Université de Mentouri-Constantine
Faculté des Sciences Exactes
Département de Mathématique

MÉMOIRE DE MAGISTÈRE EN MATHEMATIQUE

Option : Mathématiques appliqués

N° d'ordre :……
Série :……

Intitulé :

Sur L'estimation de la Fonction de Régression

Présenté par : ***M^{elle} Hanane BEGHRICHE***

Membre de Jury composé de :

Dr. Z. Mohdeb	Prof Université de Constantine	Président
Dr. F. Messaci	M.C Université de Constantine	Rapporteur
Dr. Z. Gheribi	M.C Université de Constantine	Examinateur
Dr. N. Namouchi	M.C Université de Constantine	Examinateur

Remerciements

Au terme de ce travail, Je tiens à remercier :

Dr Fatiha Messaci, mon encadrante, qui m'a fait bénéficier de son savoir, de ses compétences scientifiques et de sa passion pour la recherche. Je vous remercie également madame de m'avoir appris à aller jusqu'au bout de mes idées.

Mes vifs remerciements aux Pr Zahir Mohdeb, Dr Zoubida Gheribi et Dr Nahima Namouchi de l'honneur qu'ils m'ont fait en acceptant de siéger à mon jury de magistère.

Mes remerciements vont à tous les membres de ma famille à qui je dois beaucoup, sans leurs aides, ce travail n'aurait pu voir le jour.

Merci à tous ceux qui m'ont aidé sans ménager ni leurs temps, ni leurs encouragements, ni leurs savoirs.

Et enfin, merci à tous les chercheurs que j'ai pu rencontrer et qui se sont intéressés à mes travaux.

Hanane BEGHRICHE

Table des matières

Introduction

La théorie de l'estimation est une des branches les plus basiques de la statistique. Cette théorie est divisée en deux volets principaux, à savoir l'estimation paramétrique et l'estimation non-paramétrique qui consiste à estimer à partir des observations, une fonction inconnue, élément d'une certaine classe fonctionnelle. Une procédure non-paramétrique est définie indépendamment de la loi de l'échantillon d'observation. Plus particulièrement, on parle de méthode 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. Plus généralement un modèle de statistique semi paramétrique comporte à la fois une composante paramétrique et une autre non paramétrique (exemple : le modèle de Cox).

Un des problèmes centraux en statistique est celui de l'estimation de caractéristiques fonctionnelles associées à la loi des observations, comme par exemple, la fonction de répartition ou la fonction de régression. Dans le modéle de régression non-paramétrique, on suppose l'existence d'une fonction $r(x)$ qui exprime la valeur moyenne de la variable réponse Y en fonction de la variable d'entrée X.

Les estimateurs non paramétriques (à noyau) de la régression ont été introduits simultanément par Nadaraya (1964) et Watson(1964), nous les étudions au chapitre 2. Ce problème a suscité un grand intérêt et a conduit à la proposition de plusieurs estimateurs sur la base de l'observation d'un échantillon du couple (X, Y). En analyse de survie, il est connu que l'observation de Y n'est pas toujours possible. Y peut être le temps de survie à une maladie . Certains sujets, sous étude, peuvent disparaître de l'étude fortuitement (accident) ou d'une manière planifiée (fin de l'étude). Y est alors censuré à droite : on ne connaît pas sa valeur exacte, on sait seule-

ment qu'elle dépasse l'observation recueillie. Dans ce contexte de censure à droite Beran (1981) a proposé une classe d'estimateurs de la fonction de survie conditionnelle permettant d'en déduire des estimateurs non paramétriques de la fonction de régression difficilement calculables. Ensuite Carbonez (1992) et Carbonez, Györfi et van der Meulen (1995) ont introduit un estimateur à partitions consistant pour Y bornée et censurée à droite. Améliorant ce travail, Kohler, Máthé et Pintér (2002) ont simplifié la preuve du résultat précédent et ont même étendu le travail en proposant d'autres estimateurs non paramétriques (à noyau, plus proches voisins, moindres carrés et spline de lissage). Cependant d'autres types de censure existent. Pour certaines valeurs de Y, on peut seulement savoir qu'elles sont inférieures aux observations. C'est la censure à gauche. Cela peut être le cas lorsqu'on s'intéresse à l'âge auquel des personnes ont commencé à accomplir une tâche. Certains individus peuvent seulement se rappeller qu'ils ont commencé cette tâche avant un certain âge sans savoir exactement lequel. Nous détaillons au chapitre 3 l'étude de l'estimateur à noyau de Kohler et autres (2002), introduit dans le cadre de la censure à droite. IL se caractérise par sa forme très simple à calculer et il a suscité de nombreux travaux parmi lesquels Ould-said et Cai (2005), Kohler et al. (2006), Brunel et Comte (2006a), Brunel et Comte (2006b) et Guessoum et Ould Said (2009). Puis nous étendons le travail précédent à un modèle de censure à gauche.

Par ailleurs, les modèles de régression se subdivisent en deux familles selon le type de régresseur que l'on considère. La première famille correspond à celle dont les régresseurs sont à valeurs dans un espace de dimension finie (variable aléatoire complète ou censurée) alors que la seconde correspond à celle dont les régresseurs sont à valeurs dans un espace de dimension infinie (variable aléatoire fonctionnelle).

C'est pourquoi on s'intéresse au traitement des variables aléatoires fonctionnelles (cas de la régression d'une variable réelle sur une variable fonctionnelle). Ce domaine de recherche de la statistique a suscité un engouement certain auprès des chercheurs vu l'ampleur des publications dans ce domaine. Les premiers résultats ont été obtenus à partir de l'étude d'un estimateur à noyau généralisé introduit par Ferraty et vieu (2000). On pourra se référer à Ferraty et vieu (2004, 2006a, 2006b) pour plus de détails sur la méthode à noyau en statistique fonctionnelle. Les résultats présentés dans ce mémoire s'inscrivent dans ce cadre et sont présentés dans le chapitre 4.

Au chapitre 5, un travail de simulation permet de calculer les estimateurs étudiés, pour des modèles choisis, afin de vérifier la qualité de ces estimateurs et de confronter les résultats pratiques à ceux attendus par la théorie.

Chapitre 1

Rappels

Rappelons les résultats suivants vu leurs utilités dans la suite.

1) Théorème de Bochner

Théorème (1.1) *Soit* $K : (\mathbb{R}^m, B^m) \to (\mathbb{R}, B)$ *une fonction mesurable, où* B^m *est la tribu borélienne de* $\mathbb{R}^m$*, vérifiant :*

$$\exists\ M \text{ constante } /\ \forall z \in \mathbb{R}^m,\ \ |K(z)| \leq M,$$

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

et

$$||z||^m |K(z)| \to 0 \ \textit{quand}\ \ ||z|| \to \infty.$$

Par ailleurs, soit

$g : (\mathbb{R}^m, B^m) \to (\mathbb{R}, B)$ une fonction mesurable telle que

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

On définit :

$$g_n(x) = \frac{1}{h_n^m} \int_{\mathbb{R}^m} K(\frac{z}{h_n}) g(x - z) dz,$$

où $0 < h_n \to 0$ quand $n \to \infty$.

Si g est continue, alors

$$\lim_{n\to\infty} g_n(x) = g(x) \int_{\mathbb{R}^m} K(z) dz$$

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

2) Inégalité de type Bernstein

Il existe plusieurs versions d'inégalités de ce type. Nous nous contentons de rappeler dans le lemme ci dessous une version simplifiée, qui nous suffit dans ce travail et dont la preuve est donnée dans l'article de Hoeffding (1963).

Lemme 1.1 *Soit $\Delta_1, ..., \Delta_n$ des v.a.r.centrées, indépendantes et de même loi, telles qu'il existe deux réels positifs d et δ^2 vérifiant :*

$$|\Delta_1| \leq d \text{ et } E\Delta_1^2 \leq \delta^2.$$

Alors, pour tout $\varepsilon \in \left]0, \frac{\delta^2}{d}\right[$ on a

$$P\left[n^{-1}\left|\sum_{i=1}^{n} \Delta_i\right| > \varepsilon\right] \leq 2 \exp\left\{-\frac{n\varepsilon^2}{4\delta^2}\right\}.$$

3) Convergence du maximumum de variables aléatoires i.i.d.

Soit $Z_1, Z_2, ..., Z_n$ n v.a. i.i.d., de fonction de répartition F, posons

$$T_{k_n} = \max_{1 \leq i \leq n} \{Z_1, Z_2, ..., Z_n\}$$

et

$$T_k = \sup \{t : F(t) < 1\}.$$

En supposant que $T_k < \infty$, on montre que $T_{k_n} \to T_k$ $(n \to \infty)$ $p.s.$

En effet, on a d'une part

$$P(T_{k_n} \leq t) = P(Z_1 \leq t, Z_2 \leq t, ..., Z_n \leq t) = F^n(t) \to \begin{cases} 0 & \text{si } F(t) < 1 \\ 1 & \text{si } F(t) = 1 \end{cases},$$

qui est la fonction de la répartition de la v.a. presque sûrement égale à T_k, donc $T_{k_n} \to T_k$ en loi et comme T_k est constante alors $T_{k_n} \to T_k$ en probabilité.

D'autre part, on a T_{k_n} est croissante et presque sûrement majorée par T_k, elle converge donc, presque sûrement, vers X. Mais

$$\begin{cases} T_{k_n} \to T_k & \text{en probabilité} \\ T_{k_n} \to X & \text{en probabilité} \end{cases} \Rightarrow T_k = X \quad p.s.$$

4) Convergence presque complète

Définition 1.1 *La suite de v.a.r.* $(X_n)_{n\in\mathbb{N}}$ *converge presque complètement vers la v.a.r.* X *si*

$$\forall \varepsilon > 0, \qquad \sum_{n\in\mathbb{N}} P(|X_n - X| > \varepsilon) < \infty,$$

on note $\lim_{n\to\infty} X_n = X \;\; p.co.$

Proposition 1.1 *Si* $\lim_{n\to\infty} X_n = X \;\; p.co.$, *alors* $(X_n) \to X$ *en probabilité et presque sûrement vers* X.

Preuve 1] La convergence en probabilité est claire puisque $P(|X_n - X| > \varepsilon)$ est le terme général d'une série convergente.

2] On déduit du lemmme de Borel Cantelli que

$\forall \varepsilon > 0,\ P(\limsup_n |X_n - X| > \varepsilon) = 0$. Or $\lim_{n\to\infty} X_n(\omega) \neq X(\omega) \Rightarrow \exists\ \varepsilon > 0$ tel que $\limsup_n |X_n(\omega) - X(\omega)| > \varepsilon$, ce qui montre que $P(\lim_{n\to\infty} X_n = X) = 1$, autrement dit $(X_n) \to X$ *p.s.* ■

Chapitre 2

Régression dans le cas de la dimension finie

2.1 Présentation des estimateurs

Estimateur à noyau

Soient $(X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n)$ des couples aléatoires indépendants et identiquement distribués (i.i.d.) à valeurs dans $\mathbb{R}^2$. Le couple de variable aléatoire (X, Y) est supposé admettre une densité jointe sur $\mathbb{R}^2$ notée $f_{X,Y}(.,.)$ et nous désignons par $f_X(.)$ la densité marginale (par rapport à la mesure de Lebesgue sur $\mathbb{R}$), donnée par

$$f_X(x) = g(x) = \int_{-\infty}^{+\infty} f(x, y)dy.$$

La densité conditionnelle de Y en $X = x$ est

$$f_{Y/X}(y/x) = \frac{f(x, y)}{g(x)} \qquad \text{si} \quad g(x) \neq 0.$$

L'éspérance conditionnelle ou la fonction de régression de Y en $X = x$ s'écrit

$$\begin{aligned} r(x) &= E(Y/X = x) = \int_{-\infty}^{+\infty} y f_{Y/X}(y/x)dy \\ &= \int_{-\infty}^{+\infty} \frac{yf(x,y)}{g(x)}dy. \end{aligned}$$

Soit $J(x, y)$ une fonction de densité sur $\mathbb{R}^2$, on pose

$$K(x) = \int_{-\infty}^{+\infty} J(x, y)dy$$

Soit $h_n \to 0$ quand $n \to \infty$, on a alors

$$f_n(x, y) = \frac{1}{nh_n^2} \sum_{j=1}^{n} J\left[\frac{(x - X_j)}{h_n}, \frac{(y - Y_j)}{h_n}\right]$$

est un estimateur de $f(x, y)$.

On a aussi

$$g_n(x) = \frac{1}{nh_n} \sum_{j=1}^{n} K\left[\frac{(x - X_j)}{h_n}\right] = \int_{-\infty}^{+\infty} f_n(x, y)dy$$

est un estimateur de $g(x)$.

Un estimateur naturel de $r(x)$ est donné par

$$\hat{r}_n(x) = \int_{-\infty}^{+\infty} \frac{yf_n(x,y)}{g_n(x)}dy.$$

En posant $z = \frac{y-Y_j}{h_n}$, il vient

$$\hat{r}_n(x) = h_n \sum_{j=1}^{n} m\left(\frac{(x - X_j)}{h_n}\right) / K\left(\frac{(x - X_j)}{h_n}\right) + \sum_{j=1}^{n} Y_j K\left(\frac{(x - X_j)}{h_n}\right) / K\left(\frac{(x - X_j)}{h_n}\right),$$

où

$$m(x) = \int_{-\infty}^{+\infty} yJ(x,y)dy.$$

En choisissant

$$J(x,y) = K(x)L(y),$$

avec $\int_{\mathbb{R}} yL(y)dy = 0,$ nous obtenons $m(x) = 0.$

D'où l'expression de l'estimateur de la régression :

$$\hat{r}_n(x) = \frac{\sum_{i=1}^{n} Y_i K\left(\frac{x-X_i}{h_n}\right)}{\sum_{i=1}^{n} K\left(\frac{x-X_i}{h_n}\right)},$$

appelé estimateur à noyau, K étant le noyau et h_n la fenêtre.

Généralement, les noyaux suivants sont utilisés

$$K_1(u) = \begin{cases} \frac{1}{2}, & \text{si } |u| \leq 1, \\ 0, & \text{si } |u| > 1. \end{cases} \quad \text{(noyau rectangulaire)},$$

$$K_2(u) = \begin{cases} (1-|u|), & \text{si } |u| \leq 1, \\ 0, & \text{si } |u| > 1. \end{cases} \quad \text{(noyau triangulaire)},$$

$$K_3(u) = \begin{cases} \frac{3}{4}(1-u^2), & \text{si } |u| \leq 1, \\ 0, & \text{si } |u| > 1. \end{cases} \quad \text{(noyau parabolique ou d'Epanechnikov)},$$

$$K_4(u) = \begin{cases} \frac{15}{16}(1-u^2), & \text{si } |u| \leq 1, \\ 0, & \text{si } |u| > 1. \end{cases} \quad \text{(noyau "bixeight")},$$

$$K_5(u) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{u^2}{2}) \quad \text{(noyau gaussien)},$$

$$K_6(u) = \frac{1}{2} \exp(-|u|/\sqrt{2}) \sin(|u|/\sqrt{2} + \pi/4) \quad \text{(noyau de Silverman)}.$$

Signalons que des études ont montré que $\hat{r}_n(x)$ est peu sensible à la variation du noyau, alors que le choix de la fenêtre est primordial et a donné lieu à de nombreux travaux concernant des choix "optimaux".

