République algérienne démocratique et populaire
Ministère de l'enseignement supérieur et de la recherche scientifique
==========
Université des frères Mentouri, Constantine
Faculté des sciences exactes
==========
Département de mathématiques

N° d'ordre : 64/DS/2015
N° de série : 03/MATH/2015

THÈSE

PRÉSENTÉE POUR L'OBTENTION
DU
DIPLÔME DE DOCTORAT EN SCIENCES
EN
MATHÉMATIQUES

« Processus empiriques dans un modèle de censure et estimation fonctionnelle »

Par
Kitouni Abderrahim

OPTION
Probabilités et Statistique

Devant le jury :

Président	M.	Z. Mohdeb	Prof.	Université des frères Mentouri
Directrice de thèse	M^{me}	F. Messaci	Prof.	Université des frères Mentouri
Examinatrice	M^{me}	O. Sadki	Prof.	Université d'Oum El Bouaghi
Examinatrice	M^{me}	Z. Guessoum	M.C. A.	U.S.T.H.B., Alger
Examinateur	M.	A. Tatachak	M.C. A.	U.S.T.H.B., Alger

Soutenue le : 14 juin 2015

Remerciements

Je tiens en premier lieu à remercier sincèrement ma directrice de thèse, madame F. Messaci pour toue l'aide qu'elle m'a apportée, pour tout le temps qu'elle m'a consacré et l'extraordinaire dévouement qu'elle a dans son travail avec moi et avec tous ses étudiants.

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

Merci infiniment à madame Z. Guessoum et à madame O. Sadki, pour l'honneur qu'elles me font d'examiner mon travail, et d'avoir accepté de faire le déplacement malgré leurs multiples responsabilités.

Je suis très honoré et très reconnaissant que monsieur A. Tatachak ait accepté d'examiner mon travail, et accepté de faire le déplacement malgré ses multiples responsabilités qui laissent peu de temps.

Table des matières

Introduction

L'estimation de la fonction de répartition, qui caractérise complètement la loi de probabilité de toute variable aléatoire, est fondamentale en statistique. Elle permet, par exemple, d'estimer la probabilité d'appartenance à un intervalle. Pour les variables aléatoires à densité, cette dernière présente un avantage visuel puisqu'elle permet de mettre en évidence certaines caractéristiques de la loi comme le (ou les) mode(s), les creux et les asymétries. Si l'estimation de la fonction de densité est en elle même intéressante, elle sert aussi à estimer le taux de hasard (ou de panne) qui représente mieux les caractéristiques recherchées en analyse de survie et en fiabilité. Il n'est donc pas étonnant qu'une littérature riche et abondante ait été consacrée à l'estimation des ces trois fonctions, et notre travail s'inscrit dans ce cadre.

Sous des conditions restrictives sur le modèle sous-jacent, l'estimation peut se faire par des méthodes paramétriques. Les méthodes non paramétriques que nous adoptons dans cette thèse sont plus flexibles du fait qu'elles ne supposent aucune forme pour la loi à estimer.

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.

En analyse de survie et en fiabilité, les variables aléatoires d'intérêt représentent une durée : temps qui s'écoule jusqu'à la réalisation d'un certain événement. Ce temps est appelé temps de défaillance, durée de vie ou durée de survie, et se caractérise par la présence d'observations incomplètes. Le cas d'incomplétude le plus courant et le plus étudié aussi est la censure à droite. Il y a censure à droite lorsque la durée de survie d'intérêt est supérieure à la durée observée. C'est le cas par exemple, dans des études de fiabilité lorsque la panne d'un appareil ne permet pas de poursuivre l'observation de l'appareil objet de l'étude. La censure à droite n'est pas la seule censure que l'on peut rencontrer avec des données de survie, beaucoup d'autres mécanismes peuvent causer des censures diverses.

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

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

Dans ce contexte, Patilea et Rolin (2006) donnent un estimateur produit-limite de la fonction de survie de la durée d'intérêt X qui généralise le célèbre estimateur de Kaplan et Meier (1958) et démontrent sa convergence uniforme presque sûre ainsi que sa normalité asymptotique sous certaines conditions d'identifiabilité. Messaci et Nemouchi (2011) précisent la vitesse de cette convergence. D'autres travaux ont ciblé ce modèle, citons Shen (2011, 2012) qui propose deux estimateurs de la fonction de survie alternatifs à celui de Patilea et Rolin (2006), et les travaux concernant l'estimation de la régression rapportés dans Messaci (2010), Kebabi *et al.* (2011) et Kebabi et Messaci (2012). Dans cette thèse nous démontrons la convergence presque complète uniforme, en précisant la vitesse, pour des estimateurs de la fonction de hasard et de la fonction de survie. Puis nous déduisons le taux de convergence pour des estimateurs à noyau de la densité et du taux de hasard introduit dans cette thèse.

Nous rappelons cette notion de convergence qui est légèrement plus forte que la convergence presque sûre et qui est bien moins connue et moins utilisée.

Convergence presque complète La notion de convergence presque complète a été introduite par Hsu et Robbins (1947). Elle implique la convergence presque sûre et la convergence en probabilité. On dit que la suite de variables aléatoires réelles $(X_n)_{n\in\mathbb{N}}$ converge presque complètement vers une variable aléatoire X lorsque $n \to \infty$ si

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

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

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

Cette définition du taux a été introduite par Ferraty et Vieu (2006). Elle implique les deux vitesses de convergence classiques en probabilité et presque sûre.

Puis en seconde partie, nous nous intéressons aux processus empiriques, dont la théorie joue un rôle important en statistique, puisqu'elle concerne l'ensemble des résultats limites généraux se rapportant aux échantillons aléatoires. En particulier, des lois du logarithme ont permis de montrer un grand nombre de résultats asymptotiques. Dans cette thèse nous donnons une loi fonctionnelle du logarithme itéré pour les accroissements du processus empirique dans un modèle de censure mixte. De plus, nous déduisons des lois fortes pour des estimateurs à noyau de la fonction de densité et du taux de hasard du temps de survie.

Organisation du document Cette thèse, comprend donc deux idées, la première porte sur l'estimation non paramétrique de la fonction de répartition, de la fonction de hasard, de la fonction de densité et du taux de hasard : c'est l'objet du chapitre 4. La deuxième porte sur les processus empiriques : c'est l'objet du chapitre 5. Le chapitre 1 est consacré au produit intégral et à l'équation de Duhamel, outils précieux du transfert des propriétés entre fonction de survie et fonction de hasard. Dans ce chapitre est détaillé le travail de Gill et Johansen (1990) et Gill (1994). Le chapitre 2 contient des notions préliminaires sur les processus empiriques et sur les données censurées. Le chapitre 6 réunit des études par simulations des différents estimateurs abordés.

Chapitre 1

Produit-intégral et équation de Duhamel

1.1 Introduction

La correspondance entre une fonction de survie et sa fonction de hasard (ou son taux de hasard) est une idée centrale en analyse de survie. Cette correspondance, comme cela a été montré dans Gill et Johansen (1990), est un cas particulier d'une correspondance plus générale entre mesures additives et mesures multiplicatives à valeurs matricielles réelles. L'intégrale additive de la fonction de survie donne la fonction de hasard, et l'intégrale multiplicative (appelée produit intégral) de la fonction de hasard donne la fonction de survie. La théorie du produit intégral est un outil mathématique très utile en analyse de survie. Cette théorie a une longue histoire dans les mathématiques pures et appliquées.

La définition du produit intégral est apparue pour la première fois dans les travaux de Vito Volterra à la fin du 19^{e} siècle. Le premier travail de Volterra consacré au produit intégral a été publié en 1887 et a été écrit en italien. Il introduit les deux concepts de base du calcul multiplicatif, à savoir la dérivée d'une fonction de matrice et le produit intégral. Le traitement final du produit intégral par Volterra est représenté par le livre *Opérations infinitésimales linéaires* écrit en collaboration avec un mathématicien tchèque, Bohuslav Hostinskỳ. Ce livre (Volterra et Hostinskỳ, 1938) est apparu dans la série *Collection de monographies sur la théorie des fonctions* dirigée par Émile Borel.

Une revue extensive incluant un peu de l'histoire du produit intégral est

donné dans Gill et Johansen (1990). Une autre revue avec beaucoup plus de références et d'applications mais prenant une approche différente est donnée par Dollard et Friedman (1984). Une théorie abstraite du produit intégral est donnée dans Mac Nerney (1963). Un des intérêts du produit intégral est de fournir le formalisme qui permet d'unifier facilement le cas discret et le cas continu. Notons la contribution intéressante dans ce sens de Helton (1966, 1975).

1.2 Motivation

Soit T une variable aléatoire positive de fonction de répartition F, représentant le temps d'occurrence d'un certain événement. La fonction de survie S est définie par la probabilité de survivre jusqu'à (et y compris au) temps t

$$S(t) = P(X > t),$$

et on a bien sûr $S = 1 - F$. La fonction de hasard est définie par

$$\Lambda(t) = \int_0^t \frac{F(ds)}{S(s^-)}.$$

Nous avons

$$\Lambda(t) = \lim \sum_i \left(1 - \frac{S(t_i)}{S(t_{i-1})}\right) = \sum_i P(T \leqslant t_i / T > t_{i-1}),$$

où $0 = t_0 < t_1 \cdots < t_n = t$ est une subdivision de $]0, t]$ et la limite est prise sur les subdivisions pour lesquelles $\max_i |t_i - t_{i-1}|$ converge vers zéro. En effet, pour tout i, $0 \leqslant i \leqslant n-1$, nous avons

$$\begin{aligned}\int_{t_i}^{t_{i+1}} \frac{F(ds)}{S(s^-)} - \left(1 - \frac{S(t_{i+1})}{S(t_i)}\right) &= \int_{t_i}^{t_{i+1}} \frac{F(ds)}{S(s^-)} - \int_{t_i}^{t_{i+1}} \frac{F(ds)}{S(t_i)} \\ &= \int_{t_i}^{t_{i+1}} \left(\frac{1}{S(s^-)} - \frac{1}{S(t_i)}\right) F(ds)\end{aligned}$$

et $|1/S(s^-) - 1/S(t_i)| \leqslant 2/S(t^-)$, qui est intégrable par rapport à $F(ds)$. Le résultat s'ensuit par le théorème de la convergence dominée.

On peut aussi considérer Λ comme une mesure, $\Lambda(dt) = F(dt)/S(t^-)$. Habituellement les intervalles de temps sont ouverts à gauche et fermés à droite, et dans ce cas on écrit $\Lambda(s, t) = \Lambda(]s, t]) = \Lambda(t) - \Lambda(s)$ pour le hasard total

de l'intervalle. Ce qui fait de Λ une fonction d'intervalles additive, autrement dit pour $s \leqslant t \leqslant u$, on a

$$\Lambda(s,u) = \Lambda(s,t) + \Lambda(t,u).$$

La fonction de survie S produit une autre fonction d'intervalles $\mathrm{S}(s,t)$ en posant pour $s \leqslant t$,

$$\mathrm{S}(s,t) = \frac{\mathrm{S}(t)}{\mathrm{S}(s)} = \mathrm{P}(\mathrm{T} > t/\mathrm{T} > s),$$

qui est la probabilité de survivre à la fin de l'intervalle $]s,t]$ sachant que l'on a survécu à sa borne inférieure s. Nous avons pour $s \leqslant t \leqslant u$

$$\mathrm{S}(s,u) = \mathrm{S}(s,t)\mathrm{S}(t,u),$$

ce qui veut dire que S est une fonction d'intervalles multiplicative.

1.3 Produit-intégral

Dans la suite de ce chapitre α et μ représentent deux fonctions d'intervalles telles que α est additive et μ est multiplicative, à valeurs matricielles $p \times p$. La matrice identité est notée 1 et la matrice des zéros 0.

Le cas particulier $p = 1$, et $\alpha \geqslant 0$ ou $\mu \geqslant 1$, est appelé « cas réel positif » et nous écrivons alors α_0 et μ_0 au lieu de α et μ pour le mettre en évidence.

Nous voulons montrer la dualité

$$\mu = \prod(1 + d\alpha), \quad \alpha = \int (d\mu - 1).$$

et tirer d'autres propriétés du produit intégral.

L'approche que nous considérons consiste à étudier d'abord le cas réel positif, pour lequel les résultats s'obtiennent simplement par la monotonie. Puis nous montrons comment le cas général en découle en utilisant la propriété de domination et quelques identités algébriques sur les sommes et les produits.

Pour la relation entre la fonction de survie et la fonction de hasard, on a $\alpha = -\Lambda$ et $\mu = \mathrm{S}$. Alors $\alpha \leqslant 0$ et $\mu \leqslant 1$; nous ne sommes donc pas dans le cas réel positif, bien que Λ et S soient scalaires.

Identités algébriques fondamentales Le lemme suivant donne cinq relations qui, lorsqu'elles sont généralisées au produit continu, donneront des propriétés fondamentales du produit intégral. En particulier, la relation (1.4) donnera l'équation de Duhamel, qui est un outil puissant pour exprimer la différence entre deux produits-intégral en fonction de la différence des intégrandes.

Lemme 1.1. *Soit* $A_1, \dots, A_n$ *et* $B_1, \dots, B_n$ *des matrices* $p \times p$*, alors*

$$\prod_{j=1}^{n}(1+A_j)-1=\sum_{j=1}^{n}\left(\prod_{i=1}^{j-1}(1+A_i)\right)A_j, \tag{1.1}$$

$$\prod_{j=1}^{n}(1+A_j)-1=\sum_{j=1}^{n}A_j\left(\prod_{k=j+1}^{n}(1+A_k)\right), \tag{1.2}$$

$$\prod_{i=1}^{n}(1+A_i)-1-\sum_{i=1}^{n}A_i=\sum_{k=1}^{n}\sum_{i=1}^{k-1}A_i\left(\prod_{j=i+1}^{k-1}(1+A_j)\right)A_k, \tag{1.3}$$

$$\prod_{j=1}^{n}(1+A_j)-\prod_{j=1}^{n}(1+B_j)=\sum_{j=1}^{n}\left(\prod_{i=1}^{j-1}(1+A_i)(A_j-B_j)\prod_{k=j+1}^{n}(1+B_k)\right), \tag{1.4}$$

$$\prod_{i=1}^{n}(1+A_i)=1+\sum_{m=1}^{n}\sum_{i_1<i_2<\cdots<i_m}A_{i_1}\dots A_{i_m}, \tag{1.5}$$

où une somme vide est égale à 0 et un produit vide est égal à 1.

Démonstration.

— (1.1) Découle de (1.4) si on prend tous les B_j égaux à la matrice nulle.

— (1.2) Découle de (1.4) si on prend tous les A_j égaux à la matrice nulle.

— De (1.2) nous avons

$$\prod_{j=1}^{n}(1+A_j)=\sum_{j=1}^{n}A_j\left(\prod_{k=j+1}^{n}(1+A_k)\right)+1.$$

En remplaçant dans (1.1) nous obtenons

$$\begin{aligned}\prod_{j=1}^{n}(1+A_j)-1&=\sum_{j=1}^{n}\left(\sum_{i=1}^{j-1}A_i\left(\prod_{k=i+1}^{j-1}(1+A_k)\right)+1\right)A_j\\&=\sum_{j=1}^{n}\sum_{i=1}^{j-1}A_i\left(\prod_{k=i+1}^{j-1}(1+A_k)\right)A_j+\sum_{j=1}^{n}A_j,\end{aligned}$$

d'où le résultat (1.3).

—
$$\begin{aligned}
&\sum_j \left(\prod_{i<j} (1+A_i)(A_j - B_j) \prod_{k>j} (1+B_k) \right) \\
&\quad = \sum_j \left(\prod_{i<j} (1+A_i) \left((1+A_j) - (1+B_j) \right) \prod_{k>j} (1+B_k) \right) \\
&\quad = \sum_j \left(\prod_{i\leqslant j} (1+A_i) \prod_{k>j} (1+B_k) - \prod_{i<j} (1+A_i) \prod_{k\geqslant j} (1+B_k) \right) \\
&\quad = \prod_i (1+A_i) - \prod_k (1+B_k)
\end{aligned}$$

qui est (1.4).

— Pour $n = 2$, nous avons $(1+A_1)(1+A_2) = 1 + A_1 + A_2 + A_1A_2$. Ce résultat peut être généralisé par récurrence pour obtenir (1.5). □

Maintenant considérons les fonctions d'intervalles α (additive) et μ (multiplicative), toutes les deux continues à droite :

$$\begin{aligned}
&\alpha(s,t) \to \alpha(s,s) = 0 \quad \text{quand } t \downarrow s, \\
&\mu(s,t) \to \mu(s,s) = 1 \quad \text{quand } t \downarrow s.
\end{aligned}$$

Notons α_0 (respectivement μ_0) une fonction d'intervalles additive (respectivement multiplicative), continue à droite et vérifiant $\alpha_0 \geqslant 0$ (respectivement $\mu_0 \geqslant 1$). Nous supposons que α est dominée par α_0 et que $\mu - 1$ est dominée par $\mu_0 - 1$, c'est-à-dire que $\|\alpha\| \leqslant \alpha_0$ et $\|\mu - 1\| \leqslant \mu_0 - 1$ où $\|A\|$ est la norme matricielle $\max_i \sum_j |a_{ij}|$. Nous avons donc aussi

$$\|A + B\| \leqslant \|A\| + \|B\|, \quad \|AB\| \leqslant \|A\|.\|B\|, \quad \|1\| = 1.$$

Soit $]s,t]$ un intervalle de temps fixé et $\mathcal{T} = \{s = t_0 < t_1 \cdots < t_n = t\}$ une subdivision de $]s,t]$, et notons pour toute fonction d'intervalles f

$$\sum_{\mathcal{T}} f = \sum_{i=1}^{n} f(t_{i-1}, t_i) \qquad et \qquad \prod_{\mathcal{T}} f = \prod_{i=1}^{n} f(t_{i-1}, t_i).$$

Regardons les inégalités suivantes

$$\begin{aligned}
1 + a + b \leqslant (1+a)(1+b) \leqslant \exp(a+b) \qquad & a, b \geqslant 0, \\
\log(xy) \leqslant (x-1) + (y-1) \leqslant xy - 1 \qquad & x, y \geqslant 1.
\end{aligned}$$

La première inégalité montre que $\prod_{\mathcal{T}} (1 + \alpha_0)$ est majoré par $\exp \alpha(s,t)$ et croît avec le raffinement de la subdivision $\mathcal{T}$. De même, $\sum_{\mathcal{T}} (\mu_0 - 1)$ est

minoré par $\log \mu_0(s,t)$ et décroît avec le raffinement de la subdivision. Nous pouvons donc définir

$$\prod_{]s,t]}(1+d\alpha_0) = \lim_{\mathcal{T}} \prod_{\mathcal{T}}(1+\alpha_0) \tag{1.6}$$

$$\int_{]s,t]}(d\mu_0 - 1) = \lim_{\mathcal{T}} \sum_{\mathcal{T}}(\mu_0 - 1) \tag{1.7}$$

où les limites sont prises sur des subdivisions de plus en plus fines de $]s,t]$.

Proposition 1. *Pour α_0 donnée posons $\mu_0 = \prod(1+d\alpha_0)$. Alors $\mu_0 \geqslant 1$ est une fonction d'intervalles multiplicative, continue à droite et vérifie $\alpha_0 = \int (d\mu_0 - 1)$. Réciproquement, pour μ_o donnée posons $\alpha_0 = \int (d\mu_0 - 1)$. Alors $\alpha_0 \geqslant 0$ est une fonction d'intervalles additive, continue à droite et vérifie $\mu_0 = \prod(1+d\alpha_0)$.*

Démonstration. Les bornes suivantes sont faciles à vérifier : pour α_0 donnée, $\mu_0 = \prod(1+d\alpha_0)$ vérifie

$$\exp(\alpha_0) - 1 \geqslant \mu_0 - 1 \geqslant \alpha_0 \geqslant 0. \tag{1.8}$$

De même, pour μ_o donnée, $\alpha_0 = \int (d\mu_0 - 1)$ vérifie

$$0 \leqslant \log \mu_0 \leqslant \alpha_0 \leqslant \mu_0 - 1. \tag{1.9}$$

La continuité à droite, la multiplicativité et l'additivité sont facile à démontrer.

Soit α_0 donnée et posons $\mu_0 = \prod(1+\alpha_0)$. Soit α_j et μ_j les valeurs de α_0 et de μ_0 sur les intervalles de la subdivision $\mathcal{T}$. Alors

$$\begin{aligned}
0 \leqslant \sum_{\mathcal{T}}(\mu_0 - 1) - \alpha_0(s,t) &= \sum_j (\mu_j - 1 - \alpha_j) \\
&\leqslant \sum_j \prod_{i<j}(1+\alpha_i)(\mu_j - 1 - \alpha_j)\prod_{k>j}\mu_k \qquad (\text{car } 1+\alpha_i \geqslant 1 \text{ et } \mu_k \geqslant 1) \\
&= \prod_j \mu_j - \prod_j (1+\alpha_j) \qquad (\text{d'après (1.4)}) \\
&= \mu_0(s,t) - \prod_{\mathcal{T}}(1+\alpha_0). \qquad (\text{par la multiplicativité de } \mu_0)
\end{aligned}$$

Comme $\prod_{\mathcal{T}}(1+\alpha_0) \to \mu_0(s,t)$, nous obtenons $\sum_{\mathcal{T}}(\mu_0 - 1) \to \alpha_0(s,t)$.

