Ministère de l'enseignement superieur et de la recherche scientifique
Université Abderrahmane Mira Béjaia
Faculté des sciences exactes
Département de Recherche Opérationnelle

جامعة بجاية
Tasdawit n Bgayet
Université de Béjaïa

Projet de fin d'étude

En vu d'obtention d'un Master en Mathématiques Appliquées

Option :

Modélisation Mathématique et Techniques de Décision

Thème

L'influence Du Choix Du Noyau Discret Dans L'Estimation Non Paramétrique De La Fonction De Régression

Réalisé par :

✓ **IAMARENE** Hamza

✓ **ISSAADI** Ouissem

Devant le jury

Présidente	M^{me}	**S.AMROUN**	**M.C.B**	**Université de Béjaia**
Promotrice	M^{me}	**L.DJERROUD**	**M.C.B**	**Université de Béjaia**
Examinatrice	M^{me}	**Y.ZIANE**	**M.C.B**	**Université de Béjaia**
Examinatrice	M^{me}	**L.HARFOUCHE**	**Docteur**	**Université de Béjaia**

Année universitaire : 2019/2020

Remerciements

En premier lieu, nous remercions le dieu qui nous accorder la santé et la patience afin accomplir ce travail.

Nous tenons tout d'abord à exprimer notre profonde gratitude à monsieur Smail ADJABI Professeur de l'université A. MIRA de Bejaia, nous le remercie pour son soutien et pour la confiance qu'il nous a accordée tout au long de ce travail.

Nous voudrions exprimer toute notre gratitude à M^{me} DJRROUD Lamia la directrice de ce mémoire pour sa patience, sa disponibilité et surtout ces judicieux conseils qui ont contribué à alimenter notre réflexion.

Nous remercions aussi les membre du jury d'avoir accepté de juger notre travail.Nous tenons aussi a exprimer notre reconnaissance a tout les enseignant du département de recherche opérationnelle qui nous ont suivis tout notre cursus universitaire.

Enfin dans le souci de n'oublier personne , nous tenons à remercier tout personne ayant participé ou aidé de prés ou de loin à la réalisation de ce modeste travail.

Table des matières

Liste des tableaux

Table des figures

Introduction générale

Les modèles de régression sont très utiles pour modéliser la liaison entre une variable à expliquer y et une variable explicative x. Ils sont appliqués à de nombreux domaines tels que l'économie, la bio statistique ou encore les sciences de l'environnement.

Dans la littératures statistique, deux grandes classes de modèles de régression sont omniprésentes : les modèles paramétriques et les modèles non paramétriques. L'estimation non-paramétrique de la régression revêt un grand intérêt en statistique mathématique et trouve diverses applications, notamment dans les problèmes de prévision. Le principal avantage de la régression non paramétrique est qu'elle ne suppose aucune forme spécifique pour l'estimateur, ce qui lui donne beaucoup plus de flexibilités. Elle peut donc être utilisée pour décrire la relation entre deux variables lorsque le modèle linéaire ne s'applique pas.

Il existe plusieurs méthodes de l'estimation non-paramétriques de la fonction de régression, Nadaraya [1964][1] et Watson [1964][2] proposent une méthode plus générale, bien connue lorsqu'elle est utilisée pour estimer une densité de probabilité et appelée habituellement "méthode du noyau ", la méthode de régressograme, introduit par Tukey[1961][3], la méthode d'estimation par les séries orthogonales proposée pour estimer des densités continues a été développée à partir des travaux de Cencov[1962][4], et étudiée ensuite par plusieurs auteurs ; voir par exemple, Schwartz[1967][5], Kronmal and Tarter [1968][6], Wahba [1981][7], Bosq [2005][8] et Saadi and Adjabi [2009][9], la méthode de lissage par les fonction splines a été dénveloppée et utilisée par plusieurs auteurs nous citons les travaux de Reinsch [1967] [10],Silverman [1985][11], Wahba [1990][12] et Eubank [1999][13].

La méthode qui a rencontré beaucoup plus de succés auprès de la communauté est la méthode d'estimation par noyau. Ce succés peut s'expliquer par au moins trois raisons : d'abord, l'expression théorique de l'estimateur est trés simple puisque il s'écrit comme la somme de n variables aléatoires indépendantes et identiquement distribuées, en utilisant la fonction noyau K et le paramètre de lissage h. Ensuite, il est convergeant en de nombreux sens. Enfin, l'estimateur à noyau est flexible, car il laisse à l'utilisateur une grande latitude dans le choix du noyau K et du paramètre

de lissage h.
L'estimation non paramétrique de la fonction de régression discrète a été beaucoup moins étudiée par rapport à l'estimation de la fonction de régression dans le cas continu. Aitchison and Aitken [1976][14] ont proposé un estimateur à noyau discret pour estimer des fonctions discrètes où les données sont de type catégorielles. Récemment, deux classes de noyaux discrets, à savoir les noyaux discrets standards et les noyaux triangulaires discrets, ont été proposées par Kokonendji et al. [2007b][15], Senga Kiessé [2008][16] et Kokonendji and Senga Kiessé [2011][17] pour estimer des fonctions discrètes à support discret . La sélection du paramètre de lissage par la méthode classique de validation croisé a été étudiée par Senga Kiessé [2008][16].

L'objectif de mémoire est d'étudier l'influence du choix du noyau discret dans l'estimation non paramétrique de la fonction de régression. Les noyaux utulisés sont le noyau binomial, binomial négative, poisson et triangulaire qui ont été étudié par Senga Kiessé [2008][16] et les noyaux asociés discret dirac uniforme, Wang vang Ryzan et le noyau Optimal qui ont été étudier par senga kiessé and Gilles durrieu [2020][18] dans l'estimation de la fonction de densité.

Dans ce mémoire nous exposerons en détail la méthode du noyau associé discret pour l'estimation de la fonction de régression . Il est composé principalement de trois chapitres : Après l'introduction générale vient Le premier chapitre qui est composée de deux parties :
✓ La première partie porte sur des généralités sur la fonction de régression non paramétrique .
✓ La deuxième patrie du chapitre propose des méthodes d'estimations de la fonction de régression : la régressograme, la méthode des séries orthogonales, la méthode de lissage par les fonction splines et plus particulièrement la méthode du noyau et ses propriétés statistiques.

Dans le deuxième chapitre, nous présentons la méthode du noyau associé pour l'estimation non paramétrique de la fonction de régression dans le cas discret .Ensuite, nous donnons les propriétés statistiques (biais, variance, l'erreur quadratique) de l'estimateur et ses propriétés asymptotiques. Enfin, nous citons des exemples sur les noyaux discrets et la méthode de validation croisée pour le choix du paramètre de lissage.

Le dernier chapitre sera consacré à des applications numérique. A l'aide d'un étude de simulation et trois jeux de données nous essayons d' illustrer l'influence du choix du noyau discret et leur performance dans l'estimation non paramétrique de la fonction de régression.

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

1

Estimation non parmétrique de la fonction de régression

1.1 Introduction

Les méthodes non-paramétriques prennent en compte les échantillons (D), et leur répartition spatiale dans l'espace des paramètres, afin de produire une estimation de $f(x|D)$ plus proche de la réalité. Elles estiment la densité de probabilité et la fonction de régression à partir des données directement, sans se donner aucune hypothèse apriori sur la distribution des données. Il s'agit alors de calculer la probabilité en chaque point directement à partir des observations, sans chercher à construire une fonction de distribution définie sur tout le domaine [19].Ce chapitre est consacrée à la présentation des principales méthodes de régression non paramétrique univariée à savoir, la méthode de régressogramme, la méthode des séries orthogonales, la méthode des splines et particulièrement la méthode du noyau. Ces méthodes sont très utiles lorsque l'on veut décrire la relation entre une variable dépendante Y et une variable explicative X, sans supposer une forme particulières.

1.2 Définition de la fonction de régression

On dispose d'un échantillon, composé de n couples indépendants de variables aléatoires $(X_1, Y_1), \ldots, (X_n, Y_n)$. Soit y_i une réalisation de la variable aléatoire Y_i et x_i une réalisation d'une variable explicative X_i pour $i = \{1, \ldots, n\}$.Le modéle de régression non paramétrique est :

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

où, ϵ_i correspond aux résidus telle que $E(\epsilon_i) = 0$ et $Var(\epsilon_i) = \sigma^2$, m est fonction inconnue de régression.
Il existe deux cas principaux pour le modèle (1.1)[20] :

1. **Dispositif expérimental à effets fixes** (où fixed design) :il correspond à la situation où les $X_i = x_i$ sont fixés (c'est à dire, des constantes où, de manière équivalente, déterministes où dégénérées).
2. **Dispositif expérimental à effets aléatoires**(où random design) : désigne le modèle où les données $\{X_i : 1 \leq i \leq n\}$ sont strictement aléatoires (où non dégénérées).

Définition 1.2.1 *On appelle fonction de régression, la fonction $m(x)$ qui a pour toute réalisation x de la variable explicative X associe la quantité :*

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

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

Soient les observations bivariées (x_i, y_i), $i = 1, \ldots, n$, ou les x_i représentent les valeurs observées de la variable aléatoire explicative X et les y_i représentent celles de la variable aléatoire 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

$$Y = \beta_0 + \beta_1 x_1 + \epsilon_i \tag{1.3}$$

où les erreurs aléatoires ϵ_i sont non corrélées, de moyenne nulle et de variance σ^2. Le modèle (1.3) 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 (1.3) peut alors être employé.
la définition formelle de la moyenne conditionnelle d'une variable aléatoire de nouvelle expression(1.3) est[21] :

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

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

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

$$\Rightarrow a = E(Y|X=x)$$

Le fait que la dérivée seconde, qui se chiffre à 2, positive donc a est un minimum, et non un maximum.

1.3 Méthodes de l'estimation non paramétrique de la fonction de régression

1.3.1 Méthode régressogramme

La régressogramme représentent la méthode d'estimation de la fonction de régréssion la plus ancienne. Cet méthode a été proposé par Tukey[1961][3] et l' estimateur appelé le plus souvent "régressogramme" est construit de la façon suivante :
soit $[u,v]$ un intervalle compact et soit une partition de $[u,v]$ en k intervalles $C_{j=1,\ldots,k}$ de même longueur et k entier positif. On estime la fonction de régression par la fonction en escalier qui, dans chaque intervalle de la partition est constante et est égale à la moyenne arithmétique des ordonnées Y_i, $i=1,2,\ldots,n$ des points de l'échantillon dont l'abcisse X appartient à l'intervalle considéré. Si aucun point X de l'échantillon ne prend sa valeur dans cet intervalle on prend alors 0 comme valeur de la fonction[22]. Donc pour tout $x \in C_j$ on estime $m(x)$ comme suite[23] :

$$\tilde{m}_n(x) = \frac{\sum_{i=1}^{n} 1_{C_j}(X_i)Y_i}{\sum_{i=1}^{n} 1_{C_j}(X_i)} \tag{1.4}$$

Il est clair, que le régressogramme $\tilde{m}_n(x)$ est constant sur chaque classe C_j. la première amélioration consiste à centrer la classe sur le point x où l'on estime la fonction

de régression. Ce régressogramme mobile s'écrit

$$\tilde{m}_n(x) = \frac{\sum_{i=1}^{n} 1_{[x-h;x+h[}(X_i)Y_i}{\sum_{i=1}^{n} 1_{[x-h;x+h[}(X_i)}$$

Il en découle que :

$$\tilde{m}_n(x) = \frac{\sum_{i=1}^{n} 1_{[-1;1[}(X_i - x/h_n)Y_i}{\sum_{i=1}^{n} 1_{[-1;1[}(X_i - x/h_n)}$$

1.3.2 Méthode de noyau

Description de la méthode

La méthode du noyau est une méthode qui est communément utilisée pour faire de la régression non paramétrique. Elle donne pour estimateur de (1.2) une moyenne pondérée des valeurs y_i pour les i dont le point x_i est prés du point d'estimation. Pour appliquer cette méthode, il faut suivre les six étapes ci-dessous :

1. Tout d'abord, 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.
2. Dans un autre temps, une fonction de noyau symétrique autour de 0 et unimodale doit être choisie. Cette fonction est maximale en 0 à l'exception du noyau gaussien, est non nulle uniquement dans la région [-1,1]. La figure (1.1) donne un bref aperçu de certaines fonctions de noyau pouvant être employées .Voir Hastie et Tibshirani[1990][24] et Schimek [2000] [25].

3. Le choix d'un paramètre de lissage h, qui peut uniquement prendre des valeurs positive. Dans Simonoff [1996][21], une manière théorique basée sur la validation croisée est présentée afin de déterminer la valeur que devrait prendre le paramètre h.

4. Par la suite, le poids associé à chacune des observations doit être calculé. Ces poids sont d'ailleurs obtenus comme suit :

$$\omega_i(x_0) = K_h(X_i - x_0) = \frac{1}{h} K\left(\frac{X_i - x_0}{h}\right)$$

5. L'estimateur de $m(x_0)$ est la moyenne pondérée des valeurs de Y

$$\hat{m}(x_0) = \frac{\sum_{i=1}^{n} \omega_i(x_0)Y_i}{\sum_{i=1}^{n} \omega_i(x_0)}$$

L'estimateur du noyau qui est ainsi obtenu est :

$$\hat{m}_{NW}(x_0) = \frac{\sum_{i=1}^{n} K_h(X_i - x_0)Y_i}{\sum_{i=1}^{n} K_h(X_i - x_0)}$$

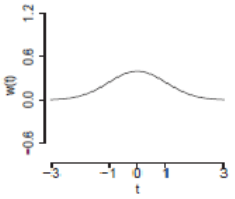
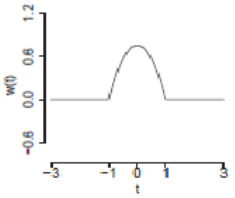
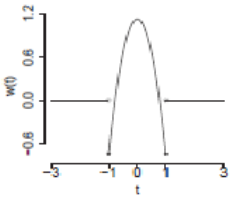
Noyau	Définition	Illustration
Normale(0,1)	$d(t) = \frac{1}{2\pi}\exp(\frac{1}{2}t^2)$	
Epanechnicov	$d(t) = \begin{cases} \frac{3}{4}(1-t^2), & \text{pour} \lvert t\rvert \leq 1 \\ 0 & \text{sinon.} \end{cases}$	
Variance Minimale	$d(t) = \begin{cases} \frac{3}{8}(3-5t^2), & \text{pour} \lvert t\rvert \leq 1 \\ 0 & \text{sinon.} \end{cases}$	

FIGURE 1.1 – Définition de différents noyaux.

6. En général, $\hat{m}_{NW}(x_0)$ est estimé pour plusieurs valeurs de x_0 sur une fine grille afin d'obtenir une courbe de $\hat{m}(x)$ en fonction de x. En conclution, l'estimateur à noyau introduit par Nadaraya-Watson est

$$\hat{m}_{NW}(x) = \frac{\sum_{i=1}^{n} K_h(X_i - x)Y_i}{\sum_{i=1}^{n} K_h(X_i - x)} \tag{1.5}$$

Les propriétés

Lorsque on veut comparer plusieurs estimateurs, il faut calculer certaines mesures permettant d'évaluer leurs qualités, telles que le biais, la variance et L'erreur quadratique moyenne.Nous représentons dans cette section les résultats Simonoff [1996][21] sur les propriétés de l'estimateur(1.5) .

•Biais

$$\begin{aligned} biais\{\hat{m}_{NW}(x)\} &= E[\hat{m}_{NW}(x) - m(x)] \\ &= h^2 \left[\frac{m^{'}(x)f_X^{'}(x)}{f_X(x)} + \frac{m^{"}(x)}{2}\right]\mu_2(K_{(0)}) + o_p(h^2) \end{aligned}$$

où

$$\mu_q(K_{(p)}) = \int u^q K_{(p)}(u)du$$

$f_X(x)$ est la fonction de densité des données exogènes x_i et $K_{(p)}$ est le noyau d'ordre $(p+1)$ lorsque p est impair et le noyau d'ordre $(p+2)$ lorsque p est pair .voir Simonoff [1996][21].

