République algérienne démocratique et populaire
Ministère de l'enseignement supérieur et de la recherche scientifique
==========

Université Constantine 1
Faculté des sciences exactes
==========

Département de mathématiques

N° d'ordre : 66/DS/2014
N° de série : 05/mat/2014

THÈSE

présentée pour l'obtention
du
diplôme de doctorat
en
mathématiques

« Estimation non-paramétrique de la fonction de régression : cas d'un modèle de censure mixte »

Par
Khedidja Kebabi

OPTION
Probabilités et Statistique

Devant le jury :

Président	M.	Z. Mohdeb	Professeur	Université Constantine 1
Directrice de thèse	M^{me}	F. Messaci	Professeur	Université Constantine 1
Examinatrice	M^{me}	N. Seddik Ameur	Professeur	Université BM Annaba
Examinatrice	M^{me}	A. Chadli	M. C.	Université BM Annaba

Soutenue le : 18 juin 2014

Remerciements

Je tiens en premier lieu à remercier sincèrement ma directrice de thèse, madame F. Messaci dont je salue les qualités humaines et scientifiques.

J'adresse mes remerciements sincères et chaleureux à monsieur Z. Mohdeb qui me fait l'honneur de présider le jury de soutenance.

Merci infiniment à madame A. Chadli, et à madame N. Seddik, pour l'honneur qu'elles me font d'examiner mon travail, et d'avoir accepté de faire le déplacement malgré leurs multiples responsabilités.

Je remercie grandement ma collègue et amie Nahima Nemouchi pour toute l'aide qu'elle m'a apportée avec autant de gentillesse et de générosité.

Merci aussi à mon jeune collègue et grand ami Abderrahim Kitouni, tout particulièrement pour ses remarques pertinentes et ses apports à mes programmes informatiques.

Table des matières

Introduction

Le but de la régression est de déterminer la manière dont l'espérance d'une variable expliquée réelle Y dépend d'une variable explicative X scalaire, vectorielle ou même fonctionnelle (ce qui veut dire qu'elle prend ses valeurs dans un espace de dimension infinie). Nous cherchons le lien entre X et Y modélisé par la fonction r vérifiant

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

où ε est l'erreur supposée centrée et indépendante de X, ce qui permet de montrer que $r(X) = E(Y/X)$. Le problème consiste donc à déterminer (ou plutôt à estimer) pour chaque réalisation x de la variable X, la valeur de la fonction $r(x)$.

Pour caractériser cette fonction, une première approche consiste à utiliser un modèle de régression paramétrique. On suppose que cette fonction peut s'écrire comme une fonction explicite des valeurs de X. Cette dernière peut être linéaire, par exemple

$$r(x) = \alpha + \beta x,$$

et on cherche alors à déterminer les meilleures valeurs des paramètres α et β compte tenu d'un critère, par exemple celui des moindres carrés. Nous nous ramenons alors à l'estimation d'un nombre fini de paramètres. Dans certains cas nous pouvons disposer pour cette estimation d'un échantillon $\{(X_i, Y_i), i = 1, \ldots, n\}$ de couples indépendants et ayant chacun la même

loi que (X, Y). Souvent, l'utilisation d'un modèle paramétrique n'est pas justifiée ; il est alors possible de se suffire de la seule donnée de l'échantillon pour réaliser une estimation. Ce sera à l'aide d'un modèle nonparamétrique. Dans ce cas on ne dispose d'aucune forme paramétrique pour r mais seulement d'hypothèses générales de régularité comme la dérivabilité.

Etudier le lien entre deux variables a généralement pour but de prédire l'une d'entre elles (variable réponse) étant donné une valeur de l'autre (variable explicative). Il y a plusieurs façons d'aborder un problème de prévision et l'une des plus utilisées est la régression qui est basée sur l'espérance conditionnelle. Pour des raisons de robustesse, deux techniques alternatives sont la médiane conditionnelle et le mode conditionnel. La prédiction au moyen de la médiane conditionnelle nécessite l'estimation préalable de la fonction de répartition conditionnelle. Celle du mode conditionnel nécessite l'estimation de la densité conditionnelle. Ces deux méthodes de prévision ont été largement étudiées. Citons, parmi les innombrables travaux qui leur ont été consacrés Gannoun *et al.* (2003),Samanta et Thavaneswaran (1990),Khardani *et al.* (2011) et Collomb *et al.* (1987).

Plusieurs paradigmes d'estimation non-paramétrique de la régression sont disponibles dont, l'estimation à poids, l'estimation des moindres carrés, et l'estimation des moindres carrés pénalisés ou spline de lissage. L'estimation à poids se décline en plusieurs modèles dont l'estimateur à partition, l'estimateur des k plus proches voisin et l'estimateur à noyau de Nadaraya-Watson. Ce dernier, qui a été introduit indépendamment par Nadaraya (1964) et Watson (1964), est l'un des plus populaires des modèles de régression non-paramétriques. Son expression est

$$r_{NW}(x) = \sum_{i=1}^{n} Y_i \frac{K((x - X_i)/h_n)}{\sum_{i=1}^{n} K((x - X_i)/h_n)},$$

où K est une fonction de $\mathbb{R}$ dans $\mathbb{R}$ et h_n est un paramètre réel strictement positif, appelé paramètre de lissage et dont le choix est essentiel.

Malheureusement dans la pratique, il n'est pas toujours possible d'avoir à disposition des données complètes. La fixation du temps de l'étude par exemple peut empêcher l'observation de la variable d'intérêt pour laquelle on saura seulement qu'elle dépasse la valeur observée, c'est le cas de la censure à droite. C'est pourquoi cet estimateur a été généralisé au le cas où la variable réponse est censurée à droite par Kohler *et al.* (2002). Guessoum et Ould Saïd (2008, 2010, 2012) ont étudié cet estimateur (qu'ils ont légèrement modifié) aussi bien dans le cas où les X_i sont iid que dans le cas où les X_i sont α-mélangeantes.

Un phénomène de censure à gauche (symétrique du précédent) peut aussi empêcher l'observation du phénomène d'intérêt pour lequel on saura seulement qu'il est inférieur à la valeur observée. Généralement, la censure à gauche s'accompagne de la censure à droite comme cela est le cas pour la censure mixte à laquelle nous nous intéressons dans cette thèse.

Un exemple d'un tel modèle, donné dans Patilea et Rolin (2006), est de considérer un système de fiabilité qui consiste en 3 composants C_1, C_2 et C_3 avec C_1 et C_2 en série et C_3 en parallèle avec le système en série. Les variables X, R et L, respectivement les durées de vie de C_1, C_2 et C_3, sont indépendantes et on peut déterminer quel composant est tombé en panne en même temps que le système. Un autre exemple a été donné par Morales *et al.* (1991) concernant la mort d'arbres plantés par parcelle dans une ferme. Le modèle mathématique est présenté au chapitre 1.

Dans ce contexte, Patilea et Rolin (2006) donnent un estimateur produit-limite de la fonction de survie de la durée d'intérêt X qui généralise le célèbre estimateur de Kaplan et Meier (1958) et démontrent sa convergence uniforme presque sûre ainsi que sa normalité asymptotique sous certaines conditions d'identifiabilité. Messaci et Nemouchi (2011) précisent la vitesse de cette convergence qui est d'ordre $\sqrt{\log\log n/n}$. Shen (2011, 2012) propose deux estimateurs alternatifs à celui de Patilea et Rolin (2006). Volgushev (2009) a donné des estimateurs du quantile conditionnel.

En ce qui concerne la régression, Messaci (2010) a introduit des estimateurs à poids de la régression, dont l'estimateur à noyau. Kebabi *et al.* (2011) ont donné des estimateurs des moindres carrés et montré leur convergence vers la valeur optimale presque sûrement. Dans cette thèse, nous donnons un taux de convergence presque complète ponctuelle et uniforme de l'estimateur à noyau. Nous y montrons aussi sa normalité asymptotique.

Organisation du document

Dans le chapitre 1, nous présentons le modèle de censure mixte et l'estimation de la fonction de survie. Dans le chapitre 2, nous exposons les lois du logarithme itéré dont nous avons besoin dans la suite et essentiellement le résultat récent de Messaci et Nemouchi (2011) relatif à une loi du logarithme itéré de l'estimateur de la fonction de survie sous censure mixte. Le chapitre 3 présente des estimateurs à poids de la fonction de régression lorsque la variable réponse est soumise à une censure mixte élaborés par Messaci (2010) dont l'estimateur à noyau, objet d'intérêt de cette thèse. Les chapitres 4 et 5, sont dédiés à l'étude de ce dernier et constituent l'apport de cette thèse. Plus précisément au chapitre 4 , nous donnons une vitesse de convergence presque complète aussi bien ponctuelle qu'uniforme de l'estimateur à noyau. Au chapitre 5 est montrée sa normalité asymptotique. Le chapitre 6 contient les différentes simulations qui confortent les résultats théoriques obtenus.

Chapitre 1

Estimation de la fonction de survie

1.1 Données censurées

Cette thèse s'intéressant à l'estimation dans un modèle de censure, commençons par présenter cette notion de censure.

En analyse de survie et en fiabilité, on s'intéresse au temps qui s'écoule jusqu'à la réalisation d'un certain événement. On appelle ce temps un temps de défaillance, la durée de vie, la durée de survie, ou simplement une durée. C'est une variable aléatoire positive et souvent supposée bornée. Cela peut être la durée de vie d'un patient après un traitement, la durée de chômage, le temps de panne d'un appareil, l'âge auquel un enfant apprend à accomplir une tâche donnée, etc. Il arrive souvent, pour diverses raisons, que la durée d'intérêt ne puisse pas être observée. Cela peut être dû à la perte de vue d'un patient, au début ou à la fin de la période d'étude, etc. Ces valeurs sont censurées. Les valeurs censurées, bien qu'inconnues, doivent être prises en compte pour obtenir des estimations correctes et des conclusions précises. Selon la situation spécifique, la littérature statistique

contient un grand nombre de procédures qui permettent de tenir compte des observations censurées. Il existe plusieurs types de censure.

Censure à droite Il y a censure à droite lorsque la durée de survie d'intérêt est supérieure à la durée observée. Un exemple typique est celui où l'événement d'intérêt est le décès d'un patient malade et la durée d'observation est une durée totale d'hospitalisation (on n'a plus de nouvelles après l'hospitalisation) . On peut aussi observer ce genre de phénomène dans des études de fiabilité quand la panne d'un appareil ne permet pas de poursuivre l'observation pour l'appareil objet de notre étude.

Pour ce type de censure, tout ce que l'on sait est que la vraie durée est supérieure à la durée observée.

Censure à gauche Il y a censure à gauche lorsque la durée de survie est inférieure à la durée observée. En fiabilité, l'exemple d'une telle situation est celui d'un composant électronique monté en parallèle avec un ou plusieurs autres composants. Une panne de ce composant n'entraîne pas nécessairement l'arrêt du système : le système peut continuer à fonctionner (bien que pouvant présenter certaines anomalies) jusqu'à ce que cette panne soit détectée (par exemple lors d'un contrôle ou en cas de l'arrêt du système). La durée observée pour ce composant est alors censurée à gauche.

Il y a des situations où les données d'un même échantillon peuvent être censurées soit à droite, soit à gauche. Par exemple, Herbert Leiderman *et al.* (1973) ont étudié l'âge auquel les enfants d'une communauté africaine apprennent à accomplir certaines tâches. Au début de l'étude, certains enfants savaient déjà effectuer les tâches étudiées, on sait seulement alors que l'âge où ils ont appris est inférieur à leur âge à la date du début de l'étude. A la fin de l'étude, certains enfants ne savaient pas encore accomplir ces tâches et on sait alors seulement que l'âge auquel ils ont appris

est supérieur à leur âge à la fin de l'étude.

Dans cet exemple, on trouve dans un même échantillon des données censurées à gauche aussi bien que des données censurées à droite. Selon le processus qui génère la censure, on est en présence de censure double (Turnbull, 1974) ou mixte (Patilea et Rolin, 2006).

Le modèle de Patilea et Rolin (2006) et celui de Turnbull (1974), bien que similaires, ne sont pas identiques. Dans le modèle de Turnbull (1974), on suppose que T est indépendante du couple (R, L) et que $P(L < R) = 1$ alors que dans le modèle de Patiléa, T,R et L sont indépendantes, si on appelle T la variable d'intérêt, L la variable de censure à gauche, et R la variable de censure à droite. La similatité vient du fait qu 'on observe $Z = \max(min(T, R), L)$.

Censure par intervalle Dans le cas de la censure par intervalle, on observe à la fois une borne inférieure et une borne supérieure de la durée d'intérêt. Ceci arrive dans des études de suivi médical où les patients sont contrôlés périodiquement, si un patient ne se présente pas à un ou plusieurs contrôles et se présente ensuite après que l'événement d'intérêt se soit produit. 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. Turnbull (1976) présente ce genre de censure avec plus de détails.

Les catégories de censure décrites ci-dessus peuvent se décliner en fonction du mode ou du mécanisme de censure. On obtient alors les types suivants :

Censure de type I L'expérimentateur fixe une date (non aléatoire) de fin d'expérience. La durée de participation maximale est alors fixée (non aléatoire) et vaut, pour chaque observation, la différence entre la date de fin d'expérience, et la date d'entrée du patient dans l'étude. Le nombre d'événements observés est, quant à lui, aléatoire. Ce modèle est souvent utilisé dans les études épidémiologiques.

Censure de type II L'expérimentateur fixe a priori le nombre d'événements à observer. La date de fin d'expérience devient alors aléatoire, le nombre d'événements étant, quant à lui, non aléatoire. Ce modèle est souvent utilisé dans les études de fiabilité.

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, par exemple, à la durée d'hospitalisation du patient). Ici, le nombre d'événements observés et la durée totale de l'expérience sont aléatoires.

Un problème important en statistique est celui de l'estimation de la fonction de répartition qui décrit complètement la loi de probabilité des observations. Ce qui revient à estimer la fonction de survie qui est le complément à 1 de la fonction de répartition. L'étude d'un tel estimateur est aussi justifiée par le fait qu'il intervient explicitement dans l'expression de l'estimateur de la régression que nous nous proposons d'étudier.

Si nous disposons d'un échantillon $(X_1, \ldots, X_n)$ de vraies réalisations de la variable X de fonction de répartition F, alors un estimateur naturel de F est la fonction de répartition empirique, donnée par

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} 1_{]-\infty,x]}(X_i).$$

C'est un estimateur sans biais et de variance minimale $\sigma^2(x) = F(x)(1 - F(x))$. Le théorème central limite assure sa normalité asymptotique et le théorème de Donsker (1952) donne un résultat plus fort qui est sa convergence en tant que processus vers un gaussien.

Le théorème de Glivenko-Cantelli donne sa convergence uniforme presque sûre, autrement dit $\sup_{x\in\mathbb{R}} |F_n(x) - F(x)| \to 0$ presque sûrement. Pour les données α-mélangeantes, ce résultat est amélioré par Collomb *et al.* (1985) en convergence presque complète sous certaines hypothèses.

Ceci est valable quand on dispose d'un échantillon $X_1, \ldots, X_n$. Mais dans la pratique, un échantillon est souvent censuré. Dans le cas de la censure à droite, Kaplan et Meier (1958) ont proposé un estimateur de la fonction de survie $(1 - F(x))$, qui se réduit au complément à 1 de la fonction de répartition empirique lorsque les observations sont complètes.

1.2 Estimateur de Kaplan-Meier

Soit $X_1, \ldots, X_n$ un échantillon représentant les durées d'intérêt (ces variables sont donc supposées positives), de fonction de répartition F, et $C_1, \ldots, C_n$ un échantillon représentant les temps de censure, que l'on suppose indépendants des durées d'intérêt, de fonction de répartition G. Dans le modèle de censure aléatoire à droite, on observe non pas la durée d'intérêt X_i mais plutôt la plus petite des deux valeurs $Z_i = \min(X_i, C_i)$, ainsi que l'indicateur de censure δ_i qui vaut 1 si la durée d'intérêt est observée, et 0 si elle est censurée, i.e. $\delta_i = 1_{\{X_i \leq C_i\}}$.

Dans ce genre de données, qui sont souvent des durées de survie ou des données de fiabilité, la fonction de répartition F est estimée par l'estimateur introduit par Kaplan et Meier (1958), donné pour $z < Z_{(n)}$ où

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

$$F_n(z) = 1 - \prod_{i:Z_i \leq z} \left(\frac{N_n(Z_i) - 1}{N_n(Z_i)} \right)^{\delta_i}$$

avec $N_n(x) = \sum_{i=1}^{n} 1_{\{Z_i \geq x\}}$. Pour $z \geq Z_{(n)}$, il y a plusieurs conventions pour définir $F_n(z)$: Soit on le définit par $F_n(Z_{(n)})$, ce qui fait que F_n peut ne pas être une fonction de répartition si $Z_{(n)}$ est une donnée censurée, soit on le définit par 0, soit on le laisse non défini.

Cet estimateur a des propriétés assez similaires à celles la fonction de répartition empirique : la convergence uniforme presque sûre (Stute et Wang, 1993; Winter *et al.*, 1978), la normalité asymptotique (Breslow et Crowley, 1974; Gill, 1983), et la loi du logarithme itéré (Földes et Rejtő, 1981).

Le cas de la censure mixte fera l'objet des sections suivantes.

1.3 Modèle de censure mixte

Considérons trois variables aléatoires positives indépendantes X, L et R de fonctions de répartition respectives F_X, F_L et F_R, et de fonctions de survie respectives S_X, S_L et S_R, où X représente la durée d'intérêt et L et R sont les durées de censure à gauche et à droite respectivement. Dans le modèle I de Patilea et Rolin (2006), au lieu d'observer un échantillon de X on observe un échantillon du couple (Z, A) où $Z = \max(\min(X, R), L)$ et

$$A = \begin{cases} 0, & \text{si } L < X \leq R, \\ 1, & \text{si } L < R < X, \\ 2, & \text{si } \min(X, R) \leq L. \end{cases}$$

Ce modèle considère la censure à droite et la censure à gauche comme deux phénomènes qui agissent indépendamment l'un de l'autre mais que

l'un peut censurer l'autre. Patilea et Rolin (2006) ont proposé d'estimer S_X, fonction de survie de la variable d'intérêt X comme suit.

1.4 Estimation

Considérons H la fonction de répartition de Z, elle peut s'écrire $\sum_{k=0}^{2} H^{(k)}(t)$ où

$$H^{(k)}(t) = P(Z \leq t, A = k), \quad \text{pour } k = 0, 1, 2.$$

En notant pour toute application R de $\mathbb{R}$ dans $\mathbb{R}$, $R(t_-)$ la limite de R à gauche de t, lorsque cette limite existe, ces fonctions s'écrivent

$$\begin{aligned}
H^{(0)}(t) &= \int_0^t F_L(u_-) S_R(u_-)\, dF_X(u), \\
H^{(1)}(t) &= \int_0^t F_L(u_-) S_X(u)\, dF_R(u), \\
H^{(2)}(t) &= \int_0^t \{1 - S_X(u) S_R(u)\}\, dF_L(u),
\end{aligned}$$

et c'est à partir de ces équations que l'estimateur est obtenu.

L'idée est de considérer dans un premier temps $Y = \min(X, R)$ et L dans un modèle de censure à gauche (c'est-à-dire que l'on considère une donnée complète si $A = 0$ ou $A = 1$ et censurée à gauche si $A = 2$), et d'estimer la fonction de répartition de Y, puis l'utiliser pour estimer la fonction de répartition de la variable d'intérêt X en considérant un modèle de censure à droite.

L'estimateur de la fonction de survie S_X ainsi obtenu, en remplaçant à la fin les fonctions $H^{(0)}$, $H^{(1)}$ et $H^{(2)}$ par leurs estimateurs empiriques, obtenus à partir d'un échantillon $(Z_i, A_i)_{1 \leq i \leq n}$, est donné par :

$$\hat{S}_n(Z'_j) = 1 - F_n(Z'_j) = \prod_{1 \leq l \leq j} \left\{ 1 - \frac{D_{0l}}{U_{l-1} - N_{l-1}} \right\},$$

où $(Z'_j)_{1\leq j\leq M}$ sont les valeurs distinctes des Z_i prises dans l'ordre croissant, et

$$D_{kj} = \sum_{1\leq i\leq n} 1_{\{Z_i=Z'_j, A_i=k\}}, \quad N_j = \sum_{1\leq i\leq n} 1_{\{Z_i\leq Z'_j\}},$$

$$U_{j-1} = n \prod_{j\leq l\leq M} \left\{1 - \frac{D_{2l}}{N_l}\right\}$$

pour $0 \leq l \leq 2$ et $1 \leq j \leq M$.

