UNIVERSITÉ KASDI MERBAH
OUARGLA

Faculté des mathématiques et sciences de la matière

DÉPARTEMENT DE MATHÉMATIQUES

MASTER

Spécialité : Mathématiques

Option : Probabilités et statistiques

Par : Henka Soufaine

Thème

l'estimation de Régression non paramétrique dans cas données censure à droite

Version de : 15/11/2020

Devant le jury composé de
:

Abassi.H .	Université KASDI Merbah- Ouargla.	Président
Halili Rachide .	Université KASDI Merbah- Ouargla.	Examinateur
Agoune Rachid M.A.	Université KASDI Merbah- Ouargla.	Rapporteur

Contents

Dédicace

C'est avec la volonté de Dieu que je suis arrivée â ce rang pour faire ce modeste travail. Je le dédie avec tout mon amour et mon respect, avec toute la tendresse a trés chers parents : mon père et ma mère

À mes frères "**elaid** ", " **mohammed** ", avec mes meilleurs voeux comme je les remercie pour leurs sacrifices, leur patience et leur encouragements.

À toutes la familles " **HENKA** "

À mes chères amies :" **emhemmed** ", " **noreddine** ",

À mes chers professeurs et toutes mes amies.

Sans oublier ma promotion, dont je garderai de très bons souvenirs.

Remerciement

Avant tout je remercie, le Dieu tout puissant de je accordée la volonté et la patience pour accomplir ce modeste travail.

Je tenie d'abord à remercier très chaleureusement mon encadreur de mémoire de fin d'études monsieur : **Agoune Rachid** , pour ses précieux conseils et son orientation tout au long de mon recherche. Tous les membres de jury d'avoir participé à la commission des examinateurs en vue dune évaluation prompte et à sa juste valeur.

Je
tenons à remercier a nos enseignements de faculté des Mathématiques et sciences de la matière.

A la fin,

Nous tenons également à remercier toutes les personnes qui participé de près ou de loi à la réalisation de ce travail.

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 et l'estimation paramétrique , Une option non-paramétrique qui consiste à estimer Collection à partir des observations , une fonction inconnue, Une action non paramétrique est spécifiée une procédure non paramétrique est déterminée par la dépendance de la loi 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 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), 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 â une maladie . Y est alors censuré à droite : on ne connaît pas sa valeur exacte, on sait seulement qu'elle dépasse l'observation recueillie. Dans ce contexte de censure à droite dans lequel permettant d'en déduire ,des estimateurs non paramétriques de la fonction de régression difficilement calculables .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 . on peut seulement savoir qu'elles sont inférieures aux observations. C'est la censure à gauche.Cela peut être le cas lorsque l'on regarde l'âge ,Par ailleurs, les modèles de régression se subdivisent en deux familles selon le type,C'est pourquoi on s'intéresse au traitement des variables aléatoires fonctionnelles (cas de la régression d'une variable réelle sur une variable fonctionnelle)

et en le dernier, 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.

Chapter 1

Rappels

1.1 Données censurées :

L'accent doit être mis le concentrer sur l'estimation dans le modèle de censure dans plusieurs études et applications statistiques,En analyse de survie et de fiabilité,On dit que c'est le moment de l'échec si un événement spécifique survient au cours de cette étude et que ce temps est pris en compte, Où nous en disons la durée de survie, ou simplement une durée. C'est une variable aléatoire positive et souvent supposée bornée. peut être la durée de vie d'un patient après un traitement, la durée de chômage, le temps de panne d'un appareil, le dent auquel un enfant apprend à accomplir Tâche particulière, il arrive souvent, pour diverses raisons, que la durée d'intérêt ne puisse pas être observée. Cela peut être dû à la perte de vision du patient, au début ou à la fin de la période d'étude, et ces valeurs sont surveillées. Les valeurs contrôlées, bien qu'inconnues, doivent être prises en compte pour obtenir des estimations valides et des conclusions précises. En fonction de la situation spécifique, la littérature statistique contient un grand nombre de procédures qui permettent de tenir compte des observations censurées

1.2 Les type des données censurées :

Dans cette étude de la censure, il existe trois classes de censure appelées censure de droite, censure de gauche et censure entre elles (lorsque nous connaissons les limites supérieure et inférieure d'un événement). Il existe

différents types de censure dans ces trois catégories :

A)- Censure de type I :

Ainsi soit-il c est une constante positive et un n-échantillon $X_1.....X_n$ on observe :

$$T_i = X_i \wedge X\left(r\right) \text{ et } \delta_i = I_{\{X_i = T_i\}}$$

Telque :$X_i \wedge X\left(r\right)$ représente le minimum $\left(X_i, X\left(r\right)\right)$ Le temps de censure est fixé par le chercheur comme étant la fin de l'étude.

B)- Censure de type II :

Soit i tel qu'à chaque $i = 1, ..., n$ est associé un couple de variables aléatoiresnon nul X_i, C_i ou seul le minimum est observé c'est-à-dire qu'on observe :

$$T_i = X_i \wedge C_i \text{ et } \delta_i = I_{\{X_i = C_i\}}$$

Où est un indicateur de censure tel que :

$$\delta_i = \begin{cases} 1 & si\ X_i \leq C_i\ ; \\ 0 & si\ X_i > C_i\ . \end{cases}$$

Où X_i est l'instant de l'événement. C_i est l'instant de censure.

C)- Censure de type III :

Etant donné un entier positif r fixé, un n-échantillon $X_1.....X_i$ d'une variable aléatoire positive X et les statistiques d'ordre $X(1), ..., X(n)$ on observe :

$$T_i = X_i \wedge X\left(r\right) \text{ et } \delta_i = I_{\{X_i = T_i\}}$$

autrement dit, ce genre de censure se caractérise par le fait que l'étude cesse aussitôt qu'a eu lieu un nombre d'événements prédéterminés par lexpérimentateur.

1.3 Censure à droite :

On dit que le temps de survie est subjugué censure à droite lorsque le temps de survie est supérieur au temps utilisé.la durée de survie d'un évènement est définie par le couple$(X;\delta)$ où :

$$X = inf\ (T;C)$$

et

$$\delta = \begin{cases} 1 & si\ T \leq C\ ; \\ \\ 0 & si\ \ T > C\ . \end{cases}$$

1.4 Censure par intervalle :

Un état de censure plus général qui se produit lorsque sa duréeLa survie est limitée à une zone de valeur inconnue

On a aussi pour ce genre d'expériences des données qui sont censurées à droite ou, plus rarement, à gauche.Un avantage de ce type est qu'il permet de représenter les données censurées à droite ou à gauche par des intervalles du type [a, $+\infty$[et [0, a] respectivement, ce qui permet de considérer ce modèle comme étant plus générique.

1.5 Censure aléatoire :

C'est typiquement ce modèle qui est utilisé pour les essais thérapeutiques. Dans ce type d'expériences, la date d'inclusion du patient dans l'étude est fixée, mais la date de fin d'observation est inconnue celle-ci correspond, à la durée d'hospitalisation du patient. Ici, le nombre d'événements observés et la durée totale de l'expérience sont aléatoires.