• **Variance**

$$\begin{aligned} Var\{\hat{m}_{NW}(x)\} &= E\left[(\hat{m}_{NW}(x) - E[\hat{m}_{NW}(x)])^2\right] \\ &= \frac{R(K_{(0)})\sigma^2(x)}{nhf_X(x)} + o_p[(nh^{-1})] \end{aligned}$$

où

$$R(K_{(p)}) = \int K_{(p)}(u)^2 du$$

La formule pour l'erreur quadratique moyenne peut ainsi être obtenue :

$$MSE\{\hat{m}_{NW}(x)\} = Var\{\hat{m}_{NW}(x)\} + Biais^2\{\hat{m}_{NW}(x)\}$$

Choix du paramètre de lissage

Le choix du paramètre de lissage est un point important de la méthode de noyau. Voir Hall[1984] [26] et Hardle et Marron [27].

La validation croisée généralisée(GCV)

l'approche de la fonction validation croisée généralisée est :

$$\hat{\delta}^2(h) = \frac{1}{n}\sum_{i=1}^{n}\{Y_i - \hat{m}_i(X_i;h)\}^2 \tag{1.6}$$

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

$$\tilde{h} = Arg\min_h \hat{\delta}^2(h)$$

La validation croisée

La fonction validation croisée est défnie par la quantité :

$$CV(h) = \frac{1}{n}\sum_{i=1}^{n}\{Y_i - \hat{m}_{-i}(X_i;h)\}^2 \tag{1.7}$$

La seule diférence avec le critère précédent réside dans l'utilisation de l'indice $\hat{m}_{-i}$ Cet indice signife que pour chaque $i = 1; ...; n$; la valeur de $m(x_i)$ est obtenue en enlevant. la i^{eme} 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 $\hat{m}_{-i}$.

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

1.3.3 Méthode des séries orthogonales

Supposons que la fonction de régression peut être représentée comme une série de Fourier[Szego] [28] :

$$m(x) = \sum_{j=0}^{\infty} \beta_j \varphi_j(x), \tag{1.8}$$

où

$\{\varphi_j\}_{j=0}^{\infty}$:est une base connue de fonctions.

$\{\beta_j\}_{j=0}^{\infty}$: sont des coefficients de Fourier inconnus .

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 abordée 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 (1.8). Alors, étant donnée 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$\{y_j\}_{j=1}^{n}$ sont prises en des points $\{x_j\}_{j=1}^{n}$ équidistants sur cet intervalle.

Supposons que le système de fonctions $\{\varphi_j\}$ constituent une base orthonormale sur $[-1; 1]$ tel que :

$$\int_{-1}^{1} \varphi_k(x)\varphi_j(x)dx = \delta_{jk} = \begin{cases} 1 & si \quad j = k \\ 0 & sinon, \end{cases}$$

Alors, les coefficients de Fourier β_j peuvent être calculés par

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

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

$$x_i \in A_i, \qquad i = 1, \ldots, n$$

Maintenant, la formule des coefficients de Fourier peut être écrite comme suit :

$$\begin{aligned} \beta_j &= \sum_{i=1}^{n} \int_{A_i} m(x)\varphi_j(x)dx \\ &\simeq \sum_{i=1}^{n} m(x) \int_{A_i} \varphi_j(x)dx \end{aligned}$$

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 :

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

Si de plus, on choisit les A_i de la même largeur ,alors $\tilde{\beta}_j$ peut être s'approximer par :

$$\hat{\beta}_j = \frac{1}{n} \sum_{i=1}^{n} y_i(x)\varphi_j(x)dx \tag{1.9}$$

Puisque seulement un nombre d'observations est disponible, les coefficients de Fourier ne peuvent pas être estimés tous a la fois. Si $N(n)$ termes dans la représentation (1.8) sont considérées, la fonction de régression est approximée par :

$$\hat{m}_N(x) = \sum_{j=0}^{N(n)} \hat{\beta}_j\varphi_j(x), \tag{1.10}$$

Le choix de $N = N(n)$ reste un entier.Une manière de choisire N pourrait être en minimisant l'erreur quadratique moyenne.Cela présuppose,bien entendu que l'information majeurs contenue dans m est contenue dans les n premiers termes de la série [29] .

Définition 1.3.1 *l'estimateur de m par la méthode des séries orthogonales est :*

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

avec

$$\hat{\beta}_j = \frac{1}{n}\sum_{k=0}^{n} y_i(x)\varphi_j(x)dx$$

Les propriétés statistiques sont défini dans [Cenzov][4], [Wahba][7], [Walter [1977][30].

1.3.4 Méthode des fonctions splines

Les splines est très utilisée dans le domaine de l'analyse numérique, 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 du spline est de diviser l'intervalle $[a; b]$ où la fonction de spline est définie, en plusieurs sous intervalles.$[a, t_1]; [t_1, t_2]; \ldots; [t_k, b]$. Les points $t_1; \ldots; t_k$ sont appelés les noeuds.

Splines de lissage

Les splines de lissage sont une façon d'utiliser les fonctions splines pour estimer la fonction de régression du modèle (1.2). 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, ce qui donne :

$$S(m) = \frac{1}{n}\sum_{i=1}^{n}(y_i - m(x_i))^2 + \lambda \int (m^r(x))^2 dt \tag{1.11}$$

où λ est le paramètre de lissage prenant ses valeurs dans $[0; \infty)$ et où r est fixé et sert a définir le degré des polynômes ajustés . La valeur du paramètre λ permet de déterminer la flexibilité de l'estimateur. Plus la valeur de λ est proche de 0, plus l'estimateur est flexible, car on diminue l'apport de la quantité de lissage dans le critère (1.11). Par contre, lorsque l'on augmente la valeur de λ on donne plus d'importance à la deuxième partie du critère (1.11), ce qui oblige l'intégrales à être plus petite et donc l'estimateur à être plus lisse.Voir Ebank[1999][13].
Généralement on utilise $r = 2$, ce qui permet d'obtenir des splines cubiques. L'estimateur de spline revient à résoudre le problème de minimisation suivant : Trouver la fonction m qui minimise

$$S(m) = \frac{1}{n}\sum_{i=1}^{n}(y_i - m(x_i))^2 + \lambda \int (m^{"}(x))^2 dt \tag{1.12}$$

Soit $Y = (y_1, \cdots, y_n)^t$, $M = (M_1, \cdots, M_n)^t$ et $M_i = \tilde{m}_\lambda(x_i)$. donc on a :

$\frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{m}_\lambda(x_i))^2 = (Y - M)^t(Y - M)$ et $\int (m^{"}(x))^2 dt = M^t K M$

Définition 1.3.2 *Cao[31]*

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

$$\tilde{m}_\lambda(x_i) = (I + \lambda K)^{-1} Y = A_\lambda Y \tag{1.13}$$

Choix du paramètre de lissage

Méthode de la validation croisée

Soit $(A_\lambda)_{ii}$ le i^{me} élément de la diagonale associée à la matrice de lissage A_λ. Le score de la validation croisée vérifie :

$$CV(\lambda) = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{y_i - \tilde{m}_\lambda(x_i)}{1 - (A_\lambda)_{ii}} \right)^2 \tag{1.14}$$

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

Méthode de la validation croisée généralisée (GCV)

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

$$CV(\lambda) = \frac{1}{n} \frac{\sum_{i=1}^{n} (y_i - \tilde{m}_\lambda(x_i))^2}{(1 - \frac{1}{n} tr(A_\lambda))^2} \tag{1.15}$$

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

1.4 Conclusion

Dans ce chapitre, nous avons rappelé quelques méthodes d'estimation de la fonction de régression de la moyenne à savoir la méthode de régressogramme, la méthode des séries orthogonales , la méthode des séries splines et particulièrement la méthode du noyau. Nousallons présenter dans le chapitre suivant la méthode d'estimation de la fonction de régression par noyau associé discret univariée. Nous donnerons l'estimateur obtenu par cette méthode, ses principales propriétés statistiques et nous citons les différents noyaux discretes.

2

Estimation de la fonction de régression par noyau discret univarié

2.1 Introduction

Considérons un échantillon formé d'une suite de n couples (x_i, y_j), $i = \{1, \ldots, n\}$, à valeurs dans $\mathbb{N}\times\mathbb{R}$. Sans aucune spécification de distribution sur les y_i et en absence d'une forme de relation évidente entre y_i et x_i, le modèle classique de régression non paramétrique de y_i sur x_i s'impose à travers :

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

où y_i est une réalisation de la variable aléatoire réelle (v.a.r.) Y_i , x_i est une réalisation d'une variable explicative de dénombrement X_i, ϵ_i correspond aux résidus tels que $E(\epsilon_i) = 0$ et $Var(\epsilon_i) = \sigma^2$, $m : \mathbb{N} \to \mathbb{R}$ est fonction discrète inconnue de régression.La fonction m peut être exprimée comme $m(x_i) = E(Y_i|X_i = x_i)$.
Dans ce chapitre, nous nous intéressons à l'estimation non paramétrique par noyaux associés discrets univaries de cette fonction de régression discrète m en prenant en compte sa structure discrète.

2.2 Définition de l'estimateur de la fonction de régression

Avant de définir l'estimateur, nous donnons la définition d'un noyau discret et un noyau associé.

Définition 2.2.1 *On appelle type de noyau K_θ discret toute fonction de masse de probabilité paramétrée par $\theta \in \Theta \in R^2$, de support $\aleph_\theta \subseteq \mathbb{Z}$ et de carré sommable [32].*

Définition 2.2.2 *Soit $x \in \aleph$ et $h > 0$. On appelle noyau associé $K_{x,h}$ toute fonction de masse de probabilité(fmp) à une variable aléatoire discrète de $\mathcal{K}_{x,h}$ de support $\aleph_x$ contenant au moins x et indépendant de h, vérifiant les quatres conditions suivantes [16] :*

$$\bigcup_x \aleph_x \supseteq \aleph \tag{2.2}$$

$$\lim_{h\to 0} E\{\mathcal{K}_{x,h}\} = x \tag{2.3}$$

$$Var\{\mathcal{K}_{x,h}\} < \infty \tag{2.4}$$

$$\lim_{h\to 0} Var\{\mathcal{K}_{x,h}\} = 0 \tag{2.5}$$

Définition 2.2.3 *(l'estimateur) On considère le modèle non-paramétrique de régression discrète défini en (2.1). L'analogue discret de l'estimateur de Nadaraya [1964] [1] et Watson [1964][2] pour la fonction discrète inconnue m de (2.1) est défini par*

$$\hat{m}_n(x) = \sum_{i=1}^{n} \omega_x(X_i) Y_i, x \in \mathbb{N} \tag{2.6}$$

où

$$\omega_x(X_i) = \frac{K_{x,h}(X_i)}{\sum_{i=1}^{n} K_{x,h}(X_i)} = \omega_{x,h}(X_i) \tag{2.7}$$

représente la fonction des poids telle que $\sum_{i=1}^{n} \omega_{x,h}(X_i) = 1$ en convenant $(0/0) = 0$. $K_{x,h}(\cdot)$ est un noyau (associé) discret. La fenêtre $h \equiv h(n, K)$ a pour rôle de déterminer le lissage discret de l'estimation.

2.3 propriétés de l'estimateur

Pour l'étude du biais et de la variance de l'estimateur $\hat{m}_n(x) = \hat{m}_n(x,h)$ de m , il est commode d'écire $\hat{m}_n(x)$ comme le rapport

$$\hat{m}_n(x) = \frac{N_n(x;h)}{D_n(x;h)}$$

avec
$D_n(x,h) = \frac{1}{n}\sum_{i=1}^{n} K_{x,h}(X_i)$ et $N_n(x;h) = \frac{1}{n}\sum_{i=1}^{n} Y_i K_{x,h}(X_i)$
Daprés Sengua Kiessé[2008][16] ,

$$E\{\hat{m}_n(x)\} = \frac{E\{N_n(x;h)\}}{E\{D_n(x;h)\}}, \qquad Var\{\hat{m}_n(x)\} = \frac{Var\{N_n(x;h)\}}{E^2\{D_n(x;h)\}}$$

• **Biais**

On a $E\{D_n(x;h)\} = E\{f(\mathcal{K}_{x,h}\}$ tel que f (f.m.p). En utulisent un dévloppement limité de taylor de $f(\mathcal{K}_{x,h})$ au point moyen $E\{\mathcal{K}_{x,h}\} = m_{x,h}$ on obtient

$$f(\mathcal{K}_{x,h}) = f(m_{x,h}) + (\mathcal{K}_{x,h} - m_{x,h})f^{(1)}(m_{x,h}) + \frac{1}{2}(\mathcal{K}_{x,h} - m_{x,h})^2 f^{(2)}(m_{x,h})$$

par equivalant asymptotique on aura

$$\begin{aligned} E\{f(\mathcal{K}_{x,h})\} &= f(m_{x,h}) + \frac{1}{2}E\{(\mathcal{K}_{x,h} - m_{x,h})^2\}f^{(2)}(x) + o(h) \\ &= f\{E(\mathcal{K}_{x,h})\} + \frac{Var(\mathcal{K}_{x,h})}{2}f^{(2)}(x) + o(h) \\ &= f\{E(\mathcal{K}_{x,h})\} + \frac{Var(\mathcal{K}_{x,h})}{2}f^{(2)}(x) + o(h) \\ &= E\{D_n(x;h)\} \end{aligned}$$

En ce qui concerne le numérateur $N_n(x;h)$

$$\begin{aligned} E\{N_n(x;h)\} &= E\{Y_1 K_{x,h}(X_1)\} \\ &= \sum_{z\in\aleph_x} m(z)f(z)Pr(\mathcal{K}_{x,h} = z) \\ &= E\{(mf)\{\mathcal{K}_{x,h}\}\} \\ &= (mf)\{E(\mathcal{K}_{x,h})\} + \{Var(\mathcal{K}_{x,h})/2\}(mf)^{(2)}(x) + o(h) \end{aligned}$$

avec $(mf)^2 = m^2 f + 2m^1 f^1 + mf^2$.
Les différences finies d'une (f.m.p) g sont telles que, pour $k \in \mathbb{N}\backslash\{0\}$, on ait :
$g^{(k)}(x) = \{g^{(k-1)}(x)\}^{(1)}$ et

$$g^{(1)}(x) = \begin{cases} \{g(x+1) - g(x-1)\}/2 & si \quad x \in \mathbb{N}\backslash\{0\} \\ g(1) - g(0) & si \quad x = 0 \end{cases} \tag{2.8}$$

De la, la différence finie d'ordre 2 est

$$g^{(2)}(x) = \begin{cases} \{g(x+2) - 2g(x) + g(x-2)\}/4 & si \quad x \in \mathbb{N}\backslash\{0,1\} \\ \{g(3) - 3g(1) + g(0)\}/4 & si \quad x = 1 \\ \{g(2) - 2g(1) + g(0)\}/2 & si \quad x = 0 \end{cases} \tag{2.9}$$

L'expression de biais est :

$$\begin{aligned} \textbf{biais}\{\hat{m}_n(x)\} &= E\{\hat{m}_n(x)\} - m(x) \\ &= \frac{(mf)\{E(\mathcal{K}_{x,h})\} + \{Var(\mathcal{K}_{x,h})/2\}(mf)^{(2)}(x)}{f\{E(\mathcal{K}_{x,h})\} + \{Var(\mathcal{K}_{x,h})/2\}f^{(2)}(x)} - m(x) \end{aligned}$$