Réciproquement, soit $\mu_0 \geqslant 1$ donnée et posons $\alpha_0 = \int (d\mu_0 - 1)$.

$$\begin{aligned}
0 &\leqslant \mu_0(s,t) - \prod_{\mathcal{T}}(1+\alpha_0) \\
&= \prod_j \mu_j - \prod_j (1+\alpha_j) \\
&= \sum_j \prod_{i<j} \mu_j(\mu_j - 1 - \alpha_j) \prod_{k>j}(1+\alpha_k) && \text{(d'après (1.4))} \\
&\leqslant \sum_j \prod_{i<j} \mu_i(\mu_j - 1 - \alpha_j) \prod_{k>j} \mu_k && \text{(d'après (1.9))} \\
&\leqslant \sum_j \prod_i \mu_i(\mu_j - 1 - \alpha_j) && \text{(car } \mu_j \geqslant 1\text{)} \\
&\leqslant \mu_0(s,t) \left(\sum_{\mathcal{T}} (\mu_0 - 1) - \alpha_0(s,t) \right).
\end{aligned}$$

Comme $\sum_{\mathcal{T}}(\mu_0 - 1) \to \alpha_0$ nous obtenons $\mu_0(s,t) = \lim_{\mathcal{T}} \prod_{\mathcal{T}} (1+\alpha_0)$. □

Remarque. La preuve de la dualité établit la suite suivante d'inégalités, qui donne une idée de pourquoi la dualité à lieu :

$$\begin{aligned}
0 &\leqslant \sum_{\mathcal{T}} (\mu_0 - 1) - \alpha_0(s,t) \\
&\leqslant \mu_0(s,t) - \prod_{\mathcal{T}} (1+\alpha_0) \\
&\leqslant \mu_0(s,t) \left(\sum_{\mathcal{T}} (\mu_0 - 1) - \alpha_0(s,t) \right)
\end{aligned} \tag{1.10}$$

Théorème 1.1. *Soit α additive, continue à droite et dominée par α_0. Alors*

$$\mu = \prod (1 + d\alpha) = \lim_{\mathcal{T}} \prod_{\mathcal{T}} (1+\alpha)$$

existe, est multiplicative, continue à droite, et $\mu - 1$ est dominée par $\mu_0 - 1$ où $\mu_0 = \prod(1 + d\alpha_0)$.

Réciproquement si μ est multiplicative, continue à droite, et $\mu - 1$ est dominée par $\mu_0 - 1$, alors

$$\alpha = \int (d\mu - 1) = \lim_{\mathcal{T}} \sum_{\mathcal{T}} (\mu - 1)$$

existe, est additive, continue à droite et est dominée par $\alpha_0 = \int (d\mu_0 - 1)$.

De plus, $\mu = \prod(1 + d\alpha)$ si et seulement si $\alpha = \int (d\mu - 1)$.

Démonstration. Soit $\mathcal{S}$ un raffinement de $\mathcal{T}$. Notons α_i et μ_j les valeurs de α et de μ sur les éléments de $\mathcal{T}$. Soit $\mathcal{T}_j$ la subdivision du i-ème intervalle de $\mathcal{T}$ induite par $\mathcal{S}$, et soit α_{ij} les valeurs de α sur cette subdivision.

Pour α donné, remarquons que (à l'aide, en particulier, de (1.3) et (1.4) du lemme 1.1)

$$\begin{aligned}\prod_{\mathcal{S}}(1+\alpha) - \prod_{\mathcal{T}}(1+\alpha) &\\ &= \prod_j \left(\prod_{\mathcal{T}_j}(1+\alpha) \right) - \prod_j (1+\alpha_j) \\ &= \sum_j \prod_{i<j} \prod_{\mathcal{T}_i} (1+\alpha) \left(\prod_{\mathcal{T}_j}(1+\alpha) - 1 - \alpha_j \right) \prod_{k>j}(1+\alpha_k)\end{aligned}$$

$$= \sum_j \prod_{i<j} \prod_{\mathcal{T}_i} (1+\alpha) \left(\sum_{l,n:l<n} \alpha_{jl} \prod_{m,l:l<m<n} (1+\alpha_{jm})\alpha_{jn} \right) \prod_{k>j}(1+\alpha_j)$$

La dernière ligne de cette suite d'égalités est une somme de produits de α_{ij} et de α_i ; sa norme est majorée par la même expression en remplaçant les matrices par leurs normes, qui sont bornées par α_{0ij} et α_{0i}. Mais cette suite d'égalités est aussi valable pour α_0 elle-même. Nous avons donc montré que

$$0 \leqslant \left\| \prod_{\mathcal{S}}(1+\alpha) - \prod_{\mathcal{T}}(1+\alpha) \right\| \leqslant \prod_{\mathcal{S}}(1+\alpha) - \prod_{\mathcal{T}}(1+\alpha).$$

Par conséquent l'existence du produit intégral de α_0 implique l'existence du produit intégral de α. En outre, si on prend pour subdivision $\mathcal{T}$ la subdivision triviale (ayant pour seul élément $]s,t]$) et on rend $\mathcal{S}$ de plus en plus fine, on obtient que $\prod(1+d\alpha)-1-\alpha$ est dominée par $\prod(1+d\alpha_0)-1-\alpha_0$.

De même, pour μ donnée telle que $\mu - 1$ est dominée par $\mu_0 - 1$, nous

constatons que (en utilisant (1.3) du lemme 1.1)

$$\begin{aligned}\sum_{\mathcal{T}}(\mu-1)-\sum_{\mathcal{S}}(\mu-1) &= \sum_i(\mu_i-1)-\sum_i\sum_{\mathcal{T}_i}(\mu-1)\\ &= \sum_i\left(\mu_i-1-\sum_{\mathcal{T}_i}(\mu-1)\right)\\ &= \sum_i\left(\prod_{\mathcal{T}_i}(1+(\mu-1))-1-\sum_{\mathcal{T}_i}(\mu-1)\right)\\ &= \sum_i\left(\sum_{j,l:j<l}(\mu_{ij}-1)\prod_{k:j<k<l}\mu_{ik}(\mu_{il}-1)\right).\end{aligned}$$

Nous avons $\|\mu-1\| \leqslant \mu_0-1$, donc la norme de la dernière ligne est bornée par la même expression en μ_0. L'existence de la somme intégrale $\int(d\mu_0-1)$ implique donc l'existence de $\int(d\mu-1)$. Ici aussi si on prend $\mathcal{T}$ égale à la subdivision triviale et $\mathcal{S}$ de plus en plus fine, on obtient que $\mu-1-\int(d\mu-1)$ est dominée par $\mu_0-1-\int(d\mu_0-1)$.

Pour α donnée la domination de $\mu-1$ par μ_0-1 et pour μ donnée la domination de α par α_0 sont toutes les deux obtenues facilement. Ceci implique que si $\mu=\prod(1+d\alpha)$ avec α dominée par α_0 alors $\int(d\mu-1)$ existe, et il en est de même si on commence par μ avec $\mu-1$ dominée par μ_0-1.

Il reste à montrer que $\mu=\prod(1+d\alpha)$ si et seulement si $\alpha=\int(d\mu-1)$. Dans les deux directions nous avons que $\mu-1-\alpha$ est dominée par $\mu_0-1-\alpha_0$. Pour l'implication directe, remarquons que $\sum_{\mathcal{T}}(\mu-1)-\alpha=\sum_{\mathcal{T}}(\mu-1-\alpha)$, qui est dominée par $\sum_{\mathcal{T}}(\mu_0-1-\alpha_0)$ et par passage à la limite sur les raffinements de $\mathcal{T}$ nous obtenons $\alpha=\int(d\mu_0-1)$. Réciproquement supposons avoir $\alpha=\int(d\mu_0-1)$. Alors $\mu=\prod_{\mathcal{T}}(1+\alpha)=\sum_j\prod_{i<j}\mu_i(\mu_j-1-\alpha_j)\prod_{k>j}(1+\alpha_k)$ qui est dominée par la même expression en μ_0 et α_0. Le passage à la limite donne le résultat. □

La notion de domination a une interprétation en terme de théorie de la mesure, proche de l'habituel notion de variation bornée. Nous disons qu'une fonction d'intervalles β (éventuellement à valeurs matricielles) est à variation bornée si et seulement si sa variation, la fonction d'intervalles $|\beta|$ définie par $|\beta|=\sup_{\mathcal{T}}\sum_{\mathcal{T}}\|\beta\|$ est finie et continue à droite. Le sup est pris sur toutes les subdivision d'un intervalle donné. Il est assez facile de vérifier que β est à variation bornée si et seulement si β est bornée par une fonction

d'intervalles additive continue à droite α_0. La condition suffisante est triviale. La condition nécessaire découle en posant $\alpha_0(s,t) = |\beta|(0,t) - |\beta|(0,s)$. Alors $|\beta|(0,t) \geqslant |\beta|(0,s) + \|\beta(s,t)\|$ qui nous donne comme requis que $\|\beta\| \leqslant \alpha_0$. Le résultat suivant concernant les fonctions d'intervalles multiplicatives et aussi utile.

Proposition 2. *$\mu - 1$ est dominée par $\mu_0 - 1$ si et seulement si $\mu - 1$ est à variation bornée.*

Démonstration. $\mu - 1$ est à variation bornée implique $\mu - 1$ est dominé par une fonction additive d'intervalles α_0 ce qui implique que $\mu - 1$ est dominé par $\mu_0 - 1 = \prod(1 + d\alpha_0) - 1 \geqslant \alpha_0$.

Réciproquement, $\mu - 1$ est dominée par $\mu_0 - 1$ implique $\sum_{\mathcal{T}} \|\mu - 1\| \leqslant \sum_{\mathcal{T}} (\mu - 1)$. Mais cette dernière somme diminue avec les raffinements ; par conséquent, elle est finie (bornée par $\mu - 1$ elle-même) et $\mu - 1$ est à variation bornée. □

1.4 Quelques équations intégrales

Soit a_j et b_j les valeurs de deux fonctions additives d'intervalles α et β, dominées et continues à droite, sur le j-ème intervalle d'une subdivision $\mathcal{T}$ d'un intervalle $]s,t]$ donné. Soit $\mathcal{T}$ une subdivision dont le pas converge vers zéro. Dans les équations (1.1)–(1.5), les sommes peuvent être vues comme des intégrales par rapport aux mesures α et $\alpha - \beta$ d'une certaine fonction en escalier constante sur les sous-intervalles de la subdivision. En fait, puisque nous nous intéressons aux matrices, nous avons, composante par composante, des sommes finies de ces intégrales réelles, mais cela ne change pas le raisonnement.

Par notre résultat d'uniformité les intégrandes sont uniformément proches des produits-intégral de α ou de β, pris jusqu'à ou à partir d'une borne à partir d'un point d'extrémité de ce sous-intervalle de la subdivision sur lequel se fait l'intégration. La seule véritable difficulté est que (1.5) comprend une somme de plus en plus de termes. Cependant, le m-ième terme de la somme est bornée uniformément par le m-ième terme de la suite sommable $\alpha_0^m/m!$ et ne pose donc pas de problèmes. Tout cela signifie qu'on peut prendre la limite quand $|\mathcal{T}| \to 0$ dans (1.1)–(1.5) et obtenir les équations suivantes :

— l'équation intégrale progressive « forward integral equation »

$$\prod_{]s,t]}(1 + d\alpha) - 1 = \int_{]s,t]} \prod_{]s,u[}(1 + d\alpha)\alpha(du)\,; \tag{1.11}$$

— l'équation intégrale régressive « backward integral equation »

$$\prod_{]s,t]}(1+d\alpha)-1=\int_{]s,t]}\alpha(du)\prod_{]u,t]}(1+d\alpha)\,; \tag{1.12}$$

—

$$\prod_{]s,t]}(1+d\alpha)-1-\alpha(s,t)=\int_{s<u<v\leqslant t}\alpha(du)\prod_{]u,v[}(1+d\alpha)\alpha(dv)\,; \tag{1.13}$$

— l'équation de Duhamel

$$\prod_{]s,t]}(1+d\alpha)-\prod_{]s,t]}(1+d\beta)$$
$$=\int_{]s,t]}\prod_{]s,u[}(1+d\alpha)(\alpha(du)-\beta(du))\prod_{]u,t]}(1+d\beta)\,; \tag{1.14}$$

— la série de Peano

$$\prod_{]s,t]}(1+d\alpha)=1+\sum_{m=1}^{\infty}\int_{s<u_1<\cdots<u_m\leqslant t}\alpha(du1)\ldots\alpha(du_m). \tag{1.15}$$

Remarquons comment les produits-intégral à l'intérieur des intervalles ordinaires sont maintenant sur des intervalles comme $]s,u[$, $]u,v[$ ou $]v,t]$. Il est facile de déduire encore plus de relations à partir de (1.11)–(1.14).

Une équation que nous allons rencontrer dans la section suivante est obtenue à partir de l'équation (1.14) en la réécrivant comme

$$\prod(1+d\alpha+d\beta)=\prod(1+d\alpha)+\int\prod(1+d\alpha+d\beta)d\beta\prod(1+d\alpha),$$

et en remplaçant plusieurs fois $\prod(1+d\alpha+d\beta)$ dans le membre de droite. On voit alors apparaître les termes d'une suite infinie. Il est facile de montrer que le reste converge vers zéro, et nous obtenons une généralisation de la série de Peano :

$$\prod(1+d\alpha+d\beta)=\prod(1+d\alpha)+$$
$$\sum_{m=1}^{\infty}\int_{s<u_1<\cdots<u_m\leqslant t}\prod_{]s,u_1[}(1+d\alpha)\left(\prod_{i=1}^{m-1}\beta(du_i)\prod_{]u_i,u_{i+1}[}(1+d\alpha)\right)\beta(du_m)\prod_{]u_m,t[}(1+d\alpha) \tag{1.16}$$

Cette équation est en fait une forme de la formule du produit de Trotter de la théorie des semi-groupes (voir Masani, 1981). Si $\prod_{]s,u]}(1+d\alpha)$ est

inversible pour tout u, on peut remplacer chaque facteur $\prod_{]u_i,u_{i+1}[}(1+d\alpha)$ du terme de droite de (1.16) par $\left(\prod_{]s,u_i]}(1+d\alpha)\right)^{-1}\prod_{]s,u_{i+1}[}(1+d\alpha)$. En mettant en facteur (à droite) $\prod_{]s,t]}(1+d\alpha)$ nous obtenons la série de Peano ordinaire pour la mesure $\beta'(ds) = \prod_{]s,u[}(1+d\alpha)\beta(du)(\prod_{]s,u]}(1+d\alpha))^{-1}$; on obtient ainsi la formule de Trotter généralisée :

$$\prod_{]s,t]}(1+d\alpha+d\beta) = \prod_{u\in(s,t]}\left(1+\prod_{]s,u[}(1+d\alpha)\beta(du)(\prod_{]s,u]}(1+d\alpha)^{-1}\right)\prod_{]s,t]}(1+d\alpha).$$

Masani (1981) a mis en évidence l'analogie entre cette formule pour l'intégrale multiplicative d'une somme la formule usuelle d'intégration par parties d'un produit, bien qu'il travaille avec une définition différente du produit intégrale ($\prod \exp(d\alpha)$ plutôt que $\prod(1+d\alpha)$).

On peut considérer (1.11) et (1.12) comme des équations intégrales de Volterra en remplaçant le produits-intégrales dans les deux côtés par une fonction d'intervalles inconnue. Il s'avère que la solution est unique ; cela peut être prouvé par la méthode standard (considérer la différence de deux solutions, qui satisfait la même équation sans le (-1) et remplacer plusieurs fois le terme de gauche dans le terme de droite). Ainsi pour s donné, la solution unique β de

$$\beta(s,t) - 1 = \int_{]s,t]}\beta(s,u^-)\alpha(du) \tag{1.17}$$

est $\beta(s,t) = \prod_{]s,t]}(1+d\alpha)$, et pour t donné la solution unique β de

$$\beta(s,t) - 1 = \int_{]s,t]}\alpha(du)\beta(s,t) \tag{1.18}$$

est la même. Plus généralement si ψ est une fonction matricielle $q\times p$ càdlàd alors la solution unique ϕ de

$$\phi(t) = \psi(t) + \int_{]0,t]}\phi(s^-)\alpha(ds) \tag{1.19}$$

est

$$\phi(t) = \psi(t) + \int_{]0,t]}\psi(s^-)\alpha(ds)\prod_{]s,t]}(1+d\alpha). \tag{1.20}$$

1.5 Propriétés analytiques du produit intégral

Quand nous nous intéressons à des problèmes statistiques, nous devons nous demander si et éventuellement comment les propriétés statistiques

d'un estimateur de la fonction de hasard, comme la convergence, peuvent être transférées aux estimateurs correspondants de la fonctions de survie. Ces résultats ne dépendent souvent que de la continuité ou de la dérivabilité du produit intégral. L'équation de Duhamel conduit naturellement à certaines propriétés de continuité et de différentiabilité.

Soit $[0, \tau]$ un intervalle fixé et considérons les deux normes des fonctions d'intervalles continues à droite à valeurs matricielles, suivantes : la norme du sup

$$\|\beta\|_\infty = \sup_{s,t} \|\beta(s,t)\|;$$

et la norme de variation

$$\|\beta\|_V = \sup_{\mathcal{T}} \|\beta\| = \alpha_0(0,\tau)$$

où $\mathcal{T}$ parcourt l'ensemble des subdivisions de $[0, \tau]$ et α_0 est la plus petite mesure réelle dominant β. Notons $\xrightarrow{V}$ et $\xrightarrow{\infty}$ la convergence par rapport à ces deux normes.

Continuité Soit α et β deux fonctions d'intervalles additives, et soit $h = \beta - \alpha$. Considérons la différence

$$\begin{aligned}\prod(1 + d\beta) - \prod(1 + \mathrm{d}\alpha) &= \int \prod(1 + d\beta)(d\beta - d\alpha) \prod(1 + d\alpha) \\ &= \int \prod(1 + d\beta) dh \prod(1 + d\alpha). \qquad (1.21)\end{aligned}$$

Nous omettons les variables dans les intégrales et de produit-intégrale. Nous devons démontrer que cette différence est petite lorsque h est petit, en norme du sup. Utilisons l'intégration par parties. Nous remplaçons $\prod(1 + d\alpha)$ et $\prod(1 + d\beta)$ par des intégrales (les équations de Volterra) puis, utilisant le théorème de Fubini, nous inversons l'ordre d'intégration.

Grâce à (1.11) et (1.12) nous obtenons

$$\begin{aligned}&\int \prod(1 + d\beta) dh \prod(1 + d\alpha) \\ &\qquad = \int dh + \iint \prod(1 + d\beta) d\beta dh + \iint dh d\alpha \prod(1 + d\alpha)) \\ &\qquad\quad + \iiint \prod(1 + d\beta) d\beta dh d\alpha \prod(1 + d\alpha).\end{aligned}$$

Ensuite, nous pouvons inverser l'ordre de toutes les intégrations, et effectuer l'intégration par rapport à h avant celle par rapport à α ou β. Une intégration disparaît et on retrouve h comme une fonction d'intervalles :

$$\begin{aligned}\int \prod(1+d\beta)\,\mathrm{d}h \prod(1+d\alpha) = h &+ \int \prod(1+d\beta) d\beta\, h \\ &+ \int h d\alpha \prod(1+d\alpha)) \\ &+ \iint \prod(1+d\alpha)\,\mathrm{d}\alpha\, h\,\mathrm{d}\beta \prod(1+d\beta).\end{aligned} \tag{1.22}$$

Remarquons que cette identité ne dépend pas de la relation $h = \beta - \alpha$ entre α, β et h. Écrivons le dernier terme de (1.22) dans son intégralité :

$$\iint_{s<u<v\leqslant t} \prod_{s}^{u^-}(1+d\alpha)\alpha(du)h(u,v^-)\beta(dv) \prod_{v}^{t}(1+d\beta).$$

Remarquons que la bornitude de α et β pour la norme de variation implique la bornitude pour la norme du sup de leurs produits-intégral. Par conséquent si nous avons $h = \beta - \alpha$ alors de (1.21) on obtient :

$$\left\| \prod(1+d\beta) - \prod(1+d\alpha) \right\|_\infty \leqslant C\|h\|_\infty$$

Différentiabilité Nous allons maintenant montrer une différentiabilité continue au sens de Hadamard par rapport à la norme du sup, sous une condition de bornitude pour la norme de variation. La différentiabilité au sens de Hadamard, aussi appelée différentiabilité compacte, est une notion intermédiaire entre la différentiabilité au sens de Fréchet (ou différentiabilité bornée) et la différentiabilité au sens de Gâteaux (ou différentiabilité directionnelle). C'est exactement ce qu'il faut pour diverses applications statistiques ; en particulier, la méthode delta fonctionnelle a besoin de cette différentiabilité. Le résultat de différentiabilité pour le produit intégrale que nous donnons ici est dû à Gill et Johansen (1990).

