux (ASE spline $<$ ASE noyaux),

Graphiquement :

On remarque que la meilleur méthode qui ajuste bien les données est la méthode des splines.
On conclue que la méthode des splines est meilleure que celle du noyau.

3.3.2 Modèle $f_2(x)$

1.x est uniforme sur $[0, 3]$;
2.e sont les résidus qui sont normalement distribués de moyenne 0 et de variance $\sigma^2 = 0.0225$;

ASE associé à f_2

n	ASE noyau	ASE spline
50	0.516535	0.08642
100	0.405451	0.07123
200	0.065675	0.00684
500	0.045677	0.00478
1000	0.033895	0.00474
2000	0.017639	0.00356

TABLE 3.2 – ASE donnée par les deux méthodes associée au modèle $f_2(x)$ en fonction de la taille de l'échantillon n.

Pour $n = 50$

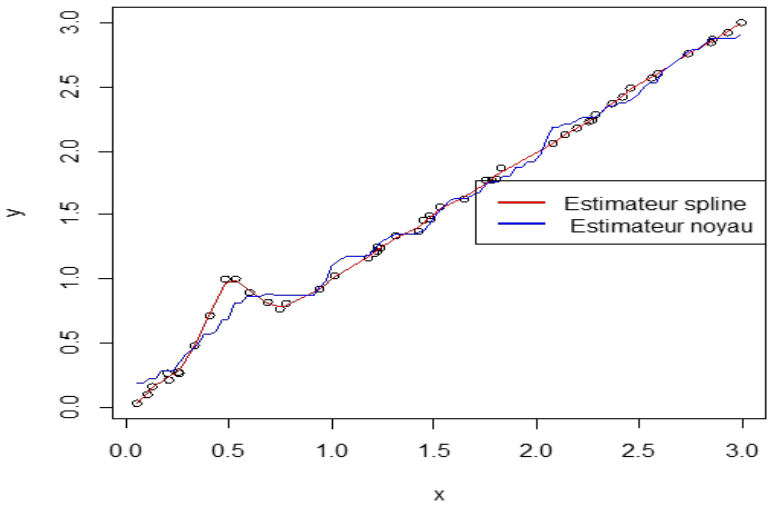

FIGURE 3.7 – Estimation de f_2, $n = 50$.

Pour $n = 100$

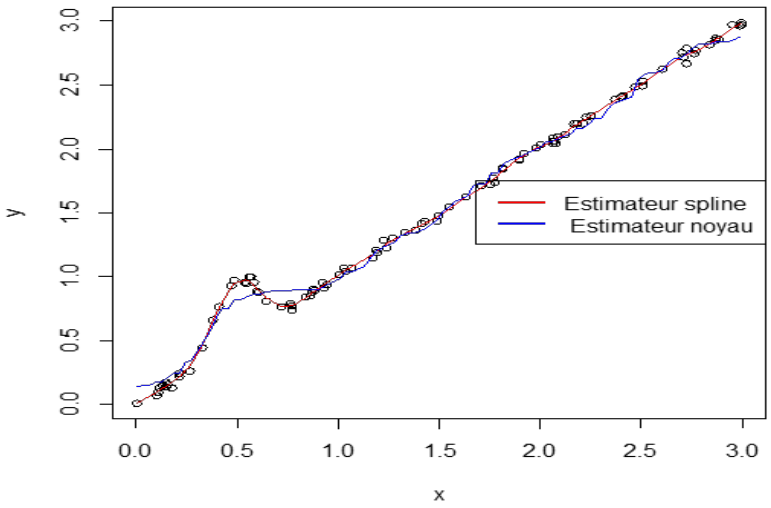

FIGURE 3.8 – Estimation de f_2, $n = 100$.