•Variance

A l'aide de l'espérance conditionnelle de Y^2 sachant X la variance $N_n(x;h)$ s'exprime :

$$\begin{aligned} Var\{N_n(x;h)\} &= \frac{1}{n}E\{Y_1^2 K_{x,h}^2(X_1)\} - \frac{1}{n}E^2\{Y_1 K_{x,h}(X_1)\} \\ &= \frac{1}{n}\sum_{y\in\aleph_x} E(Y_1^2|X_1 = y)f(y)\{Pr(\mathcal{K}_{x,h} = y)\}^2 \\ &- \frac{1}{n}\left\{\sum_{z\in\aleph_x} E(Y_1|X_1 = z)f(z)Pr(\mathcal{K}_{x,h} = z)\right\}^2 \\ &= \frac{1}{n}E(Y_1^2|X_1 = x)f(x)\{Pr(\mathcal{K}_{x,h} = x)\}^2 \\ &- \frac{1}{n}\{E(Y_1|X_1 = x)f(x)Pr(\mathcal{K}_{x,h} = x)\}^2 + r_n(x;h) \\ &= \frac{1}{n}[\{E(Y_1^2|X_1 = x) - f(x)E^2(Y_1|X_1 = x)\}f(x)\{Pr(\mathcal{K}_{x,h} = x)\}^2] \\ &+ r_n(x;h) \end{aligned}$$

avec

$$\begin{aligned} r_n(x;h) &= \frac{1}{n}\sum_{y\in\aleph_x|\{x\}} E(Y_1^2|X_1 = y)f(y)\{Pr(\mathcal{K}_{x,h} = y)\}^2 \\ &+ \frac{1}{n}\{E(Y_1|X_1 = x)f(x)Pr(\mathcal{K}_{x,h} = x)\}^2 \\ &- \frac{1}{n}\left\{\sum_{z\in\aleph_x} E(Y_1|X_1 = z)f(z)Pr(\mathcal{K}_{x,h} = z)\right\}^2 \end{aligned}$$

Donc la variance de $\hat{m}_n(x)$ s'ècrit comme suite :

$$\begin{aligned}\mathbf{Var}\{\hat{m}_n(x)\} &= Var\left\{\frac{N_n(x;h)}{D_n(x;h)}\right\}\\ &= \frac{Var\{N_n(x;h)\}}{E^2\{D_n(x;h)\}} + O\left(\frac{1}{n}\right)\\ &= \frac{1}{n}\times\frac{\{E(Y_1^2|X_1=x) - f(x)E^2(Y_1|X_1=x)\}f(x)}{[f\{E(\mathcal{K}_{x,h}\}\{Var(\mathcal{K}_{x,h})/2\}f^{(2)}(x)]^2}\{Pr(\mathcal{K}_{x,h}=x)\}^2\\ &+ \frac{r_n(x;h)}{E^2\{D_n(x;h)\}} + O\left(\frac{1}{n}\right)\end{aligned}$$

- **Risque global :**

On peut étudier le risque global à travers l'erreur quadratique moyenne intégrée

$$\mathbf{MISE}(h) = \sum_{x\in\aleph} Var\{\hat{m}_n(x;h)\} + \sum_{x\in\aleph} biais^2\{\hat{m}_n(x;h)\}$$

Risque asymptotique ponctuel

Proposition 2.3.1 *Selon Senga Kiessé[2008][16], Pour $x \in \mathbb{N}$, soient $m(x) = E(Y|X=x)$ et $f(x) = Pr(X=x)$ définies de $\mathbb{N}\to\mathbb{R}$. Soit $\hat{m}_n(x)$ l'estimateur de $m(x)$ à noyau associé discret $K_{x,h}$. Quand $n\to+\infty$ et $h = h(n)\to 0$, en tout point x où $f(x)\neq 0$ on a les développements asymptotiques*

$$\begin{aligned}\boldsymbol{biais}\{\hat{m}_n(x)\} &= E\{\hat{m}_n(x)\} - m(x)\\ &= \left\{ m^{(2)}(x) + 2m^{(1)}(x)\left(\frac{f^{(1)}}{f}\right)(x)\right\}\frac{Var(\mathcal{K}_{x,h})}{2} + O(1/n)^2 + o(h)\end{aligned}$$

$$\boldsymbol{Var}\{\hat{m}_n(x)\} = \frac{E(Y_1^2|X_1=x) - f(x)E^2(Y_1|X_1=x)}{nf(x)}\{Pr(\mathcal{K}_{x,h}=x)\}^2 + o\left(\frac{1}{n}\right)$$

Par conséquent, l'erreur quadratique moyenne s'écrit

$$\begin{aligned}\boldsymbol{MSE(x)} &= \left\{ m^{(2)}(x) + 2m^{(1)}(x)\left(\frac{f^{(1)}}{f}\right)(x)\right\}^2\frac{Var^2(\mathcal{K}_{x,h})}{4}\\ &+ \frac{E(Y_1^2|X_1=x) - f(x)E^2(Y_1|X_1=x)}{nf(x)}\{Pr(\mathcal{K}_{x,h}=x)\}^2 + o\left(h^2 + \frac{1}{n}\right)\end{aligned}$$

2.4 Exemples de noyaux associés univariés discrets

2.4.1 Noyaux associés discrets

Nous présentons dans cette section trois noyaux associés discrets :

Noyau Dirac Uniforme Discret

La version unidimensionnelle du noyau Dirac Uniforme Discret iroduit par Aitchison and Aitken [1976] [14] et Racine and Li [2004][33] est donné comme suit :

$$K_{x,h}(y) = DirDu_{x,h,c}(y) = (1-h)1_{y=x} + \left(\frac{h}{c-1}\right)1_{y\neq x}$$

où 1_A est la fonction indicatrice de A. dont le support est $\aleph_{x,c} = \{0, 1, ..., c-1\}$, $c \in \{2, 3, ...\}$ fixé et $h \in (0, 1]$.
avec $E(\mathcal{K}_{x,h}) = x + h\left(1 - x - \frac{x}{c-1} + \frac{hc}{2}\right)$
et $Var(\mathcal{K}_{x,h}) = -\left(\frac{c^2(-2x+c-1)^2}{4(c-1)^2}\right)h\left(\frac{c(6x^2+2c^2-3c+1-6xc+6x)}{6(c-1)}\right)h^2$.

Wang-VanRyzin

Une extension du noyau discret de Aitchison and Aitken [1976] dans $\mathbb{Z}$ a été proposé par Wang and Ryzin [1981][34]. Le noyau de Wang-VanRyzin est definit dans le support $\mathbb{S}_x = \mathbb{Z}$, pour $x \in \mathbb{T}_1 = \mathbb{Z}$ et $h \in (0, 1]$, comme suit :

$$K_{x,h}(y) = (1-h)\mathbf{1}_{y=x} + \frac{1}{2}(1-h)h^{|y-x|}\mathbf{1}_{|y-x|\geq 1}, y \in \mathbb{S}_x. \tag{2.10}$$

avec $E(\mathcal{K}_{x,h}) = x$ et $Var(\mathcal{K}_{x,h}) = h(1+h)/(1-h)^2$.

Noyau Optimal

Pour $x \in \mathbb{S} = \mathbb{Z}$, $h \in (0, 1]$ et $k \in \mathbb{N}$, Aitchison and Aitken[1981] proposé un noyau discret exprimé par[18] :

$$K_{k,x,h}(z) = \begin{cases} 1-h & si \quad x = z \\ 1/2(h/k) & si \quad |x-z| = 1...k \\ 0 & si \quad |x-z| \geq k+1 \end{cases} \quad y \in \mathbb{S}_x = \mathbb{Z}$$

avec $E(\mathcal{K}_{x,h}) = x$ et $Var(\mathcal{K}_{x,h}) = h(k+1)(2k+1)/6$.

2.4.2 Noyaux associés discrets standards

Nous présentons dans cette section la deuxième classe des noyaux associés discrets, dite classe des noyaux discrets standards ou de premier ordre .Ce type de noyaux ne vérifient pas la condition (2.5). Voir Sengua Kiessé[2008][16].

Noyau Poissonien

Pour une loi de Poisson $P(\lambda)$, on considère le noyau discret $P_{x,h}$ de loi $P(x+h)$ sur $\aleph_x = \mathbb{N}$ avec $x \in \mathbb{N}$ et $h > 0$. Le noyau de Poisson est donné par

$$P_{x,h}(y) = \frac{(x+h)^y \exp^{-(x+h)}}{y!}, y \in \mathbb{N}. \tag{2.11}$$

Notons que les noyau de Poisson est dit équi-dispersé de moyenne $x+h$ égale à la variance $x+h$ .la condition (2.5) n'est pas vérifiée ici car la variance tend vers x quand h tend vers 0.

Noyau Binomial

On considére une loi binomial $B(n,p)$, on lui associe le noyau $B_{x,h}$ de loi $B(x+1, x+h/x+1)$ sur $\aleph_x = \{0,1,...,x+1\}$ pour $x \in \mathbb{N}$ et $h \in]0,1]$ avec $\bigcup_x \aleph_x = \mathbb{N}$. Le noyau binomial est donné comme suit

$$B_{x,h} = \frac{(x+1)!}{y!(x+1-y)!} \left(\frac{x+h}{x+1} \right)^y \left(\frac{1-h}{x+1} \right)^{x+1-y}, y \in \aleph_x \subseteq \mathbb{N}. \tag{2.12}$$

Le noyau discret binomial $B_{x,h}$ du support $\aleph_x = \{0,1,...,x+1\}$ (dependant de x), est dit sous-dispersé de moyenne $x+h$ et de variance $(x+h)(1-h)/(x+1)$. La condition (2.5) n'est pas vérifiée, car quand $h \longrightarrow 0$ la variance tend vers $x/(x+1) < 1$.

Noyau Binomial négatif

Dans le cas de la loi binomial négative $\mathcal{BN}(\lambda, p)$, on considère le noyau discret $\mathcal{BN}_{x,h}$ de loi $\mathcal{BN}_{x,h}(x+1, (x+1)/2x+h+1)$ sur $\aleph_x = \mathbb{N}$ pour $x \in N$ et $h > 0$ .La forme du noyau binomial négatif est

$$\mathcal{BN}_{x,h} = \frac{(x+y)!}{x!y!} \left(\frac{x+h}{2x+h+1} \right)^y \left(\frac{x+1}{2x+h+1} \right)^{x+1}, y \in \aleph \subseteq \mathbb{N}. \tag{2.13}$$

Le noyau discret binomial négatif $\mathcal{BN}_{x,h}$ est dit sur-dispersé de moyenne $x+h$ et de variance $(x+h)(1+(x+h)/(x+1))$. La condition (2.5) n'est pas vérifiée car lorsque h tend vers 0,la variance tend vers $x(2x+1)/(x+1) \neq 0$.

2.4.3 Noyaux associés discrets de deuxième ordre

Nous présentons ici la troisième classe des noyaux associés discrets, dite classe des noyaux associés discrets de deuxième ordre, c'est à dire, les condition (2.2) (2.5) sont vérifiées.

Noyau associé discret Triangulaire

Les noyaux associés discrets triangulaires ont été proposés par Kokonendji et al[2007b][15] et Kokonendji and Zocchi [2010] [35]. Nous donnons d'abord la définition d'une variable aléatoire Triangulaire symétrique avant de donner la définition d'un noyau associé discret Triangulaire.

Définition 2.4.1 *Une variable aléatoire discrète $\mathcal{T}_{a,c}$ est dite triangulaire symétrique de centre $c \in \mathbb{N}$ et de bras $a \in \mathbb{N}$, si pour y dans son support $\aleph_c = \{c, c \pm 1, ...c \pm a\}$ sa probabilité individuelle sécrit*

$$Pr(\mathcal{T}_{a,c} = y) = \frac{a + 1 - \mid y - c \mid}{(a+1)^2}$$

Définition 2.4.2 *Soit $h > 0$ et $(a, c) \in \mathbb{N}^2$. Une variable aléatoire discrète $\mathcal{T}_{a;c,h}$ est dite triangulaire d'ordre h, de centre $c \in N$ et de bras $a \in \mathbb{N}$, si pour y dans son support $\aleph_c = \{c, c \pm 1, ..., c \pm a\}$ sa fonction de masse de probabilité sécrit*

$$Pr(\mathcal{T}_{a;c,h} = y) = \frac{(a+1)^h - \mid y - c \mid^h}{P(a,h)}$$

où
$P(a,h) = (2a+1)(a+1)^h - 2\sum_{k=0}^{a} k^h$
est la constante de normalisation.

Définition 2.4.3 *Soit f une fonction de masse de probabilité sur $\aleph$ Soit $h > 0$ le paramètre de lissage et $a \in \mathbb{N}$ un entier fixé. Le noyau discret triangulaire $\mathbf{T}_{a;x,h}$ associée à la variable aléatoire $\mathcal{T}_{a;x,h}$ d'ordre h, de centre x et de bras a définie sur $\aleph_x = \{x, x \pm 1, ...x \pm a\}$ est donné par*

$$Pr(\mathcal{T}_{a;X,h} = y) = \frac{(a+1)^h - \mid y - x \mid^h}{(2a+1)(a+1)^h - 2\sum_{k=0}^{a} k^h}, \forall y \in \aleph_x$$

Le noyau triangulaire discret vérifie les conditions (2.2) et (2.5) d'un noyau associée discret. avec $E(\mathcal{T}_{a;x,h}) = x$ et $Var(\mathcal{T}_{a;x,h}) \simeq [\{a(2a^2 + 3a + 1)/3\} log(a+1) - 2\sum_{k=1}^{n} k^2 log(k)]h + o(h^2)$.

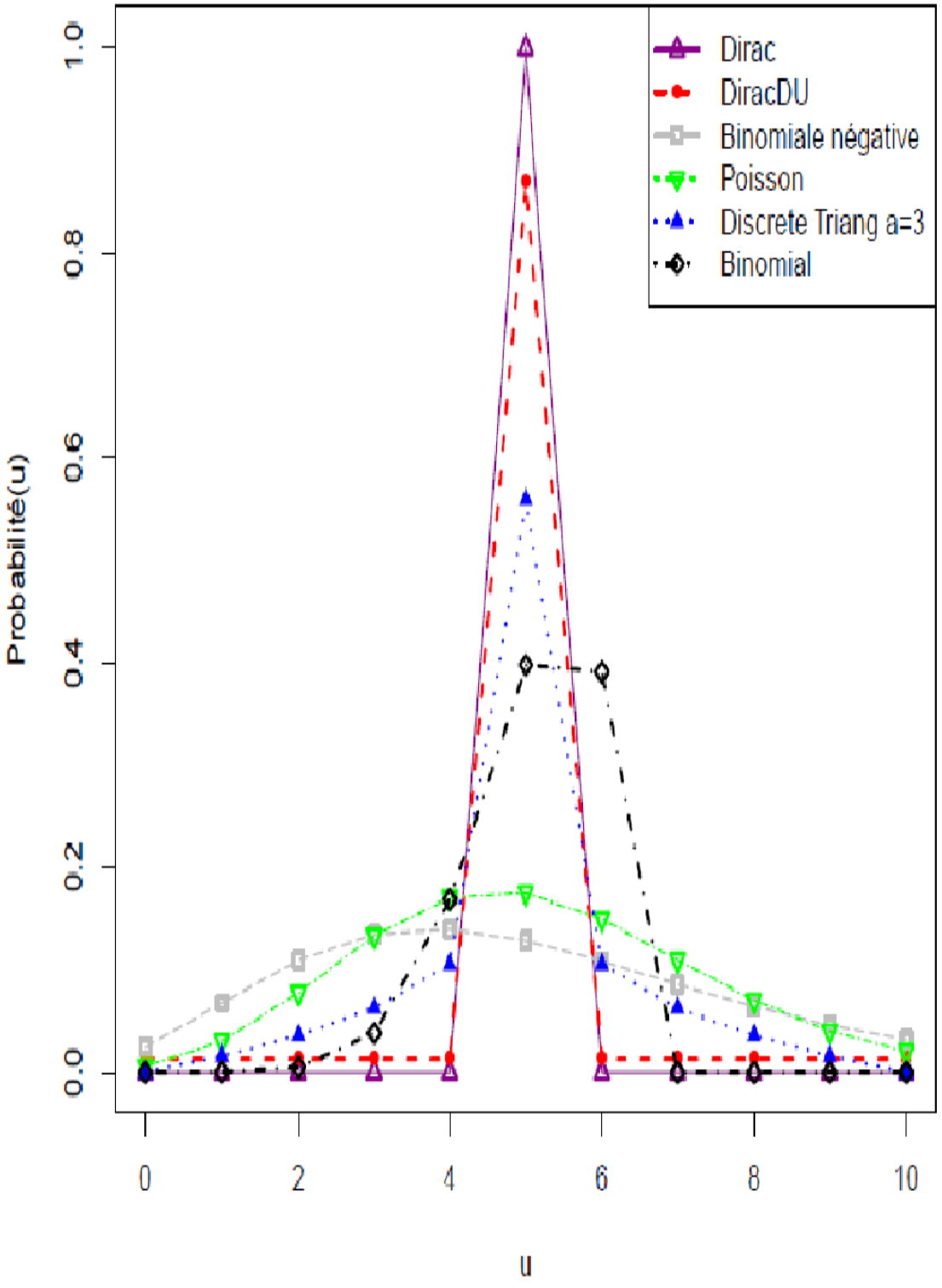


FIGURE 2.1 – Formes de quelques noyaux associés univariés discrets : Dirac, DiracDU, binomiale négative, Poisson, triangulaire discret $a = 3$ et binomial de même cible $x = 5$ et fenêtre de lissage $h = 0,13$.