Au lieu d'écrire $\beta = \alpha + h$ écrivons $\alpha + th$ où t est réel et proche de zéro. La différentiabilité au sens de Hadamard signifie que

$$\frac{1}{t}\left(\prod(1+d\beta) - \prod(1+d\alpha)\right)$$

peut être approchée, pour t petit, par une application linéaire continue de h, uniformément en h sur tout compact. Par différentiabilité compacte continue, nous entendons que l'approximation est également uniforme en α et β. La technique d'intégration par parties que nous avons utilisée pour la continuité constitue une partie de la démonstration. Pour la suite, nous avons besoin d'une autre technique, tirée de la preuve du théorème de Helly-Bray.

Avec $\beta = \alpha + th$, l'équation de Duhamel donne (voir (1.21)),

$$\frac{1}{t}\left(\prod(1+d\beta) - \prod(1+d\alpha)\right) = \int \prod(1+d\beta)dh\,\prod(1+d\alpha) \qquad (1.23)$$

Cette relation peut être réécrite de la même manière que (1.22) ; le membre de droite de cette dernière, considéré comme une application de l'ensemble des fonctions d'intervalles h muni de la norme du sup dans lui-même, est continue en h uniformément en α et β uniformément bornés pour la norme de variation. Nous pouvons donc définir $\int \prod(1+d\beta)dh\,\prod(1+d\alpha)$ pour des h qui ne sont pas à variation bornée par le membre de droite de (1.22).

Pour établir la différentiabilité continue au sens de Hadamard, nous devons montrer que $\prod(1+d\beta)dh\,\prod(1+d\alpha)$ est conjointement continu en α, β et h par rapport à la norme sup, pour α et β uniformément bornés pour la norme de variation.

Soit une suite de triplets (α_n, β_n, h_n) qui converge en norme du sup vers (α, β, h), et considérons la différence entre (1.22) au rang n et à la limite. Supposons que α_n et β_n (et donc aussi α, β) sont uniformément bornées pour la norme de variation. La technique de Helly-Bray consiste à insérer deux paires intermédiaires (α_n, β_n, h^*) et (α, β, h^*) tel que h^* est à variation bornée. Nous avons alors

$$\begin{aligned}
&\int \prod(1+d\beta_n)dh_n\,\prod(1+d\alpha_n) - \int \prod(1+d\beta)dh\,\prod(1+d\alpha)\\
&\quad= \int \prod(1+d\beta_n)(dh_n - dh^*)\,\prod(1+d\alpha_n)\\
&\qquad+ \left(\int \prod(1+d\beta_n)dh^*\,\prod(1+d\alpha_n) - \int \prod(1+d\beta)dh^*\,\prod(1+d\alpha)\right)\\
&\qquad+ \int \prod(1+d\beta)(dh^* - dh)\,\prod(1+d\alpha).
\end{aligned}$$

Dans le membre de droite, nous avons maintenant trois termes. L'intégration par parties du premier et du troisième termes (transformation en

quelque chose comme (1.22)) et l'hypothèse de variation bornée montrent que ces termes sont bornées en norme du sup par une constante multipliée par $\|h_n - h^*\|$ et $\|h - h^*\|$ respectivement. Le terme du milieu converge vers zéro quand $n \to \infty$ puisque le produit intégrale convergent en norme du sup et h^* est à variation bornée. Par conséquent, puisque $\|h_n - h^*\|_\infty \to \|h - h^*\|_\infty$ la limite supérieure de la norme du sup du membre de gauche est bornée par une constante multipliée par $\|h - h^*\|_\infty$, ce qui peut être rendu arbitrairement petit par le choix de h^*. Cela nous donne le résultat requis.

En résumé, nous avons obtenu le résultat de différentiabilité compacte continue suivant : pour $\alpha'_n = \alpha_n + t_n h_n$ avec $\alpha_n \overset{\infty}{\to} \alpha$, $h_n \overset{\infty}{\to} h$, $t_n \to 0$, et α_n et α à variation uniformément bornée, nous avons

$$\frac{1}{t_n}\left(\prod(1 + d\alpha'_n) - \prod(1 + d\alpha_n)\right) \overset{\infty}{\to} \int \prod(1 + d\alpha) dh \prod(1 + d\alpha) \quad (1.24)$$

où le membre de droite est une application linéaire de h continue en norme du sup, que l'on interprète comme dans (1.22) si h n'est pas à variation bornée. Elle est également conjointement continue en α et h.

Chapitre 2

Données censurées et processus empirique

2.1 Introduction aux données censurées

Dans de nombreuses applications statistiques, on est amené à étudier une variable de durée, c'est-à-dire un délai séparant un instant initial de l'instant où un événement d'intérêt est observé. C'est souvent le cas pour les applications médicales liées à l'étude des durées de survie. Pour cette raison, et bien que le champ d'application des méthodes que nous allons exposer ne se limite pas à la médecine ou à la biologie, nous utilisons la terminologie associées aux durées de survie. Celle-ci est la plus commune dans la littérature statistique, puisque c'est précisément dans le domaine médical que les avancées méthodologiques liées à ces variables ont été développées en premier.

Nous parlons de données censurées lorsque la durée de survie n'est connue que lorsqu'elle est dans les limites des durées d'observation. Ces limites pouvant être imposée par le type d'observation (par exemple, la durée d'hospitalisation d'un malade) ou par des événements fortuits (comme un accident ou une migration du patient au cours de son suivi). Par exemple, dans le cas dit de censure à droite, seul le minimum entre la durée de survie et une durée limite supérieure d'observation est connu, ainsi que l'indicateur exprimant si la durée de survie a été censurée ou non. Ce minimum constitue, en quelque sorte, une durée de participation qui est observée en ne donnant qu'une information partielle sur la vraie durée de survie. Il existe plusieurs autres catégories de modèles de censure, obtenus par des variations du principe de la censure à droite décrit ci-dessus. Parmi ceux-ci,

mentionnons les suivants, en y incluant le modèle de censure à droite.

Censure à droite Il y a censure à droite lorsque la durée de survie est supérieure à la durée de participation. Un exemple typique est celui où l'événement considéré est le décès d'un patient, et la durée d'observation est une durée totale d'hospitalisation. On peut aussi observer ce genre de phénomène dans des études de fiabilité quand la panne d'un appareil ou d'un composant électronique ne permet pas de continuer l'observation pour un autre appareil ou composant. L'expérimentateur peut également fixer une date de fin d'expérience et les observations pour les individus pour lesquels on n'a pas observé l'événement d'intérêt avant cette date seront censurées à droite.

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

Censure à gauche Il y a censure à gauche lorsque la durée de survie est inférieure à la durée observée.

Supposons par exemple que nous étudions la fiabilité d'un certain composant électronique qui est branché en parallèle avec un ou plusieurs autres composants. Une panne de ce composant n'entraîne pas nécessairement l'arrêt du système : le système peut continuer à fonctionner, quoique de façon aberrante, jusqu'à ce que cette panne soit détectée (par exemple lors d'un contrôle ou en cas de l'arrêt du système). La durée observée pour ce composant est alors censurée à gauche.

Un autre exemple : si l'on s'intéresse à l'âge à partir duquel une personne commence à accomplir une certaine tâche. Certaines personnes peuvent ne pas se rappeler, et donner juste une valeur supérieure (le cas extrême est que l'on prenne le début de l'étude comme observation). Cette donnée est donc censurée à gauche.

Dans ce dernier exemple, ainsi que dans beaucoup de cas, on trouve des données censurées à gauche dans un même échantillon que des données censurées à droite, ce qui conduit à la définition suivante.

Censure double (ou mixte) On dit qu'on a une censure double si on a des données censurées à droite et des données censurées à gauche dans le même échantillon.

Censure par intervalle Dans le cas de la censure par intervalle, on observe à la fois une borne inférieure et une borne supérieure de la durée d'intérêt. Ceci arrive dans des études de suivi médical où les patients sont contrôlés périodiquement, si un patient ne se présente pas à un ou plusieurs contrôles et se présente ensuite après que l'événement d'intérêt se soit produit. On a aussi pour ce genre d'expériences des données qui sont censurées à droite ou, plus rarement, à gauche. Un avantage de ce type est qu'il permet de représenter les données censurées à droite ou à gauche par des intervalles du type $[a, +\infty[$ et $[0, a]$ respectivement, ce qui permet de considérer ce modèle comme étant plus générique. Turnbull (1976) présente ce genre de censure avec plus de détails.

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

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

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

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

Dans ce travail, nous nous restreignons à l'étude de la censure aléatoire, principalement pour les données censurées à droite et à gauche selon le modèle de Patilea et Rolin (2006).

2.2 L'estimateur de Kaplan-Meier

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

Dans ce genre de données, qui sont souvent des durées de survie ou des données de fiabilité, la fonction de répartition F est estimée par l'estimateur introduit par Kaplan et Meier (1958), donné pour $z < Z_{(n)} = \max\{Z_1, \dots, Z_n\}$ par

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

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

2.3 Présentation du modèle de Patilea et Rolin

Plusieurs modèles non paramétriques ont été proposés pour l'étude de la censure double. Par exemple, le modèle de Turnbull (1974) est le plus utilisé, et plusieurs travaux sont basés sur ce modèle. Cependant, bien que la définition de ce modèle soit plus intuitive, l'estimateur qui est proposé n'est pas pour autant facile à utiliser, car il est donné par une équation intégrale dont la solution n'est pas connue explicitement.

D'autres modèles (qui englobent parfois la censure par intervalles) ont été proposés par Peto (1973), Samuelsen (1989) ou encore Huang (1999). Dans des rapports techniques, Patilea et Rolin (2001, 2004) discutent les avantages et les inconvénients de ces modèles, et en proposent d'autres.

Dans ce travail, nous nous concentrons sur le modèle de censure double proposé par Patilea et Rolin (2006), auquel nous nous référons par modèle de censure mixte dans la suite.

Considérons trois variables aléatoires positives indépendantes X, L et R de

fonctions de répartition respectives F_X, F_L et F_R, et de fonctions de survie [1] respectives S_X, S_L et S_R, où X représente la durée d'intérêt et L et R sont les durées de censure à gauche et à droite respectivement. Dans le modèle I de Patilea et Rolin (2006), au lieu d'observer un échantillon de X on observe un échantillon du couple (Z, A) où $Z = \max(\min(X, R), L)$ et

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

Ce modèle considère la censure à droite et la censure à gauche comme deux phénomènes qui agissent indépendamment l'un de l'autre mais que l'un peut censurer l'autre. Un exemple de ce modèle est donné par un système formé par trois composants, dont deux sont placés en série (le composant dont le temps de fonctionnement nous intéresse et un autre). Un troisième est placé en parallèle avec ce système en série.

Le modèle II proposé par les mêmes auteurs est similaire, mais le rôle de la censure à droite et à gauche est inversé. On observe un échantillon du couple (Z, A) où $Z = \min(\max(X, L), R)$ et

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

Le traitement des deux modèles étant très similaire, nous nous contentons de parler du premier. Considérons H la fonction de répartition de Z, elle se décompose en $\sum_{k=0}^{2} H^{(k)}(t)$ où

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

Ces fonctions peuvent s'écrire :

$$\begin{aligned} H^{(0)}(t) &= \int_0^t F_{L-}(t) S_{R-}(t)\, dF_X(t), \\ H^{(1)}(t) &= \int_0^t F_{L-}(t) S_X(t)\, dF_R(t), \\ H^{(2)}(t) &= \int_0^t \{1 - S_X(t) S_R(t)\}\, dF_L(t), \end{aligned} \tag{2.1}$$

1. Si F est la fonction de répartition d'une variable aléatoire X, alors sa fonction de survie est $S = 1 - F$.

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

Considérons la variable aléatoire $S = \min(X, R)$, sa fonction de répartition est $F_S(t) = 1 - S_X(t)F_R(t)$, et en considérant $H^{(01)} = H^{(0)} + H^{(1)}$, on obtient :

$$H^{(01)}(t) = \int_0^t F_L(t^-)\, dF_S(t)$$
$$H^{(2)}(t) = \int_0^t F_S(t)\, dF_L(t)$$

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

Rappelons que la fonction de hasard (ou mesure de hasard) d'une fonction de répartition F est définie par :

$$\Lambda(t) = \int_0^t \frac{dF(u)}{1 - F(u^-)}$$

et que

$$S(t) = 1 - F(t) = \prod_{[0,t]} (1 - d\Lambda(s))$$

où $\prod$ désigne ici le produit intégral.

De même, on peut définir la mesure de hasard inverse par :

$$M(t) = \int_0^t \frac{dF(u)}{F(u)}$$

et on a alors :

$$F(t) = \prod_{]t,\infty[} (1 - dM(s))$$

On définit les mesures de hasard inverses $M^{(2)}$ et $M^{(01)}$ par :

$$M^{(2)}(t) = \int_0^t \frac{dH^{(2)}(t)}{H(t)} \quad ; \quad M^{(01)}(t) = \int_0^t \frac{dH^{(01)}(t)}{H(t^-) + \Delta H^{(01)}(t)}$$

où $\Delta H^{(01)}(t) = H^{(01)}(t) - H^{(01)}(t^-)$, et on considère les fonctions de répartition associées $F^{(2)}$ et $F^{(01)}$. Nous avons alors

$$H(t) = F^{(2)}(t)F^{(01)}(t). \tag{2.2}$$

Les équations (2.1) et la définition de S_X impliquent que

$$\frac{dH^{(0)}(t)}{F_L(t^-)S_Y(t^-)} = \frac{dF_X(t)}{S_X(t^-)},$$

ce qui suggère de définir la mesure de hasard suivante

$$\Lambda(dt) = \frac{dH^{(0)}(t)}{F^{(2)}(t^-)S^{(01)}(t^-)}. \tag{2.3}$$

Soit F_X^* sa fonction de répartition associée. Remarquons que d'après (2.2), $F^{(2)}(t^-)S^{(01)}(t^-) = F^{(2)}(t^-) - H(t^-)$.

Considérons maintenant un échantillon $(Z_i, A_i)_{1\leqslant i\leqslant n}$ de même loi que (Z, A). Pour calculer l'estimateur de F_X, considérons H_n, $H_n^{(0)}$, $H_n^{(1)}$ et $H_n^{(2)}$ les versions empiriques de H, $H^{(0)}$, $H^{(1)}$ et $H^{(2)}$ respectivement, données par

$$H_n(t) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{Z_i\leqslant t\}}, \quad H_n^{(k)}(t) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{Z_i\leqslant t,\delta_i=k\}}, \text{ for } k = 0, 1, 2, \tag{2.4}$$

et soit $F_n^{(2)}$, $S_n^{(01)}$, Λ_n et F_n les fonctions obtenues en remplaçant $H^{(0)}$, $H^{(1)}$ et $H^{(2)}$ par leurs versions empiriques dans l'expression de $F^{(2)}$, $S^{(01)}$, Λ et F_X^I respectivement. F_n est l'estimateur de F_X proposé par Patilea et Rolin (2006). Son expression est rappelée ci-dessous.

Notons par $\{Z'_j, 1 \leqslant j \leqslant M\}$ les valeurs distinctes de $\{Z_i, 1 \leqslant i \leqslant n\}$ rangées dans l'ordre croissant et posons $D_{kj} = \sum_{i=1}^{n} 1_{\{Z_i=Z'_j, A_i=k\}}$. Alors,

$$1 - F_n(t) = \prod_{j/Z'_j\leqslant t} \left\{1 - D_{0j}/(U_{j-1} - nH_n(Z'_{j-1})\right\}, \tag{2.5}$$

où $U_{j-1} = n\prod_{j\leqslant l\leqslant M}\{1 - D_{2l}/(nH_n(Z'_l))\}$.

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

Comme nous allons le voir, l'estimateur proposé converge vers F_X^* au lieu de F_X ; il se pose alors un problème d'identifiabilité : Sous quelles conditions a-t-on $F_X^* = F_X$? Une condition suffisante pour cela est que $I_L < I_X$ et $T_X < T_R$, où pour toute variable aléatoire V, I_V et T_V sont le point initial et le point terminal du support de la distribution de V.

2.4 Propriétés de l'estimateur de Patilea et Rolin

Patilea et Rolin (2006) étudient en particulier la convergence forte (presque sûre) uniforme de leurs estimateurs ainsi que leur convergence faible vers un processus gaussien. Comme ces estimateurs s'écrivent de façon explicite comme des fonctionnelles des fonctions de répartition empiriques, leurs propriétés asymptotiques peuvent être déduites de ces dernières.

Soit T_0 le point terminal de la sous-distribution $H^{(0)}$. Nous avons le résultat suivant.

Théorème 2.1 (Patilea et Rolin, 2006, Corollaire 6.4)**.**

1. *Si* $\Lambda(T_0^-) < \infty$ *alors presque sûrement*

$$\sup_{0 \leqslant t < T_0} |\Lambda_n(t) - \Lambda(t)| \to 0$$

et $\Delta\Lambda_n(T_0) \to \Delta\Lambda(T_0)$. *Et si* $\Lambda(T_0) = \infty$ *alors presque sûrement*

$$\sup_{0 \leqslant s \leqslant t} |\Lambda_n(s) - \Lambda(s)| \to 0$$

pour tout $t < T_0$ *et* $\Lambda_n(T_0) \to \infty$.

2. *Si* $I_L < I_X$ *et* $T_X < T_R$ *alors nous avons presque sûrement*

$$\|F_n - F_X\| = \sup_{t \in [0,\infty]} |F_n(t) - F_X(t)| \to 0.$$

Notons $\mathcal{D}([a,b])$ l'ensemble des fonctions càdlàg (continues à droite et ayant des limites à gauche en tout point) définies sur $[a,b]$, muni de la norme de la convergence uniforme. Pour la convergence faible, nous considérons la tribu engendrée par les boules. Soit I_0 le point initial de la sous-distribution $H^{(0)}$. Nous avons besoin de la condition suivante.

$$\int_{]I_0,\infty]} \frac{M_2(du)}{H(u)} = \int_{]I_0,\infty]} \frac{H^{(2)}(du)}{(H(u))^2} < \infty. \tag{2.6}$$

Nous avons alors le résultat suivant.

Théorème 2.2 (Patilea et Rolin, 2006, Théorème 7.3)**.** *Supposons que la condition* (2.6) *est vérifiée. Soit* $\tau > I_0$ *tel que* $H^{(01)}(\tau) < 1$. *Alors* $\sqrt{n}(\Lambda_n - \Lambda)$ *considéré comme un processus converge faiblement vers* V *dans* $\mathcal{D}([0,\tau])$, *où*

$$V_t = \int_{]0,t]} \frac{dG_{0u}}{(F^{(2)} - H)(u^-)} - \int_{]0,t]} \frac{G_{3u-} - G_{u-}}{(F^{(2)} - H)^2(u)} H^{(0)}(du), \qquad t \in [0,\tau],$$

est un processus gaussien centré. De plus, $\sqrt{n}(F_n - F_X)$ *considéré comme un processus converge faiblement vers le processus gaussien centré* W *défini par*

$$W_t = (1 - F_X(t)) \int_{]0,t]} \frac{dV_u}{1 - \Delta\Lambda(u)}.$$

Ces propriétés sont similaires à celles de la fonction de répartition empirique (Théorèmes de Glivenko-Cantelli et de Donsker), ce qui nous permet d'utiliser les outils de la théorie des processus empiriques en construisant un processus empirique basé sur cet estimateur au lieu de la fonction de répartition empirique. Nous utiliserons cela au chapitre 5, mais d'abord rappelons quelques propriétés des processus empiriques.

2.5 Rappels sur les processus empiriques

2.5.1 Définitions

Intuitivement, un processus empirique est un processus aléatoire qui dépend d'un échantillon. Par exemple, la fonction de répartition empirique.

Plus précisément, si l'on considère un espace probabilisable $(\mathcal{X}, \mathcal{A})$, et un échantillon $X_1, X_2, \dots, X_n$ i.i.d de loi de probabilité P_X, on définit la mesure empirique P_n par :

$$P_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i},$$

où δ_x est la mesure de Dirac au point x.

Pour une famille $\mathcal{S}$ d'ensembles mesurables, on définit alors le processus empirique $\{P_n(A), A \in \mathcal{S}\}$. Un processus empirique ainsi défini est dit *indexé par des ensembles*. Dans le cas réel, la fonction de répartition empirique peut s'écrire ainsi en prenant $\mathcal{S} = \{]-\infty, t], t \in \mathbb{R}\}$.

Pour une classe $\mathcal{F}$ de fonctions mesurables, on peut définir un processus empirique $\{P_n f, f \in \mathcal{F}\}$ où

$$P_n f = \int f dP_n = \frac{1}{n} \sum_{i=1}^{n} f(X_i).$$

Un tel processus est dit *indexé par des fonctions*. Cette définition est plus générale car elle permet de décrire toutes les fonctions mesurables de l'échantillon. La définition précédente est obtenue en se limitant à des fonctions indicatrices.

De plus, pour peu que la classe $\mathcal{F}$ soit dénombrable (resp. ait la puissance du continu), le processus empirique peut être vu comme un processus "classique" à indices entiers (resp. réels).

Souvent, ce que l'on appelle processus empirique est $\{\sqrt{n}(P_n(A) - P_X(A)), A \in \mathcal{S}\}$ ou bien $\{\sqrt{n}(P_n f - P_X f), f \in \mathcal{F}\}$. C'est-à dire qu'au lieu de prendre la

mesure empirique, on prend la différence entre cette dernière et la mesure de probabilité théorique.