Plus généralement, des estimateurs de $r(x)$ peuvent s'écrire

$$\hat{r}_n(x) = \sum_{i=1}^{n} W_{n,i}(x) Y_i,$$

où $W_{n,i}(x)$ sont des poids dépendant de x et de $(X_1, ..., X_n)$. La valeur de $W_{n,i}$ dépend du type d'estimateur considéré, nous citons ci-dessous des exemples connus.

Estimateur des plus proches voisins

Soit k_n un paramètre de l'estimation, on pose

$$W_{n,i}(x) = \begin{cases} \frac{1}{k_n}, & \text{si } X_i \text{ est parmi le voisin le plus proche de } k_n \text{ de } x \text{ dans } \{X_1, ..., X_n\}, \\ 0, & \text{sinon.} \end{cases}$$

Pour plus des détails (voir Devroye et all., 1994).

Estimateur à partitions

On utilise une partition $\pi_n = \{A_{n,j} : j\}$ de $\mathbb{R}^2$ et on pose

$$W_{n,i}(x) = \sum_j \frac{I_{A_{n,j}}(X_i)}{\sum_{k=1}^{n} I_{A_{n,j}}(X_k)} I_{A_{n,j}}(x)$$

Pour plus des détails (voir Devroye et Györfi., 1983).

Signalons que d'autres types d'estimateur existent, parmi les quels nous citons l'estimateur des moindres carrés, ou plus généralement les estimateurs spline lissés (cf, par exemple Kohler and Krzyżak (2001), Kohler (1997, 1999) et Györfi et al. (1997)).

On s'intéresse principalement dans ce mémoire à l'estimateur à noyau de la fonction de régression qui jouit de nombreuses propriétés. Nous commençons par reprendre les premiers résultats le concernant, entre autres sa consistance. Ensuite nous rappellons le théorème très intéressant de Devroye et Krzyżak (1989), que nous utilisons au chapitre suivant. Il montre, entre autres, que sous des conditions sur le noyau et sur la fenêtre (qui ne sont donc pas restrictives, dans le sens que nous maîtrisons leurs choix), son erreur quadratique moyenne converge vers la valeur optimale, presque sûrement, résultat dont nous motivons l'importance.

2.2 Propriétés de l'estimateur à noyau

Les propriétés de consistance et d'absence de biais sont reprises de Nadaraya (1989).

Posons sur la fonction noyau $K : \mathbb{R} \to \mathbb{R}$ les hypothèses suivantes.

$$\begin{cases} K_1)\ K \text{ est bornée, } i.e. \sup_{u\in\mathbb{R}} |K(u)| \leq M < \infty; \\ K_2)\ \lim_{|u|\to\infty} |u|\, K(u) = 0; \\ K_3)\ \int_{\mathbb{R}} |K(u)|\, du < \infty; \\ K_4)\ \int_{\mathbb{R}} K(u)du = 1; \\ K_5)\ \forall u \in \mathbb{R}, K(u) = K(-u); \\ K_6)\ \int_{-\infty}^{+\infty} u^2 K(u)du < \infty. \end{cases}$$

Rappelons que $\hat{r}_n(x) = \frac{\hat{g}_n(x)}{\hat{f}_n(x)}$, où $\hat{g}_n(x) = \frac{1}{nh_n}\sum_{i=1}^{n} Y_i K\left(\frac{x-X_i}{h_n}\right)$et $\hat{f}_n(x) = \frac{1}{nh_n}\sum_{i=1}^{n} K\left(\frac{x-X_i}{h_n}\right)$ est l'estimateur à noyau de la densité f de$X$.

2.2.1 Consistance

Théorème 2.2.1 *Supposons que le noyau K vérifie les hypothèses $K_1 - K_6$, que $E(Y^2) < \infty$ et que $f_x(x)$ est strictement positive. Si $h_n \to 0$, $nh_n \to +\infty$ (quand $n \to \infty$), alors $\hat{r}_n(x)$ est un estimateur consistant de $r(x)$.*

Preuve Nous déduisons du théorème de Bochner que, lorsque $h_n \to 0$

$$\begin{aligned} E\left[\hat{f}_n(x)\right] &= E\left[\frac{1}{nh_n}\sum_{i=1}^{n} K\left(\frac{x-X_i}{h_n}\right))\right] \\ &= \frac{1}{h_n} E\left[K\left(\frac{x-X}{h_n}\right)\right] \\ &= \frac{1}{h_n}\int_{\mathbb{R}} K(\frac{x-t}{h_n}) f_X(t)dt \to f_X(x)\int_{\mathbb{R}} K(t)dt = f_X(x). \end{aligned}$$

Donc $\hat{f}_n(x)$ est un estimateur asymptotiquement sans biais.

D'autre part, comme les X_i sont indépendantes et identiquement distribuées, il vient que

$$\begin{aligned} var(\hat{f}_n(x)) &= \frac{1}{n} var\left[\frac{1}{h_n} K\left(\frac{x-X}{h_n}\right)\right] \\ &\leq \frac{1}{n} E\left[\frac{1}{h_n} K\left(\frac{x-X}{h_n}\right)\right]^2 \\ &= \frac{1}{nh_n} \int_{-\infty}^{+\infty} \frac{1}{h_n} K^2\left(\frac{x-t}{h_n}\right) f_X(t)dt. \end{aligned}$$

D'après le théorème de Bochner

$$\int_{-\infty}^{+\infty} \tfrac{1}{h_n} K^2\left(\tfrac{x-t}{h_n}\right) f_X(t)dt \to f_X(x) \int_{\mathbb{R}} K^2(t)dt < \infty \quad \text{quand } n \to \infty,$$

donc

$$var(\hat{f}_n(x)) \to 0 \quad \text{quand } nh_n \to \infty.$$

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

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

Nous avons

$$\begin{aligned} E(\hat{g}_n(x)) &= E\left[\frac{1}{nh_n} \sum_{i=1}^{n} Y_i K\left(\frac{x-X_i}{h_n}\right)\right] \\ &= \frac{1}{h_n} E\left[Y K\left(\frac{x-X}{h_n}\right)\right] \\ &= \frac{1}{h_n} \int_{\mathbb{R}} E(Y/X=x) K(\frac{x-t}{h_n}) f_X(t)dt \\ &= \frac{1}{h_n} \int_{\mathbb{R}} r(t) K(\frac{x-t}{h_n}) f_X(t)dt \\ &= \frac{1}{h_n} \int_{\mathbb{R}} K(\frac{x-t}{h_n}) r(t) f_X(t)dt. \\ &= \frac{1}{h_n} \int_{\mathbb{R}} K(\frac{x-t}{h_n}) g(t)dt \to g(x) \end{aligned}$$

par le théorème de Bochner

De plus, si $m(x) = \int y^2 f(x,y)dy$

$$var(\hat{g}_n(x)) = E[\hat{g}_n(x)]^2 - E^2[\hat{g}_n(x)]\tilde{} \frac{1}{h_n} m(x) \int_{\mathbb{R}} K^2(t)dt \to 0 \text{ quand } n \to \infty,$$

donc

$$\lim_{n\to\infty} var(\hat{g}_n(x)) = 0.$$

Ceci implique que

$$\hat{g}_n(x) \to g(x) \text{ en probabilité.}$$

D'où

$$\hat{r}_n(x) = \frac{\hat{g}_n(x)}{\hat{f}_n(x)} \to \frac{g(x)}{f(x)} \text{ en probabilité}$$

■

2.2.2 Absence de biais asymptotique

Théorème 2.2.2 *Sous les conditions $K_1 - K_6$ et si $f_x(x)$ est strictement positive, il vient*

a) Lorsque Y est bornée p.s., $h_n \to 0$ et $nh_n \to \infty$(quand $n \to \infty$) alors

$$E\hat{r}_n(x) = E[\hat{g}_n(x)]/E[\hat{f}_n(x)] + o((nh_n)^{-1}). \tag{2.1}$$

b) Lorsque $E[Y^2] < \infty$, $h_n \to 0$ et $nh_n^2 \to \infty$ (quand $n \to \infty$) alors

$$E\hat{r}_n(x) = E[\hat{g}_n(x)]/E[\hat{f}_n(x)] + o((n^{\frac{1}{2}}h_n)^{-1}). \tag{2.2}$$

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

Preuve En utilisant l'identité suivante

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

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

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

où

$$a_n(x) = E\left[\{\hat{g}_n(x) - E[\hat{g}_n(x)\}\{\hat{f}_n(x) - E[\hat{f}_n(x)]\}\right],$$

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

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

$$\begin{aligned} Var[\hat{g}_n(x)] &= \frac{1}{nh_n}\int_{\mathbb{R}} K^2(u)m(x - uh_n)du - \frac{1}{n}\left\{\int_{\mathbb{R}} K(u)g(x - uh_n)du\right\}^2 \\ &\approx \frac{1}{nh_n}m(x)\int_{\mathbb{R}} K^2(u)du. \end{aligned}$$

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

En utilisant l'inégalité de Cauchy-Schwartz combinée aux formules ci-dessus, on obtient

$$a_n(x) = o((nh_n)^{-1}) \tag{2.3}$$

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

$$\frac{\hat{g}_n(x)}{\hat{f}_n(x)} = \frac{\sum_{i=1}^{n} Y_i K\left(\frac{x-X_i}{h_n}\right)}{\sum_{i=1}^{n} K(\frac{x-X_i}{h_n})} \leq \frac{\sum_{i=1}^{n} MK\left(\frac{x-X_i}{h_n}\right)}{\sum_{i=1}^{n} K(\frac{x-X_i}{h_n})} = M. \tag{2.4}$$

Cette dernière ingalité (2.4) permet de borner $b_n(x)$,

$$b_n(x) \leq M \times E\left[\{\hat{f}_n(x) - E[\hat{f}_n(x)]\}^2\right]$$

$$\approx \frac{M}{nh_n} f_X(x) \int_{\mathbb{R}} K^2(u)du = o((nh_n)^{-1}). \tag{2.5}$$

Les relation (2.3) et (2.5) entraînent (2.1).

Pour démontrer le cas b), il suffit de remarquer que la relation (2.3) est toujours valable mais la relation (2.5) devient

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

$$\begin{aligned} |b_n(x)| &\leq E\left[\max_{\leq 1 i \leq n} |Y_i| \left\{\hat{f}_n(x) - E[\hat{f}_n(x)]\right\}^2\right] \\ &\leq \left\{\sum_{i=1}^{n} Y_i^2\right\}^{1/2} \times \left\{E\left[\{\hat{f}_n(x) - E[\hat{f}_n(x)]\}^4\right]\right\}^{1/2} \\ &= \sqrt{n}\left\{E[Y_i^2]\right\}^{1/2} \times o((nh_n)^{-1}) = o((n^{\frac{1}{2}}h_n)^{-1}) \end{aligned} \tag{2.6}$$

Les relations (2.3) et (2.6) impliquent (2.2), d'ou le résultat énoncé ■

2.2.3 Résultat de Devroye et Krzyżak (1989)

Motivation

Soit X un vecteur aléatoire à valeurs dans $\mathbb{R}^d$ et Y une variable aléatoire réelle de carré intégrable ($EY^2 < \infty$).

L'objet de l'analyse de régression est d'estimer Y, après l'observation de X. On cherche une fonction f telle que $f(x)$ soit la plus proche possible de Y, dans le sens qu'on veut trouver une fonction f^* qui minimise $E|f(x) - Y|^2$, i.e., f^* doit vérifier

$$E\left|f^*(x) - Y\right|^2 = \min_f E\left|f(x) - Y\right|^2$$

En notant $r(x) = E(Y/X = x)$ et μ la loi de X, on a

$$E\left|f(x) - Y\right|^2 = E\left|r(x) - Y\right|^2 + \int_{\mathbb{R}^d} E\left|f(x) - r(x)\right|^2 \mu d(x)$$

On en déduit que $f^* = r$. On cherche donc à minimiser $\int_{\mathbb{R}^d} E\left|f(x) - r(x)\right|^2 \mu d(x)$ pour que $E\left|f(x) - Y\right|^2$ soit la plus proche de sa valeur optimale $E\left|r(x) - Y\right|^2$.

Dans le cadre de l'estimation non paramétrique, la loi du couple (X, Y) est inconnue.

Sur la base d'un échantillon $(X_1, Y_1), ..., (X_n, Y_n)$ de la loi de (X, Y), on veut construire un estimateur $\hat{r}_n(x)$ de r telle que $\int_{\mathbb{R}^d} E\left|\hat{r}_n(x) - r(x)\right|^2 \mu d(x)$ soit petite. Pour cela nous avons besoin que le noyau soit régulier, notion rappelée ci dessous.

Définition 2.2.1 *On dit que le noyau positif K est régulier si $K(x) \geq BI_{S_r}$, pour des constantes positives B et r où S_r est la boule de rayon r centrée à l'origine, et*

$$\int_{y \in x + S_r} K(y)dx < \infty.$$

Enoncé du résultat

Théorème 2.2.3 *Supposons que K est un noyau régulier. Soit $\hat{r}_n(x)$ l'estimateur à noyau de la fonction de régression $r(x)$. Alors les propositions suivantes sont équivalentes.*

(*A*) *Pour chaque loi de* (X,Y) *avec* $|Y| \leq M < \infty$, *et pour toute* $\varepsilon > 0$, *il existe deux constantes* c *et* n_0 *tels que pour tout* $n \geq n_0$,

$$P\left[\int_{\mathbb{R}^d} [\hat{r}_n(x) - r(x)]\,\mu(dx) > \varepsilon\right] \leq e^{-cn}.$$

(*B*) *Pour toute loi de* (X,Y) *avec* $|Y| \leq M < \infty$,

$$\int_{\mathbb{R}^d} [\hat{r}_n(x) - r(x)]\,\mu(dx) \to 0 \quad (n \to \infty)\ p.s.$$

(*C*) *Pour toute loi de* (X,Y) *avec* $|Y| \leq M < \infty$,

$$\int_{\mathbb{R}^d} [\hat{r}_n(x) - r(x)]\,\mu(dx) \to 0 \quad (n \to \infty)\ \textit{en probabilité.}$$

(*D*)

$$\lim_{n\to\infty} h_n = 0, \qquad \lim_{n\to\infty} nh_n^d = \infty.$$

Chapitre 3

Estimation de la fonction de régression dans le modèle de censure

Pour toute variable aléatoire V, on note

$$\begin{aligned} T_V &= \sup\{t : F(t) < 1\}, \\ I_V &= \inf\{t : F(t) \neq 0\}, \end{aligned}$$

où F est la fonction de répartition de V.