2.5 Les estimateurs de régression associés aux noyaux discrets

a)Noyau Dirac Uniforme Discret

L'estimateur non-paramétrique de régression associé au noyau Dirac Uniforme Discret est :

$$\begin{aligned}\hat{m}_n^D(x) &= \frac{\sum_{i=1}^n Y_i K_{x,h}(X_i)}{\sum_{i=1}^n K_{x,h}(X_i)} \\ &= \frac{\sum_{i=1}^n Y_i(1-h)1_{X_i=x} + \left(\frac{h}{c-1}\right)1_{X_i\neq x}}{\sum_{i=1}^n (1-h)1_{X_i=x} + \left(\frac{h}{c-1}\right)1_{X_i\neq x}}\end{aligned}$$

Nous rappellons que $E(\mathcal{K}_{x,h}) = x + h\left(1 - x - \frac{x}{c-1} + \frac{hc}{2}\right)$,
$Var(\mathcal{K}_{x,h}) = -\left(\frac{c^2(-2x+c-1)^2}{4(c-1)^2}\right)h\left(\frac{c(6x^2+2c^2-3c+1-6xc+6x)}{6(c-1)}\right)h^2$ et $E\{\hat{m}_n^D(x)\} = \frac{E\{N_n(x;h)\}}{E\{D_n(x;h)\}}$.
On a :

$$\begin{aligned}E\{D_n(x;h)\} &= f\{E(\mathcal{K}_{x,h})\} + \frac{Var(\mathcal{K}_{x,h})}{2}f^{(2)}(x) \\ &= f\left\{x + h\left(1 - x - \frac{x}{c-1} + \frac{hc}{2}\right)\right\} + \frac{Var(\mathcal{K}_{x,h})}{2}f^{(2)}(x)\end{aligned}$$

On utilise un dévloppement limité de $f\left\{x + h\left(1 - x - \frac{x}{c-1} + \frac{hc}{2}\right)\right\}$ au point $x_0 = x$ on aura

$$E\{D_n(x;h)\} = f(x) + h\left(1 - x - \frac{x}{c-1} + \frac{hc}{2}\right)f^{(1)}(x) + \frac{Var(\mathcal{K}_{x,h})}{2}f^{(2)}(x)$$

et pour le numérateur $N_n(x;h)$ nous avons

$$\begin{aligned}E\{N_n(x;h)\} &= (mf)\{E(\mathcal{K}_{x,h})\} + \frac{Var(\mathcal{K}_{x,h})}{2}(mf)^{(2)}(x) \\ &= (mf)\left\{x + h\left(1 - x - \frac{x}{c-1} + \frac{hc}{2}\right)\right\} + \frac{Var(\mathcal{K}_{x,h})}{2}(mf)^{(2)}(x)\end{aligned}$$

On utilise un dévloppement limité de $(mf)\left\{x + h\left(1 - x - \frac{x}{c-1} + \frac{hc}{2}\right)\right\}$ au point $x_0 = x$ on aura

$$E\{N_n(x;h)\} = (mf)(x) + h\left(1 - x - \frac{x}{c-1} + \frac{hc}{2}\right)(mf)^{(1)}(x) + \frac{Var(\mathcal{K}_{x,h})}{2}(mf)^{(2)}(x)$$

Considerons l'approximation $E\{D_n(x;h)\} = f(x)$. le biais devient alors :

Biais

$$
\begin{aligned}
biais\{\hat{m}_n^D(x)\} &= E\{\hat{m}_n^D(x)\} - m(x) \\
&= h\left(1 - x - \frac{x}{c-1} + \frac{hc}{2}\right)\frac{(mf)^{(1)}(x)}{f(x)} \\
&+ \frac{1}{2}\left(-\left(\frac{c^2(-2x+c-1)^2}{4(c-1)^2}\right)h\left(\frac{c(6x^2+2c^2-3c+1-6xc+6x)}{6(c-1)}\right)h^2\right)\frac{(mf)^{(2)}(x)}{f(x)}
\end{aligned}
$$

Variance

La variance s'écrit :

$$
\begin{aligned}
Var\{\hat{m}_n^D(x)\} &= \frac{E(Y_1^2|X_1=x) - f(x)E^2(Y_1|X_1=x)}{nf(x)}\{(1-h)\}^2 \\
&+ \frac{r_n(x;h)}{E^2\{D_n(x,h)\}} + O\left(\frac{1}{n}\right)
\end{aligned}
$$

L'erreur quadratique moyenne

$$
\begin{aligned}
MSE\{\hat{m}_n^D(x)\} &= biais^2\{\hat{m}_n^D(x)\} + Var\{\hat{m}_n^D(x)\} \\
&= \left\{\left(1 - x - \frac{x}{c-1} + \frac{hc}{2}\right)\frac{(mf)^{(1)}(x)}{f(x)} + \left(\frac{Var(\mathcal{K}_{x,h})}{2}\right)\frac{(mf)^{(2)}(x)}{f(x)}\right\}^2 \\
&+ \frac{E(Y_1^2|X_1=x) - f(x)E^2(Y_1|X_1=x)}{nf(x)}\{(1-h)\}^2 \\
&+ \frac{r_n(x;h)}{E^2\{D_n(x,h)\}} + O\left(\frac{1}{n}\right)
\end{aligned}
$$

b)Noyau Wang-VanRyzin

L'estimateur non-paramétrique de régression associé au noyau Wang-VanRyzin est :

$$
\begin{aligned}
\hat{m}_n^W(x) &= \frac{\sum_{i=1}^n Y_i K_{x,h}(X_i)}{\sum_{i=1}^n K_{x,h}(X_i)} \\
&= \frac{\sum_{i=1}^n Y_i\{(1-h)\mathbf{1}_{X_i=x} + \left(\frac{1}{2}(1-h)h^{|X_i-x|}\right)\mathbf{1}_{|X_i-x|\geq 1}\}}{\sum_{i=1}^n (1-h)\mathbf{1}_{X_i=x} + \left(\frac{1}{2}(1-h)h^{|X_i-x|}\right)\mathbf{1}_{|X_i-x|\geq 1}}
\end{aligned}
$$

Nous rappellons que $E(\mathcal{K}_{x,h}) = x$ et $Var(\mathcal{K}_{x,h}) = h(1+h)/(1-h)^2$ donc on aura :

$$E\{D_n(x;h)\} = f(x) + \frac{Var(\mathcal{K}_{x,h})}{2} f^{(2)}(x)$$
$$E\{N_n(x;h)\} = (mf)(x) + \frac{Var(\mathcal{K}_{x,h})}{2} (mf)^{(2)}(x)$$

Considerons l'approximation $E\{D_n(x;h)\} = f(x)$, le biais devient alors :

Biais

$$\begin{aligned} biais\{\hat{m}_n^W(x)\} &= E\{\hat{m}_n^W(x)\} - m(x) \\ &= \frac{1}{2}\left(\frac{h(1+h)}{(1-h)^2}\right)\left(\frac{(mf)^{(2)}(x)}{f(x)}\right) \end{aligned}$$

Variance

La variance s'écrit :

$$\begin{aligned} Var\{\hat{m}_n^W(x)\} = &\frac{E(Y_1^2|X_1=x) - f(x)E^2(Y_1|X_1=x)}{nf(x)} \{1-h\}^2 \\ &+ \frac{r_n(x;h)}{E^2\{D_n(x,h)\}} + O\left(\frac{1}{n}\right) \end{aligned}$$

L'erreur quadratique moyenne

$$\begin{aligned} MSE\{\hat{m}_n^W(x)\} &= biais^2\{\hat{m}_n^W(x)\} + Var\{\hat{m}_n^W(x)\} \\ &= \left\{\frac{1}{2}\left(\frac{h(1+h)}{(1-h)^2}\right)\left(\frac{(mf)^{(2)}(x)}{f(x)}\right)\right\}^2 \\ &+ \frac{E(Y_1^2|X_1=x) - f(x)E^2(Y_1|X_1=x)}{nf(x)} \{1-h\}^2 \\ &+ \frac{r_n(x;h)}{E^2\{D_n(x,h)\}} + O\left(\frac{1}{n}\right) \end{aligned}$$

c)Noyau Optimal

L'estimateur non-paramétrique de régression associé au noyau Optimal est :

$$\begin{aligned} \hat{m}_n^O(x) &= \frac{\sum_{i=1}^n Y_i K_{x,h}(X_i)}{\sum_{i=1}^n K_{x,h}(X_i)} \\ &= \frac{\sum_{i=1}^n Y_i\{(1-h)\mathbf{1}_{X_i=x} + \frac{1}{2}\left(\frac{h}{k}\right)\mathbf{1}_{|X_i-x|=1\cdots k}\}}{\sum_{i=1}^n (1-h)\mathbf{1}_{X_i=x} + \frac{1}{2}\left(\frac{h}{k}\right)\mathbf{1}_{|X_i-x|=1\cdots k}} \end{aligned}$$

Nous rappellons que $E(\mathcal{K}_{x,h}) = x$ et $Var(\mathcal{K}_{x,h}) = h(k+1)(2k+1)/6$ donc on aura :

$$E\{D_n(x;h)\} = f(x) + \frac{Var(\mathcal{K}_{x,h})}{2} f^{(2)}(x)$$
$$E\{N_n(x;h)\} = (mf)(x) + \frac{Var(\mathcal{K}_{x,h})}{2} (mf)^{(2)}(x)$$

Considerons l'approximation $E\{D_n(x;h)\} = f(x)$, le biais devient alors :

Biais

$$\begin{aligned} biais\{\hat{m}_n^O(x)\} &= E\{\hat{m}_n^O(x)\} - m(x) \\ &= \frac{1}{2}\left(\frac{h(k+1)(2k+1)}{6}\right)\left(\frac{(mf)^{(2)}(x)}{f(x)}\right) \end{aligned}$$

Variance

La variance s'écrit :

$$\begin{aligned} Var\{\hat{m}_n^O(x)\} = &\frac{E(Y_1^2|X_1 = x) - f(x)E^2(Y_1|X_1 = x)}{nf(x)} \{1-h\}^2 \\ &+ \frac{r_n(x;h)}{E^2\{D_n(x,h)\}} + O\left(\frac{1}{n}\right) \end{aligned}$$

L'erreur quadratique moyenne

$$\begin{aligned} MSE\{\hat{m}_n^O(x)\} &= biais^2\{\hat{m}_n^O(x)\} + Var\{\hat{m}_n^O(x)\} \\ &= \left\{\frac{1}{2}\left(\frac{h(k+1)(2k+1)}{6}\right)\left(\frac{(mf)^{(2)}(x)}{f(x)}\right)\right\}^2 \\ &+ \frac{E(Y_1^2|X_1 = x) - f(x)E^2(Y_1|X_1 = x)}{nf(x)} \{1-h\}^2 \\ &+ \frac{r_n(x;h)}{E^2\{D_n(x,h)\}} + O\left(\frac{1}{n}\right) \end{aligned}$$

d)Noyau binomial

L'estimateur non-paramétrique de régression associé au noyau binomial $B_{x,h}$ est :

$$\begin{aligned}\hat{m}_n^B(x) &= \frac{\sum_{i=1}^n Y_i B_{x,h}(X_i)}{\sum_{i=1}^n B_{x,h}(X_i)} \\ &= \frac{\sum_{i=1}^n Y_i \frac{(x+1)!}{X_i!(x+1-X_i)} \left(\frac{x+h}{x+1}\right)^{X_i} \left(\frac{1-h}{x+1}\right)^{x+1-X_i}}{\sum_{i=1}^n \frac{(x+1)!}{X_i!(x+1-X_i)} \left(\frac{x+h}{x+1}\right)^{X_i} \left(\frac{1-h}{x+1}\right)^{x+1-X_i}}\end{aligned}$$

Nous rappellons que $E(\mathcal{B}_{x,h}) = x + h$ et $Var(\mathcal{B}_{x,h}) = V^b(x,h) + o(h)$,avec $V^b(x,h) = \{(x+h) - xh\}/(x+1)$ donc on aura :

$$\begin{aligned}E\{D_n(x;h)\} &= f(x) + hf^{(1)}(x) + \frac{V^b(x,h)}{2} f^{(2)}(x) \\ E\{N_n(x;h)\} &= (mf)(x) + h(mf)^{(1)}(x) + \frac{V^b(x,h)}{2}(mf)^{(2)}(x)\end{aligned}$$

Considerons l'approximation $E\{D_n(x;h)\} = f(x)$. le biais devient alors :

Biais

$$\begin{aligned}biais\{\hat{m}_n^B(x)\} &= E\{\hat{m}_n^B(x)\} - m(x) \\ &= h\frac{(mf)^{(1)}(x)}{f(x)} + \frac{1}{2}\left(\frac{x+h-xh}{x+1}\right)\left(\frac{(mf)^{(2)}(x)}{f(x)}\right)\end{aligned}$$

Variance

La variance s'écrit :

$$\begin{aligned}Var\{\hat{m}_n(x)\} = &\frac{E(Y_1^2|X_1=x) - f(x)E^2(Y_1|X_1=x)}{nf(x)} \left\{(1-h)\left(\frac{x+h}{x+1}\right)^x\right\}^2 \\ &+ \frac{r_n(x;h)}{E^2\{D_n(x,h)\}} + O\left(\frac{1}{n}\right)\end{aligned}$$

L'erreur quadratique moyenne

$$\begin{aligned} MSE\{\hat{m}_n(x)\} &= biais^2\{\hat{m}_n(x)\} + Var\{\hat{m}_n(x)\} \\ &= \left\{ h\frac{(mf)^{(1)}(x)}{f(x)} + \frac{1}{2}\left(\frac{x+h-xh}{x+1}\right)\left(\frac{(mf)^{(2)}(x)}{f(x)}\right)\right\}^2 \\ &+ \frac{E(Y_1^2|X_1=x) - f(x)E^2(Y_1|X_1=x)}{nf(x)}\left\{(1-h)\left(\frac{x+h}{x+1}\right)^x\right\}^2 \\ &+ \frac{r_n(x;h)}{E^2\{D_n(x,h)\}} + O\left(\frac{1}{n}\right) \end{aligned}$$

e)Noyau binomial négatif

L'estimateur non-paramétrique de régression associé au noyau binomial négatif $BN_{x,h}$ est :

$$\begin{aligned} \hat{m}_n^{BN}(x) &= \frac{\sum_{i=1}^n Y_i BN_{x,h}(X_i)}{\sum_{i=1}^n BN_{x,h}(X_i)} \\ &= \frac{\sum_{i=1}^n Y_i \frac{(x+X_i)!}{x!X_i!}\left(\frac{x+h}{2x+h+1}\right)^{X_i}\left(\frac{x+1}{2x+h+1}\right)^{x+1}}{\sum_{i=1}^n \frac{(x+X_i)!}{x!X_i!}\left(\frac{x+h}{2x+h+1}\right)^{X_i}\left(\frac{x+1}{2x+h+1}\right)^{x+1}} \end{aligned}$$