Soulignons le fait que si $L \equiv 0$ (pas de censure à gauche), $\hat{S}_n$ se réduit à l'estimateur de Kaplan-Meier qui lui-même se réduit au complément à 1 de la fonction de répartition empirique si $R \equiv \infty$.

Patilea et Rolin (2006) ont introduit cet estimateur et montré sa convergence uniforme presque sûre et sa convergence en tant que processus vers un gaussien sous des conditions d'identifiabilité du modèle.

$\hat{S}_n$ intervenant plus loin au dénominateur dans les expressions d'estimateurs de la fonction de régression, le résultat suivant donne une condition nécessaire et suffisante pour qu'il s'annule.

Dans un souci de clarté, nous notons dans la suite de ce chapitre D_{kj} par $D_{k,j}$.

Lemme 1.1. *i) Une condition nécessaire et suffisante pour que $\hat{S}_n(Z'_{k_0}) = 0$ pour la première fois et reste nul est : $D_{0,k_0} \neq 0$, $D_{1,k_0} = 0$ et $\forall j > k_0, D_{0,j} = D_{1,j} = 0$ si $k_0 \neq M$.*

ii) $\hat{S}_n$ s'annule pour la première fois en Z'_M si et seulement si $D_{0,M} \neq 0$ et $D_{1,M} = 0$

Démonstration. 1. Commençons par montrer que pour tout $k : 0 \leq k \leq M-1$, nous avons

$$U_k \geq N_k. \tag{1.1}$$

En effet

$$\begin{aligned} U_k &= n\left(\frac{N_{k+1}-D_{2,(k+1)}}{N_{k+1}}\right)\left(\frac{N_{k+2}-D_{2,(k+2)}}{N_{k+2}}\right)\cdots\left(\frac{N_M-D_{2,M}}{N_M}\right) \\ &= \left(\frac{N_k+D_{0,(k+1)}+D_{1,(k+1)}}{N_{k+1}}\right)\cdots\left(\frac{N_{M-2}+D_{0,(M-1)}+D_{1,(M-1)}}{N_{M-1}}\right) \\ &\quad \times (N_{M-1}+D_{0,M}+D_{1,M}) \\ &\geq \frac{N_k}{N_{k+1}}\times\cdots\times\frac{N_{M-2}}{N_{M-1}}\times N_{M-1}=N_k. \end{aligned}$$

Remarquons que s'il existe j tel que $j>k$ avec $D_{1,j}\neq 0$ ou $D_{0,j}\neq 0$ alors $U_k > N_k$.

2. Soit k_0 le premier indice k tel que $D_{0,k}=U_{k-1}-N_{k-1}$ (le premier k pour lequel $\hat{S}_n(Z'_k)=0$). Nous avons alors :

$$D_{0,k_0}\neq 0 \text{ et } D_{0,k_0}=U_{k_0-1}-N_{k_0-1}. \tag{1.2}$$

Par ailleurs

$$U_{k_0-1}=n\left(1-\frac{D_{2,k_0}}{N_{k_0}}\right)\cdots\left(1-\frac{D_{2,M}}{N_M}\right)=\left(1-\frac{D_{2,k_0}}{N_{k_0}}\right)U_{k_0}. \tag{1.3}$$

D'après (1.2) et (1.3), il vient

$$\begin{aligned} D_{0,k_0}+N_{k_0-1} &= \left(\frac{N_{k_0}-D_{2,k_0}}{N_{k_0}}\right)U_{k_0} \\ &= \left(\frac{N_{k_0-1}+D_{0,k_0}+D_{1,k_0}}{N_{k_0}}\right)U_{k_0}, \end{aligned}$$

et en vertu de (1.1), nous devons avoir $D_{1,k_0}=0$, donc $U_{k_0=N_{k_0}}$ alors $D_{0,k_0}\neq 0$, $D_{1,k_0}=0$ et $\forall j>k_0, D_{1,j}=D_{0,j}=0$, (au-delà de k_0, ce qui montre que la condition énoncée est nécessaire

Montrons que la condition nécessaire est aussi suffisante. Supposons que $D_{1,k_0}=0$, $D_{0,k_0}\neq 0$ et $\forall j>k_0, D_{1,j}=D_{0,j}=0$, et montrons que $\hat{S}_n(Z'_{k_0})=$

0. Nous avons

$$\begin{aligned} U_{k_0-1} &= n\left(\frac{N_{k_0}-D_{2,k_0}}{N_{k_0}}\right)\left(\frac{N_{k_0+1}-D_{2,(k_0+1)}}{N_{k_0+1}}\right)\cdots\left(\frac{N_M-D_{2,M}}{N_M}\right) \\ &= n\left(\frac{N_{k_0-1}+D_{0,k_0}}{N_{k_0}}\right)\left(\frac{N_{k_0}}{N_{k_0+1}}\right)\cdots\left(\frac{N_{M-1}}{N_M}\right) \\ &= N_{k_0-1}+D_{0,k_0}, \end{aligned}$$

avec $D_{0,k_0} \neq 0$, autrement dit $\hat{S}_n(Z'_{k_0}) = 0$. □

Chapitre 2

Loi du logarithme itéré pour l'estimateur de la fonction de survie

La loi du logarithme itéré (notée LIL) pour une somme $S_n = \sum_{i=1}^{n} X_i$ de variables aléatoire indépendantes et de même loi, remonte à Khinchine et Kolmogorov dans les années 1920. Depuis, un grand nombre de travaux ont porté sur des lois du logarithme itéré pour différents estimateurs.

2.1 Loi du logarithme itéré classique

On considère une suite (X_n) de variables aléatoires i.i.d. centrées de variance 1, et $S_n = \sum_{i=1}^{n} X_i$. La loi forte des grands nombres donne une convergence presque sûre de $\frac{S_n}{n}$ vers 0. Le théorème de la limite centrale, donne par contre la convergence en loi de la suite $\frac{S_n}{\sqrt{n}}$ vers la loi normale $\mathcal{N}(0,1)$. La loi du logarithme itéré, donne un résultat pour une suite intermédiaire : $\frac{S_n}{\sqrt{n \log\log n}}$, d'où le nom qu'on lui attribue.

Théorème 2.1 (Loi du logarithme itéré)**.** *Soient (X_n) une suite de v.a. i.i.d. centrées de variance 1, et $S_n = \sum_{i=1}^{n} X_i$. Alors :*

$$\limsup_{n\to\infty} \frac{S_n}{\sqrt{n \log\log n}} = \sqrt{2} \quad p.s.$$

Ce résultat, dû à Hartman et Wintner (1941) a été d'abord prouvé pour des v.a de Bernoulli par Khinchine (1924), puis par Kolmogorov (1929) pour des v.a. de loi normale. Une preuve plus moderne de ce résultat est donné par de Acosta (1983).

La loi du logarithme itéré pour les fonctions de répartition empiriques a été démontrée indépendamment par Chung (1949) et Smirnov dans le cas des échantillons dans $\mathbb{R}$, et par Kiefer (1961) pour $\mathbb{R}^n$. D'autres lois du logarithme itéré relatives à différentes statistiques ont fait l'objet de plusieurs travaux. Citons, sans prétendre à l'exhaustivité, Hall (1981) qui a montré une LIL pour des estimateurs non-paramétriques de la densité, Hardle (1984) qui a montré une LIL pour des estimateurs non-paramétriques de la régression, et Földes et Rejtő (1981) qui ont montré une LIL pour l'estimateur de Kaplan-Meier de la fonction de survie pour des données censurées à droite, résultat que nous rappelons ci-dessous avec la LIL de Kiefer (1961) puisque nous allons nous en servir.

2.2 Lois du logarithme itéré pour des données complètes ou censurées à droite

Dans le cas où F est la fonction de répartition d'un vecteur de $\mathbb{R}^m$ et F_n est la fonction de répartition empirique associée (cas de données complètes), nous avons

Théorème 2.2.

$$P\left(\limsup_{n\to\infty}\frac{\sqrt{n}\sup_{x\in\mathbb{R}}|F_n(t)-F(t)|}{\sqrt{\frac{1}{2}\log\log n}}\leq 1\right)=1. \tag{2.1}$$

Dans le cas de la censure à droite, nous définissons l'estimateur de Kaplan-Meier, que nous notons par $\overline{F}_n$, en utilisant la deuxième convention page 12, c'est à dire en posant $F_n(z)=0$ pour $z>Z_{(n)}$. Pour toute fonction de répartition W, on note par $T_W=\sup\{t:W(t)<1\}$ le point terminal du support de W et par $I_W=\inf\{t:W(t)\neq 0\}$ son point initial.

En notant par F (resp. G) la fonction de répartition de la variable d'intérêt (resp. de la variable de censure), nous pouvons énoncer le résultat suivant

Théorème 2.3 (Földes et Rejtő (1981))**.** *On suppose que F et G sont continues, et que $T_F<T_G$, alors*

$$P\left(\sup_{-\infty<u<+\infty}|\overline{F}_n(u)-F(u)|=O\left(\sqrt{\frac{\log\log n}{n}}\right)\right)=1.$$

La condition $T_F<T_G$ pouvant paraître restrictive, citons le théorème autrement.

Corollaire 2.1 (Földes et Rejtő (1981))**.** *Si F et G sont continues, et si le réel T est tel que $G(T)<1$, alors*

$$P\left(\sup_{-\infty<u<T^*}|\overline{F}_n(u)-F(u)|=O\left(\sqrt{\frac{\log\log n}{n}}\right)\right)=1,$$

où $T^*=\min\{T,T_F\}$.

Remarquons que dans le cas de la censure à gauche, on observe $X\vee C$ et $\delta=1_{\{X\geq C\}}$. Dans ce cas, l'estimateur de la fonction de répartition F, noté $\hat{F}_n$, se déduit de celui de Kaplan-Meier en inversant le temps

$$\hat{F}_n(z)=\prod_{i:Z_i>z}\left(\frac{\tilde{N}_n(Z_i)-1}{\tilde{N}_n(Z_i)}\right)^{\delta_i},$$

avec $\tilde{N}_n(x) = \sum_{i=1}^n 1_{\{Z_i \leq x\}}$. Il vérifie le résultat suivant.

Corollaire 2.2. *Si F et G sont continues et si le réel T vérifie $G(T) > 0$, alors*

$$P\left(\sup_{T^*<u<\infty} |\hat{F}_n(u) - F(u)| = O\left(\sqrt{\frac{\log\log n}{n}}\right)\right) = 1,$$

où $T^* = \max\{T, I_F\}$.

Passons maintenant au cas de la censure mixte.

2.3 Loi du logarithme itéré de la fonction de survie en présence d'une censure mixte

Le modèle étudié ici est celui de la censure mixte introduit au chapitre 1, nous utilisons donc les mêmes notations, et nous noterons pour toute variable aléatoire U, F_U (resp. S_U) sa fonction de répartition (resp. de survie). Nous noterons aussi $T_U = \sup\{t : F_U(t) < 1\}$ et $I_U = \inf\{t : F_U(t) \neq 0\}$ les points terminaux du support de F_U, et nous supposons que les fonctions de répartition de X, R et L sont continues. Posons

$$\bar{S}_n(t) = \prod_{j/Z_j \leq t} \{1 - D_{0j}/(U_{j-1} - N_{j-1} + 1)\}. \tag{2.2}$$

C'est une modification nécessaire de $\hat{S}_n$. Soit $H(t) = P(Z < t)$ la fonction de répartition continue de l'observation Z, sa fonction de répartition empirique est $H_n(t) = \sum_{i=1}^n 1_{\{Z_i<t\}}/n$. La sous-distribution de Z,

$$\widetilde{F}(t) = P(Z \leq t, A = 0) = \int_0^t F_L(u) S_R(u) dF_X(u),$$

est la fonction de répartition du vecteur aléatoire à trois dimensions $(X, X-R, L-X)$ au point $(t, 0, 0)$. Sa fonction de répartition empirique est $\widetilde{F}_n(t) = \sum_{i=1}^n 1_{\{Z_i \leq t, A_i=0\}}/n$.

La LIL de Kiefer (théorème 2.2) s'applique à $\widetilde{F}_n$ et donne

$$P\left(\limsup_{n\to\infty} \frac{\sup_{u\in\mathbb{R}} |\widetilde{F}_n(u) - \widetilde{F}(u)|}{\sqrt{\log\log n/2n}} \leq 1\right) = 1. \tag{2.3}$$

En inversant le temps, l'estimateur produit-limite de F_L, qui est continue, est donné par $G_n(u) = \prod_{j/Z_j \geq u} \{1 - D_{2j}/N_j\}$. Remarquons que la LIL de Kiefer (1961) s'applique à H_n et que de plus le corollaire 2.2 s'applique à G_n (sous l'hypothèse $\sup(I_R, I_L) < I_X$) pour obtenir que pour presque tout ω il existe n_1 et un nombre fixé A tel que pour tout $n > n_1$

$$\sup_{I_X \leq u} |G_n(u) - F_L(u) - (H_n(u) - H(u))| \leq A\sqrt{\frac{\log\log n}{2n}}.$$

Maintenant en tenant compte du fait que $(F_L(u) - H(u)) \geq F_L(I_X)S_R(T_X)S_X(u)$ dès que $I_X \leq u \leq T_X$, nous déduisons sous les hypothèses $I_L < I_X$ et $T_R < T_X$ que pour tout $n > n_1$,

$$G_n(u) - H_n(u) \geq (F_L(u) - H(u))/2 \quad \text{p.s.}, \tag{2.4}$$

pour tout $I_X \leq u \leq u_n$ avec

$$u_n = S_X^{-1}\left(\frac{2A}{F_L(I_X)S_R(T_X)}\sqrt{\frac{\log\log n}{2n}}\right), \tag{2.5}$$

où $S_X^{-1}(s) = \sup\{x/S_X(x) > s\}$.

La mesure de hasard associée à X est $d\Lambda(t) = dF_X(t)/S_X(t)$ qui peut être écrite $d\widetilde{F}(t)/(F_L(t) - H(t))$ pour tout t tel que $I_X \leq t < T_X$. Pour $I_X \leq u < T_X$, on pose

$$T(u) = \int_{I_X}^{u} d\Lambda(t) = -\log(S_X(u)), \quad T_n(u) = \int_{I_X}^{u} d\tilde{F}_n(t)/(G_n(t) - H_n(t)),$$

où $T_n(u)$ est obtenue en remplaçant $\widetilde{F}$, F_L et H par leurs estimateurs dans l'expression de $T(u)$. Le théorème suivant donne la loi du logarithme itéré pour $\hat{S}_n$, estimateur de Patilea et Rolin (2006).

Théorème 2.4 (Messaci et Nemouchi (2011, 2013)). *Si S_X, S_R et S_L sont des fonctions de survies continues, et si* $\sup(I_L, I_R) < I_X$ *et* $T_X < T_R$. *Alors*

$$P\left(\sup_{-\infty<u<\infty}|\hat{S}_n(u) - S(u)| = O\left(\sqrt{\frac{\log\log n}{n}}\right)\right) = 1.$$

Remarquons que l'hypothèse $\sup(I_L, I_R) < I_X$ et $T_X < T_R$ assure l'identifiabilité du modèle étudié (cf. Patilea et Rolin, 2006).

La preuve du théorème est basée sur la décomposition suivante

$$\begin{aligned}|\hat{S}_n(u) - S_X(u)| \leq |\hat{S}_n(u) - \bar{S}_n(u)| \\ + |\bar{S}_n(u) - S_X(u)|,\end{aligned} \tag{2.6}$$

et nous avons

$$\begin{aligned}\bar{S}_n(u) - S_X(u) = (\exp\log\bar{S}_n(u) - \exp(-T_n(u))) \\ + (\exp(-T_n(u)) - \exp(-T(u))).\end{aligned}$$

En appliquant le développement de Taylor aux deux derniers termes de l'expression précédente, nous obtenons

$$\bar{S}_n(u) - S_X(u) = \exp(-\theta_n(u))(\log\bar{S}_n(u) + T_n(u)) \tag{2.7}$$

$$+ S_X(u)(T(u) - T_n(u)) \tag{2.8}$$

$$+ \frac{1}{2}\exp(-\theta'_n(u))(T(u) - T_n(u))^2, \tag{2.9}$$

où

$$\min\{-\log\bar{S}_n(u),\, T_n(u)\} \leq \theta_n(u) \leq \max\{-\log\bar{S}_n(u),\, T_n(u)\}, \tag{2.10}$$

et

$$\min\{T(u),\, T_n(u)\} \leq \theta'_n(u) \leq \max\{T(u),\, T_n(u)\}. \tag{2.11}$$

Nous allons maintenant énoncer et démontrer les quatre lemmes suivants. Le lemme 2.1 nous fournit un outil pour démontrer les lemmes 2.2, 2.3 et

2.4, nous permettant de traiter le premier terme du membre de droite de (2.6), et les membres de droite (2.7) et (2.8) respectivement.

Lemme 2.1. *Pour presque tout ω, il existe $n_0(\omega)$ tel que si $n > n_0$, alors pour tout $I_X \leq u \leq u_n$, $k_1 > 0$ et $k_2 \geq 0$ où $k = k_1 + k_2 > 1$, nous avons*

$$\int_{I_X}^{u} \frac{d\tilde{F}_n(t)}{(G_n(t) - H_n(t))^{k_1}(F_L(t) - H(t))^{k_2}} = O\left(\frac{n}{\log\log n}\right)^{\frac{k-1}{2}}.$$

Démonstration. Nous avons par (2.4), pour tout $n > n_1$, pour tout $I_X \leq u \leq u_n$ et tout $I_X \leq t \leq u$

$$(G_n(t) - H_n(t))^{k_1} \geq \left(\frac{(F_L(t) - H(t)}{2}\right)^{k_1} \quad \text{p.s.},$$

c'est à dire,

$$\begin{aligned}\frac{1}{(G_n(t) - H_n(t))^{k_1}} \frac{1}{(F_L(t) - H(t))^{k_2}} &\leq \frac{2^{k_1}}{(F_L(t) - H(t))^{k_1}} \frac{1}{(F_L(t) - H(t))^{k_2}} \\ &= \frac{2^{k_1}}{(F_L(t) - H(t))^{k_1+k_2}}.\end{aligned}$$

Posons $k = k_1 + k_2$, nous obtenons alors

$$\int_{I_X}^{u} \frac{2^{k_1} d\tilde{F}_n(t)}{(F_L(t) - H(t))^{k}} = \int_{I_X}^{u} \frac{2^{k_1} d\tilde{F}(t)}{(F_L(t) - H(t))^{k}} + \int_{I_X}^{u} \frac{2^{k_1} d(\tilde{F}_n(t) - \tilde{F}(t))}{(F_L(t) - H(t))^{k}}.$$

Étudions chacun des deux termes du membre de droite de l'expression précédente.

i) Rappelons que

$$d\tilde{F}(t) = -F_L(t)S_R(t)dS_X(t).$$

De plus, un calcul élémentaire montre que

$$F_L(t) - H(t) = F_L(t)S_R(t)S_X(t).$$

Nous pouvons donc majorer le premier terme comme suit

$$\begin{aligned}\int_{I_X}^{u}\frac{2^{k_1}d\tilde{F}(t)}{(F_L(u)-H(u))^k} &= \int_{I_X}^{u}\frac{-2^{k_1}F_L(t)S_R(t)d(S_X(t))}{(F_L(t)S_R(t)S_X(t))^k}\\ &\leq -\frac{2^{k_1}}{F_L^{k-1}(I_X)S_R^{k-1}(T_X)}\int_{I_X}^{u}\frac{d(S_X(t))}{S_X^k(t)}\\ &\leq \frac{2^{k_1}}{(k-1)F_L^{k-1}(I_X)S_R^{k-1}(T_X)S_X^{k-1}(u)}\end{aligned}$$

ii) Quant au second terme, nous avons

$$\begin{aligned}\int_{I_X}^{u}\frac{2^{k_1}d(\tilde{F}_n(t)-\tilde{F}(t))}{(F_L(t)-H(t))^k} &\leq \frac{2^{k_1}}{F_L^k(I_X)S_R^k(T_X)S_X^k(u)}\int_{I_X}^{u}\left|d(\tilde{F}_n(t)-\tilde{F}(t))\right|\\ &\leq \frac{2^{k_1+1}}{F_L^k(I_X)S_R^k(T_X)S_X^k(u)}\sup_{t\in\mathbb{R}}\left|\tilde{F}_n(t)-\tilde{F}(t)\right|.\end{aligned}$$

L'application de (2.5) montre que

$$S_X(u_n)\geq\frac{2A}{F_L(I_X)S_R(T_X)}\sqrt{\frac{\log\log n}{2n}}, \tag{2.12}$$

Puisque $u\leq u_n$, tenant compte de (2.2) et regroupant les deux termes, il vient

$$\begin{aligned}&\int_{I_X}^{u}\frac{d\tilde{F}_n(t)}{(G_n(t)-H_n(t))^{k_1}(F_L(t)-H(t))^{k_2}}\\ &\quad\leq\frac{2^{k_1}}{F_L^{k-1}(I_X)S_R^{k-1}(T_X)}\frac{1}{S_X^{k-1}(u)}\times\left(\frac{2}{A}+\frac{1}{k-1}\right) \qquad (2.13)\\ &\quad\leq\frac{2^{k_1}}{F_L^{k-1}(I_X)S_R^{k-1}(T_X)}\frac{F_L^{k-1}(I_X)S_R^{k-1}(T_X)}{2^{k-1}A^{k-1}}\times\left(\frac{2n}{\log\log n}\right)^{\frac{k-1}{2}}\left(\frac{2}{A}+\frac{1}{k-1}\right)\\ &\quad= O\left(\frac{2n}{\log\log n}\right)^{\frac{k-1}{2}},\end{aligned}$$

en tenant compte encore une fois de la relation (2.12).