3.1 Censure à droite

Dans plusieurs études, il n'est pas possible d'observer un échantillon de (X, Y). Ainsi si la variable Y est le temps de survie d'un patient, à une maladie, ce patient peut décéder d'une autre cause pendant l'étude ou être toujours vivant à la fin de celle-ci.

Dans ce cas Y n'est pas observé mais l'observation est le minimum entre Y et une variable de censure C.

Plus précisément soit Y une variable d'intérêt positive et bornée et C une variable aléatoire de censure positive.

Nous observons l'échantillon $(X_1, Z_1, \delta_1), ..., (X_n, Z_n, \delta_n)$ où $Z_i = Y_i \wedge C_i$ et

$\delta_i = I_{\{Y_i \leq C_i\}}$ (δ_i est l'indicatrice de censure).

Nous nous proposons d'estimer $r(x) = E(Y/X = x)$, les estimateurs donnés au chapitre précédent ne peuvent plus être utilisés puisque Y n'est pas toujours observé.

3.1.1 Principe de l'estimation

L'idée, introduite par (Carbonez et all (1995)), et reprise par (Kohler, Màthé et Pintér (2002)) est de remplacer Y par une estimation de sa moyenne.

Soient $S(t) = P(Y > t)$ et $H(t) = P(C > t)$ les fonctions de survie respectives de Y et C.

On suppose que

$$(H_1) : \begin{cases} (H_{1.1}) & C \text{ et } (X, Y) \text{ sont indépendants et } H \text{ est continue} \\ (H_{1.2}) & T_Y < \infty \text{ et } H(T_Y) > 0. \end{cases}$$

Remarquons que la condition $H(T_Y) > 0$ implique $T_Y < T_C$.

Soit h une fonction de $\mathbb{R}^d X\ \mathbb{R} \to \mathbb{R}$. On se propose d'estimer la moyenne $E\{h(X, Y)\}$ sur la base de l'échantillon des données censurées à droite.

Un "estimateur" sans biais de $E\{h(X, Y)\}$ est donné par :

$$\frac{1}{n} \sum_{i=1}^{n} \frac{\delta_i h(X_i, Z_i)}{H(Z_i)}.$$

En utilisant l'indépendance entre (X, Y) et C avec les propriétés de l'espérance conditionnelle, il vient

$$\begin{aligned} E\left\{\frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i h(X_i, Zi)}{H(Z_i)}\right\} &= E\left\{\frac{I_{\{Y_1\leq C_1\}}h(X_1,Y_1)}{H(Y_1)}\right\} \\ &= E\left(\frac{h(X_1,Y_1)}{H(Y_1)}E\left(I_{\{Y_1\leq C_1\}}/(X,Y)\right)\right) \\ &= E\left(h(X_1,Y_1)\right). \end{aligned}$$

Le problème est que H est inconnu. On l'estime par l'estimateur de Kaplan – Meier (1958), donné par :

$$\hat{H}_n(t) = \begin{cases} \prod_{i=1}^{n}\left[1-\frac{1-\delta_{(i)}}{n-i+1}\right]^{I_{\left[Z_{(i)}\leqslant t\right]}}, & \text{si } t < T_{K,n}, \\ \lim_{s\to T_{K,n}, s<T_{K,n}} \hat{H}_n(s), & \text{si } t \geqslant T_{K,n}. \end{cases},$$

où $T_{K,n} = \max\{Z_1, ..., Z_n\}$ et les paires $(Z_{(i)}, \delta_{(i)})$, $i = 1, \ldots, n$ sont les n paires observées (Z_i, δ_i) ordonnées en $Z_{(i)}$, i.e. $Z_{(1)} \leqslant Z_{(2)} \leqslant ... \leqslant Z_{(n)} = T_{K,n}$.

Remarquons que $\hat{H}_n$ a été légérement modifié afin de ne jamais s'annuler.

Cela suggère d'estimer $r(x)$ par

$$\hat{r}_n(x) = \sum_{i=1}^{n} W_{n,i}(x)\frac{\delta_i Z_i}{\hat{H}_n(Z_i)}, \tag{3.1}$$

avec la fonction poids $W_{n,i}(x)$ définie comme suit,

$$W_{n,i}(x) = \frac{K(\frac{(x-X_i)}{h_n})}{\sum_{j=1}^{n} K(\frac{(x-X_j)}{h_n})}.$$

Nous utilisons les notations suivantes.

Pour tout t / $0 \leqslant t \leqslant \infty$ et $x \in \mathbb{R}$, on définit

$$T_{[0,t]}(x) = \begin{cases} t, & \text{si } x > t, \\ x, & \text{si } 0 \leqslant x \leqslant t, \\ 0, & \text{si } x < 0. \end{cases}$$

Pour $f : \mathbb{R}^d \to \mathbb{R}$ définissons $T_{[0,t]}f : \mathbb{R}^d \to \mathbb{R}$ par $(T_{[0,t]}f)(x) = T_{[0,t]}(f(x))$.

Du fait que $0 \leqslant Y \leqslant T_Y < \infty$ p.s, on a $0 \leqslant r(x) \leqslant T_Y$, on estime donc $r(x)$plutôt par

$$r_n(x) = \begin{cases} T_{K,n}, & \text{si } \hat{r}_n(x) > T_{K,n}, \\ \hat{r}_n(x), & \text{si } 0 \leqslant \hat{r}_n(x) \leqslant T_{K,n}, \\ 0, & \text{si } \hat{r}_n(x) < 0. \end{cases} \quad . \tag{3.2}$$

Par analogie avec (3.2), posons

$$r_n^*(x) = \begin{cases} T_Y, & \text{si } \hat{r}_n(x) > T_Y, \\ \hat{r}_n(x), & \text{si } 0 \leqslant \hat{r}_n(x) \leqslant T_Y, \\ 0, & \text{si } \hat{r}_n(x) \leq 0. \end{cases}$$

3.1.2 Propriétés de l'estimateur

Le résultat suivant est montré dans Gill et Johansen.

Lemme 3.1.1 *On a*

$$\sup_{t \leq T_Y} |\hat{H}_n(t) - H(t)| \to 0 \quad n \to \infty \ p.s$$

Remarquons que Kohler et Máthé ont utilisé le résultat de (Stute et Wang (1993)) qui exige la continuité de H pour avoir le résultat précédent.

Lemme 3.1.2 *Sous l'hypothèse* $(H_{1.2})$, *on a*

$$\int_{\mathbb{R}^d} |r_n(x) - r(x)|^2 \mu(dx) \to 0 \quad (n \to \infty) \ p.s.$$

si et seulement si

$$\int_{\mathbb{R}^d} |r_n^*(x) - r(x)|^2 \mu(dx) \to 0 \qquad (n \to \infty) \ \ p.s.$$

Preuve $|r_n^*(x) - r(x)|^2 = |r_n^*(x) - r_n(x) + r_n(x) - r(x)|^2$

$$\leqslant 2|r_n^*(x) - r_n(x)|^2 + 2|r_n(x) - r(x)|^2$$

$$\int_{\mathbb{R}^d} |r_n^*(x) - r(x)|^2 \mu(dx) \leqslant 2\int_{\mathbb{R}^d} |r_n^*(x) - r_n(x)|^2 \mu(dx) + 2\int_{\mathbb{R}^d} |r_n(x) - r(x)|^2 \mu(dx).$$

On a : $\int_{\mathbb{R}^d} |r_n(x) - r(x)|^2 \mu(dx) \to 0 \quad (n \to \infty) \ \ p.s.$

Donc il suffit de montrer que : $\int_{\mathbb{R}^d} |r_n^*(x) - r_n(x)|^2 \mu(dx) \to 0 \quad (n \to \infty) \ \ p.s.$

On a : $T_{K,n} \leqslant T_Y$ p.s.

$$|r_n^*(x) - r_n(x)| = |T_{[0,T_Y]}\hat{r}_n(x) - T_{[0,T_{K,n}]}\hat{r}_n(x)|$$

$$\leqslant T_Y - T_{K,n}$$

$$\int_{\mathbb{R}^d} |r_n^*(x) - r_n(x)|^2 \mu(dx) \leqslant \int_{\mathbb{R}^d} (T_Y - T_{K,n})^2 \mu(dx)$$
$$= (T_Y - T_{K,n})^2,$$

or $T_{K,n} \to T_Y \quad (n \to \infty) \ \ p.s.$ ■

Théorème 3.1.1 *(Kohler, Màthé et Pintér (2002))*

Sous l'hypothèse (H_1) *et si* K *est un noyau régulier,* $\lim_{n\to\infty} h_n = 0$, $\lim_{n\to\infty} nh_n^d = \infty$, *alors l'estimateur* $r_n(x)$ *défini par* (3.1), (3.2) *vérifie :*

$$\int_{\mathbb{R}^d} |r_n(x) - r(x)|^2 \mu(dx) \to 0 \quad (n \to \infty) \ p.s.$$

Preuve D'après le lemme précédent, il suffit de montrer que :

$$\int_{\mathbb{R}^d} |r_n^*(x) - r(x)|^2 \mu(dx) \to 0 \ \ (n \to \infty) \ p.s.$$

Posons

$$\bar{r}_n(x) = T_{[0,T_Y]} \left(\sum_{i=1}^{n} \frac{K(\frac{(x-X_i)}{h_n})}{\sum_{j=1}^{n} K(\frac{(x-X_j)}{h_n})} \frac{\delta_i Z_i}{H(Z_i)} \right).$$

$$|r_n^*(x) - r(x)|^2 = |r_n^*(x) - \bar{r}_n(x) + \bar{r}_n(x) - r(x)|^2$$

$$\leqslant 2|r_n^*(x) - \bar{r}_n(x)|^2 + 2|\bar{r}_n(x) - r(x)|^2$$

$$\int_{\mathbb{R}^d} |r_n^*(x) - r(x)|^2 \, \mu(dx) \leq 2 \int_{\mathbb{R}^d} |r_n^*(x) - \bar{r}_n(x) \mu(dx) + 2 \int_{\mathbb{R}^d} |\bar{r}_n(x) - r(x)|^2 \mu(dx). \tag{3.3}$$

Commençons par majorer le second terme de l'inégalité précédente.

$$\int_{\mathbb{R}^d} |\bar{r}_n(x) - r(x)|^2 \mu(dx) \leqslant \int_{\mathbb{R}^d} |\sum_{i=1}^{n} \frac{K(\frac{(x-X_i)}{h_n})}{\sum_{j=1}^{n} K(\frac{(x-X_j)}{h_n})} \cdot \frac{\delta_1 Z_1}{H(Z_1)} - r(x)|^2 \mu(dx).$$

De plus $0 \leqslant \delta_1 Z_1 / H(Z_1) \leqslant T_Y / H(T_Y)$ *p.s.* et

$$\begin{aligned} E\left\{ \frac{\delta_1 Z_1}{H(Z_1)} / X_1 \right\} &= E\left\{ \frac{I_{\{Y_1 \leq C_1\}} Y_1}{H(Y_1)} / X_1 \right\} \\ &= E\left(\frac{Y_1}{H(Y_1)} E\left(I_{\{Y_1 \leq C_1\}} / (X_1, Y_1) \right) / X_1 \right) \\ &= E\left(Y_1 / X_1 \right) = r(X_1). \end{aligned}$$

Donc d'aprés le théorème de (Devroye et Krzyżak (1989)) on obtient

$$\int_{\mathbb{R}^d} |\bar{r}_n(x) - r(x)|^2 \mu(dx) \leqslant$$

$$\int_{\mathbb{R}^d} |\sum_{i=1}^{n} \frac{K(\frac{(x-X_i)}{h_n})}{\sum_{j=1}^{n} K(\frac{(x-X_j)}{h_n})} \cdot \frac{\delta_1 Z_1}{H(Z_1)} - r(x)|^2 \mu(dx) \to 0 \ (n \to \infty) \ p.s \tag{3.4}$$

Reste à majorer le premier terme de l'inégalité donnée à la formule (3.3)

$$
\begin{aligned}
&\int_{\mathbb{R}^d} |r_n^*(x) - \bar{r}_n(x)|^2 \mu(dx) \\
\leq\ & T_Y \int_{\mathbb{R}^d} |\sum_{i=1}^{n} \frac{K(\frac{(x-X_i)}{h_n})}{\sum_{j=1}^{n} K(\frac{(x-X_j)}{h_n})} \frac{\delta_1 Z_1}{\hat{H}_n(Z_1)} - \sum_{i=1}^{n} \frac{K(\frac{(x-X_i)}{h_n})}{\sum_{j=1}^{n} K(\frac{(x-X_j)}{h_n})} \frac{\delta_1 Z_1}{H(Z_1)} | \mu(dx) \\
\leq\ & T_Y \int_{\mathbb{R}^d} \sum_{i=1}^{n} \frac{K(\frac{(x-X_i)}{h_n})}{\sum_{j=1}^{n} K(\frac{(x-X_j)}{h_n})} T_Y | \frac{1}{\hat{H}_n(Zi)} - \frac{1}{H(Z_i)} | \mu(dx) \\
\leq\ & T_Y^2 \frac{1}{\hat{H}_n(T_Y) H(T_Y)} \sup_{t \leq T_Y} |H(t) - \hat{H}_n(t)| \to 0 \quad (n \to \infty)\ p.s.
\end{aligned}
$$

à cause de $(H_{1,2})$et du lemme 3.1.1. ■

Signalons le fait que Guessoum et Ould said (2009) ont modifié légérement $\hat{H}_n$ en lui imposant de s'annuler à partir de l'observation la plus grande. Ils ont alors, d'une part établi la convergence presque sûre uniforme sur des compacts de l'estimateur ainsi obtenu, donné des vitesses de convergence et prouvé d'autre part sa normalité asymptotique.

Nous nous devons aussi de faire remarquer que le résultat donné au théorème précédent a été aussi prouvé dans Kohler, Màthé et Pintér (2002) pour des estimateurs à poids (plus proches voisins et à partitions) dans un modèle de censure à droite.

Dans la suite de ce chapitre, en nous inspirant de la démarche précédente, nous proposons un résultat similaire dans un modèle de censure à gauche.

3.2 Censure à gauche

Dans ce cas, l'évenement d'intérêt peut se produire avant le début de l'étude, au temps C_i. Si c'est le cas, on ne connaît pas le moment exact de l'événement.

Donc les observations peuvent être représentées par le maximum entre Y et une variable de censure C.