1.6 Estimation :

Parmi les fonctions que nous utilisons dans notre étude se trouve le processus d'approximation ou estimation est le processus qui consiste à trouver une estimation,ou une approximation qui est une valeur utilisablendans un certain

but même si les données d'entrée peuvent être incomplètes incertaines ou instables la valeur est néanmoins utilisables car elle est dérivée des meilleures informations disponibles.en règle générale,lestimation implique l'utilisation de la valeur d'une statistique dérivée d'un paramètre de population correspondant l'échantillon fournit des informationsqui peuvent être projetées, par le biais de divers processus formels ou informels,pour déterminern une fourchette la plus susceptible de décrire l'information manquante.une estimation qui s'avère incorrecte sera une surestimation si l'estimation dépasse le résultat réel, et une sous-estimation si l'estimation est inférieure au résultat réel

1.7 Régression :

en psychanalyse,la régression est le processus inhérent à l'organisation libidinale quifait en sorte que les fonctionsparvenues plus loin dans leur organisation,peuvent facilement aussi,enréponse à une frustration de la satisfaction-libidinale recherchée,revenir à l'un de ces stades antérieurs,en mouvement rétrograde.Ce processus produit et anime le retour d'un fonctionnement ou d'un état psychique plus avancé à un niveau dépassé,à des modalités défensives dépassées ou encore le retour aux premiers objets de la libido.

1.8 méthode du noyau

Introduction

L'estimation par noyau est une méthode non paramétrique d'estimation de la densité d'une variable aléatoire.

Cette méthode permet d'obtenir une densité continue et constitue en ce sens une généralisation de la méthode de l'histogramme.

En effet, la fonction indicatrice utilisée pour l'histogramme est ici remplacée par une fonction continue (le noyau) et une somme de fonctions continues reste continue.

Définition 01.: Soit $K : R \to R$, on dit que K est un noyau si et seulement si : $\int K(u)du = 1$ Alors K est appelé noyau. Pour tout $h_n > 0$ petit et $n \in \mathbb{N}^*$, on peut définir $x \in \mathbb{R}$ par :

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

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

- K est dit positif si $K(u) \geq 0 \ \forall \ u$
- K est dit symétrique si $K(u) = K(-u) \ \forall u$

On commence par remarque que la densité est la dérivée de la fontion de répartition,ce qui permet d'ecrire pour tout x :

$$f(x) = F'(x) = \lim_{h \to 0}\frac{F(x+h) - F(x)}{2h}$$

$$= \lim_{h \to 0}\frac{F(x+h) - F(x-h)}{2h}$$

Donc pour un $h > 0$, on peut penser â estimer $f(x)$ par :

$$\hat{f}(x) = \frac{1}{2h}(F_n(x+h) - F_n(x-h))$$

$$= \frac{1}{2nh}\sum_{i=1}^{n}1_{\{x-h,x+h\}}(X_i)$$

Exemple de noyaux :

Voici quelques exemples de noyaux les plus communément utilisés:

- Noyau rectangulaire :

$$K_1(x) = \begin{cases} \dfrac{1}{2}, & \text{si } |x| \leq 1; \\ 0\ , & \text{si } |x| > 1. \end{cases}$$

- Noyau triangulaire :

$$K_2(x) = \begin{cases} 1 - |x|\,, \text{si } |x| \leq 1; \\ 0\ , \quad \text{si } |x| > 1. \end{cases}$$

- Noyau d'Epanechnikov ou parabolique :

$$K_3(x) = \begin{cases} \dfrac{3}{4}\left(1 - x^2\right)\ , \text{ si } x \in [-1, 1]\,; \\ 0, \qquad\qquad \textit{sinon.} \end{cases}$$

- Noyau quadratique :

$$K_4(x) = \begin{cases} \dfrac{15}{16}\left(1 - x^2\right)^2\ , \text{ si } x \in [-1, 1]\,; \\ 0, \qquad\qquad \textit{sinon.} \end{cases}$$

- Noyau gaussien :

$$K_5(x) = \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-1}{2}x^2\right), \quad x \in R.$$

Les deux premiers ont l'avantage d'être simples, le noyau triangulaire étant continu partout et conduisant à une estimation f_n continue. Le troisième doit sa notoriété à une propriété d'optimalité théeorique mais sans grand intérêt pratique. Le quatrième est, à notre sens, le plus intéressant car donnant une estimation dérivable partout, tout enétant simple à mettre en oeuvre. En fait il s'agit du noyau le plus simple parmi les noyaux de forme polynomiale dérivables partout. Ainsi il assure le lissage local de la fonction f_n . Ce noyau est d'une forme très proche du noyau Gaussien et il est donc préférable. Notons que plus la valeur de h est elevee plus on elargit la fenêtre, ce qui donne un effet de lissage globale de f_n plus important.

Voici quelques courbes de noyaux usuels présentées ci-dessous

Code R .

```
K1=function(t){(1-abs(t))*ifelse(abs(t)<=1,1,0)}
K2=function(t){(15/16)*((1-t^2)^2)*ifelse(abs(t)<=1,1,0)}
K3=function(t){dnorm(t)}
K4=function(t){ifelse(abs(t)<1,(3/4)*(1-t^2),0)}
op=par(mfrow=c(2,2))
curve(K1(x),-1,1,ylab="K(x)",main="Triangulaire")
curve(K2(x),-1,1,ylab="K(x)",main="Biweight")
curve(K3(x),-4,4,ylab="K(x)",main="gaussien")
curve(K4(x),-1,1,ylab="K(x)",main="Epanechikov")
par(op)
```

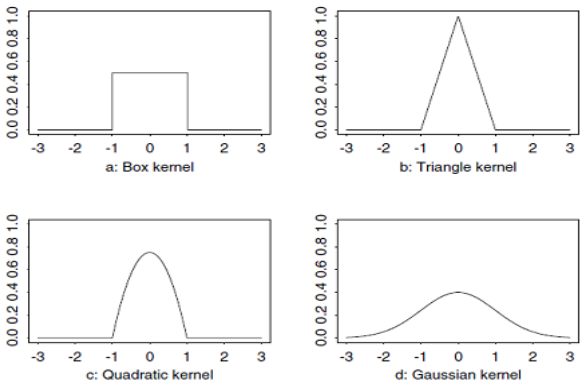

*FIG.*1—Les courbes des noyaux les plus communs

1.9 Estimateur à noyau

L'estimateur à noyau est probablement l'estimateur le plus utilisé et certainement le plus étudié mathématiquement, car il possède des propriétés qui le rendent fort intéressant.

Définition 02 : Un estimateur à noyau noté f_n de la fonction$f$ est défini par :

$$f_n(x) = \frac{1}{nh_n}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h_n}\right) .... (*)$$

où $\{h_n\}$ $n \geq 1$ est une suite de réels positifs appelés paramètres de lissage ou largeur de la fenêtre, qui tend vers 0 quand n tend vers l'infini.

