RÉPUBLIQUE ALGÉRIENNE DÉMOCRATIQUE ET POPULAIRE
MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR ET DE LA RECHERCHE SCIENTIFIQUE

UNIVERSITÉ FRÈRES MENTOURI CONSTANTINE 1
FACULTÉ DES SCIENCES EXACTES
DÉPARTEMENT DE MATHÉMATIQUES

N° d'ordre : 41/D3C/2019
N° de série : 01/math/2019

THÈSE

présentée pour l'obtention du diplôme de

Doctorat 3ème Cycle en Mathématiques

intitulée

Estimation non paramétrique dans un modèle de censure et de dépendance

---

présentée par

Nour El Houda ROUABAH

OPTION : Probabilités et Statistique

Devant le jury :

Président :	Z.	MOHDEB	Prof.	Ecole Polytechnique de Constantine
Encadreur :	N.	NEMOUCHI	Prof.	Université Frères Mentouri Constantine 1
Co-encadreur :	F.	MESSACI	Prof.	Université Salah Boubnider Constantine 3
Examinateur :	S.	BELALOUI	M.C(A)	Université Frères Mentouri Constantine 1
Examinateur :	S.	KHARFOUCHI	M.C(A)	Université Salah Boubnider Constantine 3

Soutenue le : 05 Mai 2019

"Nous entendons souvent dire
que les mathématiques consistent à
prouver des théorèmes.
Le travail d'un écrivain serait-il
d'écrire des phrases ?
L'œuvre d'un mathématicien est surtout
un enchevêtrement de conjectures,
d'analogies,
de souhaits et de frustrations ;
la démonstration,
loin d'être le noyau de la découverte,
n'est souvent que le moyen de s'assurer
que notre esprit ne nous joue pas des tours."
Gian Carlo Rota

Remerciements

Je tiens, tout d'abord, à exprimer ma profonde reconnaissance aux directrices de thèse, Professeur Fatiha Messaci et Professeur Nahima Nemouchi, qui m'ont accordé leur confiance en m'offrant la possibilité de réaliser ce travail de recherche. Je tiens à leur témoigner ma gratitude pour avoir accepté de m'initier à la recherche et de m'avoir insufflé la rigueur mathématique indispensable à l'exercice délectable pour la rédaction du manuscrit. Je les remercie particulièrement pour leur disponibilité en toutes circonstances et leur soutien dans les moments difficiles. Elles demeureront pour moi un modèle de compétences scientifiques et de qualité humaine. J'espère pouvoir transmettre dans l'avenir ne serait-ce qu'une partie de tout ce qu'elles m'ont appris, meilleure façon pour moi de leur rendre hommage.
Je remercie Professeur Z. Mohdeb de me faire l'honneur de présider le jury. Mes vives gratitudes au Professeur S. Belaloui et le Professeur S. Kharfouchi d'avoir accepté d'être examinateurs de ce travail.
Mes remerciements à toute l'équipe LAMASD. Qu'il me soit permis de remercier Professeur K. Kebabi et Professeur S. Kharchoufi pour leurs conseils, leur sympathie et disponibilité. Mes pensées éminentes vont aussi à l'ensemble de mes collègues pour leur soutien et leur encouragement lorsque le besoin s'en faisait sentir.
J'exprime toute ma gratitude et mes considérations pour mes parents. À Maman ! Merci mille fois ! je te suis reconnaissante de tout et encore plus. Tu es la personne que j'admire le plus parce que je te dois tout dans la vie. À mon Papa, tu restes un modèle et un jalon essentiel de mon existence. Je vous serai toujours reconnaissante pour m'avoir poussé et orienté vers la voie de la recherche.
Un éternel remerciement à mon époux, qui a tout le temps supporté mes caprices et soutenu dans les moments difficiles afin d'atteindre mon objectif.
Mes vives pensées vont aussi à mes adorables sœurs, Yasmine et Sara, pour leur appui et leur patience ainsi, qu'à mon frère ainé Taha, qui m'a toujours guidé. Je tiens à les remercier intensément.
Mes remerciements vont également à mes amies et à toute la grande famille sans oublier ma belle famille.
Petit clin d'œil à mon neveu Ilyes - Med Larbi "Lissou". Au regard pétillant de malice et au sourire ravageur qui me font fondre de bonheur !
Cette thèse est dédiée particulièrement à mes Grands Parents, qui espéraient assister en ces moments de réussite, appelés auprès du Miséricordieux, sans pouvoir m'accompagner à terme.

Que Dieu ait leurs âmes, reposez en paix ! ! !

Article

Rouabah, N. E. H., Nemouchi, N. & Messaci, F. (2018). A rate of consistency for non-parametric estimators of the distribution function based on censored dependent data. *Statistical Methods & Applications*, 1–22.

Communications

1. Rouabah, N. E. H., Nemouchi, N. & Messaci, F. Une loi du logarithme itéré pour l'estimateur de la fonction de répartition basé sur des données dépendantes et doublement censurées. CMA2014. Tlemcen, mai 2014.

2. Rouabah, N. E. H., Nemouchi, N. & Messaci, F. Étude d'un estimateur à noyau de la régression pour des données censurées et dépendantes. CISAA2014. Constantine, 30 novembre et 1 décembre 2014.

Lexique

Abréviations

v.a.r.	variable(s) aléatoire(s) réelle(s).
i.i.d.	indépendantes et identiquement distribuées.
p.s.	presque sûrement.
p.c.	presque complètement.
$\log_2 n$	$\log \log n$.
FE	Station de Fort Erie.
NOTL	Station de Niagara on the Lake.
ICI	Intervalle de Confiance Inférieur.
ICS	Intervalle de Confiance Supérieur.

Notations

Soient U, V deux variables aléatoires, L une fonction de $\mathbb{R} \to \mathbb{R}$ et $(u_n)_{(n\geq 1)}$, $(v_n)_{(n\geq 1)}$ deux suites réelles positives. Nous considérons les notations suivantes :

$\mathbb{1}_A$	Indicatrice de l'ensemble A.
$\mathbb{P}(V \in A)$	La probabilité que V appartient à l'ensemble A.
$\mathbb{E}(V)$	Espérance de V.
$\sigma(V)$	σ-algèbre des événements engendrés par V.
$\text{cov}(U, V)$	Covariance entre U et V.
$\text{var}(V)$	Variance de V.
F_V	Fonction de répartition de V.
S_V	Fonction de survie de V.
f_V	Densité de probabilité de V.
λ_V	Taux de hasard de V.
Λ_V	Taux de hasard cumulé de V.
I_V	Le point initial du support de V :
	$I_V := \inf\{t \in \mathbb{R} / F_V(t) > 0\}$.
T_V	Le point terminal du support de V :
	$T_V := \sup\{t \in \mathbb{R} / F_V(t) < 1\}$.
$L(t^-)$	$\lim_{x \overset{<}{\to} t} L(x)$.
$L(t^+)$	$\lim_{x \overset{>}{\to} t} L(x)$.
$L(+\infty)$	$\lim_{x \to +\infty} L(x)$.
$u_n = o(v_n)$	$\forall \epsilon > 0$, $\exists$ un entier N / $\forall n \geq N : u_n \leq \epsilon v_n$, pour un n assez grand.
$u_n = O(v_n)$	$\exists \gamma > 0 /$ $u_n \leq \gamma v_n$, pour un n assez grand.

Table des matières

Liste des figures

Liste des tableaux

Introduction Générale

L'analyse des données de survie voit le jour au $XVII^{\text{e}}$ siècle, dans le domaine de la démographie. L'objectif des analystes de ce siècle est l'estimation, à partir des registres de décès, de diverses caractéristiques de la population, son effectif, sa longévité, ...etc. Ce n'est qu'à partir du XIX^{e} siècle, qu' apparaîssent les premières modélisations concernant la probabilité de mourir à un certain âge, probabilité qui sera par la suite désignée sous le terme de "fonction de risque". Ce qui explique que le terme de "survie" soit le plus communément usité dans la littérature statistique. Enfin, l'analyse des données de survie commence de dépasser le cadre stricte de la démographie pour investir, au XX^{e} siècle, toutes les disciplines susceptibles d'avoir recours à de tels types de données. En effet, les données de survie ne sont pas l'apanage des biostatisticiens et sont aussi présentes dans des domaines comme la fiabilité, l'économie, l'assurance, la psychologie, la sociologie, ...etc. En fiabilité industrielle, les événements d'intérêt sont par exemple les durées de vies des composants d'un système, les économistes s'intéressent à des durées d'épisodes de chômage, les assureurs au temps d'arrivée d'un sinistre tandis que les psychologues mesurent le temps qu'il faut à un sujet pour accomplir une tâche donnée. En sociologie, le choix est vaste avec les successions des événements de vie : mariage, naissance du premier enfant, divorce ...etc. Ces études ne représentent que quelques exemples illustrant le grand nombre et la diversité des données de nature censurée auxquelles les statisticiens peuvent être confrontés. Raison pour laquelle l'analyse des données de survie joue un rôle central en statistique.

Le modèle statistique de données de survie exploite toute l'information qui peut être recueillie, qu'elle soit ou non censurée. Cela a suscité un nouveau champ de la Statistique. Depuis cette époque, de nombreux auteurs, ont introduit des techniques nouvelles pour aborder ces modèles. Dans cette perspective, les idées de Kaplan et Meier ont conduit à une innovation majeure. Le travail de Kaplan et Meier [1958] présente d'importants résultats concernant l'estimation non-paramétrique de la fonction de survie pour des variables aléatoires censurées à droite. Il a en particulier permis l'estimation de caractéristiques fonctionnelles associées à la loi des observations concernant, par exemple, l'estimateur à noyau de la densité, introduit par Földes et Rejtő [1981a], du taux de hasard, élaboré par Földes *et al.* [1981] et Diehl et Stute [1988], ou de la régression, donné dans Kohler *et al.* [2002], dédiés à l'étude de modèles impliquant des variables aléatoires censurées à droite. Ces estimateurs à noyau restent parmi les plus utilisés, leurs propriétés asymptotiques ont beaucoup enrichi la littérature statistique, ils ont été étudiés de manière intensive pour des données indépendantes ainsi que pour des données soumises à un certain type de dépendance. Nous donnons dans le chapitre qui suit une brève revue bibliographique

des travaux portant sur l'étude de chacun de ces estimateurs.

Bien que la censure à droite soit la plus courante dans la pratique, d'autres types de censure, généralisant la seule censure à droite, peuvent aussi intervenir.
Un autre résultat majeur de l'étude des données censurées est introduit par Turnbull [1974], étant le premier à s'intéresser à l'estimateur non-paramétrique de la fonction de survie, dans un modèle de censure double. Plus tard, Patilea et Rolin [2006] proposent un estimateur non-paramétrique de la fonction de survie dans un autre modèle de censure, où l'on suppose, contrairement au modèle de Turnbull [1974], l'indépendance des variables latentes. Son modèle est dit de censure mixte et il sera défini plus tard dans cette thèse. Par ailleurs surgit l'idée de proposer des estimateurs non-paramétriques des fonctions d'intérêt en analyse de survie pour ces deux modèles de censure. Ce sujet a donné lieu à beaucoup de travaux.

Par ailleurs, l'indépendance des observations est souvent imposée dans les études de statistique mathématique. Or, cette hypothèse s'avère dans certains cas inadmissible. En effet, un échantillon peut présenter, du fait même de sa constitution, une structure de dépendance. Au XX^{e} siècle des phénomènes étudiés dans la physique, la chimie, la biologie, l'économie et la fiabilité ont permis le développement des modèles stochastiques à variables aléatoires dépendantes. Depuis, la théorie des processus stochastiques a connu un fort engouement et de très nombreux auteurs se sont intéressés aux différentes structures de dépendance concernant une grande variété de domaines. Une vaste littérature est consacrée à l'étude de la dépendance faible, elle a été modélisée par plusieurs notions parmi lesquelles la notion de mélange fort, appelée aussi α-mélange, bénéficie d'un intérêt particulier. Cela est dû au fait qu'elle permet la modélisation de phénomènes plus généraux. Ce qui justifie amplement que les résultats présentés dans ce travail de recherche s'appliquent directement à des échantillons de variables aléatoires de nature α-mélangeantes.

Le comportement asymptotique des estimateurs non-paramétriques construits à partir de l'observation d'un processus défini dans un cadre de dépendance constitue un sujet qui a suscité une intense activité de recherche et permis l'introduction de nombreux outils et méthodes considérant des contextes multiples et variés. La motivation constante qui a dicté le choix de la problématique de la thèse consiste clairement à une contribution à la gestion de la dépendance. Dépendance forte ou mémoire longue, dépendance directionnelle ou causalité, mélange α ou β, géométrique ou arithmétique, stationnarité : ce sont différentes facettes d'un même problème. Dans tous les cas, les questions qui se posent sont directement liées à cette dépendance à savoir, comment la modéliser, comment l'estimer et/ou comment estimer des fonctions malgré l'absence d'indépendance.

Les travaux de la thèse portent sur les estimateurs non-paramétriques basés sur un procédé indépendant de la loi des données observées. Contrairement à l'approche paramétrique qui se restreint à l'estimation d'un certain nombre fini de paramètres liés à la loi de l'échantillon. De ce fait, l'estimation non-paramétrique offre une très grande flexibilité de modélisation pour les applications réelles. À cet effet, nous considérons le modèle I de Patilea et Rolin [2006] et nous entreprenons d'étendre certaines propriétés asymptotiques de quelques estimateurs non-paramétriques basés sur des données censurées et

indépendantes. La motivation essentielle est d'établir des propriétés asymptotiques tout en considérant un cadre de dépendance des données assez général qui puisse être facilement utilisé en pratique. Ainsi, notre contribution est dispensée dans quatre chapitres.

Le Chapitre 1 est un chapitre introductif dans lequel nous définissons les principales notions mathématiques utilisées tout au long de notre travail. Nous y rappelons les différents modèles et types de censure et de convergence et nous présentons également la définition du mélange fort puis les expressions de certains estimateurs non paramétriques.

Dans le Chapitre 2 nous abordons le cas de l'estimateur de Patilea et Rolin qui prolonge la notion de processus empirique standard au cas de censure mixte. Sous la condition du α-mélange, nous aboutissons à une loi du logarithme itéré. Ce résultat est une extension du travail antérieur de Messaci et Nemouchi [2011] établi dans le cas de données indépendantes.

Le Chapitre 3 traite de la consistance presque complète ponctuelle et uniforme des estimateurs non-paramétriques de la fonction de distribution et du taux de hasard cumulé, avec des données α-mélangeantes complètes puis censurées à gauche, avec une vitesse de convergence, afin d'en déduire le cas de la censure mixte de Patilea et Rolin [2006]. Nous concluons ce chapitre avec une étude de simulation et une application sur des données réelles. Ce travail a fait l'objet d'un article publié dans "*Statistical Methods & Applications*".

Dans la suite de ces idées, nous présentons, dans le Chapitre 4, des propriétés asymptotiques similaires pour les estimateurs à noyau de la fonction de densité et de la fonction de hasard dans un cadre de censure mixte. Ces résultats généralisent ceux de Kitouni *et al.* [2015], prouvés pour des données indépendantes, au cas du α-mélangeant.

Nous concluons ce manuscrit avec quelques perspectives de recherche à venir et une annexe regroupant un ensemble d'outils classiques utilisés pour les démonstrations de nos résultats.

Chapitre 1

Préliminaires

Sommaire

Dans ce chapitre nous introduisons des définitions et des notations intervenant dans la suite, et nécessaires pour une meilleure compréhension de ce travail de recherche. Nous commençons par définir le concept général de la censure ainsi que ses différents modèles et types. Nous évoquerons par la suite les formes de dépendance faible connues dans la littérature afin de motiver et d'introduire la notion des données fortement mélangeantes. Aussi, nous rappelons les outils nécessaires à l'estimation non-paramétrique dans des modèles de données complètes ou censurées.

1.1 Censure

Les données censurées proviennent du fait qu'on n'a pas accès à toute l'information, c'est-à-dire que le temps de survie n'est pas exactement connu. Il est à noter que les données censurées sont différentes des données manquantes, car les observations censurées fournissent des informations partielles permettant de fixer une borne inférieure (censure à droite) et (ou) une borne supérieure (censure à gauche), alors que les observations manquantes ne fournissent aucune information sur la variable d'intérêt. La censure à droite est le phénomène le plus couramment rencontré lors du recueil de données de survie, c'est la forme de censure qui reçoit le plus d'attention dans la littérature. Néanmoins, une situation de censure plus complexe peut être rencontrée dans la pratique comme celle de la censure mixte que nous détaillerons un peu plus loin.

1.1.1 Modèles de censure

I. **Censure à droite**

Dans la plupart des études prospectives, les individus sont suivis pendant une durée d'observation fixée à l'avance. Pour les sujets pour lesquels l'événement d'intérêt a lieu pendant la période d'observation, on dispose du moment exact d'apparition de cet événement d'intérêt, mesuré depuis une date initiale qu'il faut spécifier sans ambiguité (date de randomisation dans les essais cliniques, par exemple). Cependant, à la fin de la période d'observation, l'événement d'intérêt n'est pas réalisé chez certains sujets. On n'aura alors pour ces sujets qu'une information incomplète, à savoir que le délai d'apparition de l'événement est plus grand que la durée d'observation. De telles données sont dites censurées à droite et la durée d'observation constitue la variable de censure. Comme les sujets rentrent dans l'étude à des dates différentes, chaque sujet a un temps d'apparition et un temps de censure qui lui sont propres. L' observation est donc le minimum entre le temps d'apparition de l'événement et celui de la censure. Ce mécanisme de censure constitue la censure aléatoire à droite, qui est le cas le plus courant dans les études prospectives. Les sujets pour lesquels l'événement d'intérêt n'a pas eu lieu pendant la période d'observation sont dits "exclus vivants" à la fin de l'étude. Une autre cause de censure à droite, qu'on essaie de limiter au maximum dans les études, correspond aux sujets dits "perdus de vue", qui sont les individus qui ont quitté l'étude avant l'apparition de l'événement d'intérêt et dont on n'a plus de nouvelles à la date de fin d'étude.

De façon formelle, la durée de survie du i^{me} sujet, Y_i, est dite censurée à droite si ce dernier n'a pas subi l'événement à sa date de dernière nouvelle R_i, c'est-à-dire que la seule information dont on dispose est que $Y_i > R_i$. On associe à chaque sujet i la variable aléatoire D_i :

$$D_i = Y_i \wedge R_i = \min(Y_i, R_i),$$

qui est la durée réellement observée et un indicateur $\delta_i = \mathbb{1}_{\{Y_i \leq R_i\}}$ tel que :

$$\delta_i = \begin{cases} 0 & \text{si la durée est censurée à droite} \quad (\text{dans ce cas } D_i = R_i) \\ 1 & \text{si la vraie durée est observée} \quad (\text{dans ce cas } D_i = Y_i) \end{cases}$$

II. **Censure à gauche**

La censure à gauche est moins fréquente dans la pratique, elle correspond au cas où le sujet a déjà subi l'événement avant qu'il ne soit observé. On sait uniquement que la date de l'événement est inférieure à une certaine date connue. Ce type de censure est fréquemment rencontré dans les études épidémiologiques. La quantification des concentrations de polluants dans les études d'exposition environnementale en est un exemple, ou encore la mesure de la charge virale dans les études sur le virus de l'immunodéficience humaine (VIH) ou de l'hépatite C (VHC). En effet, En présence d'une sensibilité insuffisante de la technique de dosage de la mesure biologique, il existe un seuil de quantification analytique, en dessous duquel la valeur exacte de la mesure n'est pas connue. Ces données non détectées par la technique de dosage sont dites censurées à gauche à la valeur du seuil de quantification.

Dans ce cadre, on peut associer, pour chaque sujet i, un couple de variables aléatoires (G_i, δ_i) :

$$G_i = Y_i \vee L_i = \max(Y_i, L_i), \text{et } \delta_i = \mathbb{1}_{\{Y_i \geq L_i\}}.$$

Dans le cas où l'échantillon ne contient que des données censurées à gauche, l'étude peut se faire de manière tout à fait symétrique au cas de donnés censurés à droite, ceci explique le fait que très peu de travaux s'intéressent à la seule censure à gauche. Cependant, Les choses se compliquent considérablement lorsque les deux types de censure cohabitent dans le même échantillon comme c'est le cas dans les exemples suivants.

III. **Censure par intervalle**

La durée de survie X_i est dite censurée par intervalle si au lieu de l'observer de façon exacte, la seule information dont on dispose est qu'elle soit comprise entre deux dates connues $(C_1 < X_i < C_2)$. La censure par intervalle se rencontre généralement lorsque les sujets ne sont pas observés en temps continu mais par intermittence

lors de visites. Par exemple, si l'on s'intéresse à l'âge de survenue d'une maladie et que le sujet i est diagnostiqué malade au cours d'une visite, on sait seulement que $X_i \in [C_1, C_2]$ où C_2 est l'âge à la visite de diagnostic et C_1 est l'âge à la visite précédente.

IV. **Censure mixte**

Il y a censure mixte lorsque deux phénomènes de censure, l'un à gauche et l'autre à droite, peuvent empêcher l'observation du phénomène d'intérêt sans qu'on puisse nécessairement déterminer un intervalle auquel il appartient. Ce modèle de censure a été appliqué par Morales *et al.* [1991] pour rechercher la cause de la mort des arbres, qui sont morts avant la fin de l'étude et la date de l'infection était inconnue. Ce modèle correspond au modèle I de Patilea et Rolin [2006], où selon eux la variable d'intérêt Y est censurée à droite par une variable R. De plus, $\min(Y, R)$ est censurée à gauche par une variable L, si bien que les observations consistent en un échantillon $(Z_i, \delta_i)_{1 \leq i \leq n}$ du couple (Z, δ) avec $Z = (Y \wedge R) \vee L$ où les variables Y, R et L sont des variables aléatoires positives et indépendantes, l'indicateur de censure δ, est une variable discrète à valeur $\{0, 1, 2\}$, où 0 correspond à l'observation de Y, tandis que 1 et 2 indique l'observation de R et L, respectivement. Plus précisément,

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

Notons que l'indépendance des variables Y, L et R assure l'identifiabilité du modèle.

À l'intérieur de ces modèles, il existe différents types de censure :

1.1.2 Types de censure

I. **La censure de type I (fixe)**

Le temps de censure est fixé par le chercheur comme étant la fin de l'étude. Soit C une valeur fixée, au lieu d'observer les variables $Y_1, \ldots, Y_n$ qui nous intéressent, on n'observe Y_i que lorsque $Y_i \leq C$, sinon on sait uniquement que $Y_i > C$. On utilise la notation suivante : $D_i = Y_i \wedge C = \min(Y_i, C)$.
Ce mécanisme de censure est fréquemment rencontré dans les applications industrielles. Par exemple, on peut tester la durée de vie de n objets identiques (ampoules) sur un intervalle d'observation fixé $[0, C]$.

II. **La censure de type II (attente)**

Elle se caractérise par le fait que l'étude cesse aussi-tôt qu'a eu lieu un nombre d'événements prédéterminé par l'expérimentateur. Par exemple, quand on décide

d'observer les durées de survie des n patients jusqu'à ce que k d'entre eux soient décédés et d'arrêter l'étude à ce moment là.

III. **La censure de type III (ou censure aléatoire de type I)**

C'est le cas lorsque le moment de censure n'est plus sous le contrôle du chercheur et/ou que le temps d'entrée varie aléatoirement, c'est le type de censure le plus courant. Par exemple, lors d'un essai thérapeutique, elle peut être engendrée par la perte de vue (le patient quitte l'étude en cours et on ne le revoit plus), l'exclusion de l'étude (suite à un changement ou un arrêt du traitement).

Dans les modèles classiques d'analyse de survie (cas de censures aléatoires), une hypothèse fondamentale sur les censures consiste à supposer que la censure est indépendante de l'événement. Cela signifie que les perdus de vue ne le sont pas pour des raisons liées à l'événement étudié. Cette hypothèse est importante d'un point de vue mathématique et régle le problème d'identifiabilité.