Pour $n = 200$

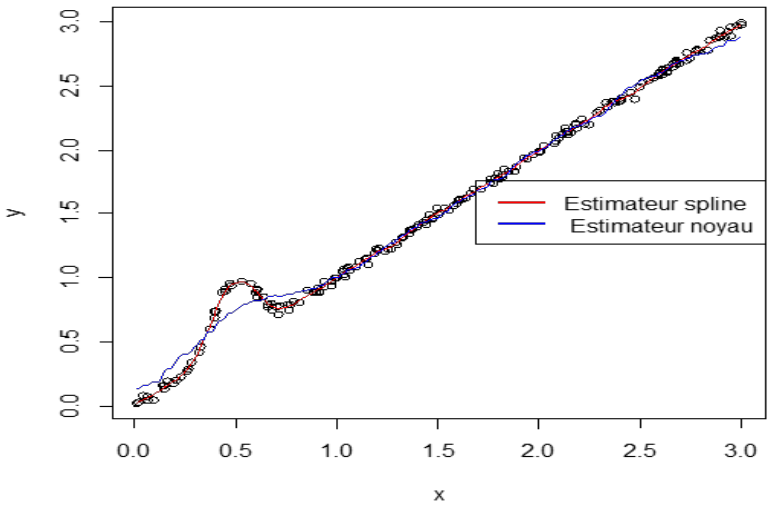

FIGURE 3.9 – Estimation de f_2, $n = 200$.

Pour $n = 500$

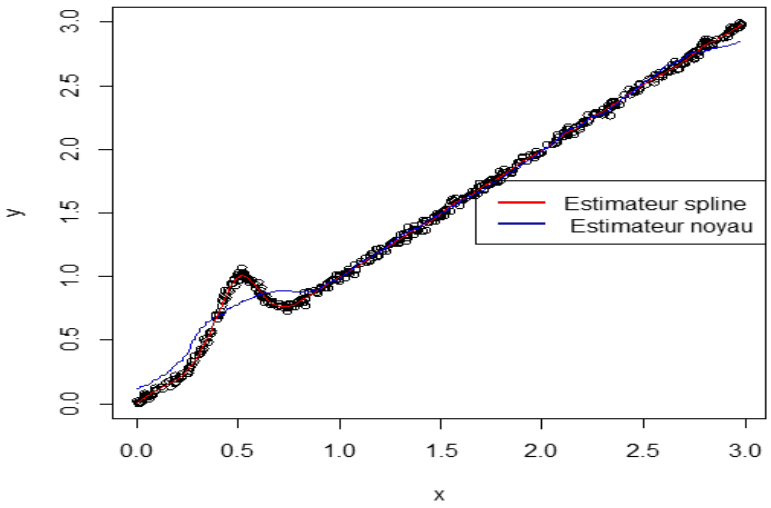

FIGURE 3.10 – Estimation de f_2, $n = 500$

Pour $n = 1000$

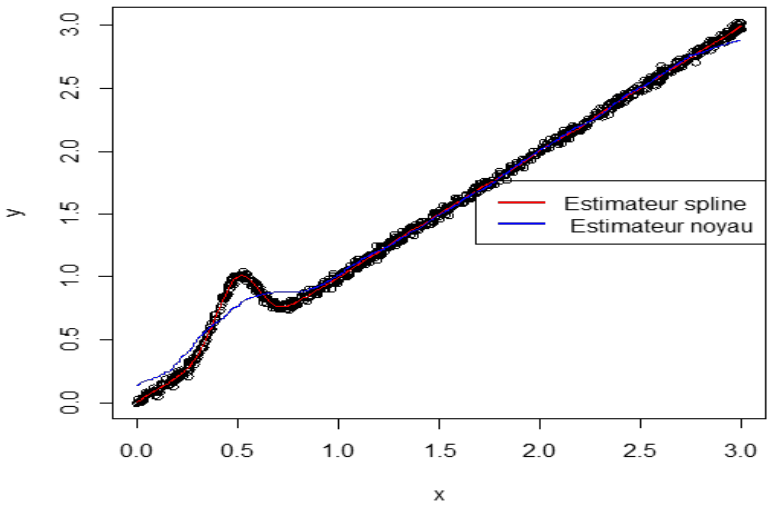

FIGURE 3.11 – Estimation de f_2, $n = 1000$.