Comme nous allons le voir par la suite, si le noyau Kest une fonction de densité alors l'estimateur à noyau f_n est lui aussi une fonction de densité.

De plus, ce dernier possède les propriétés de continuité et de différentiabilité. De sorte que si, par exemple, K est la densité normale alors f_n possède des dérivées de tout ordre.

Propriété 01 : Un estimateur à noyau est une densité

Démonstration :

$$\int_{-\infty}^{+\infty} f_n(x)\ dx = \frac{1}{nh_n}\sum_{i=1}^{n}\int_{-\infty}^{+\infty} K\left(\frac{x-X_i}{h_n}\right)dx$$

en posent $u = \left(\frac{x-X_i}{h_n}\right)$

$$= \frac{1}{nh_n}\sum_{i=1}^{n}\int_{-\infty}^{+\infty} Kh_n du$$

$$= \frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{+\infty} K(u)\,du = \tfrac{1}{n}(n) = 1$$

1.10 Théorème de Bochner :

Théorème 01 : Soit K : (R ^{m} , B ^{m}) $\rightarrow$ (R, B) une fonction mesurable, où B ^{m} est la tribu borélienne de R ^{m} , vérifant :

$\exists\ M$ (*constante*) telle que, $\forall\ z\ \in\ R^m\ ,\ |\ K(z)\ | \leq M,$

$$\int_{R^d} |\ k\,(z)\ |\ \ dz < \infty$$

et

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

ainsi que nous rappelons , soit

$g : (R^m, B^m)\ \rightarrow\ (R, B)$ une fonction mesurable telle que

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

On définit :

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

$$\text{où } \ 0 \ < \ h_n \ \rightarrow \ 0 \text{ quand } \ n \ \rightarrow \ \infty.$$

Si g est continue, alors

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

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

Chapter 2

Estimation de la fonction de régression dans le modèle de censure droite

Pour toute variable aléatoire V , on note

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

$$I_V = \inf\{t : F(t) \neq 0\}.$$

où F est la fonction de répartition de V.

2.1 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 lé'tude 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, δ_1),...., (X_n, Z_n, δ_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 chapitreprécédent ne peuvent plus être utilisés puisque Y n'est pas toujours observé.

2.2 Principe de l'estimation

L'idée, introduite par (Carbonez et all (1995)), et reprise par (Kohler, M·athé 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

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

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

Soit h une fonction de $R^d X \quad R \to R$. 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 :

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

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

$$E\left\{\frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i h(X_i,Y_i)}{H(Z_i)}\right\} = E\left\{\frac{I_{\{Y_1\le C_1\}}h(X_1,Y_1)}{H(Y_1)}\right\}$$

$$= E\left(\frac{h(X_1,Y_1)}{H(Y_1)}E\left(I_{\{Y_1\le C_1\}}/(X,Y)\right)\right)$$

$$= E(h(X_1,Y_1)).$$

•Le problème est que H est inconnu.

l'estimateur de Kaplan -Meier (1958) donné par :

$$\hat{H}_n(t) = \begin{cases} \prod_{i=1}^{n}\left[1-\frac{1-\delta_{(i)}}{n-i+1}\right]^{I_{\left[Z_{(i)}\le t\right]}} & , si \quad t < T_{k,n,} \\ \lim s \to T_{k,n,} s \ < \ T_{k,n,}\hat{H}_n(s) \ , & si \quad t \ge T_{k,n,} \end{cases}$$

où $T_{k,n} = \max\{Z_1, ..., Z_n\}$ et les paires $\left(Z_{(i)}, \delta_{(i)}\right)$, $i = 1, ..., n$ sont les n paires observées $\left(Z_{(i)}, \delta_{(i)}\right)$ ordonnées en $Z_{(i)}$,i.e. $Z_{(1)} \leq Z_{(2)} \leq ... \leq Z_{(n)} = T_{k,n,}$.

Remarquons que $\hat{H}_n$ a été légérement modifié afin de ne jamais s'annuler. Cela suggère d'estimer $r(x)$ par

$$\hat{r}_n(x) = \sum_{i=1}^{n} W_{n;i}(x) \frac{\delta_i Z_i}{\hat{H}_n(Z_i)}, \qquad (2.1)$$

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

$$W_{n;i}(x) = \frac{K\left(\frac{(x - X_i)}{h_n}\right)}{\sum_{j=1}^{n} K\left(\frac{(x - Xj)}{h_n}\right)}.$$

$\forall t$ / $0 \leq t \leq \infty$ et $x \in \mathsf{R}$, on définit

$$T_{[0,1]}(x) = \begin{cases} t\ , & si \quad x > t, \\ x\ , & si \quad 0 \leq x \leq t, \\ 0\ , & si \quad x < 0. \end{cases}$$

Pour $f : R^d \to R$ définissons $T_{[0,1]} f : R^d \to R$ par

$$\left(T_{[0,1]} f\right)(x) = T_{[0,1]}\left(f\left(x\right)\right)$$

.

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

$$r_n(x) = \begin{cases} T_{k,n}\ , & si \quad \hat{r}_n(x) > T_{k,n}, \\ r_n(x) = \hat{r}_n(x), & si \quad 0 \leq \hat{r}_n(x) \leq T_{k,n}, \\ 0\ , & si \quad \hat{r}_n(x) < 0. \end{cases} \tag{2.2}$$

Par analogie avec (2.2) , posons

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

2.3 Propriétés de l'estimateur

Le résultat Gill et Johansen.

Lemma 2.1.1 : On a

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

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

Lemma 2.1.2 : Sous l'hypothèse $(H_{1:2})$, on a

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

si et seulement si

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

preuve $\left|r_n^*(x)-r(x)\right|^2=\left|r_n^*(x)-r_n(x)+r_n(x)-r(x)\right|^2$

$$\leq 2\left|r_n^*(x)-r_n(x)\right|^2+2\left|r_n(x)-r(x)\right|^2$$

$$\int_{R^d} |r_n^*(x) - r(x)|^2 \mu(dx) \leq 2 \int_{R^d} |r_n^*(x) - r_n(x)|^2 \mu(dx)$$

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

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

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

On a: $T_{k,n} \leq T_Y$ p.s.