Nous obtenons bien le résultat annoncé dans le lemme. □

Lemme 2.2. *Nous avons*

$$\sup_{I_X \leq u \leq u_n} |\hat{S}_n(u) - \bar{S}_n(u)| = O\left(\sqrt{\frac{1}{n \log\log n}}\right) \ p.s.$$

Démonstration. Rappelons l'inégalité suivante, dont nous allons nous servir pour majorer $|\hat{S}_n(u) - \bar{S}_n(u)|$. Si pour tout $1 \leq i \leq n$, $|a_i| \leq 1$ et $|b_i| \leq 1$ alors

$$\left|\prod_{i=1}^{n} a_i - \prod_{i=1}^{n} b_i\right| \leq \sum_{i=1}^{n} |a_i - b_i|,$$

Nous avons donc

$$\begin{aligned}
|\hat{S}_n(u) - \bar{S}_n(u)| &= \left|\prod_{j/Z_j \leq u} \left\{\frac{1 - D_{0j}}{(U_{j-1} - N_{j-1})}\right\} - \prod_{j/Z_j \leq u} \left\{\frac{1 - D_{0j}}{(U_{j-1} - N_{j-1} + 1)}\right\}\right| \\
&\leq \sum_{j/Z_j \leq u} \frac{D_{0,j}}{(U_{j-1} - N_{j-1})^2} \\
&= \sum_{j/Z_j \leq u} \frac{n\tilde{F}_n(Z_j)}{(nG_n(Z_j) - nH_n(Z_j))^2} \\
&= \int_{I_X}^{u} \frac{nd\tilde{F}_n(t)}{(nG_n(t) - nH_n(t))^2} \\
&= O\left(\sqrt{\frac{1}{(n \log\log n)}}\right) \quad p.s.,
\end{aligned}$$

en appliquant le lemme 2.1 pour $k_1 = 2$ et $k_2 = 0$. □

Lemme 2.3. *Nous avons*

$$\sup_{I_X \leq u \leq u_n} |\log \bar{S}_n(u) + T_n(u)| = O\left(\sqrt{\frac{1}{n \log\log n}}\right) \ p.s.$$

Démonstration. De (2.2), nous déduisons que

$$\log \bar{S}_n(u) = \int_{I_X}^{u} n \log(1 - \frac{1}{nG_n(t) - nH_n(t) + 1}) d\tilde{F}_n(t).$$

En outre le développement logarithmique implique que

$$\begin{aligned}
|\log \bar{S}_n(u) + T_n(u)| &= \left| \int_{I_X}^{u} \frac{d\tilde{F}_n(t)}{G_n(t) - H_n(t)} - \int_{I_X}^{u} n \sum_{l=1}^{\infty} \frac{1}{l} (nG_n(t) - nH_n(t) + 1)^{-l} d\tilde{F}_n(t) \right| \\
&\leq \left| \int_{I_X}^{u} \frac{d\tilde{F}_n(t)}{G_n(t) - H_n(t)} - \frac{d\tilde{F}_n(t)}{\frac{1}{n} + G_n(t) - H_n(t)} \right| \\
&\quad + \left| \int_{I_X}^{u} -n \sum_{l=2}^{\infty} \frac{1}{l} (nG_n(t) - nH_n(t) + 1)^{-l} d\tilde{F}_n(t) \right| \\
&\leq 2 \int_{I_X}^{u} \frac{d\tilde{F}_n(t)}{(nG_n(t) - nH_n(t))^2}.
\end{aligned}$$

Il reste à appliquer le lemme 2.1 pour obtenir le résultat visé. □

Lemme 2.4. *Nous avons*

$$\sup_{I_X \leq u \leq u_n} S_X(u) |T_n(u) - T(u)| = O\left(\sqrt{\frac{\log \log n}{n}} \right) \ p.s.$$

Démonstration. Remarquons que,

$$\begin{aligned}
|T_n(u) - T(u)| &\leq \left| \int_{I_X}^{u} \frac{(G_n(t) - H_n(t)) - (F_L(t) - H(t))}{(G_n(t) - H_n(t))(F_L(t) - H(t))} d\tilde{F}_n(t) \right| \\
&\quad + \left| \int_{I_X}^{u} \frac{d(\tilde{F}_n(t) - \tilde{F}(t))}{F_L(t) - H(t)} \right| \\
&\leq \sup_{I_X \leq t} |(G_n(t) - H_n(t)) - (F_L(t) - H(t))| \int_{I_X}^{u} \frac{d\tilde{F}_n(t)}{(G_n(t) - H_n(t))(F_L(t) - H(t))} \\
&\quad + \frac{2 \sup |\tilde{F}_n(t) - \tilde{F}(t)|}{F_L(I_X) S_R(T_X) S_X(u)}.
\end{aligned}$$

En vertu de (2.3) et (2.13), il s'ensuit que pour n assez grand

$$|T_n(u) - T(u)| \leq 2\sqrt{\log \log n / 2n} \frac{2A(\frac{2}{A} + 1) + (1 + \epsilon)}{F_L(I_X) S_R(T_X) S_X(u)}. \tag{2.14}$$

Compte tenu de (2.5), nous voyons qu'il existe une constante K, telle que

$$\sup_{I_X \leq u \leq u_n} |T_n(u) - T(u)| \leq K \text{ p.s.}$$

En revenant encore à (2.14), nous déduisons que

$$\sup_{I_X \leq u \leq u_n} S_X(u)|T_n(u) - T(u)| = O\left(\sqrt{\frac{\log\log n}{n}}\right) \text{ p.s.} \qquad \square$$

Nous sommes maintenant en mesure de donner la démonstration du Théorème 2.4.

Démonstration du Théorème 2.4. En vertu de (2.11), nous voyons que

$$\begin{aligned}\frac{1}{2}\exp(-\theta_n'(u))|T_n(u) - T(u)|^2 &\leq \frac{1}{2}S_X(u)|T_n(u) - T(u)|^2 \exp(|T_n(u) - T(u)|) \\ &\leq \frac{K}{2}S_X(u)|T_n(u) - T(u)| \exp(K).\end{aligned}$$

L'application immédiate du lemme 2.4 donne alors

$$\frac{1}{2}\exp(-\theta_n'(u))|T_n(u) - T(u)|^2 = O(\sqrt{(\log\log n)/n}) \text{ p.s.} \tag{2.15}$$

Par ailleurs, tenant compte de (2.10), il vient

$$\exp(-\theta_n(u))|\log \bar{S}_n(u) + T_n(u)| \leq |\log \bar{S}_n(u) + T_n(u)|.$$

Combinant les relations (2.9) et (2.15) avec les lemmes 2.3 et 2.4, nous pouvons conclure que, pour $I_X \leq u \leq u_n$

$$|S_X(u) - \bar{S}_n(u)| = O\left(\sqrt{(\log\log n)/n}\right) \text{ p.s.}$$

Cette dernière relation, combinée avec l'inégalité

$$|\hat{S}_n(u) - S_X(u)| \leq |\hat{S}_n(u) - \bar{S}_n(u)| + |S_X(u) - \bar{S}_n(u)|,$$

montre que

$$\sup_{I_X \leq u \leq u_n} |\hat{S}_n(u) - S_X(u)| = O(\sqrt{(\log\log n)/n})$$

pour n suffisamment grand.

La preuve du théorème est maintenant immédiate en combinant le dernier résultat avec la relation suivante

$$\sup_{u_n<u<+\infty} |\hat{S}_n(u) - S_X(u)| \leq |S_X(u_n)| + |\hat{S}_n(u_n) - S_X(u_n)|. \qquad \square$$

Nous terminons ce chapitre en donnant un taux de convergence presque complète de $\hat{S}_n$ (se reporter à l'appendice pour un rappel sur la convergence presque complète et le taux associé).

Sous les mêmes conditions que le théorème 2.4, nous avons

$$\sup_{-\infty<u<\infty} |\hat{S}_n(u) - S_X(u)| = O_{a.co}(\sqrt{(\log n)/n}). \tag{2.16}$$

Démonstration. Rappelons que $F_n = 1 - \hat{S}_n$ et $\Lambda(t) = \int_0^t \frac{dF_X(u)}{S_X(u)}$.
Pour tout t tel que $I_L < t < T_R$, $\Lambda(t)$ peut s'écrire

$$\Lambda(t) = \int_0^t \frac{d\widetilde{F}(u)}{F_L(u) - H(u)},$$

et peut donc s'estimer par

$$\Lambda_n(t) = \int_0^t \frac{d\widetilde{F}_n(u)}{G_n(u) - H_n(u)}.$$

L'utilisation de l'équation de Duhamel permet alors d'écrire pour tout $t \leq \theta < \min(T_R, T_X)$

$$|F_n(t) - F_X(t)| = (1 - F_X(t)) \left| \int_{I_X}^t \frac{1 - F_n(u^-)}{1 - F_X(u)}\, d(\Lambda_n - \Lambda)(u) \right|.$$

Posons $M_n(t) = \int_{I_X}^t d(\Lambda_n - \Lambda)(u)$ et intégrons par parties, nous obtenons

$$\begin{aligned}
|F_n(t) - F_X(t)| &\leq \left| \int_{I_X}^t \frac{1 - F_n(u^-)}{1 - F_X(u^-)}\, dM_n(u) \right| \\
&\leq \left| \frac{1 - F_n(t)}{1 - F_X(t)} M_n(t) - \int_{I_X}^t M_n(u)\, d\left(\frac{1 - F_n(u)}{1 - F_X(u)} \right) \right| \\
&\leq \frac{1}{1 - F_X(\theta)} \left| M_n(t) \right| + \left| \int_{I_X}^t M_n(u) \frac{dF_n(u)}{1 - F_X(u)} \right| \\
&\quad + \left| \int_{I_X}^t M_n(u) \left(1 - F_n(u^-)\right) \frac{dF_X(u)}{(1 - F_X(u))\,(1 - F_X(u))} \right| \\
&\leq \frac{2\,(1 - F_X(\theta)) + 1}{(1 - F_X(\theta))^2} \sup_{I_X \leq u \leq \theta} |M_n(u)| \,. \qquad (2.17)
\end{aligned}$$

Il reste donc à montrer que $\sup_{I_X \leq u \leq \theta} |\Lambda_n(t) - \Lambda(t)| = O_{a.co.}\left(\sqrt{\frac{\log n}{n}} \right)$.

$$\begin{aligned}
|\Lambda_n(t) - \Lambda(t)| &\leq \int_{I_X}^t \left| \frac{1}{G_n(u) - H_n(u)} - \frac{1}{F_L(u) - H(u)} \right| d(u) \\
&\quad + \left| \int_{I_X}^t \frac{1}{F_L(u) - H(u)} d(\widetilde{F}_n - \widetilde{F})(t) \right| \\
&=: B_{n,1}(t) + B_{n,2}(t). \qquad (2.18)
\end{aligned}$$

Etudions ces deux termes.

– Puisque $F_L(u) - H(u) = F_L(u) S_R(u) S_X(u)$, nous avons

$$\begin{aligned}
B_{n,1}(t) &\leq \frac{\sup_{I_X \leq u \leq \theta} |F_L(u) - G_n(u) + H_n(u) - H(u)|}{F_L(I_X) S_R(\theta) S_X(\theta)} \int_{I_X}^t \frac{d\widetilde{F}_n}{G_n(u) - H_n(u)} \\
&\leq 2 \frac{\sup_{I_X \leq u \leq \theta} |F_L(u) - G_n(u) + H_n(u) - H(u)|}{F_L(I_X) S_R(\theta) S_X(\theta) \inf_{I_X \leq u \leq \theta} |G_n(u) - H_n(u)|}.
\end{aligned}$$

Puisque $I_R < I_X$, le théorème 1 de Bitouzé *et al.* (1999) permet d'avoir

$$\sup_{I_X < t} |G_n(t) - F_L(t)| = O_{a.co.}\left(\sqrt{\frac{\log n}{n}} \right).$$

Par ailleurs, l'application de l'inégalité DKW (voir Dvoretzky *et al.* (1956)) donne

$$\sup_{t\in\mathbb{R}} |H_n(t) - H(t)| = O_{a.co.}\left(\sqrt{\frac{\log n}{n}}\right).$$

De plus, pour $\varepsilon_0 \in]0, F_L(I_X)S_R(\theta)S_X(\theta)[$, nous avons

$$P\left(\inf_{I_X\leq u\leq\theta} |G_n(u) - H_n(u)| < \varepsilon_0\right) \leq P\left(\sup_{I_X\leq u\leq\theta} |F_L(u) - G_n(u) + H_n(u) - H(u)| > \varepsilon\right),$$

où $\varepsilon = F_L(I_X)S_R(\theta)S_X(\theta) - \varepsilon_0$. Le terme à droite de l'inégalité est le terme général d'une série convergente. Nous pouvons donc affirmer que

$$\sup_{I_X\leq t\leq\theta} B_{n,1}(t) = O_{a.co.}\left(\sqrt{\frac{\log n}{n}}\right). \tag{2.19}$$

– Intégrons encore par parties, il vient

$$\begin{aligned}
B_{n,2}(t) &\leq \left|\frac{\widetilde{F}_n(t) - \widetilde{F}(t)}{F_L(t) - H(t)}\right| + \left|\frac{\widetilde{F}_n(I_X) - \widetilde{F}(I_X)}{F_L(I_X) - H(I_X)}\right| \\
&\quad + \left|\int_{I_X}^{t} \widetilde{F}_n(u) - \widetilde{F}(u))d\left(\frac{1}{F_L(u) - H(u)}\right)\right| \\
&\leq \frac{2}{F_L(I_X)S_R(\theta)S_X(\theta)} \sup_{I_X\leq u\leq\theta} \left|\widetilde{F}_n(u) - \widetilde{F}(u)\right| \\
&\quad + \left|\int_{I_X}^{t} \frac{\widetilde{F}_n(u) - \widetilde{F}(u)}{F_L(u)} d\left(\frac{1}{S_R(u)S_X(u)}\right)\right| \\
&\quad + \left|\int_{I_X}^{t} \frac{\widetilde{F}_n(u) - \widetilde{F}(u)}{S_R(u)S_X(u)} d\left(\frac{1}{F_L(u)}\right)\right| \\
&\leq D \sup_{I_X\leq u\leq\theta} \left|\widetilde{F}_n(u) - \widetilde{F}(u)\right|,
\end{aligned}$$

où D est une constante déterministe.

Il reste à appliquer le théorème $1 - m$ de Kiefer (1961) pour obtenir

$$\sup_{I_X\leq u\leq\theta} \left|\widetilde{F}_n(u) - \widetilde{F}(u)\right| = O_{a.co.}\left(\sqrt{\frac{\log n}{n}}\right). \qquad \square$$

Chapitre 3

Estimateurs non paramétriques à poids de la fonction de régression

Ce chapitre est le développement de l'article de Messaci (2010) dans lequel est introduit, entre autres, l'estimateur à noyau de la fonction de régression lorsque la variable réponse est soumise à une censure mixte, objet principal d'étude dans cette thèse.

Etant donné une covariable aléatoire X à valeurs dans $\mathbb{R}^d$, une variable réponse Y positive et deux variables de censure R et L positives, nous nous plaçons dans le contexte de la censure mixte exposé au chapitre 1. Plus précisément, Y est censurée et nous disposons seulement d'observations du triplet (X, Z, A) où $Z = \max(\min(Y, R), L)$ et

$$A = \begin{cases} 0, & \text{si } L < Y < R, \\ 1, & \text{si } L < R \leq Y, \\ 2, & \text{si } \min(Y, R) \leq L. \end{cases}$$

Notre but est de construire des estimateurs $r_n(x)$ de la fonction de régression $r(x) = E(Y/X = x)$, qui minimisent l'erreur quadratique moyenne,

ce qui revient à faire tendre $\int |r(x) - r_n(x)|^2 \mu(dx)$ vers zéro, où μ est la loi de probabilité de X.

Il est bien connu que lorsque Y n'est pas censurée, il existe des estimateurs (estimateurs à noyau, à partition, des k plus proches voisins, des moindres carrés, spline de lissage) pour lesquels $\int |r(x) - r_n(x)|^2 \mu(dx)$ converge en probabilité ou presque sûrement vers zéro. Voir par exemple Stone (1977), Devroye *et al.* (1994), Lugosi et Zeger (1995), Kohler (1997, 1999), Kohler et Krzyżak (2001), Györfi et Walk (1997) et Devroye *et al.* (1980). Les estimateurs à noyau, à partition et des k plus proches voisins font partie d'une classe d'estimateurs appelés *estimateurs à poids*, sur les quels nous revenons ci-dessous.

3.1 Estimateurs à poids

Une classe bien connue et très utilisée d'estimateurs non-paramétriques est la classe des estimateurs dits *à poids* qui s'écrivent

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

où les poids $W_{n,i}(x) = W_{n,i}(x, X_1, \ldots, X_n) \in \mathbb{R}$ dépendent de $X_1, \ldots, X_n$. Habituellement les poids sont positifs, leur somme ne dépasse pas 1 et $W_{n,i}(x)$ est petit si x est « éloigné » de X_i. L'idée d'estimer la régression de cette façon vient du raisonnement suivant : si plusieurs observations sont disponibles, on peut estimer $E(Y/X = x_0)$ par la moyenne des Y_i pour lesquelles les X_i correspondantes sont "assez proches" de x_0.

Estimateur des plus proches voisins Soit k_n un paramètre de l'estimation, nous posons

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

Estimateur à partition Nous utilisons une partition $\pi_n = \{A_{n,j} : j\}$ de $\mathbb{R}^d$ et nous posons

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

Estimateur à noyau Ici K étant un noyau et h_n une fenêtre, nous posons

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

3.1.1 Cas des données censurées à droite

Lorsque Y est censurée à droite c'est à dire qu'on a seulement des observations du triplet $\big(X_i, Z_i = \min(X_i, C_i), \delta_i = 1_{\{Y_i \leq C_i\}}\big)$, une idée de Carbonez *et al.* (1995) (qui a été reprise par Kohler *et al.* (2002)) est de remplacer $W_{n,i}(x)Y_i$ dans l' expressions des estimateurs par

$$W_{n,i}(x)\frac{\delta_i Z_i}{G_n(Z_i)}, \tag{3.1}$$

où G_n est l'estimateur de Kaplan-Meier de la fonction de survie G de C_i. Ceci est motivé par le fait que $\frac{1}{n}\sum_{i=1}^n W_{n,i}(x)\frac{\delta_i Z_i}{G(Z_i)}$ a pour moyenne $EW_{n,i}Y_i$ mais la fonction G étant inconnue, elle a été estimée par G_n. Ce qui a permis à Carbonez *et al.* (1995) de définir et d'étudier l'estimateur à partition puis à Kohler *et al.* (2002) d'introduire et d'étudier des estimateurs à poids, les estimateurs des moindres carrés ainsi que estimateurs spline de lissage.