Pour $n = 2000$

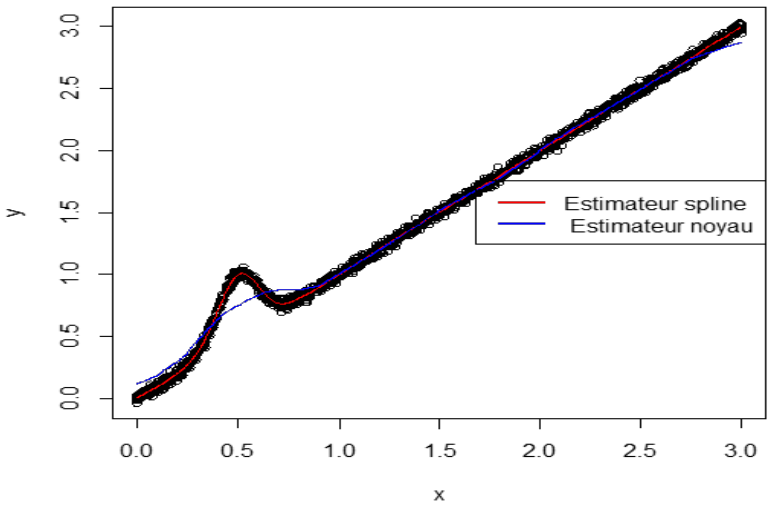

FIGURE 3.12 – Estimation de f_2, $n = 2000$.

Discutez les résultats du modèle f_2 :

D'aprés les résultats du tableau(3.2) on peut observée que :
Les valeur de ASE diminue lorsque la taille d'échantillon n augmente,ce résultat est vrai pour les deux méthodes .
On trouve les mêmes résolution que le modèle f_1 .Donc la méthode optimale pour nos données est toujours la méthode spline .Et cette constatation est vérifiée graphiquement.

3.3.3 Cas réel

Soit l'ensemle données réelles concerne les données de l'éolienne, voir [Ebank(1999)][12] :
•La variable dépendante Y est la sortie du courant continu.
•La variable indépendante X est la vitesse du vent (miles/h).

x	y	x	y	x	y
2.45	0.123	2.70	0.500	2.90	0.653
3.05	0.558	3.40	1.057	3.60	1.137
3.95	1.144	4.10	1.194	4.60	1.562
5.00	1.582	5.45	1.501	5.80	1.737
6.00	1.822	6.20	1.866	6.35	1.930
7.00	1.800	7.40	2.088	7.85	2.179
8.15	2.166	8.80	2.112	9.10	2.303
9.55	2.294	9.40	2.386	10.00	2.236
10.20	2.310				

TABLE 3.3 – Données sur les éoliennes.

ASE associée au cas réel

ASE noyau	ASE spline
0.751359	0.386751

TABLE 3.4 – ASE associée au cas réel.

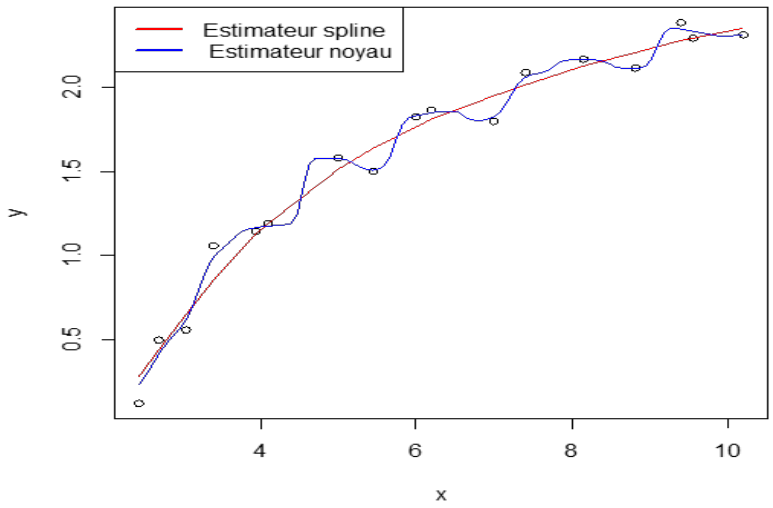

FIGURE 3.13 – Estimation de la courbe d'éoliennes par la méthode du noyau et la méthode des fonctions splines.