Plus précisément soit Y une variable d'intérêt positive et bornée et C une variable aléatoire de censure positive.

Nous observons l'échantillon $\mathcal{D}_n = \{(X_i, Z_i, \delta_i), i = 1, ..., n\}$, où $Z_i = Y_i \vee C_i$ et $\delta_i = I_{\{Y_i \geq C_i\}}$ (δ_i est l'indicatrice de censure).

Nous nous proposons d'estimer $r(x) = E\left(Y/X = x\right)$.

3.2.1 Définition de l'estimateur

Soient $F(t) = P(Y \leq t)$ et $G(t) = P(C \leq t)$ les fonctions de répartition respectives de Y et C.

On a besoin de l'hypothèse (H_3) englobant les hypothèses suivantes

$$(H_3) : \begin{cases} (H_{3.1}) & C \text{ et } (X, Y) \text{ sont indépendants,} \\ (H_{3.2}) & T_Y \vee T_C < \infty \text{ et } G(I_Y) > 0. \end{cases}$$

Ici la condition $G(I_Y) > 0$ implique $I_C \leq I_Y$.

Soit l une fonction de $\mathbb{R}^d X\ \mathbb{R} \to \mathbb{R}$. On se propose d'estimer la moyenne $E\left\{l(X, Y)\right\}$ à partir de l'échantillon $\mathcal{D}_n = \{(Z_i, \delta_i, X_i), i = 1, ..., n\}$.

Sous l'hypothèse d'indépendance, un "estimateur" sans biais de cette quantité est donné par :

$$E(l(X, Y)) = \frac{1}{n} \sum_{i=1}^{n} \frac{\delta_i l(X_i, Z_i)}{G(Z_i)}.$$

En effet,

$$\begin{aligned} E\left\{\frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i l(X_i,Z_i)}{G(Z_i)}\right\} &= E\left\{\frac{\delta_1 l(X_1,Z_1)}{G(Z_1)}\right\} \\ &= E\left\{\frac{1_{\{Y_1\geq C_1\}} l(X_1,Y_1)}{G(Z_1)}\right\} \\ &= E\left\{\frac{l(X_1,Y_1)}{G(Z_1)} E\left[1_{\{Y_1\geq C_1\}}\backslash X_1,Y_1\right]\right\} \\ &= E(l(X_1,Y_1)). \end{aligned}$$

En pratique, G est inconnue. On ne peut donc pas utiliser "l'estimateur" tel quel, il est alors naturel de remplacer G par un estimateur, obtenu par adaptation au contexte de censure à gauche de l'estimateur de Kaplan Meier et donné par

$$\hat{G}_n(t) = \begin{cases} \prod_{i=1}^{n}\left[1-\frac{1-\delta_{(i)}}{i}\right]^{I_{\left[Z_{(i)}>t\right]}}, & \text{si } t > T_{L,n} \\ \prod_{i=1}^{n}\left[1-\frac{1-\delta_{(i)}}{i}\right]^{I_{\left[Z_{(i)}>T_{L,n}\right]}}, & \text{si } t \leq T_{L,n} \end{cases},$$

où $T_{L,n} = \inf\{Z_1,\ldots,Z_n\}$. Remarquons que cet estimateur ne s'annule jamais.

On peut se reporter à Kebabi et Messaci (2009) pour des détails sur cet estimateur, il y est en particulier montré le résultat suivant.

Lemme 3.2.1 *On a*

$$\sup_{t\geq I_Y} |\hat{G}_n(t) - G(t)| \to 0 \quad n\to\infty \ p.s$$

Un estimateur possible de $E(l(X,Y))$ est alors donné par

$$E(l(X,Y)) = \frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i Z_i}{\hat{G}_n(Z_i)}.$$

Cela suggère d'estimer $r(x)$ par :

$$\hat{r}_{n,g}(x) = \sum_{i=1}^{n} \frac{K(\frac{(x-X_i)}{h_n})}{\sum_{j=1}^{n} K(\frac{(x-X_j)}{h_n})} \frac{\delta_i Z_i}{\hat{G}_n(Z_i)}. \tag{3.5}$$

Du fait que $0 \leqslant Y \leqslant T_Y < \infty$ p.s, on a $0 \leqslant r(x) \leqslant T_Y$, on l'estime donc plutôt par

$$r_{n,g}(x) = \begin{cases} Z_{(n)}, & \text{si } \hat{r}_{n,g}(x) > Z_{(n)}, \\ \hat{r}_n(x), & \text{si } 0 \leqslant \hat{r}_{n,g}(x) \leqslant Z_{(n)}, \\ 0, & \text{si } \hat{r}_{n,g}(x) < 0. \end{cases} , \tag{3.6}$$

où $Z_{(n)} = \max(Z_1, \ldots, Z_n)$.

Par analogie avec (3.6), posons

$$r^*_{n,g}(x) = T_{[0,T_Y \vee T_C]}(\hat{r}_{n,g}(x)) = \begin{cases} T_Y \vee T_C, & \text{si } \hat{r}_{n,g}(x) > T_Y \vee T_C, \\ \hat{r}_{n,g}(x), & \text{si } 0 \leqslant \hat{r}_{n,g}(x) \leqslant T_Y \vee T_C, \\ 0, & \text{si } \hat{r}_{n,g}(x) < 0. \end{cases}$$

3.2.2 Propriétés de l'estimateur

Lemme 3.2.2 *Sous l'hypothèse* $(H_{3,2})$*, on a*

$$\int_{\mathbb{R}^d} |r_{n,g}(x) - r(x)|^2 \mu(dx) \to 0 \quad (n \to \infty) \quad p.s.$$

si et seulement si

$$\int_{\mathbb{R}^d} |r^*_{n,g}(x) - r(x)|^2 \mu(dx) \to 0 \quad (n \to \infty)\ p.s.$$

Preuve $|r^*_{n,g}(x) - r(x)|^2 = |r^*_{n,g}(x) - r_{n,g}(x) + r_{n,g}(x) - r(x)|^2$

$$\leqslant 2|r^*_{n,g}(x) - r_{n,g}(x)|^2 + 2|r_{n,g}(x) - r(x)|^2$$

$$\int_{\mathbb{R}^d} |r^*_{n,g}(x) - r(x)|^2 \mu(dx) \leqslant 2 \int_{\mathbb{R}^d} |r^*_{n,g}(x) - r_{n,g}(x)|^2 \mu(dx) + 2 \int_{\mathbb{R}^d} |r_{n,g}(x) - r(x)|^2 \mu(dx).$$

Il suffit de montrer que

$$\int_{\mathbb{R}^d} |r^*_{n,g}(x) - r_{n,g}(x)|^2 \mu(dx) \to 0 \quad (n \to \infty) \ p.s.$$

$$\begin{aligned} |r^*_{n,g}(x) - r_{n,g}(x)| &= |T_{[0,T_Y \vee T_C]} \hat{r}_{n,g}(x) - T_{[0,Z_{(n)}]} \hat{r}_{n,g}(x)| \\ &\leqslant T_Y \vee T_C - Z_{(n)} \end{aligned}$$

$$\begin{aligned} \int_{\mathbb{R}^d} |r^*_{n,g}(x) - r_{n,g}(x)|^2 \mu(dx) &\leqslant \int_{\mathbb{R}^d} (T_Y \vee T_C - Z_{(n)})^2 \mu(dx) \\ &= (T_Y \vee T_C - Z_{(n)})^2 \end{aligned}$$

et $Z_{(n)} \to T_Y \vee T_C \quad (n \to \infty) \ \ p.s.$ Alors :

$$\int_{\mathbb{R}^d} |r^*_{n,g}(x) - r_{n,g}(x)|^2 \mu(dx) \to 0 \quad (n \to \infty) \ p.s.$$

D'ou le résultat. ■

Théorème 3.2.1 *Sous l'hypothèse* (H_3) *et si* K *est un noyau régulier,* $\lim_{n\to\infty} h_n = 0$, $\lim_{n\to\infty} nh_n^d = \infty$,

alors l'estimateur $r_{n,g}$ *défini par* (3.5) *et* (3.6) *vérifie :*

$$\int_{\mathbb{R}^d} |r_{n,g}(x) - r(x)|^2 \mu(dx) \to 0 \quad (n \to \infty) \ p.s.$$

Preuve D'après le lemme précédent, il suffit de montrer que

$$\int_{\mathbb{R}^d} |r^*_{n,g}(x) - r(x)|^2 \mu(dx) \to 0 \quad (n \to \infty) \ p.s.$$

Posons

$$\bar{r}_{n,g}(x) = T_{[0,T_Y \vee T_C]} \left(\sum_{i=1}^{n} \frac{K(\frac{(x-X_i)}{h_n})}{\sum_{j=1}^{n} K(\frac{(x-X_j)}{h_n})} . \frac{\delta_i Z_i}{G(Z_i)} \right).$$

On a

$$\int_{\mathbb{R}^d} \left|r^*_{n,g}(x) - r(x)\right|^2 \mu(dx) \leq 2\int_{\mathbb{R}^d} |r^*_{n,g}(x) - \bar{r}_{n,g}(x)\mu(dx) + 2\int_{\mathbb{R}^d} |\bar{r}_{n,g}(x) - r(x)|^2\mu(dx). \qquad (3.7)$$

Commençons par majorer le second terme de l'inégalité précédente.

Remarquons que $\int_{\mathbb{R}^d} |\bar{r}_{n,g}(x) - r(x)|^2\mu(dx) \leqslant \int_{\mathbb{R}^d} |\sum_{i=1}^n \frac{K(\frac{(x-X_i)}{h_n})}{\sum_{j=1}^n K(\frac{(x-X_j)}{h_n})} . \frac{\delta_1 Z_1}{G(Z_1)} - r(x)|^2\mu(dx).$

De plus $0 \leqslant \delta_1 Z_1/G(Z_1) \leqslant T_Y \vee T_G/G(I_Y)$ $p.s.$

et

$$\begin{aligned} E\left\{\frac{\delta_1 Z_1}{G(Z_1)}/X_1\right\} &= E\left\{\frac{I_{\{Y_1\geq C_1\}}Y_1}{G(Y_1)}/X_1\right\} \\ &= E\left(\frac{Y_1}{G(Y_1)}E\left(I_{\{Y_1\geq C_1\}}/(X_1,Y_1)\right)/X_1\right) \\ &= E\left(Y_1/X_1\right) \\ &= r(X_1). \end{aligned}$$

Le théorème de (Devroye et Krzyżak (1989)) est donc appliquable et on obtient

$$\int_{\mathbb{R}^d} |\sum_{i=1}^n \frac{K(\frac{(x-X_i)}{h_n})}{\sum_{j=1}^n K(\frac{(x-X_j)}{h_n})} . \frac{\delta_1 Z_1}{G(Z_1)} - r(x)|^2\mu(dx) \to 0 \ (n \to \infty) \ p.s.$$

Il reste à majorer le premier terme de l'inégalité donnée à la formule (3.7)

$$\begin{aligned} &\int_{\mathbb{R}^d} |r^*_{n,g}(x) - \bar{r}_{n,g}(x)|^2\mu(dx) \\ \leq\ & T_Y \vee T_C \int_{\mathbb{R}^d} |\sum_{i=1}^n \frac{K(\frac{(x-X_i)}{h_n})}{\sum_{j=1}^n K(\frac{(x-X_j)}{h_n})} . \frac{\delta_1 Z_1}{\hat{G}_n(Z_1)} - \sum_{i=1}^n \frac{K(\frac{(x-X_i)}{h_n})}{\sum_{j=1}^n K(\frac{(x-X_j)}{h_n})} . \frac{\delta_1 Z_1}{G(Z_1)}|\mu(dx) \\ \leq\ & T_Y \vee T_C \int_{\mathbb{R}^d} \sum_{i=1}^n \frac{K(\frac{(x-X_i)}{h_n})}{\sum_{j=1}^n K(\frac{(x-X_j)}{h_n})} . T_Y \vee T_C |\frac{1}{\hat{G}_n(Z_1)} - \frac{1}{G(Z_1)}|\mu(dx) \\ \leq\ & (T_Y \vee T_C)^2 \frac{1}{\hat{G}_n(I_Y)G(I_Y)} \sup_{t>I_Y} |\hat{G}_n(t) - G(t)| \to 0 \ \ (n \to \infty) \ p.s., \end{aligned}$$

par le lemme 3.2.1 et par le fait que $G(I_Y) > 0$ et $\hat{G}_n(I_Y) > 0$. ■

Remarquons que contrairement au cas de la censure à droite, nous n'avons pas utilisé l'hypothèse de la continuité de G.

Chapitre 4

Régression fonctionnelle

4.1 Introduction

Définition 4.1.1 *Une variable aléatoire est dite fonctionnelle si elle est à valeurs dans un espace de dimension infinie.*

4.1.1 Motivation

De très nombreux travaux concernent l'étude de modèles sur des variables aléatoires réelles, censurées, multivariées et c'est un domaine de la statistique toujours très étudié. Cependant, les récentes innovations réalisées sur les appareils de mesure et les méthodes d'acquisition ainsi que l'utilisation intensive de moyens informatiques permettent souvent de récolter des données discrétisées sur des grilles de plus en plus fines, ce qui les rend intrinsèquement fonctionnelles. Les courbes de croissance, les enregistrements sonores, les images satellites, les séries chronologiques, les courbes spectrométriques sont des exemples de données de nature fonctionnelle que le statisticien peut être amené à étudier. C'est pourquoi un nouveau champ de la statistique est

dédié à l'étude de données fonctionnelles.

Ce travail est basé sur les résultats de Ferraty et Vieu (2000).

4.1.2 Présentation du modèle

Nous allons nous intéresser dans ce chapitre à l'estimation de la fonction de régression

$r(x) = E(Y/X = x)$

où Y est une variable aléatoire réelle et X est une variable aléatoire à valeurs dans un espace vectoriel semi normé $(H, ||.||)$, X et Y étant définies sur le même espace de probabilité $(\Omega, \mathcal{A}, P)$.

Dans la suite, il apparaîtra l'importance de mesurer la proximité des observations, ce qui se fera par l'utilisation de la semi norme de H.

L'hypothèse imposée à la variable fonctionnelle concerne la dimension fractale de sa mesure de probabilité, dont voici la définition.

Définition 4.1.2 *La dimension fractale de la variable fonctionnelle X (cf. eg. Bardet (1997) Page 34), est le réel positif $\delta(x)$ donnée par*

$$\lim_{\alpha \to 0^+} \frac{\log P(X \in B(x, \alpha))}{\log \alpha} = \delta(x)$$

où $B(x, \alpha)$ est la boule de centre x et de rayon α, associée à la semi-norme $||.||$.