Dans la suite, nous nous restreignons au cas réel, et nous appelons processus empirique un processus de la forme $\sqrt{n}(F_n(x) - F(x))$ où F_n est la fonction de répartition empirique ou un autre estimateur de la fonction de répartition. Par exemple, dans le cas des données censurées à droite, en remplaçant F_n par l'estimateur de Kaplan-Meier on obtient le *processus empirique de Kaplan-Meier*. Pour plus de détails concernant la théorie générale des processus empiriques, voir Shorack et Wellner (1986, chap. 26), Kosorok (2008) et van der Vaart et Wellner (1996).

2.5.2 Processus empirique uniforme

Un résultat important, qui est souvent utilisé pour simplifier les preuves, est la réduction au cas uniforme. Ceci consiste à introduire un processus empirique uniforme (c'est à dire un processus empirique basé sur un échantillon de loi uniforme sur $[0,1]$) qui est plus facile à étudier. L'extension au cas d'une loi arbitraire est souvent possible sans trop d'hypothèses, mais nous devons parfois supposer la continuité de la fonction de répartition.

Soit F une fonction de répartition et soit F^{inv} son inverse généralisée définie pour $t \in]0,1[$ par :

$$F^{inv}(t) = \inf\{x \in \mathbb{R} : F(x) \geqslant t\}.$$

On a les propriétés suivantes :

- $\forall t \in [0,1] : F\left(F^{inv}(t)\right) \geqslant t$;
- $\forall x \in \mathbb{R} : F^{inv}\left(F(x)\right) \leqslant x$;
- $F(x) \geqslant t \iff F^{inv}(t) \leqslant x$;
- Si X suit la loi F, alors $\forall t \in [0,1] : P(F(X) \leqslant t) \leqslant t$, et si t appartient à l'image de F alors $P(F(X) \leqslant t) = t$. En particulier, si F est continue alors $F(X)$ suit la loi uniforme sur $[0,1]$.

Théorème 2.3. *Soit ξ une variable aléatoire de loi uniforme sur $[0,1]$. Et soit P_X une loi de probabilité sur $\mathbb{R}$, de fonction de répartition F. On définit la variable aléatoire X par $X = F^{inv}(\xi)$.*

Alors X suit la loi P_X.

Démonstration. D'après les propriétés précédentes et la croissance de F, si $\xi \leqslant F(x)$ alors $X = F^{inv}(\xi) \leqslant F^{inv}(F(x)) \leqslant x$. Et si $X \leqslant x$, alors $F(x) \geqslant F(X) = F\left(F^{inv}(\xi)\right) \geqslant \xi$.

Ce qui montre que les événements $\{X \leqslant x\}$ et $\{\xi \leqslant F(x)\}$ sont équivalents, et on a alors $P(X \leqslant x) = P(\xi \leqslant F(x)) = F(x)$. □

Ce résultat élémentaire peut se généraliser au cas des processus empiriques : Considérons une suite $(U_n)_{n\geqslant 1}$ de v.a. i.i.d. de loi uniforme sur $[0,1]$. Nous notons respectivement $\mathbb{U}_n$ et α_n la fonction de répartition empirique et le processus empirique associés

$$\alpha_n(t) = \sqrt{n}(\mathbb{U}_n(t) - t) \quad \text{et} \quad \mathbb{U}_n(t) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{U_i \leqslant t\}}.$$

Théorème 2.4. *Les suites de processus* $\{F_n(t), t \in \mathbb{R}\}$ *et* $\{\mathbb{U}_n(F(t)), t \in \mathbb{R}\}$ *ont les mêmes lois de probabilité conjointes en n (c'est-à-dire que pour tout* $k \in \mathbb{N}^*$ *et pour tout* $n_1, n_2, \dots, n_k \in \mathbb{N}$, $(F_{n_1}, F_{n_2}, \dots F_{n_k})$ *et* $(\mathbb{U}_{n_1}(F), \mathbb{U}_{n_2}(F), \dots \mathbb{U}_{n_k}(F)$ *ont la même loi). De même pour les processus* $\{\sqrt{n}(F_n(t) - F(t)), t \in \mathbb{R}\}$ *et* $\{\alpha_n(F(t)), t \in \mathbb{R}\}$.

Démonstration. voir Shorack et Wellner (1986, Théorème 2, page 4) □

Ceci nous permet de limiter le travail probabiliste au cas des échantillons de loi uniforme, et de généraliser ces résultats en insérant la fonction de répartition F dans le résultat (On a parfois besoin de supposer que F est continue pour faire cela).

Si F est continue, et si on choisit $U_n = F(X_n)$ pour tout n dans le théorème précédent, alors non seulement on a égalité des distributions mais également l'égalité presque sûre. Ce résultat peut être généralisé sous des hypothèses assez faibles au cas où F n'est pas continue (voir par exemple Shorack et Wellner, 1986, p. 102).

2.5.3 Inégalité de Dvoretzky-Kiefer-Wolfowitz

Une inégalité importante concernant le processus empirique est l'inégalité de Dvoretzky-Kiefer-Wolfowitz (ou inégalité DKW). Elle a été donnée pour la première fois par Dvoretzky *et al.* (1956). Plusieurs auteurs ont essayé de préciser la constante dans le terme de droite. La version suivante est due à Massart (1990), qui montre que 2 est la meilleure constante possible.
Théorème 2.5 (Massart, 1990)**.** *Pour tout* $\lambda > 0$, *nous avons*

$$P\left(\sup_{x\in\mathbb{R}} \sqrt{n}\,|F_n(x) - F(x)| > \lambda\right) \leqslant 2\exp(-2\lambda^2)$$

2.5.4 Processus empirique pour des données censurées

L'estimateur de Kaplan-Meier a des propriétés assez similaires à la fonction de répartition empirique, par exemple la convergence uniforme presque sûre (Stute et Wang, 1993 ; Winter *et al.*, 1978), la normalité asymptotique (Breslow et Crowley, 1974 ; Gill, 1983), et la loi du logarithme itéré (Földes et Rejtő, 1981a). Ceci justifie que l'on s'intéresse à généraliser la théorie des processus empiriques au cas des données censurées.

Aussi, on définit *le processus empirique de Kaplan-Meier* par $a_n(t) = \sqrt{n}(\hat{\mathrm{F}}_n(t) - \mathrm{F}(t))$.

Une version de l'inégalité DKW pour le processus empirique de Kaplan-Meier à été donnée par Bitouzé *et al.* (1999).

Théorème 2.6 (Bitouzé *et al.*, 1999)**.** *Il existe une constante absolue* C *telle que pour tout* λ *positif*

$$\mathrm{P}\left(\sup_{x\in\mathbb{R}} \sqrt{n}\left|(1-\mathrm{G})\left(\hat{\mathrm{F}}_n(x) - \mathrm{F}(x)\right)\right| > \lambda\right) \leqslant 2,5\exp(-2\lambda^2 + \mathrm{C}\lambda),$$

où G est la fonction de répartition de la variable de censure.

De la même manière, on peut définir le processus empirique basé sur l'estimateur de Patilea-Rolin.

Chapitre 3

Lois du logarithme itéré

3.1 Loi du logarithme itéré classique

La loi du logarithme itéré (ou LIL pour "Law of the Iterated Logarithm") est un des théorème limites importants de la statistique. Sa version initiale, donnée pour une somme de variables aléatoires indépendantes et de même loi, remonte à Khinchine et Kolmogorov dans les années 1920. Depuis, un grand nombre de travaux ont porté sur des lois du logarithme itéré pour la fonction de répartition empirique et d'autres estimateurs de la fonction de répartition.

D'autres lois du logarithme itéré relatives à différentes statistiques ont fait l'objet de plusieurs travaux. Citons, sans prétendre à l'exhaustivité, Hall (1981) qui a montré une LIL pour des estimateurs non-paramétriques de la densité, Hardle (1984) qui a montré une LIL pour des estimateurs non-paramétriques de la régression, et Földes et Rejtő (1981a) qui ont montré une LIL pour l'estimateur de Kaplan-Meier de la fonction de survie pour des données censurées à droite, résultat que nous rappelons ci-dessous avec la LIL de Kiefer (1961) puisque nous allons nous en servir.

Sous sa forme la plus simple, le théorème peut être exprimé ainsi.
Théorème 3.1 (Loi du logarithme itéré)**.** *Soit* (X_n) *une suite de v.a. i.i.d. centrées de variance 1, et* $S_n = \sum_{i=1}^{n} X_i$*. Alors :*

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

3.2 Loi du logarithme itéré pour les processus empiriques

3.2.1 Cas des données complètes

La loi du logarithme itéré pour les fonctions de répartition empiriques a été démontrée indépendamment par Chung (1949) et Smirnov dans le cas des échantillons dans $\mathbb{R}$, et par Kiefer (1961) pour $\mathbb{R}^n$.

Dans cette section, nous allons exposer des résultats qui seront, ainsi que leur généralisation au cas des données censurées, essentiels pour la suite. Considérons d'abord le cas du processus empirique uniforme, la généralisation au cas d'une loi continue quelconque étant immédiate.

Soit $U_1, \dots, U_n$ des v.a. i.i.d. de loi $\mathcal{U}_{[0,1]}$ et soit la fonction de répartition empirique $\mathbb{U}(t) = \frac{1}{n}\sum 1_{\{U_i \leqslant t\}}$ et le processus empirique $\alpha_n(t) = \sqrt{n}(\mathbb{U}(t) - t)$ associés.

Théorème 3.2 (Smirnov)**.** *Soit $b_n = \sqrt{2\log_2 n}$. Alors :*

$$\limsup_{n\to\infty} \frac{\|\alpha_n\|}{b_n} = \limsup_{n\to\infty} \frac{\|n(\mathbb{U}_n - \mathrm{I})\|}{\sqrt{n}b_n} = \frac{1}{2} \quad p.s.$$

où $\|x\| = \sup_{t\in[0,1]} |x(t)|$ est la norme de la convergence uniforme.

Le résultat suivant, qui a été prouvé indépendamment par Chung (1949), est plus fort au sens qu'il implique le résultat précédent (voir Shorack et Wellner, 1986, Chapitre 13).

Théorème 3.3. *Soit (λ_n) une suite croissante de nombres positifs, alors :*

$$\mathrm{P}\left(\limsup_{n\to\infty} \{\|\alpha_n\| \geqslant \lambda_n\}\right) = \begin{cases} 0 \textit{ si } \sum \frac{\lambda_n}{n}\exp(-2\lambda_n^2) < \infty \\ 1 \textit{ sinon} \end{cases}$$

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

Théorème 3.4 (Kiefer, 1961)**.**

$$\mathrm{P}\left(\limsup_{n\to\infty} \frac{\sqrt{n}\sup_{x\in\mathbb{R}} |F_n(t) - F(t)|}{\sqrt{\frac{1}{2}\log\log n}} \leqslant 1\right) = 1. \tag{3.1}$$

3.2.2 Cas des données censurées à droite — Estimateur de Kaplan-Meier

Le résultat suivant est une loi du logarithme itéré pour l'estimateur de Kaplan-Meier de la fonction de répartition, qui est d'une certaine manière similaire au théorème de Chung-Smirnov.

Dans le cas de la censure à droite, nous définissons l'estimateur de Kaplan-Meier, que nous notons par $\hat{\mathrm{F}}_n$, en posant $\hat{\mathrm{F}}_n(z) = 0$ pour $z > \mathrm{Z}_{(n)}$. Pour toute variable aléatoire V, on note par I_{V} et T_{V} le point initial et le point terminal du support de la loi de V.

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

Théorème 3.5 (Földes et Rejtő (1981a))**.** *On suppose que* F *et* G *sont continues, et que* $\mathrm{T}_{\mathrm{X}} < \mathrm{T}_{\mathrm{C}}$*, alors*

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

La condition $\mathrm{T}_{\mathrm{X}} < \mathrm{T}_{\mathrm{C}}$ pouvant paraître restrictive, citons le théorème autrement.

Corollaire 3.1 (Földes et Rejtő (1981a))**.** *Si* F *et* G *sont continues, et si le réel* T *est tel que* $\mathrm{G}(\mathrm{T}) < 1$*, alors*

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

où $\mathrm{T}^* = \min\{\mathrm{T}, \mathrm{T}_{\mathrm{X}}\}$.

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

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

avec $\tilde{\mathrm{N}}_n(x) = \sum_{i=1}^{n} 1_{\{\mathrm{Z}_i\leqslant x\}}$. Nous avons alors le résultat suivant.

Corollaire 3.2. *Si* F *et* G *sont continues et si le réel* I *vérifie* $G(I) > 0$, *alors*

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

où $I^* = \max\{I, I_X\}$.

Passons maintenant au cas de la censure mixte.

3.2.3 Cas des données censurées à droite et à gauche — Estimateur de Patilea et Rolin

Le modèle étudié ici est celui de la censure mixte introduit à la section 2.3, nous utilisons donc les mêmes notations, et nous noterons pour toute variable aléatoire U, F_U (resp. S_U) sa fonction de répartition (resp. de survie). Nous noterons aussi $T_U = \sup\{t : F_U(t) < 1\}$ et $I_U = \inf\{t : F_U(t) \neq 0\}$ les points terminaux du support de F_U, et nous supposons que les fonctions de répartition de X, R et L sont continues.

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

$$H^{(0)}(t) = P(Z \leqslant t, A = 0) = \int_0^t F_L(u) S_R(u) dF_X(u),$$

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

La LIL de Kiefer (théorème 3.4) s'applique à $H_n^{(0)}$ et donne

$$P\left(\limsup_{n\to\infty} \frac{\sup_{u\in\mathbb{R}} |H_n^{(0)}(u) - H^{(0)}(u)|}{\sqrt{\log\log n/2n}} \leqslant 1\right) = 1. \tag{3.2}$$

En inversant le temps, l'estimateur produit-limite de F_L, qui est continue, est donné par $\tilde{F}_n(u) = \prod_{j/Z_j\geqslant u} \left\{1 - D_{2j}/N_j\right\}$. Remarquons que la LIL de Kiefer (1961) s'applique à H_n et que de plus le corollaire 3.2 s'applique à $\tilde{F}_n$ (sous l'hypothèse $\sup(I_R, I_L) < I_X$) pour obtenir que pour presque tout ω il existe n_1 et un nombre fixé A tel que pour tout $n > n_1$

$$\sup_{I_X\leqslant u} |\tilde{F}_n(u) - F_L(u) - (H_n(u) - H(u))| \leqslant A\sqrt{\frac{\log\log n}{2n}}.$$

Maintenant, comme $(F_L(u) - H(u)) \geqslant F_L(I_X)S_R(T_X)S_X(u)$ dès que $I_X \leqslant u \leqslant T_X$, nous déduisons sous les hypothèses $I_L < I_X$ et $T_R < T_X$ que pour tout $n > n_1$,

$$\tilde{F}_n(u) - H_n(u) \geqslant (F_L(u) - H(u))/2 \quad \text{p.s.}, \tag{3.3}$$

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

$$u_n = S_X^{-1}\left(\frac{2A}{F_L(I_X)S_R(T_X)}\sqrt{\frac{\log\log n}{2n}}\right), \tag{3.4}$$

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

La mesure de hasard associée à X est $d\Lambda(t) = dF_X(t)/S_X(t)$ qui peut être écrite $dH^{(0)}(t)/(F_L(t) - H(t))$ pour tout t tel que $I_X \leqslant t < T_X$. Pour $I_X \leqslant u < T_X$, on pose

$$T(u) = \int_{I_X}^{u} d\Lambda(t) = -\log(S_X(u)), \quad T_n(u) = \int_{I_X}^{u} dH_n^{(0)}(t)/(\tilde{F}_n(t) - H_n(t)),$$

où $T_n(u)$ est obtenue en remplaçant $H^{(0)}$, F_L et H par leurs estimateurs dans l'expression de $T(u)$. Le théorème suivant donne la loi du logarithme itéré pour S_n, estimateur de Patilea et Rolin (2006).

Théorème 3.6 (Messaci et Nemouchi, 2011, 2013)**.** *Si* S_X, S_R *et* S_L *sont des fonctions de survies continues, et si* $\sup(I_L, I_R) < I_X$ *et* $T_X < T_R$. *Alors*

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

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

Posons

$$\bar{S}_n(t) = \prod_{j/Z_j \leqslant t}\left\{1 - D_{0j}/(U_{j-1} - N_{j-1} + 1)\right\}. \tag{3.5}$$

La preuve du théorème est basée sur la décomposition suivante

$$\begin{aligned}|S_n(u) - S_X(u)| \leqslant{}& |S_n(u) - \bar{S}_n(u)| \\ &+ |\bar{S}_n(u) - S_X(u)|,\end{aligned} \tag{3.6}$$

et nous avons

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

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

$$\bar{S}_n(u) - S_X(u) = \exp(-\theta_n(u))(\log \bar{S}_n(u) + T_n(u)) \tag{3.7}$$
$$+ S_X(u)(T(u) - T_n(u)) \tag{3.8}$$
$$+ \frac{1}{2}\exp(-\theta'_n(u))(T(u) - T_n(u))^2, \tag{3.9}$$

où

$$\min\{-\log \bar{S}_n(u),\, T_n(u)\} \leqslant \theta_n(u) \leqslant \max\{-\log \bar{S}_n(u),\, T_n(u)\}, \tag{3.10}$$

et

$$\min\{T(u),\, T_n(u)\} \leqslant \theta'_n(u) \leqslant \max\{T(u),\, T_n(u)\}. \tag{3.11}$$

Nous allons maintenant énoncer et démontrer les quatre lemmes suivants. Le lemme 3.1 nous fournit un outil pour démontrer les lemmes 3.2, 3.3 et 3.4, nous permettant de traiter le premier terme du membre de droite de (3.6), et les membres de droite (3.7) et (3.8) respectivement.

Lemme 3.1. *Pour presque tout ω, il existe $n_0(\omega)$ tel que si $n > n_0$, alors pour tout $I_X \leqslant u \leqslant u_n$, $k_1 > 0$ et $k_2 \geqslant 0$ où $k = k_1 + k_2 > 1$, nous avons*

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

Démonstration. Nous avons par (3.3), pour tout $n > n_1$, pour tout $I_X \leqslant u \leqslant u_n$ et tout $I_X \leqslant t \leqslant u$

$$\left(\tilde{F}_n(t) - H_n(t)\right)^{k_1} \geqslant \left(\frac{F_L(t) - H(t)}{2}\right)^{k_1} \quad \text{p.s.},$$

c'est à dire,

$$\frac{1}{\left(\tilde{F}_n(t) - H_n(t)\right)^{k_1}} \frac{1}{\left(F_L(t) - H(t)\right)^{k_2}} \leqslant \frac{2^{k_1}}{\left(F_L(t) - H(t)\right)^{k_1}} \frac{1}{\left(F_L(t) - H(t)\right)^{k_2}}$$
$$= \frac{2^{k_1}}{\left(F_L(t) - H(t)\right)^{k_1+k_2}}.$$

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

$$\int_{I_X}^{u} \frac{2^{k_1} dH_n^{(0)}(t)}{\left(F_L(t) - H(t)\right)^{k}} = \int_{I_X}^{u} \frac{2^{k_1} dH^{(0)}(t)}{\left(F_L(t) - H(t)\right)^{k}} + \int_{I_X}^{u} \frac{2^{k_1} d(H_n^{(0)}(t) - H^{(0)}(t))}{\left(F_L(t) - H(t)\right)^{k}}.$$

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

i) Rappelons que

$$d\mathrm{H}^{(0)}(t) = -\mathrm{F_L}(t)\mathrm{S_R}(t)d\mathrm{S_X}(t).$$

De plus, un calcul élémentaire montre que

$$\mathrm{F_L}(t) - \mathrm{H}(t) = \mathrm{F_L}(t)\mathrm{S_R}(t)\mathrm{S_X}(t).$$

Nous pouvons donc majorer le premier terme comme suit

$$\begin{aligned}\int_{\mathrm{I_X}}^{u} \frac{2^{k_1} d\mathrm{H}^{(0)}(t)}{(\mathrm{F_L}(u) - \mathrm{H}(u))^k} &= \int_{\mathrm{I_X}}^{u} \frac{-2^{k_1}\mathrm{F_L}(t)\mathrm{S_R}(t)d(\mathrm{S_X}(t))}{(\mathrm{F_L}(t)\mathrm{S_R}(t)\mathrm{S_X}(t))^k} \\ &\leqslant -\frac{2^{k_1}}{\mathrm{F_L^{k-1}}(\mathrm{I_X})\mathrm{S_R^{k-1}}(\mathrm{T_X})} \int_{\mathrm{I_X}}^{u} \frac{d(\mathrm{S_X}(t))}{\mathrm{S_X^k}(t)} \\ &\leqslant \frac{2^{k_1}}{(k-1)\mathrm{F_L^{k-1}}(\mathrm{I_X})\mathrm{S_R^{k-1}}(\mathrm{T_X})\mathrm{S_X^{k-1}}(u)}\end{aligned}$$

ii) Quant au second terme, nous avons

$$\begin{aligned}\int_{\mathrm{I_X}}^{u} \frac{2^{k_1} d(\mathrm{H}_n^{(0)}(t) - \mathrm{H}^{(0)}(t))}{(\mathrm{F_L}(t) - \mathrm{H}(t))^k} &\leqslant \frac{2^{k_1}}{\mathrm{F_L^k}(\mathrm{I_X})\mathrm{S_R^k}(\mathrm{T_X})\mathrm{S_X^k}(u)} \int_{\mathrm{I_X}}^{u} \left|d(\mathrm{H}_n^{(0)}(t) - \mathrm{H}^{(0)}(t))\right| \\ &\leqslant \frac{2^{k_1+1}}{\mathrm{F_L^k}(\mathrm{I_X})\mathrm{S_R^k}(\mathrm{T_X})\mathrm{S_X^k}(u)} \sup_{t\in\mathbb{R}} \left|\mathrm{H}_n^{(0)}(t) - \mathrm{H}^{(0)}(t)\right|.\end{aligned}$$

L'application de (3.4) montre que

$$\mathrm{S_X}(u_n) \geqslant \frac{2\mathrm{A}}{\mathrm{F_L}(\mathrm{I_X})\mathrm{S_R}(\mathrm{T_X})}\sqrt{\frac{\log\log n}{2n}}, \tag{3.12}$$

Puisque $u \leqslant u_n$, tenant compte de (3.5) et regroupant les deux termes, il vient

$$\begin{aligned}&\int_{\mathrm{I_X}}^{u} \frac{d\mathrm{H}_n^{(0)}(t)}{(\tilde{\mathrm{F}}_n(t) - \mathrm{H}_n(t))^{k_1}(\mathrm{F_L}(t) - \mathrm{H}(t))^{k_2}} \\ &\quad\leqslant \frac{2^{k_1}}{\mathrm{F_L^{k-1}}(\mathrm{I_X})\mathrm{S_R^{k-1}}(\mathrm{T_X})} \frac{1}{\mathrm{S_X^{k-1}}(u)} \times \left(\frac{2}{\mathrm{A}} + \frac{1}{k-1}\right) \\ &\quad\leqslant \frac{2^{k_1}}{\mathrm{F_L^{k-1}}(\mathrm{I_X})\mathrm{S_R^{k-1}}(\mathrm{T_X})} \frac{\mathrm{F_L^{k-1}}(\mathrm{I_X})\mathrm{S_R^{k-1}}(\mathrm{T_X})}{2^{k-1}\mathrm{A}^{k-1}} \times \left(\frac{2n}{\log\log n}\right)^{\frac{k-1}{2}} \left(\frac{2}{\mathrm{A}} + \frac{1}{k-1}\right) \\ &\quad= \mathrm{O}\left(\frac{2n}{\log\log n}\right)^{\frac{k-1}{2}},\end{aligned} \tag{3.13}$$

en tenant compte encore une fois de la relation (3.12).