1.2 Mesures de dépendance

L'hypothèse fondamentale dans la construction de la plupart des modèles et tests statistiques est en générale l'indépendance des observations. Cependant, plusieurs exemples pratiques prouvent que cette hypothèse d'indépendance n'est pas réaliste. Si nous nous intéressons au taux de pollution de l'air dans la ville de Constantine, par exemple, le taux de pollution enregistré durant ce mois dépendra clairement de celui enregistré durant le mois précédent. Ces données présentent alors une forme de dépendance, modélisée en mathématique par des données mélangeantes. Les notions de mélange font parties des conditions de dépendance faible les plus étudiées.

Les techniques des suites mélangeantes constituent un moyen de contrôler la dépendance entre les éléments d'une suite, en obtenant une indépendance asymptotique de blocs de suites de variables qui, lorsqu'ils sont éloignés l'un de l'autre, se comportent de manière indépendante. En effet, La dépendance faible est facturée en terme de coefficients de dépendance entre les tribus engendrées par les variables de la suite avant un instant t et les tribus engendrées par les variables après l'instant $t + n$. Les suites ont des propriétés d'autant plus proches de celles des suites indépendantes que les coefficients de mélange décroissent rapidement vers 0 quand n tend vers l'infini.

1.2.1 Mélangeance

Dans la théorie classique, il existe cinq notions importantes de mélange fort introduites pour étudier les suites de variables aléatoires, depuis Rosenblatt [1956] pour le α-mélange, la mesure du β-mélange par Volkonskii et Rozanov [1959], Ibragimov [1959] pour les suites ϕ-mélangeantes, la notion de ψ-mélange introduite par Blum *et al.* [1963] et Philipp [1969a]

et le ρ-mélange par Kolmogorov et Rozanov [1960].

Les travaux de Doukhan [1994] et Bradley [2005] proposent un point de vue global sur ces différentes formes de dépendance faible et décrivent la relation entre elles comme suit :

$$\phi - \text{mélange} \implies \rho - \text{mélange} \implies \alpha - \text{mélange},$$
$$\psi - \text{mélange} \implies \phi - \text{mélange} \implies \beta - \text{mélange} \implies \alpha - \text{mélange}.$$

L'exploitation des comportements asymptotiques de ces mesures, pour n grand, conduit à une sorte d'indépendance asymptotique à partir de laquelle une théorie asymptotique intéressante est développée.

Il est clair que le coefficient du α-mélange est notamment plus faible que les autres coefficients de mélange et donc le moins restrictif. En ce sens, les résultats obtenus concernent une classe plus large de processus. Dans ce travail de recherche, nous modélisons la dépendance au sein de l'échantillon étudié en considérant des processus α-mélangeantes. Ce choix a été motivé par le fait que cette notion est assez générale et que l'on dispose de nombreux résultats permettant d'étudier ce type de processus.

1.2.2 Le mélange fort

Soit $(X_i)_{i\in\mathbb{Z}}$ une suite de variables aléatoires définie sur un espace de probabilitlé $(\Omega, \mathcal{F}, \mathbb{P})$ à valeurs dans espace mesuré $(\Omega', \mathcal{F}')$. On note $\mathcal{F}_j^k$ avec $-\infty \leq j \leq k \leq \infty$, la σ-algèbre engendrée par $\{X_s, j \leq s \leq k\}$.
Le coefficient α du mélange fort a été introduit par Rosenblatt [1956] de la manière suivante :

Définition 1. *La suite $(X_i)_{i\in\mathbb{Z}}$ est dite fortement mélangeante, ou α-mélangeante si*

$$\alpha(n) = \sup_k \sup_{A\in\mathcal{F}_{-\infty}^k} \sup_{B\in\mathcal{F}_{k+n}^{\infty}} |\mathbb{P}(A\cap B) - \mathbb{P}(A)\mathbb{P}(B)| \underset{n\to+\infty}{\longrightarrow} 0.$$

Les ARCH et les GARCH sont des exemples de processus α-mélangeants.

Nous nous intéressons en particulier aux types suivants de α-mélange : arithmétique et géométrique.

Définition 2. — *La suite $(X_i)_{i\in\mathbb{Z}}$ est dite géométriquement α-mélangeante si :*

$$\exists\, C > 0,\ \exists\, t \in]0,1[,\ \alpha(n) \leq Ct^n.$$

Les processus AR, ARMA et les processus bilinléaires (sous certaines conditions d'érgodicité) sont des exemples de processus géométriquement α-mélangeants.

— *La suite $(X_i)_{i\in\mathbb{Z}}$ est arithmétiquement α-mélangeante d'orde $\nu > 0$, si existe une constante $C > 0$ telle que*

$$\alpha(n) \leq Cn^{-\nu}.$$

Remarque. *On peut remarquer facilement que si la suite $(X_i)_{i\in\mathbb{Z}}$ est géométriquement α-mélangeante, elle est arithmétiquement α-mélangeante d'ordre ν, pour tout ν. Nous pourrons donc appliquer les résultats donnés dans les chapitres suivants pour des observations arithmétiquement α-mélangeantes à des données géométriquement α-mélangeantes.*

On trouve dans la littérature un nombre conséquent de travaux consacrés à l'étude des suites de variables aléatoires α-mélangeantes dont les monographies de Doukhan [1994], Bosq [1996], Rio [2000], Yoshihara [2004] et Bradley [2007]. On peut également se référer à Doukhan *et al.* [1994], Rio [1995a] et Liebscher [1996, 2001a] pour des résultats de type théorème central limit. Bosq [1975, 993b], Carbon [1983] et Rhomari [2002] proposent des inégalités exponentielles et des inégalités concernant les covariances. D'autres travaux se sont intéressés à la nature faiblement dépendante de chaînes de Markov et de processus autorogressifs. On peut par exemple citer à ce sujet les travaux de Davydov [1973], Chanda [1974], Gorodetskii [1977], Withers [1981], Mokkadem [1990] et les références qu'ils contiennent.

Tous ces résultats constituent des outils primordiaux dans l'étude des échantillons à base de processus α-mélangeants et permettent de généraliser certains résultats obtenus avec des échantillons indépendants. L'utilisation de ces outils nous permet d'établir les résultas présentés dans cette thèse étudiant le comportement asymptotique des estimateurs non-paramétriques sous la condition du α-mélange.

1.3 Estimation non-paramétrique pour des données complètes

1.3.1 Mode de convergence

Nous rappelons ici les définitions des modes de convergence que nous allons aborder dans les chapitres à venir, à savoir, la convergence presque sûre et la convergence presque complète.

Soit $(\Omega, \mathcal{F}, \mathbb{P})$ un espace de probabilité, (X_n) une suite de variables aléatoires réelles sur Ω et X une variable aléatoire réelle définie sur Ω.

Convergence presque sûre

La notion de convergence presque-sûre correspond à la convergence simple sur tout le complémentaire d'un ensemble négligeable. .

Définition 3. *(X_n) converge vers X presque sûrement s'il existe un ensemble $\mathbb{P}$ négligeable $\mathcal{N}$ tel que*

$$\forall \omega \in \mathcal{N}^c,\ |X_n(\omega) - X(\omega)| \underset{n\to\infty}{\to} 0.$$

On note $X_n \xrightarrow{p.s.} X$.

Cette définition signifie que pour presque tout ω, il est possible de rendre la "distance" entre $X_n(\omega)$ et $X(\omega)$ aussi petite que l'on veut dès que n dépasse un certain N, autrement dit :

$$\forall \omega \in \mathcal{N}^c, \mathbb{P}(\mathcal{N}) = 0; \forall \epsilon > 0, \exists N : \forall n \geq N \Rightarrow |X_n(\omega) - X(\omega)| \leq \epsilon].$$

Quant à la convergence presque complète, nous pouvons l'expliquer comme suit. La suite (X_n) converge presque complètement vers X veut dire que la probabilité que la distance entre X_n et X excède un nombre strictement positif (aussi petit soit il) tend vers zéro assez vite pour être le terme général d'une série convergente

Convergence presque complète

Définition 4. *On dit que la suite $(X_n)_{n\in N}$ converge presque complètement vers X lorsque $n \to \infty$ si*

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

et on note $X_n \xrightarrow{p.c.} X$.

Définition 5. *Soit $(U_n)_{n\in N}$ une suite de nombres réels positifs qui tend vers 0, on dit que la vitesse de convergence presque complète de la suite $(X_n)_{n\in N}$ vers X est d'ordre (U_n) si*

$$\exists\, \epsilon_0 > 0, \quad \sum_{n\in N} \mathbb{P}(|X_n - X| > \epsilon_0 U_n) < \infty,$$

et on note

$$X_n - X = O_{p.c.}(U_n).$$

Nous renvoyons à Ferraty et Vieu [2006] pour un point de vue complet sur ce type de convergence.

Remarque. *Remarquons que le lemme de Borel-Cantelli montre que cette convergence est plus forte que la convergence presque sûre, elle entraîne donc aussi la convergence en probabilité.*

1.3.2 Distribution de la durée de survie

La durée de survie X est une variable aléatoire positive. La distribution de X est caractérisée par l'une des cinq fonctions suivantes définies pour $x \geq 0$, chacune pouvant être obtenue à partir de l'une des autres.

- ❋ La fonction de répartition $F_X(x)$ est la probabilité de subir l'événement avant le temps x :

$$F_X(x) = \mathbb{P}(X \leq x).$$

❋ La fonction de survie $S_X(x)$ qui est la probabilité de survie jusqu'au temps x, elle est définie comme :

$$S_X(x) = \mathbb{P}(X > x) = 1 - F_X(x).$$

C'est donc une fonction continue à droite décroissante et telle que $S_X(0) = 1$ et $\lim_{x\to\infty} S_X(x) = 0$.

❋ Si X est absolument continue, on peut également pour préciser cette distribution recourir à la densité de probabilité $f_X(x)$ qui est la dérivée de Radon Nykodim de la loi de X par rapport à la mesure de Lebesgue. Si de plus f_X est continue, alors :

$$f_X(x) = \lim_{\Delta x\to 0} \frac{\mathbb{P}(x < X \leq x + \Delta x)}{\Delta x} = F'_X(x) = -S'_X(x),$$

et elle est telle que

$$F_X(x) = \int_0^x f(u)du.$$

❋ Un concept important est celui du risque caractérisé par la fonction du taux de hasard $\lambda_X(x)$. C'est le taux de survenue de l'événement durant l'intervalle de temps $[x,\, x + \Delta x[$ sachant qu'il ne s'était pas réalisé avant x :

$$\lambda_X(x) = \lim_{\Delta x\to 0} \frac{\mathbb{P}(x < X \leq x + \Delta x / X > x)}{\Delta x}.$$

Cette fonction est liée aux précédentes puisque :

$$\lambda_X(x) = \begin{cases} \frac{f_X(x)}{S_X(x)} & \text{si} \quad S_X(x) \neq 0, \\ 0 & \text{sinon} \end{cases} \tag{1.1}$$

❋ Il est encore possible de définir la fonction de hasard cumulé selon :

$$\Lambda_X(x) = \int_0^x \lambda_X(u)du.$$

Avec l'égalité suivante entre fonction de survie et fonction de hasard cumulé :

$$\Lambda_X(x) = -\ln(S_X(x)).$$

On peut déduire de cette équation une expression de la fonction de survie en fonction du taux de hasard cumulé (ou du risque instantané) :

$$S_X(x) = \exp(-\Lambda_X(x)) = \exp\left(-\int_0^x \lambda_X(u)du\right).$$

On en déduit que

$$f_X(x) = \lambda_X(x)\exp\left(-\int_0^x \lambda_X(u)du\right).$$

Toutes ces fonctions sont donc liées entre elles, si on se donne une seule de ces fonctions, alors les autres sont dans le même temps également définies. En particulier, un choix de spécification sur la fonction de hasard implique la sélection d'une certaine distribution des données de survie.

1.3.3 Estimation de la fonction de répartition

Soit $(X_i)_{1\leq i\leq n}$ une suite de variables aléatoires réelles indépendantes identiquement distribuées, de même loi que X, de fonction de répartition $F_X(x)$. L'estimation de cette dernière tient une place importante dans l'étude de nombreux phénomènes de nature aléatoire. Le point de départ de l'estimation non-paramétrique de la fonction de répartition fut l'introduction de la fonction de répartition empirique qui se calcule sur la base de véritables observations de la variable d'intérêt X. C'est une fonction en escalier, limitée à gauche et continue à droite qui met un poids $1/n$ sur chaque point X_i telle que :

$$\widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{X_i\leq x\}}. \tag{1.2}$$

Cette estimation est d'excellente qualité. On se place pour commencer dans le cas où les données sont indépendantes. La loi forte des grands nombres nous montre que cet estimateur est fortement consistant sur tout $\mathbb{R}$. Le théorème de Glivenko-Contelli (Voir Annexe A. Théorème 11) améliore ce résultat en donnant la convergence uniforme. Chang [1949] introduit la loi du logarithme itéré pour cet estimateur dans $\mathbb{R}$. Puis Kiefer [1961] a précisé le taux de cette convergence dans $\mathbb{R}^m$.

On se place à présent dans le cas où les données sont α-mélangeantes. Il semble que les premiers travaux consacrés à ce type de situation soient ceux de Collomb *et al.* [1985] et Cai et Roussas [1992]. Dans le premier travail, le résultat de Glivenko-Contelli est amélioré en donnant la convergence presque complète de l'estimateur $\widehat{F}_n(x)$. Dans le deuxième travail, les auteurs ont montré que la fonction de répartition empirique $\widehat{F}_n(x)$ converge uniformément presque surement vers $F_X(x)$ ou $x \in \mathbb{R}$, ils ont aussi établi la loi du logarithme itéré de cet estimateur où l'unique condition imposée sur la suite (X_n) est que son coefficient de mélange soit de la forme $\alpha(n) = O(n^{-\nu})$ où $\nu > 3$.

1.3.4 Estimation de la densité de probabilité

Supposons que la loi de X est absolument continue, par rapport à la mesure de Lebesgue, de densité de probabilité f_X, que nous cherchons à estimer. Ce problème tire son intérêt du fait qu'on peut déduire de l'estimateur des éléments concernant la symétrie ou la multimodalité de la loi étudiée. En effet, même si les fonctions de répartition et de densité caractérisent toutes les deux la loi de probabilité d'une variable, la densité a un net avantage sur le plan visuel. Elle permet d'avoir un aperçu très rapide des principales caractéristiques de la distribution (pics, creux, asymétries, ...), ce qui explique le volume important de littérature qui lui est consacré. La fonction de répartition contient bien sûr cette information mais de manière moins visible.

La méthode du noyau a d'abord été décrite en 1951 dans un rapport non publié par Fix et Hodges (voir Silverman et Jones [1989]). L'une des motivations principales est de construire un estimateur lisse de la densité en remplaçant les indicatrices utilisées dans la construction de l'histogramme mobile par des fonctions continues. Rappelons que la densité de probabilité f_X est égale à la dérivée de la fonction de répartition F_X (si f_X

est continue). Une des premières idées intuitives est donc de considérer pour $h > 0$ assez petit

$$\widehat{f}_n(x) = \frac{\widehat{F}_n(x+h) - \widehat{F}_n(x-h)}{2h} = \frac{1}{2nh} \sum_{i=1}^{n} \mathbb{1}_{\{x-h \leq X_i \leq x+h\}}.$$

Cet estimateur, appelé estimateur de Rosenblatt (Rosenblatt [1956]), est la première version d'estimateur à noyau construit à l'aide du noyau uniforme $K(u) = \frac{1}{2}\mathbb{1}_{\{-1<u\leq 1\}}$.

Parzen [1962] et Rosenblatt [1956] ont suggéré une généralisation de cet estimateur pour une vaste classe de noyaux, en posant pour tout $x \in \mathbb{R}$:

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

où $K : \mathbb{R} \to \mathbb{R}$ est une fonction intégrable, vérifiant la condition $\int K(u)du = 1$ connue depuis sous le nom du noyau et h, la fenêtre de lissage, est un paramètre tendant avec une certaine vitesse vers 0 lorsque n tend vers l'infini. En chaque observation X_i on place une "bosse" (la densité de probabilité K). L'estimateur qui en résulte est simplement la somme de ces "bosses". Le noyau K détermine la forme de la "bosse" et la fenêtre h détermine sa largeur.

La fonction K est une fonction mesurable supposée satisfaire certaines hypothèses basiques de régularité. Les exemples les plus utilisés de noyaux sont : le noyau uniforme, le noyau gaussien, le noyau triangulaire, noyau rectangulaire et le noyau d'Epanechnicov. le Tableau 1.1 nous donne un aperçu de ces fonctions à noyau.

Tableau 1.1 – Exemples de noyaux usuels

Noyau	**Fonction**
noyau uniforme	$K(x) = \mathbb{1}_{\{\lvert x\rvert \leq 1/2\}}$
noyau gaussien	$K(x) = \frac{1}{\sqrt{2\pi}}\, e^{-x^2/2}$
noyau triangulaire	$K(x) = (1 - \lvert x\rvert)\ \mathbb{1}_{\{\lvert x\rvert \leq 1/21\}}$
noyau rectangulaire	$K(x) = \frac{1}{2}\ \mathbb{1}_{\{\lvert x\rvert \leq 1/21\}}$
noyau d'Epanechnicov	$K(x) = \frac{3(1-x^2/5)}{4}\ \mathbb{1}_{\left\{\lvert x\rvert \leq \sqrt{5}\right\}}$

Des travaux ont montré qu'en pratique le choix de la fonction noyau n'influence que peu les résultats d'estimation. La seule exception notable étant liée à l'utilisation d'une fonction

noyau uniforme qui peut donner des résultats sensiblement différents des autres noyaux, les fonctions noyaux triangulaires ou gaussiennes donnent plus de poids au voisinage de zéro. Berlinet [1993] a proposé une méthode automatique du choix du noyau. Pour des résultats et références récentes sur le choix du noyau, nous renvoyons à Vieu [1999].

Le paramètre de lissage h_n a une grande influence sur la performance de l'estimateur. Un h trop petit résulte en un estimateur avec une "bosse" en chaque observation. Un h trop grand résulte en un estimateur qui montre peu de détails. Il faut donc essayer de choisir un h qui fasse un compromis entre le biais et la variance. En particulier, pour que $\widehat{f_n}$ converge simplement dans $\mathbb{L}^2$ vers f_X, il faut choisir $h = h_n$ une suite de nombres réels strictement positifs tel que $h_n \to 0$ et $nh_n \to \infty$ de manière à faire tendre simultanément le biais et la variance vers zéro.

On dispose, dans la plupart des cas, de l'expression de la fenêtre h_n minimisant l'erreur quadratique moyenne (MSE), qui est un critère de selection local, ou bien l'erreur quadratique moyenne intégrée (MISE) qui représente un critère de selection global. Cependant, cette largeur de fenêtre "idéale" (relativement au critère d'erreur retenu) n'est pas directement calculable puisqu'elle dépend de paramètres inconnus, entre autres la dérivée seconde de la fonction à estimer. Partant de ce fait, l'étude des méthodes de sélection du paramètre de lissage a nourri une littérature abondante. Les méthodes reposant sur la validation croisée introduite par Rudemo [1982] et Bowman [1984] sont parmi les plus populaires. D'un point de vue pratique, un des principaux intérêts de cette technique est son caractère "direct". On pourra se référer à Hall [1984], Hardle et Marron [1985] et Hardle [1990] pour approfondir encore plus ce type de méthodes.

Parzen [1962] a établi la normalité asymptotique de $\widehat{f_n}$. Le cas multivarié a été traité par Cacoullos [1966]. Pour la première fois, Foldes [1974] et Rûschendorf [1977] ont étudié la convergence presque sûre de l'estimateur de la densité pour un échantillon ϕ-mélangeant. Le travail de Sarda et Vieu [1989] traite aussi ce sujet et propose des taux de cette covergence. Par la suite Delecroix [1979] a étudié la convergence presque sûre et le convergence en moyenne quadratique de l'estimateur de la densité pour de processus strictement stationnaires et α-mélangeants. La consistence forte de $\widehat{f_n}$ pour des processus stationnaires et α-mélangeantes est prouvée par Roussas [1988], Tran [1990], Nze et Rios [1995] et Liebscher [1998, 2001b]. Yu [1993] a abordé le même problème en présence de processus stationnaires β-mélangeantes. Sa normalité asymptotique pour un échantillon α-mélangeant est établie par Robinson [1983] et Roussas [1990]. Kim et Lee [2005] ont étudié la consistence et le théorème central limit de $\widehat{f_n}$ avec des réalisations de processus non-stationnaires et α-mélangeantes.

1.3.5 Estimation de la fonction de hasard

L'estimation du taux de hasard, de part la variété de ses possibilités d'application, est une question importante en statistique. Une des techniques les plus courantes pour construire des estimateurs de λ_X est basée sur sa définition, donnée par la relation (1.1), et consiste à étudier un quotient entre un estimateur de f_X et un estimateur de S_X. L'article de Patil

et al. [1994] fait une présentation générale de ces techniques d'estimation. Les méthodes non-paramétriques basées sur les idées de noyau, qui sont connues pour leur bon comportement dans les problèmes d'estimation de densité, sont ainsi abondamment utilisées en estimation non-paramétrique de la fonction de hasard. Un large éventail de la littérature dans ce domaine est fourni par les revues bibliographiques de Singpurwalla et Wong [1983], Hassani *et al.* [1986], Izenman [1991], Gefeller et Michels [1992] et Pascu et Vaduva [2003].

Il est donc tout naturel de construire un estimateur de la fonction λ_X en s'inspirant de ces idées de la manière suivante :

$$\widehat{\lambda}_n(x) = \begin{cases} \frac{\widehat{f}_n(x)}{\widehat{S}_n(x)}, & \text{si} \qquad \widehat{S}_n(x) \neq 0, \\ 0 & \text{sinon.} \end{cases}$$

Les propriétés de l'estimateur de la fonction de hasard s'obtiennent relativement facilement à partir de la littérature connue en matière d'estimation des fonctions de répartition et de densité.

1.4 Estimation non-paramétrique dans un modèle de censure à droite

Dans cette section nous nous placerons dans le cadre le plus fréquent d'une censure à droite aléatoire définie dans la section 1.1.1.I.. De tels modèles requièrent l'utilisation de techniques adaptées pour prendre en compte nos observations censurées sans perdre trop d'information sur Y.

1.4.1 L'estimateur de Kaplan-Meier de la fonction de survie

Dans le cas où les données sont censurées, il est impossible d'utiliser la fonction $\widehat{F}_n(t)$ puisqu'elle fait intervenir des quantités non observées (tous les (Y_i) censurées ne sont pas observées). Kaplan et Meier [1958] sont les premiers à considérer le problème de l'estimation non-paramétrique de la fonction de répartition d'une variable aléatoire Y censurée à droite par une variable aléatoire R. ils fournissent un bon estimateur de la fonction de survie $S_Y(t) = 1 - F_Y(t)$ ayant la forme suivante :

$$\bar{S}_n(t) = \prod_{j/Z'_j \leq t} \left(1 - \frac{M(Z'_j)}{R(Z'_j)}\right) = \prod_{j/Z'_j \leq t} \left(\frac{n-j}{n-j+1}\right)^{\delta_j}, \tag{1.3}$$

où Z'_j $(1 \leq j \leq M)$ sont les valeurs distinctes de $D_i = \min(Y_i, R_i)$ rangées dans l'ordre croissant, $M(Z'_j) = \sum_{i=1}^{n} \delta_i \mathbb{1}_{\{D_i = Z'_j\}}$ représente le nombre de sujets subissant l'événement au temps Z'_j et $R(Z'_j) = \sum_{i=1}^{n} \mathbb{1}_{\{D_i \geq Z'_j\}}$ est le nombre de sujets à risque de mourir juste

avant l'instant Z'_j .

On a aussi l'estimateur de Kaplan-Meier de la fonction de survie $S_R(t)$ de la variable de censure R, donné par :

$$\mathring{S}_n(t) = \prod_{j/Z'_j \leq t} \left(\frac{n-j}{n-j+1} \right)^{1-\delta_j} . \tag{1.4}$$

L'estimateur de Kaplan-Meier, est également appelé Produit Limite. C'est une fonction en escalier (dont la valeur change uniquement aux temps correspondant à des événements observés), décroissante, continue à droite. Notons que quand il n'y a pas de censure, l'estimateur de Kaplan-Meier se réduit à la fonction de survie empirique $1 - \widehat{F}_n(x)$.