4.2 Estimation

4.2.1 Présentation de l'estimateur à noyau fonctionnel

Soit $(X_i, Y_i)_{i=1,...,n}$ un échantillon de n observations indépendantes du couple(X, Y). L'estimateur $\hat{r}_n(x)$ de $r(x) = E(Y/X = x)$ a été défini par Ferraty et Vieu, en adaptant au cas

fonctionnel l'estimateur classique de Nadaraya-Watson, en posant

$$\hat{r}_n(x) = \sum_{i=1}^{n} W_{i,n}(x) Y_i,$$

avec

$$W_{i,n}(x) = \frac{K\left(\frac{||X_i - x||}{h_n}\right)}{\sum_{i=1}^{n} K\left(\frac{||X_i - x||}{h_n}\right)},$$

où h_n est une suite de nombres positifs (fenêtre) et K est un noyau décroissant permettant d'attribuer aux observations X_i les plus proches de x, les poids les plus grands.

4.2.2 Propriétés de l'estimateur

Introduisons les hypothèses nécessaires à l'obtention des résultats asymptotiques.

(H_4)

$$\lim_{\alpha \to 0^+} \frac{P(X \in B(x, \alpha))}{\alpha^{\delta(x)}} = c(x),$$

où $\delta(x)$ et $c(x)$ sont deux réels strictement positifs et $B(x, \alpha)$ désigne la boule de centre x et de rayon α. Remarquons que $\delta(x)$ est la dimension fractale de la variable X.

(H_5)englobe les 3 hypothèses $(H_{5.1})$, $(H_{5.2})$ et $(H_{5.3})$.

$\cdot(H_{5.1})$ la largeur de la fenêtre h_n est telle que :

$$\lim_{n \to \infty} h_n = 0 \quad \text{et} \quad \lim_{n \to \infty} \frac{n h_n^{\delta(x)}}{\log n} = \infty,$$

$\cdot(H_{5.2})$ le noyau K est Lipschitzien d'ordre 1 et de support $[0, \zeta]$ avec $\zeta \in \mathbb{R}_*^+$,

$\cdot(H_{5.3})$ la variable aléatoire réelle Y est bornée presque sûrement.

(H_6) $P(X \in B(x, \alpha)) = \alpha^{\delta(x)} c(x) + O(\alpha^{\delta(x)+b(x)})$,

où $b(x)$ est un réel strictement positif.

Théorème 4.2.1 *Soit x un point fixé de H. Sous les hypothèses $(H_4), (H_5)$ et si r est continue en x, alors on a :*

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

Preuve Posons

$$\hat{f}_n(x) = \frac{1}{nh_n^{\delta(x)}} \sum_{i=1}^{n} K\left(\frac{||X_i - x||}{h_n}\right) \text{ et } \hat{g}_n(x) = \frac{1}{nh_n^{\delta(x)}} \sum_{i=1}^{n} Y_i K\left(\frac{||X_i - x||}{h_n}\right)$$

de sorte que

$$\hat{r}_n(x) = \frac{\hat{g}_n(x)}{\hat{f}_n(x)}.$$

La démonstration de ce résultat est alors basée sur la décomposition suivante :

$$\hat{r}_n(x) - r(x) = \frac{1}{\hat{f}_n(x)} \{\hat{g}_n(x) - r(x)C_\delta(x)\} - \frac{r(x)}{\hat{f}_n(x)} \left\{\hat{f}_n(x) - C_\delta(x)\right\}, \tag{4.1}$$

où $C_\delta(x)$ est un réel strictement positif (dépendant de x ainsi que de la dimension fractale $\delta(x)$ qui sera précisé en cours de la démonstration). ■

Afin de démontrer ce théorème nous avons besoin des lemmes suivants :

Lemme 4.2.1 *Sous les hypothèses $(H_{5.1})$ et $(H_{5.2})$, et si*

i) (H_4) est vérifiée, alors on a :

$$\lim_{n\to\infty} E\hat{f}_n(x) = C_\delta(x). \tag{4.2}$$

ii) (H_6) est vérifiée, alors on a :

$$E\hat{f}_n(x) = C_\delta(x) + O(h_n^{b(x)})$$

Preuve *i*) Comme K est Lipschitzien d'ordre 1, K est en particulier continu. On peut donc écrire :

$$K(t) = \lim_{J\to\infty} K_J(t) \text{ où } K_J(t) = \sum_{j=0}^{J-1} K(j\zeta/J) I_{[j\zeta/J,(j+1)\zeta/J]}.$$

D'où

$$E\hat{f}_n(x) = \frac{1}{nh_n^{\delta(x)}} \sum_{i=1}^{n} E\left\{ \lim_{J\to\infty} K_J\left(\frac{||X_i - x||}{h_n}\right)\right\} = \frac{1}{h_n^{\delta(x)}} E\left\{ \lim_{J\to\infty} K_J\left(\frac{||X-x||}{h_n}\right)\right\}.$$

Le théorème de la convergence dominée permet alors d'écrire

$$E\hat{f}_n(x) = \frac{1}{h_n^{\delta(x)}} \lim_{J\to\infty} E\left\{ K_J\left(\frac{||X-x||}{h_n}\right)\right\}$$

$$= \lim_{J\to\infty} \sum_{j=0}^{J-1} K(j\zeta/J) \left\{ \left(\frac{(j+1)\zeta}{J}\right)^{\delta(x)} \frac{P(X\in B(x, h_n(j+1)\zeta/J))}{\left(\frac{(j+1)\zeta h_n}{J}\right)^{\delta(x)}} - \left(\frac{j\zeta}{J}\right)^{\delta(x)} \frac{P(X\in B(x,h_n j\zeta/J))}{\left(\frac{j\zeta h_n}{J}\right)^{\delta(x)}} \right\}.$$

Posons

$$C_\delta(x) = \lim_{J\to\infty} c(x) \sum_{j=0}^{J-1} K(j\zeta/J) \left\{ \left(\frac{(j+1)\zeta}{J}\right)^{\delta(x)} - \left(\frac{j\zeta}{J}\right)^{\delta(x)} \right\}.$$

Il vient

$$E\hat{f}_n(x) - C_\delta(x) = \lim_{J\to\infty} \sum_{j=0}^{J-1} K(j\zeta/J) \left\{ \left(\frac{(j+1)\zeta}{J}\right)^{\delta(x)} \left(\frac{P(X\in B(x,h_n(j+1)\zeta/J))}{\left(\frac{(j+1)\zeta h_n}{J}\right)^{\delta(x)}} - c(x)\right) - \left(\frac{j\zeta}{J}\right)^{\delta(x)} \left(\frac{P(X\in B(x,h_n j\zeta/J))}{\left(\frac{j\zeta h_n}{J}\right)^{\delta(x)}} - c(x)\right)\right\} \tag{4.3}$$

D'où, grâce à l'hypothèse (H_4)

$E\hat{f}_n(x) - C_\delta(x) =$

$$\lim_{J\to\infty} \sum_{j=0}^{J-1} \left\{ \left(K((j+1)\zeta/J) \left(\tfrac{(j+1)\zeta}{J}\right)^{\delta(x)} - K(j\zeta/J) \left(\tfrac{j\zeta}{J}\right)^{\delta(x)} \right) o(1) \right\}$$
$$- \lim_{J\to\infty} \sum_{j=0}^{J-1} \left\{ \left(K((j+1)\zeta/J) - K(j\zeta/J)\right) \left(\tfrac{(j+1)\zeta}{J}\right)^{\delta(x)} o(1) \right\}.$$

La première quantité de la partie droite de cette inégalité tend vers 0 (les termes de la somme s'annullent 2 à 2). De plus, K Lipschitzien d'ordre 1 implique :

$$\left| \sum_{j=0}^{J-1} \left\{ \left(K((j+1)\zeta/J) - K(j\zeta/J)\right) \left(\tfrac{(j+1)\zeta}{J}\right)^{\delta(x)} o(1) \right\} \right|$$
$$\leq \zeta^{\delta(x)} \sum_{j=0}^{J-1} \left|\left(K((j+1)\zeta/J) - K(j\zeta/J)\right)\right| o(1)$$
$$\leq \zeta^{\delta(x)} \sum_{j=0}^{J-1} \tfrac{\zeta}{J} o(1).$$

Conclusion : pour tout $\varepsilon > 0$, il existe un rang n_0 à partir duquel

$$|E\hat{f}_n(x) - C_\delta(x)| < \varepsilon.$$

Il suffit maintenant de préciser $C_\delta(x)$: en posant $H(u) = K(u^{\frac{1}{\delta(x)}})$, on obtient

$$C_{\delta(x)} = c(x) \int_0^{\zeta^{\delta(x)}} H(u)du = c(x)\delta(x) \int_0^{\zeta} K(v) v^{\delta(x)-1} dv$$

ii) Les conditions (4.3) et (H_6) permettent d'écrire :

$$|E\hat{f}_n(x) - C_\delta(x)| \leq$$
$$\lim_{J\to\infty} \left| \sum_{j=0}^{J-1} \left\{ \left(K((j+1)\zeta/J) \left(\tfrac{(j+1)\zeta}{J}\right)^{\delta(x)} - K(j\zeta/J) \left(\tfrac{j\zeta}{J}\right)^{\delta(x)} \right) O(h_n^{b(x)}) \right\} \right|$$
$$+ \lim_{J\to\infty} \left| \sum_{j=0}^{J-1} \left\{ \left(K((j+1)\zeta/J) - K(j\zeta/J)\right) \left(\tfrac{(j+1)\zeta}{J}\right)^{\delta(x)} O(h_n^{b(x)}) \right\} \right|.$$

Il suffit alors de reprendre les arguments développés précédemment pour conclure. ■

Lemme 4.2.2 *Sous les hypothèses* (H_4), $(H_{5.1})$ *et* $(H_{5.2})$ *on a :*

$$\hat{f}_n(x) - E\hat{f}_n(x) = O\left(\sqrt{\frac{\log n}{n h_n^{\delta(x)}}}\right) \quad p.co. \tag{4.4}$$

Preuve On a

$$\hat{f}_n(x) - E\hat{f}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \Delta_{i,x},$$

où

$$\Delta_{i,x} = \frac{1}{h_n^{\delta(x)}}\left\{K\left(\frac{\|X_i - x\|}{h_n}\right) - EK\left(\frac{\|X_i - x\|}{h_n}\right)\right\}.$$

Il est clair que l'on a $|\Delta_{i,x}| \leq C_1/h_n^{\delta(x)}$.

De plus, en reprenant les calculs faits précédemment, on a, d'une part,

$$\lim_{n\to\infty}\left\{\frac{1}{h_n^{\delta(x)}}EK\left(\frac{\|X - x\|}{h_n}\right)\right\}^2 = \lim_{n\to\infty}\left(E\hat{f}_n(x)\right)^2 = C_\delta^2(x),$$

et d'autre part, en remplaçant K par K^2,

$$\lim_{n\to\infty}\frac{1}{h_n^{\delta(x)}}K^2\left(\frac{\|X - x\|}{h_n}\right) = D_\delta(x),$$

où

$$D_\delta(x) = c(x)\delta(x)\int_0^\zeta K^2(v)v^{\delta(x)-1}dv.$$

Comme

$$E\Delta_{i,x}^2 = \frac{1}{h_n^{\delta(x)}}\left\{\frac{1}{h_n^{\delta(x)}}EK^2\left(\frac{\|X - x\|}{h_n}\right)\right\} - \left\{\frac{1}{h_n^{\delta(x)}}EK\left(\frac{\|X - x\|}{h_n}\right)\right\}^2,$$

il vient que

$$E\Delta_{i,x}^2 \ \leq C_2\frac{1}{h_n^{\delta(x)}}.$$

On obtient alors en appliquant le Lemme (1.1)

$$P\left(\left|\hat{f}_n\ (x) - E\hat{f}_n(x)\right| > \eta\sqrt{\frac{\log n}{nh_n^{\delta(x)}}}\right) \leq C_3 \exp\left\{-n\eta^2 \frac{\log n}{nh_n^{\delta(x)}} \frac{h_n^{\delta(x)}}{C_4}\right\},$$

ce qui équivaut à

$$P\left(\left|\hat{f}_n(x) - E\hat{f}_n(x)\right| > \eta\sqrt{\frac{\log n}{nh_n^{\delta(x)}}}\right) \leq Cn^{-C\eta^2}.$$

Ainsi, il existe η_0 tel que

$$\sum_{n=1}^{\infty} P\left[\left|\hat{f}_n(x) - E\hat{f}_n(x)\right| > \eta_0\sqrt{\frac{\log n}{nh_n^{\delta(x)}}}\right] < \infty.$$

■

Lemme 4.2.3 *Les conditions* (H_4), $(H_{5.1})$, $(H_{5.2})$ *et la continuité de* r *entraînent*

$$\lim_{n\to\infty} E\hat{g}_n(x) = r(x)C_\delta(x). \tag{4.5}$$

Preuve On a

$$\hat{g}_n(x) = \frac{1}{nh_n^{\delta(x)}} \sum_{i=1}^{n} Y_i K\left(\frac{||X_i - x||}{h_n}\right),$$

et

$$\begin{aligned} E\hat{g}_n(x) &= \frac{1}{h_n^{\delta(x)}} E\left(YK\left(\frac{||X-x||}{h_n}\right)\right), \\ &= \frac{1}{h_n^{\delta(x)}} E\left(r(X)K\left(\frac{||X-x||}{h_n}\right)\right). \end{aligned}$$

Ceci se reécrit

$$E\hat{g}_n(x) = r(x)E\hat{f}_n(x) + \frac{1}{h_n^{\delta(x)}} E\left(\left(r(X) - r(x)\right)K\left(\frac{\|X-x\|}{h_n}\right)\right),$$

et implique que

$$|E\hat{g}_n(x) - r(x)C_\delta(x)| \leq |r(x)| \left| E\hat{f}_n(x) - C_\delta(x) \right| + \frac{1}{h_n^{\delta(x)}} E\left(|r(X) - r(x)| \left| K\left(\frac{\|X - x\|}{h_n} \right) \right| \right).$$

Par le lemme(4.2.1), le première terme de la partie droite de cette inégalité tend vers 0. Le second terme tend vers 0 à cause de la continuité de r et par application une nouvelle fois du lemme (4.2.1) au noyau $|K|$. ■

Lemme 4.2.4 *Sous les conditions* (H_4) *et* (H_5) *on a*

$$\hat{g}_n(x) - E\hat{g}_n(x) = O\left(\sqrt{\frac{\log n}{n h_n^{\delta(x)}}} \right) \ p.co. \tag{4.6}$$

Preuve Utilisons la décomposition suivante

$$\hat{g}_n(x) - E\hat{g}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \Delta_{i,x}^{'} \ ,$$

où

$$\Delta_{i,x}^{'} = \frac{1}{h_n^{\delta(x)}} \left\{ Y_i K\left(\frac{||X_i - x||}{h_n} \right) - E\left(Y_i K\left(\frac{||X_i - x||}{h_n} \right) \right) \right\}.$$

Il est clair, puisque Y est bornée, qu'en reprenant les arguments développés pour prouver (4.4), on a

$$|\Delta_{i,x}^{'}| \leq C \frac{1}{h_n^{\delta(x)}} \quad \text{et} \quad E\Delta_{i,x}^{'2} \leq C \frac{1}{h_n^{\delta(x)}}.$$

Par analogie avec la fin de la preuve de (4.4), il suffit d'utiliser le Lemme (1.1)

$$P\left(|\hat{g}_n\ (x) - E\hat{g}_n\ (x)| > \eta \sqrt{\frac{\log n}{n h_n^{\delta(x)}}} \right) \leq C n^{-C\eta^2}.$$

Ainsi, il existe η_0 tel que

$$\sum_{n=1}^{\infty} P\left[|\hat{g}_n\ (x) - E\hat{g}_n\ (x)| > \eta \sqrt{\frac{\log n}{n h_n^{\delta(x)}}} \right] < \infty.$$

■

Enfin, pour achever la démonstration du théorème (4.2.1), il suffit de montrer qu'il existe un réel τ strictement positif tel que

$$\sum_{n=1}^{\infty} P\left(\left|\hat{f}_n(x)\right| \leq \tau\right) < \infty. \tag{4.7}$$

Or les résultats (4.2) et (4.4) entraînent en particulier la convervenge presque complète de $\hat{f}_n(x)$ vers $C_\delta(x)$.