Nous obtenons bien le résultat annoncé dans le lemme. □

Lemme 3.2. *Nous avons*

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

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

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

Nous avons donc

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

en appliquant le lemme 3.1 pour $k_1 = 2$ et $k_2 = 0$. □

Lemme 3.3. *Nous avons*

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

Démonstration. De (3.5), nous déduisons que

$$\log \bar{S}_n(u) = \int_{I_X}^{u} n \log(1 - \frac{1}{n\tilde{F}_n(t) - nH_n(t) + 1}) dH_n^{(0)}(t).$$

En outre le développement logarithmique implique que

$$
\begin{aligned}
|\log \bar{S}_n(u) + T_n(u)| &= \left| \int_{I_X}^{u} \frac{dH_n^{(0)}(t)}{\tilde{F}_n(t) - H_n(t)} - \int_{I_X}^{u} n \sum_{l=1}^{\infty} \frac{1}{l} (n\tilde{F}_n(t) - nH_n(t) + 1)^{-l} dH_n^{(0)}(t) \right| \\
&\leqslant \left| \int_{I_X}^{u} \frac{dH_n^{(0)}(t)}{\tilde{F}_n(t) - H_n(t)} - \frac{dH_n^{(0)}(t)}{\frac{1}{n} + \tilde{F}_n(t) - H_n(t)} \right| \\
&\quad + \left| \int_{I_X}^{u} -n \sum_{l=2}^{\infty} \frac{1}{l} (n\tilde{F}_n(t) - nH_n(t) + 1)^{-l} dH_n^{(0)}(t) \right| \\
&\leqslant 2 \int_{I_X}^{u} \frac{dH_n^{(0)}(t)}{(n\tilde{F}_n(t) - nH_n(t))^2}.
\end{aligned}
$$

Il reste à appliquer le lemme 3.1 pour obtenir le résultat visé. □

Lemme 3.4. *Nous avons*

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

Démonstration. Remarquons que,

$$
\begin{aligned}
|T_n(u) - T(u)| &\leqslant \left| \int_{I_X}^{u} \frac{(\tilde{F}_n(t) - H_n(t)) - (F_L(t) - H(t))}{(\tilde{F}_n(t) - H_n(t))(F_L(t) - H(t))} dH_n^{(0)}(t) \right| \\
&\quad + \left| \int_{I_X}^{u} \frac{d(H_n^{(0)}(t) - H^{(0)}(t))}{F_L(t) - H(t)} \right| \\
&\leqslant \sup_{I_X \leqslant t} \left| (\tilde{F}_n(t) - H_n(t)) - (F_L(t) - H(t)) \right| \\
&\quad \times \int_{I_X}^{u} \frac{dH_n^{(0)}(t)}{(\tilde{F}_n(t) - H_n(t))(F_L(t) - H(t))} \\
&\quad + \frac{2 \sup |H_n^{(0)}(t) - H^{(0)}(t)|}{F_L(I_X) S_R(T_X) S_X(u)}.
\end{aligned}
$$

En vertu de (3.2) et (3.13), il s'ensuit que pour n assez grand

$$
|T_n(u) - T(u)| \leqslant 2\sqrt{\log\log n / 2n} \frac{2A(\frac{2}{A} + 1) + (1 + \varepsilon)}{F_L(I_X) S_R(T_X) S_X(u)}. \tag{3.14}
$$

Compte tenu de (3.4), nous voyons qu'il existe une constante K, telle que

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

En revenant encore à (3.14), nous déduisons que

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

Nous sommes maintenant en mesure de donner la démonstration du Théorème 3.6.

Démonstration du Théorème 3.6. En vertu de (3.11), nous voyons que

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

L'application immédiate du lemme 3.4 donne alors

$$\frac{1}{2}\exp(-\theta_n'(u))|T_n(u) - T(u)|^2 = O(\sqrt{(\log\log n)/n}) \text{ p.s.} \qquad (3.15)$$

Par ailleurs, tenant compte de (3.10), il vient

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

Combinant les relations (3.9) et (3.15) avec les lemmes 3.3 et 3.4, nous pouvons conclure que, pour $I_X \leqslant u \leqslant u_n$

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

Cette dernière relation, combinée avec l'inégalité

$$|S_n(u) - S_X(u)| \leqslant |S_n(u) - \bar{S}_n(u)| + |S_X(u) - \bar{S}_n(u)|,$$

montre que

$$\sup_{I_X \leqslant u \leqslant u_n} |S_n(u) - S_X(u)| = O(\sqrt{(\log\log n)/n})$$

pour n suffisamment grand.

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

$$\sup_{u_n < u < +\infty} |S_n(u) - S_X(u)| \leqslant |S_X(u_n)| + |S_n(u_n) - S_X(u_n)|. \qquad \square$$

3.3 Lois fonctionnelles du logarithme itéré

Un autre résultat très important est celui de Finkelstein (1971), il consiste à prouver une loi du logarithme itéré pour le processus empirique considéré comme un élément de $\mathbb{B}$, l'ensemble des fonctions réelles bornées, et pas seulement pour la norme uniforme de ce processus, comme c'est le cas pour les résultats précédents. Un tel résultat est parfois appelé loi fonctionnelle du logarithme itéré.

Avant d'aller plus loin, donnons quelques définitions

Définition. Soit $(X_n)_{n\geqslant 0}$ une suite d'éléments aléatoires d'un espace métrique (S, d) définies sur l'espace de probabilité $(\Omega, \mathcal{A}, P)$. On dit que (X_n) est presque sûrement relativement compacte dans (S, d) avec pour ensemble limite H, s'il existe $\Omega_0 \in \mathcal{A}$ avec $P(\Omega_0) = 1$ tel que pour tout $\omega \in \Omega_0$:

1. toute suite n' de nombres entiers admet une sous suite n'' telle que $X_{n''}(\omega)$ converge dans (S, d) ;
2. toute les valeurs d'adhérence de $X_n(\omega)$ appartiennent à H ;
3. pour tout $h \in H$, il existe une suite $n' = n_{h,\omega}$ telle que $X_{n'}(\omega)$ converge vers h.

Définition. Soit h une fonction définie sur un intervalle I et à valeurs réelles. On dit que h est absolument continue si $\forall \varepsilon > 0, \exists \delta > 0, \forall (]x_i, y_i[)_{1\leqslant i\leqslant N}$ intervalles disjoints de I :

$$\sum_{i=1}^{N} (y_i - x_i) < \delta \implies \sum_{i=1}^{N} |h(y_i) - h(x_i)| < \varepsilon.$$

On dit que h est absolument continue par rapport à une mesure μ (ou une fonction de répartition en sous entendant que c'est par rapport à la mesure de probabilité liée à cette fonction), si $\forall \varepsilon > 0, \exists \delta > 0, \forall (]x_i, y_i[)_{1\leqslant i\leqslant N}$ intervalles disjoints de I :

$$\sum_{i=1}^{N} \mu(]x_i, y_i[) < \delta \implies \sum_{i=1}^{N} |h(y_i) - h(x_i)| < \varepsilon.$$

Considérons l'ensemble

$$\mathcal{H} = \left\{ \begin{array}{c} h : h \text{ est absolument continue sur } [0,1] \text{ avec} \\ h(0) = h(1) = 0 \quad \text{et} \quad \int_0^1 [h'(t)]^2 \, dt \leqslant 1 \end{array} \right\},$$

où h' est la dérivée au sens de Lebesgue de h.

On donne d'abord le théorème pour le cas uniforme.

Théorème 3.7. *Soit $b_n = \sqrt{2 \log \log n}$. Alors la suite $\left\{\frac{\alpha_n}{b_n}\right\}$ est presque sûrement relativement compacte dans $\mathbb{B}([0,1])$ avec pour ensemble limite l'ensemble $\mathcal{H}$ ci-dessus.*

Il peut se généraliser au cas de variables aléatoires de lois continues quelconques :

Théorème 3.8. *Soit $(X_n)_{n \geqslant 1}$ une suite de variables aléatoires i.i.d. ayant une fonction de répartition F définie et continue sur un intervalle $[a,b]$, et soit a_n le processus empirique associé. Alors la suite $\left\{\frac{a_n}{b_n}\right\}$ est presque sûrement relativement compacte dans $\mathbb{B}([0,1])$ avec pour ensemble limite l'ensemble $\mathcal{H}_F$ des fonctions f définies sur $[a,b]$ et vérifiant :*

- *$f(a) = f(b) = 0$,*
- *f est absolument continue par rapport à F,*
- *$\int_a^b (df/dF)^2 \, dF \leqslant 1$ où df/dF est la dérivée de f par rapport à F.*

Chapitre 4

Vitesse de convergence presque complète pour des estimateurs non paramétriques dans un modèle de censure mixte

Plusieurs travaux dans la littérature statistique traitent de l'estimation non-paramétrique lorsque la variable d'intérêt est complète ou censurée à droite. Cependant, dans certaines études de fiabilité et d'analyse de survie, on peut se trouver face à des types de censure plus complexes.

Le modèle que nous nous proposons de traiter est le modèle I de Patilea et Rolin (2006). Dans ce modèle, la variable d'intérêt X est censurée à droite par une variable R, et $\min(X, R)$ est encore censurée à gauche par une variable L, ces variables étant supposés indépendantes.

Dans ce chapitre, nous montrons la convergence uniforme presque complète de l'estimateur de Patilea-Rolin, avec un taux $O\left(\sqrt{\log n/n}\right)$ sous les mêmes conditions que le théorème 3.6 mais en se passant de l'hypothèse de continuité des lois des variables latentes. Nous montrons aussi le même taux de convergence pour un estimateur de la fonction de hasard. Nous appliquons ces résultats pour obtenir les taux de convergence de quelques estimateurs à noyau de la densité et du taux de hasard.

Ces résultats ont été publiés dans la revue *Statistics & Probability Letters* (Kitouni *et al.*, 2015).

4.1 Estimation de la fonction de répartition

Nous commençons par rappeler l'estimateur produit-limite de la fonction de répartition d'une variable censurée à gauche. Soit L et Y deux variables aléatoires indépendantes positives représentant respectivement une durée d'intérêt et une variables de censure à gauche. Rappelons que dans ce modèle, au lieu d'observer L, on observe seulement un échantillon $(Z_i, \delta_i)_{1 \leqslant i \leqslant n}$ de (Z, δ) où $Z = \max(L, Y)$ et $\delta = 1_{\{L \geqslant Y\}}$. Notons par H la fonction de répartition de Z et par N la fonction de répartition de la sous distribution des données non-censurées, donnée par $N(t) = P(Z \leqslant t, \delta = 1)$. Considérons les versions empiriques de H et N données respectivement par

$$H_n(t) = \frac{1}{n} \sum_{i=1}^{n} 1_{\{Z_i \leqslant t\}} \text{ et } N_n(t) = \frac{1}{n} \sum_{i=1}^{n} 1_{\{Z_i \leqslant t, \delta_i = 1\}}. \tag{4.1}$$

Les relations $F_L(t) = \prod_{x>t}(1 - d\Gamma(x))$ et $\Gamma(t) = -\int_t^\infty \frac{dF_L}{F_L} = -\int_t^\infty \frac{dN}{H}$, suggèrent d'estimer F_L par

$$\tilde{F}_n(t) = \prod_{j/Z'_j > t} (1 - \Delta\Gamma_n(Z'_j)), \tag{4.2}$$

où $\Gamma_n(t) = -\int_t^\infty \frac{dN_n}{H_n}$ et $\left\{Z'_j,\ 1 \leqslant j \leqslant m\right\}$ sont les valeurs distinctes de $\{Z_i,\ 1 \leqslant i \leqslant n\}$ rangées dans l'ordre croissant.

Notons que ce même estimateur peut être retrouvé à partir de l'estimateur de Kaplan-Meier en inversant le temps, ce qui nous permet d'adapter le théorème 2.6 (voir aussi Bitouzé *et al.*, 1999) pour avoir

$$P\left(\sup_{I_Y < t} |\tilde{F}_n(t) - F_L(t)| > \sqrt{\frac{\log n}{n}}\right) \leqslant 2.5\, e^{-2\log n + C\sqrt{\log n}},$$

où C est une constante absolue. En choisissant $\alpha < 1$, nous obtenons pour n suffisamment grand $e^{C\sqrt{\log n}} \leqslant n^\alpha$, ce qui implique que

$$\sup_{I_Y < t} \left|\tilde{F}_n(t) - F_L(t)\right| = O_{a.co.}\left(\sqrt{\frac{\log n}{n}}\right). \tag{4.3}$$

Remarquons aussi qu'en vertu de l'inégalité DKW (théorème 2.5), nous avons

$$\sup_{t \in \mathbb{R}} |H_n(t) - H(t)| = O_{a.co.}\left(\sqrt{\frac{\log n}{n}}\right). \tag{4.4}$$

Considérons maintenant trois variables aléatoires indépendantes positives X, R et L, qui représentent respectivement la durée d'intérêt, une durée de censure à droite et une durée de censure à gauche (qui censure $\min(X, R)$). Nous sommes dans la situation où on peut seulement observer un échantillon $(\max(\min(X_i, R_i), L_i), A_i)$, $1 \leqslant i \leqslant n$ de $(\max(\min(X, R), L), A)$ où $A = 1_{\{L<R<X\}} + 2 \times 1_{\{\min(X,R)\leqslant L\}}$. C'est le modèle de censure double étudié dans Patilea et Rolin (2006).

Posons $Y = \min(X, R)$ et $Z = \max(Y, L) = \max(\min(X, R), L)$, et définissons la fonction de répartition de la sous distribution de H des données non censurées par $H^{(0)}(t) = P(Z \leqslant t, A = 0)$. Comme $H^{(0)}(t) = \int_0^t F_L(x^-)S_R(x^-)\, dF_X(x)$ et $H(t) = F_L(t)F_Y(t)$, la fonction de hasard cumulé Λ de X peut s'écrire

$$\Lambda(t) = \int_0^t \frac{dF_X(u)}{S_X(u^-)} = \int_0^t \frac{dH^{(0)}(u)}{F_L(u^-) - H(u^-)},$$

pour tout t vérifiant $I_L < t < T_R$. Soit $H_n^{(0)}$ la version empirique de $H^{(0)}$ donnée par

$$H_n^{(0)}(t) = \frac{1}{n} \sum_{1\leqslant i\leqslant n} 1_{\{Z_i\leqslant t, A_i=0\}}.$$

En utilisant le théorème 3.2 (voir aussi Kiefer, 1961), nous obtenons

$$\sup_{t\in\mathbb{R}} |H_n^{(0)}(t) - H^{(0)}(t)| = O_{a.co.}\left(\sqrt{\frac{\log n}{n}}\right). \tag{4.5}$$

Posons $D_{kj} = \sum_{i=1}^n 1_{\{Z_i=Z'_j, A_i=k\}}$, pour tout $1 \leqslant j \leqslant m$, et tout $k \in \{0, 1, 2\}$. D'après (4.2), F_L peut être estimée par

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

où H_n est donnée par (4.1). Ceci suggère d'estimer $\Lambda(t)$ par

$$\Lambda_n(t) = \int_0^t \frac{dH_n^{(0)}(u)}{\tilde{F}_n(u^-) - H_n(u^-)}, \tag{4.6}$$

ce qui conduit à estimer S_X par

$$1 - F_n(t) = \prod_{j/Z'_j\leqslant t} \left(1 - \frac{D_{0j}}{n\tilde{F}_n(Z'_{j-1}) - nH_n(Z'_{j-1})}\right).$$

On retrouve l'estimateur introduit par Patilea et Rolin (2006). Notre résultat concerne son taux de convergence presque complète sous l'hypothèse d'identifiabilité suivante

H1 $\max(I_L, I_R) < I_X$.

Comme L est une variable de censure à gauche, l'hypothèse $I_L < I_X$ est naturelle. Nous avons aussi besoin d'estimer F_L, et l'hypothèse $I_R < I_X$ est introduite pour permettre d'appliquer la formule (4.3) sur tout le support de la variable d'intérêt X. Remarquons que cette hypothèse a déja été utilisée dans Messaci and Nemouchi (2013).

Théorème 4.1. *Sous* **H1** *nous avons pour tout* $\theta < \min(T_X, T_R)$,

i) $\sup_{t\leqslant\theta} |\Lambda_n(t) - \Lambda(t)| = O_{a.co.}\left(\sqrt{\frac{\log n}{n}}\right)$,

ii) $\sup_{t\leqslant\theta} |F_n(t) - F_X(t)| = O_{a.co.}\left(\sqrt{\frac{\log n}{n}}\right)$.

Földes et Rejtő (1981b) donnent le même taux de convergence pour l'estimateur de Kaplan-Meier (cas de la seule censure à droite) sous la condition supplémentaire de la continuité de la fonction de répartition des variables latentes.

Démonstration. i) Nous avons

$$\begin{aligned}
|\Lambda_n(t) - \Lambda(t)| &= \left| \int_{I_X}^t \frac{dH_n^{(0)}(u)}{\tilde{F}_n(u^-) - H_n(u^-)} - \int_{I_X}^t \frac{dH^{(0)}(u)}{F_L(u^-) - H(u^-)} \right| \\
&\leqslant \int_{I_X}^t \left| \frac{1}{\tilde{F}_n(u^-) - H_n(u^-)} - \frac{1}{F_L(u^-) - H(u^-)} \right| dH_n^{(0)}(u) \\
&\quad + \left| \int_{I_X}^t \frac{1}{F_L(u^-) - H(u^-)} d\left(H_n^{(0)} - H^{(0)}(t)\right) \right| \\
&=: B_{n,1}(t) + B_{n,2}(t). \qquad (4.7)
\end{aligned}$$

Étudions ces deux termes séparément.