3.1.2 Cas de la censure mixte

Lorsque Y est soumise à une censure mixte, nous estimons pour toute fonction $h : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}$, $Eh(X, Y)$ par

$$\frac{1}{n}\sum_{i=1}^{n}\frac{1_{\{A_i=0\}}Z_i}{\hat{S}_n(Z_i)\hat{F}_n(Z_i)}, \tag{3.2}$$

où $\hat{S}_n$ et $\hat{F}_n$ sont respectivement les estimateurs de S_R et F_L dont nous rappelons les expressions à la section 3.2.2. Ceci est motivé par le fait que

$$\begin{aligned} E\left(\frac{1_{\{A=0\}}h(X,Z)}{S_R(Z)F_L(Z)}\right) &= E\left(E\left(\frac{1_{\{A_i=0\}}h(X,Y)}{S_R(Z)F_L(Z)}\right)\Big/(X,Y)\right) \\ &= E\left(\frac{h(X,Y)}{S_R(Y)F_L(Y)}E\left(1_{\{A_i=0\}}/(X,Y)\right)\right) \\ &= E\left(h(X,Y)\right) \end{aligned}$$

car

$$E\left(1_{\{A_i=0\}}/X,Y\right) = S_R(Y)F_L(Y), \tag{3.3}$$

si les hypothèses $H_{1,1}$, $H_{1,2}$, $H_{1,4}$ ci-dessous sont satisfaites. En effet

Pour tout B dans $\sigma(X,Y)$ (tribu engendrée par le couple (X,Y)) il existe un borélien C tel que $B = (X,Y)^{-1}(C)$. L'indépendance de (X,Y) et (L,R) permet d'ecrire

$$\begin{aligned} \int_B (1_{A=0})\,dP &= \int_B (1_{L<Y<R})\,dP \\ &= \int_{C\times\mathbb{R}_+^2} (1_{l<y<r})\,dP_{(X,Y,L,R)} \\ &= \int_{C\times\mathbb{R}_+^2} (1_{l<y<r})\,dP_{(X,Y)}\otimes dP_{(L,R)}. \end{aligned}$$

Maintenant par le théorème de Fubini et l'indépendance de R et L, nous

obtenons

$$\begin{aligned}\int_B (1_{A=0})\, dP &= \int_C \left(\int_{\mathbb{R}_+^2} (1_{l<y<r})\, dP_{(L,R)} \right) dP_{(X,Y)} \\ &= \int_C \left(\int_{\mathbb{R}+} (1_{l<y})\, dP_L \times \int_{\mathbb{R}+} (1_{y<r})\, dP_R \right) dP_{(X,Y)} \\ &= \int_C (F_L(y) S_R(y))\, dP_{(X,Y)} \\ &= \int_B F_L((Y_1) S_R(Y_1) dP,\end{aligned}$$

car F est continue. De plus, $F_L(Y)S_R(Y)$ étant clairement mesurable par rapport à $\sigma(X,Y)$, le résultat en découle.

La suite de ce chapitre est vouée à l'étude des estimateurs à poids que nous définirons pour Y dans ce cadre de censure.

3.2 Hypothèses et estimation

3.2.1 Hypothèses

Comme dans la section 2.3, nous notons pour toute variable aléatoire U, F_U (resp. S_U) sa fonction de répartition (resp. survie, $S_U = 1 - F_U$). $T_U = \sup\{t : F_U(t) < 1\}$ et $I_U = \inf\{t : F_U(t) \neq 0\}$ dénotent les points terminaux du support de U.

Soit H_1 l'hypothèse comprenant les cinq conditions suivantes
- $H_{1,1}$: Y, R et L sont indépendantes,
- $H_{1,2}$: (R, L) et (X, Y) sont indépendantes,
- $H_{1,3}$: $\exists\, T < T_R$ et $I > I_L$ tels que
  $\forall n \in \mathbb{N},\ \forall i\ (1 \leq i \leq n), \quad A_i = 0 \Rightarrow I \leq Z_i \leq T$ a.s.,
- $H_{1,4}$: F_L est continue sur $]0, +\infty[$,

– $H_{1,5} : T_R \vee T_L < \infty$.
Nous aurons aussi besoin de l'hypothèse d'identifiabilité suivante
– $H_2 : I_Y \leq I_L < I_R \quad \text{et} \quad T_R \leq T_Y$.
Remarquons que puisque $I_L < Z_i < T_R$ dès que $A_i = 0$, l'hypothèse $H_{1,3}$ semble ne pas être trop restrictive.

3.2.2 Estimation

Notons $\{W_j,\, 1 \leq j \leq M\}$ les valeurs distinctes de $\{Z_i,\, 1 \leq i \leq n\}$, rangées par ordre croissant. Posons $D_{kj} = \sum_{i=1}^{n} 1_{\{Z_i = W_j, A_i = k\}}$ et $N_j = \sum_{i=1}^{n} 1_{\{Z_i \leq W_j\}}$. L'estimateur produit-limite de S_R donné dans Patilea et Rolin (2006) est

$$\hat{S}_n(W_j) = \prod_{1 \leq l \leq j} \left\{ 1 - \frac{D_{1l}}{U_{l-1} - N_{l-1}} \right\} \text{ où } U_{j-1} = n \prod_{j \leq l \leq M} \left\{ 1 - \frac{D_{2l}}{N_l} \right\}. \quad (3.4)$$

Patilea et Rolin (2006) ont montré que sous les hypothèses H_2 et $H_{1,1}$

$$\lim_{n \mapsto \infty} \sup_{t \in \mathbb{R}^+} |\hat{S}_n(t) - S_R(t)| = 0. \quad \text{p.s.} \quad (3.5)$$

Par inversion du temps, nous obtenons l'estimateur inversé de celui de Kaplan- Meier, qui est un estimateur de F_L (censure à gauche) et il est donné par

$$\hat{F}_n(W_j) = \prod_{j < l \leq M} \left\{ 1 - \frac{1_{\{A_l = 2\}}}{l} \right\}. \quad (3.6)$$

En adaptant le théorème, de type Glivenko-Cantelli, de Stute et Wang (1993), nous obtenons

$$\lim_{n \mapsto \infty} \sup_{I_L < t} |\hat{F}_n(t) - F_L(t)| = 0 \quad \text{p.s.}, \quad (3.7)$$

si F_L est continue ce qui est supposé dans l'hypothèse $H_{1,4}$.

Nous pouvons déduire de l'hypothèse $H_{1,3}$ que

$$S_R(T) > 0 \quad \text{et} \quad F_L(I) > 0. \quad (3.8)$$

En vertu de (3.5), (3.7) et (3.8), nous obtenons pour n suffisamment grand

$$\hat{S}_n(T) > 0 \quad \text{et} \quad \hat{F}_n(I) > 0 \quad \text{p.s.} \tag{3.9}$$

En vertu de (3.2), nous proposons comme estimateurs poids de $r(x)$

$$r_n(x) = \sum_{i=1}^{n} W_{n,i}(x) 1_{\{A_i=0\}} \frac{Z_i}{\hat{S}_n(Z_i)\hat{F}_n(Z_i)} \quad \left(\frac{0}{0} := 0\right). \tag{3.10}$$

En appliquant le lemme 1.1 à l'estimateur de S_R, on voit que k_0 est le plus petit entier naturel k tel que $\hat{S}_n(Z_k) = 0$ ssi

$$D_{1k_0} \neq 0, D_{0k_0} = 0 \quad \text{et} \quad \forall j > k_0 \; D_{1j} = D_{0j} = 0,$$

ce qui implique que $\hat{S}_n(Z_i) \neq 0$ dès que $A_i = 0$. Nous pouvons aussi remarquer que $\hat{F}_n(Z_i) \neq 0$ dans l'expression de $r_n(x)$ donnée en (3.10) dès que $A_i = 0$.

3.3 Convergence de r_n

Théorème 3.1. *Si les poids $W_{n,i}$ sont positifs, si la somme des poids est au plus égale à un et si pour toutes les lois de (X, Y) avec $|Y|$ bornée presque sûrement*

$$\lim_{n\to\infty} \int_{\mathbb{R}^d} |\sum_{i=1}^{n} W_{n,i} Y_i - r(x)|^2 \mu(dx) = 0 \; p.s. \tag{3.11}$$

alors sous les hypothèses H_1 et H_2, les estimateurs r_n satisfont à

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

Démonstration. Introduisons les quantités

$$\hat{r}_n(x) = \sum_{i=1}^{n} W_{n,i}(x) 1_{\{A_i=0\}} \frac{Z_i}{S_R(Z_i)\hat{F}_n(Z_i)}$$

et

$$\bar{r}_n(x) = \sum_{i=1}^{n} W_{n,i}(x) 1_{\{A_i=0\}} \frac{Z_i}{S_R(Z_i)F_L(Z_i)}.$$

Puisque

$$\int_{\mathbb{R}^d} |r_n(x) - r(x)|^2 \mu(dx) \leq 2 \int_{\mathbb{R}^d} |r_n(x) - \hat{r}_n(x)|^2 \mu(dx) + 4 \int_{\mathbb{R}^d} |\hat{r}_n(x) - \bar{r}_n(x)|^2 \mu(dx) + 4 \int_{\mathbb{R}^d} |\bar{r}_n(x) - r(x)|^2 \mu(dx),$$

la preuve est divisée en trois étapes.

Dans la première étape nous montrons que

$$\int_{\mathbb{R}^d} |r_n(x) - \hat{r}_n(x)|^2 \mu(dx) \to 0 \text{ p.s.} \tag{3.12}$$

Utilisant (3.8), (3.9), l'hypothèse H_2 et le fait que les poids soient positifs et que leur somme soit au plus égale à un, nous trouvons pour n suffisamment grand

$$\begin{aligned} |r_n(x) - \hat{r}_n(x)| &\leq \left| \sum_{i=1}^{n} W_{n,i}(x) 1_{\{A_i=0\}} Z_i \frac{\hat{S}_n(Z_i) - S_R(Z_i)}{S_R(Z_i)\hat{S}_n(Z_i)\hat{F}_n(Z_i)} \right| \\ &\leq T_L \vee T_R \frac{\sup_{t\in\mathbb{R}} |\hat{S}_n(t) - S_R(t)|}{S_R(T)\hat{S}_n(T)\hat{F}_n(I)}. \end{aligned}$$

Donc (3.12) a lieu en vertu de (3.5) et de l'hypothèse $H_{1,5}$.

Dans la deuxième étape nous montrons que

$$\int_{\mathbb{R}^d} |\bar{r}_n(x) - \hat{r}_n(x)|^2 \mu(dx) \to 0 \text{ p.s.},$$

dont la preuve se fait de la même manière que la précédente mais en utilisant (3.7) au lieu de (3.5).

Finalement, dans la dernière étape il reste à prouver que

$$\int_{\mathbb{R}^d} |\bar{r}_n(x) - r(x)|^2 \mu(dx) \to 0 \text{ p.s.} \tag{3.13}$$

Nous pouvons déduire des hypothèses $H_{1,3}$ et H_2

$$0 \leq 1_{\{A_1=0\}} \frac{Z_1}{S_R(Z_1)F_L(Z_1)} \leq \frac{T_L \vee T_R}{S_R(T)F_L(I)} \text{ p.s.}$$

D'autre part, la relation (3.3) nous permet d'écrire

$$E(\frac{1_{\{A_1=0\}}Z_1}{S_R(Z_1)F_L(Z_1)}/X_1) = E[\frac{Y_1}{S_R(Y_1)F_L(Y_1)}E(1_{\{A_1=0\}}/(X_1,Y_1)/X_1)] = r(X_1).$$

Nous pouvons donc appliquer l'hypothèses (3.11) pour aboutir à (3.13). □

3.4 Application aux différents estimateurs à poids

Si dans l'expression de r_n

a) $W_{n,i}(x) = \begin{cases} \frac{1}{k_n} & \text{si } X_i \text{ est parmi les } k_n \text{ plus proches voisin de } x, \\ 0 & \text{sinon} \end{cases}$, où k_n est un paramètre de l'estimation. Nous obtenons l'estimateur des k_n plus proches voisins noté par $r_{n,1}$. Le théorème 3.1 et l'article de Devroye *et al.* (1994), nous permettent d'énoncer le résultat suivant.

Corollaire 3.1. *Si $k_n \to \infty$, $\frac{k_n}{n} \to 0$ et si les hypothèses H_1 et H_2 sont satisfaites alors*

$$\lim_{n\to\infty} \int_{\mathbb{R}^d} |r_{n,1}(x) - r(x)|^2 \mu(dx) = 0 \ \ p.s.,$$

si $\|X - x\|$ est absolument continue pour tout $x \in \mathbb{R}^d$.

b) $W_{n,i}(x) = \sum_j \frac{1_{A_{n,j}}(X_i)}{\sum_{k=1}^n 1_{A_{n,j}}(X_k)} 1_{A_{n,j}}(x)$, où $(A_{n,j})_j$ est une partition de $\mathbb{R}^d$ et 1_A représente la fonction indicatrice de l'ensemble A, nous obtenons l'estimateur à partition noté par $r_{n,2}$. D'après le théorème 3.1 et l'article de Devroye et Györfi (1983) nous obtenons le résultat suivant.

Corollaire 3.2. *Soit $\{A_{n,j}\}$ une partition de $\mathbb{R}^d$ telle que pour toute sphère S centrée à l'origine*

$$\lim_{n\to\infty} \max_{j:A_{n,j}\cap S\neq\emptyset} diam(A_{n,j}) = 0 \text{ et } \lim_{n\to\infty} \|\{j : A_{n,j} \cap S \neq \emptyset\}\|/n = 0.$$

Alors si les hypothèses H_1 et H_2 sont satisfaites

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

c) $W_{n,i}(x) = \frac{K((x-X_i)/h_n)}{\sum_{i=1}^n K((x-X_i)/h_n)}$, nous obtenons l'estimateur à noyau noté par $r_{n,3}$, où K est une fonction noyau et h_n est une fenêtre positive. Le théorème 3.1 et l'article de Devroye et Krzyżak (1989) impliquent le résultat suivant.

Corollaire 3.3. *Soit K un noyau régulier, c'est à dire $K \geq 0$, il existe une boule $S_{0,r}$ centrée à l'origine et de rayon $r > 0$, et une constante b telle que pour tout $x \in S_{0,r}$, $K(x) \geq b$ et $\int_{\mathbb{R}^d} \sup_{u\in x+S_{0,r}} K(u)dx < \infty$.*
Si $h_n \to 0$, $nh_n^d \to \infty$, alors sous les hypothèses H_1 et H_2, nous obtenons

$$\lim_{n\to\infty} \int_{\mathbb{R}^d} |r_{n,3}(x) - r(x)|^2 \mu(dx) = 0 \text{ p.s.}$$

Chapitre 4

Convergence presque complète de l'estimateur à noyau de la fonction de régression

Au chapitre 3, nous avons présenté l'estimateur à noyau de la régression en présence d'une censure mixte et démontré que l'erreur quadratique integrée entre cet estimateur et la régression converge vers 0 presque sûrement. Dans ce chapitre nous montrons la convergence presque complète aussi bien ponctuelle qu'uniforme de cet estimateur en précisant les vitesses de convergence (se référer à l'appendice A pour un rappel sur ces notions). Ce travail a fait l'objet d'une publication dans la revue *Statistics and Probabilty Letters* (Kebabi et Messaci, 2012). Dans toute la suite, nous appelerons r_n l'estimateur à noyau de la regression défini dans la section 3.4 et considéré pour une variable aléatoire X à valeurs dans $\mathbb{R}$, de densité f dont nous supposons l'existence. Pour le reste nous gardons les notations du chapitre 3. Rappelons que

$$r_n(x) = \sum_{i=1}^{n} W_{n,i}(x) 1_{\{A_i=0\}} \frac{Z_i}{\hat{S}_n(Z_i)\hat{F}_n(Z_i)} \tag{4.1}$$

où $W_{n,i}$ est donné par

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

4.1 Hypothèses

h_1 : Les fonctions de répartition des variables Y, R et L sont continues, $I_Y \leq I_L < I_R$ et $T_R < T_Y$.

$h_{2,1}$: r et f sont ℓ fois continuement dérivables dans un voisinage de x.

$h_{2,2}$: $f(x) > 0$,

$h_{2,3}$: $\lim_{n\to\infty} h_n = 0, \quad \lim_{n\to\infty} \frac{nh_n}{\log n} = +\infty$.

$h_{2,4}$: K est bornée, intégrable, à support compact et $\int K(t)\,dt = 1$.

$h_{2,5}$: $\int t^j K(t)\,dt = 0,\ \forall j = 1, \ldots, \ell - 1$ et $0 < |\int t^\ell K(t)| < \infty$.

Soit S un sous-ensemble compact de $\mathbb{R}$.

$h'_{2,1}$: r et f sont ℓ fois continûment dérivables autour de S

$h'_{2,2}$: $\exists \theta > 0, \inf_{x\in S} f(x) > \theta$,

$h'_{2,6}$: $\exists \beta > 0, \exists C < \infty, \forall x \in S, \forall y \in S : |K(x) - K(y)| \leq C|x - y|^\beta$

Nous terminons la liste des hypothèses par les conditions introduites au chapitre 3 en vue de traiter notre type de censure.

$H_{1,2}$: (R, L) et (X, Y) sont indépendants.

$H_{1,3}$: $\exists\, T < T_R$ et $I > I_L$ tels que $\forall n \in \mathbb{N},\ \forall i\ (1 \leq i \leq n),\ A_i = 0 \implies I \leq Z_i \leq T$ p.s.

4.2 Convergence presque complète ponctuelle

Le théorème suivant donne le taux de convergence presque complète ponctuelle de r_n.