Discution des résultats :

Pour les résultats du tableau (3.3) en conclus que l'erreur associée à la méthode des fonctions splines est petite par rapport à celle donnée par la méthode du noyau. Ce qui signifie que la méthode des spline estime bien les données même dans le cas réel.

Conclusion

Dans ce dérnier chapitre, nous avons comparé la méthode du noyau et la méthode des splines pour éstimer la fonction de régréssion. Les résultats obtenus montrent que la méthode des splines donne de meilleurs résultats que celle du noyau et ces résultats sont confirmés graphiquement.

Conclusion générale

Dans ce travail nous avons présenté deux méthodes d'estimation non paramétrique de la fonction de régression, les méthodes étudiées sont la méthode du noyau et la méthode des fonctions des splines . Le but principal est de comparer ces deux méthodes en s'appuyant sur des exemples de fonctions de régressions simulé 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 le premier chapitre nous avons défini quelques méthodes d'éstimation non paramétrique de la fonction de régréssion, la méthode de noyau, les k plus proche voisin, la méthode des séries orthogonales et la méthode des splines .
Dans le deuxième chapitre, nous nous sommes intéressé à la méthode du noyau et la méthode spline, nous avons défini les estimateurs et leurs différentes propriétés statistiques et asymptotiques.

La deuxième partie de ce travail, est la partie simulation ,l'utilisent le logiciel R qui nous a permis de comparer par simulation, sur des fonctions cibles et un cas réel les deux méthode. Cette comparaison basé sur le critère ASE .
On conclut :

. Le meilleur résultat est obtenu avec la méthode de spline.

. l'erreur assocer à la méthode des spline est la plus petite à celle associer a la méthode du noyaux, (ASE spline $<$ ASE noyau). Donc le meilleur résulltat est obtenu par la méthode des spline.

Bibliographie

[1] Adjabi, S., Amroun, S., et al. *L'estimation de la courbe de régression de la moyenne.* PhD thesis, Universié de bejaia, 2011.

[2] Barbillon, P. Méthodes d'interpolation à noyau pour l'approximation de fonction type boît noire coûteuses. *Université Paris Sud-Paris* (2010).

[3] Bosq, D. Estimation sur optimale de la densité par projection. *Canadian Journal of Statistics 33* (2005), 21–37.

[4] Cao, Y. Inégalités d'oracle pour l'estimation de la régression. *Thèse Doctorat, Université de Provence* (2008).

[5] Cao, Y., and Golubev, Y. On oracle iniqualities related to smoothing spline. *Mathematical Methods of Statistics 15* (2006), 398–414.

[6] Cencov, N. Evaluation of an unkown distribution density fromobservations. *Jornal of the Institute of Mathematics and Its Aplications 18* (1559-1962).

[7] Collomb, G. 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*, 15 (1977), 24–46.

[8] Cover, TM et Hart, P. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory 13* (1967), 21–27.

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

[10] Epanechnikov, V. A. Nonparametric estimation of a multidimensional probability density. *Theory Probab. Appl* (1969).

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

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

[13] Ferraty, F., and Vieu, P. Statistique fonctionnelle : Modèl non paramétrique de régression. *Note de coursde DEA* (2002).

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

[15] Hardle, W. Robust regression function estimation. *Journal of Multivariate Analysis 14* (1984), 169–80.

[16] Hardle, W. Applied nonparametric regression. *Cambridge University Press, Cambridge* (1990).

[17] Hardle, W. Applied nonparametric regression. *Humboldt-Universit at zu Berlin.* (1994).

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

[19] KHADRAOUI, R. Estimation ā noyau de la fonction de régression.

[20] Kronmal, R., and M.Tarter. the estimation of probability densities and cumulatives by fourier series methodes. *J.Amer. Statist. Assoc 63* (1968), 925–952.