Le comportement asymptotique de cet estimateur a suscité l'intérêt d'un grand nombre d'auteurs. Pour des variables aléatoires indépendantes, Breslow et Crowley [1974] furent les premiers à traiter de sa convergence et de sa normalité asymptotique. En imposant la continuité de la fonction de répartition de la variable d'intérêt et celle de la variable de censure, Földes et Rejtő [1981b] trouvent un taux de convergence uniforme presque complète pour $\bar{S}_n(t)$ de l'ordre de $\sqrt{\log n/n}$, ces derniers ont aussi introduit la loi du logarithme itéré pour l'estimateur de Kaplan-Meier durant la même année de 1981 (Voir Földes et Rejtő [1981a]). Stute et Wang [1993] traitent à leur tour sa convergence uniforme presque sûre. La loi fonctionnelle du logarithme itéré pour des données censurées à droite ou tronquées est déduite par Gu et Lai [1990].

En modélisant la dépendance par la notion de ϕ-mélangeant Ying et Wei [1994] ont établi la normalité asymptotique de l'estimateur de Kaplan-Meier.Sous l'hypothèse de dépendance forte des variables d'intérêts Cai [1998] a montré la consistance de cet estimateur, en précisant sa vitesse de convergence. Dans son article de 2001 (Voir Cai [2001]) , il généralise le résultat de Cai et Roussas [1992] à l'estimateur de Kaplan-Meier à savoir, la loi du logarithme itéré, sous certaines conditions de régularités et de mélange fort.

Leurs travaux connaissent par la suite un développement intensif tant en théorie qu'en pratique.

1.4.2 Estimateur à noyau de la densité pour des modèles de censure à droite

En supposant que la variable aléatoire positive et censurée à droite Y admet une densité de probabilité f_Y, Földes *et al.* [1981] ont proposé une extension à l'estimateur de Rosenblatt [1956] et Parzen [1962] qui s'exprime comme suit :

$$\bar{f}_n(y) = \frac{1}{h_n} \int K\left(\frac{y-z}{h_n} \right) d\bar{F}_n(z), \tag{1.5}$$

où $\bar{F}_n = 1 - \bar{S}_n$ est l'estimateur de Kaplan-Meier donné par la relation (1.3).

Dans ce même travail ils ont prouvé sa convergence presque complète. Sa normalité asymptotique est établie par Mielniczuk [1986] et améliorée dans Diehl et Stute [1988]. Il a aussi été étudié par Hentzschel et Liebscher [1990] et Xiang [1994]. Par la suite, Kagba [2004] a donné sa convergence en moyenne quadratique. Peu de travaux traitent l'estimateur de la fonction de densité $\bar{f}_n$ dans le cas de données dépendantes. On peut citer le travail de Cai [998b] proposant un taux de convergence presque sûre pour des processus stationnaires et α-mélangeants. Plus tard, Liebscher [2002] a amélioré ce résultat.

1.5 Estimation non-paramétrique dans un modèle de censure mixte

Dans cette section nous nous intéressons à une nouvelle classe d'estimateur où les observations Y_i sont soumises à un mécanisme de censure mixte tel qu'étudié par Patilea et Rolin [2006] et défini dans 1.1.1.IV.. Au cours de la dernière décennie, beaucoup d'intérêt est porté sur l'estimation non-paramétrique dans ce modèle.

1.5.1 L'estimateur de Patilea et Rolin de la fonction de survie

Rappelons que nous observons un échantillon $(Z_i, \delta_i)_{1\leq i\leq n}$ du couple (Z, δ) avec $Z = (Y \wedge R) \vee L = \max(T, L)$ où Y, L et R sont des variables aléatoires positives et indépendantes représentant, repectivement, la variable d'intérêt, la variable de censure à gauche et la variable de censure à droite.

Soit $H(t)$ la fonction de répartition de Z et $H^{(0)}(t)$ sa sous-distribution pour les observations non censurées, ayant les expressions suivantes :

$$H(t) = P(Z \leq t) = F_L(t)F_T(t) = F_L(t)(1 - S_R(t)S_Y(t)), \tag{1.6}$$

où $T = \min(Y, R)$ et

$$H^{(0)}(t) = P(Z \leq t, \delta = 0) = \int_0^t F_L(u)S_R(u)dF_Y(u). \tag{1.7}$$

Ainsi que leurs versions empiriques, données respectivement par :

$$H_n(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{Z_i\leq t\}}, \tag{1.8}$$

et

$$H_n^{(0)}(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{Z_i\leq t, \delta_i=0\}} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{Z_i\leq t,\ Y_i-R_i\leq 0,\ L_i-Y_i\leq 0\}}. \tag{1.9}$$

On note Z'_j $(1 \leq j \leq M)$ les valeurs distinctes de Z_i rangées dans l'ordre croissant et pour $k \in \{0, 1, 2\}$:

$$D_{kj} = \sum_{i=1}^{n} \mathbb{1}_{\{Z_i = Z'_j, \delta_i = k\}}$$

L'estimateur non-paramétrique, noté $\widetilde{S}_n$, de S_Y est l'estimateur produit limit donné par Patilea et Rolin [2006] sous la forme :

$$\widetilde{S}_n(t) = 1 - \widetilde{F}_n(t) = \prod_{j/Z'_j \leq t} \left(1 - \frac{D_{0j}}{n\mathring{F}_n(Z'_{j-1}) - nH_n(Z'_{j-1})}\right), \tag{1.10}$$

où $\mathring{F}_n$ est l'estimateur de Kaplan-Meier de la fonction de répartition F_L, défini en inversant le temps par

$$\mathring{F}_n(t) = \prod_{j/Z'_j > t} \left(1 - \frac{D_{2j}}{nH_n(Z'_j)}\right). \tag{1.11}$$

La convergence presque sûre uniforme de l'estimateur $\widetilde{S}_n$ de Patilea-Rolin, est prouvée dans ce même article.

Nous étudions cet estimateur dans les deux chapitres à venir. L'étude de données α-mélangeantes dans ce cadre de censure mixte est très récente. Nous allons généraliser des résultats obtenus pour des échantillons de v.a. indépendantes à des échantillons de processus fortement mélangeants, en ajoutant des hypothèses sur le coefficient de mélange. Puis, nous établirons dans le dernier chapitre de cette thèse la convergence uniforme presque complète sur un compact des estimateurs non paramétriques de la densité et du taux de hasard pour des échantillions basés sur des observations α-mélangeantes soumises à une censure mixte.

Chapitre 2

La loi du logarithme itéré dans un modèle de dépendance et de censure mixte

Sommaire

Dans ce chapitre, nous nous proposons, après une brève introduction sur l'origine de la loi du logarithme itéré, de généraliser les résultats obtenus sous un modèle de censure mixte dans le cadre indépendant aux processus fortement mélangeants.

2.1 Introduction

Soit (T_n) une suite de variables aléatoires indépendantes et identiquement distribuées, on pose

$$S_n = \sum_{i=1}^{n} T_i.$$

A l'instar du théorème central limite (voir Annexe A. Théorème 10 du document), la loi du logarithme itéré illustre le fait fascinant que même l'aléa complet obéit à des lois précises. Ces deux derniers représentent des résultats clés en théorie des probabilités.

Les premières études sont issues du cas classique où les T_n sont des variables aléatoires de Bernoulli centrées. La loi forte des grands nombres de Borel fut le premier résultat obtenu en 1909 qui affirme que, presque sûrement, $S_n/n \to 0$, une propriété peu apte à décrire le comportement de la suite $(S_n)_{n>0}$ de façon satisfaisante. La loi du logarithme itéré a été introduite dans la théorie de probabilité afin de perfectionner le théorème de Borel et obtenir son taux de convergence exact. Elle peut être considérée comme un résultat intermédiaire entre la loi forte des grands nombres et le théorème central limit. Cependant, les résultats obtenus dans cette direction étaient de plus en plus précis. Citons, en particulier, le travail de Hausdorff [1913] montrant que pour tout $\epsilon > 0$,

$$S_n = O(n^{\epsilon+1/2}) \quad p.s. \text{ quand } n \to +\infty.$$

Puis, Hardy et Littlewood [1914] ont donné une conséquence directe de l'estimation des grands écarts avec un taux de l'ordre de $\sqrt{n \log n}$, un résultat qui a été affiné par Steinhaus [1922] en prouvant qu'on a presque sûrement,

$$\limsup_{n\to\infty} \frac{|S_n|}{\sqrt{2n \log n}} \leq 1.$$

Ce n'est qu'en 1924 que le meilleur taux possible a été énoncé par Khintchine [1924]. Ce fut le premier résultat de la loi du logarithme itéré dans la théorie de la probabilité. Le résultat est donné par le théorème ci-dessous.

Théorème 1. *(Khintchine [1924]). Soit (X_n) une suite de v.a. i.i.d. de loi de Bernoulli centrées. Alors, presque sûrement :*

$$\limsup_{n\to\infty} \frac{|S_n|}{\sqrt{2n\log_2 n}} = +1$$

et

$$\liminf_{n\to\infty} \frac{|S_n|}{\sqrt{2n\log_2 n}} = -1.$$

Ce résultat a été généralisé par Kolmogorov [1929] pour une classe plus large de variables aléatoires (indépendantes mais pas nécessairement identiquement distribuées). Des recherches ultérieures considèrent des suites de variables aléatoires suivant des lois autres que la loi de bernoulli. Parmi eux nous mentionnons la généralisation suivante de la loi du logarithme itéré de Khintchine [1924] aux variables aléatoires i.i.d..

Théorème 2. *(Hartmann et Wintner [1941]). Soit* (T_n) *une suite de variables aléatoires i.i.d., centrées d'écart-type* 1*. Alors on a presque sûrement :*

$$\limsup_{n\to\infty} \frac{S_n}{\sqrt{n\log_2 n}} = \sqrt{2}.$$

Les premiers travaux s'intéressant à la loi du logarithme itéré de la fonction de répartition empirique, définie par (1.2), pour des suites de variables aléatoires indépendantes et identiquement distribuées, sont dus à Smirnov [1939] et Chang [1949] ainsi que la loi fonctionnelle du logarithme itéré de Kiefer [1961]. Dans un modèle de censure à droite, on trouve dans l'article de Földes et Rejtő [1981a] la loi du logarithme itéré de l'estimateur de Kaplan-Meier, donné par la relation (1.3). Messaci et Nemouchi [2011] introduisent ce résultat pour l'estimateur de Patilea-Rolin, donné par la relation (1.10), sous le modèle I de Patilea et Rolin [2006] (voir 1.1.1.IV.).

Les résultats obtenus sous l'hypothèse d'indépendance ont servi de point de départ à de nombreuses recherches sur l'application de la loi du logarithme itéré à des suites de variables et de vecteurs aléatoires dépendantes. Walter Philipp a été un des premiers à mener des recherches dans cette optique. Par une série de travaux à partir de 1967, il a étudié la loi du logarithme itéré pour des sommes partielles de processus faiblement dépendants (voir Philipp [1967, 1969b,c, 1977], pour ne citer que ceux-là). Indépendamment, Iosifescu [1968] et Reznik [1968] ont étudié le même problème, Oodaira et Yoshihara [1971] ont affaibli leurs conditions. Phillip et Berkes [1978] ont inventé une nouvelle technique permettant de prouver des principes d'invariance presque sûr pouvant également être utilisée pour des processus à valeurs vectorielles. La technique d'approximation de Berkes-Philipp a été la base de la plupart des travaux sur les principes d'invariance et de la loi du logarithme itéré dans les décennies suivantes. Pour les sommes partielles de processus fortement mélangeants, les résultats les plus précis actuellement disponibles sont dus à Rio [1995b].

En l'absence de censure, Cai et Roussas [1992] s'intéressent à la loi du logarithme itéré de la fonction de répartition empirique basée sur un processus α-mélangeant. Cai [2001] a considéré le cas des processus fortement mélangeants soumis à un mécanisme de censure à droite, il a donc proposé dans son article une loi du logarithme itéré de l'estimateur de Kaplan-Meier. Dans ce même état d'esprit, notre contribution consiste à généraliser le résultat obtenu par Messaci et Nemouchi [2011] à des processus α-mélangeants, en considérant le même modèle de censure mixte que ces dernières.

2.2 Hypothèses et résultat

Notre travail rentre donc dans le cadre de la censure mixte correspondant au modèle I de Patilea et Rolin [2006] abordé dans la partie IV. de 1.1.1. Autrement dit, nous observons l'échantillon $(Z_i, \delta_i)_{1\leq i\leq n}$ du couple (Z, δ) où $Z = (Y \wedge R) \vee L = \max(T, L)$, avec les variables aléatoires positives et indépendantes Y, R et L représentant, respectivement, la variable d'intérêt, la variable de censure à droite et la variable de censure à gauche. L'indicateur de censure δ est de la forme suivante :

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

Nous proposons d'étudier dans ce chapitre, l'estimateur de S_Y introduit précédemment par $\widetilde{S}_n$ dans (1.10) et rappelé ci dessous, sous la condition du mélange fort. Notons que nous considérons les mêmes notations de la section 1.5 du chapitre précédent avec :

$$\widetilde{S}_n(t) = 1 - \widetilde{F}_n(t) = \prod_{j/Z'_j\leq t} \left(1 - \frac{D_{0j}}{n\mathring{F}_n(Z'_{j-1}) - nH_n(Z'_{j-1})}\right),$$

avec $D_{0j} = \sum_{i=1}^{n} \mathbb{1}_{\{Z_i=Z'_j,\delta_i=0\}}$,

Z'_j $(1 \leq j \leq M)$ les valeurs distinctes de Z_i rangées dans l'ordre croissant,

$\mathring{F}_n$ est l'estimateur de Kaplan-Meier de la fonction de répartition F_L, défini en inversant le temps par

$$\mathring{F}_n(t) = \prod_{j/Z'_j>t} \left(1 - \frac{D_{2j}}{nH_n(Z'_j)}\right),$$

et H_n est la fonction de répartition empirique de Z, donnée par

$$H_n(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{Z_i\leq t\}}.$$

En préalable à l'énoncé de notre résultat, nous allons préciser les hypothèses qui seront imposées afin d'y aboutir.

- ❋ (**C.0**) : Y, R et L, sont des v.a.r. continues,
- ❋ (**C.1**) : (Y_i), (R_i) et (L_i), $1 \leq i \leq n$, sont des suites indépendantes de v.a. positives stationnaires et α-mélangeantes, de coefficient de mélange $\alpha_1(n)$, $\alpha_2(n)$ et $\alpha_3(n)$, respectivement,
- ❋ (**C.2**) : $\max(I_L, I_R) < I_Y$ et $T_Y < T_R$.

Remarque. *Sous l'hypothèse* (***C.1***)*, l'utilisation du Lemme (2) de Cai [2001], nous permet de déduire que les suites* $(Y_i, L_i, R_i)_{i\geq 1}$ *et* $(Z_i)_{i\geq 1}$ *sont des suites stationnaires et* α*-mélangeantes de coefficient de mélange* $\alpha(n) = 4\max(4\max(\alpha_1(n), \alpha_2(n)), \alpha_3(n))$*, avec* $\alpha(n) = O(n^{-\nu})$ *où* $\nu > 4$.

Nous avons le résultat suivant qui établit la loi du logarithme itéré de l'estimateur de Patilea-Rolin $\widetilde{S}_n$, en présence de processus α-mélangeants.

Théorème 3. *Sous les hypothèses* (***C.0***)–(***C.2***)*, nous avons presque sûrement, pour* $\nu > 4$ *:*

$$\sup_{t\in\mathbb{R}} |\widetilde{S}_n(t) - S_Y(t)| = O\left(\sqrt{\frac{\log_2 n}{n}}\right).$$

2.3 Preuve

Avant de développer le détail de notre démonstration, nous allons introduire des quantités qui nous seront indispensables.

Posons, pour $I_Y \leq t < T_Y$, la quantité Λ_Y définie par

$$\Lambda_Y(t) = \int_{I_Y}^{t} d\Lambda_Y(u) = -\log\left(S_Y(t)\right), \tag{2.1}$$

avec $d\Lambda_Y(u)$ la mesure de hasard de la variable d'intérêt Y, donnée par définition, pour tout $I_Y \leq t < T_Y$, par :

$$\Lambda_Y(t) = \int_{I_Y}^{t} \frac{dF_Y(u)}{S_Y(u)} = \int_{0}^{t} \frac{dH^{(0)}(u)}{F_L(u) - H(u)}. \tag{2.2}$$

Il vient de (2.1) et (2.2), en remplaçant les lois inconnues par leurs estimateurs, que :

$$\widetilde{\Lambda}_n(t) = \int_{I_Y} \frac{dH_n^{(0)}(u)}{\mathring{F}_n(u) - H_n(u)}. \tag{2.3}$$

Posons également

$$\ddot{S}_n(t) = \prod_{j/Z'_j \leq t} \left(1 - \frac{D_{0j}}{U_{j-1} - N_{j-1} + 1}\right), \tag{2.4}$$

où $N_j = \sum_{i=1}^{n} \mathbb{1}_{\{Z_i \leq Z'_j\}}$ et $U_{j-1} = n \prod_{j \leq l \leq M} \left\{1 - \frac{D_{2l}}{N_l}\right\}$.

La preuve de notre résultat s'inspire largement de celle de Messaci et Nemouchi [2011]. En procédant donc de la même manière, nous utilisons la décomposition suivante :

$$|\widetilde{S}_n(t) - S_Y(t)| \leq |\widetilde{S}_n(t) - \ddot{S}_n(t)| + |\ddot{S}_n(t) - S_Y(t)|, \tag{2.5}$$

avec

$$\begin{aligned} |\ddot{S}_n(t) - S_Y(t)| &= |\exp(\log \ddot{S}_n(t)) - \exp(-\Lambda_Y(t))| \\ &\leq |\exp(\log \ddot{S}_n(t)) - \exp(-\widetilde{\Lambda}_n(t))| + |\exp(-\widetilde{\Lambda}_n(t)) - \exp(-\Lambda_Y(t))|. \end{aligned}$$

Le développement de Taylor des deux termes de la dernière ligne de cette inégalité entraîne que :

$$\begin{aligned} |\ddot{S}_n(t) - S_Y(t)| \leq |\exp(-\dot{\Lambda}_n(t))(\log \ddot{S}_n(t) + \widetilde{\Lambda}_n(t))| + |S_Y(t)(\Lambda_Y(t) - \widetilde{\Lambda}_n(t))| \\ + \left|\frac{1}{2}\exp(-\ddot{\Lambda}_n(t))(\Lambda_Y(t) - \widetilde{\Lambda}_n(t))^2\right|, \end{aligned} \tag{2.6}$$

où

$$\min\left(-\log \ddot{S}_n(t), \widetilde{\Lambda}_n(t)\right) \leq \dot{\Lambda}_n(t) \leq \max\left(-\log \ddot{S}_n(t), \widetilde{\Lambda}_n(t)\right), \tag{2.7}$$

et

$$\min\left(\Lambda_Y(t), \widetilde{\Lambda}_n(t)\right) \leq \ddot{\Lambda}_n(t) \leq \max(\Lambda_Y(t), \widetilde{\Lambda}_n(t)). \tag{2.8}$$

L'inégalité (2.5) devient alors :

$$\begin{aligned}|\widetilde{S}_n(t) - S_Y(t)| \leq |\widetilde{S}_n(t) - \ddot{S}_n(t)| + |\exp(-\dot{\Lambda}_n(t))(\log \ddot{S}_n(t) + \widetilde{\Lambda}_n(t))| \\ + |S_Y(t)(\Lambda_Y(t) - \widetilde{\Lambda}_n(t))| + \left|\frac{1}{2}\exp\left(-\ddot{\Lambda}_n(t)\right)(\Lambda_Y(t) - \widetilde{\Lambda}_n(t))^2\right|.\end{aligned} \tag{2.9}$$

Pour la suite, nous ferons usage du Théorème 3.2 de Cai et Roussas [1992] qui nous fournit, sous l'hypothèse (**C.1**), les deux résultats suivants :

$$\limsup_{n\to+\infty}\left[\left(\frac{n}{2\log_2 n}\right)^{1/2}\sup_{t\in\mathbb{R}}|H_n(t) - H(t)|\right] = 1 \quad p.s. \tag{2.10}$$

et

$$\limsup_{n\to+\infty}\left[\left(\frac{n}{2\log_2 n}\right)^{1/2}\sup_{t\in\mathbb{R}}|H_n^{(0)}(t) - H^{(0)}(t)|\right] = 1 \quad p.s. \tag{2.11}$$

Aussi, en se servant de la loi du logarithme itéré de Cai [2001], nous avons presque sûrement, sous les hypothèses (**C.0**)–(**C.2**) :

$$\sup_{t\geq I_Y}|\mathring{F}_n(t) - F_L(t)| = O\left(\sqrt{\frac{\log_2 n}{n}}\right). \tag{2.12}$$

Celà dit que pour presque tout w, il existe $n_1(w)$ et un nombre fixe A tel que pour tout $n > n_1$, la combinaison des résultats donnés par (2.10) et (2.12) implique que

$$\sup_{I_Y\leq t}|(\mathring{F}_n(t) - F_L(t)) - (H_n(t) - H(t))| \leq A\sqrt{\frac{\log_2 n}{n}}. \tag{2.13}$$

D'autre part, puisque pour tout $I_Y \leq t \leq T_Y$:

$$F_L(t) - H(t) \geq F_L(I_Y) S_R(T_Y) S_Y(t), \tag{2.14}$$

il vient alors que si $F_L(I_Y) > 0$ et $S_R(T_Y) > 0$, nous avons presque sûrement, pour tout $n > n_1$ et $I_Y < t < t_n$:

$$\mathring{F}_n(t) - H_n(t) \geq \frac{1}{2}(F_L(t) - H(t)), \tag{2.15}$$

avec

$$t_n = S_Y^{-1}\left(\frac{2A}{F_L(I_Y) S_R(T_Y)} \sqrt{\frac{\log\log n}{2n}}\right), \tag{2.16}$$

où A est une constante et $S_Y^{-1}(s) = \sup\{y / S_Y(y) > s\}$.

Pour atteindre le résultat du Théorème 3 nous proposons de traiter chacun des termes de la décomposition (2.9) dans une série de lemmes.