Nous rappellons que $E(\mathcal{BN}_{x,h}) = x$ et $Var(\mathcal{BN}_{x,h}) = V^{bn}(x,h)+o(h)$,avec $V^{bn}(x,h) = (x+h)(1+\frac{x}{x+1}) + \frac{xh}{(x+1)}$ donc on aura :

$$\begin{aligned} E\{D_n(x;h)\} &= f(x) + hf^{(1)}(x) + \frac{V^{bn}(x,h)}{2}f^{(2)}(x) \\ E\{N_n(x;h)\} &= (mf)(x) + h(mf)^{(1)}(x) + \frac{V^{bn}(x,h)}{2}(mf)^{(2)}(x) \end{aligned}$$

Considerons l'approximation $E\{D_n(x;h)\} = f(x)$, le biais devient alors :

Biais

$$\begin{aligned} biais\{\hat{m}_n^{BN}(x)\} &= E\{\hat{m}_n^{BN}(x)\} - m(x) \\ &= \frac{h(mf)^{(1)}(x)}{f(x)} + \frac{1}{2}\left(\frac{(x+h)(2x+1)+xh}{x+1}\right)\left(\frac{(mf)^{(2)}(x)}{f(x)}\right) \end{aligned}$$

Variance

La variance s'écrit :

$$Var\{\hat{m}_n(x)\} = \frac{E(Y_1^2|X_1=x) - f(x)E^2(Y_1|X_1=x)}{nf(x)} \left\{ \frac{2x!}{(x!)^2} \left(\frac{x+h}{2x+h+1} \right)^x \left(\frac{x+1}{2x+h+1} \right)^{x+1} \right\}^2 + \frac{r_n(x;h)}{E^2\{D_n(x,h)\}} + O\left(\frac{1}{n}\right)$$

L'erreur quadratique moyenne

$$\begin{aligned} MSE\{\hat{m}_n(x)\} &= biais^2\{\hat{m}_n(x)\} + Var\{\hat{m}_n(x)\} \\ &= \left\{ \frac{h(mf)^{(1)}(x)}{f(x)} + \frac{1}{2} \left(\frac{(x+h)(2x+1)+xh}{x+1} \right) \left(\frac{(mf)^{(2)}(x)}{f(x)} \right) \right\}^2 \\ &+ \frac{E(Y_1^2|X_1=x) - f(x)E^2(Y_1|X_1=x)}{nf(x)} \left\{ \frac{2x!}{(x!)^2} \left(\frac{x+h}{2x+h+1} \right)^x \left(\frac{x+1}{2x+h+1} \right)^{x+1} \right\}^2 \\ &+ \frac{r_n(x;h)}{E^2\{D_n(x,h)\}} + O\left(\frac{1}{n}\right) \end{aligned}$$

f)Noyau de poisson

L'estimateur non-paramétrique de régression associé au noyau de poisson $P_{x,h}$ est :

$$\begin{aligned} \hat{m}_n^B(x) &= \frac{\sum_{i=1}^n Y_i P_{x,h}(X_i)}{\sum_{i=1}^n P_{x,h}(X_i)} \\ &= \frac{\sum_{i=1}^n Y_i \frac{(x+h)^{X_i} \exp^{-(x+h)}}{X_i!}}{\sum_{i=1}^n \frac{(x+h)^{X_i} \exp^{-(x+h)}}{X_i!}} \end{aligned}$$

Nous rappellons que $E(\mathcal{P}_{x,h}) = x+h$ et $Var(\mathcal{P}_{x,h}) = V^p(x,h) = x+h$ donc on aura :

$$\begin{aligned} E\{D_n(x;h)\} &= f(x) + hf^{(1)}(x) + \frac{V^p(x,h)}{2} f^{(2)}(x) \\ E\{N_n(x;h)\} &= (mf)(x) + h(mf)^{(1)}(x) + \frac{V^p(x,h)}{2} (mf)^{(2)}(x) \end{aligned}$$

Considerons l'approximation $E\{D_n(x;h)\} = f(x)$. le biais devient alors :

Biais

$$\begin{aligned} biais\{\hat{m}_n^P(x)\} &= E\{\hat{m}_n^P(x)\} - m(x) \\ &= h\frac{(mf)^{(1)}(x)}{f(x)} + \frac{1}{2}(x+h)\left(\frac{(mf)^{(2)}(x)}{f(x)}\right) \end{aligned}$$

Variance

La variance s'écrit :

$$\begin{aligned} Var\{\hat{m}_n^P(x)\} = &\frac{E(Y_1^2|X_1=x) - f(x)E^2(Y_1|X_1=x)}{nf(x)}\left\{\frac{(x+h)^x \exp^{-(x+h)}}{x!}\right\}^2 \\ &+ \frac{r_n(x;h)}{E^2\{D_n(x,h)\}} + O\left(\frac{1}{n}\right) \end{aligned}$$

L'erreur quadratique moyenne

$$\begin{aligned} MSE\{\hat{m}_n^P(x)\} &= biais^2\{\hat{m}_n^P(x)\} + Var\{\hat{m}_n^P(x)\} \\ &= \left\{h\frac{(mf)^{(1)}(x)}{f(x)} + \frac{1}{2}(x+h)\left(\frac{(mf)^{(2)}(x)}{f(x)}\right)\right\}^2 \\ &+ \frac{E(Y_1^2|X_1=x) - f(x)E^2(Y_1|X_1=x)}{nf(x)}\left\{\frac{(x+h)^x \exp^{-(x+h)}}{x!}\right\}^2 \\ &+ \frac{r_n(x;h)}{E^2\{D_n(x,h)\}} + O\left(\frac{1}{n}\right) \end{aligned}$$

g)Noyau Triangulaire

L'estimateur non-paramétrique de régression associé au noyau triangilaire $T_{x,h}$ est :

$$\begin{aligned} \hat{m}_n^T(x) &= \frac{\sum_{i=1}^n Y_i T_{a;x,h}(X_i)}{\sum_{i=1}^n T_{a;x,h}(X_i)} \\ &= \frac{\sum_{i=1}^n Y_i\{(a+1)^h - \mid X_i - x \mid^h\}}{\sum_{i=1}^n (a+1)^h - \mid X_i - x \mid^h} \end{aligned}$$

Nous rappellons que $E(\mathcal{T}_{a;x,h}) = x$ et $Var(\mathcal{T}_{a;x,h}) = V^T(a;x,h) + o(h^2)$ avec $V^T(a;x,h) = [\{a(2a^2+3a+1)/3\}log(a+1) - 2\sum_{k=1}^n k^2 log(k)]h$ donc on aura :

$$E\{D_n(x;h)\} = f(x) + \frac{V^T(a;x,h)}{2} f^{(2)}(x)$$
$$E\{N_n(x;h)\} = (mf)(x) + \frac{V^T(a;x,h)}{2} (mf)^{(2)}(x)$$

Considerons l'approximation $E\{D_n(x;h)\} = f(x)$. le biais devient alors :

Biais

$$\begin{aligned} biais\{\hat{m}_n^T(x)\} &= E\{\hat{m}_n^T(x)\} - m(x) \\ &= \left(\frac{V^T(a;x,h)}{2}\right)\left(\frac{(mf)^{(2)}(x)}{f(x)}\right) \end{aligned}$$

Variance

La variance s'écrit :

$$\begin{aligned} Var\{\hat{m}_n^T(x)\} = &\frac{E(Y_1^2|X_1=x) - f(x)E^2(Y_1|X_1=x)}{nf(x)} \left\{\frac{(a+1)^h}{P(a;h)}\right\}^2 \\ &+ \frac{r_n(x;h)}{E^2\{D_n(x,h)\}} + O\left(\frac{1}{n}\right) \end{aligned}$$

L'erreur quadratique moyenne

$$\begin{aligned} MSE\{\hat{m}_n^T(x)\} &= biais^2\{\hat{m}_n^T(x)\} + Var\{\hat{m}_n^T(x)\} \\ &= \left\{\left(\frac{V^T(a;x,h)}{2}\right)\left(\frac{(mf)^{(2)}(x)}{f(x)}\right)\right\}^2 \\ &+ \frac{E(Y_1^2|X_1=x) - f(x)E^2(Y_1|X_1=x)}{nf(x)} \left\{\frac{(a+1)^h}{P(a;h)}\right\}^2 \\ &+ \frac{r_n(x;h)}{E^2\{D_n(x,h)\}} + O\left(\frac{1}{n}\right) \end{aligned}$$

Remarque :

Le biais de l'estimateur non-paramétrique de régression associé aux noyaux discrets de première ordre (binomial, binomial négative et poisson) dépend des différences finies d'ordre 1 et 2 de la fonction produit mf. Tandis que pour les noyaux discrets triangulaires et les noyaux associés (dirac uniforme discret, Wang Van-Ryzan

et le noyau optimal), les biais dépendent uniquement de la différence finie du second ordre $(mf)^{(2)}$ de mf.

2.6 Choix du paramètre de lissage

La qualité de l'estimateur par noyau dépend crucialement du choix du paramètre de lissage. Dans le cas univariée, ceci revient à choisir un paramètre scalaire h strictement positif qui contrôle le degré de lissage.

2.6.1 La méthode de validation croisée

Cette approche a été adaptée par Kokonendji et al[2009][16].Pour obtenir le paramètre de lissage optimal dans l'estimation des fonctions discrètes de régression. Pour un noyau associée, le paramètre de lissage optimal est obtenu en minimisant le critère $CV(h)$

$$h_{cv} = \min_h CV(h)$$

avec

$$CV(h) = \frac{1}{n}\sum_{i=1}^{n}\{Y_i - \hat{m}_{-i}(X_i;h)\}^2\omega(X_i) \tag{2.14}$$

$\omega(X_i)$ est défini en (2.7).
$\hat{m}_{-i}(X_i;h) = \sum_{j\neq i}^{n} Y_j K_{X_i,h}(X_j)/\sum_{j\neq i}^{n} K_{X_i,h}(X_j)$ est l'estimateur de la fonction régression calculé à partir de toutes les observations sauf l'observation x_i [36].

2.7 Conclution

Dans ce chapitre, nous nous sommes intéressé au cas de l'estimation non-paramétrique de la fonction de régression par la méthode de noyau associe discret univarié.Ainsi nous avons rappelé leur propriétés (l'espérance, la variance, l'erreur quadratique). Ensuit, nous avons utilisé la méthode de validation croisée pour le choix du paramètre de lissage .
Dans le chapitre suivant nous allons comparer par simulation et sur des données réelles la perfermance de différents noyaux discrets dans l'estimation de la fonction de régression

3

Applications numériques

3.1 Introduction

Nous présentons dans ce chapitre, une simulation et trois exemples de jeux de données pour l'étude du l'influence du choix de noyau discret binomial, binomial négative, poisson, triangulaire, dirac uniforme discret, Wang-VanRyzin et le noyau Optimal dans l'estimation non paramétrique de la fonction de régression.

3.2 Etude de simulation

Dans cette section, nous présentons une étude de simulation pour évaluer l'utilisation de différents noyaux discrets pour l'estimation non paramétrique de la fonction de régression sur des données de comptage simulées selon le modèle de régression suivant :

$$y_i = m(x_i) + \epsilon_i \tag{3.1}$$

avec
$m(x_i) = \frac{2^{x_i}}{x_i!}$ la fonction de régression discret et l'erreur ϵ_i de modèle gaussien $N(0, \sigma^2)$ avec $(\sigma = 0.2)$.
D'abord, nous utilisons la méthode de validation croisée pour le choix de paramètre de lissage. Ensuite, Pour faire une comparaison sur la performance des éstimateurs à noyaux, nous utilisons le critère de l'herreur quadratique moyenne (ASE_{moy}) à

chaque échantillon de taille $n = (20, 50, 100, 200, 500)$ basé sur le nombre de simulation $N_{sim} = 100$, tell que :

$$ASE = \frac{1}{n}\sum_{i=1}^{n}[\hat{m}(x_i, h) - m(x_i, h)]^2 \tag{3.2}$$

d'où

$$ASE_{moy} = \frac{1}{N_{sim}}\sum_{j=1}^{N_{sim}}(ASE)_j \tag{3.3}$$

Les résultats de (ASE_{moy}) sont représenté dans le tableau (Table (3.1)).

		Binomial	Binomial négative	poisson
	n=20	0.009613863	0.06101554	0.01124089
	n=50	0.003840419	0.02434837	0.004621298
ASE_{moy}	n=100	0.0003267324	0.01217418	0.002240418
	n=200	0.0001633662	0.00608709	0.001120209
	n=500	0.0001248296	0.003296508	0.0006823661
		Triangulaire(a=1)	Triangulaire(a=2)	Triangulaire(a=4)
	n=20	0.002386133	0.01060954	0.05763107
	n=50	0.0008390217	0.004232443	0.01630222
ASE_{moy}	n=100	0.0005034131	0.002116222	0.008151111
	n=200	0.0004242893	0.001058111	0.002369881
	n=500	0.0004656127	0.0007770595	0.001630222
		dirac uniforme(c=2)	Wang-VanRyzin	Noyau Optimal (k=2)
	n=20	0.1486337	0.0000000034	0.00490154
	n=50	0.06446813	0.0000000013	0.002037373
ASE_{moy}	n=100	0.03281537	0.0000000006	0.001488525
	n=200	0.01654742	0.0000000003	0.0005093433
	n=500	0.006651808	0.0000000001	0.0001859341

TABLE 3.1 – Estimation de critère ASE moyenne pour $N_{sim} = 100$.

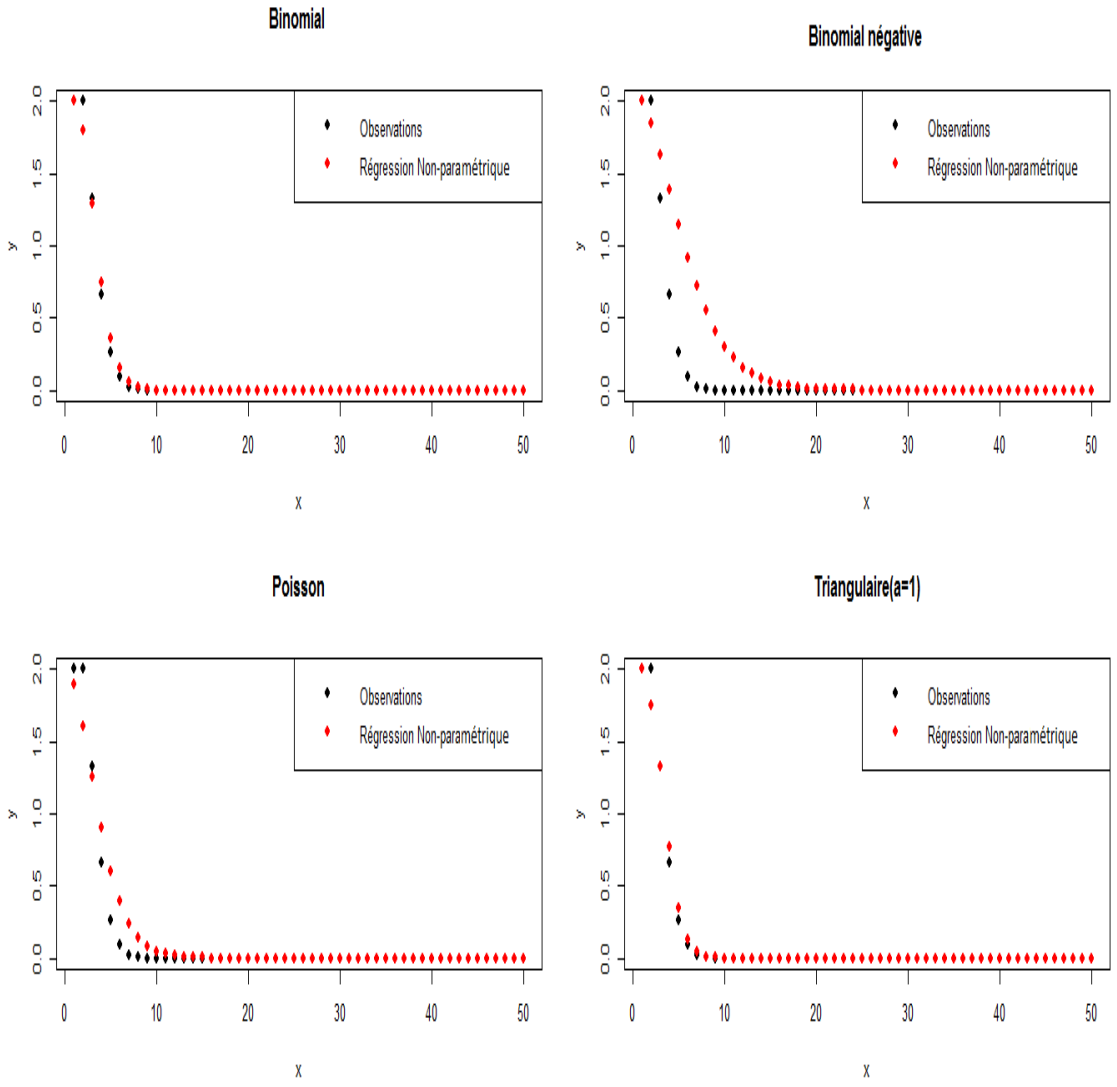


FIGURE 3.1 – Estimation non paramétrique de la fonction de régression par noyau associé univarié binomial, binomial négative, poisson et triangulaire avec($n = 50$) pour des donnés simulés.