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

$$\leq T_Y - T_{k,n}$$

$$\int_{R^d} |r_n^*(x) - r_n(x)|^2 \mu(dx) \leq \int_{R^d} (T_Y - T_{k,n})^2 \mu(dx)$$

$$= (T_Y - T_{k,n})^2,$$

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

Théorème 2.1.1: (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 $(2.1), (2.2)$ vérifie :

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

Preuve

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

Posons

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

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

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

$$\int_{R^d} |r_n^*(x) - r(x)|^2 \, \mu\,(dx)$$

$$\leq 2 \int_{R^d} |r_n^*(x) - \bar{r}_n(x)| \, \mu\,(dx) + 2 \int_{R^d} |\bar{r}_n(x) - r(x)|^2 \, \mu\,(dx)\,. \qquad (2.3)$$

nous substituons $\bar{r}_n(x)$

$$\int_{R^d} |\bar{r}_n(x) - r(x)|^2 \, \mu\,(dx) \leq \int_{R^d} \left| \sum_{i=1}^n \frac{K\left(\frac{(x - X_i)}{h_n}\right)}{\sum_{j=1}^n K\left(\frac{(x - Xj)}{h_n}\right)} \frac{\delta_1 Z_1}{H_1\,(Z_1)} - r(x) \right|^2 \mu\,(dx)\,.$$

De plus $0 \leq \delta_1 \qquad Z_1 = H(Z_1) \;\leq\; T_Y \;= H\;\; (T_Y)$. et

$$E\left\{\frac{\delta_1 Z_1}{H\,(Z_1)}/X_1\right\} = \; E\left\{\frac{I_{\{Y_1 \leq C_1\}} Y_1}{H\,(Y_1)}/X_1\right\}$$

$$= E\left(\frac{Y_1}{H\,(Y_1)} E\left(I_{\{Y_1 \leq C_1\}}/(X_1, Y_1)\right)\right.$$

$$= \; E\,(Y_1/X_1) = r(X_1)$$

Donc d'aprés le théorème de (Devroye et Krzy· zak (1989)) on obtient

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

$$\int_{R^d} \sum_{i=1}^{n} \frac{K\left(\frac{(x-X_i)}{h_n}\right)}{\sum_{j=1}^{n} K\left(\frac{(x-Xj)}{h_n}\right)} \cdot \frac{\delta_1 Z_1}{H(Z_1)} - r(x)^2 \mu(dx) \to 0 \ ,(n \to \infty) \text{ p.s.} \tag{2.4}$$

Reste à majorer le premier terme de l'inégalité donnée à la formule (2.2)

$$\int_{R^d} |r_n^*(x) - \bar{r}(x)|^2 \, \mu(dx)$$

$$\leq T_Y \int_{R^d} \sum_{i=1}^{n} \frac{K\left(\frac{(x-X_i)}{h_n}\right)}{\sum_{j=1}^{n} K\left(\frac{(x-Xj)}{h_n}\right)} \frac{\delta_1 Z_1}{\hat{H}_n(Z_1)} - \sum_{i=1}^{n} \frac{K\left(\frac{(x-X_i)}{h_n}\right)}{\sum_{j=1}^{n} K\left(\frac{(x-Xj)}{h_n}\right)} \frac{\delta_1 Z_1}{H(Z_1)} \mu(dx)$$

$$\leq T_Y \int_{R^d} \sum_{i=1}^{n} \frac{K\left(\frac{(x-X_i)}{h_n}\right)}{\sum_{j=1}^{n} K\left(\frac{(x-Xj)}{h_n}\right)} T_Y \frac{1}{\hat{H}_n(Z_i)} - \frac{1}{H(Z_i)} \mu(dx)$$

$$\leq T_Y^2 \frac{1}{\hat{H}_n(T_Y) H(T_Y)} \sup_{t \leq T_Y} \left| H(t) - \hat{H}_n(t) \right| \to 0 \quad (n \to \infty) \text{ p.s.}$$

à cause de ($H_{1,2}$)et du lemme 2.1.1.

Signalons le fait que Guessoum et Ould said (2009) ont modifié légérement $\hat{H}_n$ en lui imposant de s'annuler à partir de léobservation 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 U64aautre part sa normalité asymptotique.

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

Chapter 3

Etude asymptotique

3.1 Convergence en loi

Définition 3.1.1 : Soit (X_n) et X des vecteurs aléatoires â valeurs dans l'espace probabilisable (R^P, B_{R^P}). On dit que la suite(X_n) converge en loi vers X si, pour toute fonction h de R^p vers R, continue et bornée, on a

$$\lim_{n\to\infty} E\left[h\left(x_n\right)\right] = Eh\left(x\right).$$

On note $X_n \xrightarrow{L} X$ et on dit aussi parfois que la loi de X_n converge vers celle de X .

Proposition 3.1.1 : Soit (X_n) et X des v.a.r. de fonction de répartition (F_n) et F respectivement. La suite (X_n) converge en loi vers X si, et seulement si,

$$\lim_{n\to\infty} F_n\left(x\right) = F\left(x\right).$$

en tout point x où F est continue.

Exemple On considère la suite (X_n) de v.a.r. telle que, pour tout n, la v.a.r. X_n ait pour loi

$$P\left(X_n = 2 + \tfrac{1}{n}\right) = 1$$

i.e. la loi de X_n est la dirac en $1 + \frac{2}{n}(P_{X_n} = \delta_{2+\frac{1}{n}})$. En raison de la convergence de la suite $(2 + \frac{1}{n})$ vers 2, on a :

$$\forall x > 2, \exists n_0 : \forall n > n_0, 2 + \tfrac{1}{n} < x.$$

$$\forall x > 2, \exists n_0 : \forall n > n_0, F_n(x) = P(X_n \leq x) = 1$$

Par ailleurs, pour tout x ≤ 2, on a :

$$F_n(x) = P(X_n \leq 2) = 0.$$

Définissons alors X la v.a.r. de loi $\delta 2$. Sa fonction de répartition est alors :

$$F_n(x) = \begin{cases} 0 & si\ x < 2; \\ \\ 1 & si\ \ x \geq 2. \end{cases}$$

On remarque que la fonction F_X est continue sur $R \backslash \{2\}$ et que, sur cet ensemble, on a :

$$\lim_{n \to \infty} F_n(x) = F(x).$$

Ainsi, d'après la proposition précédente, on a la convergence de $X_n$vers X . Il est intéressant de noter que la convergence des fonctions de répartition n'a pas lieu au point de discontinuité de F puisque l'on a, pour tout n,

$$F_n(2) = 0 \neq F(2) = 1$$

Théorème 3.1.1: Soit (X_n) et X des vecteurs aléatoires de R^p, absolument continus dedensité $(f_{n,X})$ et f par rapportâ la mesure de Lebesgue dans R^p. Si on a, λ_p-presquepartout,

$$\lim_{n\to\infty} f_{n,x} = f$$

alors

$$X_n \xrightarrow{L} X$$

Théorème (Théorème de Paul Lévy) 3.1.2 :

1. Si (X_n) est une suite de variables aléatoires dans R^p convergeant en loi vers une variable aleatoire X dans R^p , alors la suite (φ_{X_n}) des fonctions caractéristiques associée â la suite (X_n) converge en tout point vers la fonction caractéristique φ_X de X , i.e.

$$X_n \xrightarrow{L} X \Rightarrow \forall x \in R^p, \varphi_{X_n}(x) \to \varphi_X(x).$$

2. Soit (X_n) est une suite de variables aléatoires dans R^p. Si la suite (φ_{X_n}) de ses fonctions caractéristiques converge simplement vers une fonction φ continue en 0,alors ' est la fonction caractéristique d'une variable aléatoire X et X_n converge en loi vers X , i.e.