Lemme 2.1. *Pour presque tout w, $\exists\ n_0(w)$ tel que si $n > n_0$ alors : $\forall\ I_Y \leq t \leq t_n$, $k_1 > 0$ et $k_2 \geq 0$ où $k = k_1 + k_2 > 1$, nous avons :*

$$\int_{I_Y}^{t} \frac{dH_n^{(0)}(u)}{(\mathring{F}_n(u) - H_n(u))^{k_1} (F_L(u) - H(u))^{k_2}} = O\left(\frac{n}{\log_2 n}\right)^{\frac{k-1}{2}}.$$

Démonstration. Suite à la relation (2.15), on a $\forall n > n_1$ et pour tout $I_Y \leq t \leq t_n$ et $I_Y \leq u \leq t$:

$$\int_{I_Y}^{t} \frac{dH_n^{(0)}(u)}{(\mathring{F}_n(u) - H_n(u))^{k_1} (F_L(t) - H(t))^{k_2}} \leq \left| \int_{I_Y}^{t} \frac{2^{k_1}}{(F_L(u) - H(u))^{k}}\, dH^{(0)}(u) \right| + \left| \int_{I_Y}^{t} \frac{2^{k_1}}{(F_L(u) - H(u))^{k}}\, d(H_n^{(0)}(u) - H^{(0)}(u)) \right|.$$

En vu des expressions de $H(t)$ et $H^{(0)}(t)$ données dans le Chapitre 1 par les relations (1.6) et (1.7), il vient par intégration :

$$\begin{aligned}
\int_{I_Y}^{t} \frac{dH_n^{(0)}(u)}{(\mathring{F}_n(u) - H_n(u))^{k_1}(F_L(u) - H(u))^{k_2}} &\leq -\int_{I_Y}^{t} 2^{k_1} \frac{F_L(u)S_R(u)d(S_Y(u))}{(F_L(u)S_R(u)S_Y(u))^k} \\
&\quad + \int_{I_Y}^{t} \frac{2^{k_1}}{(F_L(u)S_R(u)S_Y(t))^k} |d(H_n^{(0)}(u) - H^{(0)}(u))| \\
&\leq -\frac{2^{k_1}}{F_L^{k-1}(I_Y)S_R^{k-1}(T_Y)} \int_{I_Y}^{t} \frac{d(S_Y(u))}{S_Y^k(u)} \\
&\quad + \frac{2^{k_1}}{(F_L(I_Y)S_R(T_Y)S_Y(t))^k} \sup_{x\in\mathbb{R}} |H_n^{(0)}(x) - H^{(0)}(x)| \\
&\quad + 2^{k_1} \sup_{x\in\mathbb{R}} |H_n^{(0)}(x) - H^{(0)}(x)| \left(\frac{1}{(F_L(I_Y)S_R(T_Y)S_Y(t))^k} \right) \\
&\leq \frac{2^{k_1}}{(k-1)F_L^{k-1}(I_Y)S_R^{k-1}(T_Y)S_Y^{k-1}(t)} \\
&\quad + \frac{2^{k_1+1} \sup_{x\in\mathbb{R}} |H_n^{(0)}(x) - H^{(0)}(x)|}{(F_L(I_Y)S_R(T_Y)S_Y(t))^k}.
\end{aligned}$$

Ainsi, en combinant le résultat donné par (2.11) avec l'équation (2.16) et le fait que $t \leq t_n$, nous aboutissons à :

$$\begin{aligned}
\int_{I_Y}^{t} \frac{dH_n^{(0)}(u)}{(\mathring{F}_n(u) - H_n(u))^{k_1}(F_L(u) - H(u))^{k_2}} &\leq \frac{2^{k_1}}{(k-1)F_L^{k-1}(I_Y)S_R^{k-1}(T_Y)S_Y^{k-1}(t)} \\
&\quad + \frac{2^{k_1+2}}{(F_L(I_Y)S_R(T_Y)S_Y(t))^k} \sqrt{\frac{\log_2 n}{2n}} \\
&\leq \frac{2^{k_1}}{(k-1)F_L^{k-1}(I_Y)S_R^{k-1}(T_Y)S_Y^{k-1}(t)} \\
&\quad + \frac{2^{k_1+2}}{F_L^{k}(I_Y)S_R^{k}(T_Y)S_Y^{k-1}(t)S_Y(t_n)} \sqrt{\frac{\log_2 n}{2n}} \\
&\leq \frac{2^{k_1}}{F_L^{k-1}(I_Y)S_R^{k-1}(T_Y)} \frac{1}{S_Y^{k-1}(t)} \times \left(\frac{2}{A} + \frac{1}{k-1} \right).
\end{aligned} \tag{2.17}$$

Nous déduisons le résultat du Lemme 2.1 à partir de la dernière ligne de cette majoration, en utilisant la relation (2.16) qui implique que pour $t \geq t_n$:

$$S_Y(t_n) \geq \frac{2A}{F_L(I_Y)S_R(T_Y)}\sqrt{\frac{\log\log n}{2n}}.$$

□

Il nous faut à présent traiter les termes de la décomposition (2.9). On s'intéresse dans le Lemme suivant à son premier terme.

Lemme 2.2. *Nous avons,*

$$\sup_{I_Y \leq t \leq t_n} |\widetilde{S}_n(t) - \ddot{S}_n(t)| = O\left(\sqrt{\frac{1}{n\log_2 n}}\right) \quad p.s.$$

Démonstration. Pour la preuve de ce lemme, nous ferons usage de l'inégalité suivante :

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

avec $|a_i| \leq 1$ et $|b_i| \leq 1$, pour tout $1 \leq i \leq n$.

L'application de cette dernière nous donne :

$$\begin{aligned}
|\widetilde{S}_n(t) - \ddot{S}_n(t)| &= \left|\prod_{j/Z'_j \leq t}\left(1 - \frac{D_{0j}}{U_{j-1} - N_{j-1}}\right) - \prod_{j/Z'_j \leq t}\left(1 - \frac{D_{0j}}{U_{j-1} - N_{j-1} + 1}\right)\right| \\
&\leq \sum_{j/Z'_j \leq t} \frac{D_{0j}}{(U_{j-1} - N_{j-1})^2} \\
&= \sum_{j/Z'_j \leq t} \frac{n\, H_n^{(0)}(Z'_j)}{(n\mathring{F}_n(Z'_j) - nH_n(Z'_j))^2} \\
&= \int_{I_X}^{t} \frac{n\, dH_n^{(0)}(u)}{(n\mathring{F}_n(u) - nH_n(u))^2}.
\end{aligned}$$

On en déduit aisément du Lemme 2.1, que le Lemme 2.2 est vérifié pour $k_1 = 2$ et $k_2 = 0$.
□

Intéressons nous maintenant au second terme de la décomposition (2.9).

Lemme 2.3. *Nous avons,*

$$\sup_{I_Y \leq t \leq t_n} |\log \ddot{S}_n(t) + \widetilde{\Lambda}_n(t)| = O\left(\sqrt{\frac{1}{n \log_2 n}}\right) \quad p.s.$$

Démonstration. Nous partons du fait que

$$\begin{aligned}
\log \ddot{S}_n(t) &= \log \prod_{j/Z'_j \leq t} \left(1 - \frac{D_{0j}}{U_{j-1} - N_{j-1} + 1}\right) \\
&= \sum_{j/Z'_j \leq t} \log \left(1 - \frac{D_{0j}}{U_{j-1} - N_{j-1} + 1}\right) \\
&= \int_{I_Y}^{t} n \log \left(1 - \frac{1}{n\mathring{F}_n(u) - nH_n(u) + 1}\right) dH_n^{(0)}(u).
\end{aligned}$$

Nous obtenons par le développement de la fonction logarithme que :

$$
\begin{aligned}
|\log \ddot{S}_n(t) + \widetilde{\Lambda}_n(t)| &= \left| \int_{I_Y}^t n \log \left(1 - \frac{dH_n^{(0)}(u)}{n\mathring{F}_n(u) - nH_n(u) + 1} \right) + \int_{I_Y}^t \frac{dH_n^{(0)}(u)}{\mathring{F}_n(u) - H_n(u)} \right| \\
&= \left| \int_{I_Y}^t \left(n \left[-\sum_{l=1}^{\infty} \frac{1}{l} (n\mathring{F}_n(u) - nH_n(u) + 1)^{-l} \right] + \frac{1}{\mathring{F}_n(u) - H_n(u)} \right) dH_n^{(0)}(u) \right| \\
&\leq \left| \int_{I_Y}^t \left(\frac{1}{\mathring{F}_n(u) - H_n(u)} - \frac{1}{\frac{1}{n} + \mathring{F}_n(u) - H_n(u)} \right) dH_n^{(0)}(u) \right| \\
&\quad + \frac{1}{2} \left| \int_{I_Y}^t -n \sum_{l=2}^{\infty} \frac{1}{l} (n\mathring{F}_n(u) - nH_n(u) + 1)^{-l} dH_n^{(0)}(u) \right| \\
&= \left| \int_{I_Y}^t \frac{1}{n(\mathring{F}_n(u) - H_n(u))(\frac{1}{n} + \mathring{F}_n(u) - H_n(u))} dH_n^{(0)}(u) \right| \\
&\quad + \frac{1}{2} \left| \int_{I_Y}^t \frac{n(1 + n\mathring{F}_n(u) - nH_n(u))}{(n\mathring{F}_n(u) - nH_n(u) + 1)^2 (n\mathring{F}_n(u) - nH_n(u))} dH_n^{(0)}(u) \right| \\
&\leq 2 \left| \int_{I_Y}^t \frac{dH_n^{(0)}(u)}{n(\mathring{F}_n(u) - H_n(u))^2} \right|.
\end{aligned}
$$

La preuve de ce lemme peut donc être achevée de manière similaire à celle du Lemme 2.2. □

Lemme 2.4. *Nous avons,*

(i)

$$\sup_{I_Y \leq t \leq t_n} |\Lambda_Y(t) - \widetilde{\Lambda}_n(t)| \leq K \quad p.s. \text{ où } K \text{ est une constante.}$$

(ii)

$$\sup_{I_Y \leq t \leq t_n} S_Y(t) |\Lambda_Y(t) - \widetilde{\Lambda}_n(t)| = O\left(\sqrt{\frac{\log_2 n}{n}} \right) \quad p.s.$$

(iii)

$$\frac{1}{2} \exp\left(-\ddot{\Lambda}_n(u)\right) |\Lambda_Y(t) - \widetilde{\Lambda}_n(t)|^2 = O\left(\sqrt{\frac{\log_2 n}{n}} \right) \quad p.s.$$

Démonstration. Remarquons pour commencer que :

$$|\Lambda_Y(t) - \widetilde{\Lambda}_n(t)| = \left| \int_{I_Y}^{t} \frac{dH_n^{(0)}(u)}{\mathring{F}_n(u) - H_n(u)} - \int_{I_Y}^{t} \frac{dH^{(0)}(u)}{F_L(u) - H(u)} \right|$$

$$\leq \sup_{I_Y \leq x} |F_L(x) - H(x) - \mathring{F}_n(x) + H_n(x)| \int_{I_Y}^{t} \frac{dH_n^{(0)}(u)}{(\mathring{F}_n(u) - H_n(u))(F_L(u) - H(u))}$$

$$+ \frac{2 \sup_{x \in \mathbb{R}} (H_n^{(0)}(x) - H^{(0)}(x))}{F_L(I_Y) S_R(T_Y) S_Y(t)}.$$

En tenant compte des résultats (2.11), (2.13) et de la relation (2.17), nous avons pour n suffisamment grand :

$$|\widetilde{\Lambda}_n(t) - \Lambda_Y(t)| \leq 2\sqrt{\frac{\log_2 n}{2n}} \, \frac{2A(2/A+1) + (1+\epsilon)}{F_L(I_Y) S_R(T_Y) S_Y(t)}. \tag{2.18}$$

Par conséquant d'après la relation (2.16), il existe une constante K tel que :

$$\sup_{I_Y \leq t \leq t_n} |\widetilde{\Lambda}_n(t) - \Lambda_Y(t)| \leq K \quad p.s.$$

Ceci achève la preuve de (i).

Nous déduisons aisément de la majoration (2.18) que :

$$\sup_{I_Y \leq t \leq t_n} S_Y(t) |\widetilde{\Lambda}_n(t) - \Lambda_Y(t)| = O\left(\sqrt{\frac{\log_2 n}{n}} \right) \quad p.s \tag{2.19}$$

Nous passons à présent à la preuve de (iii).

En utilisant la relation (2.8), nous pouvons voir que :

$$\exp(-\ddot{\Lambda}_n(t)) \leq S_Y(t) \exp |\widetilde{\Lambda}_n(t) - \Lambda_Y(t)|,$$

d'où

$$\frac{1}{2} \exp\left(-\ddot{\Lambda}_n(u)\right) |\widetilde{\Lambda}_n(t) - \Lambda_Y(t)|^2 \leq \frac{1}{2} S_Y(t) |\widetilde{\Lambda}_n(t) - \Lambda_Y(t)|^2 \exp(|\widetilde{\Lambda}_n(t) - \Lambda_Y(t)|)$$

$$\leq \frac{K}{2} S_Y(t) |\widetilde{\Lambda}_n(t) - \Lambda_Y(t)| \exp(K).$$

Suite au résultat (2.19), on en conclut que :

$$\frac{1}{2}\exp\left(-\ddot{\Lambda}_n(t)\right)|\widetilde{\Lambda}_n(t)-\Lambda_Y(t)|^2 = O\left(\sqrt{\frac{\log_2 n}{n}}\right) \quad p.s. \tag{2.20}$$

□

De plus des Lemmes précédents, nous pouvons déduire à partir de la relation (2.7), la majoration suivante :

$$\exp\left(-\dot{\Lambda}_n(t)\right)|\log\ddot{S}_n(t)+\widetilde{\Lambda}_n(t)| \leq |\log\ddot{S}_n(t)+\widetilde{\Lambda}_n(t)|. \tag{2.21}$$

En vu de la décomposition (2.9), la combinaison des Lemmes 2.2, 2.3 et 2.4 avec la majoration précédente (2.21), nous permet de conclure que pour un n assez grand :

$$\sup_{I_Y\leq t\leq t_n}|\widetilde{S}_n(t)-S_n(t)| = O\left(\sqrt{\frac{\log_2 n}{n}}\right) \quad p.s.$$

Ce dernier résultat associé à la majoration suivante

$$\sup_{t_n<t<\infty}|\widetilde{S}_n(u)-S_X(u)| \leq |S_Y(t_n)|+|\widetilde{S}_n(t_n)-S_Y(t_n)|,$$

implique le résultat recherché et achève la preuve du Théorème 3.

Chapitre 3

Taux de consistance des estimateurs non-paramétriques pour des données censurées et dépendantes

Sommaire

Dans ce chapitre, nous nous intéressons aux estimateurs non-paramétriques de la fonction de distribution et de la fonction de hasard cumulé lorsque les données sont éventuellement censurées et satisfont la condition de mélange fort. Plus précisément, la convergence presque complète uniforme sera établie, sous la condition du α-mélange, pour des données complètes, puis dans le cas d'une seule censure pour finir avec le cas d'une censure mixte. Pour étayer davantage nos résultats théoriques, une étude de simulation est réalisée afin d'illustrer la bonne performance de la méthode étudiée, pour des échantillons de taille relativement petite. Enfin, une application aux processus censurés et α-mélangeants est fournie via l'analyse des teneurs en Chrome des échantillons collectés à partir de deux stations de la rivière de Niagara au Canada. Ce travail a fait l'objet d'une publication dans la revue "*Statistical Methods & Applications*".

À partir d'un échantillon α-mélangeant, un résultat donnant le taux de convergence presque complète uniforme de la fonction de distribution empirique sera donné dans la première section, consacrée au cas de données complètes. Ce résultat sera étendu, dans la section 3.2, au cas de la censure simple à gauche. Ces deux premières sections, nous permettrons de déduire, dans la section 3.3, les taux de convergence presque complète uniforme pour les estimateurs de la fonction de hasard cumulé et de la fonction de distribution. Dans la section 3.4, une étude de simulation apportera un soutien à nos résultats théoriques. A la suite de cette étude nous allons appliquer notre méthode sur des données réelles dans la section 3.5 qui clôturera ce chapitre.

3.1 Données complètes et α-mélangeantes

Dans cette section, $(X_i)_{i\geq 1}$ est supposée être une suite de variables aléatoires réelles, stationnaire et α-mélangeante de coefficent de mélange $a(n)$ arithmétique d'ordre ν. Chaque variable aléatoire réelle X_i est supposée avoir une fonction de distribution inconnue F_X qui est naturellement estimée par la fonction de distribution empirique basée sur le segment $X_1, \ldots, X_n$ et donnée précédemment dans le chapitre 1 par la relation (1.2), tel que :

$$\widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{X_i \leq x\}}.$$

Notre objectif est de dériver le taux de convergence presque complète uniforme de $\widehat{F}_n$, comme indiqué dans le résultat suivant.

Théorème 4. *Si $a(n) = O(n^{-\nu})$ avec $\nu > 4$, nous avons*

$$\sup_{x\in\mathbb{R}} |\widehat{F}_n(x) - F_X(x)| = O_{p.c.}\left(\sqrt{\frac{\log n}{n}}\right).$$

Démonstration. Soit $x \in \mathbb{R}$, remarquons dans un premier temps, que si $x < I_X$ (respectivement, $x \geq T_X$), il s'ensuit que $\widehat{F}_n(x) = F_X(x) = 0$ (respectivement, $\widehat{F}_n(x) = F_X(x) = 1$), le résultat du théorème 4 s'obtient alors immédiatement.

Pour $x \in]I_X, T_X[$, $F_X(x)(1 - F_X(x)) \neq 0$. Introduisons alors la v.a.r. bornée et identiquement distribuée $U_i = 1_{\{X_i \leq x\}} - F_X(x)$, d'où $\text{var}(U_1) = E(U_1^2) = F_X(x)(1 - F_X(x))$.

Notons que pour tout x de $\mathbb{R}$ et tout i de $\{1, \ldots, n\}$, $|U_i| \leq 1$. De plus, nous constatons que la suite (U_i) est arithmétiquement α-mélangeante avec le même taux ν, du fait que U_i et X_i appartiennent à la même tribu engendrée par X_i. On peut alors appliquer l'inégalité de Bernstein pour des données α-mélangeantes (voir Lemme A.4 en annexe A).

Tout d'abord, calculons σ_n^2 où

$$\begin{aligned}
\sigma_n^2 &= \sum_{i=1}^{n}\sum_{j=1}^{n} \text{cov}(U_i, U_j) \\
&= \sum_{i \neq j}\sum \text{cov}(U_i, U_j) + n \ \text{var}(U_1) \\
&= \sum_{i=1}^{n} \sum_{j/\,|i-j|=1} \text{cov}(U_i, U_j) + \sum_{i=1}^{n} \sum_{j/\,2 \leq |i-j| \leq n-1} \text{cov}(U_i, U_j) + nF_X(x)(1 - F_X(x)) \\
&\leq \frac{2(n-1)}{4} + \sum_{i=1}^{n} \sum_{j/\,|i-j| \geq 2} |\text{cov}(U_i, U_j)| + \frac{n}{4}.
\end{aligned}$$

Au vu de la Proposition A.10.i dans Ferraty et Vieu [2006], il existe une constante C_1 tel que :

$$\sigma_n^2 \leq \frac{3n}{4} + C_1 \sum_{i=1}^{n} \sum_{j/\,|i-j| \geq 2} |i - j|^{-\nu}.$$

En utilisant l'inégalité $\sum_{k \geq 2} k^{-\nu} \leq \int_1^\infty x^{-\nu} dx < \infty$, nous obtenons que :

$$\sigma_n^2 \leq C_2 n,$$

où C_2 est une constante.

En injectant cette majoration dans l'inégalité exponentielle donnée par le Lemme A.4 (voir annexe A du document) avec $\epsilon = \epsilon_0 \sqrt{\frac{\log n}{n}}$ et $r = (\log n)^2$, il apparaît que, pour une

constante C_3 :

$$\mathbb{P}\left(\frac{1}{n}\left|\sum_{i=1}^{n} U_i\right| > \epsilon_0\sqrt{\frac{\log n}{n}}\right) \leq C_3\left(\left(1+\frac{\epsilon_0^2\log n}{r}\right)^{\frac{-r}{2}} + nr^{-1}\left(\frac{r\sqrt{n}}{n\epsilon_0\sqrt{\log n}}\right)^{\nu+1}\right)$$
$$\leq C_3\left(\left(1+\frac{\epsilon_0^2}{\log n}\right)^{\frac{-(\log n)^2}{2}} + n^{\frac{1-\nu}{2}}(\log n)^{\frac{3\nu-1}{2}}\right).$$

Sachant que $\log(1+u) = u - \frac{u^2}{2} + o(u^2)$ quand $u \to 0$, nous pouvons écrire, $\forall x \in \mathbb{R}$,

$$\mathbb{P}\left(|\widehat{F}_n(x) - F_X(x)| = \frac{1}{n}\left|\sum_{i=1}^{n} U_i\right| > \epsilon_0\sqrt{\frac{\log n}{n}}\right) \leq C_4\left(n^{\frac{-\epsilon_0^2}{2}} + n^{\frac{1-\nu}{2}}(\log n)^{\frac{3\nu-1}{2}}\right), \tag{3.1}$$

où C_4 est une constante.

Donc pour $\nu > 4$, le choix d'un ϵ_0 adéquat conduit à

$$\sum_{n=1}^{\infty}\mathbb{P}\left(\frac{1}{n}\left|\sum_{i=1}^{n} U_i\right| > \epsilon_0\sqrt{\frac{\log n}{n}}\right) < \infty,$$

ceci permet de conclure que

$$\widehat{F}_n(x) - F_X(x) = O_{p.c.}\left(\sqrt{\frac{\log n}{n}}\right). \tag{3.2}$$

Remarque. *Ce résultat est applicable à* $\widehat{F}_n(x^-) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{\{X_i<x\}}$ *en prenant* $U_i = \mathbb{1}_{\{X_i<x\}} - F_X(x^-)$.

Nous allons maintenant donner la version uniforme de la relation précédente. Pour ce faire, nous procéderons de façon identique à la preuve du théorème de Glivenko-Cantelli (donnée dans Laha et Rohatgi [1979]).

Posons pour tout entier $N \geq 2$ et $k \in \{1, \ldots, N-1\}$,

$$t_{N,k} = F_X^{-1}\left(\frac{k}{N}\right),\ t_{N,0} = -\infty \text{ et } t_{N,N} = +\infty,$$

où pour tout $u \in]0,1[: F_X^{-1}(u) = \inf\{x \in \mathbb{R}/F_X(x) \geq u\}$ est l'inverse généralisée de F_X.

Soit $t \in \mathbb{R}$, $\exists\, k \in \{1, \ldots, N-1\}$ tel que $t \in [t_{N,k}, t_{N,k+1}[$.

Compte tenu des propriétés de F_X^{-1} on a, pour $t < t_{N,k}$,

$$F_X(t) < \frac{k}{N} \Rightarrow \lim_{t \stackrel{<}{\to} t_{N,k}} F_X(t) \leq \frac{k}{N},$$

i.e. $F_X(t_{N,k}^-) \leq \dfrac{k}{N}$. On a donc

$$F_X(t_{N,k}) \geq \frac{k}{N} \geq F_X(t_{N,k}^-),$$

ce qui implique

$$F_X(t_{N,k}) + \frac{1}{N} \geq \frac{k+1}{N} \geq F_X(t_{N,k+1}^-),$$

d'où

$$F_X(t_{N,k+1}^-) - F_X(t_{N,k}) \leq \frac{1}{N}. \tag{3.3}$$

D'autre part, $\widehat{F}_n$ et F_X étant croissantes,

$$\widehat{F}_n(t_{N,k}) - F_X(t_{N,k+1}^-) \leq \widehat{F}_n(t) - F_X(t) \leq \widehat{F}_n(t_{N,k+1}^-) - F_X(t_{N,k}).$$

Cette dernière majoration combinée à celle dans (3.3), nous donne

$$\begin{aligned}
\widehat{F}_n(t_{N,k}) - F_X(t_{N,k}) - \frac{1}{N} &\leq \widehat{F}_n(t) - F_X(t) \\
&\leq \widehat{F}_n(t_{N,k+1}^-) - F_X(t_{N,k+1}^-) + F_X(t_{N,k+1}^-) - F_X(t_{N,k}) \\
&\leq \widehat{F}_n(t_{N,k+1}^-) - F_X(t_{N,k+1}^-) + \frac{1}{N}.
\end{aligned}$$

Ainsi,

$$|\widehat{F}_n(t) - F_X(t)| \leq \max_{1\leq k\leq N-1}\left\{|\widehat{F}_n(t_{N,k}) - F_X(t_{N,k})| + \frac{1}{N}, |\widehat{F}_n(t^-_{N,k+1}) - F_X(t^-_{N,k+1})| + \frac{1}{N}\right\}$$

$$= \max_{1\leq k\leq N-1}\{|\widehat{F}_n(t_{N,k}) - F_X(t_{N,k})|, |\widehat{F}_n(t^-_{N,k+1}) - F_X(t^-_{N,k+1})|\} + \frac{1}{N}$$

$$\leq \max_{1\leq k\leq N-1}|\widehat{F}_n(t_{N,k}) - F_X(t_{N,k})| + \max_{1\leq k\leq N-1}|\widehat{F}_n(t^-_{N,k+1}) - F_X(t^-_{N,k+1})| + \frac{1}{N}. \tag{3.4}$$

En choisissant, pour $n \in \mathbb{N}/\ n \geq 2$,

$$N = \left[\sqrt{\frac{n}{\log n}}\right] + 1, \tag{3.5}$$

où $[x]$ désigne la partie entière de x, la majoration (3.4) devient

$$\sup_{t\in\mathbb{R}}|\widehat{F}_n(t) - F_X(t)| \leq \max_{1\leq k\leq N-1}|\widehat{F}_n(t_{N,k}) - F_X(t_{N,k})|$$
$$+ \max_{1\leq k\leq N-1}|\widehat{F}_n(t^-_{N,k+1}) - F_X(t^-_{N,k+1})| + \sqrt{\frac{\log n}{n}}.$$

Nous déduisons alors que

$$P\left(\sup_{t\in\mathbb{R}}|\widehat{F}_n(t) - F_X(t)| > (1+\epsilon_0)\sqrt{\frac{\log n}{n}}\right)$$

$$\leq \sum_{k=1}^{N-1} P\left(|\widehat{F}_n(t_{N,k}) - F_X(t_{N,k})| > \frac{\epsilon_0}{2}\sqrt{\frac{\log n}{n}}\right)$$

$$+ \sum_{k=1}^{N-1} P\left(|\widehat{F}_n(t^-_{N,k+1}) - F_X(t^-_{N,k+1})| > \frac{\epsilon_0}{2}\sqrt{\frac{\log n}{n}}\right).$$

L'association de ce résultat aux relations (3.1) et (3.5) et la remarque 3.1, implique que, pour $\nu > 4$,

$$\sum_{n\geq 2} P\left(\sup_{t\in\mathbb{R}}|\widehat{F}_n(t) - F_X(t)| > (1+\epsilon_0)\sqrt{\frac{\log n}{n}}\right) < \infty,$$

ce qui conclut la démonstration du Théorème 4. □

3.2 Modèle de censure à gauche α-mélangeant

Soient $(L_i)_{1\leq i\leq n}$ et $(T_i)_{1\leq i\leq n}$ deux suites indépendantes de répliques aléatoires des v.a. positives L et T respectivement, que nous considérons comme le temps de survie et le temps de censure à gauche, respectivement. Elles sont supposées être des suites stationnaires et arithmétiquement α-mélangeantes avec des coefficients de mélange respectifs $\alpha_1(n)$ et $\alpha_2(n)$. Dans le contexte de la censure à gauche, on ne peut observer que les paires $(Z_i,\delta_i)_{1\leq i\leq n}$ où $Z_i=\max(L_i,T_i)$ et $\delta_i=\mathbb{1}_{\{L_i\geq T_i\}}$. Nous pouvons montrer, en appliquant le Lemme 2 de Cai [2001], que $(L_i,T_i)_{i\geq 1}$ et $(Z_i)_{i\geq 1}$ sont arithmétiquement α-mélangeantes de coefficient $4\max(\alpha_1(n),\alpha_2(n))$.