3.2.1 Interprétation des résultats

Daprés les résultats du tableau(3.1) on peut observée que :

• La valeur de ASE_{moy} diminue lorsque la taille d'échantillon n augmente pour tous les noyaux.

•La valeur de ASE_{moy} de noyau triangulaire augmente avec l'augmentation de bras a.

• Le meilleur résultat est obtenu avec le noyau associé discret Wang-VanRyzin quel que soit la taille de l'échantillon.

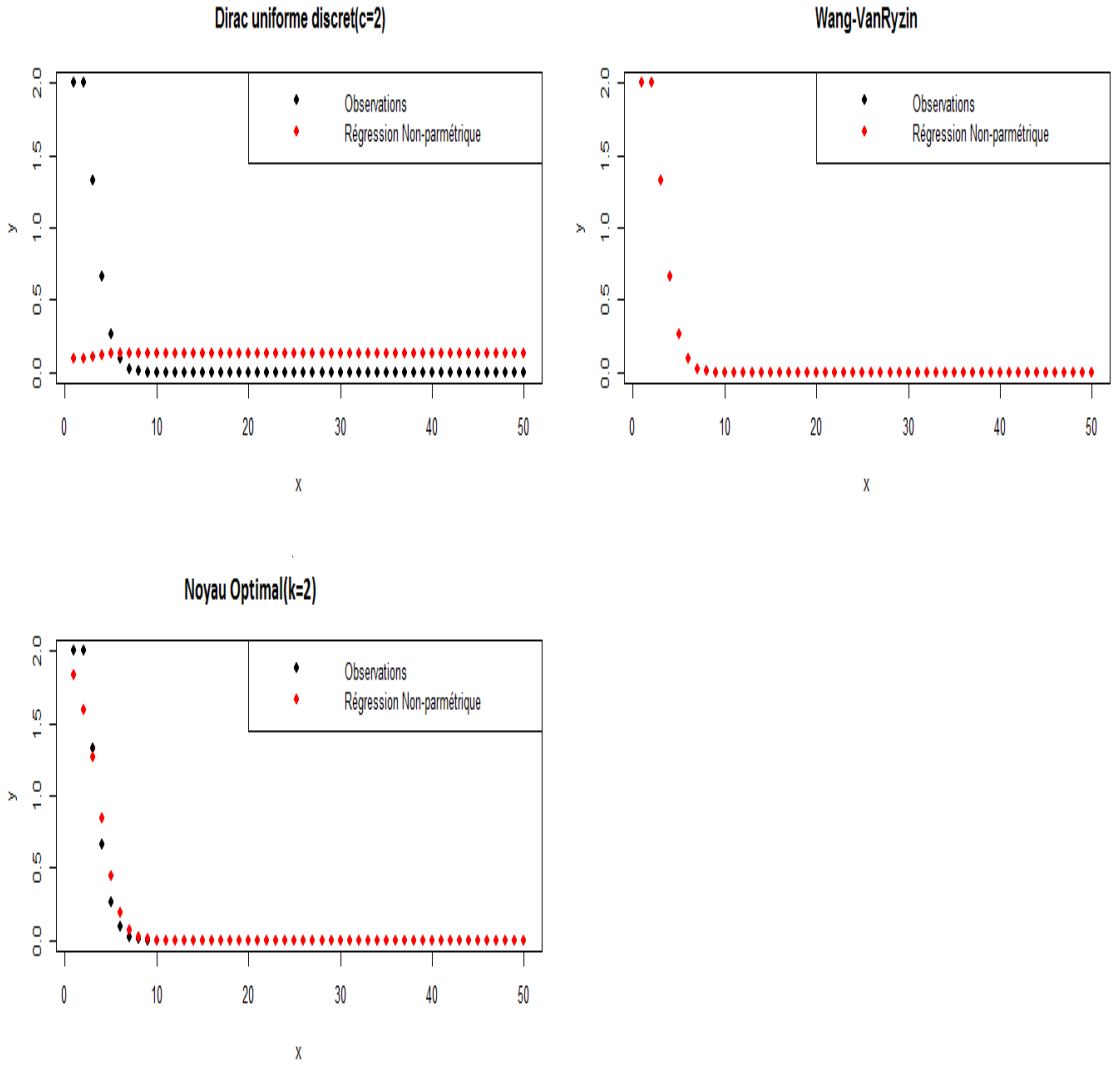


FIGURE 3.2 – Estimation non paramétrique de la fonction de régression par noyau associé univarié dirac uniforme, Wang-VanRyzin et le noyau Optimal avec($n = 50$) pour des donnés simulés.

• Les plus mauvais résultats au sens de ASE_{moy} sont obtenu on utilisant les noyaux dirac uniforme discret(c=2), binomial négative et poisson.
• Les noyaux Optimal, triangulaire (a=1) et binomial donnent des bons résulats au sens de ASE_{moy} .

3.3 Application sur des données réelles

Dans la nature, il existe de nombreux exemples de données de comptage provenant de domaines très divers comme l'agriculture, l'économie, la médecine, l'assurance, le sport, etc. La diversité des données ouvre un large champ d'application des estimateurs à noyau discret de la fonction de régression et renforce l'intérêt de cette section sur trois jeux de données.
D'abord, nous utilisons la méthode de validation croisée pour le choix de paramètre de lissage. Ensuit, Pour évaluer la performance de ses estimateurs non-paramétrique on utilise le coefficient de détermination R^2 et RMSE (Root Mean Squared Error), qui sont défini comme suite :

Définition 3.3.1 *(R^2)*

Le coefficient de détermination est une mesure de la qualité de la prédition est défini par le rapport de la variance expliquée par la régression sur la variance totale.

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{3.4}$$

où
$\hat{m}(x_i, h) = \hat{y}_i$, $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ *et* $\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
Plus la valeur de R^2 *est proche de* 1 *plus le modèle évalué est meilleur .*

Définition 3.3.2 *Critère RMSE (Root Mean Square Error) : la caractérisation de la taille des écarts entre les observations y_i et les péditions $\hat{y}_i$– Critère exactitude.*
Le biais nous indique des écarts, mais il ne nous donne pas d'information sur l'amplitude de ces écarts, vu que les valeurs positives et négatives de ϵ_i se compensent dans la moyenne. Le critère RMSE (Root Mean Square Error) permet de faire ce calcul.
L'amplitude des écarts ϵ_i peut se caractériser par la moyenne des carrés des écarts ϵ_i afin de les rendre positifs. Le calcul est le suivant :

$$\begin{aligned} RMSE &= \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\epsilon_i)^2} \\ &= \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \end{aligned}$$

Lorsque l'on utilise l'indicateur sans la racine carrée, nous obtenons un autre indicateur, que l'on l'appelle MSE (Mean Square Error).
Plus la valeur des critères RMSE où MSE est proche de zéro plus le modèle évalué est meilleur en terme d'exactitude [37].

3.3.1 Exemples sur les jeux de données

Le premier jeu de données qui est défini dans le tableau (3.2)représente des données de croissance des garçons Eubank[1999][13]. Les mesures exposées y sont des tailles établies en millimètres en fonction d'âge en année x.

x_i	1	2	3	4	5	6	7	8	9	10
y_i	745	859	940	1007	1065	1121	1183	1238	1298	1348
x_i	11	12	13	14	15	16	17	18	19	20
y_i	1391	1470	1578	1664	1708	1727	1727	1729	1738	1738

TABLE 3.2 – Données numériques de croissance.

	Binomial	Binomial négative	Poisson
h_{cv}	0.02779361	0.001066041	0.188379
R^2	0.9677335	0.8918081	0.7796691
RMSE	11.11786	290.4173	50.73348
	Triangulaire(a=1)	Triangulaire(a=2)	Triangulaire(a=4)
h_{cv}	0.618416	0.001066041	0.001066041
R^2	0.9620942	0.9146751	0.7746481
RMSE	16.51403	28.17793	55.34636
	dirac uniforme(c=2)	Wang-VanRyzin	Noyau Optimal (k=2)
h_{cv}	0.7642139	0.001063522	0.3828139
R^2	0.001282457	0.9999627	0.9658732
RMSE	334.8187	0.01688948	10.94204

TABLE 3.3 – Résultas de coefficient de détermination R^2 et $RMSE$ des régressions sur les données de croissance tableau (3.2) par les estimateurs à noyaux discrets.

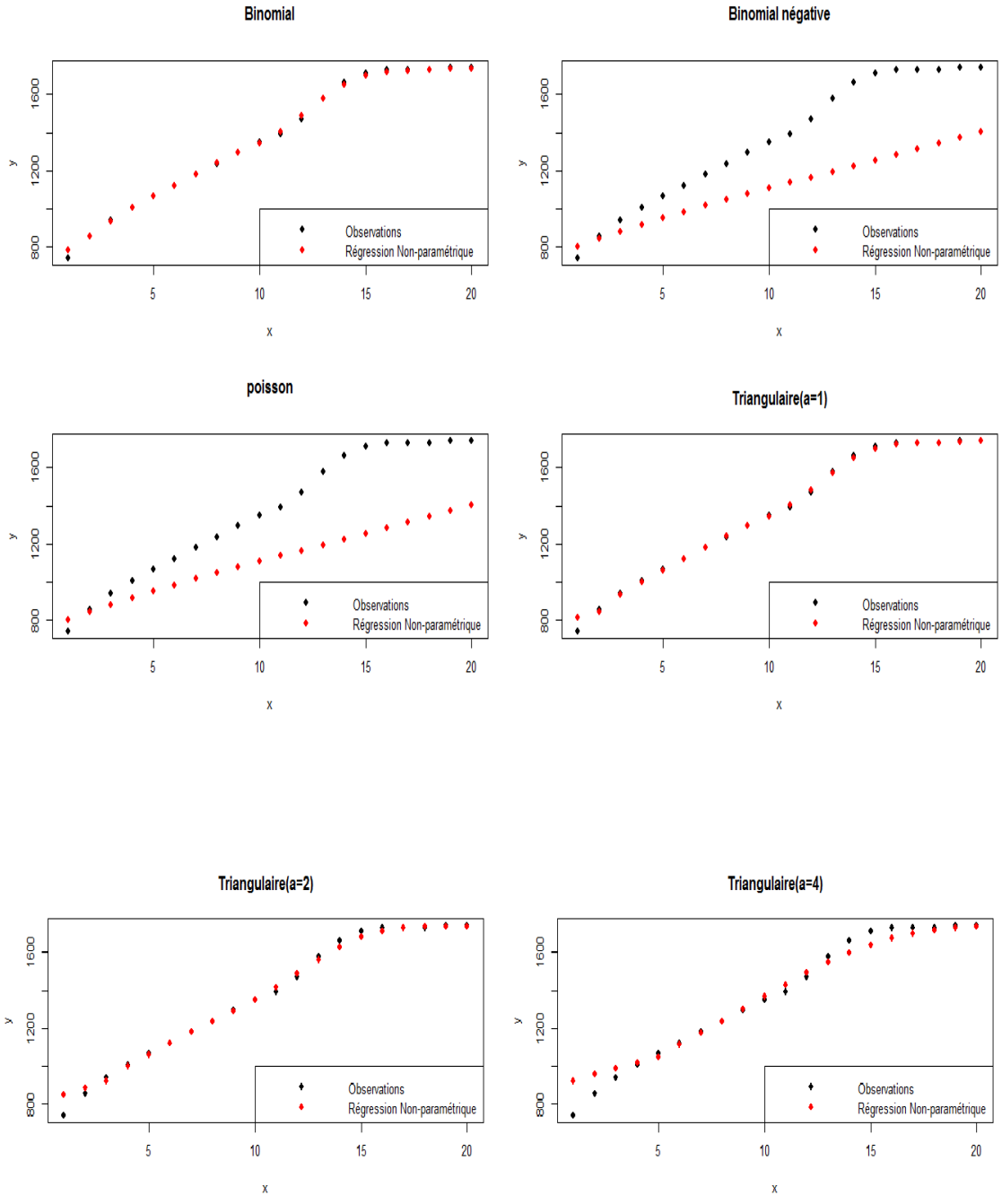


FIGURE 3.3 – Régressions sur les données de croissance (Table (3.2)) par les estimateurs à noyau discret binomial, binomial négative, poisson et triangulaire.

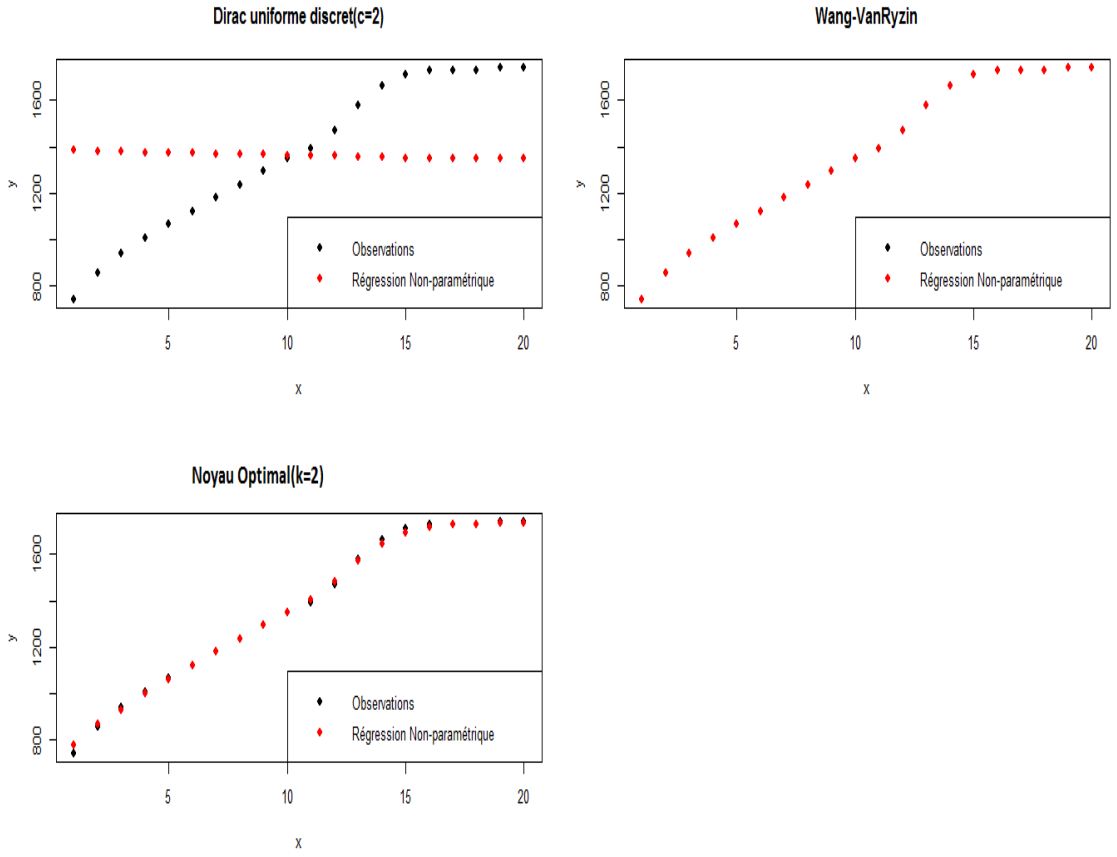


FIGURE 3.4 – Régressions sur les données de croissance (Table (3.2)) par les estimateurs à noyau discret dirac uniforme, Wang-VanRyzin et le noyau optimal.