Comme $F_L(u^-) - H(u^-) = F_L(u^-)S_R(u^-)S_X(u^-)$, nous obtenons

$$\begin{aligned}
B_{n,1}(t) &= \int_{I_X}^{t} \frac{\left|F_L(u^-) - H(u^-) - \tilde{F}_n(u^-) + H_n(u^-)\right|}{|F_L(u^-) - H(u^-)|} \times \frac{dH_n^{(0)}(u)}{\left|\tilde{F}_n(u^-) + H_n(u^-)\right|} \\
&\leqslant \frac{\sup_{I_X \leqslant u \leqslant \theta} \left|F_L(u^-) - \tilde{F}_n(u^-) + H_n(u^-) - H(u^-)\right|}{F_L(I_X^-)S_R(\theta)S_X(\theta)} \\
&\qquad \times \int_{I_X}^{t} \frac{dH_n^{(0)}}{\tilde{F}_n(u^-) - H_n(u)} \\
&\leqslant 2\frac{\sup_{I_X \leqslant u \leqslant \theta} \left|F_L(u^-) - \tilde{F}_n(u^-) + H_n(u^-) - H(u^-)\right|}{F_L(I_X^-)S_R(\theta)S_X(\theta) \inf_{I_X \leqslant u \leqslant \theta} \left|\tilde{F}_n(u^-) - H_n(u^-)\right|}.
\end{aligned}$$

Pour $\varepsilon_0 \in]0, F_L(I_X^-)S_R(\theta)S_X(\theta)[$ donné, posons $\varepsilon = F_L(I_X^-)S_R(\theta)S_X(\theta) - \varepsilon_0$, alors

$$\begin{aligned}
&P\left(\inf_{I_X \leqslant u \leqslant \theta} |\tilde{F}_n(u^-) - H_n(u^-)| < \varepsilon_0\right) \\
&\qquad \leqslant P\left(\sup_{I_X \leqslant u \leqslant \theta} |F_L(u^-) - \tilde{F}_n(u^-) + H_n(u^-) - H(u^-)| > \varepsilon\right). \quad (4.8)
\end{aligned}$$

En effet, si nous supposons que $\inf_{I_X \leqslant t \leqslant \theta} \left|\tilde{F}_n(u^-) - H_n(u^-)\right| < \varepsilon_0$, alors il existe $t_0 \in [I_X, \theta]$ tel que $\tilde{F}_n(t_0^-) - H_n(t_0^-) \leqslant \varepsilon_0$, ce qui implique que

$$\begin{aligned}
F_L(t_0^-) - H(t_0^-) - \tilde{F}_n(t_0^-) + H_n(t_0^-) & \\
&= S_X(t_0^-)S_R(t_0^-)F_L(t_0^-) - \tilde{F}_n(t_0^-) + H_n(t_0^-) \\
&\geqslant S_X(\theta)S_R(\theta)F_L(I_X^-) - \varepsilon_0 = \varepsilon,
\end{aligned}$$

donc

$$\sup_{I_X \leqslant t \leqslant \theta} \left|F_L(t^-) - H(t^-) - \tilde{F}_n(t^-) + H_n(t^-)\right| > \varepsilon.$$

Sous l'hypothèse **H1**, le membre de droite de (4.8) est le terme général d'une série convergente ; ceci permet, avec (4.3) et (4.4), de conclure que

$$\sup_{I_X \leqslant t \leqslant \theta} B_{n,1}(t) = O_{a.co.}\left(\sqrt{\frac{\log n}{n}}\right). \quad (4.9)$$

D'autre part, en intégrant par parties, nous obtenons

$$\begin{aligned}
\mathrm{B}_{n,2}(t) &= \left| \int_{\mathrm{I_X}} \frac{1}{\mathrm{F_L}(u^-) - \mathrm{H}(u^-)} d\left(\mathrm{H}_n^{(0)} - \mathrm{H}^{(0)}(t)\right) \right| \\
&\leqslant \left| \frac{\mathrm{H}_n^{(0)}(t) - \mathrm{H}^{(0)}(t)}{\mathrm{F_L}(t) - \mathrm{H}(t)} \right| + \left| \frac{\mathrm{H}_n^{(0)}(\mathrm{I_X}) - \mathrm{H}^{(0)}(\mathrm{I_X})}{\mathrm{F_L}(\mathrm{I_X}) - \mathrm{H}(\mathrm{I_X})} \right| \\
&\quad + \left| \int_{\mathrm{I_X}}^{t} (\mathrm{H}_n^{(0)}(u) - \mathrm{H}^{(0)}(u))\, d\left(\frac{1}{\mathrm{F_L}(u) - \mathrm{H}(u)} \right) \right|
\end{aligned}$$

De plus, en utilisant le fait que $\mathrm{F_L}(u) - \mathrm{H}(u) = \mathrm{F_L}(u)\mathrm{S_R}(u)\mathrm{S_X}(u)$, nous obtenons

$$\begin{aligned}
\mathrm{B}_{n,2}(t) &\leqslant \frac{2}{\mathrm{F_L}(\mathrm{I_X})\mathrm{S_R}(\theta)\mathrm{S_X}(\theta)} \sup_{\mathrm{I_X} \leqslant u \leqslant \theta} \left| \mathrm{H}_n^{(0)}(u) - \mathrm{H}^{(0)}(u) \right| \\
&\quad + \left| \int_{\mathrm{I_X}}^{t} \frac{\mathrm{H}_n^{(0)}(u) - \mathrm{H}^{(0)}(u)}{\mathrm{F_L}(u^-)} d\left(\frac{1}{\mathrm{S_R}(u)\mathrm{S_X}(u)} \right) \right| \\
&\quad + \left| \int_{\mathrm{I_X}}^{t} \frac{\mathrm{H}_n^{(0)}(u) - \mathrm{H}^{(0)}(u)}{\mathrm{S_R}(u)\mathrm{S_X}(u)} d\left(\frac{1}{\mathrm{F_L}(u)} \right) \right| \\
&\leqslant \mathrm{D} \sup_{\mathrm{I_X} \leqslant u \leqslant \theta} \left| \mathrm{H}_n^{(0)}(u) - \mathrm{H}^{(0)}(u) \right|,
\end{aligned}$$

où D est une constante déterministe. En combinant cela avec (4.5), (4.7) et (4.9), nous obtenons le résultat désiré.

ii) En utilisant l'équation de Duhamel (1.14), nous avons pour tout $t \leqslant \theta < \min(\mathrm{T_R}, \mathrm{T_X})$,

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

Posons $\mathrm{M}_n(t) = \int_{\mathrm{I_X}}^{t} \frac{1-\mathrm{F_X}(u^-)}{1-\mathrm{F_X}(u)} d(\Lambda_n - \Lambda)(u)$; en intégrant par parties nous obtenons

$$\begin{aligned}
|\mathrm{F}_n(t) - \mathrm{F_X}(t)| &\leqslant \left| \int_{\mathrm{I_X}}^{t} \frac{1 - \mathrm{F}_n(u^-)}{1 - \mathrm{F_X}(u^-)}\, d\mathrm{M}_n(u) \right| \\
&= \left| \frac{1 - \mathrm{F}_n(t)}{1 - \mathrm{F_X}(t)} \mathrm{M}_n(t) - \int_{\mathrm{I_X}}^{t} \mathrm{M}_n(u)\, d\left(\frac{1 - \mathrm{F}_n(u)}{1 - \mathrm{F_X}(u)} \right) \right|.
\end{aligned}$$

En utilisant le fait que

$$d\left(\frac{1 - \mathrm{F}_n(u)}{1 - \mathrm{F_X}(u)} \right) = \frac{(1 - \mathrm{F_X}(u))\,(-\mathrm{dF}_n(u)) - (1 - \mathrm{F}_n(u))\,(-\mathrm{dF_X}(u^-))}{(1 - \mathrm{F_X}(u))\,(1 - \mathrm{F_X}(u^-))},$$

nous obtenons

$$
\begin{aligned}
|\mathrm{F}_n(t) - \mathrm{F}_X(t)| &\leqslant \frac{1}{1-\mathrm{F}_X(\theta)} |\mathrm{M}_n(t)| + \left| \int_{\mathrm{I}_X}^{t} \mathrm{M}_n(u) \frac{d\mathrm{F}_n(u)}{1-\mathrm{F}_X(u)} \right| \\
&\quad + \left| \int_{\mathrm{I}_X}^{t} \mathrm{M}_n(u) \left(1-\mathrm{F}_n(u^-)\right) \frac{d\mathrm{F}_X(u)}{\left(1-\mathrm{F}_X(u)\right)\left(1-\mathrm{F}_X(u^-)\right)} \right| \\
&\leqslant \frac{1}{1-\mathrm{F}_X(\theta)} |\mathrm{M}_n(t)| + \frac{1}{1-\mathrm{F}_X(\theta)} \sup_{\mathrm{I}_X \leqslant u \leqslant \theta} |\mathrm{M}_n(u)| \, \mathrm{F}_n(t) \\
&\quad + \frac{1}{\left(1-\mathrm{F}_X(\theta)\right)^2} \sup_{\mathrm{I}_X \leqslant u \leqslant \theta} |\mathrm{M}_n(u)| \, \mathrm{F}_X(t) \\
&\leqslant \frac{2\left(1-\mathrm{F}_X(\theta)\right)+1}{\left(1-\mathrm{F}_X(\theta)\right)^2} \sup_{\mathrm{I}_X \leqslant u \leqslant \theta} |\mathrm{M}_n(u)| \, .
\end{aligned}
$$

Il reste à traiter le terme $\sup_{\mathrm{I}_X \leqslant u \leqslant \theta} |\mathrm{M}_n(u)|$.

$$
\begin{aligned}
|\mathrm{M}_n(t)| &\leqslant |\Lambda_n(t) - \Lambda(t)| + \left| \sum_{u \leqslant t / \Delta \mathrm{F}_X(u) > 0} \frac{\Delta \mathrm{F}_X(u)}{1-\mathrm{F}_X(u)} \left[\Delta\Lambda_n(u) - \Delta\Lambda(u)\right] \right| \\
&\leqslant |\Lambda_n(t) - \Lambda(t)| + 2 \sup_{\mathrm{I}_X \leqslant u \leqslant \theta} |\Lambda_n(u) - \Lambda(u)| \frac{1}{1-\mathrm{F}_X(\theta)} \mathrm{F}_X(t).
\end{aligned}
$$

Donc

$$
\sup_{\mathrm{I}_X \leqslant t \leqslant \theta} |\mathrm{M}_n(t)| \leqslant \frac{3-\mathrm{F}_X(\theta)}{1-\mathrm{F}_X(\theta)} \sup_{\mathrm{I}_X \leqslant t \leqslant \theta} |\Lambda_n(t) - \Lambda(t)| \, .
$$

Le résultat s'ensuit directement en utilisant i). □

4.2 Estimation de la densité et du taux de hasard

4.2.1 Estimation à noyau de la densité

Dans cette section nous supposons que la variable doublement censurée X a une densité de probabilité notée f. Nous l'estimons, par analogie avec le cas des données complètes et des données censurées à droite (voir Földes *et al.*, 1981 ; Parzen, 1962 ; Rosenblatt, 1956), comme suit

$$
f_n(x) = \frac{1}{h_n} \int \mathrm{K}\left(\frac{x-y}{h_n}\right) d\mathrm{F}_n(y),
$$

où K est le noyau et h_n la fenêtre. Nous précisons le taux de convergence uniforme presque complète de f_n sur un compact $\mathrm{C} \subset [0, \min(\mathrm{T}_X, \mathrm{T}_R)[$,

sous des conditions usuelles dans le cadre de l'estimation non paramétrique.

H2 Il existe un entier $r \geqslant 2$ tel que f est r fois continûment différentiable autour de C.

H3 $\exists\, \alpha > 0, \beta > 0, \varepsilon > 0,\ \forall x \in \mathrm{C},\ \forall y \in\ ]x-\varepsilon, x+\varepsilon[,\ |f(x)-f(y)| \leqslant \alpha|x-y|^{\beta}$.

H4 $h_n \to 0$ and $nh_n^2/\log n \to \infty$.

H5 K est une fonction continue à droite, à variations bornées, vérifiant $\int \mathrm{K}(t)\,dt = 1$, et telle que $\exists \mathrm{M} > 0,\ \forall u \in \mathbb{R},\ |u| \geqslant \mathrm{M} \Rightarrow \mathrm{K}(u) = 0$.

H6 $\forall 1 \leqslant j \leqslant r-1, \int u^j \mathrm{K}(u)\,du = 0$.

Le comportement asymptotique de f_n est décrit dans le théorème suivant.

Théorème 4.2. *i) Sous* **H1**, **H2** *et* **H4–H6**, *nous avons*

$$\sup_{x\in\mathrm{C}} |f_n(x) - f(x)| = \mathrm{O}_{a.co.}\left(h_n^r + \frac{1}{h_n}\sqrt{\frac{\log n}{n}}\right).$$

ii) Sous **H1** *et* **H3–H5**, *nous avons*

$$\sup_{x\in\mathrm{C}} |f_n(x) - f(x)| = \mathrm{O}_{a.co.}\left(h_n^{\beta} + \frac{1}{h_n}\sqrt{\frac{\log n}{n}}\right).$$

Démonstration. Considérons la décomposition usuelle

$$|f_n(x) - f(x)| \leqslant |f_n(x) - \mathbb{E}f_n(x)| + |\mathbb{E}f_n(x) - f(x)|,$$

où $\mathbb{E}f_n(x) = \frac{1}{h_n}\int \mathrm{K}(\frac{x-y}{h_n})d\mathrm{F}(y)$.

Étudions chaque terme séparément. Premièrement nous avons pour tout $x \in \mathrm{C}$

$$|f_n(x) - \mathbb{E}f_n(x)| = \frac{1}{h_n}\left|\int \mathrm{K}\left(\frac{x-y}{h_n}\right) d(\mathrm{F}_n(y) - \mathrm{F}(y))\right|.$$

L'intégration par partie donne

$$|f_n(x) - \mathbb{E}f_n(x)| \leqslant \frac{\mathrm{V}_\mathrm{K}}{h_n} \sup_{u>-\mathrm{M}} |\mathrm{F}_n(x-uh_n) - \mathrm{F}(x-uh_n)|,$$

où V_K est la variation totale de K sur $\mathbb{R}$. Posons $\theta = \max(\mathrm{C})$, et $\theta^* \in\ ]\theta, \min(\mathrm{T}_\mathrm{X}, \mathrm{T}_\mathrm{R})[$. Comme $h_n \to 0$, il s'ensuit que $h_n < \frac{\theta^*-\theta}{\mathrm{M}}$ pour n suffisamment grand, d'où

$$\sup_{x\in\mathrm{C}} |f_n(x) - \mathbb{E}f_n(x)| \leqslant \frac{\mathrm{V}_\mathrm{K}}{h_n} \sup_{t<\theta^*} |\mathrm{F}_n(t) - \mathrm{F}(t)| = \mathrm{O}_{a.co.}\left(\frac{1}{h_n}\sqrt{\frac{\log n}{n}}\right),$$

d'après le théorème 4.1, ii).

D'autre part, d'après l'hypothèse **H5** nous avons

$$\mathbb{E}f_n(x) - f(x) = \int_{-M}^{M} K(u)\,(f(x - uh_n) - f(x))\,du.$$

i) Sous les hypothèse **H2** et **H6**, le développement de Taylor donne

$$|\mathbb{E}f_n(x) - f(x)| = \left|\frac{h_n^r}{r!}\int_{-M}^{M} f^{(r)}(\eta)u^r K(u)\,du\right|,$$

où η est compris entre x and $x - uh_n$. Donc $\sup_{x\in C}|\mathbb{E}f_n(x) - f(x)| = O(h_n^r)$, en combinant les hypothèses **H2**, **H4** et la compacité de C.

ii) Sous les hypothèses **H3** et **H5**

$$\begin{aligned}\sup_{x\in C}|\mathbb{E}f_n(x) - f(x)| &\leqslant \alpha\int_{-M}^{M}|K(u)||u|^\beta h_n^\beta du\\ &\leqslant \alpha M^\beta h_n^\beta\int |K(u)|\,du = O(h_n^\beta).\end{aligned}$$ □

4.2.2 Estimation du taux de hasard

Le taux de hasard de X est défini par $\lambda(x) = \frac{f(x)}{S_X(x)}$ si $S_X(x) \neq 0$ et $\lambda(x) = 0$ sinon. Plusieurs estimateurs de λ ont été proposés dans la littérature (Diehl et Stute, 1988 ; Földes *et al.*, 1981 ; Xiang, 1994). Nous avons choisi d'étudier deux estimateurs de λ.

Le premier est donné par $\lambda_n(x) = f_n(x)/(1 - F_n(x) + u_n)$, où $(u_n)_{n\in\mathbb{N}}$ est une suite de nombres réels strictement positifs tendant vers zéro. Remarquons que Földes *et al.* (1981) ont étudié un estimateur analogue dans le cadre de la censure à droite avec $u_n = 1/n$.

Le second estimateur est obtenu en suivant la même idée que pour estimer la densité dans la sous section précédente, et est une adaptation de l'estimateur étudié dans (Diehl et Stute, 1988) au modèle de censure double. Il est donné par

$$\tilde{\lambda}_n(x) = \frac{1}{h_n}\int K\left(\frac{x-y}{h_n}\right)d\Lambda_n(y),$$

où Λ_n est définie en (4.6).

Le théorème suivant donne la convergence de λ_n ; il est obtenu à partir des théorèmes 4.1 et 4.2.

Théorème 4.3. *i) Sous les hypothèses* **H1**, **H2** *et* **H4–H6** *et si on choisit* $u_n = \mathrm{O}\left(h_n^r + h_n^{-1}\sqrt{\log n/n}\right)$, *alors*

$$\sup_{x\in\mathrm{C}} |\lambda_n(x) - \lambda(x)| = \mathrm{O}_{a.co.}\left(h_n^r + \frac{1}{h_n}\sqrt{\frac{\log n}{n}}\right).$$

ii) Sous **H1** *et* **H3–H5** *et si on choisit* $u_n = \mathrm{O}\left(h_n^\beta + h_n^{-1}\sqrt{\log n/n}\right)$, *alors*

$$\sup_{x\in\mathrm{C}} |\lambda_n(x) - \lambda(x)| = \mathrm{O}_{a.co.}\left(h_n^\beta + \frac{1}{h_n}\sqrt{\frac{\log n}{n}}\right).$$

Démonstration. Nous avons pour tout $x \in \mathrm{C}$

$$\begin{aligned}
|\lambda_n(x) - \lambda(x)| &\leqslant \frac{1}{1 - \mathrm{F}_n(x) + u_n}|f_n(x) - f(x)| \\
&\quad + |1 - \mathrm{F}_n(x) - \mathrm{S}_\mathrm{X}(x) + u_n| \frac{f(x)}{\mathrm{S}_\mathrm{X}(x)(1 - \mathrm{F}_n(x) + u_n)} \\
&\leqslant \frac{1}{\inf_{x\in\mathrm{C}}(1 - \mathrm{F}_n(x) + u_n)} \sup_{x\in\mathrm{C}} |f_n(x) - f(x)| \\
&\quad + \frac{\sup_{x\in\mathrm{C}} f(x) \sup_{x\in\mathrm{C}} |1 - \mathrm{F}_n(x) - \mathrm{S}_\mathrm{X}(x) + u_n|}{\mathrm{S}_\mathrm{X}(\theta) \inf_{x\in\mathrm{C}}(1 - \mathrm{F}_n(x) + u_n)}, \qquad (4.10)
\end{aligned}$$

où $\theta = \max(\mathrm{C})$. De plus, pour $\gamma \in \left]0, \mathrm{S}_\mathrm{X}(\theta)\right[$ nous avons

$$\mathrm{P}\left(\inf_{x\in\mathrm{C}}(1 - \mathrm{F}_n(x) + u_n) \leqslant \frac{\gamma}{2}\right) \leqslant \mathrm{P}\left(\sup_{x\in\mathrm{C}} |1 - \mathrm{F}_n(x) - \mathrm{S}_\mathrm{X}(x) + u_n| > \frac{\gamma}{2}\right),$$

d'où

$$\sum_{n=1}^{\infty} \mathrm{P}\left(\inf_{x\in\mathrm{C}}(1 - \mathrm{F}_n(x) + u_n) \leqslant \frac{\gamma}{2}\right) < \infty. \qquad (4.11)$$

Les résultats du théorèmes découlent alors directement de (4.10), (4.11) et des théorèmes 4.1 et 4.2. □

Le taux de convergence presque complète de $\tilde{\lambda}_n$ peut être obtenu de la même manière que pour f_n. Pour ce faire, nous avons besoin de l'hypothèse suivante

H'3 $\exists\, \alpha > 0, \beta > 0, \varepsilon > 0,\ \forall x \in \mathrm{C},\ \forall y \in \left]x - \varepsilon, x + \varepsilon\right[,\ |\lambda(x) - \lambda(y)| \leqslant \alpha|x - y|^\beta$.

En procédant de la même manière que pour la démonstration du théorème 4.2, en remplaçant $\mathbb{E}f_n$ par $\mathbb{E}\tilde{\lambda}_n(x) = \frac{1}{h_n}\int \mathrm{K}\left(\frac{x-y}{h_n}\right) d\Lambda(y)$, et en utilisant la partie *i)* du théorème 4.1 au lieu de *ii)*, nous obtenons le résultat suivant.

Théorème 4.4. *i) Sous* **H1**, **H2** *et* **H4–H6** *nous avons*

$$\sup_{x\in C}|\tilde{\lambda}_n(x)-\lambda(x)| = O_{a.co.}\left(h_n^r + \frac{1}{h_n}\sqrt{\frac{\log n}{n}}\right).$$

ii) Sous **H1**, **H'3**, **H4** *et* **H5** *nous avons*

$$\sup_{x\in C}|\tilde{\lambda}_n(x)-\lambda(x)| = O_{a.co.}\left(h_n^\beta + \frac{1}{h_n}\sqrt{\frac{\log n}{n}}\right).$$

Dans la partie *i)* de chacun des théorèmes 4.2, 4.3 et 4.4, et pour le choix optimal de h_n donné par $h_n = (\log n/n)^{1/2r+2}$, nous obtenons un taux de convergence presque complète de l'ordre de $(\log n/n)^{r/2r+2}$. Remarquons que pour les données censurées à droite, Diehl et Stute (1988) et Xiang (1994) ont obtenu, sous des conditions plus fortes que les nôtres, des taux de convergence presque sûre de l'ordre de $(\log n/n)^{r/2r+1}$.