Notons H la fonction de distribution de $Z=\max(L,T)$ et $N^{(1)}$ sa fonction de sous-distribution définie par :

$$N^{(1)}(t)=\mathbb{P}(Z\leq t,\delta=1)=\int_0^t F_T(u)dF_L(u),$$

où $\delta=\mathbb{1}_{\{L\geq T\}}$.

Les versions empiriques de H et $N^{(1)}$ sont données respetivement par :

$$H_n(t)=\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{Z_i\leq t\}} \text{ et } N_n^{(1)}(t)=\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{Z_i\leq t,\delta_i=1\}}.$$

La relation $F_L(t)=\pi_{]t,\infty[}(1-d\Gamma)$ où $d\Gamma=\dfrac{dF_L}{F_L}=\dfrac{dN^{(1)}}{H}$, suggère d'estimer F_L par :

$$\mathring{F}_n(t)=\pi_{]t,\infty[}(1-d\Gamma_n)=\prod_{j/Z'_j>t}\left(1-\Gamma_n(Z'_j)\right), \tag{3.6}$$

où $\{Z'_j,\ 1\leq j\leq m\}$ sont les valeurs distinctes de $\{Z_i,\ 1\leq i\leq n\}$ rangées dans l'ordre croissant et $d\Gamma_n=\dfrac{dN_n^{(1)}}{H_n}$.

Remarque. $\pi_{]t,\infty[}(1-d\Gamma):=\prod_{x>t}[1-\Gamma(\{x\})]\exp\left[-\Gamma^c(]t,\infty[)\right]$, *où* Γ^c *est la prtie continue de* Γ. *Rappelons que* π *est l'integrale produit (voir Gill et Johansen [1990]).*

Dans cette section nous allons étendre le résultat du Théorème 4 à $\mathring{F}_n$. Seulement, de par le mécanisme de censure, la convergence n'est plus sur tout $\mathbb{R}$, comme le confirme le Théorème ci-dessous.

Théorème 5. *Si* $\max(\alpha_1(n),\alpha_2(n))=O(n^{-\nu})$ *avec* $\nu>4$, *nous avons, pour tout* $\theta>I_Z$,

$$\sup_{t\geq\theta}|\mathring{F}_n(t)-F_L(t)|=O_{p.c.}\left(\sqrt{\frac{\log n}{n}}\right).$$

Remarque. *En inversant le temps, nous pouvons facilement déduire le même taux de convergence pour l'estimateur de Kaplan-Meier.*

Démonstration. Selon l'équation de Duhamel (voir Théorème 6, page 1519 de Gill et Johansen [1990]) et comme l'intégrale produit est une fonction multiplicative, nous pouvons écrire :

$$\begin{aligned} F_L(t) - \mathring{F}_n(t) &= \pi_{]t,\infty[}(1 - d\Gamma) - \pi_{]t,\infty[}(1 - d\Gamma_n) \\ &= F_L(t) \int_{]t,+\infty[} \frac{\mathring{F}_n(u)}{F_L(u^-)} \, d(\Gamma_n(u) - \Gamma(u)) \\ &= F_L(t) \int_{]t,+\infty[} \frac{\mathring{F}_n(u)}{F_L(u)(1 - \Delta\Gamma(u))} \, d(\Gamma_n(u) - \Gamma(u)). \end{aligned}$$

Celà implique que :

$$|\mathring{F}_n(t) - F_L(t)| \leq \left| \int_{]t,+\infty[} \frac{\mathring{F}_n(u)}{F_L(u)} \, d\mathring{M}_n(u) \right|,$$

où $\mathring{M}_n(t) = \int_{]t,+\infty[} \dfrac{d(\Gamma_n(u) - \Gamma(u))}{1 - \Delta\Gamma(u)}$.

En intégrant par parties, nous pouvons voir que :

$$\begin{aligned} |\mathring{F}_n(t) - F_L(t)| &\leq \left| \frac{\mathring{F}_n(t)}{F_L(t)} \, \mathring{M}_n(t) \right| + \left| \int_{]t,+\infty[} \mathring{M}_n(u^-) \, d\left(\frac{\mathring{F}_n(u)}{F_L(u)} \right) \right| \\ &\leq \frac{1}{F_L(\theta)} \sup_{u \geq \theta} |\mathring{M}_n(u)| + \left| \int_{]t,+\infty[} \mathring{M}_n(u^-) \mathring{F}_n(u^-) \, d\left(\frac{1}{F_L(u)} \right) \right| \\ &\quad + \left| \int_{]t,+\infty[} \frac{\mathring{M}_n(u^-)}{F_L(u)} \, d\mathring{F}_n(u) \right| \\ &\leq \frac{1}{F_L(\theta)} \sup_{u \geq \theta} |\mathring{M}_n(u)| + \sup_{u \geq \theta} |\mathring{M}_n(u)| \left(\frac{1}{F_L(t)} - 1 \right) \\ &\quad + \frac{1}{F_L(\theta)} \sup_{u \geq \theta} |\mathring{M}_n(u)| |\mathring{F}_n(t) - 1| \\ &\leq \frac{3 - F_L(\theta)}{F_L(\theta)} \sup_{u \geq \theta} |\mathring{M}_n(u)|. \end{aligned}$$

Par conséquent, il suffit de traiter $\mathring{M}_n$. De par sa définition, nous avons :

$$\begin{aligned}
|\mathring{M}_n(u)| &= \left| \int_{]u,+\infty[} \frac{F_L(x)}{F_L(x^-)}\, d(\Gamma_n(x) - \Gamma(x)) \right| \\
&= \left| \int_{]u,+\infty[} \left(1 + \frac{\Delta F_L(x)}{F_L(x^-)}\right)\, d(\Gamma_n(x) - \Gamma(x)) \right| \\
&\leq \left| \int_{]u,+\infty[} d(\Gamma_n(x) - \Gamma(x)) \right| + \left| \int_{]u,+\infty[} \frac{\Delta F_L(x)}{F_L(x^-)}\, d(\Gamma_n(x) - \Gamma(x)) \right|,
\end{aligned}$$

d'où

$$\begin{aligned}
|\mathring{M}_n(u)| &\leq \sup_{u\geq\theta} |\Gamma_n(u) - \Gamma(u)| + \sum_{\substack{x>u \\ \Delta_{F_L}(x)>0}} \left|\frac{\Delta F_L(x)}{F_L(x^-)}\right| |\Delta\Gamma_n(x) - \Delta\Gamma(x)| \\
&\leq \sup_{u\geq\theta} |\Gamma_n(u) - \Gamma(u)| + \frac{1}{F_L(\theta^-)} \sup_{u\geq\theta} |\Delta\Gamma_n(x) - \Delta\Gamma(x)| \sum_{\substack{x>u \\ \Delta_{F_L}(x)>0}} |\Delta F_L(x)| \\
&\leq \left(\frac{F_L(\theta^-) + 2}{F_L(\theta^-)}\right) \sup_{u\geq\theta} |\Gamma_n(u) - \Gamma(u)|.
\end{aligned}$$

On en déduit que

$$\sup_{t\geq\theta} |\mathring{F}_n(t) - F_L(t)| \leq \frac{(3 - F_L(\theta))(2 + F_L(\theta^-))}{F_L(\theta)F_L(\theta^-)} \sup_{t\geq\theta} |\Gamma_n(t) - \Gamma(t)|.$$

Ainsi, notre preuve se réduit au traitement du terme $\sup_{t\geq\theta}|\Gamma_n(t) - \Gamma(t)|$.

À cet effet, en intégrant par partie nous obtenons :

$$
\begin{aligned}
|\Gamma_n(t) - \Gamma(t)| &= \left| \int_{]t,+\infty[} \frac{dN_n^{(1)}(u)}{H_n(u)} - \int_{]t,+\infty[} \frac{dN^{(1)}(u)}{H(u)} \right| \\
&= \left| \int_{]t,+\infty[} \frac{dN_n^{(1)}(u)}{H_n(u)} - \int_{]t,+\infty[} \frac{dN^{(1)}(u)}{H(u)} + \int_{]t,+\infty[} \frac{dN_n^{(1)}(u)}{H(u)} - \int_{]t,+\infty[} \frac{dN_n^{(1)}(u)}{H(u)} \right| \\
&\leq \left| \int_{]t,+\infty[} \left(\frac{1}{H_n(u)} - \frac{1}{H(u)} \right) dN_n^{(1)}(u) \right| + \left| \int_{]t,+\infty[} \frac{1}{H(u)}\, d(N_n^{(1)}(u) - N^{(1)}(u)) \right| \\
&\leq \frac{1}{H_n(\theta)H(\theta)} \sup_{u\geq\theta} |H(u) - H_n(u)| (N_n^{(1)}(+\infty) - N_n^{(1)}(t)) \\
&\quad + |N_n^{(1)}(+\infty) - N^{(1)}(+\infty)| + \left| \frac{N_n^{(1)}(t) - N^{(1)}(t)}{H(t)} \right| \\
&\quad + \left| \int_{]t,+\infty[} (N_n^{(1)}(u^-) - N^{(1)}(u^-)) d\left(\frac{1}{H(u)} \right) \right| \\
&\leq \frac{1}{H_n(\theta)H(\theta)} \sup_{u\geq\theta} |H_n(u) - H(u)| + \frac{2}{H(\theta)} \sup_{u\geq\theta} |N_n^{(1)}(u) - N^{(1)}(u)|.
\end{aligned}
$$

Par ailleurs, au vu du Théorème 4, nous avons :

$$
\sup_{u\geq\theta} |H_n(u) - H(u)| = O_{p.c.}\left(\sqrt{\frac{\log n}{n}} \right), \tag{3.7}
$$

et

$$
H_n(\theta) \to H(\theta) \ p.c. \ \text{ avec } \ H(\theta) \neq 0. \tag{3.8}
$$

Donc, d'après la Proposition A6 de Ferraty et Vieu [2006] (rappelée dans l'annexe A, Proposition A.2), $(\sup_{u\geq\theta}|H_n(u)-H(u)|)/H_n(\theta)$ a le même taux de convergence que $\sup_{u\geq\theta}|H_n(u)-H(u)|$.

De plus, en remarquant que $V_i = 1_{\{Z_i\leq t,\delta_i=1\}} - N^{(1)}(t)$ est stationnaire et arithmétiquement α-mélangeante avec un coefficient $4\max(\alpha_1(n), \alpha_2(n))$, nous pouvons conclure de la preuve du Théorème 4, en remplaçant U_i par V_i, que :

$$
\sup_{u\geq\theta} |N_n^{(1)}(u) - N^{(1)}(u)| = O_{p.c.}\left(\sqrt{\frac{\log n}{n}} \right). \tag{3.9}
$$

En combinant les résultats issus de (3.7) et (3.9), nous aboutissons au résultat du Théorème 4. □

3.3 Modèle de censure mixte α-mélangeant

Dans cette section, nous disposons des variables aléatoires positives indépendantes Y, R et L, représentant respectivement la variable d'intérêt, la variable de censure à droite et la variable de censure à gauche. Plus précisément, notre observation est limitée à l'échantillon $(Z_i, \delta_i)_{1\leq i\leq n}$ de variables identiquement distribuées comme le couple $(Z, \delta) := (\max(\min(Y, R), L), \delta)$ où

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

Tout au long de cette section, nous supposons également que les suites (Y_i), (R_i) et (L_i) sont stationnaires et arithmétiquement α-mélangeantes avec des coefficients de mélange $\alpha_1(n)$, $\alpha_2(n)$ et $\alpha_3(n)$, respectivement. Au vu du Lemme 2 dans Cai [2001], les suites (Y_i, R_i, L_i) et (Z_i) sont également arithmétiquement α-mélangeantes avec le coefficient de mélange $\alpha(n) = 4\max(4\max(\alpha_1(n), \alpha_2(n)), \alpha_3(n))$.

Afin de ne pas alourdir inutilement les notations, nous gardons ceux précédemment utilisées dans la Section 1.5 du Chapitre 1, à savoir la fonction de distribution de la v.a. Z, $H(t) = F_L(t)F_T(t)$ où $T = \min(Y, R)$ et sa sous-distribution, pour les observations non censurées, $H^{(0)}(t) = \mathbb{P}(Z \leq t, \delta = 0)$ ainsi que leurs versions empirique données respectivement par $H_n(t) = 1/n \sum_{i=1}^n \mathbb{1}_{\{Z_i\leq t\}}$ et $H_n^{(0)}(t) = 1/n \sum_{i=1}^n \mathbb{1}_{\{Z_i\leq t, \delta_i=0\}}$.

Nous déduisons d'après le Théorème 4, que pour $\nu > 4$:

$$\sup_{t\in\mathbb{R}} |H_n^{(0)}(t) - H^{(0)}(t)| = O_{p.c.}\left(\sqrt{\frac{\log n}{n}}\right), \tag{3.10}$$

et

$$\sup_{t\in\mathbb{R}} |H_n(t) - H(t)| = O_{p.c.}\left(\sqrt{\frac{\log n}{n}}\right). \tag{3.11}$$

Nous reprenons également l'estimateur de Patilea-Rolin, noté $\widetilde{S}_n$, de la fonction de survie S_Y,

$$\widetilde{S}_n(t) = 1 - \widetilde{F}_n(t) = \prod_{j/Z'_j\leq t} \left(1 - \frac{D_{0j}}{n\mathring{F}_n(Z'_{j-1}) - nH_n(Z'_{j-1})}\right),$$

où Z'_j $(1 \leq j \leq M)$ sont les valeurs distinctes de Z_i rangées dans l'ordre croissant, $D_{0j} = \sum_{i=1}^n \mathbb{1}_{\{Z_i=Z'_j, \delta_i=0\}}$ et $\mathring{F}_n$ est l'estimateur de Kaplan-Meier de F_L, défini précédemment par :

$$\mathring{F}_n(t) = \prod_{j/Z'_j>t} \left(1 - \frac{D_{2j}}{nH_n(Z'_j)}\right).$$

Rappelons également que la fonction de hasard cumulé de Y, Λ_Y, est définie dans le Chapitre 2 pour tout $I_Y \leq t < T_Y$ par la relation (2.1) ainsi que son estimateur $\widetilde{\Lambda}_n$ défini dans (2.3) et rappelé ci-dessous

$$\widetilde{\Lambda}_n(t) = \int_{I_Y}^{t} \frac{dH_n^{(0)}(u)}{\mathring{F}_n(u) - H_n(u)}.$$

Les taux de convergence presque complète uniforme de $\widetilde{\Lambda}_n$ et $\widetilde{F}_n$ sont énoncés dans le Théorème suivant.

Théorème 6. *Si* $\max(I_L, I_R) < I_Y$ *et* $\max(\alpha_1(n), \alpha_2(n), \alpha_3(n)) = O(n^{-\nu})$ *avec* $\nu > 4$, *nous avons, pour tout* $\theta < \min(T_Y, T_R)$,

$$\textbf{(i)} \ \sup_{t\leq\theta} \left|\widetilde{\Lambda}_n(t) - \Lambda_Y(t)\right| = O_{p.c.}\left(\sqrt{\frac{\log n}{n}}\right),$$

$$\textbf{(ii)} \ \sup_{t\leq\theta} \left|\widetilde{F}_n(t) - F_Y(t)\right| = O_{p.c.}\left(\sqrt{\frac{\log n}{n}}\right).$$

Remarque. *Notons que ces taux de convergence sont les mêmes que ceux obtenus par Kitouni* et al. *[2015], dans le cas de données indépendantes. De plus, remarquons que par rapport à la loi du logarithme itéré du chapitre précédent (qui donne un taux presque sûr), la convergence dans ce Théorème et un peu plus forte mais la vitesse de convergence est un peu détériorée. L'avantage ici est qu'on se dispense de l'hypothèse de continuité des variables latentes.*

Démonstration. **(i)** De par les définitions de Λ_Y et $\widetilde{\Lambda}_n$ données dans (2.1) et (2.3), nous avons :

$$\begin{aligned}
|\widetilde{\Lambda}_n(t) - \Lambda_Y(t)| \leq & \int_{I_Y}^{t} \left|\frac{1}{\mathring{F}_n(u^-) - H_n(u^-)} - \frac{1}{F_L(u^-) - H(u^-)}\right| dH_n^{(0)}(u) \\
& + \left|\int_{I_Y}^{t} \frac{1}{F_L(u^-) - H(u^-)}\, d\left(H_n^{(0)}(u) - H^{(0)}(u)\right)\right| \\
=: & \ B_{n,1}(t) + B_{n,2}(t). \qquad (3.12)
\end{aligned}$$

Remarquons dans un premier temps qu'au vu de (2.14),

$$B_{n,1}(t) = \int_{I_Y}^{t} \left| \frac{1}{\mathring{F}_n(u^-) - H_n(u^-)} - \frac{1}{F_L(u^-) - H(u^-)} \right| dH_n^{(0)}(u)$$

$$\leq \frac{\sup_{I_Y \leq u \leq \theta} \left| F_L(u^-) - \mathring{F}_n(u^-) + H_n(u^-) - H(u^-) \right|}{F_L(I_Y^-) S_R(\theta) S_Y(\theta)}$$

$$\times \int_{I_Y}^{t} \frac{dH_n^{(0)}}{\mathring{F}_n(u^-) - H_n(u^-)}$$

$$\leq 2 \, \frac{\sup_{I_Y \leq u \leq \theta} \left| F_L(u^-) - \mathring{F}_n(u^-) + H_n(u^-) - H(u^-) \right|}{F_L(I_Y^-) S_R(\theta) S_Y(\theta) \inf_{I_Y \leq u \leq \theta} \left| \mathring{F}_n(u^-) - H_n(u^-) \right|}.$$

Pour un $\varepsilon_0 \in]0, F_L(I_Y^-) S_R(\theta) S_Y(\theta)[$ et un $\varepsilon = F_L(I_Y^-) S_R(\theta) S_Y(\theta) - \varepsilon_0$, supposons que $\inf_{I_Y \leq u \leq \theta} \left| \mathring{F}_n(u^-) - H_n(u^-) \right| < \varepsilon_0$, il existe donc un $t_0 \in [I_Y, \theta]$ tel que $\mathring{F}_n(t_0^-) - H_n(t_0^-) \leq \varepsilon_0$, il vient alors que

$$F_L(t_0^-) - H(t_0^-) - \mathring{F}_n(t_0^-) + H_n(t_0^-)$$

$$= S_Y(t_0^-) S_R(t_0^-) F_L(t_0^-) - \mathring{F}_n(t_0^-) + H_n(t_0^-)$$

$$\geq S_Y(\theta) S_R(\theta) F_L(I_Y^-) - \varepsilon_0 = \varepsilon,$$

d'où

$$\sup_{I_Y \leq u \leq \theta} |F_L(u^-) - \mathring{F}_n(u^-) + H_n(u^-) - H(u^-)| > \varepsilon,$$

ce qui implique que

$$\mathbb{P}\left(\inf_{I_Y \leq u \leq \theta} |\mathring{F}_n(u^-) - H_n(u^-)| < \varepsilon_0 \right)$$

$$\leq \mathbb{P}\left(\sup_{I_Y \leq u \leq \theta} |F_L(u^-) - \mathring{F}_n(u^-) + H_n(u^-) - H(u^-)| > \varepsilon \right),$$

puisque $\max(I_L, I_R) < I_Y$, le terme de droite de cette dernière inégalité est le terme général d'une série convergente.

Ceci joint aux résultats du Théorème 4 et de (3.11), nous permet de déduire que

$$\sup_{I_Y \leq t \leq \theta} B_{n,1}(t) = O_{p.c.}\left(\sqrt{\frac{\log n}{n}}\right). \tag{3.13}$$

De plus, en intégrant par partie et en sus de (2.14), nous obtenons :

$$\begin{aligned}
B_{n,2}(t) &= \left| \int_{I_Y}^{t} \frac{1}{F_L(u^-) - H(u^-)} \, d\left(H_n^{(0)}(u) - H^{(0)}(u)\right) \right| \\
&\leq \left| \frac{H_n^{(0)}(t) - H^{(0)}(t)}{F_L(t) - H(t)} \right| + \left| \frac{H_n^{(0)}(I_Y) - H^{(0)}(I_Y)}{F_L(I_Y) - H(I_Y)} \right| \\
&\quad + \left| \int_{I_Y}^{t} (H_n^{(0)}(u) - H^{(0)}(u) \, d\left(\frac{1}{F_L(u) - H(u)}\right) \right| \\
&\leq \frac{2}{F_L(I_Y) S_R(\theta) S_Y(\theta)} \sup_{I_Y \leq u \leq \theta} \left| H_n^{(0)}(u) - H^{(0)}(u) \right| \\
&\quad + \left| \int_{I_Y}^{t} \frac{H_n^{(0)}(u) - H^{(0)}(u)}{F_L(u^-)} \, d\left(\frac{1}{S_R(u) S_Y(u)}\right) \right| \\
&\quad + \left| \int_{I_Y}^{t} \frac{H_n^{(0)}(u) - H^{(0)}(u)}{S_R(u) S_Y(u)} \, d\left(\frac{1}{F_L(u)}\right) \right| \\
&\leq C \sup_{I_Y \leq u \leq \theta} \left| H_n^{(0)}(u) - H^{(0)}(u) \right|,
\end{aligned}$$

où C est une constante.
Donc, au vu de (3.10),

$$\sup_{I_Y \leq t \leq \theta} B_{n,2}(t) = O_{p.c.}\left(\sqrt{\frac{\log n}{n}}\right). \tag{3.14}$$

L'association de ce résultat à (3.12) et (3.13), achève la démonstration de la partie **(i)**.

(ii) Nous procédons de manière semblable à celle utilisée pour prouver le théorème 5. Utilisons donc l'équation de Duhamel (voir Gill et Johansen [1990]) qui nous donne pour tout $t \leq \theta < \min(T_R, T_Y)$

$$\left| \widetilde{F}_n(t) - F_Y(t) \right| = (1 - F_Y(t)) \left| \int_{I_Y}^{t} \frac{1 - \widetilde{F}_n(u^-)}{1 - F_Y(u)} \, d(\widetilde{\Lambda}_n - \Lambda_Y)(u) \right|.$$

Posons $\widetilde{M}_n(t) = \int_{I_Y}^t \frac{1-F_Y(u^-)}{1-F_Y(u)}\, d(\widetilde{\Lambda}_n - \Lambda_Y)(u)$.