Théorème 4.1. *Sous les hypothèses h_1, $h_{2,1}$–$h_{2,5}$, $H_{1,2}$ et $H_{1,3}$, nous avons*

$$r_n(x) - r(x) = O(h_n^\ell) + O_{a.co}\left(\sqrt{\frac{\log n}{nh_n}}\right).$$

Démonstration. Posons

$$\begin{aligned}\hat{r}_n(x) &= \sum_{i=1}^{n} W_{n,i}(x) 1_{\{A_i=0\}} \frac{Z_i}{S_R(Z_i)F_L(Z_i)} \\ &= \frac{\frac{1}{nh_n}\sum_{i=1}^{n} K\left(\frac{x-X_i}{h_n}\right)\frac{1_{\{A_i=0\}}Z_i}{S_R(Z_i)F_L(Z_i)}}{\frac{1}{nh_n}\sum_{i=1}^{n} K\left(\frac{x-X_i}{h_n}\right)} \\ &:= \frac{\hat{r}_{n,N}(x)}{f_n(x)}, \end{aligned} \tag{4.2}$$

où

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

est l'estimateur à noyau de f, puis effectuons la décomposition suivante

$$r_n(x) - r(x) = r_n(x) - \hat{r}_n(x) + \frac{\hat{r}_{n,N} - g(x)}{f_n(x)} + \frac{f(x) - f_n(x)}{f_n(x)} r(x), \tag{4.3}$$

où $g = rf$. La démonstration du théorème est alors une conséquence directe des lemmes suivants. □

Lemme 4.1. *Sous les hypothèses h_1, $h_{2,3}$ et $H_{1,3}$, nous avons*

$$r_n(x) - \hat{r}_n(x) = O_{p.co.}\left(\sqrt{\frac{\log n}{nh_n}}\right).$$

Démonstration. Les définitions de $r_n(x)$, de $\hat{r}_n(x)$ et l' hypothèse $H_{1,3}$ nous

permettent d'écrire

$$
\begin{aligned}
&|r_n(x) - \hat{r}_n(x)| \\
&\quad \leq \left| \sum_{i=1}^{n} W_{n,i}(x) 1_{\{A_i=0\}} \frac{Z_i}{\hat{S}_n(Z_i)\hat{F}_n(Z_i)} - \sum_{i=1}^{n} W_{n,i}(x) 1_{\{A_i=0\}} \frac{Z_i}{S_R(Z_i)F_L(Z_i)} \right| \\
&\quad \leq T \sum_{i=1}^{n} W_{n,i}(x) 1_{\{A_i=0\}} \\
&\qquad \times \left| \frac{F_L(Z_i)S_R(Z_i) - \hat{F}_n(Z_i)S_R(Z_i) + \hat{F}_n(Z_i)S_R(Z_i) - \hat{F}_n(Z_i)\hat{S}_n(Z_i)}{\hat{F}_n(I)\hat{S}_n(T)F_L(I)S_R(T)} \right| \\
&\quad \leq T \frac{\sup_{t>I_L} |\hat{F}_n(t) - F_L(t)| + \sup_{t\in\mathbb{R}} |\hat{S}_n(t) - S_R(t)|}{\hat{F}_n(I)\hat{S}_n(T)F_L(I)S_R(T)} \sum_{i=1}^{n} W_{n,i}(x).
\end{aligned}
$$

En tenant compte de h_1, nous pouvons appliquer la relation 2.16 et le théorème 1 dans Bitouzé *et al.* (1999) . De plus $\sum_{i=1}^{n} W_{n,i}(x) = 1$, nous pouvons donc conclure que

$$r_n(x) - \hat{r}_n(x) = O_{p.co}\left(\sqrt{\frac{\log n}{n}}\right).$$

De plus, puisque $\lim_{n\to\infty} h_n = 0$ (hypothèse $h_{2,3}$), nous déduisons que

$$r_n(x) - \hat{r}_n(x) = O_{p.co}\left(\sqrt{\frac{\log n}{nh_n}}\right). \qquad \square$$

Lemme 4.2. *Sous les hypothèses $h_{2,1}$, $h_{2,4}$, $h_{2,5}$ et $H_{1,2}$, nous obtenons*

$$E\hat{r}_{n,N}(x) - g(x) = O(h_n^{\ell})$$

Démonstration. Puisque (X_i, Z_i, A_i) sont i.i.d., nous avons

$$
\begin{aligned}
&E\hat{r}_{N,n}(x) - g(x) \\
&\qquad = E\left[\frac{1}{h_n} K\left(\frac{x - X_1}{h_n}\right) E\left(\frac{1_{\{A_1=0\}} Y_1}{F_L(Y_1)S_R(Y_1)} \middle/ X_1\right) - g(x)\right] \\
&\qquad = \int \frac{1}{h_n} K\left(\frac{x-u}{h_n}\right) E\left\{\frac{1_{\{A_1=0\}} Y_1}{F_L(Y_1)S_R(Y_1)} \middle/ X_1 = u\right\} f(u)du - g(x).
\end{aligned}
$$

L'utilisation de la relation (3.3) permet d'écrire

$$E\left[\frac{1_{\{A_1=0\}}Y_1}{F_L(Y_1)S_R(Y_1)}\Big/X_1\right] = E\left[\frac{Y_1}{F_L(Y_1)S_R(Y_1)}E\left(1_{\{A_1=0\}}/(X_1,Y_1)\right)\Big/X_1\right]$$
$$= r(X_1).$$

Donc, et puisque $\int K(t)dt = 1$, nous obtenons

$$E\hat{r}_{N,n}(x) - g(x) = \int \frac{1}{h_n}K\left(\frac{x-u}{h_n}\right)g(u)du - g(x)$$
$$= \int \left(g(x-zh_n) - g(x)\right)K(z)\,dz.$$

Les hypothèses $h_{2,1}$, $h_{2,4}$ et $h_{2,5}$ et le développement de Taylor au voisinage de x permettent d'écrire

$$E\hat{r}_{N,n}(x) - g(x) = (-1)^\ell h_n^\ell \int z^\ell K(z)\frac{g^\ell(x)dz}{\ell!} + o(h^\ell). \qquad \square$$

Lemme 4.3. *Sous les hypothèses $h_{2,1}$, $h_{2,3}$, $h_{2,4}$ et $H_{1,3}$, nous avons*

$$E\hat{r}_{n,N}(x) - \hat{r}_{n,N}(x) = O_{p.co.}\left(\sqrt{\frac{\log n}{nh_n}}\right)$$

Démonstration. Nous appliquons l'inégalité de Bernstein (corollaire A.1). Pour cela, posons

$$\Gamma_i = \frac{1_{\{A_i=0\}}Z_i}{h_n}\frac{K\left(\frac{x-X_i}{h_n}\right)}{S_R(Z_i)F_L(Z_i)} \text{ et } U_i = \Gamma_i - E\Gamma_i.$$

Il est clair que $\hat{r}_{N,n}(x) - E\hat{r}_{N,n}(x) = \frac{1}{n}\sum_{i=1}^n U_i$. En vertu des hypothèses $h_{2,4}$ et $H_{1,3}$, nous obtenons

$$|\Gamma_i| \leq \frac{T}{h_n S_R(T)F_L(I)} := M.$$

Nous devons maintenant borner $\sigma^2 = EU_1^2$, pour cela calculons $E\Gamma_1^2$.

$$\begin{aligned} E\Gamma_1^2 &= \frac{1}{h_n^2} E\left\{ E\frac{1_{\{A_1=0\}} Z_1^2}{S_R^2(Z_1)F_L^2(Z_1)} K^2\left(\frac{x-X_1}{h_n}\right) \Big/ X_1 \right\} \\ &= \frac{1}{h_n}\int \phi(x-yh_n) f(x-yh_n) K^2(y) dy, \end{aligned}$$

où $\phi(u) = E\left[\frac{1_{\{A_1=0\}} Z_1^2}{S_R^2(Z_1)F_L^2(Z_1)} \Big/ X_1 = u\right]$. En raison des hypothèses $h_{2,1}$, $h_{2,4}$ et $H_{1,3}$, nous concluons qu'il existe une constante C telle que $EU_1^2 \leq \frac{C}{h_n}$, donc pour $\varepsilon < \frac{\sigma^2}{M}$, il vient

$$\mathbb{P}\left[|\hat{r}_{N,n}(x) - E\hat{r}_{N,n}(x)| > \varepsilon\right] \leq 2e^{-\frac{n\varepsilon^2 h_n}{4C}}. \tag{4.4}$$

En vertu de l'hypothèse $h_{2,3}$, nous pouvons poser $\varepsilon = \varepsilon_0\sqrt{\frac{\log n}{nh_n}}$. Finalement pour ε_0 suffisamment grand, le terme de droite de (4.4) est le terme général d'une série convergente. □

Lemme 4.4. *Sous les hypothèses $h_{2,1}$-$h_{2,5}$, nous avons*

$$Ef_n(x) - f(x) = O(h_n^\ell), \quad Ef_n(x) - \hat{f}_n(x) = O_{p.co.}\left(\sqrt{\frac{\log n}{nh_n}}\right)$$

$$\textit{et } \exists \eta > 0,\ \sum_{n=1}^{\infty} P\left(\hat{f}_n(x) < \eta\right) < \infty.$$

Démonstration. La preuve des deux premières relations se fait de la même manière que celles des lemmes 4.2 et 4.3 en utilisant la constante 1 au lieu de la variable $\frac{1_{\{A_i=0\}} Z_i}{S_R(Z_i)F_L(Z_i)}$. Il reste à prouver la dernière relation. Remarquons que $\hat{f}_n(x)$ converge presque complètement vers $f(x)$, donc il suffit de poser $\eta = \frac{f(x)}{2}$ et de remarquer que

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

4.3 Convergence presque complète uniforme

Théorème 4.2. *Sous les hypothèses* h_1, $h_{2,3}$–$h_{2,5}$, $H_{1,2}$, $H_{1,3}$, $h'_{2,1}$, $h'_{2,2}$ *et* $h'_{2,6}$ *nous avons*

$$\sup_{x\in S}|r_n(x)-r(x)| = O(h_n^{\ell}) + O_{p.co.}\left(\sqrt{\frac{\log n}{nh_n}}\right).$$

Démonstration. Nous suivons la même démarche que pour la démonstration du théorème 4.1. L'utilisation de la relation (4.2), la décomposition (4.3) et les quatre lemmes suivants permettent de déduire directement le résultat du théorème. □

Lemme 4.5. *Sous les hypothèses* h_1, $h_{2,3}$ *et* $H_{1,3}$, *nous avons*

$$\sup_{x\in S}|r_n(x)-\hat{r}_n(x)| = O_{p.co.}\left(\sqrt{\frac{\log n}{nh_n}}\right).$$

Démonstration. La démonstration est la même que celle du lemme 4.1 □

Lemme 4.6. *Sous les hypothèses* $h_{2,4}$, $h_{2,5}$, $H_{1,2}$ *et* $h'_{2,1}$, *nous avons*

$$\sup_{x\in S}|E\hat{r}_{n,N}(x)-g(x)| = O(h_n^{\ell})$$

Démonstration. En tenant compte de l'hypothèse $h'_{2,1}$, la démonstration se fait de la même manière que celle du lemme 4.2. □

Lemme 4.7. *Sous les hypothèses* $h_{2,3}$, $h_{2,4}$, $H_{1,3}$, $h'_{2,1}$ *et* $h'_{2,6}$ *, nous avons*

$$\sup_{x\in S}|E\hat{r}_{n,N}(x)-\hat{r}_{n,N}(x)| = O_{p.co.}\left(\sqrt{\frac{\log n}{nh_n}}\right).$$

Démonstration. Dans cette preuve C' désigne une constante générique. La compacité de S nous permet d'écrire $S \subset \cup_{k=1}^{z_n}]t_k - l_n, t_k + l_n[$ où l_n et z_n vérifient

$$l_n = C' z_n^{-1} \text{ et } \; z_n \sim C' n^{\frac{\beta+1}{\beta}}. \tag{4.5}$$

En posant $t_x = \arg\min_{t \in \{t_1, t_2 \ldots, t_{z_n}\}} |x - t|$, nous obtenons

$$\begin{aligned} \sup_{x \in S} |E\hat{r}_{n,N}(x) - \hat{r}_{n,N}(x)| \leq & \sup_{x \in S} |\hat{r}_{n,N}(x) - \hat{r}_{n,N}(t_x)| \\ & + \sup_{x \in S} |E\hat{r}_{n,N}(x) - E\hat{r}_{n,N}(t_x)| \\ & + \sup_{x \in S} |E\hat{r}_{n,N}(t_x) - \hat{r}_{n,N}(t_x)|. \end{aligned}$$

En utilisant les hypothèses $H_{1,3}$ et $h'_{2,6}$ nous obtenons

$$\sup_{x \in S} |\hat{r}_{n,N}(x) - \hat{r}_{n,N}(t_x)| \leq C' \frac{l_n^\beta}{h_n^{1+\beta}},$$

et nous pouvons déduire que

$$\sup_{x \in S} |E\hat{r}_{n,N}(x) - E\hat{r}_{n,N}(t_x)| \leq C' \frac{l_n^\beta}{h_n^{1+\beta}}.$$

Maintenant, en vertu de (4.5) et de l'hypothèse $H_{1,3}$ nous pouvons voir que $\frac{l_n^\beta}{h_n^{1+\beta}} \sqrt{n h_n} \to 0$, nous en déduisons que pour tout $\varepsilon > 0$ et pour n suffisamment grand

$$P\left(\frac{l_n^\beta}{h_n^{1+\beta}} \sqrt{\frac{n h_n}{\log n}} > \varepsilon \right) = 0.$$

Finalement, l'utilisation de la relation (4.4) et (4.5) permet d'obtenir

$$\begin{aligned} & P\left[\sup_{x \in S} |E\hat{r}_{n,N}(t_x) - \hat{r}_{n,N}(t_x)| > \varepsilon_0 \sqrt{\frac{n h_n}{\log n}} \right] \\ & \leq P\left[\max_{k \in 1, \ldots, z_n} |E\widehat{r}_{n,N}(t_k) - \hat{r}_{n,N}(t_k)| > \varepsilon_0 \sqrt{\frac{n h_n}{\log n}} \right] \\ & \leq 2 z_n e^{\frac{-n\varepsilon_0^2 h_n}{4C}} \leq C' n^{\frac{\beta+1}{\beta} - \frac{\varepsilon_0^2}{4C}}. \end{aligned}$$

Le choix d'un ε_0 suffisamment grand permet d' avoir

$$\sum_{1}^{n} P\left[\sup_{x\in S}|E\hat{r}_{n,N}(t_x)-\hat{r}_{n,N}(t_x)|>\varepsilon_0\sqrt{\frac{nh_n}{\log n}}\right]<\infty. \qquad \square$$

Lemme 4.8. *Sous les hypothèses $h_{2,3}$–$h_{2,5}$, $h'_{2,1}$, $h'_{2,2}$ et $h'_{2,6}$, nous avons*

$$\sup_{x\in S}|Ef_n(x)-f(x)|=O(h_n^{\ell}),\ \sup_{x\in S}|Ef_n(x)-f_n(x)|=O_{p.co.}\left(\sqrt{\frac{\log n}{nh_n}}\right)$$

$$\textit{et}\quad \exists\eta>0,\sum_{n=1}^{\infty}P[\inf_{x\in S}|f_n(x)|\leq\eta]<\infty.$$

Démonstration. La preuve des deux premières relations se fait de la même manière que celles des lemmes 4.2 et 4.3 en utilisant la constante 1 au lieu de la variable $\frac{1_{\{A_i=0\}}Z_i}{S_R(Z_i)F_L(Z_i)}$. Il reste à montrer la dernière relation. Nous pouvons facilement déduire que pour tout $\epsilon>0$, $\sum P[\sup_{x\in S}|f_n(x)-f(x)|>\epsilon]<\infty$. Maintenant, en vertu de $h'_{2,2}$ $\inf_{x\in S}f_n(x)\leq\frac{\theta}{2}\Rightarrow\sup_{x\in S}|f_n(x)-f(x)|>\frac{\theta}{2}$. Il suffit de prendre $\eta=\epsilon=\frac{\theta}{2}$ pour obtenir le résultat visé. $\square$

4.4 Choix du paramètre de lissage

Par construction l'estimateur à noyau dépend de deux paramètres : Le noyau K et le paramètre de lissage h. Dans la pratique, il faut décider du choix à faire pour ces deux paramètres. Comme d'habitude le noyau n'a pas une grande influence sur l'estimateur, par contre le choix de h_n est essentiel.

Tous les résultats de convergence pour lesquels des vitesses de convergence sont précisées mettent en évidence le rôle du paramètre de lissage. Pour la convergence presque complète, nous savons que 4.1

$$r_n(x)-r(x)=O(h_n^{\ell})+O_{a.co}\left(\sqrt{\frac{\log n}{nh_n}}\right). \qquad (4.6)$$

Théoriquement, pour choisir h, il suffit de minimiser le 2e membre de (4.6). La vitesse de convergence presque complète est la même que lorsque Y est complétement observée (voir Ferraty et Vieu (2002)). La condition à imposer à h pour atteindre asymptotiquement le minimum est

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

Ceci implique que la vitesse optimale est

$$r_n(x) - r(x) = O_{p.co.}\left(\left(\frac{n}{\log n}\right)^{-\frac{\ell}{2\ell+1}}\right),$$

Il en est de même pour la vitesse de convergence presque complète uniforme et nous avons

$$\sup_{x\in S}|r_n(x) - r(x)| = O_{p.co.}\left(\left(\frac{n}{\log n}\right)^{-\frac{\ell}{2\ell+1}}\right).$$

Notons que la vitesse de convergence presque complète est la même pour les données complètes et pour les données soumises à une censure mixte. En ce qui concerne les données censurées à droite, Guessoum et Ould Saïd (2008) ont obtenu la même vitesse mais pour une convergence presque sûre.

Chapitre 5

Normalité asymptotique

La première démonstration de la normalité asymptotique de l'estimateur à noyau a été fournie par Schuster *et al.* (1972) qui a étendu le résultat de Nadaraya (1964), dans lequel celui-ci montre la normalité asymptotique sous la condition que Y est borné et que $nh_n^2 \to \infty$. D'autres démonstrations sont disponibles dans la littérature statistique (Hardle, 1990; Nadaraya, 1989).

Dans le cas où la variable réponse est censurée à droite, la normalité asymptotique de l'estimateur de type Nadaraya-Watson, introduit dans Kohler *et al.* (2002), a été prouvée par Guessoum et Ould Saïd (2008) pour des données indépendantes et identiquement distribuées et par Guessoum et Ould Saïd (2012) pour des données α-mélangeantes. Dans ce chapitre nous montrons la normalité asymptotique de l'estimateur à noyau r_n étudié aux chapitres 3 et 4 dont nous gardons le cadre et les notations, sauf mention contraire. Commençons par lister l'ensemble des hypothèses nécessaires à ce travail et qui sont standard dans le contexte de notre étude.

5.1 Hypothèses

$\mathcal{H}_1$: La fenêtre h_n vérifie : $h_n > 0$, $\lim_{n\to\infty} h_n = 0$, $\lim_{n\to\infty} \frac{nh_n}{\log n} = +\infty$, $\log\log n = o(\frac{1}{h_n^\mu})$, où $0 < \mu < 1$.

$\mathcal{H}_2$: K est borné et à support compacte, $\int K(t)\,dt = 1$, il existe $\ell : \int t^j K(t)\,dt = 0,\ \forall j = 1, \ldots, \ell - 1$ et $0 < |\int t^\ell K(t)| < \infty$.

$\mathcal{H}_3$: $f(x) > 0$, g et f sont ℓ fois différentiables dans un voisinage de x, où $g(x) = \int y f_{X,Y}(x, y)\,dy$ et $f_{X,Y}(x, y)$ est la densité du couple (X, Y) dont nous supposons l'existence.

$\mathcal{H}_4$: $\lim_{n\to 0} nh_n^{2\ell+1} = 0$, où ℓ est donnée en $\mathcal{H}_2$ et $\ell > 2$.

$\mathcal{H}_5$: $\int_{\{F_L(y)S_R(y)\neq 0\}} \frac{y^2 f_{X,Y}(x,y)}{F_L(y)S_R(y)} dy = q(x)$ deux fois différentiable.

$\mathcal{H}_6$: Les fonctions de répartition des variables Y, R et L sont continues, $\sup(I_L, I_Y) < I_R$ et $T_R < T_Y$.

Rappelons les hypothèses $H_{1,2}$ et $H_{1,3}$ déjà utilisées aux chapitres 3 et 4

$H_{1,2}$: (R, L) et (X, Y) sont indépendantes.

$H_{1,3}$: $\exists T < T_R$ et $I > I_L$ telle que $\forall n \in \mathbb{N}, \forall i\ (1 \leq i \leq n),\ A_i = 0 \implies I \leq Z_i \leq T$ p.s.