$$\varphi_{X_n}(x) \to \varphi_X(x) \ , \forall x \in R^p \Rightarrow X_n \xrightarrow{L} X \ ,$$

Théorème (Théorème de Cramer-Wold) 3.1.3 :

Soit (X_n) et X des vecteurs aléatoiresdansR^p. On a alors l'équivalence suivante :

$$X_n \xrightarrow{L} X \Leftrightarrow \forall u \in R^p : u'X_n \xrightarrow{L} u'X \ ,$$

Preuve. Supposons en premier lieu que X_n converge en loi vers X . La fonction g de R^p vers R définie par $g(x) = u\prime x$, pour u dans est une forme linéaire. Elle est donc continue. Ainsi, d'après le théorème de Slutsky, on a la convergence

$$u'X_n \xrightarrow{L} u'X$$

Réciproquement, supposons que pour tout u dans R^p, on ait

$$u'X_n \xrightarrow{L} u'$$

Le théorème de Paul Lévy, nous donne alors la convergence

$$\varphi_{u'X_n}(t) \to \varphi_{u'X}(t)$$

pour tout t dans R. Celle-ci prise en $t = 1$, nous donne :

$$\varphi_{u'X_n}(1) = \varphi_{X_n}(u) \to \varphi_X(u) \ = \varphi_{u'X}(1)$$

dont on tire, en utilisant la réciproque du théorème de Paul Lévy, la convergence $X_n \xrightarrow{L} X$.

3.2 Convergence presque sûre

Définition 3.2.1 : On dit que la suite (X_n) de v.a.r. converge presque sûrement vers X s'il existe un élément A de la tribu A tel que $P(A) = 1$ et

$$\forall \omega \in A \lim_{n \to \infty} X_n(\omega) = X(\omega)$$

On note

$$X_n \overset{p.s}{\to} X.$$

Théorème 3.2.1 : La suite de v.a.r. (X_n) converge presque sûrement vers X si la suite dev.a.r. (Y_m) définie par :

$$Y_m = \sup_{n \geq m} | \ X_n - X. |$$

alors (Y_m) converge en probabilité vers 0.

Proposition 3.2.1: Si, pour tout ε strictement positif, la série de terme général $P[|X_n| > \varepsilon]$ est convergente, i.e.

$$\forall \varepsilon > 0 \ \ \sum_n P[|X_n| > \varepsilon] < +\infty$$

alors (X_n) converge presque sûrement vers zéro.

3.3 Estimation de la foction de régression

Introduction

Le modéle de la régression est l'un des modèle les plus fréquemment rencontrés en statistique paramétrique et non paramétrique. Soient (X_1,Y_1); (X_2, Y_2),....,(X_n, Y_n) dans couples de variables aléatoire indépendantes et de même loi que (X,Y), de densité jointe f (x, y) sur $\mathbb{R}^2$et une densité marginale f (x) $>$ 0.Dans le modèle de régression non paramétrique on suppose l'existence d'une fonction "r" qui exprime la valeur moyenne de la variable à expliquer Y en fonction de la variable fonctionnelle explicative X , c'est-à-dire :

$$Y = r(X) + \varepsilon,$$

où "ε " est une variable centrée réduite et indépendante de X .

Définition 3.3.1:

Soit (X_i,Y_i)$_{i=1,...,n}$ et n paires, i.i.d. comme (X ,Y) évaluées dans $E \times \mathbb{R}$, où (E, d) est un espace semi-métrique.

On définit l'opérateur de régression "r" par :

$$r(x) = E[Y/X = x] :$$

3.3.1 Présentation de l'estimateur :

Nous proposons pour l'opérateur non linéaire r, l'estimateur de régression du noyau fonctionnel défini par :

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

où K est un noyau asymétrique et h est strictement positif réel.

On pose :

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

Il est facile de réecrire l'estimateur de noyau comme suit :

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

avec,

$$\sum_{i=1}^{n} W_{i;h}(x) = 1$$

3.3.2 Hypothèses

$$\mathrm{r} \in \mathrm{C}_E^0 \tag{3.1}$$

où

$$\mathrm{C}_E^0 = \left\{ f : E \to \mathbb{R}\ , lim_{d(x,x')\to 0} f(x') = f(x) \right\}.$$

$$\forall \epsilon > 0, P(X \in B(X, \varepsilon)) = \varphi(\epsilon) > 0. \tag{3.2}$$

$$\begin{cases} h \text{ est positivetelleque :} \\ \lim_{n\to\infty} h = 0 \text{ et } \lim_{n\to\infty} \frac{\log n}{\varphi(h)} \\ K \text{ est une noyau de type } I \\ o\grave{u} \\ K \text{ est une noyau de type } II \end{cases} \tag{3.3}$$

$\forall m \geq 2, E(|Y^m| / X = x) < \sigma_m(x) < \infty$ avec $\sigma_m(.)$ continue a x . (3.4)

Une fonction K de $\mathbb{R}$ en $\mathbb{R}^+$telle que $\int k = 1$,appelé noyau de type I s'il existe deux constantes réelles $0 < c_1 < c_2 < \infty$ tell que :

$$c_1 1_{[0;1]} \leq K \leq c_2 1_{[0;1]}$$

$$\exists\ c_3 > 0, \exists\ \epsilon_0, \forall \varepsilon < \epsilon_0, \int_0^{\varepsilon} \varphi_x(u)du > c_3 \epsilon \varphi_x(\varepsilon) \qquad (3.5)$$

$\exists\ \beta > 0$, telle que :

$$r \in \text{LipE},\beta, \qquad (3.6)$$

où

$$\text{LipE},\beta = \left\{ f : E \to \mathbb{R}, \exists C \in \mathbb{R}_*^+, \forall x' \in E, |f(x) - f(x')| < C d(x, x')^{\beta} \right\}$$

.

3.3.3 La convergence presque complète ponctuelle

Théorème 3.4.1:

Dans le modèle de type de continuité (3.1) avec la probabilité conditionnlle (3.2), si l'estimateur vérifie (3.3) et si la variable réponse Y satisfait (3.4), alors nous avons :

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

Démonstration.