L'intégration par partie entraîne :

$$\begin{aligned}
\left|\widetilde{F}_n(t) - F_Y(t)\right| &\leq \left| \int_{I_Y}^t \frac{1-\widetilde{F}_n(u^-)}{1-F_Y(u^-)}\, d\widetilde{M}_n(u) \right| \\
&= \left| \frac{1-\widetilde{F}_n(t)}{1-F_Y(t)}\, \widetilde{M}_n(t) - \int_{I_Y}^t \widetilde{M}_n(u)\, d\left(\frac{1-\widetilde{F}_n(u)}{1-F_Y(u)}\right) \right| \\
&\leq \frac{1}{1-F_Y(\theta)} \left|\widetilde{M}_n(t)\right| + \left| \int_{I_Y}^t \widetilde{M}_n(u) \frac{d\widetilde{F}_n(u)}{1-F_Y(u)} \right| \\
&\quad + \left| \int_{I_Y}^t \widetilde{M}_n(u) \left(1-\widetilde{F}_n(u^-)\right) \frac{dF_Y(u)}{(1-F_Y(u))\,(1-F_Y(u^-))} \right| \\
&\leq \frac{1}{1-F_Y(\theta)} \left|\widetilde{M}_n(t)\right| + \frac{1}{1-F_Y(\theta)} \sup_{I_Y \leq u \leq \theta} \left|\widetilde{M}_n(u)\right| \widetilde{F}_n(t) \\
&\quad + \frac{1}{(1-F_Y(\theta))^2} \sup_{I_Y \leq u \leq \theta} \left|\widetilde{M}_n(u)\right| F_Y(t) \\
&\leq \frac{2\,(1-F_Y(\theta)) + 1}{(1-F_Y(\theta))^2} \sup_{I_Y \leq u \leq \theta} \left|\widetilde{M}_n(u)\right|.
\end{aligned}$$

Il suffit donc de traiter le terme $\sup_{I_Y \leq u \leq \theta} \left|\widetilde{M}_n(u)\right|$.

$$\begin{aligned}
\left|\widetilde{M}_n(t)\right| &\leq \left|\widetilde{\Lambda}_n(t) - \Lambda_Y(t)\right| + \left| \sum_{\substack{u \leq t \\ \Delta F_Y(u) > 0}} \frac{\Delta F_Y(u)}{1-F_Y(u)} \left(\Delta\widetilde{\Lambda}_n(u) - \Delta\Lambda_Y(u)\right) \right| \\
&\leq \left|\widetilde{\Lambda}_n(t) - \Lambda_Y(t)\right| + 2 \sup_{I_Y \leq u \leq \theta} \left|\widetilde{\Lambda}_n(u) - \Lambda_Y(u)\right| \frac{1}{1-F_Y(\theta)} F_Y(t).
\end{aligned}$$

Il vient donc que

$$\sup_{I_Y \leq t \leq \theta} \left|\widetilde{M}_n(t)\right| \leq \frac{3-F_Y(\theta)}{1-F_Y(\theta)} \sup_{I_Y \leq t \leq \theta} \left|\widetilde{\Lambda}_n(t) - \Lambda_Y(t)\right|.$$

Le résultat découle alors immédiatement, au vu de **(i)**. □

3.4 Étude de simulation

Dans cette partie, afin d'illustrer ce qui précède, nous examinons la performance de l'estimateur $\widetilde{F}_n$ (donné par (1.10)), qui généralise les deux estimateurs $\hat{F}_n$ et $\mathring{F}_n$, au travers des simulations. Nous générons les données comme suit.

Soient (ϵ_i), (ϵ'_i) et (ϵ''_i) des bruits blancs indépendants gaussiens standard $\mathcal{N}(0,1)$. Nous considérons les processus suivants.

$$Y_i = |Y'_i| + 0.01,\ R_i = b|R'_i| \text{ et } L_i = c|L'_i|,$$

où

$$Y'_i = aY'_{i-1} + \epsilon_i\sqrt{1-a^2},\ \ Y'_0 \sim \mathcal{N}(0,1) \text{ et il est indépendant de } (\epsilon_i),$$

$$R'_i = aR'_{i-1} + \epsilon'_i\sqrt{1-a^2},\ \ R'_0 \sim \mathcal{N}(0,1) \text{ et il est indépendant de } (\epsilon'_i),$$

$$L'_i = aL'_{i-1} + \epsilon''_i\sqrt{1-a^2},\ \ L'_0 \sim \mathcal{N}(0,1) \text{ et il est indépendant de } (\epsilon''_i),$$

sauf pour le cas non censuré dans la Figure 3.3, pour lequel nous choisissons pour tout i, $R_i = 100$ et $L_i = 0$. Le paramètre a $(0 < a < 1)$ est sélectionné afin de contrôler la dépendance entre les variables de la population considérée, en le faisant varier, avec une taille d'échantillon et un pourcentage de censure fixés, nous étudions la performance de notre estimateur. Le Tableau 3.2 indique les valeurs des paramètres b et c choisies pour assurer différents pourcentages de censure p, afin de mieux se rendre compte de l'influence de la censure sur la qualité de l'estimateur.

Comme les (Y'_i), (R'_i) et (L'_i) sont des processus AR(1), il est clair que (Y_i), (R_i) et (L_i) sont positives, stationnaires et α-mélangeantes. En outre, le processus de génération de variables est effectué afin d'assurer leur indépendance. Notons que la condition $\max(I_L, I_R) < I_Y$ est évidemment satisfaite.

Dans un premier temps, nous examinons graphiquement la qualité de l'estimation et comment est ce qu'elle est influencée par l'index de dépendance (a), la taille de l'échantillon (n) et le pourcentage de censure (p). Ainsi, dans chaque graphe, nous fixons deux paramètres et nous varions le troisième.

D'une part, nous pouvons voir aux Figures 3.1, 3.2 et 3.3 que tous les graphes de l'estimateur présentent une forme acceptable par rapport aux vraies courbes correspondantes. D'autre part, nous remarquons sans surprise que le comportement de l'estimateur $\widetilde{F}_n$ est meilleur pour un petit a (dépendance faible), un échantillon de grande taille et un faible taux de censure. Ceci est conforté par le Tableau 3.1, où la mesure des distances entre les fonctions de distribution cumulatives théoriques et empiriques est calculée en fonction de a, p et n. Pour chaque valeur n, p et a, nous répliquons 100 fois et prenons la médiane sur $x \in [0,3]$ des erreurs quadratiques moyennes de l'estimateur $\widetilde{F}_n(x)$ par rapport à la valeur

réelle $F_Y(x)$. Plus précisément, les valeurs du Tableau 3.1 sont calculées par la formule suivante :

$$\underset{x\in[0,3]}{\text{med}}\left[\frac{1}{100}\sum_{i=1}^{100}(\widetilde{F}_n(x)-F_Y(x))^2\right]\times 10^2$$

TABLEAU 3.1 – La distance entre $\widetilde{F}_n$ et F_Y

a \ n	70			100			200		
	$p\simeq 0\%$	$\simeq 20\%$	$\simeq 40\%$	$p\simeq 0\%$	$\simeq 20\%$	$\simeq 40\%$	$p\simeq 0\%$	$\simeq 20\%$	$\simeq 40\%$
0.1	0.062	0.127	0.138	0.043	0.086	0.101	0.025	0.044	0.049
0.5	0.090	0.139	0.155	0.061	0.095	0.116	0.035	0.052	0.055
0.9	0.344	0.458	0.517	0.301	0.339	0.358	0.129	0.155	0.183

TABLEAU 3.2 – Les pourcentages de censure

p	$\simeq 0\%$	$\simeq 20\%$	$\simeq 30\%$	$\simeq 40\%$
b	1	3.5	3.2	3
c	1	0.1	0.2	0.3

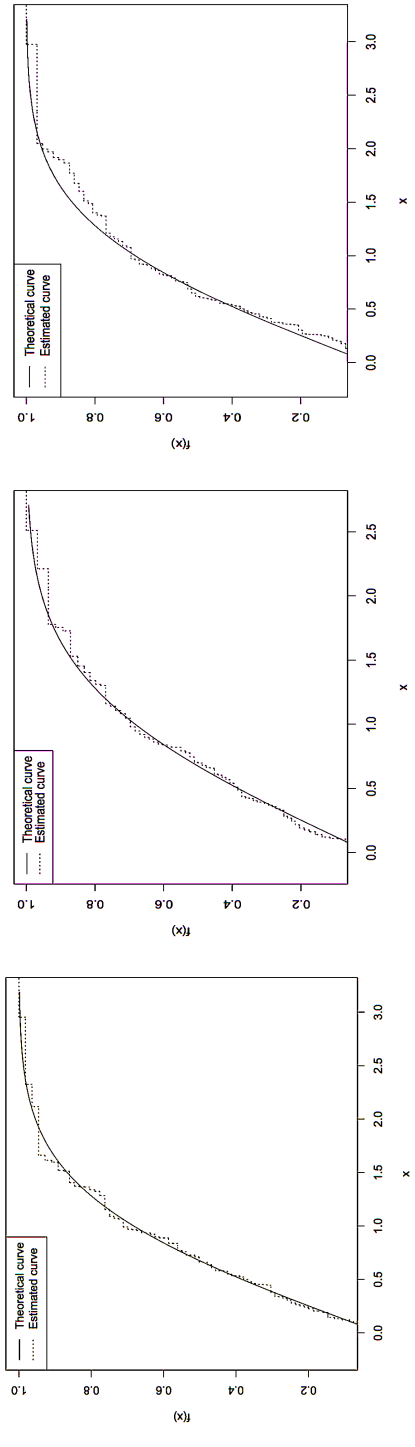

FIGURE 3.1 – Performance de l'estimateur $\widetilde{F}_n$ de Patilea et Rolin pour $p \simeq 30\%$, $n = 100$ et de gauche à droite $a = 0.1$, $a = 0.5$ et $a = 0.9$

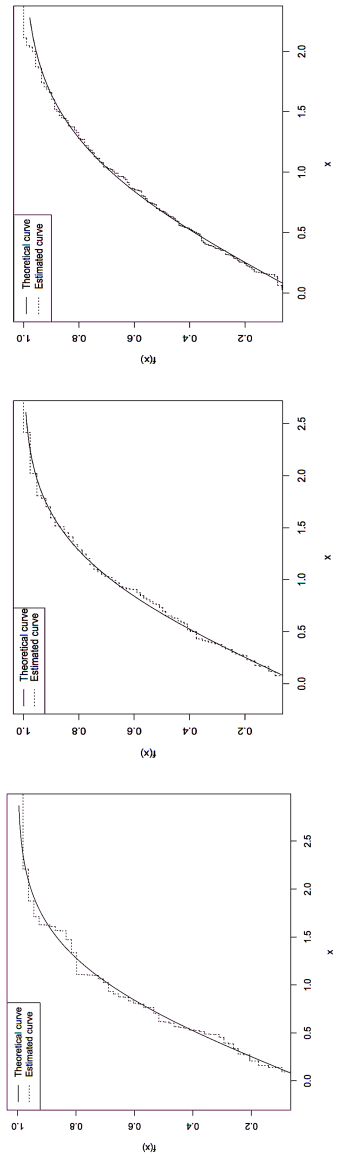

FIGURE 3.2 – Performance de l'estimateur $\widetilde{F}_n$ de Patilea et Rolin pour $a = 0.1$, $p \simeq 30\%$ et de gauche à droite $n = 70$, $n = 100$ et $n = 200$

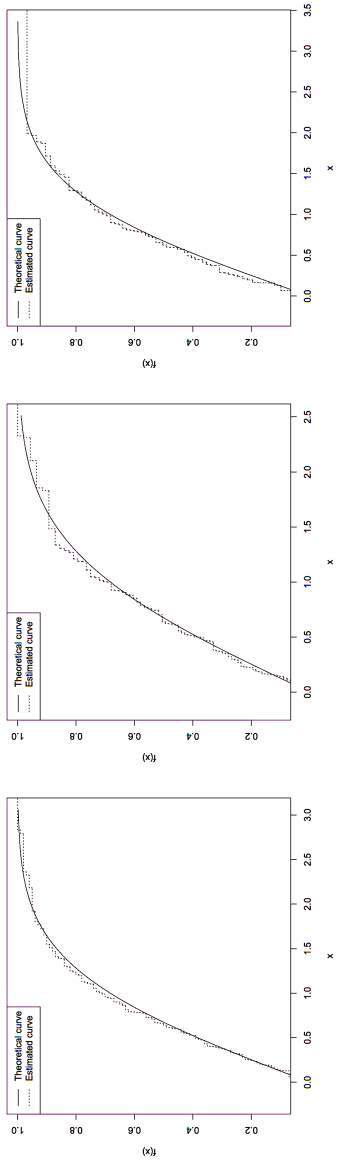

FIGURE 3.3 – Performance de l'estimateur $\widetilde{F}_n$ de Patilea et Rolin pour $a = 0.1$, $n = 100$ et de gauche à droite $p \simeq 0\%$, $p \simeq 20\%$, et $p \simeq 40\%$

3.5 Application sur données réelles

Les données censurées peuvent être rencontrées dans de nombreux domaines tels que la médecine, l'ingénierie, la biologie et l'environnement. Dans le dernier domaine, un processus analytique peut avoir un seuil minimal en dessous duquel la valeur d'une concentration de polluant ne peut être mesurée. Donc, pour certaines concentrations, nous savons seulement qu'elles se situent quelque part entre zéro et une limite de détection (DL).
Ces données sont censurées et sont très courantes dans les études sur la qualité de l'eau. C'est le cas des données provenant de deux stations de la rivière de Niagara au Canada appellées : Niagara on the Lake (NOTL) située à l'embouchure de la rivière de Niagara et Fort Erie (FE) à la tête de la rivière, comme l'indique la Figure 3.4.

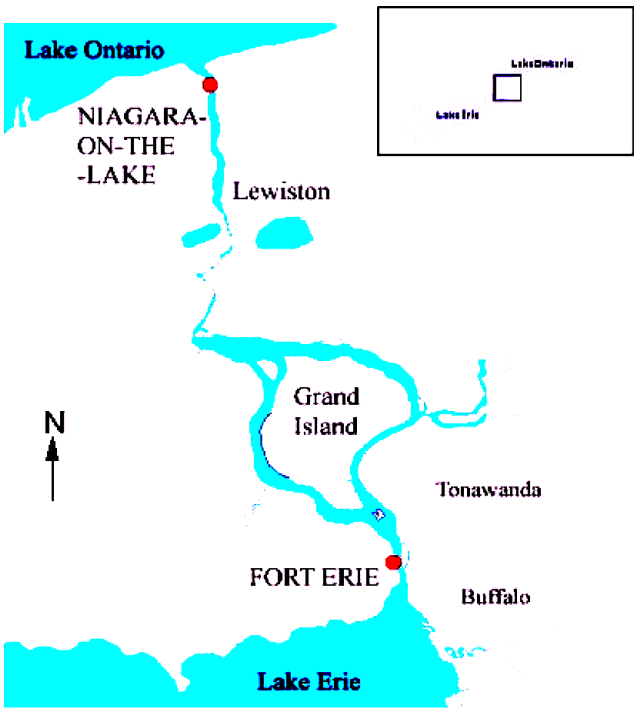

FIGURE 3.4 – Les stations de la rivière de Niagara sur lesquelles porte notre étude

Ces données ont été recueillies au cours de la période de deux ans allant du 1 er avril 1999 au 31 mars 2001 afin de "déterminer si les concentrations de produits chimiques spécifiés à la station NOTL sont statistiquement différentes des concentrations à la station FE, et pour évaluer les tendances dans le temps", comme indiqué dans le rapport de Kuntz et Klawunn [2005].

Dans ce rapport, l'estimation du maximum de vraisemblance, notée MLE a été appliquée afin d'estimer les moyennes annuelles et les intervalles de confiance des concentrations de plusieurs polluants de l'eau. Une autre méthode courante est celle de Kaplan-Meier qui n'est basée que sur l'estimateur $\mathring{F}_n$ étudié dans la section 3.2. Nous utilisons une suite censurée à gauche α-mélangeante de concentrations du Chrome, fournie dans le rapport cité et reprise en annexe B rapportée dans ce document, pour comparer les deux méthodes

mentionnées. Nous accomplissons cela avec la commande "censtats" disponible dans le package NADA du logiciel R et utilisé dans le livre de Helsel [2005]. Les résultats de cette comparaison sont présentés dans les tableaux 3.3 et 3.4 suivants.

TABLEAU 3.3 – Moyenne et écart type des concentrations du Chrome, en μg/L, durant l'année 1999-2000

	la station FE		**la station NOTL**	
Méthode	Moyenne	Écart Type	Moyenne	Écart Type
K-M	0.574	0.715	0.569	0.409
MLE	0.522	0.707	0.572	0.528

TABLEAU 3.4 – Moyenne et écart type des concentrations du Chrome, en μg/L, durant l'année 2000-2001

	la station FE		**la station NOTL**	
Méthode	Moyenne	Écart Type	Moyenne	Écart Type
K-M	0.515	0.688	0.824	0.832
MLE	0.838	3.871	1.365	5.346

Nous pouvons voir que la méthode de Kaplan-Meier a le plus faible écart type, sauf pour l'année 1999-2000 à la station FE où les deux valeurs sont bien proches. Ainsi, similairement à Helsel [2005], nous utilisons la commande "cenfit" du package NADA pour étudier les concentrations du Chrome à chaque station par la méthode de Kaplan-Meier (l'estimateur $\mathring{F}_n$). La figure 3.5 représente les courbes des fonctions de distributions estimées ($\mathring{F}_n$) des concentrations du Chrome dans chaque station durant l'année 1999-2000.

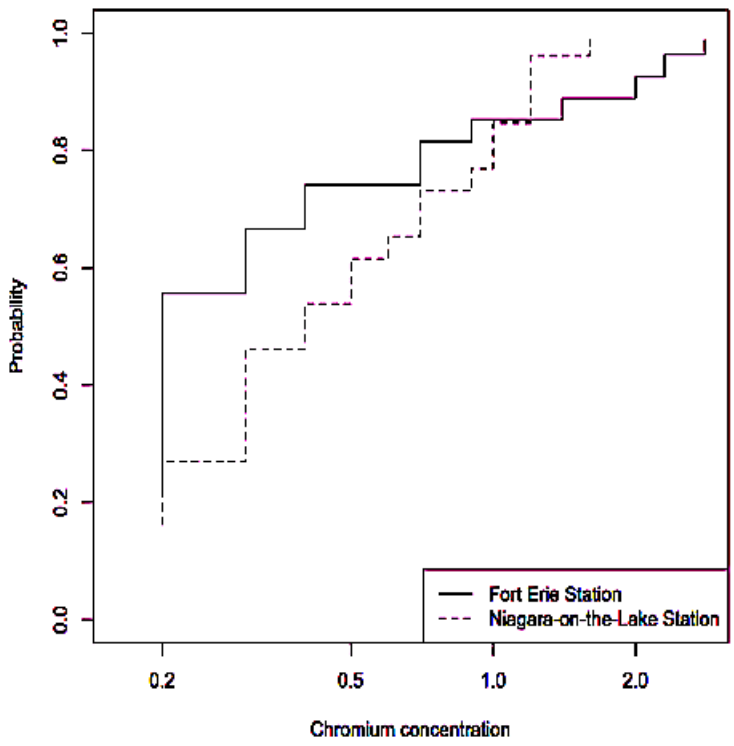

FIGURE 3.5 – Les estimateurs $\mathring{F}_n$ des fonctions de distribution des concentrations du Chrome, en μg/L, durant l'année 1999-2000

Sur cette Figure, nous remarquons d'abord que les deux courbes commencent à la même valeur, ce qui signifie que les deux stations ont la même concentration minimale égale à $0,2$ μg/L. Deuxièmement, la courbe de la station NOTL est décalée vers la droite après le 25^{e} percentile jusqu'à environ le 85^{e} percentile, ce qui indique que la concentration du Chrome apparaît plus élevée à la station de NOTL que les percentiles équivalents à la station de FE, en dessous d'environ le 85^{e} percentile. En dessous du 25^{e} percentile environ, les données des deux stations sont censurées et ne peuvent donc pas être distinguées.

Le Tableau 3.5 résume certaines statistiques descriptives ainsi que les intervalles de confiance à 90% des concentrations du Chrome durant l'année 1999-2000.

TABLEAU 3.5 – Statistique Sommaire des concentrations du Chrome, en μg/L, dans la rivière de Niagara durant l'année 1999-2000

Statistiques	**Station : FE**	**NOTL**	**Percentiles**	**Station : FE**	**NOTL**
n [a]	27	26	10[e]	NA [d]	NA
$n.cen$ [b]	6	4	25[e]	0.200	0.200
moyenne	0.574	0.569	Médiane	0.200	0.400
90% ICI	0.347	0.437	75[e]	0.700	0.900
90% ICS	0.800	0.701	90[e]	2.000	1.200
σ [c]	0.715	0.409			

a. taille de l'échantillon
b. observations censurées à gauche
e. écart type
f. indisponible

Maintenant, la Figure 3.6 montre les courbes des distributions estimées ($\mathring{F}$) des concentrations du Chrome dans la rivière de Niagara durant l'année 2000-2001.

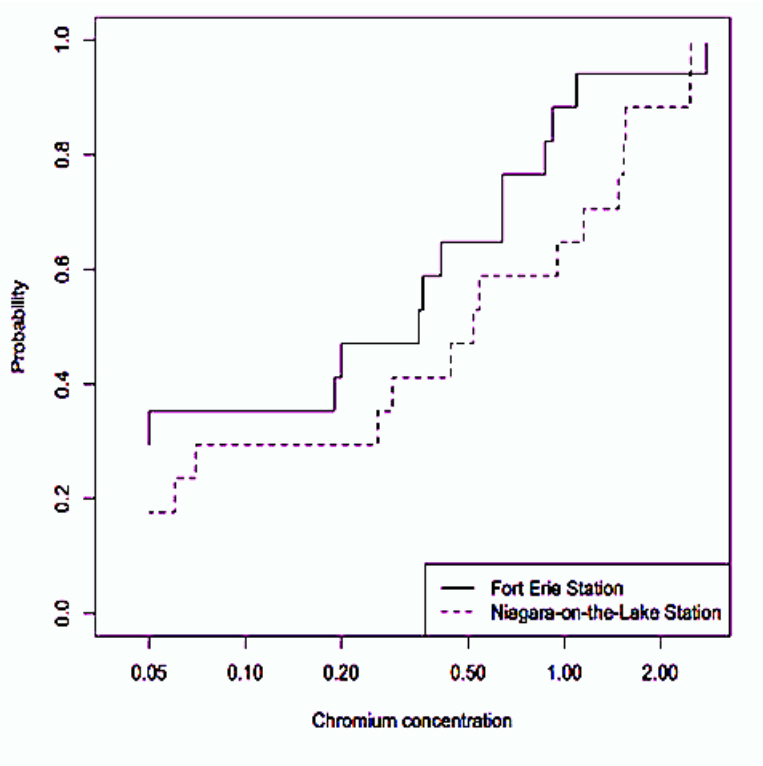

FIGURE 3.6 – Les estimateurs $\mathring{F}_n$ des fonctions de distribution des concentrations du Chrome, en μg/L, durant l'année 2000-2001

Sur cette Figure, nous remarquons que les données des deux stations sont censurées au-dessous du 25^{e} percentile où les deux courbes commencent à la même plus petite valeur ($0,05$ μg/L). Au-delà, la courbe de la station NOTL est clairement décalée vers la droite jusqu'à environ le 95^{e} percentile, ce qui indique que la station FE présente la plus faible concentration du Chrome.

Le Tableau 3.6 fournit les mêmes informations que celles du Tableau 3.5 mais pour l'année 2000-2001.

Tableau 3.6 – Statistique Sommaire des concentrations du Chrome, en μg/L, dans la rivière de Niagara durant l'année 2000-2001

Statistiques	**station : FE**	**NOTL**	**Percentiles**	**station : FE**	**NOTL**
n	17	17	10^{e}	NA	NA
$n.cen$	5	3	25^{e}	NA	0.070
moyenne	0.515	0.824	Médiane	0.350	0.520
90% ICI	0.240	0.491	75^{e}	0.640	1.480
90% ICS	0.790	1.156	90$ème$	1.090	2.480
σ	0.688	0.832			

En conclusion, les Figures 3.5 et 3.6 et les Tableaux 3.5 et 3.6 indiquent que, en gros, les percentiles des concentrations du Chrome à la station NOTL semblent être plus élevés que leur percentiles équivalents à la station FE durant les deux années. De plus, les concentrations du Chrome montrent une augmentation en 2000-2001 à la station NOTL.