Le deuxième jeu de données (voir Table (3.4)) portent sur l'étude de la moyenne journalière de graisse (kg/jour) fournit par le lait d'une vache pendant 35 semaines Mc-Culloch [2001][38]. La quantité de graisse contenue dans le lait augmente pendant les quatorze premières semaines avant de commencer à diminuer. Cela correspondrait à un cycle de production de lait qui dépendrait des périodes d'allaitement des veaux.

x_i	1	2	3	4	5	6	7	8	9	10	11	12
y_i	0.31	0.39	0.50	0.58	0.59	0.64	0.68	0.66	0.67	0.70	0.72	0.68
x_i	13	14	15	16	17	18	19	20	21	22	32	24
y_i	0.65	0.64	0.57	0.48	0.46	0.45	0.31	0.33	0.36	0.30	0.26	0.34
x_i	25	26	27	28	29	30	31	32	33	34	35	
y_i	0.29	0.31	0.29	0.20	0.15	0.18	0.11	0.07	0.06	0.01	0.01	

TABLE 3.4 – Moyenne journalière de graisse (kg/jour) dans le lait produit par une vachesur 35 semaines (McCulloch, 2001).

	Binomial	Binomial négative	Poisson
h_{cv}	0.001066041	0.001066041	0.15062
R^2	0.9624677	0.8222796	0.6739469
RMSE	0.02108807	0.2642198	0.05509005
	Triangulaire(a=1)	Triangulaire(a=2)	Triangulaire(a=4)
h_{cv}	0.6706409	0.001066041	0.001066041
R^2	0.9646832	0.9267871	0.8222796
$RMSE$	0.02521399	0.03459284	0.2642198
	dirac uniforme(c=2)	Wang-VanRyzin	Noyau Optimal (k=2)
h_{cv}	0.4239289	0.001054824	0.1477109
R^2	0.0001030187	0.9999334	0.9842728
RMSE	0.2142176	0.0000034	0.00582888

TABLE 3.5 – Résultas coefficient de détermination R^2 et $RMSE$ des régressions sur les données de graisse (Table (3.4)) par les estimateurs à noyaux discrets.

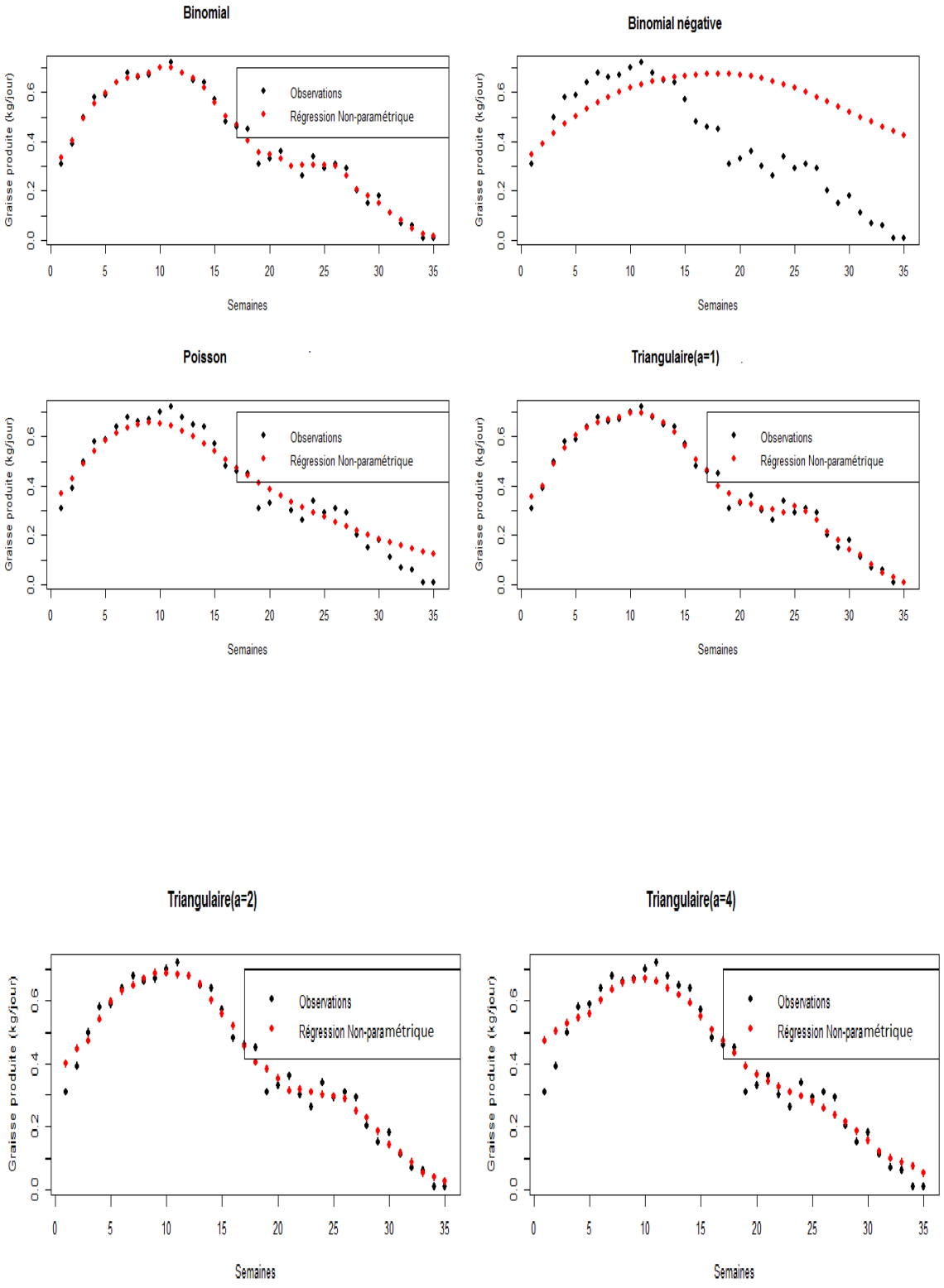


FIGURE 3.5 – Régressions sur les données de graisse (Table (3.4)) par les estimateurs à noyau discret binomial, binomial négative, poisson et triangulaire.

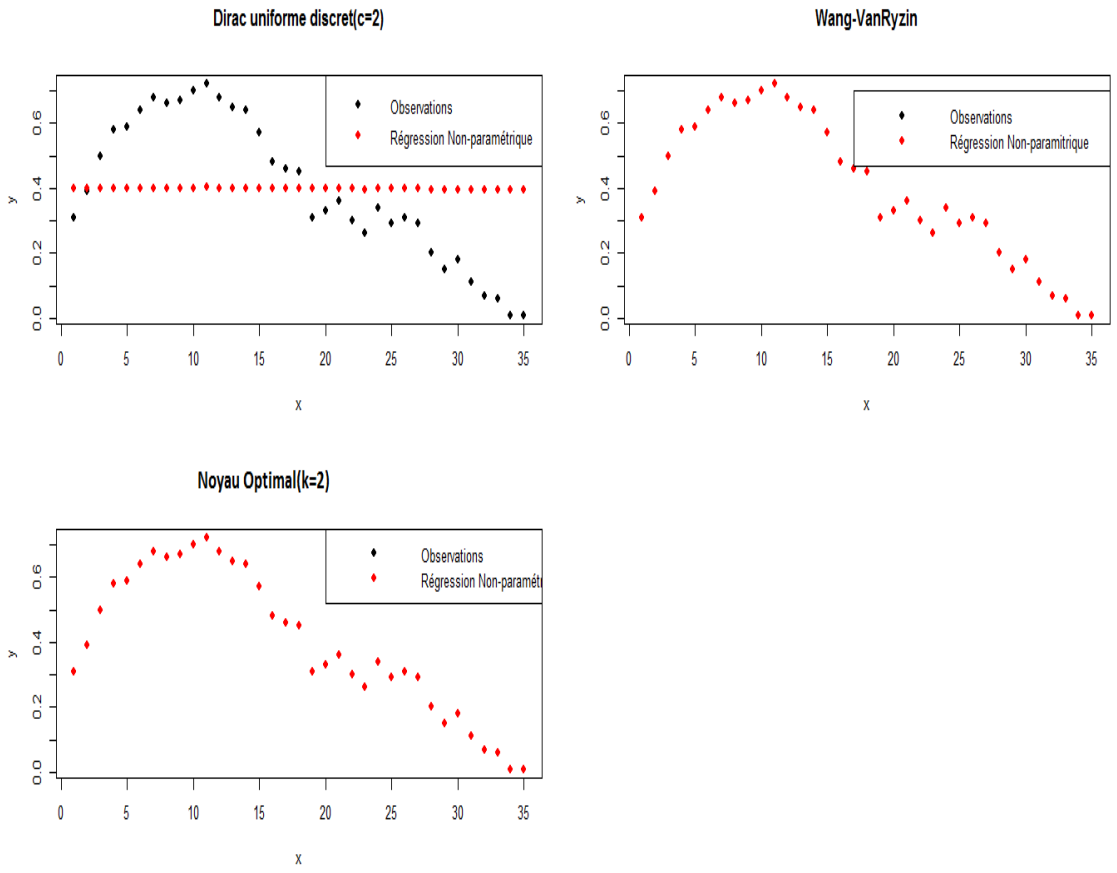


FIGURE 3.6 – Régressions sur les données de graisse (Table (3.4)) par les estimateurs à noyau discret dirac uniforme, Wang-VanRyzin et le noyau optimal.

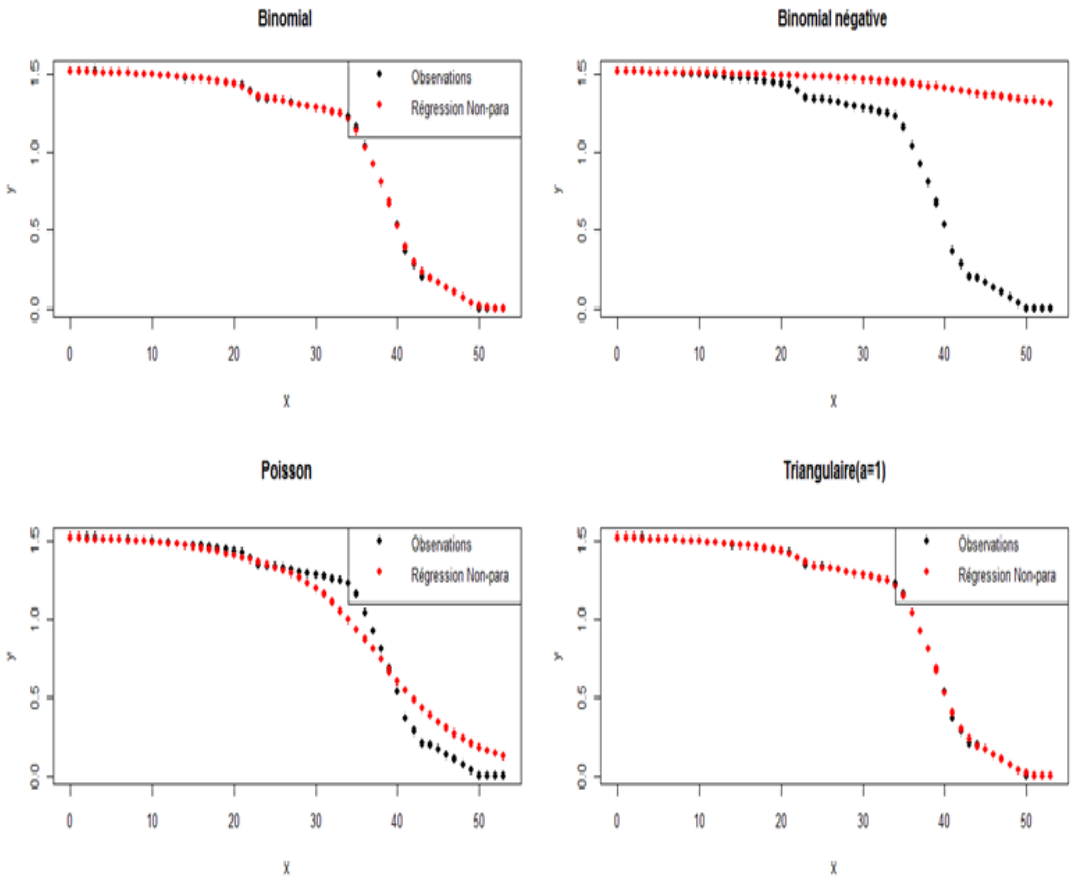
Le triosième jeu de données (Table (3.6)) concerne des données sur le volume de l'arbre de hêtre dans les forêts. Ces données sont fournies par l'agence nationale française des projets de recherche Senga Kiessé and Rivoire [2011][39]. Le volume des tiges cumulatif noté y a été calculé pour chaque diamètre $x \in \{0, 1, \cdots, 53\}$(cm) sur la base de tronc de cône. Plus précisément, à la base de l'arbre, où le diamètre est proche de 53 cm,le volume cumulatif est égal à 0, tandis qu'à l'extrémité de l'arbre, le diamètre est proche de 0, et le volume cumulatif est le volume total de la tige.

x_i	0	1	2	3	4	5	6	7
y_i	1.51625	1.51621	1.51609	1.51575	1.51510	1.51384	1.51126	1.50838
x_i	8	9	10	11	12	13	14	15
y_i	1.50535	1.50195	1.49773	1.49436	1.48930	1.48450	1.47750	1.47750
x_i	16	17	18	19	20	21	22	23
y_i	1.47102	1.46363	1.45516	1.44807	1.43756	1.42577	1.3922	1.34545
x_i	24	25	26	27	28	29	30	31
y_i	1.33754	1.33329	1.32418	1.31423	1.30341	1.29166	1.27896	1.27225
x_i	32	33	34	35	36	37	38	39
y_i	1.25805	1.24281	1.22648	1.15572	1.04385	0.92528	0.80005	0.67887
x_i	40	41	42	43	44	45	46	47
y_i	0.53924	0.36795	0.27868	0.21124	0.19722	0.16811	0.13754	0.10548
x_i	48	49	50	51	52	53		
y_i	0.07190	0.03674	0	0	0	0		

TABLE 3.6 – Donnés des arbres de hête.

	Binomial	Binomial négative	Poisson
h_{cv}	0.00278041	0.001066041	0.3648908
R^2	0.9927485	0.5461066	0.7646503
RMSE	0.005280843	0.6446578	0.1078911
	Triangulaire(a=1)	Triangulaire(a=2)	Triangulaire(a=4)
h_{cv}	0.618416	0.001066041	0.001066041
R^2	0.9942577	0.9849645	0.9519128
RMSE	0.007150092	0.01386434	0.0332244
	dirac uniforme(c=2)	Wang-VanRyzin	Noyau Optimal (k=2)
h_{cv}	0.691292	0.001066041	0.3269117
R^2	0.0003552728	0.9999915	0.9937709
RMSE	0.7093073	0.00001031	0.005608744

TABLE 3.7 – Résultas coefficient de détermination R^2 et $RMSE$ des régressions sur les données des arbres de hêtre (Table (3.6)) par les estimateurs à noyaux discrets .