Chapitre 5

Loi du logarithme itéré pour le processus empirique dans un modèle de censure mixte

L'étude de la fonction de répartition empirique est un sujet important en statistique. Sa normalisation naturelle conduit à la définition du processus empirique, et les accroissements de ce dernier constituent ce que l'on appelle le processus empirique local.

Une des raisons motivant l'étude des processus empiriques est le fait qu'ils jouent un rôle essentiel pour établir des lois limites pour des statistiques qui peuvent être exprimées comme des fonctionnelles locales de ces processus. Des exemples typiques de telles statistiques sont les estimateurs de la densité et du taux de hasard. Pour plus de détails à ce sujet pour des données complètes ou censurées à droite, voir Deheuvels et Mason (1992), Deheuvels et Einmahl (1996, 2000) et les références qu'ils citent.

Le but du présent chapitre est de donner une loi du logarithme itéré pour le processus empirique local au voisinage d'un point fixé, dans le cas de la censure mixte selon le modèle de Patilea et Rolin (2006).

Ce travail à été soumis pour publication.

5.1 Estimateur Produit-limite

Notons par $F_V(t) = P(V \leqslant t)$ la fonction de répartition d'une variable aléatoire réelle V et par $S_V = 1 - F_V$ sa fonction de survie. Notons

par ailleurs I_V ($= \inf\{t : F_V(t) > 0\}$) et T_V ($= \sup\{t : F_V(t) < 1\}$) respectivement les points initial et terminal de la distribution F_V. Pour toute fonction R, posons $R(t^-) = \lim_{\varepsilon \downarrow 0} R(t - \varepsilon)$ si cette limite existe. Pour tout ensemble $C \subseteq \mathbb{R}$, notons $\mathbb{B}(C)$ l'ensemble des fonctions réelle bornées définies sur C muni de la topologie de la convergence uniforme.

Considérons les variables aléatoires réelles positives indépendantes X, R et L, représentant respectivement une durée de survie, une variable de censure à droite et une variable de censure à gauche. Dans le modèle I de Patilea et Rolin (2006), on observe un échantillon $(Z_i, \delta_i)_{1 \leqslant i \leqslant n}$ de variables aléatoires indépendantes de même loi que (Z, δ) où $Z = \max(\min(X, R), L)$ et

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

Considérons les fonctions de répartition des sous-lois de Z définies pour $k = 0, 1, 2$ par $H^{(k)}(t) = P(Z \leqslant t, \delta = k)$. On a les relations

$$\begin{aligned} H^{(0)}(t) &= \int_0^t F_L(u_-) S_R(u^-)\, dF_X(u), \\ H^{(1)}(t) &= \int_0^t F_L(u^-) S_X(u)\, dF_R(u), \\ H^{(2)}(t) &= \int_0^t \{1 - S_X(u) S_R(u)\}\, dF_L(u), \end{aligned} \tag{5.1}$$

et la fonction de répartition de Z est $H = H^{(0)} + H^{(1)} + H^{(2)}$. Posons $Y = \min(X, R)$ et $H^{(01)} = H^{(0)} + H^{(1)}$, alors $H^{(01)}(t) = \int_0^t F_L(u^-)\, dF_Y(u)$ et $H^{(2)}(t) = \int_0^t F_Y(u)\, dF_L(u)$.

Considérons d'abord l'estimation dans un modèle de censure à gauche. Pour ce faire, considérons les mesures de hasard inverse

$$M_2(dt) = \frac{dH^{(2)}(t)}{H(t)} \quad ; \quad M_{01}(dt) = \frac{dH^{(01)}(t)}{H(t^-) + \Delta H^{(01)}(t)},$$

(avec $\Delta H^{(01)}(t) = H^{(01)}(t) - H^{(01)}(t^-)$) et soit $F^{(2)}$ et $F^{(01)}$ (resp. $S^{(2)}$ et $S^{(01)}$) les fonctions de répartition (resp. les fonctions de survie) associées, qui peuvent être directement estimées à partir des données observées [1]. Nous obtenons

$$H(t) = F^{(2)}(t) F^{(01)}(t). \tag{5.2}$$

1. Si M est une mesure de hasard inverse, alors la fonction de répartition associée est donnée par $F(t) = \prod_{]t,\infty]} (1 - M(ds))$.

Les équations (5.1) et la définition de S_X impliquent que

$$\frac{dH^{(0)}(t)}{F_L(t^-)S_Y(t^-)} = \frac{dF_X(t)}{S_X(t^-)},$$

ce qui suggère de définir la mesure de hasard suivante

$$\Lambda(dt) = \frac{dH^{(0)}(t)}{F^{(2)}(t^-)S^{(01)}(t^-)}. \tag{5.3}$$

Soit F_X^I sa fonction de répartition associée.

Pour calculer l'estimateur de F_X^I, soit H_n, $H_n^{(0)}$, $H_n^{(1)}$ et $H_n^{(2)}$ les versions empiriques de H, $H^{(0)}$, $H^{(1)}$ et $H^{(2)}$ respectivement, données par

$$H_n(t) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{Z_i \leqslant t\}}, \quad H_n^{(k)}(t) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{Z_i \leqslant t, \delta_i = k\}}, \text{ for } k = 0, 1, 2, \tag{5.4}$$

et soit $F_n^{(2)}$, $S_n^{(01)}$, Λ_n et F_n les fonctions obtenues en remplaçant $H^{(0)}$, $H^{(1)}$ et $H^{(2)}$ par leurs versions empiriques dans l'expression de $F^{(2)}$, $S^{(01)}$, Λ et F_X^I respectivement. F_n est l'estimateur de F_X proposé par Patilea et Rolin (2006) (Notons que $F_X = F_X^I$ sous certaines conditions d'identifiabilité). Son expression est rappelée ci-dessous.

Notons par $\{Z'_j, 1 \leqslant j \leqslant M\}$ les valeurs distinctes de $\{Z_i, 1 \leqslant i \leqslant n\}$ rangées dans l'ordre croissant et posons $D_{kj} = \sum_{i=1}^{n} 1_{\{Z_i = Z'_j, A_i = k\}}$. Alors,

$$1 - F_n(t) = \prod_{j/Z'_j \leqslant t} \left\{1 - D_{0j}/(U_{j-1} - nH_n(Z'_{j-1})\right\}, \tag{5.5}$$

où $U_{j-1} = n\prod_{j \leqslant l \leqslant M}\{1 - D_{2l}/(nH_n(Z'_l))\}$. On peut voir que $1 - F_n$ généralise l'estimateur de Kaplan-Meier (Kaplan et Meier, 1958).

5.2 Résultats

L'estimateur défini en (5.5) engendre un processus empirique défini en posant pour tout $t \in \mathbb{R}$,

$$a_n(t) = \sqrt{n}(F_n(t) - F_X(t)). \tag{5.6}$$

Soit $(h_n)_{n \geqslant 0}$ une suite de nombres strictement positifs vérifiant les hypothèses suivantes quand $n \to \infty$.

H1 $h_n \downarrow 0$ et $nh_n \uparrow \infty$.

H2 $nh_n / \log\log n \to \infty$.

Nous avons aussi besoin de l'hypothèse d'identifiabilité suivante.

H3 $\max(\mathrm{I_L}, \mathrm{I_R}) < \mathrm{I_X}$ et $\mathrm{T_X} < \mathrm{T_R}$.

Soit $\mathrm{M} > 0$ fixé et $z \in]\mathrm{I_X}, \mathrm{T_X}[$. Posons $b_n = \sqrt{2h_n \log\log n}$ et définissons les accorissements de (a_n), aussi appelé processus empirique local, pour tout $u \in [-\mathrm{M}, \mathrm{M}]$ par

$$\xi_n(u) = \frac{1}{b_n}(a_n(z + h_n u) - a_n(z)). \tag{5.7}$$

Le résultat principal de ce chapitre est le suivant.

Théorème 5.1. *Soit* $z \in]\mathrm{I_X}, \mathrm{T_X}[$. *Supposons que* $\mathrm{F_X}$, $\mathrm{F_L}$ *et* $\mathrm{F_R}$ *sont continues et que la dérivée* f_X *de* $\mathrm{F_X}$ *au point* z *existe. Alors sous les hypothèses* **(H1)**, **(H2)** *et* **(H3)** *la suite* $\{\xi_n, n \geqslant 1\}$ *est presque sûrement relativement compacte dans* $\mathbb{B}([-\mathrm{M}, \mathrm{M}])$ *avec pour ensemble limite l'ensemble de toutes les fonctions* h *de la forme*

$$h(u) = \int_0^u \psi(s)\, ds \quad \textit{pour } u \in [-\mathrm{M}, \mathrm{M}], \textit{ où } \int_{-\mathrm{M}}^{\mathrm{M}} \psi^2(s)\, ds \leqslant \frac{f_\mathrm{X}(z)}{\mathrm{F_L}(z)\mathrm{S_R}(z)}. \tag{5.8}$$

Ce résultat est similaire au théorème 1.2 de Deheuvels et Einmahl (1996) où les données sont censurées à droite seulement.

L'intérêt d'un tel résultat est qu'il permet d'obtenir des taux de convergences pour les estimateurs qui dépendent des accroissements de la fonction de répartition ou d'un estimateur de cette dernière. Un exemple d'un tel estimateur est l'estimateur à noyau de la densité, introduit par Rosenblatt (1956) :

$$f_n(x) = \int_{-\infty}^{+\infty} \frac{1}{h_n} \mathrm{K}\left(\frac{y-u}{h_n}\right) d\mathrm{G}_n(u),$$

où G_n est la fonction de répartition empirique, K est appelé noyau (ou fonction poids), et (h_n) est une suite de nombres strictement positifs appelé fenêtre. En remplaçant G_n par l'estimateur défini en (5.5), nous obtenons un estimateur à noyau de la densité dans le cas des données doublement censurées. L'application du théorème 5.1 pour obtenir le taux de convergence presque sûre de cet estimateur est présentée à la section 5.3.

Le reste de cette section est dédié à la démonstration du théorème 5.1. Cette démonstration est divisée en plusieurs lemmes dont les démonstrations sont présentées à la section 5.4.

Comme est souvent le cas pour les démonstrations concernant les processus empiriques, nous utilisons la réduction au cas uniforme sur $[0,1]$. En s'inspirant de l'article de Deheuvels et Einmahl (1996), nous construisons une suite de variables aléatoires de loi uniforme sur $[0,1]$ telle que la suite des observations (Z_i, δ_i) peut s'écrire en fonction de cette suite. Pour ce faire, posons $p = P(\delta = 0)$ et $q = P(\delta = 1)$, et supposons que que $p > 0$, $q > 0$ et $p + q < 1$. Remarquons que le cas $p + q = 1$ correspond au cas de la censure à droite, et $q = 0$ correspond au cas de la censure à gauche qui peut aussi être déduit du résultat de Deheuvels et Einmahl (1996) par inversion du temps.

Considérons $Q^{(0)}$, $Q^{(1)}$ et $Q^{(2)}$ les fonctions quantile de $H^{(0)}$, $H^{(1)}$ et $H^{(2)}$ respectivement ; c'est à dire

$$\begin{aligned}
Q^{(0)}(s) &= \inf\{x : H^{(0)}(x) \geqslant s\}, \qquad 0 < s < p;\\
Q^{(1)}(s) &= \inf\{x : H^{(1)}(x) \geqslant s\}, \qquad 0 < s < q;\\
Q^{(2)}(s) &= \inf\{x : H^{(2)}(x) \geqslant s\}, \qquad 0 < s < 1 - p - q.
\end{aligned}$$

Ces définitions impliquent

$$\begin{aligned}
Q^{(0)}(s) \leqslant x &\iff s \leqslant H^{(0)}(x) \qquad \text{pour } 0 < s < p;\\
Q^{(1)}(s) \leqslant x &\iff s \leqslant H^{(1)}(x) \qquad \text{pour } 0 < s < q;\\
Q^{(2)}(s) \leqslant x &\iff s \leqslant H^{(2)}(x) \qquad \text{pour } 0 < s < 1 - p - q.
\end{aligned} \tag{5.9}$$

Nous pouvons maintenant énoncer le résultat donnant la réduction au cas uniforme.

Lemme 5.1. *Sur un espace de probabilité suffisamment riche, on peut définir une variable aléatoire* U *de loi* $\mathcal{U}([0,1])$ *telle que, presque sûrement,* $\delta = 1_{\{p<U<1\}} + 1_{\{p+q<U<1\}}$ *et*

$$Z = \begin{cases} Q^{(0)}(U), & \text{si } 0 < U < p;\\ Q^{(1)}(U - p), & \text{si } p < U < p + q;\\ Q^{(2)}(U - p - q), & \text{si } p + q < U < 1. \end{cases} \tag{5.10}$$

Nous pouvons ainsi considérer une suite (U_n) de variables aléatoire de loi $\mathcal{U}([0,1])$, vérifiant pour tout $i \geqslant 1$

$$\begin{aligned}
\delta_i &= 1_{\{p<U_i<1\}} + 1_{\{p+q<U_i<1\}};\\
Z_i &= \begin{cases} Q^{(0)}(U_i), & \text{si } 0 < U_i < p;\\ Q^{(1)}(U_i - p), & \text{si } p < U_i < p + q;\\ Q^{(2)}(U_i - p - q), & \text{si } p + q < U_i < 1. \end{cases}
\end{aligned} \tag{5.11}$$

Soit $\mathbb{U}_n(s) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{U_i \leqslant s\}}$ la fonction de répartition empirique de cette suite et $\alpha_n(s) = \sqrt{n}\,(\mathbb{U}_n(s) - s)$ le processus empirique associé. D'après (5.4), (5.9) et (5.11), nous pouvons écrire $H_n^{(0)}$ comme

$$H_n^{(0)}(x) = \mathbb{U}_n\left(H^{(0)}(x)\right), \qquad \text{pour } 0 < H^{(0)}(x) < p. \tag{5.12}$$

Nous démontrons d'abord un résultat similaire au théorème 5.1 pour la sous distribution de données non censurées. En d'autres termes, nous commençons par étudier le processus empirique k_n obtenu en remplaçant F_n et F_X dans l'expression de ξ_n (voir (5.7)) par $H_n^{(0)}$ et $H^{(0)}$ respectivement. Ceci est justifié par le fait que les sauts de l'estimateur de Patilea-Rolin surviennent uniquement au points où l'on a une observations non censurée. C'est le théorème 5.2 ci-dessous. Ensuite, nous montrons à travers une succession d'approximations (Lemmes 5.5 à 5.7) que la compacité relative presque sûre de ξ_n peut être déduite de celle de k_n.

Le lemme suivant est un cas particulier du théorème 1.1 de Deheuvels et Mason (1994). C'est le point de départ de notre travail. Posons $s_0 \in]0, p[$ and $M_1 > 0$, et considérons le processus des accroissements du processus α_n défini par

$$f_n(u) = \frac{1}{b_n}\left(\alpha_n(s_0 + h_n u) - \alpha_n(s_0)\right), \text{ pour } -M_1 < u < M_1. \tag{5.13}$$

Lemme 5.2. *Sous les hypothèses* **(H1)** *et* **(H2)**, *la suite* (f_n) *est presque sûrement relativement compacte dans* $\mathbb{B}([-M_1, M_1])$ *avec pour ensemble limite l'ensemble de toutes les fonctions* f *vérifiant*

$$f(u) = \int_0^u \psi_0(s)ds; \quad -M_1 \leqslant u \leqslant M_1; \quad \text{où } \int_{-M_1}^{M_1} \psi_0^2(s)ds \leqslant 1.$$

Dans la suite, nous utilisons une version modifiée de ce lemme. Considérons $M > 0$ fixé et soit (γ_n) une suite de fonctions définies sur $[-M, M]$ vérifiant

$$\lim_{n\to\infty} \sup_{u\in[-M,M]} |\gamma_n(u) - \gamma_0 u| = 0, \tag{5.14}$$

où γ_0 est une constante strictement positive. Pour tout $u \in [-M, M]$, posons $g_n(u) = f_n\left(\gamma_n(u)\right) = \frac{1}{b_n}\left(\alpha_n\left(s_0 + h_n\gamma_n(u)\right) - \alpha_n(s_0)\right)$.

Lemme 5.3. *Sous les hypothèses* **(H1)** *et* **(H2)**, *la suite* (g_n) *est presque sûrement relativement compacte dans* $\mathbb{B}([-M, M])$ *avec pour ensemble limite l'ensemble de toutes les fonctions* g *vérifiant*

$$g(u) = \int_0^u \phi_0(s)\,ds; \text{ pour } -M \leqslant u \leqslant M, \text{ où } \int_{-M}^{M} \phi_0^2(s)\,ds \leqslant \gamma_0. \tag{5.15}$$

Posons $s_0 = \mathrm{H}^{(0)}(z)$, et pour $\mathrm{M} > 0$ fixé, considérons la suite de fonctions aléatoires définies pour $-\mathrm{M} < u < \mathrm{M}$ par

$$k_n(u) = \frac{\sqrt{n}}{b_n}\left(\mathrm{H}_n^{(0)}(z+h_n u) - \mathrm{H}^{(0)}(z+h_n u) - \mathrm{H}_n^{(0)}(z) + \mathrm{H}^{(0)}(z)\right). \quad (5.16)$$

D'après **(H1)**, il existe $n_0 \geqslant 1$ tel que pour tout $n \geqslant n_0$ et tout $u \in [-\mathrm{M}, \mathrm{M}]$, nous avons $z + h_n u \in]\mathrm{I}_X, \mathrm{T}_X[$. Le théorème suivant donne le comportement asymptotique de (k_n).

Théorème 5.2. *Supposons que la dérivée f_X de F_X existe au point z et que F_L et F_R sont continues en z. Alors sous les hypothèses* **(H1)** *et* **(H2)**, *la suite (k_n) est presque sûrement relativement compacte dans $\mathbb{B}([-\mathrm{M}, \mathrm{M}])$ avec pour ensemble limite l'ensemble des fonctions k vérifiant*

$$k(u) = \int_0^u \phi(s)\, ds \text{ pour } -\mathrm{M} \leqslant u \leqslant \mathrm{M} \text{ où } \int_{-\mathrm{M}}^{\mathrm{M}} \phi^2(s)\, ds \leqslant f_X(z)\mathrm{F}_L(z)\mathrm{S}_R(z).$$

Démonstration du théorème 5.2. En utilisant (5.12), (5.16) et (5.13), nous pouvons écrire

$$\begin{aligned} k_n(u) &= \frac{1}{b_n}\left(\alpha_n\left(\mathrm{H}^{(0)}(z+h_n u) - \mathrm{H}^{(0)}(z) + s_0\right) - \alpha_n(s_0)\right) \\ &= \frac{1}{b_n}\left(\alpha_n\left(\mathrm{H}^{(0)}(z+h_n u)\right) - \alpha_n\left(\mathrm{H}^{(0)}(z)\right)\right) \qquad (\text{car } s_0 = \mathrm{H}^{(0)}(z)) \\ &= f_n\left(h_n^{-1}\left(\mathrm{H}^{(0)}(z+h_n u) - \mathrm{H}^{(0)}(z)\right)\right) = g_n(u), \end{aligned}$$

où $\gamma_n(u) = h_n^{-1}\left(\mathrm{H}^{(0)}(z+h_n u) - \mathrm{H}^{(0)}(z)\right)$, pour $-\mathrm{M} \leqslant u \leqslant \mathrm{M}$. D'après la définition de $\mathrm{H}^{(0)}$ dans (5.1) et sous les hypothèses sur F_X, F_L et F_R nous avons

$$\begin{aligned} &\sup_{-\mathrm{M}\leqslant u\leqslant \mathrm{M}} \left|\mathrm{H}^{(0)}(z+h_n u) - \mathrm{H}^{(0)}(z) - f_X(z)\mathrm{S}_R(z)\mathrm{F}_L(z)hu\right| \\ &= \sup_{-\mathrm{M}\leqslant u\leqslant \mathrm{M}} \left|\int_z^{z+hu} \mathrm{S}_R(t^-)\mathrm{F}_L(t^-)\, d\mathrm{F}_X(t) - f_X(z)\mathrm{S}_R(z)\mathrm{F}_L(z)hu\right| \\ &\leqslant \sup_{-\mathrm{M}\leqslant u\leqslant \mathrm{M}} \left|\int_z^{z+hu} \mathrm{S}_R(t^-)\mathrm{F}_L(t^-)\, d\mathrm{F}_X(t) - \int_z^{z+hu} \mathrm{S}_R(z)\mathrm{F}_L(z)\, d\mathrm{F}_X(t)\right| \\ &\quad + \sup_{-\mathrm{M}\leqslant u\leqslant \mathrm{M}} \left|\mathrm{S}_R(z)\mathrm{F}_L(z)\left(\mathrm{F}_X(z+hu) - \mathrm{F}_X(z)\right) - \mathrm{S}_R(z)\mathrm{F}_L(z)f_X(z)hu\right| \\ &\leqslant \sup_{-\mathrm{M}\leqslant u\leqslant \mathrm{M}} \left|\int_z^{z+hu} \left(\mathrm{S}_R(t^-)\mathrm{F}_L(t^-) - \mathrm{S}_R(z)\mathrm{F}_L(z)\right)\, d\mathrm{F}_X(t)\right| \\ &\quad + \mathrm{S}_R(z)\mathrm{F}_L(z) \sup_{-\mathrm{M}\leqslant u\leqslant \mathrm{M}} \left|\mathrm{F}_X(z+hu) - \mathrm{F}_X(z) - f_X(z)hu\right| \\ &\leqslant \sup_{z-h\mathrm{M}\leqslant u\leqslant z+h\mathrm{M}} \left|\mathrm{S}_R(u^-)\mathrm{F}_L(u^-) - \mathrm{S}_R(z)\mathrm{F}_L(z)\right| \sup_{-\mathrm{M}\leqslant u\leqslant \mathrm{M}} \left|\mathrm{F}_X(z+hu) - \mathrm{F}_X(z)\right| \\ &\quad + \mathrm{S}_R(z)\mathrm{F}_L(z) \sup_{-\mathrm{M}\leqslant u\leqslant \mathrm{M}} \left|\mathrm{F}_X(z+hu) - \mathrm{F}_X(z) - f_X(z)hu\right|. \end{aligned}$$