Notons qu'il s'agit d'une étude descriptive et que certains tests peuvent vérifier si les différences observées sont significatives, mais cela sort du cadre de cette thèse.

Remarquons également que si nous utilisons l'intervalle de confiance supérieur à 90% (voir Tableaux 3.5 et 3.6), qui constitue une approche protectrice contre les dépassements des critères, les concentrations de Chrome dans les deux stations sont très inférieures à la valeur critère annuelle de 50 μg/L (comme indiqué à la page 13 du rapport de Kuntz et Klawunn [2005]).

Chapitre 4

Étude des estimateurs de la densité et du taux de hasard dans un cadre de censure mixte α-mélangeant

Sommaire

Dans ce chapitre, nous supposons que la variable Y, du précédent chapitre, admet une fonction de densité, notée f_Y, estimée par $\widetilde{f}_n$. Dans un modèle de censure mixte α-mélangeant, nous présentons des propriétés asymptotiques de consistance presque complète, ponctuelle et uniforme de cet estimateur de la fonction de densité, avec des vitesses de convergence. Nous étudions par la suite la consistance de l'estimateur de la fonction de hasard. Nous obtenons nos résultats, en faisant usage de ceux obtenus dans le chapitre précédent.

Dans ce qui suit, nous appliquons le résultat du Théorème 6, concernant le taux de consistance de l'estimateur de Patilea-Rolin $\widetilde{F}_n$, afin d'établir, dans la section 4.1, la convergence presque complète uniforme de l'estimateur de la fonction de densité f_Y et nous dérivons la vitesse de convergence de l'estimateur considéré, sur un compact. Nous déduisons par la suite, dans la section 4.2, le même taux de convergence presque complète uniforme de l'estimateur du taux de hasard de la variable Y soumise aux conditions du α-mélange et de censure mixte.

Tout au long de ce chapitre, nous nous plaçons dans le cadre du modèle I de Patilea et Rolin [2006] tout en gardant les mêmes notations précédemment utilisées dans la Section 1.5.

4.1 Taux de consistance forte de l'estimateur de la densité

En suivant le même raisonnement adopté par Rosenblatt [1956] et Parzen [1962] dans le cas de données complètes (expliqué précédemment dans 1.3.4), puis par Földes *et al.* [1981] dans le cas de censure à droite (voir 1.5), Kitouni *et al.* [2015] ont proposé d'estimer f_Y avec l'estimateur non-paramétrique suivant :

$$\widetilde{f}_n(y) = \frac{1}{h_n} \int K\left(\frac{y-z}{h_n}\right) d\widetilde{F}_n(z), \tag{4.1}$$

où K est le noyau, h_n est la fenêtre et $\widetilde{F}_n$ est l'estimateur de Patilea-Rolin donné par la relation suivante

$$\widetilde{S}_n(t) = 1 - \widetilde{F}_n(t) = \prod_{j/Z'_j \leq t} \left(1 - \frac{D_{0j}}{n\mathring{F}_n(Z'_{j-1}) - nH_n(Z'_{j-1})}\right),$$

Notons par ζ un sous-ensemble compact de $[0, \min(T_Y, T_R)[$. Notre objectif est d'étendre le résultat donné par Kitouni *et al.* [2015], concernant le taux de convergence presque complète uniforme de $\widetilde{f}_n$ sur ζ, au cas de données α-mélangeantes. Nous allons donc travailler sous la condition du α-mélange (**C.1**), définie précedemment dans le Chapitre 2, avec $\max(\alpha_1(n), \alpha_2(n), \alpha_3(n)) = O(n^{-\nu})$ et $\nu > 4$. Nous supposons, de plus, quelques hypothèses classiques en estimation non-paramétrique données ci-dessous.

❋ (**C.3**) : f_Y est r fois continûment différentiable sur ζ, avec $r \geq 2$,

❋ (**C.4**) : $\exists \beta_1 > 0, \beta_2 > 0, \varepsilon > 0, \forall y \in \zeta, \forall z \in]y-\varepsilon, y+\varepsilon[, |f_Y(y)-f_Y(z)| \leq \beta_1 |y-z|^{\beta_2}$,

❋ (**C.5**) : K est une fonction continue à droite, à variation bornées, vérifiant $\int K(u)\, du = 1$ (i.e. $\exists M > 0,\ \forall u \in \mathbb{R},\ |u| \geq M \Rightarrow K(u) = 0$) et $\forall 1 \leq j \leq r-1$, $\int u^j K(u) du = 0$,

❋ (**C.6**) : $\lim\limits_{n \to +\infty} h_n = 0$ et $\lim\limits_{n \to +\infty} n h_n^2 / \log n = \infty$.

Dans le théorème suivant, nous énonçons la convergence presque complète uniforme, avec taux, de $\widetilde{f}_n$.

Théorème 7. *Si* $\max(I_L, I_R) < I_Y$ *et* $\max(\alpha_1(n), \alpha_2(n), \alpha_3(n)) = O(n^{-\nu})$ *avec* $\nu > 4$ *alors*

(i) *Sous les hypothèses* (***C.3***), (***C.5***) *et* (***C.6***), *nous avons*

$$\sup_{y \in \zeta} |\widetilde{f}_n(y) - f_Y(y)| = O_{p.c.}\left(h_n^r + \frac{1}{h_n}\sqrt{\frac{\log n}{n}} \right).$$

(ii) *Sous les hypothèses* (***C.4***)–(***C.6***), *nous avons*

$$\sup_{y \in \zeta} |\widetilde{f}_n(y) - f_Y(y)| = O_{p.c.}\left(h_n^{\beta_2} + \frac{1}{h_n}\sqrt{\frac{\log n}{n}} \right).$$

Remarque. *Les vitesses de convergences obtenues dans ce théorème sont les mêmes que celles obtenues par Kitouni* et al. *[2015] pour le même estimateur mais avec des données indépendantes.*

4.1.1 Preuve

En s'inspirant largement de la preuve de Kitouni *et al.* [2015], nous considérons la décomposition suivante :

$$|\widetilde{f}_n(y) - f_Y(y)| \leq |\widetilde{f}_n(y) - \mathbb{E}(\widetilde{f}_n(y))| + |\mathbb{E}(\widetilde{f}_n(y)) - f_Y(y)|, \tag{4.2}$$

où

$$\mathbb{E}(\widetilde{f}_n(y)) = \frac{1}{h_n} \int K\left(\frac{y-z}{h_n}\right) dF_Y(z). \tag{4.3}$$

Prouver le Théorème 7 revient à montrer les deux Lemmes ci-dessous, traitant chacun un des termes de la décomposition 4.2. Commençons par étudier le premier terme.

Lemme 4.1. *Si* $\max(I_L, I_R) < I_Y$, $\max(\alpha_1(n), \alpha_2(n), \alpha_3(n)) = O(n^{-\nu})$ *avec* $\nu > 4$ *et les hypothèses* (***C.5***) *et* (***C.6***) *sont satisfaites, nous avons*

$$\sup_{y\in\zeta} |\widetilde{f}_n(y) - \mathbb{E}(\widetilde{f}_n(y))| = O_{p.c.}\left(\frac{1}{h_n}\sqrt{\frac{\log n}{n}}\right).$$

Démonstration. Suite aux relations (4.1) et (4.3), nous avons

$$|\widetilde{f}_n(y) - \mathbb{E}(\widetilde{f}_n(y))| = \frac{1}{h_n}\left|\int K\left(\frac{y-z}{h_n}\right) d(\widetilde{F}_n(z) - F_Y(z))\right|. \tag{4.4}$$

L'intégration par partie de cette dernière avec un changement de variable classique, utilisés conjointement avec l'hypothèse (**C.5**) impliquent que

$$\begin{aligned}
|\widetilde{f}_n(y) - \mathbb{E}(\widetilde{f}_n(y))| &= \frac{1}{h_n}\left|\int_{-M}^{M} K(u) d(\widetilde{F}_n(y-uh_n) - F_Y(y-uh_n))\right| \\
&\leq \frac{1}{h_n}\left|\int_{-M}^{M} (\widetilde{F}_n(y-uh_n) - F_Y(y-uh_n)) dK(u)\right| \\
&\leq \frac{V_K}{h_n} \sup_{u>-M} |F_n(y-uh_n) - F_Y(y-uh_n)|, \qquad (4.5)
\end{aligned}$$

où V_K est la variation totale de K sur $\mathbb{R}$.

Posons $\theta = \max(\zeta)$, et $\theta^* \in]\theta, \min(T_Y, T_R)[$. Comme $h_n \underset{n\to+\infty}{\longrightarrow} 0$, il est alors important de remarquer que, pour un n assez grand, $h_n < \dfrac{\theta^* - \theta}{M}$.

Ceci combiné à la majoration (4.5), il en résulte que

$$\begin{aligned}
\sup_{y\in\zeta} |\widetilde{f}_n(y) - \mathbb{E}(\widetilde{f}_n(y))| &\leq \frac{V_K}{h_n} \sup_{t<\theta^*} |\widetilde{F}_n(t) - F_Y(t)| \\
&= O_{p.c.}\left(\frac{1}{h_n}\sqrt{\frac{\log n}{n}}\right),
\end{aligned}$$

en vertu du Théorème 6, **(ii)**.

□

Enfin, le deuxième terme de la décomposition (4.2), se contrôle par le lemme suivant :

Lemme 4.2. **(i)** *Sous les hypothèses* (***C.3***), (***C.5***) *et* (***C.6***), *nous avons*

$$\sup_{y\in\zeta}|\mathbb{E}(\widetilde{f}_n(y)) - f_Y(y)| = O(h_n^r).$$

(ii) *Sous les hypothèses* (***C.4***)–(***C.6***), *nous avons*

$$\sup_{y\in\zeta}|\mathbb{E}(\widetilde{f}_n(y)) - f_Y(y)| = O(h_n^{\beta_2}).$$

Démonstration. L'hypothèse (**C.5**), avec un changement de variable classique, permet d'écrire

$$\begin{aligned}\mathbb{E}(\widetilde{f}_n(y)) - f_Y(y) &= \frac{1}{h_n}\int K\left(\frac{y-z}{h_n}\right) f_Y(y)dy - f_Y(y)\\ &= \int_{-M}^{M} K(u) f_Y(y-uh_n)du - f_Y(y)\int_{-M}^{M} K(u)du\\ &= \int_{-M}^{M} K(u)(f_Y(y-uh_n)du - f_Y(y))du. \qquad (4.6)\end{aligned}$$

(i) Par l'application du développement de Taylor sur (4.6), nous obtenons sous les hypothèses (**C.3**) et (**C.5**), que pour $y - uh_n \leq y_0 \leq y$,

$$|\mathbb{E}\widetilde{f}_n(y) - f_Y(y)| = \left|\frac{h_n^r}{r!}\int_{-M}^{M} f_Y^{(r)}(y_0)u^r K(u)du\right|.$$

Le résultat du Lemme 4.2.**(i)** découle alors immédiatement, des hypothèses (**C.3**) et (**C.6**) et la compacité de ζ.

(ii) En revenant à l'équation (4.6) et en utilisant l'hypothèse (**C.4**), nous obtenons

$$|f_Y(y-uh_n) - f_Y(y)| \leq \beta_1|u|^{\beta_2}h_n^{\beta_2},$$

d'où

$$\begin{aligned}\sup_{y\in\zeta}|\mathbb{E}\widetilde{f}_n(y) - f_Y(y)| &\leq \beta_1\int_{-M}^{M}|K(u)||u|^{\beta_2}h_n^{\beta_2}du\\ &\leq \beta_1 M^{\beta_2}h_n^{\beta_2}\int|K(u)|du\\ &= O(h_n^{\beta_2}),\end{aligned}$$

au vu de l'hypothèse (**C.5**).

□

4.2 Taux de consistance forte de l'estimateur du taux de hasard

Dans cette section, nous nous intéressons à la vitesse de convergence presque complète uniforme de l'estimateur du taux de hasard de la variable aléatoire Y, défini par

$$\lambda_Y(y) = \begin{cases} \frac{f_Y(y)}{S_Y(y)} & \text{si} \quad S_Y(y) \neq 0, \\ 0 & \text{sinon.} \end{cases}$$

Etendant le cas des données censurées à droite, étudié par Földes *et al.* [1981], Kitouni *et al.* [2015] ont proposé d'estimer λ_Y comme suit :

$$\widetilde{\lambda}_n(y) = \frac{\widetilde{f}_n(y)}{1 - \widetilde{F}_n(y) + v_n}, \tag{4.7}$$

où $(v_n)_{n\in\mathbb{N}}$ est une suite de nombres réels strictement positifs qui converge vers 0, introduite afin d'éviter la division par 0.

Nous donnons maintenant un résultat uniforme sur la convergence presque complète de l'estimateur du taux de hasard $\widetilde{\lambda}_n$ avec les mêmes vitesses que l'estimateur de la fonction de densité $\widetilde{f}_n$.

Théorème 8. *Si* $\max(I_L, I_R) < I_Y$ *et* $\max(\alpha_1(n), \alpha_2(n), \alpha_3(n)) = O(n^{-\nu})$ *avec* $\nu > 4$ *alors*

(i) *Sous les hypothèses* (***C.3***), (***C.5***) *et* (***C.6***), *nous avons*

$$\sup_{y\in\zeta} |\widetilde{\lambda}_n(y) - \lambda_Y(y)| = O_{p.c.}\left(h_n^r + \frac{1}{h_n}\sqrt{\frac{\log n}{n}}\right).$$

(ii) *Sous les hypothèses* (***C.4***)–(***C.6***), *nous avons*

$$\sup_{y\in\zeta} |\widetilde{\lambda}_n(y) - \lambda_Y(y)| = O_{p.c.}\left(h_n^{\beta_2} + \frac{1}{h_n}\sqrt{\frac{\log n}{n}}\right).$$

Remarque. *Notons que dans le Théorème ci-dessus établi dans le cas* α*-mélangeant, nous atteignons les mêmes vitesses obtenues dans le Théorème 3 de Kitouni* et al. *[2015], pour des données indépendantes.*

4.2.1 Preuve

Pour la preuve de ce Théorème, nous procédons de manière analogue à Kitouni *et al.* [2015] en écrivant, pour tout $y \in \zeta$:

$$|\widetilde{\lambda}_n(y) - \lambda_Y(y)| \leq \frac{1}{1-\widetilde{F}_n(y)+v_n}|\widetilde{f}_n(y) - f_Y(y)| + \left|1-\widetilde{F}_n(y) - S_Y(y) + v_n\right| \times \frac{f_Y(y)}{S_Y(y)(1-\widetilde{F}_n(y)+v_n)},$$

on en déduit alors, en notant $\theta = \max(\zeta)$, que :

$$|\widetilde{\lambda}_n(y) - \lambda_Y(y)| \leq \frac{1}{\inf\limits_{y\in\zeta}(1-\widetilde{F}_n(y)+v_n)} \sup\limits_{y\in\zeta}|\widetilde{f}_n(y) - f_Y(y)| + \sup\limits_{y\in\zeta} f_Y(y) \frac{\sup\limits_{y\in\zeta}|1-\widetilde{F}_n(y) - S_Y(y) + v_n|}{S_Y(\theta)\inf\limits_{y\in\zeta}(1-\widetilde{F}_n(y)+v_n)}. \tag{4.8}$$

D'autre part, nous avons pour $\xi \in]0, S_Y(\theta)[$

$$\mathbb{P}\left(\inf_{y\in\zeta}(1-\widetilde{F}_n(y)+v_n) \leq \frac{\xi}{2}\right) \leq \mathbb{P}\left(\sup_{y\in\zeta}|1-\widetilde{F}_n(y) - S_Y(y) + v_n| > \frac{\xi}{2}\right),$$

or, d'aprés le Théorème 6.(ii), $\sup\limits_{y\in\zeta}|1-\widetilde{F}_n(y) - S_Y(y) + v_n| > \frac{\xi}{2}$ est le terme général d'une série convergente, ceci implique que

$$\sum_{n=1}^{\infty} \mathbb{P}\left(\inf_{y\in\zeta}(1-\widetilde{F}_n(y)+v_n) \leq \frac{\xi}{2}\right) < \infty. \tag{4.9}$$

Pour finir la preuve, il suffit de combiner (4.8) et (4.9) avec le Théorème 6.(ii) et le Théorème 7.

Remarque. *Notons que, de façon similaire à Kitouni* et al. *[2015], nous pouvons déduire le même résultat du Théorème 8 pour* $\lambda_n(y) = \frac{1}{h_n}\int K\left(\frac{y-z}{h_n}\right) d\widetilde{\Lambda}_n(z)$, *où* $\widetilde{\Lambda}_n$ *est donnée par la relation (2.3) du Chapitre 2. Sous la condition du* α*-mélange et en remplaçant l'hypothèse* (***C.4***) *par l'hypothèse* (***C.4******), *nous obtenons la même vitesse de convergence presque complète uniforme.*

❋ (***C.4******) $\exists \beta_1 > 0, \beta_2 > 0, \varepsilon > 0, \forall y \in \zeta, \forall z \in]y-\varepsilon, y+\varepsilon[, |\lambda_Y(y) - \lambda_Y(z)| \leq \beta_1 |y-z|^{\beta_2}$. *La preuve de ce résultat suit les mêmes lignes que celles de la preuve du Théorème 7, en remplaçant* $\mathbb{E}(\widetilde{f}_n(y)$ *par* $\mathbb{E}\lambda_n(y) = \frac{1}{h_n}\int K\left(\frac{y-z}{h_n}\right) d\Lambda_Y(z)$ *et en utilisant le résultat* **(i)** du Théorème 6 à la place de **(ii)**.

Perspectives de recherche

Dans cette partie, nous présentons brièvement certaines pistes à explorer dans des recherches à venir.

Il serait intéressant d'étudier le prolongement de nos résultats au cas de données fonctionnelles, sous des conditions de dépendance plus fortes.

Il est aussi envisageable de s'intéresser à l'estimation non-paramétrique de la fonction de régression dans un modèle de censure mixte α-mélangeant. L'estimateur à étudier est une généralisation des estimateurs de type Nadaraya-Watson au cas de censure mixte, il a été introduit par Messaci [2010]. Notre contribution dans ce domaine serait d'établir la convergence uniforme presque complète de cet estimateur basé sur des observations fortement mélangeantes.

D'autre part, il serait souhaitable de travailler dans le cadre d'un modèle de censure double, proposé par Turnbull [1974], qui a construit un estimateur self-consistant de la fonction de répartition. Dans ce cas aussi la variable d'intérêt est censurée à droite par une variable aléatoire R et à gauche par une variable aléatoire L. La différence entre ce type de censure et celui évoqué dans notre thèse, est que dans la censure double, nous supposons que L est presque sûrement inférieure à R, alors que dans la censure mixte nous supposons qu'elles sont indépendantes.

Annexe A

Quelques outils de probabilités

Soit $S_n = \sum_{i=1}^n X_i$ et $\bar{X} = S_n/n$.

A.1 Loi forte des grands nombres

Théorème 9. *Soit $(X_n)_{n\geq 1}$ une suite de variables aléatoires i.i.d., alors $\bar{X}$ converge presque sûrement si et seulement si $\mathbb{E}(|X_1|) < +\infty$.*

A.2 Théorème Central limit

Théorème 10. *Soit $(X_n)_{n\geq 1}$ un échantillon i.i.d. d'une loi de moyenne m et de variance σ^2. La convergence suivante a lieu en loi, lorsque $n \to \infty$*

$$\sqrt{n}\,\frac{\bar{X} - m}{\sigma} = \frac{S_n - nm}{\sigma\sqrt{n}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1),$$

où $\mathcal{L}$ désigne la convergence en loi et $\mathcal{N}(0,1)$ est la loi gaussienne centrée réduites.

A.3 Lemme de Borel-Cantelli

Lemme A.1. *Soit $(A_n)_{n\geq 1}$ une suite d'événements. Si $\sum_{n=1}^{\infty} \mathbb{P}(A_n) < \infty$, alors*

$$\mathbb{P}(\limsup_{n\to\infty} A_n) = 0.$$

Si les événements sont indépendants et $\sum_{n=1}^{\infty} \mathbb{P}(A_n) < \infty$, alors

$$\mathbb{P}(\limsup_{n\to\infty} A_n) = 1.$$

A.4 Théorème de Glivenko-Contelli

Théorème 11. *Soit $(X_n)_{n\geq 1}$ une suite de v.a.i.i.d., alors avec une probabilité* 1,

$$\sup_{t\in\mathbb{R}}|\widehat{F}_n(t) - F_X(t)| \underset{n\to\infty}{\longrightarrow} 0.$$

Pour les détails de la démonstration, consulter la page 115 de Laha et Rohatgi [1979].

A.5 Proposition A.6 de Ferraty et Vieu [2006]

Lemme A.2. *Soient $(X_n)_{n\geq 1}$, $(Y_n)_{n\geq 1}$ des suites de v.a.r. et (u_n) une suite de nombres réels positifs. Si $\lim_{n\to\infty} u_n = 0$, $X_n = O_{p.co.}(u_n)$ et $\lim_{n\to\infty} Y_n = l_Y$, p.co., où l_Y est un nombre réel, alors*

— $X_nY_n = O_{p.co.}(u_n)$,

— $\dfrac{X_n}{Y_n} = O_{p.co.}(u_n),\ \ l_Y \neq 0.$

A.6 Inégalités exponentielles

Les inégalités exponentielles sont un outil que nous avons utilisé de manière déterminante dans notre travail de recherche et que l'on peut rencontrer sous l'appellation de "inégalité de type Fuk-Nagaev".

A.6.1 Inégalité de Davydov

Soit $(X_n)_{n\in\mathbb{Z}}$ une suite de v.a. α-mélangeantes. on considère la v.a. T (resp. T') qui, pour tout $k \in \mathbb{Z}$, est $\sigma(X_i, -\infty \leq i \leq k)$-mesurable (resp. $\sigma(X_i,\ n+k \leq i \leq +\infty)$-mesurable).

Lemme A.3. *(Proposition A.10.i de Ferraty et Vieu [2006]). Si T et T' sont bornées, alors :*

$$\exists C, 0 < C < +\infty,\ \ cov(T, T') \leq C\alpha(n).$$

Nous nous en tiendrons ici à la version la plus explicite que celles qui sont effectivement disponibles dans la littérature mais, qui est largement suffisante dans notre contexte et que l'on trouve sous cette forme dans les livres de Rio [2000], p. 87 et Ferraty et Vieu [2006], p. 237.

Lemme A.4. *Soit $(U_i)_{i\in\mathbb{N}}$ une suite de v.a.r. centrées identiquement distribuées et de coefficient de mélange fort $\alpha(n) = O(n^{-\nu})$, $\nu > 1$. S'il existe $M < \infty$ telle que $|U_i| \leq M$, alors il existe un $C < \infty$ telle que pour tout $r \geq 1$ et tout $\epsilon > 0$:*

$$\mathbb{P}\left\{\left|\sum_{i=1}^{n} U_i\right| > \epsilon\right\} \leq C\left(1 + \frac{\epsilon^2}{rS_n^2}\right)^{-r/2} + nCr^{-1}\left(\frac{r}{\epsilon}\right)^{\nu+1},$$

où $S_n^2 = \sum_{1\leq i,j\leq n} |cov(U_i, U_j)|$.

Cette inégalité est une des extensions de l'inégalité de Bernstein, qu'on peut trouver sous diverses formes dans Hoeffding [1963]. Le développement des résultats de l'estimation non paramétrique sous la condition de dépendance s'est fait en parallèle avec celui de telles inégalités. Depuis, la puissance de ces inégalités a été améliorée par Bosq [993b] jusqu'à l'inégalité de Rio [2000].