5.2 Résultats et preuve

Notre but est de montrer que

$$\sqrt{nh_n}\,(r_n(x) - r(x)) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \Sigma_x),$$

où Σ_x sera explicitée plus loin et $\xrightarrow{\mathcal{D}}$ désigne la convergence en loi. Pour cela, nous utilisons la méthode delta. Nous pouvons écrire

$$r_n(x) = \sum_{i=1}^{n} W_{n,i}(x) 1_{\{A_i=0\}} \frac{Z_i}{S_n(Z_i)F_n(Z_i)} := \frac{r_{n,N}(x)}{f_n(x)}, \tag{5.1}$$

où

$$r_{n,N}(x) = \frac{1}{nh_n}\sum_{i=1}^{n} K\left(\frac{x-X_i}{h_n}\right)\frac{1_{\{A_i=0\}}Z_i}{S_n(Z_i)F_n(Z_i)};$$
$$f_n(x) = \frac{1}{nh_n}\sum_{i=1}^{n} K\left(\frac{x-X_i}{h_n}\right).$$

De (5.1), nous pouvons voir que

$$r_n(x) - r(x) = \frac{r_{n,N}(x)}{f_n(x)} - \frac{g(x)}{f(x)} = \varphi\left(r_{n,N}(x), f_n(x)\right) - \varphi\left(g(x), f(x)\right),$$

où $g(x) = \int y f_{X,Y}(x,y)\,dy$, et

$$\begin{aligned}\varphi\colon \quad \mathbb{R}\times\mathbb{R}_+^* &\to \mathbb{R}\\ (x,y) &\mapsto \tfrac{x}{y}\end{aligned}$$

est une application différentiable. Donc, pour appliquer la méthode delta nous allons commencer par déterminer la loi de probabilité de

$$\sqrt{nh_n}\left[(r_{n,N}(x), f_n(x)) - (g(x), f(x))\right].$$

A cette fin, nous avons besoin des décompositions suivantes

$$r_{n,N}(x) - g(x) = \sum_{i=1}^{3}\Lambda_{i,n}(x), \quad f_n(x) - f(x) = \sum_{i=1}^{2}\Gamma_{i,n}(x),$$

où

$$\Lambda_{1,n}(x) = r_{n,N}(x) - \hat{r}_{n,N}(x), \quad \Lambda_{2,n}(x) = \hat{r}_{n,N}(x) - E\left(\hat{r}_{n,N}(x)\right),$$
$$\Lambda_{3,n}(x) = E\left(\hat{r}_{n,N}(x)\right) - g(x),$$
$$\Gamma_{1,n}(x) = f_n(x) - Ef_n(x) \quad \text{et} \quad \Gamma_{2,n}(x) = Ef_n(x) - f(x),$$

avec

$$\hat{r}_{n,N}(x) = \frac{1}{nh_n}\sum_{i=1}^{n}\frac{1_{\{A_i=0\}}Z_i}{S_R(Z_i)F_L(Z_i)}K\left(\frac{x-X_i}{h_n}\right).$$

Dans la suite $\xrightarrow{\mathcal{P}}$ désigne la convergence en probabilité. Afin de traiter chacun des termes précédents, nous introduisons les lemmes suivants

Lemme 5.1. *Sous les hypothèses $\mathcal{H}_2$, $\mathcal{H}_3$ et $\mathcal{H}_4$, nous avons*

i) $\sqrt{nh_n}\,\Gamma_{2,n}(x) \xrightarrow{\mathcal{P}} 0$.

ii) Si de plus, $H_{1,2}$ est satisfaite, nous obtenons $\sqrt{nh_n}\,\Lambda_{3,n}(x) \xrightarrow{\mathcal{P}} 0$

Démonstration. i) découle directement de la première assertion du lemme (4.4) et de l'hypothèse $\mathcal{H}_4$.

ii) découle du lemme 4.2 et de l'hypothèse $\mathcal{H}_4$. □

Lemme 5.2. *Sous les hypothèses $\mathcal{H}_6$, $\mathcal{H}_1$ et $H_{1,3}$, nous avons*

$$\sqrt{nh_n}\Lambda_{1,n}(x) \xrightarrow{\mathcal{P}} 0.$$

Démonstration.

$$\begin{aligned}
\Lambda_{1,n}(x) &= r_{n,N}(x) - \hat{r}_{n,N}(x) \\
&= \frac{1}{nh_n}\sum_{i=1}^{n} 1_{\{Ai=0\}} \frac{Z_i}{S_n(Z_i)F_n(Z_i)} K\left(\frac{x-X_i}{h_n}\right) \\
&\quad - \frac{1}{nh_n}\sum_{i=1}^{n} 1_{\{Ai=0\}} \frac{Z_i}{S_R(Z_i)F_L(Z_i)} K\left(\frac{x-X_i}{h_n}\right) \\
&= \frac{1}{nh_n}\sum_{i=1}^{n} Z_i 1_{\{A_i=0\}} K\left(\frac{x-X_i}{h_n}\right) \\
&\quad \times \left[\frac{F_L(Z_i)\left(S_R(Z_i) - S_n(Z_i)\right) + S_n(Z_i)\left(F_L(Z_i) - F_n(Z_i)\right)}{S_n(Z_i)F_n(Z_i)S_R(Z_i)F_L(Z_i)}\right] \\
&\leq \frac{1}{nh_n}\sum_{i=1}^{n} Z_i 1_{\{A_i=0\}} K\left(\frac{x-X_i}{h_n}\right) \\
&\quad \times \left[\frac{F_L(Z_i)\sup_{t\in\mathbb{R}}|(S_R(t) - S_n(t))| + S_n(Z_i)\sup_{u\geq I_L}|F_L(u) - F_n(u)|}{S_n(Z_i)F_n(Z_i)S_R(Z_i)F_L(Z_i)}\right] \\
&\leq T\frac{\sup_{t\in\mathbb{R}}|S_R(t) - S_n(t)| + S_n(I)\sup_{u\geq I_L}|F_n(u) - F(u)|}{S_n(T)S_R(T)F_n(I)F_L(I)} \\
&\quad \times \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h_n} K\left(\frac{x-X_i}{h_n}\right)
\end{aligned}$$

et puisque nous avons
– $S_n(T) \longrightarrow S_R(T)$,
– $F_n(I) \longrightarrow F_L(I)$,
– $\sup_{t\in\mathbb{R}} |S_R(t) - S_n(t)| = O\left(\sqrt{\frac{\log\log n}{n}}\right)$, (Théorème 2.4)
– $\sup_{u\geq I_T} |F_n(u) - F(u)| = O\left(\sqrt{\frac{\log\log n}{n}}\right)$, (Corollaire 2.2)
– $\frac{1}{n}\sum_{i=1}^{n} \frac{1}{h_n} K\left(\frac{x-X_i}{h_n}\right)$ converge presque sûrement,
alors

$$\Lambda_{1,n}(x) \leq c\left(\frac{\log\log n}{n}\right)^{\frac{1}{2}},$$

et en vertu de l'hypothèse $\mathcal{H}_1$, il vient $\sqrt{nh_n}\Lambda_{1,n}(x) = O(\sqrt{h_n^{1-\mu}}) \to 0$ quand $n \to \infty$. □

Lemme 5.3. *Sous les hypothèses $\mathcal{H}_6$, $\mathcal{H}_1$–$\mathcal{H}_5$, $H_{1,2}$ et $H_{1,3}$ nous obtenons*

$$\sqrt{nh_n}\,(\Lambda_{2,n}(x), \Gamma_{1,n}(x)) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \Sigma_x)$$

où

$$\Sigma_x = \int K^2(t)dt \begin{pmatrix} q(x) & g(x) \\ g(x) & f(x) \end{pmatrix},$$

et $q(x)$ est défini à l'hypothèse $\mathcal{H}_5$.

Démonstration. Commençons par calculer les variances et les covariances nécessaires à notre preuve

1.

$$\mathrm{Var}\left(\sqrt{nh_n}\Lambda_{2,n}(x)\right) = \frac{1}{h_n} E\left(1_{\{A_1=0\}} \frac{Z_1^2 K^2\left(\frac{x-X_1}{h_n}\right)}{S_R^2(Z_1)F_L^2(Z_1)}\right) - \frac{1}{h_n} E^2\left(1_{\{A_1=0\}} \frac{Z_1 K\left(\frac{x-X_1}{h_n}\right)}{S_R(Z_1)F_L(Z_1)}\right).$$

i) D'après la relation (3.3) et les hypothèses $\mathcal{H}_2$ et $\mathcal{H}_5$ nous pouvons écrire

$$\begin{aligned}
&\frac{1}{h_n}E\left(1_{\{A_1=0\}}\frac{Z_1^2K^2\left(\frac{x-X_1}{h_n}\right)}{S_R^2(Z_1)F_L^2(Z_1)}\right)\\
&\quad=\frac{1}{h_n}E\left(K^2\left(\frac{x-X_1}{h_n}\right)E\left(\frac{Y_1^2}{S_R^2(Y_1)F_L^2(Y_1)}E\left(1_{\{A_1=0\}}/(X_1,Y_1)\right)\Big/X_1\right)\right)\\
&\quad=\frac{1}{h_n}\int K^2\left(\frac{x-u}{h_n}\right)\left(\int\frac{y^2}{S_R(y)F_L(y)}f_{Y_1/X_1=u}(y)dy\right)f(u)\,du\\
&\quad=\frac{1}{h_n}\int K^2\left(\frac{x-u}{h_n}\right)q(u)\,du,
\end{aligned}$$

où $q(u)=\int\frac{y^2f_{X,Y}(u,y)}{S_R(y)F_L(y)}dy$. Donc

$$\frac{1}{h_n}E\left(1_{\{A_1=0\}}\frac{Z_1^2K^2\left(\frac{x-X_1}{h_n}\right)}{S_R^2(Z_1)F_L^2(Z_1)}\right)=q(x)\int K^2(t)dt+\frac{h_n^2}{2!}\int t^2K^2(t)q''(c)\,dt,$$

où c est compris entre x et $x-th_n$. Nous obtenons donc

$$\frac{1}{h_n}E\left(1_{\{A_1=0\}}\frac{Z_1^2K^2\left(\frac{x-X_1}{h_n}\right)}{S_R^2(Z_1)F_L^2(Z_1)}\right)=q(x)\int K^2(t)dt+o(h_n). \quad (5.2)$$

ii) En suivant la même démarche qu'en i) nous obtenons facilement

$$\begin{aligned}
E\left(1_{\{A_1=0\}}\frac{Z_1K\left(\frac{x-X_1}{h_n}\right)}{S_R(Z_1)F_L(Z_1)}\right)&=\int K\left(\frac{x-u}{h_n}\right)g(u)du\\
&=h_ng(x)+o(h_n^2),
\end{aligned}$$

donc

$$\frac{1}{h_n}E^2\left(1_{\{A_1=0\}}\frac{Z_1K\left(\frac{x-X_1}{h_n}\right)}{S_R(Z_1)F_L(Z_1)}\right)=o(1). \quad (5.3)$$

En combinant (5.2) et (5.3), nous obtenons

$$\mathrm{Var}\left(\sqrt{nh_n}\Lambda_{2,n}(x)\right) = q(x)\int K^2(t)dt + o(1). \tag{5.4}$$

2.

$$\mathrm{Var}\left(\sqrt{nh_n}\Gamma_{1,n}(x)\right) = \frac{1}{h_n}\left[EK^2\left(\frac{x-X_1}{h_n}\right) - E^2K\left(\frac{x-X_1}{h_n}\right)\right].$$

En procédant comme en 1. i) pour le terme $EK^2\left(\frac{x-X_1}{h_n}\right)$ et comme en 1. ii) pour le terme $E^2K\left(\frac{x-X_1}{h_n}\right)$, nous pouvons déduire, en tenant compte de l'hypothèse $\mathcal{H}_3$ que

$$\mathrm{Var}\left(\sqrt{nh_n}\Gamma_{1,n}(x)\right) = f(x)\int K^2(t)dt + o(1). \tag{5.5}$$

3. Par équidistribution et indépendance des variables (X_i, Z_i, A_i)

$$\begin{aligned}
&\mathrm{Cov}\left(\sqrt{nh_n}\Lambda_{2,n}(x), \sqrt{nh_n}\Gamma_{1,n}(x)\right)\\
&\quad= \frac{1}{h_n}E\left(1_{\{A_1=0\}}Z_1\frac{K^2\left(\frac{x-X_1}{h_n}\right)}{S_R(Z_1)F_L(Z_1)}\right)\\
&\qquad- \frac{1}{h_n}E\left(1_{\{A_1=0\}}Z_1\frac{K\left(\frac{x-X_1}{h_n}\right)}{S_R(Z_1)F_L(Z_1)}\right)E\left(K\left(\frac{x-X_1}{h_n}\right)\right).
\end{aligned}$$

D'une part, en vertu de la partie 2, nous avons

$$EK\left(\frac{x-X_1}{h_n}\right) = h_nf(x) + o(h_n^2)$$

et

$$E\left(1_{\{A_1=0\}}\frac{Z_1K\left(\frac{x-X_i}{h_n}\right)}{S_R(Z_1)F_L(Z_1)}\right) = h_ng(x) + o(h_n^2),$$

ce qui nous permet de déduire que

$$\frac{1}{h_n}E\left(1_{\{A_1=0\}}\frac{Z_1K\left(\frac{x-X_1}{h_n}\right)}{S_R(Z_1)F_L(Z_1)}\right)EK\left(\frac{x-X_1}{h_n}\right) = o(1). \tag{5.6}$$

D'autre part, des calculs analogues à ceux des cas précédents donnent :

$$\frac{1}{h_n}E\left(1_{\{A1=0\}}Z_1\frac{K^2\left(\frac{x-X_1}{h_n}\right)}{S_R(Z_1)F_L(Z_1)}\right) = \frac{1}{h_n}\int K^2\left(\frac{x-X_1}{h_n}\right)g(u)d(u)$$
$$= g(x)\int K^2(t)dt + o(h_n). \quad (5.7)$$

Finalement, en combinant (5.6) et (5.7), nous obtenons

$$\mathrm{Cov}\left(\sqrt{nh_n}\Lambda_{2,n}(x), \sqrt{nh_n}\Gamma_{1,n}(x)\right) = g(x)\int K^2(t)dt + o(1). \quad (5.8)$$

Il reste à démontrer que les combinaisons linéaires de $\sqrt{nh_n}\Lambda_{2,n}(x)$ et $\sqrt{nh_n}\Gamma_{1,n}(x)$ sont asymptotiquement gaussiennes. Soient a et b deux nombres réels. Nous allons appliquer le théorème central limite de Lyapounov à la suite $\mathcal{W}_i$ où

$$\mathcal{W}_i = \frac{a}{\sqrt{nh_n}}\left\{1_{\{A_i=0\}}\frac{Z_iK\left(\frac{x-X_i}{h_n}\right)}{S_R(Z_i)F_L(Z_i)} - E\left(1_{\{A_i=0\}}\frac{Z_iK\left(\frac{x-X_i}{h_n}\right)}{S_R(Z_i)F_L(Z_i)}\right)\right\}$$
$$+\frac{b}{\sqrt{nh_n}}\left\{K\left(\frac{x-X_i}{h_n}\right)E\left(K\left(\frac{x-X_i}{h_n}\right)\right)\right\}.$$

Soit $\sigma_i^2 = \mathrm{Var}\,\mathcal{W}_i$, $s_n^2 = \sum_{i=1}^n \sigma_i^2$ et pour $\delta > 0$, posons $\beta = \delta + 2$. Par l'inégalité de Jensen, nous obtenons

$$E|\mathcal{W}_i|^\beta \le 2^{2\beta-1}\frac{a^\beta h_n^{\beta/2}}{n^{\beta/2}}E\left|\frac{1}{h_n}1_{\{A_i=0\}}\frac{Z_iK\left(\frac{x-X_i}{h_n}\right)}{S_R(Z_i)F_L(Z_i)}\right|^\beta$$
$$+ 2^{2\beta-1}\frac{b^\beta h_n^{\beta/2}}{n^2}E\left|\frac{1}{h_n}K\left(\frac{x-X_i}{h_n}\right)\right|^\beta.$$

De plus, nous obtenons par l'hypothèse $H_{1,3}$

$$E\left|\frac{1}{h_n}1_{\{A_i=0\}}\frac{Z_iK\left(\frac{x-X_i}{h_n}\right)}{S_R(Z_i)F_L(Z_i)}\right|^\beta \le \frac{T^\beta}{S_R^\beta(T)F_L^\beta(T)}E\left|K\left(\frac{x-X_i}{h_n}\right)\right|^\beta$$

et

$$E\left|K\left(\frac{x-X_i}{h_n}\right)\right|^\beta \leq \int \left|K\left(\frac{x-u}{h_n}\right)\right|^\beta f(u)du$$
$$= h_n \int |K(z)|^\beta f(x-zh_n)\,dz = O(h_n),$$

en vertu des hypothèses $\mathcal{H}_2$ et $\mathcal{H}_3$. Donc

- d'une part $\sum_{i=1}^n E|\mathcal{W}_i|^\beta = O\left((nh_n)^{1-\beta/2}\right) \to 0$, puisque $nh_n \to \infty$ et $\beta > 2$;
- d'autre part, en utilisant les relations (5.4), (5.5) et (5.8), nous obtenons

$$\begin{aligned}\sum_{i=1}^n \sigma_i^2 &= \mathrm{Var}(\sum_{i=1}^n \mathcal{W}_i)\\ &= a^2\,\mathrm{Var}\left(\sqrt{nh_n}\lambda_{2,n}(x)\right) + b^2\,\mathrm{Var}\left(\sqrt{nh_n}\Gamma_{1,n}(x)\right)\\ &\quad + 2ab\,\mathrm{Cov}\left(\sqrt{nh_n}\Lambda_{2,n}(x), \sqrt{nh_n}\Gamma_{1,n}(x)\right)\\ &= a^2 q(x)\int K^2(t)dt + b^2 f(x)\int K^2(t)dt\\ &\quad + 2abg(x)\int K^2(t)dt + o(1).\end{aligned}$$

La condition de Liapounov étant satisfaite, le lemme est donc démontré. □

Théorème 5.1. *Sous les hypothèses $\mathcal{H}_1$–$\mathcal{H}_6$, $H_{1,2}$ et $H_{1,3}$, nous avons*

$$\sqrt{nh_n}\,(r_n(x) - r(x)) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma^2(x))$$

où $\sigma^2(x) = \frac{q(x)f(x)-g^2(x)}{f^3(x)} \int K^2(t)dt$, q étant donné à l'hypothèse $\mathcal{H}_5$.

Démonstration. En combinant les résultats des lemmes précédents, nous obtenons par le théorème de Mann-Wald

$$\sqrt{nh_n}\,[(r_{n,N}(x), f_n(x)) - (g(x), f(x))] \xrightarrow{\mathcal{D}} T \sim \mathcal{N}(0, \Sigma_x).$$

Puis la méthode delta nous permet d'avoir

$$\sqrt{nh_n}\left(r_n(x) - r(x)\right) \xrightarrow{\mathcal{D}} \varphi'_{(g(x),f(x))}(T)$$

où φ' désigne la différentielle de φ. En d'autres termes,

$$\sqrt{nh_n}\left(r_n(x) - r(x)\right) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \nabla\varphi^t \Sigma_x \nabla\varphi\right),$$

où le gradient $\nabla\varphi^t = \left(\frac{\partial\varphi}{\partial x}, \frac{\partial\varphi}{\partial y}\right)$ est évalué au point $(g(x), f(x))$. Le calcul de $\nabla\varphi^t\Sigma_x\nabla\varphi$ conduit à $\nabla\varphi^t\Sigma_x\nabla\varphi = \frac{q(x)f(x)-g^2(x)}{f^3(x)} \int K^2(t)dt$. □

Chapitre 6

Simulation

6.1 Estimateur produit-limite de la fonction de survie

Dans cette section, nous calculons et représentons graphiquement la fonction de survie et son estimateur en vue de les comparer dans des situations simulées de censure mixte. Il s'agit de l'estimateur proposé par Patilea et Rolin (2006) et présenté au chapitre 1. Rappelons qu'en présence de censure mixte, au lieu d'observer un échantillon de la variable d'intérêt X, nous observons un échantillon du couple (Z, A) où $Z = \max(\min(X, R), L)$ et

$$A = \begin{cases} 0, & \text{si } L < X \leq R, \\ 1, & \text{si } L < R < X, \\ 2, & \text{si } \min(X, R) \leq L. \end{cases}$$

6.1.1 Construction de l'échantillon

Nous commençons par simuler les temps de survie d'intérêt $\{X_i,\, 1 \leq i \leq n\}$, les temps de censure à droite $\{R_i,\, 1 \leq i \leq n\}$ et les temps de censure à gauche $\{L_i,\, 1 \leq i \leq n\}$ selon deux modèles et pour trois tailles de l'échantillon ($n = 100$, $n = 200$ et $n = 300$) pour chaque modèle.

1. Modèle 1 (Figure 6.1)
   - X_i suit une loi normale $\mathcal{N}(7, 2)$,
   - R_i suit une loi normale $\mathcal{N}(9, 1)$,
   - L_i suit une loi normale $\mathcal{N}(1, 2)$.
2. Modèle 2 (Figure 6.2)
   - X_i suit une loi exponentielle $\mathcal{E}(2)$,
   - R_i suit une loi exponentielle $\mathcal{E}(10)$,
   - L_i suit une loi exponentielle $\mathcal{E}(0.2)$.

Nous prenons alors $Z_i = \max(\min(Y_i, R_i), L_i)$ et $A_i = 1_{\{L_i < R_i \leq Y_i\}} + 2 \times 1_{\{\min(Y_i, R_i) \leq L\}}$.