En utilisant la notation de Δ_i défini par :

$$\Delta_i = \frac{K\left(h^{-1}d\ \ (x - X_i)\right)}{EK\left(h^{-1}d\ \ (x - X_i)\right)},$$

Soient $\hat{r}_1(x)$ et $\hat{r}_2(x)$ les quantités suivantes :

$$\hat{r}_1(x) = \frac{1}{n}\sum_{i=1}^{n} \Delta_i$$

$$\hat{r}_2(x) = \frac{1}{n}\sum_{i=1}^{n} Y_i\Delta_i$$

Nous avons clairement $\hat{r}(x)=\frac{\hat{r}_2(x)}{\hat{r}_1(x)}$ telle que :

$$\hat{r}_1(x) = \frac{1}{nE\left[K\left(h^{-1}d\ \ (x - X_i)\right)\right]}\sum_{i=1}^{n} K\left(h^{-1}d\ \ (x - X_i)\right)$$

et

$$\hat{r}_2(x) = \frac{1}{nE\left[K\left(h^{-1}d\ \ (x - X_i)\right)\right]}\sum_{i=1}^{n} K\left(h^{-1}d\ \ (x - X_i)\right) Y_i$$

En utilisant la décomposition suivante :

$$\hat{r}(x)-r(x) = \frac{1}{\hat{r}_1(x)}\left\{(\hat{r}_2(x) - E\hat{r}_2(x)) - (r(x) - E\hat{r}_2(x))\right\} - \frac{r(x)}{\hat{r}_1(x)}\left\{\hat{r}_1(x) - 1\right\}.$$

Lemme 1:
sous les hypothéses (3.1) et (3.3), on a :

$$\lim_{n\to\infty} E\ \hat{r}_2(x) = r(x)$$

Démonstration.

$$r(x) - E\ \hat{r}_2(x) = r(x) - E\ (Y_1\Delta_1),$$

$$= r(x) - E\ (E(Y_1\Delta_1/X_1)),$$
$$= r(x) - E\ (r(X_1)\Delta_1),$$

Puis que le support de la fonction de noyau K est [0 ,1], nous avons :

$$|r(x) - r(X_1)|\Delta_1 \leq \sup_{x'\in B(x,h)} |r(x) - r(x')|\Delta_1,$$

et l'hypothése de continuité sur r permet d'obtenir le résultat revendiqué.

Lemme 2 :

solon les hypothéses(3.2) et (3.4) et (3.3), nous avons :

$$\hat{r}_2(x) - E\ \hat{r}_2(x) = O_{p.co.}\left(\sqrt{\frac{\log n}{n\varphi_x(h)}}\right).$$

Démonstration.

On pose, pour i = 1,..., n ,$K_i = K\left(h^{-1}d\ (x - X_i)\right)$,

La démonstration de ce résultat est basée sur l'utilisation d'une inégalité exponentielle de type Berntein.effectivement,

$$\mathbf{P}\left(|\hat{r}_2(x) - E\hat{r}_2(x)| > \epsilon\right) = \mathbf{P}\left(\frac{1}{n}\left|\sum_{i=1}^{n}\left(Y_i\Delta_i - E\left(Y_i\Delta_i\right)\right)\right| > \epsilon\right).$$

Et nous devons montrer qu'il existe $\epsilon > 0$ telle que :

$$\sum_{n\in\mathbb{N}^*} \mathbf{P}\left(\frac{1}{n}\left|\sum_{i=1}^{n}\left(Y_i\Delta_i - E\left(Y_i\Delta_i\right)\right)\right| > \epsilon_0\sqrt{\frac{\log n}{n\varphi_x(h)}}\right) < \infty$$

Donc, nous appliquons l'inégalité exponentielle $Z_i = Y_i\Delta_i - EY_1\Delta_1$

$$\exists C > 0, \forall m = 2, 3, ..., \left|E\left(Y_1\Delta_1 - E\left(Y_1\Delta_1\right)\right)\right| \leq C\varphi_x\left(h\right)^{-m+1}$$

Tout d'abord, nous prouvons que pour $m \geq 2$.

$$E\left|Y_1\right|^m \Delta_1^m = O\left(\varphi_x\left(h\right)^{-m+1}\right). \tag{3.7}$$

Pour cela, nous écrivons :

$$E\left|Y_1\right|^m \Delta_1^m = \frac{1}{(EK_1)^m}\left\{E\left|Y_1\right|^m K_1^m\right\},$$

$$= \frac{1}{(EK_1)^m}\left\{E\left(E\left(\left|EY_1\right|^m X\right) K_1^m\right)\right\},$$

$$= \frac{1}{(EK_1)^m}\left\{E\sigma_m\left(X\right) K_1^m\right\}$$

$$= \frac{1}{(EK_1)^m}\left\{E\left(\left(\sigma_m\left(X\right) - \sigma_m\left(x\right)\right) K_1^m\right) + \sigma_m\left(x\right) EK_1^m\right\}.$$

Ce qui implique que :

$$|E\left|Y_1\right|^m \Delta_1^m| \leq E\left|\sigma_m\left(X\right) - \sigma_m\left(x\right)\right| \Delta_1^m + \sigma_m\left(x\right) E\Delta_1^m$$

$$\leq \left(\sup_{x' \in B(x,h)} \left|\sigma_m\left(x'\right) - \sigma_m\left(x\right)\right| \right) E\Delta_1^m + \sigma_m\left(x\right) E\Delta_1^m$$

Parce que $0 < \int K^m < \infty$,si K est du type I(resp II) puis $K^m / \int K^m$ est également du type I(resp II).

Donc, en appliquant lemme on :

$$C_1 \varphi_x(h) \leq E K_1^m \leq C_2 \varphi_x(h) \tag{3.8}$$

En utilisant (3.8) peut l'écrire pour m = 2; 3,.... :

$$\frac{C_1}{\varphi_x(h)^{m-1}} \leq E\Delta_1^m \leq \frac{C_2}{\varphi_x(h)^{m-1}}$$

Ce qui implique que

$$|E\left|Y_1\right|^m \Delta_1^m| = O\left(\varphi_x(h)^{m-1}\right)$$

De plus, nous avons :

$$(Y_1\Delta_1 - E(Y_1\Delta_1))^m = \sum_{k=0}^{m} c_{k,m} (Y_1\Delta_1)^k E(Y_1\Delta_1)^{m-k} (-1)^{m-k}$$

où

$$c_{k,m} = \frac{m\,!}{(k!(m-k)!)},$$ce qui implique que

$$E|Y_1\Delta_1 - E(Y_1\Delta_1)|^m \leq C \sum_{k=0}^{m} c_{k,m} E(Y_1\Delta_1)^k (r(x))^{m-k}$$

$$\leq C \max_{k=0,1,2,\ldots,m} E(Y_1\Delta_1)^k$$

$$\leq C \max_{k=0,1,2,\ldots,m} \varphi_x(h)^{-k+1}$$

La dernière inégalité utilise (3.7) pour K $\geq$2 alors que pour K = 1 on peut montrer que $E|Y|\Delta_1 = O(1)$ qu'en suivant les mêmes étape que celles de la preuve de Lemme 1 Parce que ' (h) tend vers zéro avec n, il devient

$$E|Y_1\Delta_1 - E(Y_1\Delta_1)|^m = O\left((\varphi_x(h))^{-m+1}\right)$$

Ensuite, nous avons $u_n = (a^2 \log n)/n = \log n/(n\varphi_x(h))$ avec $a^2 = \varphi_x(h)^{-1}$, il est clair que un tend vers zéro à n en utilisant l'hypothèse (3.3)

Lemme 3

Solon les hypothéses(3.2) et (3.3), on a :

$$\hat{r}_1(x) - 1 = O_{p.co.}\left(\sqrt{\frac{\log n}{n\varphi_x(h)}}\right).$$

Démonstration.