Annexe B

Concentrations du Chrome dans la rivière de Niagara

Tableau B.1 – Concentration du Chrome dans la rivière de Niagara durant l'année 1999-2000

NOTL station	0.2 0.3 0.3 0.3 1.0 0.5 1.2 1.0 0.2 0.3 0.4 0.7 0.2^- 0.2^- 0.2^- 1.6 1.2
(μ g/L)	0.2^- 1.2 0.9 0.5 0.3 0.6 0.4 0.2 0.7
FE station	0.3 0.2 0.7 2.3 0.2 0.2 0.4 2.8 0.2 0.2 0.2 0.3 0.2 0.4 0.2 0.2^- 0.2^-
(μ g/L)	0.2^- 0.3 2.0 1.4 0.7 0.2^- 0.2^- 0.2^- 0.2 0.9

Tableau B.2 – Concentration du Chrome dans la rivière de Niagara durant l'année 2000-2001

NOTL station	1.48 0.95 1.15 0.54 0.44 0.52 0.29 0.26 0.05^- 1.56 1.54 2.48 2.49 0.05^-
(μg/L)	0.06 0.07 0.05^-
FE station	1.09 0.64 0.87 0.41 0.35 0.19 0.2 0.05^- 0.05 0.92 0.64 2.79 0.36 0.05^-
(μg/L)	0.05^- 0.05^- 0.05^-

- left censored value.

Bibliographie

Berlinet, A. (1993). Hierarchies of higher order kernels. *Probability Theory and Related Fields*, 94:489–504. 15

Blum, J., Hanson, D. et Koopmans, L. (1963). On the strong law of large numbers for a class of stochastic processes. *Wahrscheinlichkeitstheorie und Verw. Gebiete*, 2:1–11. 8

Bosq, D. (1975). Inégalité de bernstein pour les processus stationnaires et mélangeants. *Applications. C. R. Acad. Sci. Paris*, 24:1095–1098. 10

Bosq, D. (1993b). Bernstein-type large deviations inequalities for partial sums of strong mixing processes. *Statistics*, 24:59–70. 10, 69

Bosq, D. (1996). *Nonparametric statistics for stochastic processes. Estimation and prediction.* Lecture Notes in Statistics 110 Springer- Verlag, New York. 10

Bowman, A. W. (1984). An alternative method of cross-validation for smoothing of density estimates. *Biometrika*, 71(2):253–360. 15

Bradley, R. (2005). Basic properties of strong mixing conditions. *A survey and some open questions Probability Surveys*, 2:107–144. 9

Bradley, R. (2007). *Introduction to Strong Mixing Conditions, Vols. 1, 2, and 3.* Kendrick Press, Heber City (Utah). 10

Breslow, N. et Crowley, J. (1974). A large sample study of the life table and product limit estimates under random censorship. *The Annals of Statistics*, 2(3):437–453. 17

Cacoullos, T. (1966). Estimation of a multivariate density. *Ann. Inst. Statist. Math.*, 18:179–189. 15

Cai, Z. (1998). Asymptotic properties of kaplan-meier estimator for censored dependent data. *Statistics & Probability Letters*, 37:381–389. 17

Cai, Z. (1998b). Kernel density and hazard rate estimation for censored dependent data. *J. Multivariate Anal.*, 67:23–34. 18

Cai, Z. (2001). Estimating a distribution function for censored time series data. *J. Multivariate Anal*, 78:299–318. 17, 23, 24, 26, 40, 44

Cai, Z. et Roussas, G. G. (1992). Uniform strong estimation under α-mixing with rates. *Statistics & Probability Letters*, 15:47–55. 13, 17, 23, 26

Carbon, M. (1983). Inégalité de bernstein pour les processus fortement mélangeants, non nécessairement stationnaires. *C. R. Math. Acad. Sci. Paris*, 5:303–306. 10

Chanda, K. (1974). Strong mixing properties of linear stochastic processes. *J. Appl. Probability*, 11:401–408. 10

Chang, K. L. (1949). An estimate concerning the kolmogoroff limit distribution. *Transactions of the American Mathematical Society*, 67:36–50. 13, 22

Collomb, G., Hassani, S., Sarda, P. et Vieu, P. (1985). Estimation non paramétrique de la fonction de hasard pour des observations dépendantes. *Statistique et Analyse des Donneés*, 10:42–49. 13

Davydov, Y. (1973). Mixing conditions for markov chains. *Theory Probab. Appl.*, 18:312–328. 10

Delecroix, M. (1979). Sur l'estimation des densités d'un processus stationnaire et mélangeant. *Publ. U.E.R. Math. Pures Appl. IRMA*, 20(4):exp. no. I, 24. Seminar on Mathematical Statistics. 15

Diehl, S. et Stute, W. (1988). Kernel density and hazard function estimation in the presence of censoring. *Journal of Multivariate Analysis*, 25:299–310. 1, 18

Doukhan, P. (1994). *Mixing : Properties and Examples.* Lecture Notes in Statistics. Springer-Verlag, Berlin. 9, 10

Doukhan, P., Massart, P. et Rio, E. (1994). The functional central limit theorem for strongly mixing processes. *Ann. Inst. H. Poincaré Probab. Statist.*, 1:63–82. 10

Ferraty, F. et Vieu, P. (2006). *Nonparametric functional data analysis : Theory and practice.* Springer. , 11, 36, 43, 68

Foldes, A. (1974). Density estimation for dependent sample. *Studia Scientiarum Mathematicarum Hungarica*, 9:443–452. 15

Földes, A. et Rejtő, L. (1981a). A LIL type result for the product limit estimator. *Probability Theory and Related Fields*, 56(1):75–86. 1, 17, 22

Földes, A. et Rejtő, L. (1981b). Strong uniform consistency for nonparametric survival curve estimators from randomly censored data. *The Annals of Statistics*, 9:122–129. 17

Földes, A., Rejtő, L. et Winter, B. (1981). Strong consistency properties of nonparametric estimators for randomly censored data, ii : estimation of density and failure rate. *Period. Math. Hungar.*, 12(1):15–29. 1, 17, 60, 64

Gefeller, O. et Michels, P. (1992). A review on smoothing methods for the estimation of the hazard rate based on kernel functions. *In : Dodge, Y. and Whittaker, J.(Eds.), Computational Statistics, Physica-Verlaz, Switzerland*, pages 459–464. 16

Gill, R. D. et Johansen, S. (1990). A survey of product-integration with a view toward application in survival analysis. *The Annals of Statistics*, 18(4):1501–1555. 40, 41, 47

Gorodetskii, V. (1977). On the strong mixing property for linear sequences. *Theory Probab. Appl.*, 22:411–413. 10

Gu, M. G. et Lai, T. L. (1990). Functional laws of the iterated logarithm for the product limit-estimator of a distribution function under random censorship or truncation. *The Annals of Probability*, 18(1):160–189. 17

Hall, P. (1984). Asymptotic properties of integrated square error and cross validation function. *Z. Wahrsch. Verw. Gebiete*, 67:175–195. 15

Hardle, W. (1990). *Applied Nonparametric Regression.* Cambridge University Press. 15

Hardle, W. et Marron, J. (1985). Optimal bandwidth selection in nonparametric regression function estimation. *The Annals of Statistics*, 13(4):1465–1481. 15

Hardy, G. et Littlewood, J. (1914). Some problems of diophantine approximation. *Acta Math.*, 37:155–191. 21

Hartmann, P. et Wintner, A. (1941). On the law of the iterated logarithm. *Amer. J. Math.*, 63:169–176. 22

Hassani, S., Sarda, P. et Vieu, P. (1986). Approche non paramétrique en théorie de la fiabilité : revue bibliographique. *Rev. Statist. Appl.*, 35(4):27–41. 16

Hausdorff, F. (1913). *Grundzüge der Mengenlehre.* Leipzig. 21

Helsel, D. (2005). *Nondetects And Data Analysis : Statistics for censored environmental data.* John Wily & Sons, New York. 55

Hentzschel, J. et Liebscher, E. (1990). Density estimators for complete and censored samples. *Wissenschaftliche Zeitschrift der HfV Dresden*, Special Issue 57:32–57. 18

Hoeffding, w. (1963). Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.*, 58:15–30. 69

Ibragimov, I. (1959). Some limit theorems for stochastic processes stationary in the strict sense. *Dokl. Akad. Nauk SSSR*, 125:711–714. 8

Iosifescu, M. (1968). The law of the iterated logarithm for a class of dependent random variables. *Theory Prob. Appl.*, 13:304–313. 22

Izenman, A. (1991). Developments in nonparametric density estimation. *J. Amer. Statist. Assoc.*, 86:205–224. 16

Kagba, N. (2004). *On kernel density estimation for censored data.* (Ph.D. thesis), University of California, San Diego. 18

Kaplan, E. L. et Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53:457–481. 1, 16

Khintchine, A. (1924). Ueber einen satz der wahrscheinlichkeitsrechnung. *Fundamentall Mathematica*, 6:9–20. 21, 22

Kiefer, J. (1961). 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. 13, 22

Kim, T. et Lee, S. (2005). Kernel density estimator for strong mixing processes. *J. Statist. Plann.*, 133(2):273–284. 15

Kitouni, A., Boukeloua, M. et Messaci, F. (2015). Rate of strong consistency for nonparametric estimators based on twice censored data. *Statistics & Probability Letters*, 96:255–261. 3, 45, 60, 61, 64, 65, 80, 81

Kohler, M., Màthé, K. et Pintér, M. (2002). Prediction from randomly right censored data. *J. Multivariate Anal*, 80:73–100. 1

Kolmogorov, A. et Rozanov, Y. (1960). On strong mixing conditions for stationary gaussian processes. *Theor. Probab. Appl.*, 5:204–208. 9

Kolmogorov, N. (1929). Über des gesetz des iterierten logarithmus. *Mathematische Annalen*, 101:126–135. 22

Kuntz, K. et Klawunn, P. (2005). *Niagara River upstream/downstream monitoring program Report 1999-2000 & 2000-2001.* Library and Archives Canada Cataloguing in Publication. 54, 58

Laha, R. et Rohatgi, V. (1979). *Probability Theory.* John Wiley & Sons, New York. 37, 68

Liebscher, E. (1996). Central limit theorems for sums of α-mixing random variables. *Stochastics Stochastics Rep.*, 3-4:241–258. 10

Liebscher, E. (1998). Convergence or hermite series density estimators under conditions of weak dependence. *Statistics*, 31:191–214. 15

Liebscher, E. (2001a). Central limit theorems for α-mixing triangular arrays with applications to nonparametric statistics. *Math. Methods Statist.*, 2:194–214. 10

Liebscher, E. (2001b). Estimation of the density and the regression function under mixing conditions. *Int. Stat. Rev.*, 19(1):9–26. 15

Liebscher, E. (2002). Kernel density and hazard rate estimation for censored data under α-mixing condition. *Ann. Inst. Statist. Math.*, 34:19–28. 18

Messaci, F. (2010). Local averaging estimates of the regression function with twice censored data. *Statistics & Probability Letters*, 80:1508–1511. 66

Messaci, F. et Nemouchi, N. (2011). A law of the iterated logarithm for the product limit estimator with doubly censored data. *Statistics & Probability Letters*, 81(8):1241–1244. 3, 22, 23, 25

Mielniczuk, J. (1986). Some asymptotic properties of kernel estimators of a density function in case of censored data. *Int. Stat. Rev.*, 14(2):766–773. 18

Mokkadem, A. (1990). Propriétés de mélange des processus autorégressifs polynomiaux. *Ann. Inst. H. Poincaré Probab. Statist.*, 2:219–260. 10

Morales, D., L., P. et Quesada, V. (1991). Bayesian survival estimate for incomplete data when the life distribution is proportionally related to the censoring time distribution. *Comm. Statist. Theory Methods*, 20:831–850. 7

Nze, P. et Rios, R. (1995). Density estimation in the l^∞ norm for mixing processes (in french). *C.R. Acad. Sci. Paris*, 320:1259–1262. 15

Oodaira, H. et Yoshihara, K. (1971). The law of iterated logarithm for stationary processes satisfying mixing conditions. *Kodai Math. Sem. Rep.*, 23:311–334. 22

Parzen, E. (1962). On estimation of a probability density function and mode. *Ann. Math. Statist.*, 33:1065–1076. 14, 15, 17, 60

Pascu, M. et Vaduva, I. (2003). Nonparametric estimation of the hazard rate : a survey. *Roumaine Math. Pures Appl.*, 48:173–191. 16

Patil, P., Wells, M. et Maron, J. (1994). Some heuristics of kernel based estimators of ratio functions. *J. Nonparametr. Statist.*, 4:203–209. 15

Patilea, V. et Rolin, J.-M. (2006). Product limit estimators of the survival function with twice censored data. *The Annals of Statistics*, 34(2):925–938. 2, 3, 7, 18, 19, 22, 23, 60, 81

Philipp, W. (1967). Das gesetz vom iterierten logarithmus fur stark mischende stationare prozesse, z. *Wahrsch. verw. Gebiete*, 8:204–209. 22

Philipp, W. (1969a). The central limit problem for mixing sequences of random variables. *Wahrscheinlichkeitstheorie und Verw. Gebiete*, 12:155–171. 8

Philipp, W. (1969b). Das gesetz vom iterierten logarithmus mit anwendungen auf die zahlentheorie. *Mathematische Annalen*, 180:74–94. 22

Philipp, W. (1969c). The law of the iterated logarithm for mixing stochastic processes. *Ann. Math. Stat.*, 40:1985–1991. 22

Philipp, W. (1977). A functional law of the iterated logarithm for empirical distribution functions of weakly dependent random variables. *Ann. Prob.*, 5:319–350. 22

Phillip, W. et Berkes, W. (1978). Approximation theorems for independent and weakly dependent random variables. *The Annals of Probability*, 7:29–54. 22

Reznik, M. K. (1968). The law of the iterated logarithm for some classes of stationary processes. *Theory Prob. Appl.*, 8:606–621. 22

Rhomari, N. (2002). Approximation et inégalités exponentielles pour les sommes de vecteurs aléatoires dépendants. *C. R. Math. Acad. Sci. Paris*, 2:149–154. 10

Rio, E. (1995a). About the lindeberg method for strongly mixing sequences. *ESAIM Probab. Statist.*, 1:35–61. 10

Rio, E. (1995b). The functional law of iterated logarithm for stationary, strongly mixing processes. *Ann. Prob.*, 23:1188–1203. 22

Rio, E. (2000). *Théorie asymptotique des processus aléatoires faiblement dépendants.* Mathématiques & Applications Springer- Verlag, Berlin. 10, 68, 69

Robinson, P. (1983). Nonparametric estimators for time series. *J. Time Series Anal*, 4(3):185–207. 15

Rosenblatt, M. (1956). A central limit theorem and a strong mixing condition. *Proc. Nat. Acad. Sci. U. S. A.*, 42:43–47. 8, 9, 14, 17, 60

Roussas, G. (1988). Nonparametric estimation in mixing sequences of random variables. *J. Statist. Plann.*, 18:135–149. 15

Roussas, G. (1990). Asymptotic normality of the kernel estimate under dependence conditions : application to hazard rate. *J. Statist. Plann.*, 25:81–104. 15

Rudemo, M. (1982). Emperical choice of histograms and kernel density estimators. *Scand. j. Stat.*, 9:65–78. 15

Rûschendorf, L. (1977). Consistency of estimators for multivariate density functions and for the mode. *Sankhya*, 39:243–250. 15

Sarda, P. et Vieu, P. (1989). Empirical distribution function for mixing random variables. *Statistics*, 20:559–571. 15

Silverman, B. et Jones, M. (1989). an important contribution to nonparametric discriminant analysis and density estimation. *Int. Stat. Rev.*, 57(3):233–247. 13

Singpurwalla, N. et Wong, M. (1983). Estimation of the failure rate : a survey of nonparametric models. part *i* : Non-bayesian methods. *Comm. Statist. Theory Methods*, 12:559–588. 16

Smirnov, N. (1939). Sur les écarts de la courbe de distribution empirique. *Recueil Mathématique [Matematicheskii Sbornik]*, 6(48)(1):3–26. 22

Steinhaus, H. (1922). Les probabilités dénombrables et leur rapport à la théorie de la mesure. *Fund. Math.*, 4:286–310. 21

Stute, W. et Wang, J.-L. (1993). The strong law under random censorship. *The Annals of Statistics*, 21(3):1591–1607. 17

Tran, L. (1990). Kernal density estimation under dependence. *Statist. Probab. Lett.*, 10:193–201. 15

Turnbull, B. W. (1974). Nonparametric estimation of a survivorship function with doubly censored data. *Journal of the American Statistical Association.*, 69(345):169–173. 2, 66

Vieu, P. (1999). Multiple kernel procedure : an asymptotic support. *Scand. J. of Statist.*, 26:61–72. 15

Volkonskii, V. et Rozanov, Y. (1959). Some limit theorems for random functions i. *Theor. Probab. Appl.*, 4:178–197. 8

Withers, C. (1981). Conditions for linear processes to be strong mixing. *Z Wahrsch. Verw. Gebiete*, 4:477–480. 10

Xiang, X. (1994). Law of the logarithm for density and hazard rate estimation for censored data. *J. Multivariate Anal.*, 49:278–286. 18

Ying, Z. et Wei, L. J. (1994). The kaplen-meier for dependent failure time observations. *Journal of Multivariate Analysis*, 50:17–29. 17

Yoshihara, K. (2004). *Weakly dependent stochastic sequences and their applications.* Vol. XIV. Recent topics on weak and strong limit theorems. Sanseido Co., Ltd., Chiyoda-ku. 10

Yu, B. (1993). Density estimation in the l^∞ norm for dependent data with applications to the gibbs sampler. *Ann. Statist.*, 21:711–735. 15

ملخّص

دراسة الأعمال الحاليّة تُظهر أنّ العديد من النّتائج التقريبيّة التي تمّ الحصول عليها في إطار الإحصاء الغير مقياسي في حالة البيانات المحجوبة على اليمين ، تستند على خصائص مقدّر كابلان و ماير (١٩٥٨)، الذي يقدّر دالة التّوزيع . لذلك، هذا المقدّر الذي تمّ تعميمه من طرف باتيليا و رولان في ٢٠٠٦ إلى حالة الحجب المزدوج، بدا من المثير دراسة خصائص هذا الأخير (مقدّر باتيليا و رولان) ، هذا هو الهدف الرّئيسي لهذه الأطروحة. بتعبير أدقّ ، نحن مهتمّون بهذا النّموذج من الحجب مع عمليّات قويّة الإختلاط. في هذا الإطار، وبعد استنتاج قانون اللّوغارتم المكرّر لمقدّر باتيليا و رولان، نثبت التّقارب المنتظم شبه الكامل لمقدّرات دالة التّوزيع مع تحديد سرعة هذا التّقارب، أوّلًا لدالة التّوزيع التّجريبية المستندة على عمليّات ألفا-مختلطة. ثمّ ، في حالة الحجب اليساري، وتحت نفس فرضيّة الإرتباط، نحدّد سرعة هذا التّقارب لمقّدر دالة التّوزيع (الذي يستنتج من ذلك الخاص بكابلان ماير عن طريق عكس الزّمن) . نستغل بعد ذلك هاتين النّتيجتين السّابقتين لالحصول على التّقارب شبه الكامل لمقدّر باتيليا و رولان بالإضافة إلى مقدّر النّواة لنسبة المجازفة، من أجل عيّنات ذات عمليّات ألفا-مختلطة. لدعم دراستنا النّظرية ، نقدّم دراسة محاكاة مصحوبة بتطبيق على بيانات حقيقيّة.

من جهة أخرى، بدءاً من نتيجة باتيليا و رولان، تّم إقتراح مقدّر لدالّة الكثافة لهذا النّموذج، بإستعمال طريقة النّواة، من قِبل كيتوني و آل. (٢٠١٤). تطبيق نتائجنا السّابقة يتيح لنا إذا متابعة دراسة هذا المقدّر الأخير تحت شرط الاختلاط القوي. نقدّم سرعة تقاربه شبه الكامل، مع تحديد، في نفس الإطار، ذلك الخاص بمقدّر النّواة لنسبة المجازفة. وتجدر الإشارة إلى أنّ المعدّلات المقترحة في هذه الأطروحة ، تحت شرط الإختلاط القوي، هي مماثلة لتلك التي تمّ الحصول عليها من قبل في ظلّ البيانات المستقلّة.

الكلمات المفتاحيّة : الخلط القوي، الحجب، تقارب شبه كامل، التّوزيع، الكثافة، نسبة المجازفة، المقدّرات غير المقياسيّة، مقدّرات النّواة، قانون اللّوغارتم المكرّر.

ABSTRACT

The study of existing work shows that many of the asymptotic results obtained in the context of nonparametric statistics for right censored observations are based on the properties of the Kaplan Meier estimator of the survival function. So, since this estimator was generalized by Patilea and Rolin [2006] to the case of the twice censorship model, it became interesting to study the properties of the last estimator (the Patilea-Rolin estimator), this is the main purpose of this thesis. More precisely, we are interested in this type of censorship with strong mixing processes. In this framework, after deducing the law of the iterated logarithm for the Patilea-Rolin estimator, we show the uniform almost complete convergence of the distribution function estimators, with rate, first for the empirical distribution function based on α-mixing data. Then, in the case of left censorship, and under the same hypothesis of dependence, we specify the rate of this convergence for the estimator of the distribution function (which is deduced from that of Kaplan-Meier by inverting the time). We then exploit these two previous results to obtain the rate of the almost complete convergence of the Patilea-Rolin estimator as well as the kernel estimator of the cumulative failure rate, based on α- mixing data. To support our theoretical study, we present a simulation study accompanied by an application on real data.
Starting from the result of Patilea and Rolin [2006], the kernel estimation of the density function for this model, was proposed by Kitouni *et al.* [2015]. Based on our previous results, we then continue the study of this last estimator under the condition of strong mixing. We establish its rate of the uniform almost complete convergence as well as that of the kernel failure rate estimator. It should be noted that the rates proposed in this thesis, under the condition of the strong mixing, are identical to those obtained for independent data.
Keywords : α-mixing, censorship model, almost complete convergence, distribution, density, failure rate, nonparametric estimators, kernel estimators, law of the iterated logarithm.

RÉSUMÉ

L'étude des travaux existants montre que beaucoup de résultats asymptotiques obtenus dans le cadre de la statistique non paramétrique pour des observations censurées à droite, se basent sur les propriétés de l'estimateur de Kaplan Meier qui estime la fonction de survie. Par conséquent, cet estimateur ayant été généralisé par Patilea et Rolin [2006] au cas de la censure mixte, il est apparu intéressant d'étudier les propriétés de ce dernier (estimateur de Patilea-Rolin), c'est l'objet principal de cette thèse. Plus précisément, nous nous intéressons à ce type de censure avec des processus fortement mélangeants. Dans ce cadre, après avoir établi la loi du logarithme itéré pour l'estimateur de Patilea-Rolin, nous montrons la convergence uniforme presque complète de l'estimateur de la fonction de répartition tout en précisant le taux de cette convergence, d'abord pour la fonction de répartition empirique basée sur un processus α-mélangeant. Puis, dans le cas de la censure à gauche, et sous la même hypothèse de dépendance, nous précisons le taux de cette convergence pour l'estimateur de la fonction de répartition (qui peut se déduire de celui de Kaplan-Meier en inversant le temps). Nous exploitons ensuite ces deux derniers résultats pour obtenir la convergence presque complète de l'estimateur de Patilea-Rolin ainsi que l'estimateur à noyau du taux de hasard cumulé, pour des échantillons constitués d'observations α-mélangeantes. Afin d'appuyer notre étude théorique, nous présentons une étude de simulation accompagnée d'une application sur des données réelles.
Par ailleurs, partant du résultat de Patilea et Rolin [2006], l'estimation à noyau de la fonction de densité pour ce modèle, a été proposée par Kitouni *et al.* [2015]. L'application de nos résultas précédents nous permet alors de poursuivre l'étude de ce dernier estimateur sous la condition de mélange fort. Nous établissons son taux de convergence uniforme presque complète, tout en précisant, dans le même cadre, celui de l'estimateur à noyau du taux de hasard. Il est à noter que tous les taux obtenus dans cette thèse, sous la condition du mélange fort, sont identiques à ceux précédemment établis pour des données indépendantes.
Mots-clés : α-mélange, censure, convergence presque complète, distribution, densité, taux de hasard, estimateurs non paramétriques, estimateurs à noyau, loi du logarithme itéré.