Ainsi pour tout $\varepsilon > 0$ on a

$$\sum_{n=1}^{\infty} P\left[\left|\hat{f}_n\ (x) - C_\delta(x)\right| > \varepsilon\right] < \infty.$$

En remarquant que

$$\left|\hat{f}_n(x)\right| \leq \frac{C_\delta(x)}{2} \Rightarrow \left|\hat{f}_n(x) - C_\delta(x)\right| > \frac{C_\delta(x)}{2},$$

on peut écrire

$$P\left(\left|\hat{f}_n(x)\right| \leq \frac{C_\delta(x)}{2}\right) \leq P\left(\left|\hat{f}_n(x) - C_\delta(x)\right| > \frac{C_\delta(x)}{2}\right).$$

Il suffit maintenant de poser $\tau = \varepsilon = C_\delta(x)/2$ pour obtenir (4.7). Finalement, le résultat du Théorème (4.2.1) se déduit aisément de (4.1) ainsi que des cinq résultats intermédiaires précédents, à savoir (4.2), (4.4), (4.5), (4.6) et (4.7).

La vitesse de convergence presque complète peut être précisée grâce au résultat suivant.

Théorème 4.2.2 *Sous les hypothèse* $(H_6), (H_5)$ *et si*

$$|r(u) - r(v)| \leq C \left\|u - v\right\|^{\beta}, \tag{4.8}$$

et si pour h_0 indépendant de n,

$$h_n = h_0 \left(\frac{\log n}{n} \right)^{\frac{1}{2\gamma(x)+\delta(x)}},$$

où

$$\gamma(x) = \min \{b(x), \beta\},$$

alors

$$r(x) - \hat{r}_n(x) = O\left(\left(\frac{\log n}{n} \right)^{\frac{\gamma(x)}{2\gamma(x)+\delta(x)}} \right) \ p.co. \tag{4.9}$$

Preuve D'après le résultat $i)$ du Lemme (4.2.1), il vient

$$E\hat{f}_n(x) = C_\delta(x) + O(h_n^{b(x)}) \tag{4.10}$$

Par ailleurs, au cours de la preuve du Lemme(4.2.3) on a vu que

$$\begin{aligned} |E\hat{g}_n(x) - r(x)C_\delta(x)| \quad &\leq \quad |r(x)| \left| E\hat{f}_n(x) - C_\delta(x) \right| \\ &\quad + \frac{1}{h_n^{\delta(x)}} E\left(|r(X) - r(x)| \left| K\left(\frac{\|X - x\|}{h_n} \right) \right| \right). \end{aligned}$$

Grâce à la condition (4.8) sur r et (4.10) on obtient

$$E\hat{g}_n(x) = r(x)C_\delta(x) + O(h_n^{b(x)}) + O(h_n^\beta). \tag{4.11}$$

En combinant ces deux résultats avec la décomposition

$$\hat{r}_n(x) - r(x) = \frac{1}{\hat{f}_n(x)} \{\hat{g}_n(x) - r(x)C_\delta(x)\} - \frac{r(x)}{\hat{f}_n(x)} \left\{ \hat{f}_n(x) - C_\delta(x) \right\},$$

puis en utilisant (4.10) et (4.11) ainsi que les Lemmes (4.2.2) et (4.2.4), on arrive à

$$\hat{r}_n(x) - r(x) = O(h_n^{b(x)}) + O(h_n^\beta) + O\left(\sqrt{\frac{\log n}{n h_n^{\delta(x)}}} \right).$$

Il suffit de remplacer h par sa valeur pour obtenir (4.9). ■

Remarque 4.2.1 *On peut étudier la convergence presque complète uniforme sur un compact S de l'espace semi-normé $(H, ||.||)$. Il est clair que l'obtention de résultats uniformes nécessite le renforcement de certaines hypothèses utilisées dans le paragraphe précédent. En particulier, on rend uniforme l'hypothèse (H_4) de la manière suivante :*

$$\lim_{\alpha \to 0^+} \sup_{x \in S} \left\{ \frac{P(X \in B(x, \alpha)}{\alpha^\delta} - c(x) \right\} = 0, \tag{4.12}$$

et

$$\inf_{x \in S} c(x) > 0, \tag{4.13}$$

où δ est un réel strictement positif ne dépendant pas de x. Notons que dans ce cas, cette hypothèse peut s'interpréter en terme de dimension de Hausdorff de la loi de la variable aléatoire fonctionnelle X (cf Bardet, 1997, p.28). On peut alors énoncer le résultat suivant.

Sous l'hypothèses (H_5), si les relations (4.12) et (4.13) sont vérifées, et si r est uniformément continu sur S,

alors on a :

$$\lim_{n \to \infty} \sup_{x \in S} |\hat{r}_n(x) - r(x)| = 0 \ p.co.$$

Remarque 4.2.2 *De nombreux résultats de statistique fonctionnelle ont été établis en considérant des échantillons indépendants. Cependant, il est parfois intéressant de considérer et d'étudier des échantillons dépendants afin de pouvoir répondre à des situations où les données ne sont pas indépendantes. Dans ce cadre des résultats ont été obtenus sous l'hypothése de $(\alpha - mélange)$ (cf Ferraty et Vieu (2000).*

Chapitre 5

Simulation

Nous terminons ce mémoire par un travail de simulation de l'estimateur à noyau de la fonction de régression aussi bien pour des données complètes que pour des données censurées à droite ou à gauche et pour différents modèles.

Nous choisissons le noyau gaussien

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

et la fenêtre donnée par

$$h_n = (log(log(n))/n)^{(1/5)}.$$

5.1 Modèle linéaire

Soit $Y_i = \alpha X_i + \alpha_0 + b\varepsilon_i$, où X_i et ε_i sont deux suites de variables aléatoires i.i.d. de loi normale $N(0,1)$. Nous choisissons pour notre simulation

$$b = 0.2, \quad \alpha = 2 \quad \text{et } \alpha_0 = -1.$$

On a $r(x) = E(Y/X = x) = \alpha x + \alpha_0 = 2x - 1$.

Dans le cas du modèle de censure, nous simulons aussi n variables aléatoires C_i i.i.d. de loi $N(0,1)$.

Nous posons

$$Z_i = Y_i \wedge C_i, \quad \delta_i = 1_{\{Y_i \leq C_i\}} \quad \text{dans le cas de la censure à droite,}$$

et

$$Z_i = Y_i \vee C_i, \quad \delta_i = 1_{\{Y_i \geq C_i\}} \quad \text{dans le cas de la censure à gauche.}$$

5.2 Modèle non linéaire

Nous traitons les cas suivants

a) Cas exponentiel $Y_i = \exp(X_i - 0.2) + \varepsilon_i$.

b) Cas parabolique $Y_i = \frac{3}{2}X_i^2 - \frac{1}{2} + \varepsilon_i$.

c) Cas sinusoïdal $Y_i = \sin(\frac{3}{2}X_i) + \varepsilon_i$,

où X_i et ε_i sont deux suites de variables i.i.d. de loi $N(0,1)$.

La variable de censure est aussi régie par la loi $N(0,1)$ dans le cas parabolique, sinusoïdal mais elle est régie par la loi exponentielle de paramètre 2.5 dans le cas exponentiel.

Les résultats de la simulation sont présentés dans les pages suivantes et montrent la bonne performance des estimateurs étudiés aussi bien dans le cas de données complètes que censurées à droite ou à gauche.

La censure ne semble pas influer sur les résultats obtenus. Ceci est conforme aux résultats théoriques qui s'équivalent que les données soient censurées ou pas.

Données complètes

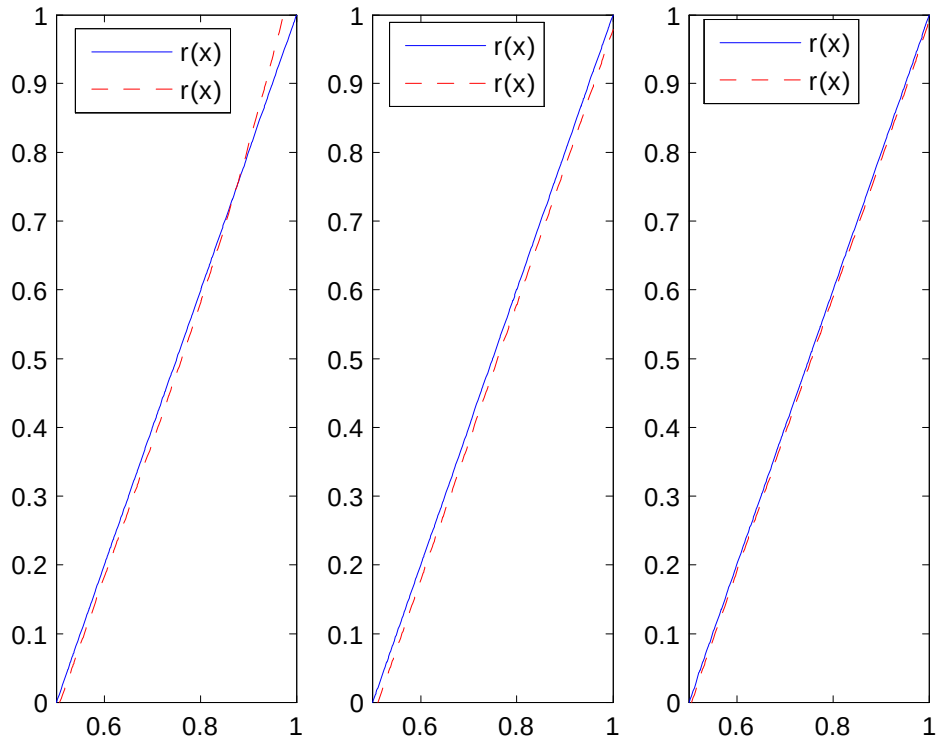

FIG. 5.1 – Cas linéaire avec n=100, 500, 1000 respectivement.

Données complètes

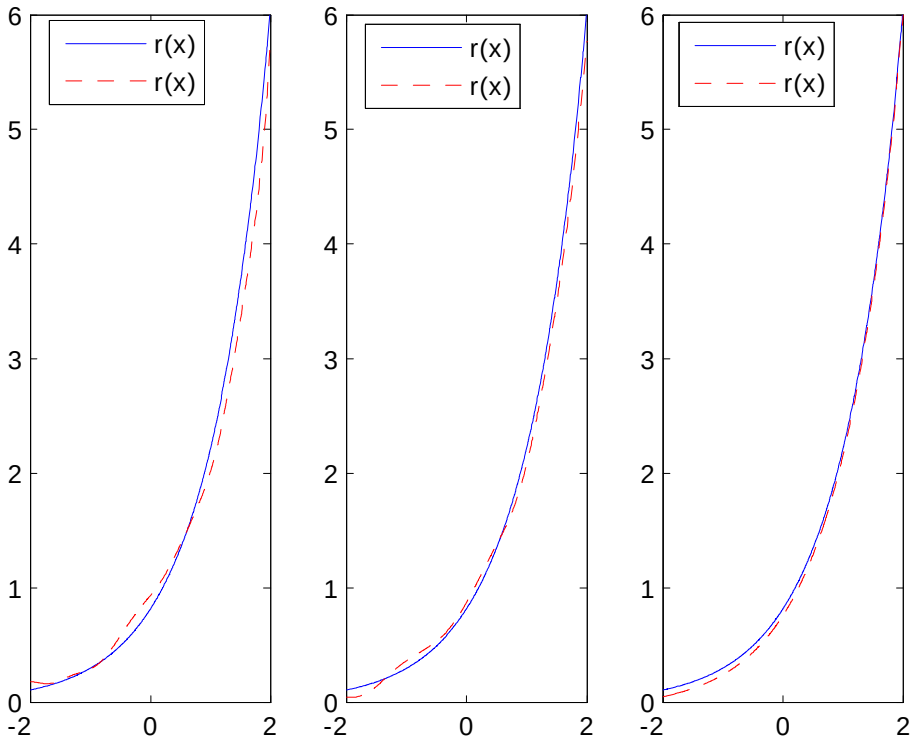

FIG. 5.2 – Cas exponentiel avec n=100, 500, 1000 respectivement.