pour i = 1,..., n , et $Y_1 = 1$ on utiliser inégalité exponentielle de type Berntein.donc,

$$\mathbf{P}\left(|\hat{r}_1(x) - 1| > \epsilon\right) = \mathbf{P}\left(\frac{1}{n}\left|\sum_{i=1}^{n} \Delta_i - E\Delta_i\right| > \epsilon\right)$$

Et nous devons montrer qu'il existe $\epsilon > 0$ telle que :

$$\sum_{n\in\mathbb{N}^*} \mathbf{P}\left(\frac{1}{n}\left|\sum_{i=1}^{n} \left(\Delta_i - E\left(\Delta_i\right)\right)\right| > \epsilon_0\sqrt{\frac{\log n}{n\varphi_x(h)}}\right) < \infty$$

Donc, nous appliquons l'inégalité exponentielle $Z_i = \Delta_i - EY_1\Delta_1$Pour ce faire, nous devons d'abord montrer :

$$\exists C > 0, \forall m = 2, 3, ..., \left|E\left(Y_1\Delta_1 - E\left(Y_1\Delta_1\right)\right)\right| \leq C\varphi_x\left(h\right)^{-m+1}$$

Tout d'abord, nous prouvons que pour $m \geq 2$.

$$E\left|Y_1\right|^m \Delta_1^m = O\left(\varphi_x\left(h\right)^{-m+1}\right). \tag{3.9}$$

Pour cela, nous écrivons :

$$E\left|Y_1\right|^m \Delta_1^m = \frac{1}{(EK_1)^m}\left\{E\left|Y_1\right|^m K_1^m\right\},$$

$$= \frac{1}{(EK_1)^m}\left\{E\left(E\left(\left|EY_1\right|^m X\right) K_1^m\right)\right\},$$

$$= \frac{1}{(EK_1)^m}\left\{E\sigma_m\left(X\right) K_1^m\right\}$$

$$= \frac{1}{(EK_1)^m}\left\{E\left(\left(\sigma_m\left(X\right) - \sigma_m\left(x\right)\right) K_1^m\right) + \sigma_m\left(x\right) EK_1^m\right\}.$$

Ce qui implique que :

$$\left|E\left|Y_1\right|^m \Delta_1^m\right| \leq E\left|\sigma_m\left(X\right) - \sigma_m\left(x\right)\right| \Delta_1^m + \sigma_m\left(x\right) E\Delta_1^m$$

$$\leq \left(\sup_{x' \in B(x,h)} \left|\sigma_m\left(x'\right) - \sigma_m\left(x\right)\right|\right) E\Delta_1^m + \sigma_m\left(x\right) E\Delta_1^m$$

Parce que $0 < \int K^m < \infty$,si K est du type I(resp II) puis $K^m / \int K^m$ est également du type I(resp II).

Donc, en appliquant lemme on :

$$C_1\varphi_x(h) \leq EK_1^m \leq C_2\varphi_x(h) \tag{3.10}$$

En utilisant (3.9) , on peut l'écrire pour m = 2; 3,.... :

$$\frac{C_1}{\varphi_x(h)^{m-1}} \leq E\Delta_1^m \leq \frac{C_2}{\varphi_x(h)^{m-1}}$$

Ce qui implique que

$$\left(Y_1\Delta_1 - E\left(Y_1\Delta_1\right)\right)^m = \sum_{k=0}^{m} c_{k,m}\left(Y_1\Delta_1\right)^k E\left(Y_1\Delta_1\right)^{m-k}(-1)^{m-k}$$

où

$c_{k,m} = \frac{m\,!}{(k!(m-k)!)}$,ce qui implique que

$$E\left|Y_1\Delta_1 - E\left(Y_1\Delta_1\right)\right|^m \leq C\sum_{k=0}^{m} c_{k,m} E\left|Y_1\Delta_1\right|^k \left|r\left(x\right)\right|^{m-k}$$

$$\leq C \max_{k=0,1,2,...,m} E\left|Y_1\Delta_1\right|^k$$

$$\leq C \max_{k=0,1,2,...,m} \varphi_x(h)^{-k+1}$$

La dernière inégalité utilise (3.7) pour K $\geq$2 alors que pour K = 1 on peut montrer que $E\left|Y\right|\Delta_1 = O(1)$ qu'en suivant les mêmes étape que celles de la preuve de Lemme 1 Parce que ' (h) tend vers zéro avec n, il devient

$$E\left|Y_1\Delta_1 - E\left(Y_1\Delta_1\right)\right|^m = O\left(\left(\varphi_x(h)\right)^{-m+1}\right)$$

Ensuite, nous avons $u_n = \left(a^2 \log n\right)/n = \log n/\left(n\varphi_x(h)\right)$ avec $a^2 = \varphi_x(h)^{-1}$, il est clair que un tend vers zéro à n en utilisant l'hypothèse (3.3)

Chapter 4

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 et pour différents modèles.

Nous choisissons le noyau gaussien

$$K(t) = \frac{1}{\sqrt{2}} \exp\left(-\frac{t^2}{2}\right)$$

4.1 Modéle linéaire

Soit$Y_i = \alpha X_i + \alpha_0 + \beta\varepsilon_i$ 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

$$\beta = 1 \ \alpha = 0.8 \ \ et \ \alpha_0 \ = 3.$$

On a $r(x) = E(Y = X = x) = \alpha x + \alpha_0 = 0.8x + 3$

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

Application sous R:

```
rn(list=ls(all=TRUE))
n=100
X=rnorm(n,0,2)
E=rnorm(n)
Y=3+0.8*X+E
K=function(t){(1/sqrt(2*pi))*exp(-0.5*t^2)}
h=n^-.2
s=100
a=min(X)
b=max(X)
x=seq(a,b,length=s)
V=numeric(n)
fn=numeric(s)
for(j in 1 :s){
for(i in 1 :n){ V[i]=K((x[j]-X[i])/h) }
fn[j]=sum(V)/(n*h)}
W=numeric(n)
Hn=numeric(s)
for(j in 1 :s){
for(i in 1 :n){ W[i]=K((x[j]-X[i])/h)*Y[i] }
Hn[j]=sum(W)/(n*h)}
Rn =Hn/fn
op=par(mfrow=c(1,3))
plot(x,Rn,xlab="x", ylab="Rn(x)", main="n=100",type='l',col=4, lwd= 2)
abline(3,.8,lwd= 2)\bigskip
```

Pour n =500

```
n=500
X=rnorm(n,0,2)
E=rnorm(n)
Y=3+0.8*X+E
h=n^-.2
V=numeric(n)
for(j in 1 :s){
for(i in 1 :n){ V[i]=K((x[j]-X[i])/h) }
fn[j]=sum(V)/(n*h)}
W=numeric(n)
for(j in 1 :s){
for(i in 1 :n){ W[i]=K((x[j]-X[i])/h)*Y[i] }
Hn[j]=sum(W)/(n*h)}
Rn =Hn/fn
plot(x,Rn,xlab="x", ylab="Rn(x)", main="n=500",type='l',col=4, lwd= 2)
abline(3,0.8,lwd= 2)\bigskip
```