6.1.2 Calcul de l'estimateur

Notons $\{Z_j, 1 \leq j \leq M\}$ les valeurs distinctes de $\{Y_i, 1 \leq i \leq n\}$ rangées par ordre croissant.

$$D_{kj} = \sum_{i=1}^{n} 1_{\{Y_i = Z_j, A_i = k\}}$$

et

$$N_j = \sum_{i=1}^{n} 1_{\{Y_i \leq Z_j\}}.$$

Nous calculons alors l'estimateur produit-limite $\hat{S}_n$ de la fonction de survie S_X de X donné par

$$\hat{S}_n(t) = \prod_{j/Z_j \leq t} \left\{1 - D_{0j}/(U_{j-1} - N_{j-1})\right\},$$

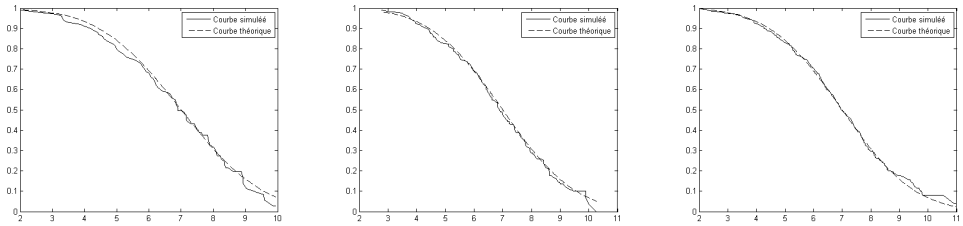

FIGURE 6.1 – Modèle 1 : pour $n = 100, 200$ et 300

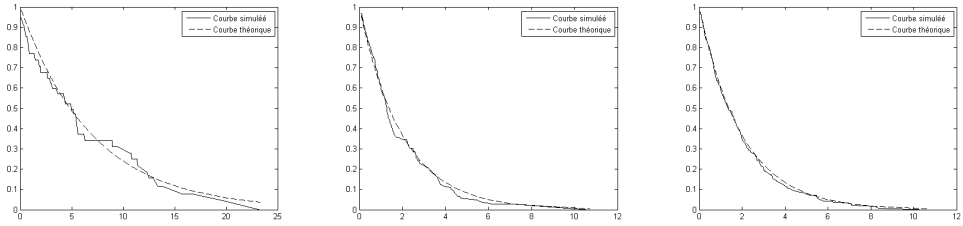

FIGURE 6.2 – Modèle 2 : pour $n = 100, 200$ et 300.

où

$$U_{j-1} = n \prod_{j \leq l \leq M} \{1 - D_{2l}/N_l\} .$$

Les graphes confirment la bonne performance de l'estimateur $\hat{S}_n$, les résultats s'améliorant avec l'augmentation de la taille de l'échantillon, sans surprise.

6.2 Estimateur à noyau de la régression

Pour donner un aperçu de la performance de l'estimateur à noyau, nous le représentons graphiquement avec la fonction de régression pour des données simulées selon le processus de censure mixte décrit au chapitre 1.

Rappelons qu'il est donné par l'expression

$$r_n(x) = \frac{\frac{1}{nh_n}\sum_{i=1}^{n} K\left(\frac{x-X_i}{h_n}\right)\frac{1_{\{A_i=0\}}Z_i}{S_n(Z_i)F_n(Z_i)}}{\frac{1}{nh_n}\sum_{i=1}^{n} K\left(\frac{x-X_i}{h_n}\right)}.$$

Il dépend du noyau K et de la fenêtre h_n qu'il faut choisir pour calculer $\hat{r}_n(x)$. Nous savons que le choix du noyau K n'est pas déterminant, et nous choisissons le noyau Gaussien. En revanche, le choix de h_n est crucial. Nous calculons la fenêtre optimale par minimisation de la distance maximale entre l'estimateur et la fonction de régression théorique, pour $h_n \in [0.1, 2]$. Nous prenons le maximum sur l'intervalle $S = [1, 5]$.

Nous considérons le même modèle de censure que nous simulons dans la section 6.1. Ici, la variable censurée est la variable d'intérêt Y, variable expliquée, alors que X est la variable explicative. Nous considérons trois modèles de régression.

1. Modèle 1 (Figure 6.3) $Y_i = 2X_i + 1 + 0.2\varepsilon_i$,
2. Modèle 2 (Figure 6.4) $Y_i = \cos(\frac{3}{2}X_i) + 0.2\varepsilon_i$,
3. Modèle 3 (Figure 6.5) $Y_i = \exp(\frac{3}{2}X_i + 1) + 0.2\varepsilon_i$.

Pour ces trois modèles, ε_i suit une loi normale centrée réduite, ε_i est indépendante de X_i qui suit une loi normale $\mathcal{N}(3, 3)$. R_i et L_i suivent respectivement les lois $\mathcal{N}(14, 1)$ et $\mathcal{N}(0, 1)$ dans le premier modèle, $\mathcal{N}(2, 2)$ et $\mathcal{N}(-2, 2)$ dans le deuxième modèle, et $\mathcal{N}(100, 8)$ et $\mathcal{N}(3, 1)$ dans le troisième modèle.

Les figures 6.3, 6.4 et 6.5 suggèrent la bonne qualité de l'ajustement pour des taux de censure autour de 30%, qui de plus augmente avec la taille de l'échantillon, ce qui est tout à fait naturel. Pour le cas linéaire et exponentiel l'ajustement est bon dès $n = 50$. Signalons le fait que les tailles des échantillons sont bien plus petites que celles utilisées par Guessoum et Ould Saïd (2008) dans le cas de la seule censure à droite.

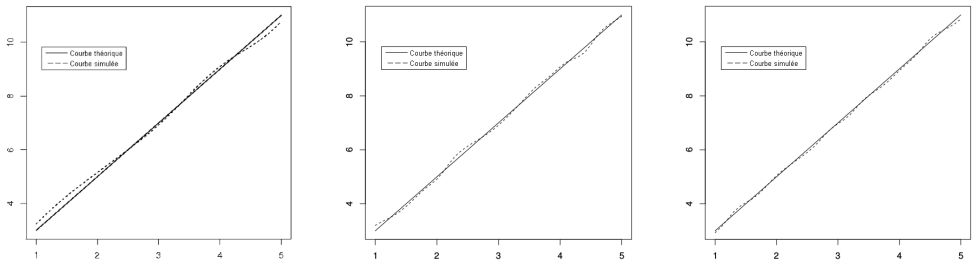

FIGURE 6.3 – $r(x) = 2x+1$, $n = 50, 100, 150$ et des taux de censure à gauche et à droite de (16% ;12%), (11% ; 12%), (8.5% ; 12.5%).

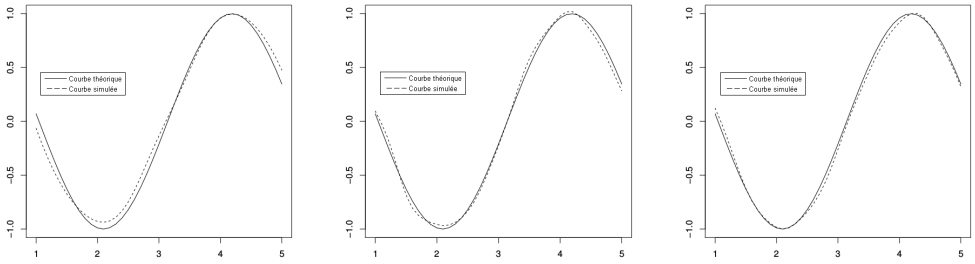

FIGURE 6.4 – $r(x) = \cos(\frac{3}{2}x)$, pour $n = 100, 200, 300$ et des taux de censure à gauche et à droite de (17% ;12%), (20% ; 11.5%), (20.6% ; 11.3%) respectivement.

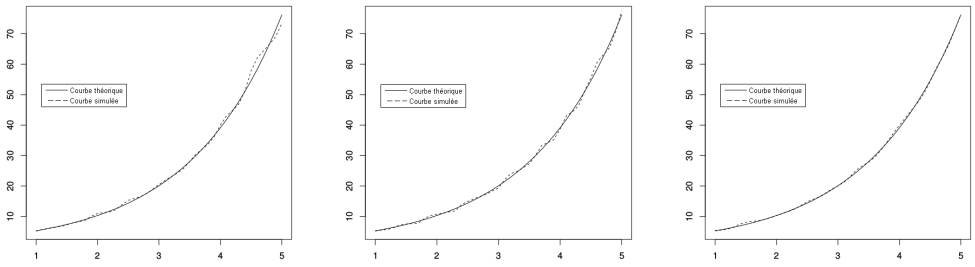

FIGURE 6.5 – $r(x) = \exp(\frac{3}{2}x + 1)$, pour $n = 50$, 100, 150 et des taux de censure à gauche et à droite de (15% ; 20%), (15% ; 27%), (16% ; 24%) respectivement.

6.3 Normalité asymptotique

Nous considérons $\frac{\sqrt{nh_n}}{\hat{\sigma}}(r_n(x) - r(x))$ pour $x = 0$ et nous posons

$$\bar{r}_n = \frac{\sqrt{nh_n}}{\hat{\sigma}}(r_n(0) - r(0)),$$

où $\hat{\sigma}$ est l'estimateur empirique de l'écart type de $(r_n(0) - r(0))$. Nous comparons la distribution d'échantillonage de $\bar{r}_n$ à une distribution normale centrée réduite dans le cas où

– $r(x) = 2x + 1$,

– $r(x) = cos\left(\frac{3}{2}x\right)$.

Nous simulons B échantillons de taille n et chaque échantillon nous permet de calculer une valeur simulée de $\bar{r}_n$. La simulation de chaque échantillon et le calcul de $r_n(x)$ sont identiques à ceux de la section 6.2. L'ensemble des valeurs de $\bar{r}_n$ ainsi obtenues constitue un échantillon R de $\bar{r}_n$.

Nous comparons alors l'histogramme et l'estimateur à noyau de la densité de $\bar{r}_n$ à la densité normale centrée réduite. Nous donnons aussi le graphe des quantiles de $\bar{r}_n$ en fonction des quantiles de la loi normale (QQ-plot).

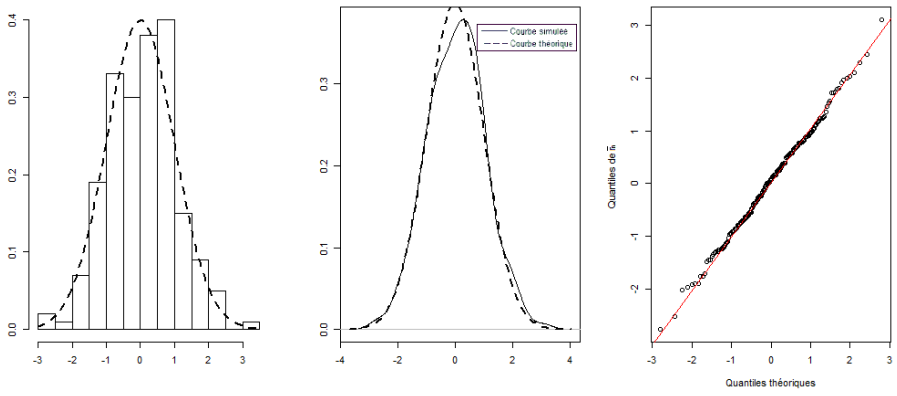

FIGURE 6.6 – $n = 100$ et $B = 200$, $r(x) = 2x + 1$.

La figure 6.6 montre, de gauche à droite, en trait discontinu la densité normale centrée réduite superposée à l'histogramme, puis la même densité superposée à l'estimateur à noyau de la densité de $\bar{r}_n$ en trait plein, et enfin le graphe des quantiles de $\bar{r}_n$ en fonction des quantiles de la loi normale centrée réduite, pour une taille $n = 100$ des échantillons et un nombre $B = 200$ d'échantillons. Le test de Kolmogarov-Smirnov appliqué à l'échantillon de $\bar{r}_n$ donne une p-value $p = 0.78$. La figure 6.7 reprend les mêmes éléments mais cette fois pour $n = 300$. Le test de Kolmogorov-Smirnov donne $p = 0,84$. Les figures 6.6 et 6.7 et les résultats du test sont en faveur d'une bonne approximation par la loi normale de $\bar{r}_n$.

Pour le deuxième modèle, le test de Kolmogorov-Smirnov donne $p = 0.29$ pour $n = 100$ et $p = 0.83$ pour $n = 300$. Les figure 6.8 et 6.9 nous permettent de tirer les même conclusions que précédemment pour ce deuxième modèle.

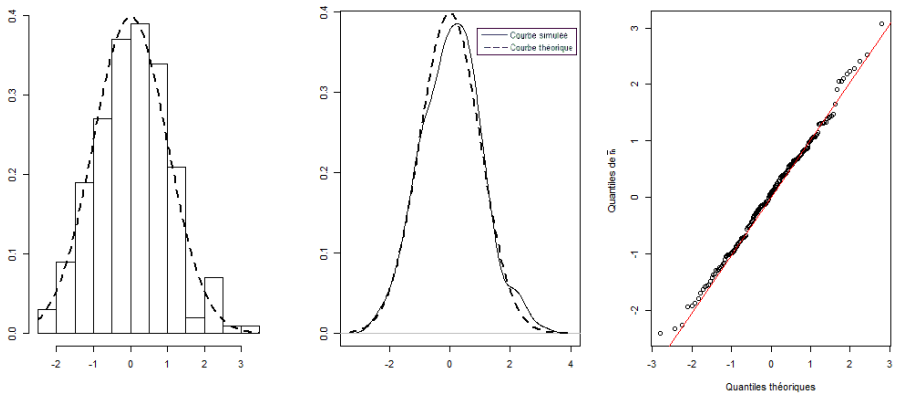

FIGURE 6.7 – $n = 300$ et $B = 200$, $r(x) = 2x + 1$.

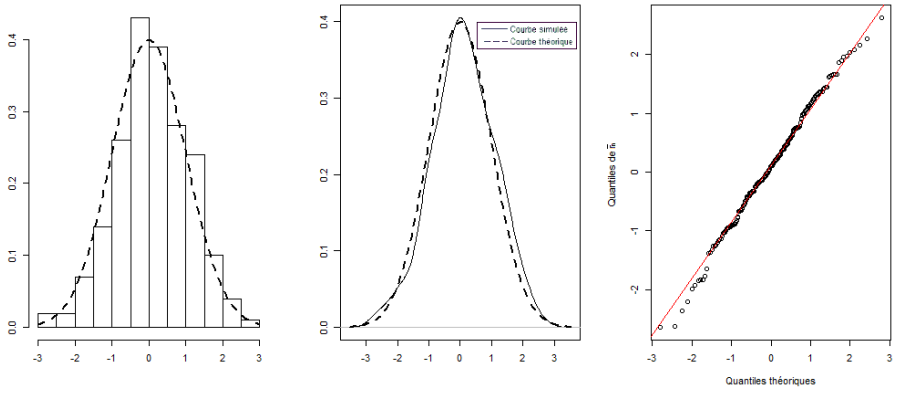

FIGURE 6.8 – $n = 100$ et $B = 200$, $r(x) = cos(\frac{3}{2}x)$.

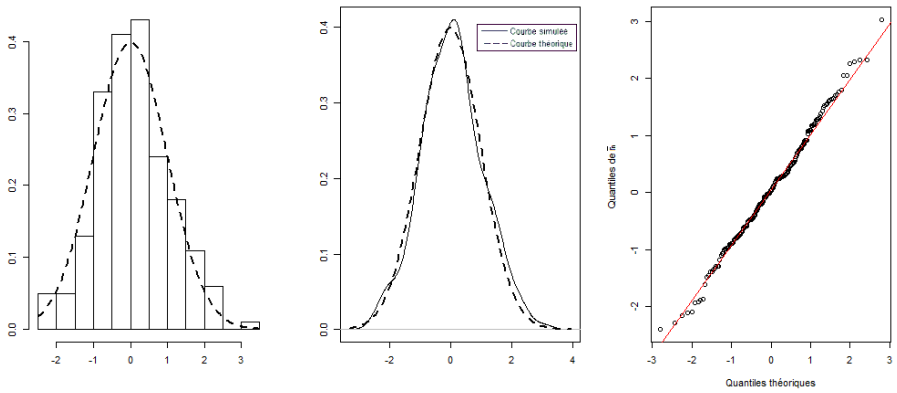

FIGURE 6.9 – $n = 300$ et $B = 200$, $r(x) = cos(\frac{3}{2}x)$.

Perspectives

Pour conclure cette thèse nous soulevons sous la forme d'une simple liste, quelques points susceptibles de faire l'objet de futurs travaux.

– Approfondissement de l'étude sur le choix de la fenêtre.
– Étude de la convergence en moyenne quadratique de r_n.
– Étude d'autres outils de prévision dans ce contexte de censure.
– Étude de la censure double.
– Réalisation d'applications.

Annexe A

Convergence presque complète et inégalités de Bernstein

Le concept de convergence presque complète a été introduit par Hsu et Robbins (1947). Elle implique la convergence presque sûre et se prête bien aux calculs faisant intervenir des sommes de variables aléatoires. Malgré cela, elle ne commence à devenir populaire dans la communauté statistique que dans les années 1980 après les travaux de Collomb. Elle est utilisée surtout en statistique non-paramétrique.

On dit que la suite de variables aléatoires réelles $(X_n)_{n\in\mathbb{N}}$ converge presque complètement vers une variable aléatoire X lorsque $n \to \infty$ si

$$\forall \varepsilon > 0, \quad \sum_{n\in\mathbb{N}} P\left[|X_n - X| > \varepsilon\right] < \infty.$$

et on dit que la vitesse de convergence presque complète de la suite de variables aléatoires réelles $(X_n)_{n\in\mathbb{N}}$ vers X est d'ordre (u_n) si

$$\exists \varepsilon_0 > 0, \quad \sum_{n\in\mathbb{N}} P[|X_n - X| > \varepsilon_0 u_n] < \infty.$$

Cette définition du taux a été introduite par Ferraty et Vieu (2006). Elle a l'avantage théorique d'impliquer les deux vitesses de convergence clas-

siques en probabilité et presque sûre, et l'avantage pratique d'être souvent plus facile à démontrer.

Dans les quinze dernières années, ce mode de convergence a été très utilisé dans des travaux concernant la statistique non-paramétrique des données fonctionnelles. Dans la section A.1, nous rappelons quelques propriétés relatives à la convergence presque complète. Dans la section A.2, nous introduisons quelques inégalité exponentielles pour des sommes de variables aléatoires, utiles à la démonstration des propriétés de convergence presque complète

A.1 Propriétés

Rappelons que la convergence presque complète entraîne à la fois la convergence presque sûre et la convergence en probabilité.