Données complètes

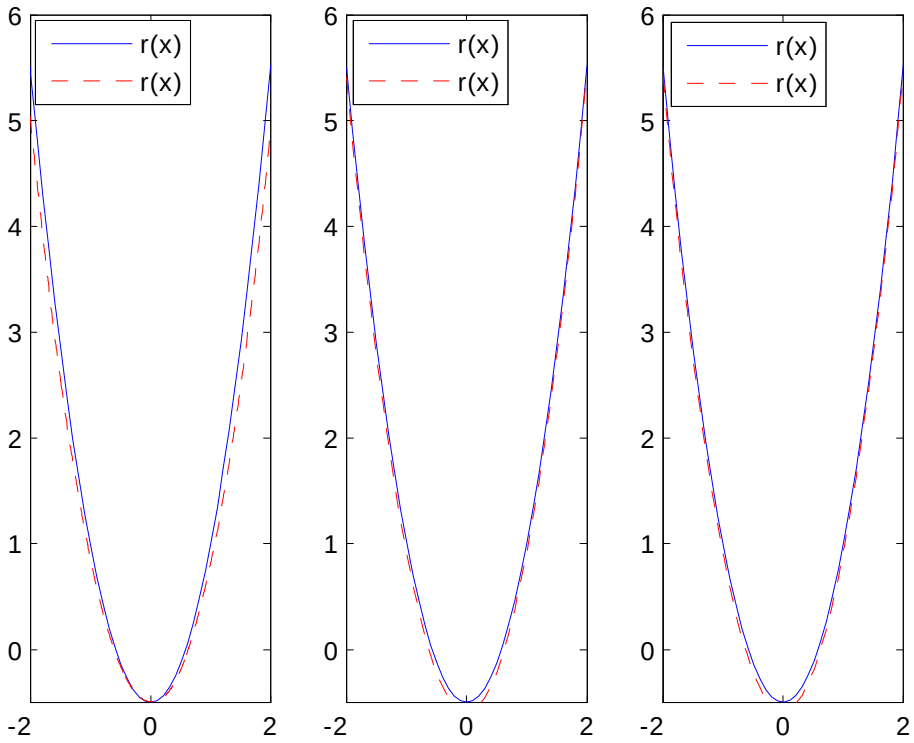

FIG. 5.3 – Cas parabolique avec n=100, 500, 1000 respectivement.

Données complètes

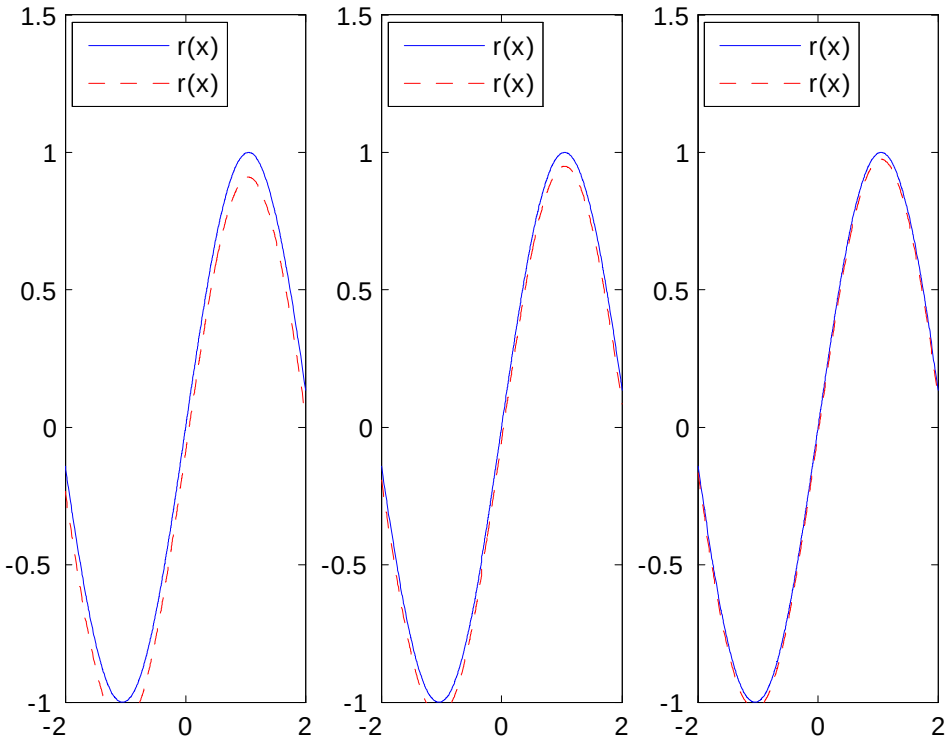

FIG. 5.4 – Cas sinusoïdal avec n=100, 500, 1000 respectivement.

Données censurées à droite

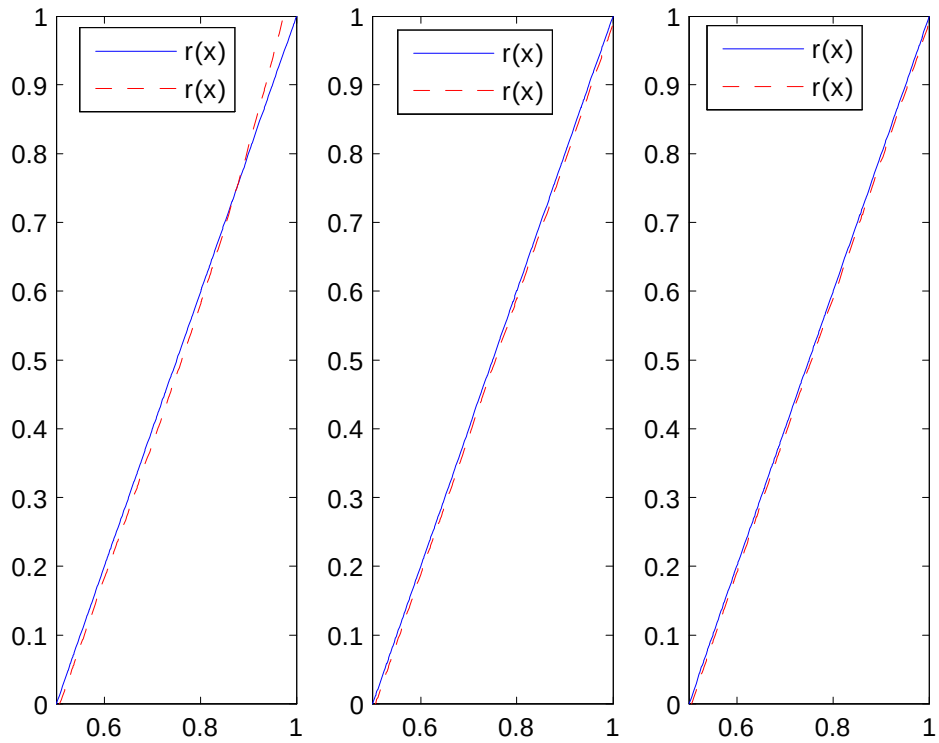

FIG. 5.5 – Cas linéaire avec n=100, 500, 1000 respectivement.

Données censurées à droite

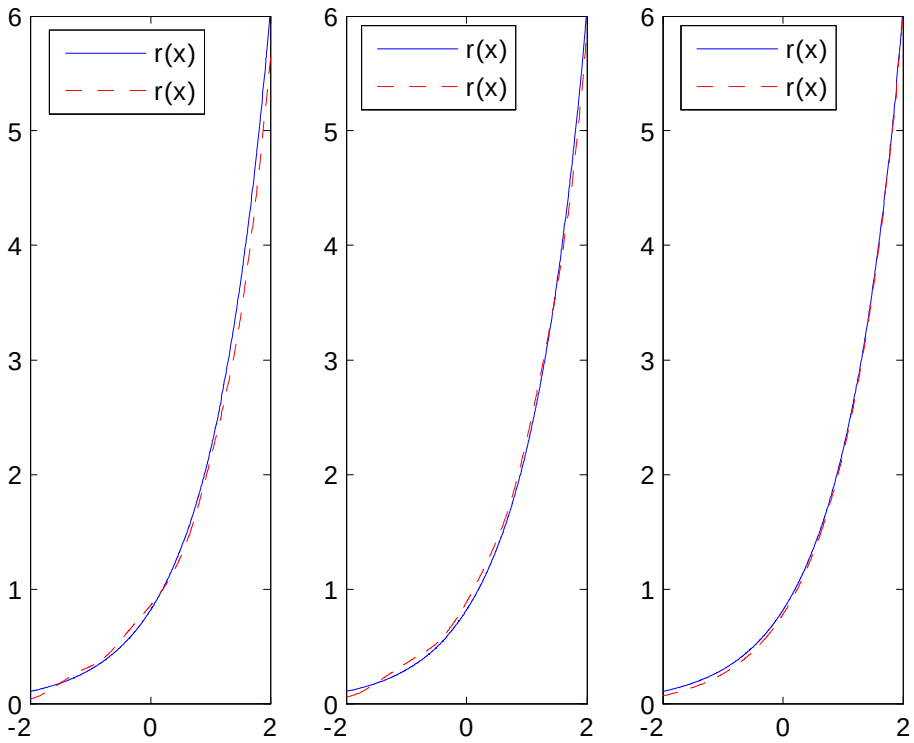

FIG. 5.6 – Cas exponentiel avec n=100, 500, 1000 respectivement.

Données censurées à droite

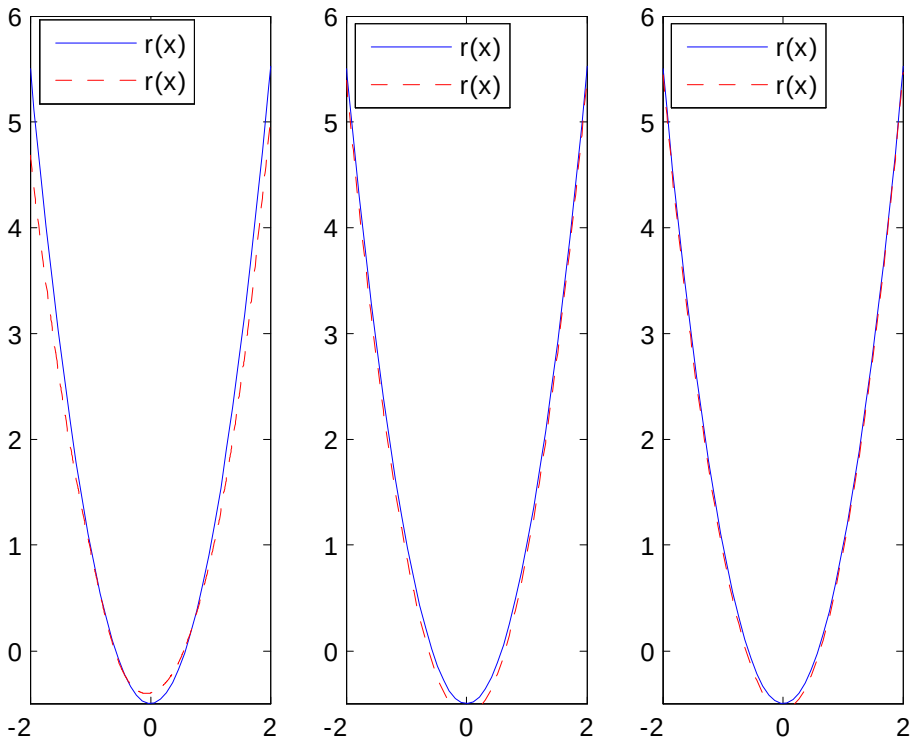

FIG. 5.7 – Cas parabolique avec n=100, 500, 1000 respectivement.

Données censurées à droite

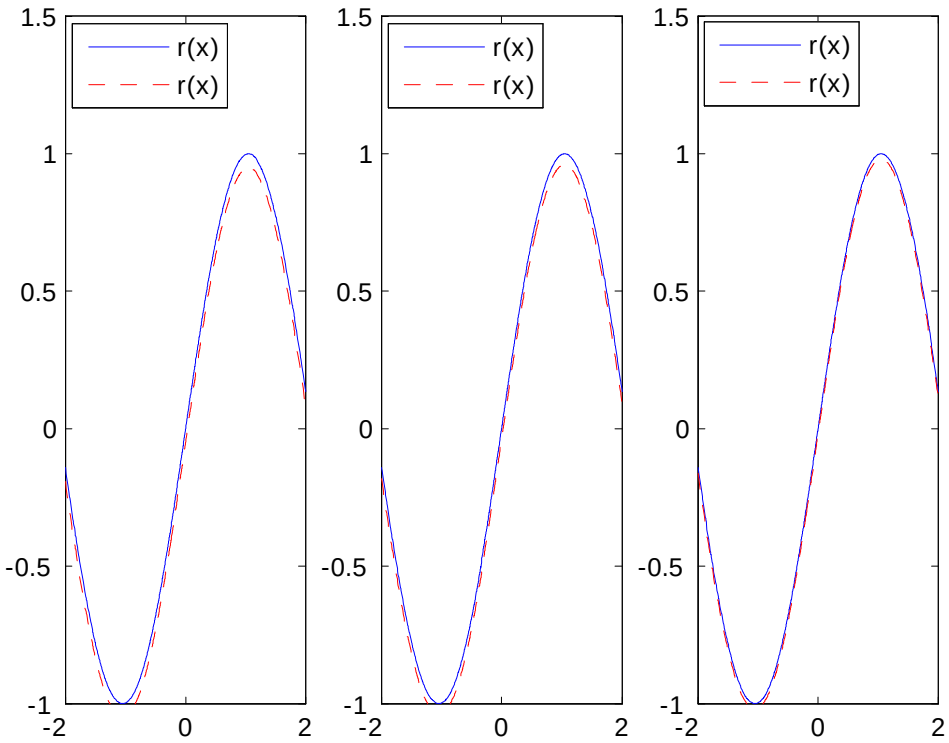

FIG. 5.8 – Cas sinusoïdal avec n=100, 500, 1000 respectivement.

Données censurées à gauche

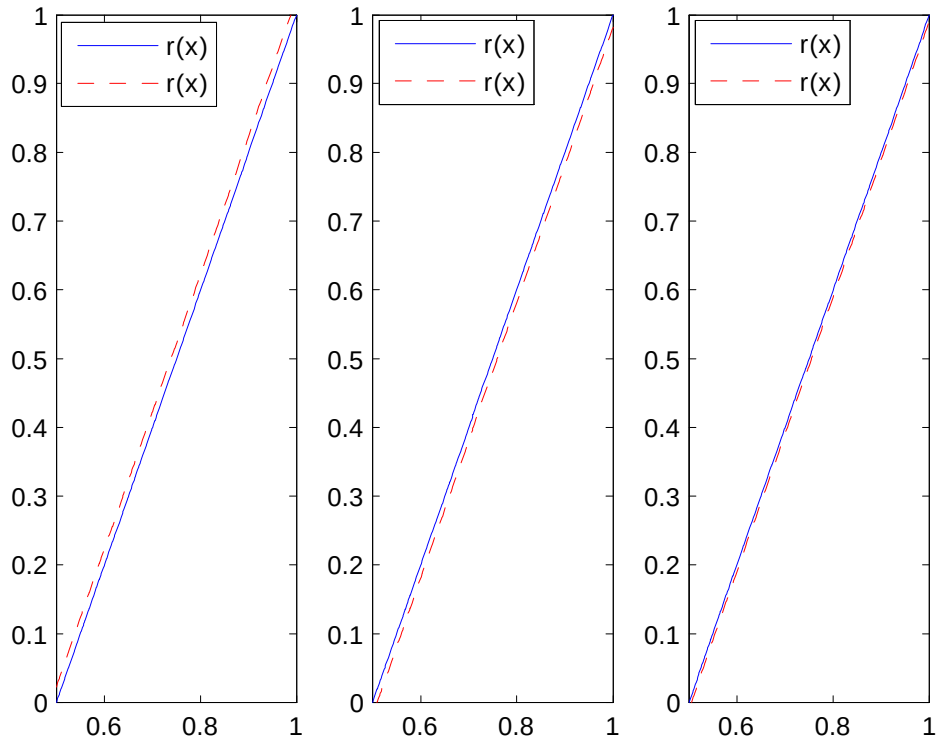

FIG. 5.9 – Cas linéaire avec n=100, 500, 1000 respectivement.