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

[22] Loftsgaarden, D. O., and Quesenberry, C. P. A nonparametric estimate of a multivariate density function. *Annals of Mathematical Statistics 36* (2005), 1049–51.

[23] M.H, S. Spline analysis. *Prentice Hall, Englewood Cliffs, N J* (1973).

[24] Moulines, M. C. Classification et sélection de caractéristiques de textures.

[25] Nadaraya, E. A. On estimating regression. *Theor. Prob. Appl. 9* (1964), 141–142.

[26] Oliveira, J. T. D. Estatéstica de densidades : resultados assintoticos. *Universidade de Lisboa. Faculdade de Ciencias de Lisboa,* (1963).

[27] Parzen, E. On estimation of a probability density function and mode. *The annals of mathematical statistics 33*, 3 (1962), 1065–1076.

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

[29] Reinsch, C. Smoothing by spline function. *Numererisch Matheematik 10* (1967), 177–183.

[30] Rosenblatt, M. Remarks on some nonparametric estimates of a density function. *Annals of Mathematical Statistic 27* (1956), 832–837.

[31] Rutkowski, L. 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* (1981), 263–44.

[32] Saadi, N., and Adjabi, S. On the estimation of the probability density by trigonometric series. *communications in Statistics-Theory and Methods 38* (2009), 3583–3595.

[33] Sarda, P., and Vieu, P. Kernal regression. smoothing and regression : Approches, comutation, and application. *Ed. M.G. Schimek, Wiley series in probability and statistics* (2000), 43–47.

[34] Schoenberg, I. J. Spline functions and the problem of graduation. *Mathematics 52* (1964), 974–50.

[35] Schuster., E. F. Joint asymptotic distribution of the estimated regression function at a finite number of points. *Annals of Mathematical Statistics 43*, 1 (1972), 84–88.

[36] Schwartz, S. C. Estimation of probability densities by an orthogonal series. *Annals of Mathematical Statistics 38* (1967), 1261–1265.

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

[38] Szego, G. Orthogonal polynomials. *Amer. Math. Soc. Coll. Publ.* (1959).

[39] t P. Vieu F. Ferraty. Statistique fonctionnelle : Modèle non-paraamétrique de régression. *Notes de cours de DEA.* (2002/2003).

[40] Thomas-Agnan, C. Spline function and filtering. *The Annals of Staistics, 19 :* (1999), 1512–152.

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

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

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

[44] Wahba, G. Data-based optimal smoothing of orthogonal series density estimates. *The Annals of Statistics 9(1)* (1981), 146–156.

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

[46] Wahba, G., and Wold., S. A completely automatic french curve : Fitting spline fuctions by cross validation. *Simulation and Computation* (1975), 1–17.

[47] Walter, G. Properties of hermite series estimation of probability density. *Annals of Statistics 5* (19), 1258–64.

[48] Watson, G. S. Smooth regression analysis. *Sankhyà Ser. A 26* (1964), 359–372.

[49] Whittaker, E. On a new method of graduation. *Proceedings of the Edinburgh Mathematical Society 41* (1923), 63–75.

Résumé

Dans ce travail nous etudiants l'estimateur non paramétrique de la fonction de régréssion pour les données réelles en utilisant la méthode de noyau et splin. Puis nous allons la comparer avec la méthode de ASE (avrege squart error) Finalement un travail de simulation on a donnée des explications graphiques des résultats théoriques appliqués sur des exemples de regression non linéaire à laide du logiciel R. Pour vérifier la bonne estimateurs entre noyaux et spline.

Mots clés : régression non-paramétrique ; Fonction ; Spline de lissage ; noyau ; VCG

Abstract

In this work we study the nonparametric estimator of the regression function. sion for real data using the kernel and spline method. Then we compare it with the ASE (avrege squart errer) method Finally, a simulation work graphical explanations of the theoretical results applied to nonlinear examples of non-linear regression using R.software to check that the estimators between kernels and splin.

words :non-parametric regression ;Function ;Smoothing spline ;kernels ;VCG