Données complètes

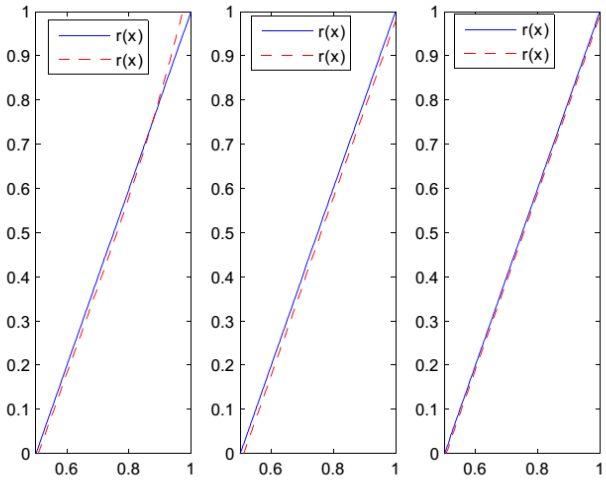

FIG 4.1−cas linéaire: vaec n = 100 ,500 ,1000

Données censurées à droite

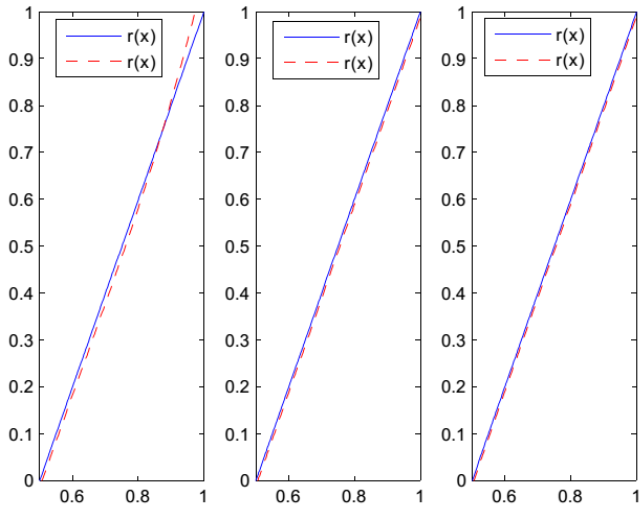

*Fig.*4.2−cas linéaire avec n = 100, 500 ,1000.

Données censurées à gauche

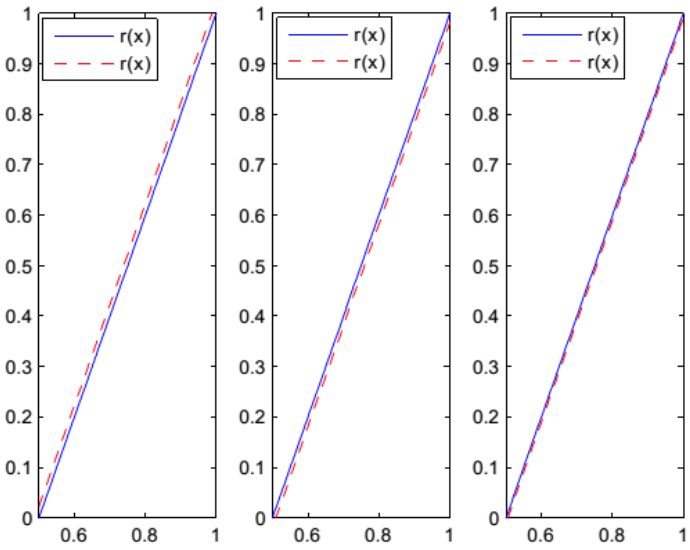

Fig.4.3-Cas linéaire avec n=100,500,1000

Bibliographie

[1] Frédéric Ferraty et philippe Vieu. Non Parametric Functional Data Analysis.Springer Series in statistics, 2006.

[2] Ferraty et Vieu. Locally modelled regression and functional data.Barrientos-Marin, 2010

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

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

[5] 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).

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

[7] Kohler M., Mathé K., Pintér M. (2002). Prediction from randomly right censored data. Journal of Multivariate Analysis, 80 : 73-100.

[8] Ould-Saïd E., Lemdani M. (2006). Asymptotic properties of a nonparametric regression function estimator with randomly truncated data. J. of the Institute of Statistical Mathematics, 58 : 357-378..

[9] Carbon, M., Francq, C., Sarda, P. (2007). Kernel regression estimation for random fields. J. Statist. Plann. Inference.

[10] FERRATY, F., RABHI, A., VIEU, P. (2008). Estimation non paramétrique de la fonction de hasard avec variable explicative fonctionnelle. Rom. J. Pure et Applied Math. 52, 1-18.

[11] Delsol, L. (2007). Régression non paramétrique fonctionnelle :expression asymptotique des moments, Ann. I.S.U.P. Vol LI, 3, 43-67.

[12] Laksaci, A., Madani, F., Rachdi, M.(2010). Kernel conditional density estimation when the regressor is valued in a semi metric space. International Statistical Review. (In press).

[13] Henriques, C., Oliveira, P. E. (2005). Exponential rates for kernel density estimation under association. Statist. Neerlandica 59 (4) : 448–466

[14]Kuczmaszewska, A. (2009). On complete convergence for arrays of rowwise negatively associated random variables. Statist. Probab. Lett. 79 : 116–124.

Conclusion

Dans ce mémoire, on a présenté la méthode d'estimation à noyau, qui permettant d'effectuer de la régression non paramétrique. Ce travail a montré que la méthode d'estimation de régression non paramétrique est simple et peut être très utile dans plusieurs situations. Par exemple, dans l'analyse des données, lorsque l'on désire comprendre et observer les relations qui existent entre les variables.

Dans la régression non paramétrique, la méthode du noyau joue un grand rôle. Pour que son soit plus utilisée par les praticiens, il est nécessaire que les programmes informatiques permettant d'appliquer ces méthodes soient facilement accessibles et assez simples d'utilisation. Cela favorise aussi les échanges entre statisticiens et utilisateur. L'estimateur à noyau de la régression non paramétrique d'pend de deux paramètres le noyau K et le paramètre de lissage h.

Dans la pratique, on utilisé le logiciel R pour présenté des exemples sur cet estimateur, et à travers les résultats obtenus, nous concluons que : le noyau K est peu influence sur l'estimateur, par contre le paramètre h est un grand influence, et dont le choix est crucial

Résumé

Ce mémoire porte sur l'étude 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.

Finalement, nous donnons des explications graphiques des résultats théoriques appliqués sur des exemples de régression linéaire à l'aide du logiciel R.

travail de simulation a permis de vérifier la bonne performance des estimateurs étudiés.

Abstract

This dissertation covers the study of kernel estimators of the regression function in different contexts, namely for complete data (real and functional) as well as for right-censored data.

Finally, we give graphic explanations of the theoretical results applied to linear regression examples using the R software.

simulation work made it possible to verify the good performance of the estimators studied.