Données censurées à gauche

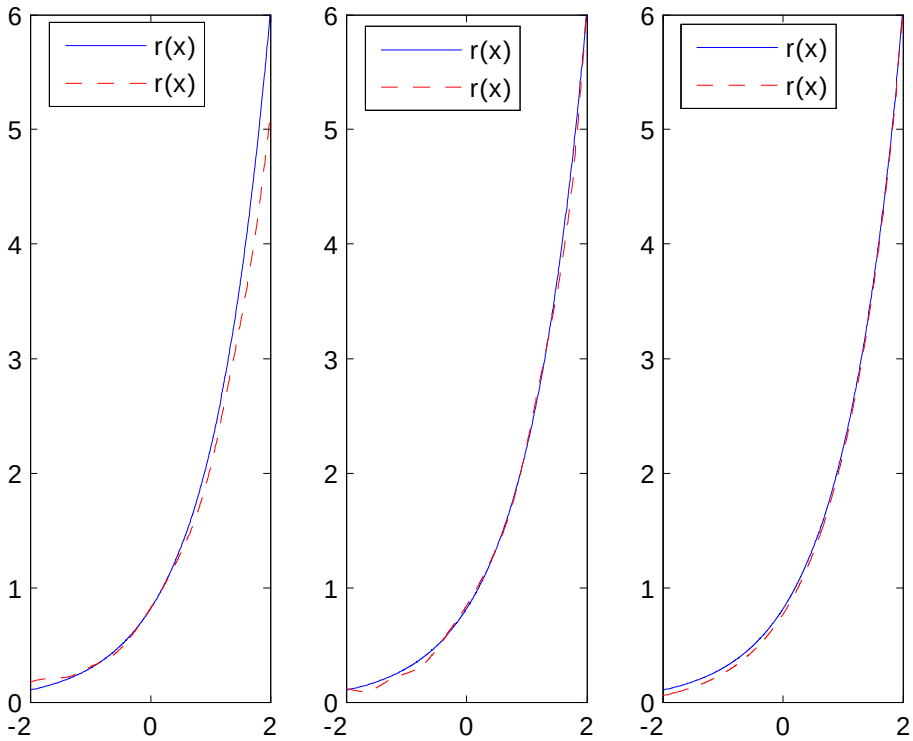

FIG. 5.10 – Cas exponentiel avec n=100, 500, 1000 respectivement.

Données censurées à gauche

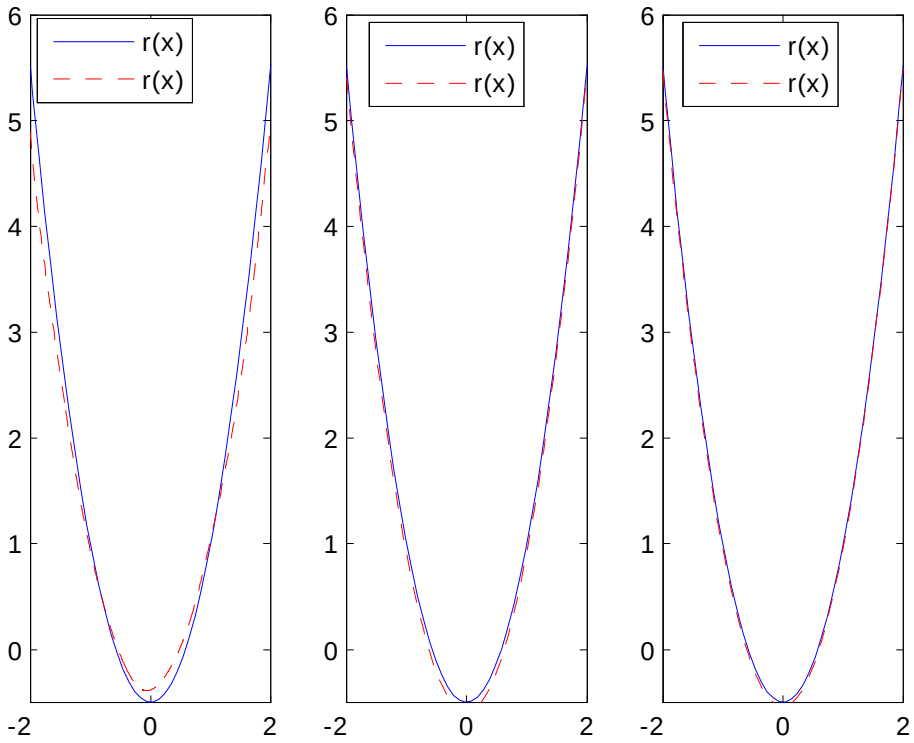

FIG. 5.11 – Cas parabolique avec n=100, 500, 1000 respectivement.

Données censurées à gauche

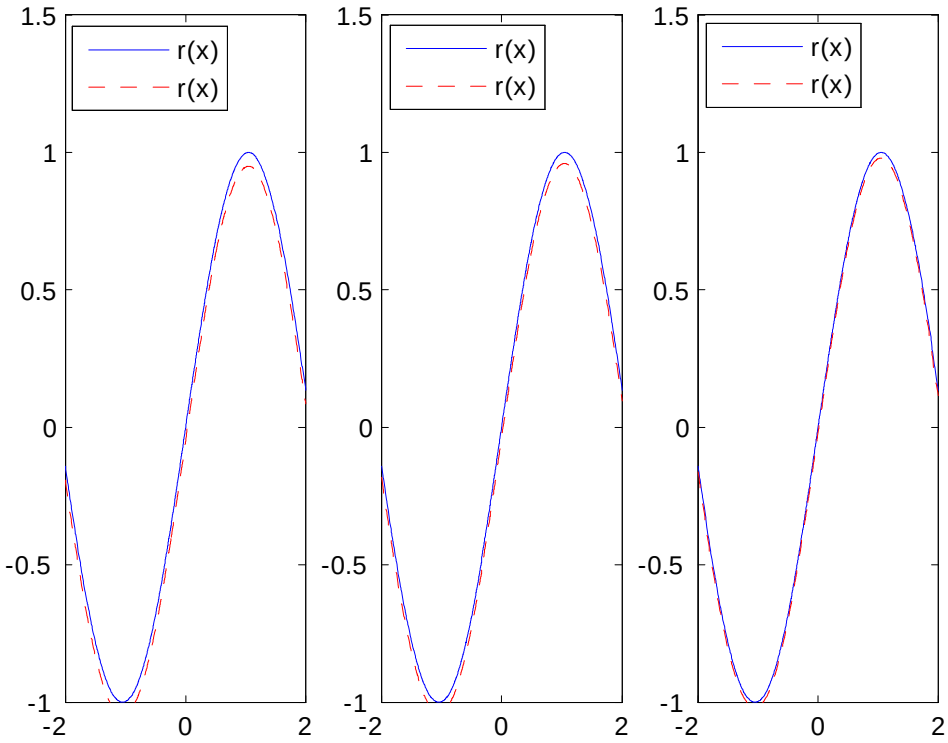

FIG. 5.12 – Cas sinusoïdal avec n=100, 500, 1000 respectivement.

Bibliographie

[1] J. M. Bardet. Tests d'autosimilarité des processus gaussiens. Dimension fractale et dimension de corrélation. *Thèse 3eme cycle, Paris-Sud,* (1997).

[2] R. Beran. Nonparametric regression with randomly censored survival data, *Technical university of Clifornia, Berkeley.* (1981).

[3] E. Brunel and F. Comte. *Model selection for additive regression inpresence of right censoring.* Preprint MAP5. (2006a).

[4] E. Brunel and F. Comte. Adaptive nonparametric regression estimation in presence of right censorig. *Math. Methods Statist,* **15**. 3. 233-255. (2006b).

[5] A. Carbonez. Nonparametric Functional Estimation under Random Censoring and a New Semiparametric Model of Random Censorship. Phd Thesis Katholieke Universiteit leuven, (1992).

[6] A. Carbonez, L. Györfi, and E. C. van der Meulen, Partition-estimates of a regression function under random censoring, *Statist. Decisions* **13**, 21-37, (1995).

[7] D. M. Dabrowska. Nonparametric regression with censored survival data. *Scand. J. Statist.*, **14** :181-197, (1987).

[8] L. Devroye, L. Györfi, Distribution-free exponential bounds on the L1 error of partitioning estimates of a regression function, in "Proceedings of the Fourth Pannonian Symposium on Mathematical Statistics" (*F. Konecny, J. Mogyorodi, and W. Wertz, Eds.), Akadémiai Kiado, Budapest, Hungary, pp.*67-76, (1983).

[9] L. Devroye, L. Györfi, A. Krzyżak, and G. Lugosi. On the strong universal consistency of nearest neighbor regression function estimates, *Ann. Statist.* **22**, 1371-1385, (1994).

[10] L. Devroye, and A. Krzyżak. An equivalence theorem for L1 convergence of the kernel regression estimates, *J. Statist. Plann. Inference* **23**, 71-82, (1989).

[11] J. Fan and I. Gijbels. Censored regression : Local linear approximations and their applications, *J. Amer. Statist. Assoc.* **89**, 560-570, (1994).

[12] F. Ferraty et P. Vieu Dimension fractale et estimation de la régression dans des espaces vectoriels semi-normés. *Compte Rendus Acad. Sci. Paris,* **330**, 139-142. (2000).

[13] F. Ferraty et P. Vieu Functional Nonparametric Model : a Now Tool for Spectrometric Data. *soumis pour publication.(2001).*

[14] F. Ferraty et P. Vieu *Modèles Non-paramétriques de Régression.* Cours de D.E. A. (2002-2003).

[15] F. Ferraty et P. Vieu Nonparametric models for function data, with application in regression, time-series prediction and curve discrimination. The International Conference on Recent Trends and Directions in Nonparametric Statistics. *J. Nonparametric. Stat.* **16** (1-2), 111-125, (2004).

[16] F. Ferraty et P. Vieu *Nonparametric function data.* Springer series statistics. (2006).

[17] F. Ferraty et P. Vieu. *Nonparametric modelling for function data*. Springer-Verlag, New York, (2006a).

[18] F. Ferraty et P. Vieu Function nonparametric statistics in action. *The art of semiparametrics.* Contrib. Statist. Physica-Verlag / Springer, Heidelberg, 112-129, (2006b).

[19] R. D. Gill and S. Johansen. A survey of product integration with a view toward application in survival analysis. *The Annals of Statistics*. Vol **18**, No 4, 1501-1555, 1990.

[20] Z. Guessoum and E. Ould Said. On the nonparametric estimation of the regression function under censorship model. *Statist.Decisions*. **26,** 159-177**,** (2009).

[21] L.Györfi and H. Walk. On the stong universal consistency of a recursive regression by Pal Reversz, *Statist. Probab. Lett.* **31**, 177-183, (1997).

[22] W. Hoeffding. Probability inequalities for sums of bounded random variable. *J. Amer. Statist. Assoc.*, **58**, 15-30. (1963).

[23] E. L. Kaplan and P. Meier. Nonparametric estimation from incomplete observation, *J. Amer. Statist. Assoc.* **53**, 457-481, (1958).

[24] K. Kebabi et F. Messaci. Estimation non paramétrique dans un modèle de censure à gauche. Communication RAMA09, Alger, (2009).

[25] M. Kohler, On the universal consistency of a least squares splines regression estimator, *Math. Methods Statist.* **6**, 349-364, (1997).

[26] M. Kohler, Universally consistent regression function estimation using heirarchical B-splines, *J. Multivariante Anal.* **67**, 138-164, (1999).

[27] M. Kohler, and A. Krzyżak, Nonparametric regression estimation using penalized least squares, to appear in *IEEE Trans. Infrom. Theory* (2001).

[28] M. Kohler, K. Mathé and M. Printér. Prediction from Randomly Right Censored Data, *J. Multivariate, Anal.* **80**, 73-100, (2002).

[29] M. Kohler, S. Kul, K. Mathé. *Least squares estimates for censored regression*, Preprint, Available at http ://www.mathematik. uni-stuttgart. de/math A/Ist3/kohler/hfm-pub-en : html, (2006).

[30] E. A. Nadaraya. : On estimating regression. *Theor. Probab. Appl.* **9**, 141-142 (1964).

[31] E. A. Nadaraya. :*Nonparametric Estimation of Probability Densities and regression Curves.* Kluver, Dordrecht, (1989).

[32] E. Ould Said and Z. Cai. Strong uniform consistency of nonparametric estimation of the consored conditional mode function. *J. Nonparametric Stat.* **17**, No. 7, 797-806. (2005).

[33] W. Stute and J. L. Wang, The strong law under random censorship, *Ann. Statist.* **21**, 1591-1607, (1993).

[34] M. P. Wand, et M. C. Jones. *Kernel Smoothing*. Chapman and Hall, London, (1995).

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

On the estimate of the function of regression

Abstract :

This memory relates to the kernel estimators of the function of regression in various contexts, namely for complete data (real and functional) like for data censored on the right. Then an extension to the model of censure on the left was proposed. Finally a work of simulation made it possible to check the good performance of the studied estimators.

Key words : Nonparametric regression estimate, kernel estimator, almost complete convergence and almost sure convergence.

Sur l'estimation de la fonction de régression

Résumé :

Ce mémoire porte sur les estimateurs à noyau de la fonction de régression dans différents contextes, à savoir pour des données complètes (réelles et fonctionnelles) ainsi que pour des données censurées à droite. Puis une extension au modèle de censure à gauche a été proposée. Finalement un travail de simulation a permis de vérifier la bonne performance des estimateurs étudiés.

Mots clé : L'estimateur non paramétrique de régression, l'estimateur à noyau, convergence presque complète et convergence presque sûre.

تقدير وظيفة الإنحدار

ملخص

هذه المذكرة تتعلق بتقدير وظيفة الإنحدار بطريقة النواة في سياقات مختلفة، وهي ما تعرف بالمعطيات الكاملة (الحقيقية والوظيفية) وبمعطيات المراقبة على اليمين.
كما قمنا كذلك باقتراح نموذج المراقبة على اليسار.
وأخيرا أكدت عملية المحاكاة التي قمنا بها بالنوعية الجيدة للتقادير المدروسة.

كلمات البحث الرئيسية: التقدير الغيرالوسيطي، مقدر النواة، التقارب الكامل تقريبا، التقارب المؤكد تقريبا.