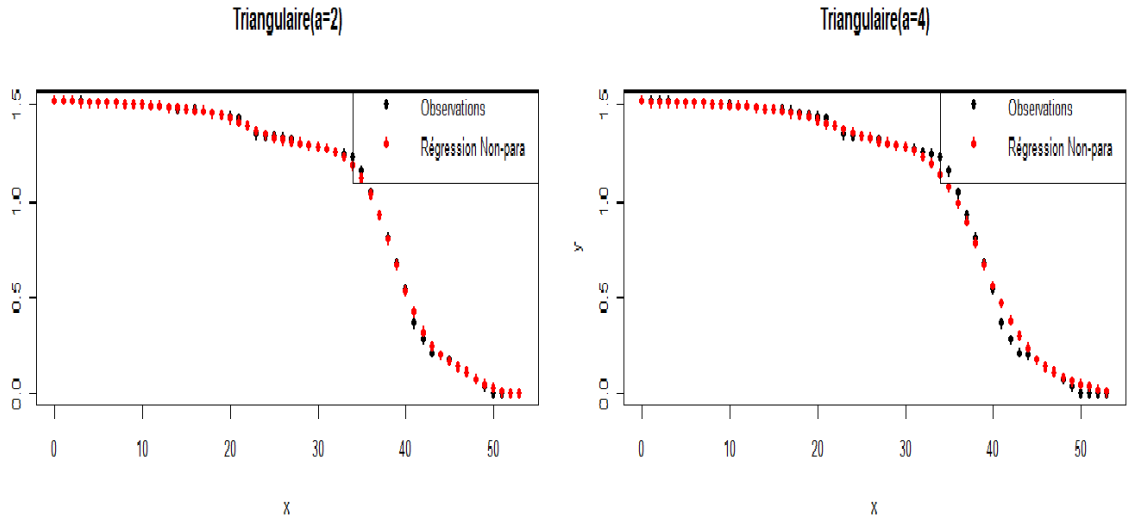


FIGURE 3.7 – Régressions sur les données des arbres hêtre(Table(3.6)) par les estimateurs à noyau discret binomial, binomial négative, poisson et triangulaire

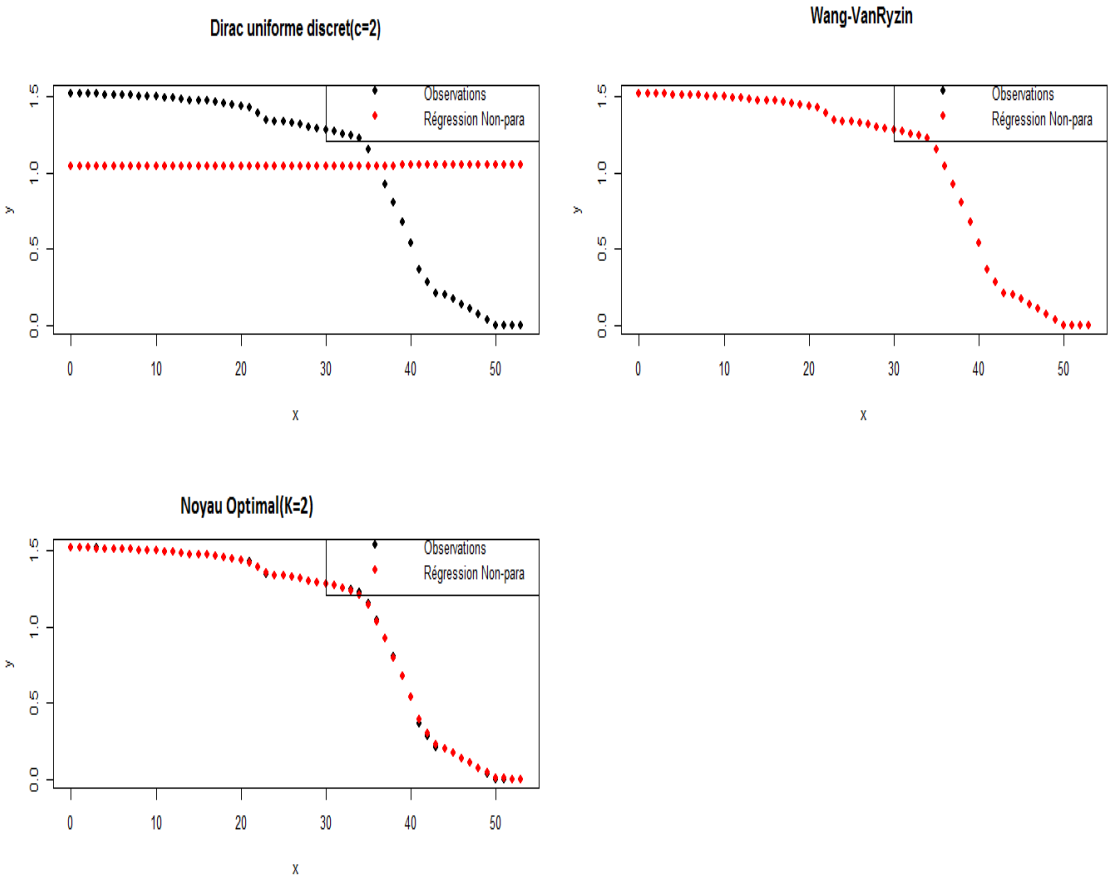


FIGURE 3.8 – Régressions sur les données des arbres hêtre(Table(3.6)) par les estimateurs à noyau discret dirac uniforme, Wang-VanRyzin et le noyau optimal.

3.3.2 Discussion de résultas

D'aprés les tableaux(3.3). (3.5).(3.7) on peut observer que :
• Le meilleur résultat est obtenu avec le noyau associé discret Wang-VanRyzin .
• Les noyaux discrets noyau optimal, binomial et triangulaire avec $(a = 1)$, produit des bons résultas au sens de R^2 et de $RMSE$ pour h_{cv} optimal.
• La valeur de $RMSE$ augmente avec l'augmentation de bras $a \in \{2, 4\}$ de noyau triangulaire et la valeur de R^2 diminué qui signifier que pour des grands valeurs de a on obtient des mauvais régressions.
• Les plus mauvais résultats sont obtenu avec le noyau binomial négative et le noyau de poisson pour h_{cv} optimal .
Selon les figures (3.3).(3.4) .(3.5).(3.6).(3.7).(3.8) on remarque que :
• Les qualités de lissage obtenues avec le noyau Wang-VanRyzin, le noyau optimal, binomial et triangulaire avec $(a = 1)$ sont plus lisses que le noyau binomial négative et poisson .

Remarque :

Selon Aitchison and Aitken [1976][14]Le noyau dirac uniforme est applicable pour estimer des fonctions discrètes où les données sont de type catégorielles.Ce qui nous avons confirmé dans les troix jeux de donnés qui ne sont pas de type catégorielles, tell que nous avons obtenu des plus mauvais résultas au sens de R^2 et de $RMSE$.

3.4 Conclution

Dans ce dernier chapitre, nous avons discuté la performance d'utilisation des noyaux discrets dans l'estimation non paramétrique de la fonction de régression sur des données simulées on utilisant le critère (ASE) moyenne qui nous permettrons de comparer entre ces différents noyaux. Ensuit, nous avons illustré cette régression sur trois jeux de données on utilisant le coefficient de détermination R^2 et $RMSE$. Nous avons constaté que Les noyaux Wang-VanRyzin, le noyau optimal, binomial et triangulaire est plus perfermants que le noyau binomial négative et poisson dans l'estimation non paramétrique de la fonction de régression.

Conclution générale

Dans ce travail nous avons présenté la méthode de noyau discret pour l'estimation non paramétrique de la fonction de régression. Les noyaux utilisés sont le noyau binomial, binomial négative, poisson, triangulaire, dirac uniforme discret , Wang vang -Ryzan et le noyau optimal. Le but principal est comparer l'influence du choix de noyau discret dans l'estimation non paramétrique de la fonction de régréssion en s'appuyant sur des exemples des fonctions de régression simulés et sur des cas réelles. 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'estimation non paramétrique de la fonction de régréssion, la méthode de noyau, la régressograme, la méthode des séries orthogonales et la méthode de lissage par les fonction splines .Dans le deuxième chapitre, nous sommes intéressé à la méthode classique de noyau discret, nous avons défini l'estimateur et ses 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 trois jeux de données la performance des noyaux discrets et leur influence dans l'estimation non paramétrique de la fonction de régression. Cette comparaison basé sur le critère ASE moyenne pour le cas simulés et le coefficient de détermination R^2 et $RMSE$ pour le cas réel. Nous avons constaté que :

• Le meilleur résultat est obtenu avec le noyau associé discret Wang-VanRyzin .

• Les noyaux discrets noyau optimal, binomial et triangulaire avec $(a = 1)$, produit des bons résultas .

• Les plus mauvais résultats sont obtenu avec le noyau binomial négative et le noyau de poisson .

Le travail réalisé offre plusieurs perspectives,nous citons :

✓ La construction d'autre noyaux associés discrèts.

✓ Utiliser les noyaux dirac uniforme discret , Wang vang -Ryzan et le noyau optimal dans l'estimation non paramétrique de la fonction de régréssion multivarié.

✓ Utiliser l'approche bayésienne pour la sélection de paramètre de lissage.

Bibliographie

[1] Nadaraya, E.A. (1964). On estimating regression. Theory of Probability and its Applications 9, 141–142.

[2] Watson, G.S. (1964). Smooth regression analysis. Sankhy a Ser. A 26, 359–372.

[3] Tukey J.W. Curves as parameters and touch estimation. Proc ; of the 4th Berkeley Sump. on Math. Stat.Prob. 681-694.

[4] N. N. Cencov. Evaluation of an unknown distribution density from observations. Soviet Mathematics, 3 :1559–1562, 1962.

[5] S. C. Schwartz. Estimation of probability densities by an orthogonal series. the Annals of Mathematical Statistics, 38 :1261–1265, 1967.

[6] R. Kronmal and M. Tarter. The estimation of probability densities and cumulatives by fourier series methods. Journal of the American Statistical Association, 63 :925–952, 1968.

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

[8] D. Bosq. Estimation sur optimale de la densité par projection. Canadian Journal of Statistics, 33 :21–37, 2005.

[9] N. Saadi and S. Adjabi. On the estimation of the probability density by trigonometric series. Communicatins in Statistics-Simulation and Computation, 38 :3583–3595, 2009.

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

[11] B W. Silverman. Some aspects of the spline smoothing approach to non-parametric regression curve fitting. 13 :1–52, 1985.

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

[13] Eubank, R. L. (1999). Nonparametric Regression and Spline Smoothing. Marcel Dekker, Inc, New York.

[14] J. Aitchison and C. G. G. Aitken. Multivariate binary discrimination by the kernel method. Biometrika, 63 :413–420, 1976.

[15] C C. Kokonendji, T. Senga Kiessé, and S S. Zocchi. Discrete triangular distributions and non-parametric estimation for probability mass function. Journal of Nonparametric Statistics, 19 :241–257, 2007b.

[16] T.Senga Kiessé.Approche non- paramétrique par noyaux associées discrets des donnés de dénombremment.Thèse de doctorat ,université de Pau,2008.

[17] C C. Kokonendji and T. Senga Kiessé. Discrete associated kernels method and extensions. Statistical Methodology, 8 :497–516, 2011.

[18] T.senga kiessé,Gilles Durrieu optimal symmetric Kernels for estimating count data distributions.2020.

[19] S.A. Kessentini. Modèles graphiques probabilistes pour l'estimation de densité en grande dimension : applications du principe perturb et Combine pour les mélanges d'arbres. Thèse de doctorat en informatique, Université de Nante, France, Décembre 2010.

[20] D. Blondin. Lois limites uniformes et estimation non-paramétrique de la régression. Thèese Doctorat, Universitée Paris 6, 1 - 26., 2004.

[21] Simonoff, J. S. (1996). Smoothing Methods in Statistics. Springer Series in Statistics. Springer-Verlag, New York.

[22] Nadji.R.Estimationnon paramétrique de la fonction de régréssion :régressogramme et méthode de spaline. Thèse de doctorat,Univirsté des science et technique de lille 1,novembre 1980.

[23] Thi Mong Ngoc. N .Estimation récursive pour des modèles semi paramétriques.Thèse de doctorat en mathématiques Appliquées Statistique,Université bordeaux 1,novembre 2010.

[24] Hastie, T. J. et Tibshirani, R. J. (1990). Generalized Additive Models. Chapman and Hall, Londre.

[25] Schimek, M.G. (2000). Smoothing and regression approaches, computation, and application. John Wiley Sons, Inc, New York.

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

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

[28] G.szego. Orthogonal polynomials. Amer.Math.Soc.Coll.Publ, 1959.

[29] Jean-Jacques Droesbeke.Approches non paramétriques en régression.1-143.

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

[31] Y. Cao. Inégalités d'oracle pour l'estimation de la régression. Thèse Doctorat, Universitè de Provence, 2008.

[32] Sobom Matthieu Some. Estimations non paramétriques par noyaux associés multivariés et applications.Statistiques [math.ST]. Université de Franche-Comté, 2015. Français.

[33] J. S. Racine and Q. Li. Nonparametric estimation of regression functions with both categorical and continuous data. Journal of Econometrics, 119 :99–130, 2004.

[34] M. C. Wang and J. V. Ryzin. A class of smooth estimators for discrete distributions. Biometrika, 68 :301–9, 1981.

[35] C. C. Kokonendji and S. S. Zocchi. Extensions of discrete triangular distributions and boundary bias in kernel estimation for discrete functions. Statistics and Probability Letters, 80 :1655–1662, 2010.

[36] Zougab, N. (2013). Approche Bayésienne dans l'Estimation Non-paramétrique de la densité de Probabilité et la Courbe de Régression de la Moyenne. Thèse à Université Abderrahmane Mira de Béjaïa, Algérie.

[37] Gy, P.M., (1998), Sampling for Analytical Purposes, John

[38] Wiley and Sons Ltd., Chichester. McCulloch, C.E. (2001). An Introduction to Generalized Linear Mixed Models. 46a Reunião Anual da RBRAS - 9o SEAGRO, University of São Paulo - ESALQ, Piracicaba.

[39] T. Senga Kiessé and M. Rivoire. Discrete semi-parametric régression models with associated kernel and applications. Journal of Nonparametric Statistics, 23(4) :927–941, 2011.

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. Dans ce travail, on présente quelques méthodes de l'estimation non paramétrique de la fonction de régression, la régressograme, la méthode des séries orthogonales la méthode de lissage par les fonctions splines et la méthode du noyau. Ensuit nous avons étudier plus particulièrement la méthode du noyau dans le cas discret ainsi que leurs propriétés statistiques et asymptotiques. le paramètre de lissage qui intervient dans la forme de l'estimateur est sélectionné par la méthode de validation croisée. Enfin, nous avons étudié l'influence du choix de noyau discret dans l'estimation non paramétrique de la fonction de régression à la base des données simulées et trois jeux de données.

Mot clés : Régression non paramétrique, estimation, noyau associé discret, validation croisé.

Abstract

The non-parametric regression is a statistical tool used to describe a relationship between a dependent variable and an explanatory variable, without specifying the form of this relationship. In this work, we present some methods of the nonparametric estimation of the regression function, the régressograme, the method of orthogonal series the method of smoothing by the spline functions and the method of the kernel. Then we study more particularly the kernel method in the discrete case as well as their statistical and asymptotic properties . The smoothing parameter that occurs in the form of the estimator is selected by the validation cross method. Finally, we studied the influence of discrete kernel choice on the non-parametric estimation of the regression function on the basis of simulated data and three datasets.

Key words : Non-parametric regression, estimation, discrete associated kernel, cross validation.