Cette convergence possède les propriétés suivantes

Proposition A.1. *Soit l_x et l_y deux nombres réels et (u_n) une suite denombres réels tels que $\lim_{n\to\infty} u_n = 0$, alors*

i) Si $\lim_{n\to\infty} X_n = l_x$ p.co. et $\lim_{n\to\infty} Y_n = l_y$ p.co., alors

a) $\lim_{n\to\infty}(X_n + Y_n) = l_x + l_y$ *p.co.,*

b) $\lim_{n\to\infty}(X_n Y_n) = l_x l_y$ *p.co.,*

c) $\lim_{n\to\infty} \frac{1}{X_n} = \frac{1}{l_x}$ *p.co. lorsque* $l_x \neq 0$.

ii) Si $X_n - l_x = O_{a.co}(u_n)$ et $Y_n - l_y = O_{p.co}(u_n)$, alors

a) $(X_n + Y_n) - l_x - l_y = O_{p.co.}(u_n)$,

b) $(X_n Y_n) - l_x l_y = O_{p.co.}(u_n)$,

c) $\frac{1}{X_n} - \frac{1}{l_x} = O_{p.co.}(u_n)$ *lorsque* $l_x \neq 0$.

iii) Si $X_n = O_{p.co.}(u_n)$ et $\lim_{n\to\infty} Y_n = l_y$ p.co., alors

a) $X_n Y_n = O_{p.co.}(u_n)$,

b) $\frac{X_n}{Y_n} = O_{p.co.}(u_n)$*, lorsque* $l_y \neq 0$*.*

Démonstration. i a) La preuve découle immédiatement de l'inégalité suivante

$$P(|(X_n + Y_n) - (l_x + l_y)| > \varepsilon) \leq P(|X_n - l_x| > \frac{\varepsilon}{2}) + P(|Y_n - l_y| > \frac{\varepsilon}{2}).$$

ii a) Il suffit d'appliquer ce résultat à $\varepsilon = \varepsilon_0 U_n$.

i b) Sans perte de généralité, on pose $l_x = 0$. La décomposition suivante

$$X_n \times Y_n = X_n(X_n - l_y) + X_n \times l_y,$$

nous donne

$$\begin{aligned} P\left(|(X_n \times Y_n)| > \varepsilon\right) &\leq P\left(|Y_n - l_y||X_n| > \frac{\varepsilon}{2}\right) + P\left(|l_y X_n| > \frac{\varepsilon}{2}\right) \\ &\leq P\left(|Y_n - l_y| > \sqrt{\frac{\varepsilon}{2}}\right) + P\left(|X_n| > \sqrt{\frac{\varepsilon}{2}}\right) + P\left(|X_n l_y| > \frac{\varepsilon}{2}\right) \\ &\leq P\left(|Y_n - l_y| > \frac{\varepsilon}{2}\right) + P\left(|X_n| > \frac{\varepsilon}{2}\right) + P\left(|X_n l_y| > \frac{\varepsilon}{2}\right). \end{aligned}$$

L'inégalité précédente et la convergence presque complète de X_n et Y_n permettent d'écrire

$$\sum_{n \in \mathbb{N}} P(|X_n Y_n| > \varepsilon) < \infty.$$

ii b) Il suffit d'appliquer le résultat précédent à $\varepsilon = \varepsilon_0 U_n$.

i c) La convergence presque complète de Y_n vers l_y implique l'existence de $\delta \geq 0$ ($\delta = \frac{l_y}{2}$)) tel que :

$$\sum_{n \in \mathbb{N}} P(|Y_n| \leq \delta) < \infty, \tag{A.1}$$

de plus nous avons

$$\begin{aligned} P\left(\frac{1}{Y_n} - \frac{1}{l_y} > \varepsilon\right) &= P(|Y_n - l_y| > \varepsilon |l_y Y_n|) \\ &\leq P(|Y_n - l_y| > \varepsilon |l_y Y_n|, |Y_n| > \delta) + P(|Y_n \leq \delta|) \\ &\leq P(|Y_n - l_y| > \varepsilon \delta |l_y|) + P(|Y_n| \leq \delta). \end{aligned}$$

En utilisant la relation (A.1) et la définition de la convergence presque complète de Y_n vers l_y, il vient

$$\forall \varepsilon > 0, \quad \sum_{n \in \mathbb{N}} P\left(\left| \frac{1}{Y_n} - \frac{1}{l_y} \right| \geq \varepsilon \right) < \infty.$$

ii c) On procède de la même manière pour $\varepsilon = \varepsilon_0 U_n$.

iii a) La définition de la convergence presque complète de Y_n vers l_y implique l'existence de $\delta > 0$ tel que

$$\sum_{n \in \mathbb{N}} P[|Y_n| > \delta] < \infty.$$

La décomposition suivante

$$\begin{aligned} P(|Y_n X_n| > \varepsilon U_n) &= P(|Y_n X_n| > \varepsilon U_n, |Y_n| \leq \delta) + P(|X_n Y_n| > \varepsilon U_n, |Y_n| > \delta) \\ &\leq P(|X_n| > \varepsilon \delta^{-1} U_n) + P(|Y_n| > \delta), \end{aligned}$$

associée à l'inégalité précédente et à l'hypothèse $X_n = O_{a.co}(U_n)$, conduisent à $X_n Y_n = O_{a.co}(U_n)$. □

A.2 Inégalités de type Bernstein

Soit $\{X_n, n \geq 1\}$ une suite de variables aléatoires centrées. Pour démontrer la convergence presque complète, nous avons besoin de trouver des bornes supérieures pour certaines probabilités concernant des sommes de variables aléatoires telles que

$$P\left(\left| \sum_{i=1}^{n} Z_i \right| > \varepsilon \right)$$

où éventuellement ε décroît avec n. Dans ce contexte, il existe de puissants outils probabilistes appelés inégalités exponentielles. On en trouve

différentes versions dans la littérature. Les inégalités diffèrent selon les hypothèses imposés aux variables aléatoires Z_i. Nous en présentons ici celles qu'on appelle inégalités de type Bernstein dont la forme convient le plus à notre travail.

Supposons que $\{X_n, n \geq 1\}$ est une suite de variables aléatoires réelles, indépendantes et centrées.

Proposition A.2. *Si*

$$\forall m \geq 2,\ |E(X_i^m)| \leq \left(\frac{m!}{2}\right)(a_i)^2 b^{m-2},$$

alors

$$\forall \varepsilon \geq 0, \qquad P\left[\sum_{i=1}^{n} \left| X_i \right| > \varepsilon A_n\right] \leq 2\exp\left\{\frac{-\varepsilon^2}{2(1+\frac{b\varepsilon}{A_n})}\right\},$$

où $(a_i)_{1\leq i\leq n}$ sont des réels positifs, $b \in \mathbb{R}^+$ et $A_n^2 = a_1^2 + a_2^2 + a_3^2 + ... + a_n^2$.

Démonstration. cf. Bernstein (1946); Uspensky (1937); Yurinskii (1976) □

Corollaire A.1. *a) S'il existe une constante positive $M < \infty$, telle que $|X_1| \leq M$, alors on a*

$$\forall \varepsilon \geq 0, P\left(\left|\sum_{i=1}^{n} X_i\right| > \varepsilon n\right) \leq 2\exp\left\{\frac{-\varepsilon^2 n}{2\sigma^2(1+\frac{M\varepsilon}{\sigma^2})}\right\},$$

où $\sigma^2 = EX_i^2$.

b) Supposons que les $(X_i)_{1\leq i\leq n}$ dépendent de n et que $\sigma_n^2 = EX_i^2$, s'il existe $M = M_n < \infty$ tel que $|X_1| \leq M$, si $\frac{M}{\sigma_n^2} \leq C < \infty$ et si $u_n = n^{-1}\sigma_n^2 \log n$, vérifie $\lim_{n\to\infty} u_n = 0$, alors nous avons

$$\frac{1}{n}\sum_{i=1}^{n} X_i = O_{a.co}(\sqrt{u_n}).$$

Démonstration. a) En appliquant la proposition A.2 à $a_i^2 = \sigma^2$, $A_n^2 = n\sigma^2$ et $b = M$ nous aboutissons à a).

b) Comme $\frac{MU_n}{\sigma_n^2}$ tend vers zéro, il suffit de rependre le résultat a) pour $\varepsilon = \varepsilon_0\sqrt{u_n}$, on arrive donc à l'existence d'une constante C' telle que

$$P\left[\frac{1}{n}\left|\sum_{i=1}^{n} X_i\right| > \varepsilon_0 U_n\right] \leq 2\exp\left\{\frac{-\varepsilon_0^2 \log n}{2\left(1+\varepsilon_0\sqrt{\frac{MU_n}{\sigma_n^2}}\right)}\right\}$$
$$\leq 2n^{-C'}\varepsilon_0^2.$$

Pour ε_0 bien choisi le terme de droite est le terme général d'une série convergente. Ainsi s'achève la preuve de ce corollaire. □

Bibliographie

S. BERNSTEIN : *Probability Theory, 4th ed. (in russian)*. M. L. Gostechizdat, 1946.

D. BITOUZÉ, B. LAURENT et P. MASSART : A Dvoretzky-Kiefer-Wolfowitz type inequality for the Kaplan-Meier estimator. *Ann. Inst. Henri Poincaré, Probab. Stat.*, 35(6):735–763, 1999. ISSN 0246-0203.

N. BRESLOW et J. CROWLEY : A large sample study of the life table and product limit estimates under random censorship. *The Annals of Statistics*, 2(3):437–453, 1974.

A. CARBONEZ, L. GYÖRFI et E. C. van der MEULEN : Partition-estimates of a regression function under random censoring. *Statist. Decisions*, 13:21–37, 1995.

K.-L. CHUNG : An estimate concerning the Kolmogoroff limit distribution. *Transactions of the American Mathematical Society*, 67(1):36–50, September 1949.

G. COLLOMB, W. HÄRDLE et S. HASSANI : A note on prediction via estimation of the conditional mode function. *Journal of Statistical Planning and Inference*, 15:227–236, 1987.

G. COLLOMB, S. HASSANI, P. SARDA et P. VIEU : Estimation non paramétrique de la fonction de hasard pour des observations dépendantes. *Statistique et analyse des données*, 10(3):42–49, 1985.

A. de ACOSTA : A new proof of the Hartman-Wintner law of the iterated logarithm. *The Annals of Probability*, 11(2):270–276, May 1983.

L. DEVROYE et L. GYÖRFI : Distribution-free exponential bounds on the l_1 error of partitioning estimates of a regression function. *In Proceedings of the Fourth Pannonian Symposium on Mathematical Statistics*, p. 67–76, 1983.

L. DEVROYE, L. GYÖRFI, A. KRZYZAK et G. LUGOSI : On the strong universal consistency of nearest neighbor regression function estimates. *The Annals of Statistics*, p. 1371–1385, 1994.

L. DEVROYE et A. KRZYŻAK : An equivalence theorem for l_1 convergence of the kernel regression estimate. *Journal of Statistical Planning and Inference*, 23(1):71–82, 1989.

L. P. DEVROYE, T. WAGNER *et al.* : Distribution-free consistency results in nonparametric discrimination and regression function estimation. *The Annals of Statistics*, 8(2):231–239, 1980.

M. D. DONSKER : Justification and extension of Doob's heuristic approach to the Kolmogorov-Smirnov theorems. *The Annals of Mathematical Statistics*, 23:277–281, 1952.

A. DVORETZKY, J. KIEFER et J. WOLFOWITZ : Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *Ann. Math. Stat.*, 27:642–669, 1956. ISSN 0003-4851.

F. FERRATY et P. VIEU : Statistique fonctionnelle : Modèles nonparamétriques de régression. Notes de cours de DEA, Université Paul Sabatier, Toulouse. France, 2002.

F. FERRATY et P. VIEU : *Nonparametric functional data analysis : Theory and practice*. Springer, 2006.

A. FÖLDES et L. REJTŐ : A LIL type result for the product limit estimator. *Probability Theory and Related Fields*, 56(1):75–86, 1981.

A. GANNOUN, J. SARACCO et K. YU : Nonparametric prediction by conditional median and quantiles. *Journal of statistical Planning and inference*, 117(2):207–223, 2003.

R. GILL : Large sample behaviour of the product-limit estimator on the whole line. *The Annals of Statistics*, 11(1):49–58, 1983.

Z. GUESSOUM et E. OULD SAÏD : On nonparametric estimation of the regression function under random censorship model. *Statistics & Decisions*, 26(3):159–177, 2008.

Z. GUESSOUM et E. OULD SAÏD : Kernel regression uniform rate estimation for censored data under α -mixing condition. *Electronic Journal of Statistics*, 4:117–132, 2010. URL `http://dx.doi.org/10.1214/08-EJS195`.

Z. GUESSOUM et E. OULD SAÏD : Central limit theorem for the kernel estimator of the regression function for censored time series. *Journal of Nonparametric Statistics*, 24(2):379–397, 2012.

L. GYÖRFI et H. WALK : On the strong universal consistency of a recursive regression estimate by Pál Révész. *Statistics & Probability Letters*, 31:177–183, 1997.

P. HALL : Laws of the iterated logarithm for nonparametric density estimators. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 56 (1):47–61, 1981.

W. HARDLE : A law of the iterated logarithm for nonparametric regression function estimators. *The Annals of Statistics*, p. 624–635, 1984.

W. HARDLE : *Applied nonparametric regression*, vol. 27. Cambridge Univ Press, 1990.

P. HARTMAN et A. WINTNER : On the law of the iterated logarithm. *American Journal of Mathematics*, 63(1):169–176, January 1941.

P. HERBERT LEIDERMAN, B. BABU, J. KAGIA, H. C. KRAEMER et G. F. LEIDERMAN : African infant precocity and some social influences during the first year. *Nature*, 242:247–249, 1973.

P. L. HSU et H. ROBBINS : Complete convergence and the law of large numbers. *Proceedings of the national academy of sciences*, 33(2), 1947.

E. L. KAPLAN et P. MEIER : Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53:457–481, 1958.

K. KEBABI, I. LAROUSSI et F. MESSACI : Least squares estimators of the regression function with twice censored data. *Statistics & Probability Letters*, 81(11):1588–1593, 2011.

K. KEBABI et F. MESSACI : Rate of the almost complete convergence of a kernel regression estimate with twice censored data. *Statistics & Probability Letters*, 82(11):1908–1913, 2012.

S. KHARDANI, M. LEMDANI et E. OULD SAÏD : Uniform rate of strong consistency for a smooth kernel estimator of the conditional mode for censored time series. *Journal of Statistical Planning and Inference*, 141(11): 3426–3436, 2011.

A. KHINCHINE : Über einen Satz der Wahrscheinlichkeitsrechnung. *Fundamenta Mathematica*, 6:9–20, 1924.

J. KIEFER : On large deviations of the empiric df of vector chance variables and a law of the iterated logarithm. *Pacific J. Math*, 11(3):649–660, 1961.

M. KOHLER : On the universal consistency of a least squares spline regression estimator. *Math. Methods Statist.*, 6:349–364, 1997.

M. KOHLER : Universally consistent regression function estimation using hierarchical b-splines. *J. Multivariate Anal.*, 67:138–164, 1999.

M. KOHLER et A. KRZYŻAK : Nonparametric regression estimation using penalized least squares. *IEEE Trans. Inform. Theory*, 47:3054–3058, 2001.

M. KOHLER, K. MÁTHÉ et M. PINTÉR : Prediction from randomly right censored data. *J. Multivariate Anal.*, 80:73–100, 2002.

A. KOLMOGOROV : Über das Gesetz des iterierten Logarithmus. *Mathematische Annalen*, 101:126–135, 1929.

G. LUGOSI et K. ZEGER : Nonparametric estimation via empirical risk minimization. *IEEE Trans. Inform. Theory*, 41:677–687, 1995.

F. MESSACI : Local averaging estimates of the regression function with twice censored data. *Statistics & Probability Letters*, 80(19-20):1508–1511, 2010.

F. MESSACI et N. NEMOUCHI : A law of the iterated logarithm for the product limit estimator with doubly censored data. *Statistics & Probability Letters*, 81(8):1241–1244, 2011.

F. MESSACI et N. NEMOUCHI : Erratum to "a law of the iterated logarithm for the product limit estimator with doubly censored data" [Statist. Probab. Lett. 81 (2011) 1241–1244]. *Statistics & Probability Letters*, 83(9):2142, 2013.

D. MORALES, L. PARDO et V. QUESADA : Bayesian survival estimation for incomplete data when the life distribution is proportionally related to the censoring time distribution. *Comm. Statist. Theory Methods*, 20:831–850, 1991.

E. A. NADARAYA : On estimating regression. *Theory of Probability & Its Applications*, 9(1):141–142, 1964.

E. A. NADARAYA : *Nonparametric estimation of probability densities and regression curves*. Springer, 1989.

V. PATILEA et J. M. ROLIN : Product-limit estimators of the survival function with twice censored data. *Ann. Statist.*, 34(2):925–938, 2006.

M. SAMANTA et A. THAVANESWARAN : Non-parametric estimation of the conditional mode. *Communications in Statistics-Theory and Methods*, 19 (12):4515–4524, 1990.

E. F. SCHUSTER *et al.* : Joint asymptotic distribution of the estimated regression function at a finite number of distinct points. *The Annals of Mathematical Statistics*, 43(1):84–88, 1972.

P.-s. SHEN : Nonparametric estimators of the survival function with twice censored data. *Annals of the Institute of Statistical Mathematics*, 63(6):1207–1219, 2011.

P.-s. SHEN : Modified self-consistent estimators of the survival function with twice censored data. *Journal of Statistical Planning and Inference*, 142 (6):1549–1556, 2012.

C. J. STONE : Consistent nonparametric regression. *The annals of statistics*, p. 595–620, 1977.

W. STUTE et J.-L. WANG : The strong law under random censorship. *The Annals of Statistics*, 21(3):1591–1607, 1993.

B. W. TURNBULL : Nonparametric estimation of a survivorship function with doubly censored data. *Journal of the American Statistical Association*, 69(345):169–173, March 1974.

B. W. TURNBULL : The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 38(3):290–295, 1976.

J. USPENSKY : *Introduction to mathematical probability*. McGraw-Hill, New York, 1937.

S. VOLGUSHEV : *Non-parametric Quantile Regression for Censored Data*. Thèse de doctorat, Ruhr-Universität Bochum, 2009.

G. S. WATSON : Smooth regression analysis. *Sankhyā : The Indian Journal of Statistics, Series A*, p. 359–372, 1964.

B. B. WINTER, A. FÖLDES et L. REJTŐ : Glivenko-Cantelli theorems for the product limit estimate. *Problems of Control and Information Theory*, 7:213–225, 1978.

V. YURINSKII : Exponential inequalities for sums of random vectors. *Journal of Multivariate Analysis*, 6:475–499, 1976.

Résumé

L'objet de cette thèse est d'établir des résultats asymptotiques d'un estimateur à noyau de la fonction de régression analogue à l'estimateur de Nadaraya-Watson mais pour une variable réponse soumise à une censure mixte. Il s'agit de la vitesse de convergence ponctuelle et uniforme et de la normalité asymptotique. Le mode de convergence utilisé est celui de la convergence presque complète. Cette notion de convergence presque complète entraîne la convergence presque sûre.

Nous y développons aussi l'article de Messaci (2010) qui introduit cet estimateur et celui de Messaci et Nemouchi (2011) qui démontre une loi du logarithme itéré de l'estimateur de Patilea et Rolin (2006) de la fonction de survie. Notons que cet estimateur intervient explicitement dans l'expression de l'estimateur à noyau de la régression qui fait l'objet de notre étude. Nous donnons aussi des illustrations de nos résultats sur des données simulées. Notre cadre de travail est celui de l'estimation non-paramétrique de la régression et des données censurées.

Mots clés : Régression non paramétrique, données censurées, estimateur à noyau, normalité asymptotique, vitesse de convergence.

Abstract

The purpose of this dissertation is to establish asymptotic results of a kernel type estimator of the regression function. This estimator is analogous to the Nadaraya-Watson estimator but the response variable is subject to twice censorship. We are conserned with the pointwise and uniform convergence rate and asymptotic normality. The used convergence mode is the almost complete convergence's. This notion of almost complete convergence leads to almost sure convergence.

We also develop the article of Messaci(2010) who introduced this estimator and the one of Messaci and Nemouchi(2011) which proves a law of the iterated logarithm of the estimator of Patilea et Rolin(2006) of the survival function. Let us note that this estimator is explicitly involved in the expression of the kernel estimator of the regression that is the subject of our study. We will also give illustrations of our results on simulated data. Our framework is that of nonparametric estimation of regression and censored data.

Key words : Nonparametric regression, censored data, kernel estimator, asymptotic normality, rate of convergence.

ملخّص

في هذه الرسالة، نقدم بعض النتائج - عند اللانهاية - لتقدير دالة الانحدار بطريقة النواة المماثلة لطريقة ندرايا-واتسون لكن في حالة المعطيات الخاضعة لحجب مزدوج.

هذه النتائج تخصّ سرعة التقارب النقطية والمنتظمة لمقدر النواة، وقانون الاحتمال اللانهائي لهذا المقدر. مفهوم التقارب المستعمل هو التقارب شبه الكامل وهو يستلزم التقارب شبه الأكيد. نقوم كذلك بشرح مقال Messaci (2010) الذي عُرّف فيه المقدّر المذكور كما نقوم بشرح مقال Messaci et Nemouchi (2011) الذي فيه بُرهن على قانون اللّوغارتم المكرّر الخاص بمقدر دالّة التوزيع الاحتمالي المعرف من طرف Patilea and Rolin (2006)، علما ان المقدر آنف الذكر يظهر صراحة في عبارة مقدر دالّة الانحدار محل دراستنا.

وفي النهاية، يتم إجراء دراسة محاكاة لاستكشاف خصائص المقدر المدروس من اجل عينات ذات حجم منته.

الكلمات المفتاحية الانحدار غير الوسيطي، معطيات محجوبة، مقدّر ذو نواة، سرعة التّقارب، التّقارب الطبيعي.