Or comme $\sup_{z-h\mathrm{M}\leqslant u\leqslant z+h\mathrm{M}} |\mathrm{S_R}(u^-)\mathrm{F_L}(u^-) - \mathrm{S_R}(z)\mathrm{F_L}(u)| \to 0$ (car $\mathrm{S_R}$ et $\mathrm{F_L}$ sont continues), $\frac{1}{h}\sup_{-\mathrm{M}\leqslant u\leqslant \mathrm{M}} |\mathrm{F_X}(z+hu) - \mathrm{F_X}(z)|$ est bornée et $\mathrm{F_X}(z+hu) = \mathrm{F_X}(z) + f_{\mathrm{X}}(z)hu + o(h)$, nous obtenons

$$\lim_{h\to 0}\frac{1}{h}\sup_{u\in[-\mathrm{M},\mathrm{M}]}\left|\mathrm{H}^{(0)}(z+hu) - \mathrm{H}^{(0)}(z) - f_{\mathrm{X}}(z)\mathrm{F_L}(z)\mathrm{S_R}(z)u\right| = 0,$$

ce qui montre que $(\gamma_n(u))$ vérifie (5.14) pour $\gamma_0 = f_{\mathrm{X}}(z)\mathrm{F_L}(z)\mathrm{S_R}(z)$. Le résultat du théorème en découle en appliquant le Lemme 5.3. □

Nous étudions maintenant la relation entre k_n et ξ_n. Nous donnons d'abord une représentation intégrale du processus empirique de Patilea et Rolin. Posons

$$\Pi_n(t) = \sqrt{n}\,\frac{\mathrm{F}_n(t) - \mathrm{F_X}(t)}{1 - \mathrm{F_X}(t)}. \tag{5.17}$$

Lemme 5.4. *Sous* **(H3)**, *pour tout* $t < \max\{\mathrm{Z}_j, 1 \leqslant j \leqslant n/\delta_j = 0 \text{ ou } \delta_j = 1\}$, *nous avons*

$$\Pi_n(t) = \frac{1}{\sqrt{n}}\int_0^t \frac{1-\mathrm{F}_n(u-)}{1-\mathrm{F_X}(u)}\frac{1}{\mathrm{F}_n^{(2)}(u-)\mathrm{S}_n^{(01)}(u-)}\,d\mathrm{M}_n(u). \tag{5.18}$$

où

$$\mathrm{M}_n(t) = n\left(\mathrm{H}_n^{(0)}(t) - \int_0^t \mathrm{F}_n^{(2)}(s^-)\mathrm{S}_n^{(01)}(s^-)d\Lambda(s)\right). \tag{5.19}$$

La fonction M_n joue le même role dans les démonstrations que la martingale de base dans le modèle de censure à droite. Considérons les fonctions d'accroissement

$$\mu_n(u) = \frac{1}{b_n\sqrt{n}}(\mathrm{M}_n(z+h_nu) - \mathrm{M}_n(z)), \tag{5.20}$$

et

$$\pi_n(u) = \frac{1}{b_n}\left(\Pi_n(z+h_nu) - \Pi_n(z)\right); \tag{5.21}$$

pour $-\mathrm{M} \leqslant u \leqslant \mathrm{M}$. Les lemmes suivants donnent les étapes nécessaires pour arriver à la relation entre ξ_n et k_n.

Lemme 5.5. *Supposons que* $\mathrm{F_X}$, $\mathrm{F_L}$ *et* $\mathrm{F_R}$ *sont continues et que la dérivée de* $\mathrm{F_X}$ *au point z existe. Alors sous* **(H1)** *et* **(H3)**

$$\sup_{u\in[-\mathrm{M},\mathrm{M}]}|\mu_n(u) - k_n(u)| = \mathrm{O}(\sqrt{h_n}) \to 0.$$

Lemme 5.6. *Sous les hypothèses du théorème 5.1*

$$\sup_{u\in[-M,M]}\left|\pi_n(u)-\frac{\mu_n(u)}{F^{(2)}(z)S^{(01)}(z)}\right|\to 0$$

Lemme 5.7. *Sous les hypothèses du lemme 5.5*

$$\sup_{u\in[-M,M]}|\xi_n(u)-S_X(z)\pi_n(u)|\longrightarrow 0 \quad p.s$$

Démonstration du théorème 5.1. De (5.2) et du fait que $z > I_X$, nous déduisons que $F^{(2)}(z)S^{(01)}(z) = F_L(z)S_R(z)S_X(z)$. D'après les Lemmes 5.5, 5.6 et 5.7, ceci implique

$$\sup_{u\in[-M,M]}\left|\xi_n(u)-\frac{k_n(u)}{F_L(z)S_R(z)}\right|\to 0. \tag{5.22}$$

Le fait que $F_L(z)S_R(z)S_X(z) > 0$ vient de l'hypothèse **(H3)** et des conditions imposées à z. D'autre part, le théorème 5.2 implique que la suite $\left(\frac{k_n(u)}{F_L(z)S_R(z)} : n \geqslant 1\right)$ de fonctions de u est presque sûrement relativement compacte dans $\mathbb{B}([-M,M])$ avec pour ensemble limite l'ensemble de toutes les fonctions h définies pour $u \in [-M,M]$, vérifiant

$$h(u)=\int_0^u \frac{\phi(s)}{F_L(z)S_R(z)}\,ds, \quad \text{où } \int_{-M}^{M}\phi^2(s)\,ds \leqslant f_X(z)F_L(z)S_R(z). \tag{5.23}$$

Pour $\Psi(s) = \frac{\phi(s)}{F_L(z)S_R(z)}$, il est clair que (5.23) est équivalente à (5.8). Le résultat du théorème en découle en utilisant (5.22). □

5.3 Application : Estimation de la densité

Le but du théorème 5.1 est de donner des résultats pour quelques estimateurs qui dépendent localement du processus empirique. L'exemple que nous allons traiter est l'estimateur à noyau de la densité, mais la méthode que nous exposerons peut être utilisée pour les autres estimateurs du même type (par exemple l'estimateur du taux de hasard).

Considérons une fonctionnelle Γ définie et continue sur un ensemble fermé $\mathcal{S}$ de $\mathbb{B}([-M,M])$ et satisfaisant la condition $\xi_n \in \mathcal{S}, \forall n \geqslant 1$. On définit alors la statistique $T_n = \Gamma(\xi_n)$.

Théorème 5.3. *Soit $z \in]0, \Theta[$ et $M > 0$ fixés. On suppose que F est continue dans un voisinage de z et dérivable en z, et que G est continue en z. Alors, sous (H1) et (H2), la suite $(T_n)_{n\geqslant 1}$ est presque sûrement relativement compacte dans $\mathbb{R}$, avec pour ensemble limite l'intervalle*

$$\left[\inf_{h\in\mathbb{L}_M} \Gamma(h), \sup_{h\in\mathbb{L}_M} \Gamma(h)\right],$$

où $\mathbb{L}_M$ est l'ensemble limite de (ξ_n) (voir le théorème 5.1).

Démonstration. Le fait que T_n soit presque sûrement relativement compacte dans $\mathbb{R}$ avec pour ensemble limite $\Gamma(\mathbb{L}_M)$ découle directement du théorème 5.1 en utilisant la continuité de Γ.

L'ensemble $\mathbb{L}_M$ est connexe dans $\mathbb{B}([-M, M])$. En effet, l'ensemble des fonctions ψ vérifiant $\int_{-M}^{M} \psi^2(s)\, ds \leqslant \frac{f(z)}{1-G(z)}$ est une boule fermée de $L^2([-M, M])$, elle est donc connexe. L'opérateur qui à à toute fonction $\psi \in L^2([-M, M])$ associe la fonction h définie pour $u \in [-M, M]$ par $h(u) = \int_0^u \psi(s)\, ds$ étant continu, $\mathbb{L}_M$ est aussi connexe.

De plus, si $h \in \mathbb{L}_M$, alors il existe $\psi \in L^2([-M, M])$ tel que $\Psi(u) = \int_0^u \psi(s)\, ds$, et on a alors, pour tout $x_1, x_2 \in [-M, M]$:

$$\begin{aligned}
|h(x_2) - h(x_1)| &= \left|\int_{x_1}^{x_2} \psi(s)\, ds\right| \\
&\leqslant |x_2 - x_1|^{1/2} \left(\int_{-M}^{M} \psi^2(s)\, ds\right)^{1/2} \\
&\leqslant \left(\frac{f(z)}{1-G(z)}\right)^{1/2} |x_2 - x_1|^{1/2};
\end{aligned}$$

ce qui montre que $\mathbb{L}_M$ est uniformément équicontinu. Et comme $\mathbb{L}_M$ est borné (c'est l'image par un opérateur borné d'une boule), il est compact d'après le théorème d'Arzelà-Ascoli.

Son image par l'application continue Γ est donc un intervalle fermé. □

En choisissant la fonctionnelle Γ, on peut montrer des taux de convergence pour quelques estimateurs qui dépendent localement du processus empirique, la preuve ci dessous est un exemple pour l'estimateur à noyau de la densité

$$f_n(x) = \int_{-\infty}^{+\infty} \frac{1}{h_n} K\left(\frac{y-u}{h_n}\right) dF_n(u),$$

Supposons que K vérifie les conditions suivantes :

K1 K est à variation bornée sur $\mathbb{R}$,

K2 K est à support compact,

K3 $\int_{-\infty}^{\infty} \mathrm{K}(u)\,\mathrm{d}u = 1$;

et définissons la quantité

$$\mathbb{E}f_n(z) = \int_{-\infty}^{+\infty} \frac{1}{h_n}\mathrm{K}\left(\frac{y-u}{h_n}\right) d\mathrm{F}(u).$$

Quand les données sont complètes $\mathbb{E}f_n(z)$ est égale à l'espérance mathématique de $f_n(z)$, mais ce n'est pas le cas en général. Remarquons que le terme $\mathbb{E}f_n(z) - f(z)$ se traite de la même manière que dans le cas de données complètes, sous des conditions appropriées sur K et h_n, et des conditions de régularité sur f_{X}, voir le théorème 4.2. Le corollaire suivant donne le taux de convergence de f_n.

Corollaire 5.1. *Soit* $z \in]\mathrm{I}_{\mathrm{X}}, \mathrm{T}_{\mathrm{X}}[$. *Supposons que* F_{X}, F_{L} *et* F_{R} *sont continues et que la dérivée* f_{X} *de* F_{X} *au point* z *existe. Alors sous les conditions* **(H1)**, **(H2)**, **(H3)**, **(K1)**, **(K2)** *et* **(K3)** *nous avons*

$$\limsup_{n\to\infty} \pm\sqrt{\frac{nh_n}{\log\log n}}\,(f_n(z) - \mathbb{E}f_n(z)) = \left(\frac{f_x(z)}{\mathrm{F}_{\mathrm{L}}(z)\mathrm{S}_{\mathrm{R}}(z)}\int \mathrm{K}^2(u)du\right)^{1/2}$$

Démonstration. En vertu de l'hypothèse (K2), il existe un réel positif M tel que $\forall u, |u| \geqslant \frac{\mathrm{M}}{2} \implies \mathrm{K}(u) = 0$. Donc,

$$\begin{aligned}
f_n(z) - \mathbb{E}f_n(z) &= \int_{-\infty}^{+\infty} \frac{1}{h_n}\mathrm{K}\left(\frac{t-z}{h_n}\right) d\mathrm{F}_n(t) - \int_{-\infty}^{+\infty} \frac{1}{h_n}\mathrm{K}\left(\frac{t-z}{h_n}\right) d\mathrm{F}(t) \\
&= \frac{1}{h_n}\int_{-\infty}^{+\infty} \mathrm{K}(u)\, d(\mathrm{F}_n(z+h_nu) - \mathrm{F}(z+h_nu)) \\
&= \frac{1}{h_n}\int_{-\mathrm{M}}^{\mathrm{M}} \mathrm{K}(u)\, d(\mathrm{F}_n(z+h_nu) - \mathrm{F}_n(z) - \mathrm{F}(z+h_nu) + \mathrm{F}(z)) \qquad (5.24) \\
&= \frac{-1}{h_n}\int_{-\mathrm{M}}^{\mathrm{M}} (\mathrm{F}_n(z+h_nu) - \mathrm{F}_n(z) - \mathrm{F}(z+h_nu) + \mathrm{F}(z))\, d\mathrm{K}(u) \qquad (5.25) \\
&= \frac{-b_n}{h_n\sqrt{n}}\int_{-\mathrm{M}}^{\mathrm{M}} \xi_n(u)\, d\mathrm{K}(u),
\end{aligned}$$

où (5.24) est due au fait que $\mathrm{F}_n(z)$ et $\mathrm{F}(z)$ sont constants, et (5.25) est obtenue en utilisant l'intégration par parties (ce qui est possible grâce à l'hypothèse **(K1)**).

Définissons la fonctionnelle Γ en posant pour toute fonction $h \in \mathbb{B}([-M, M])$ à variation bornée

$$\Gamma(h) = -\int_{-M}^{M} h(u)\, dK(u);$$

du du fait que $b_n = \sqrt{2h_n \log\log n}$ on a alors

$$T_n = \Gamma(\xi_n) = \frac{h_n\sqrt{n}}{b_n}(f_n(z) - \mathbb{E}f_n(z)) = \sqrt{\frac{nh_n}{2\log\log n}}(f_n(z) - \mathbb{E}f_n(z)).$$

Le théorème 5.3 s'applique et l'ensemble limite de la suite T_n est l'intervalle

$$\left[\inf_{h\in\mathbb{L}_M} \Gamma(h), \sup_{h\in\mathbb{L}_M} \Gamma(h)\right].$$

Il reste à calculer $\sup_{h\in\mathbb{L}_M} \pm\Gamma(h)$ (car $\inf_x f(x) = -\sup_x -f(x)$). Or par les hypothèses **(K1)** et **(K2)**, une intégration par parties permet d'écrire

$$\begin{aligned}
\sup_{h\in\mathbb{L}_M} \pm\Gamma(h) &= \sup_{h\in\mathbb{L}_M} \pm\int_{-M}^{M} K(u)\, dh(u) \\
&= \sup_{h\in\mathbb{L}_M} \left\{\pm\int_{-M}^{M} K(u)\psi(u)\, du \Big/ h(u) = \int_0^u \psi(s)\, ds\right\} \\
&\leqslant \left(\int_{-M}^{M} K^2(u)\, du\right)^{1/2} \left(\frac{f_X(z)}{F_L(z)S_R(z)}\right)^{1/2},
\end{aligned}$$

d'après l'inégalité de Schwarz. D'autre part, le choix particulier de

$$\Psi^*(u) = \frac{K^*(u)}{\left(\int_{-M}^{M}(K^*(t))^2\right)^{1/2}}$$

montre l'égalité $\sup_{h\in\mathbb{L}_M} \pm\Gamma(h) = \left(\int_{-M}^{M} (K^*(u))^2\, du\right)^{1/2}$. D'où

$$\sup_{h\in\mathbb{L}_M} \pm\Gamma(h) = \left(\frac{f^-(z)}{1-G^-(z)}\int_{-M}^{0} K^2(u)\, du + \frac{f_+(z)}{1-G(z)}\int_0^M K^2(u)\, du\right)^{1/2}.$$

Le résultat visé en découle en rappelant que pour toute suite (x_n), $\limsup x_n$ n'est autre que la borne supérieure de l'ensemble des valeurs d'adhérence de (x_n). □

Cet estimateur peut être utilisé pour définir un estimateur λ_n du taux de hasard $\lambda(z) = f(z)/(1 - F(z))$ en posant $\lambda_n(z) = f_n(z)/(1 - F_n(z))$ si $(1 - F_n(z)) \neq 0$. Le résultat suivant est une conséquence directe du Corollaire 5.1.

Corollaire 5.2. *Soit* $z \in]\mathrm{I_X}, \mathrm{T_X}[$ *et* $\mathrm{M} > 0$ *fixés. Suppposons que* $\mathrm{F_X}$, $\mathrm{F_R}$ *et* $\mathrm{F_L}$ *sont continues et que la dérivée* f_{X} *de* $\mathrm{F_X}$ *existe au point* z. *Alors sous les conditions* **(H1)**, **(H2)**, **(H3)**, **(K1)**, **(K2)** *et* **(K3)** *nous avons*

$$\left|\lambda_n(z) - \frac{\mathbb{E}f_n(z)}{1-\mathrm{F}(z)}\right| = \mathrm{O}\left(\sqrt{\frac{\log\log n}{nh_n}}\right)$$

Démonstration. Nous avons la décomposition

$$\left|\lambda_n(z) - \frac{\mathbb{E}f_n(z)}{1-\mathrm{F}(z)}\right| \leqslant \frac{|f_n(z) - \mathbb{E}f_n(z)|}{1-\mathrm{F}(z)} + |f_n(z)|\left|\frac{1}{1-\mathrm{F}_n(z)} - \frac{1}{1-\mathrm{F}(z)}\right|.$$

D'après le Corollaire 5.1, le premier terme est $\mathrm{O}\left(\sqrt{\frac{\log\log n}{nh_n}}\right)$, et d'après la loi du logarithme itéré de Messaci et Nemouchi (2011) le second terme est $\mathrm{O}\left(\sqrt{\frac{\log\log n}{n}}\right)$, d'où le résultat. □

Une comparaison des résultats de ce chapitre à ceux obtenus au chapitre 4, montre que les taux de convergence que nous venons d'établir, pour les estimateurs de la densité et du taux de hasard, sont meilleurs (une loi du logarithme itéré), mais pour un mode de convergence plus faible. Nous utilisons ici la convergence presque sûre qui est légèrement plus faible que la convergence presque complète, bien qu'utilisée beaucoup plus souvent. D'autre part, nous étudions ici la convergence ponctuelle alors que nous nous sommes intéressés à la convergence uniforme au chapitre 4.

Une perspective de recherche est alors d'essayer d'étendre les résultats de ce chapitre à la convergence uniforme en utilisant une approche similaire à celle de Deheuvels et Einmahl (2000).

5.4 Démonstration des lemmes

Démonstration du lemme 5.1. Remarquons que pour tout variable aléatoire V, de fonction de répartition $\mathrm{D}(v) = \mathrm{P}(\mathrm{V} \leqslant v)$ et de fonction quantile $\mathrm{D}^{inv}(s) = \inf\{v : \mathrm{D}(v) \geqslant s\}$, on peut définir V sur un espace de probabilité qui contient une variable aléatoire W de loi $\mathcal{U}([0,1])$ tel que $\mathrm{V} = \mathrm{D}^{inv}(\mathrm{W})$ p.s.

Comme la fonction de répartition conditionnelle de Z sachant $\delta = 0$ est $p^{-1}\mathrm{H}^{(0)}(t)$, et sa fonction quantile est $\mathrm{Q}^{(0)}(ps)$, l'application de la remarque ci-dessus montre que conditionnellement à $\delta = 0$, il existe une variable aléatoire W_0 de loi $\mathcal{U}([0,1])$ telle que $\mathrm{Z} = \mathrm{Q}^{(0)}(p\mathrm{W}_0)$ p.s.

Le même raisonnement appliqué au cas $\delta = 1$ (resp. $\delta = 2$) donne que conditionnellement à $\delta = 1$ (resp. $\delta = 2$) il existe une variable aléatoire W_1 (resp. W_2) de loi $\mathcal{U}([0,1])$ telle que $Z = Q^{(1)}(qW_1)$ p.s. (resp. $Z = Q^{(2)}((1-p-q)W_2)$ p.s.).

Posons $U = pW_0 1_{\{\delta=0\}} + (p+qW_1)1_{\{\delta=1\}} + (p+q+(1-p-q)W_2)\,1_{\{\delta=2\}}$. Alors, U vérifie (5.10) par construction, et on a

$$\begin{aligned}
P(U \leqslant u) &= P(pW_0 \leqslant u/\delta = 0)P(\delta = 0) + P(p+qW_1 \leqslant u/\delta = 1)P(\delta = 1) \\
&\quad + P\left(p+q+(1-p-q)W_2 \leqslant u/\delta = 2\right)P(\delta = 2) \\
&= \left(\frac{u}{p}1_{\{0\leqslant u\leqslant p\}} + 1_{\{u>p\}}\right)p + \left(\frac{u-p}{q}1_{\{p\leqslant u\leqslant p+q\}} + 1_{\{u>p+q\}}\right)q \\
&\quad + \left(\frac{u-p-q}{1-p-q}1_{\{p+q\leqslant u\leqslant 1\}} + 1_{\{u>1\}}\right)(1-p-q) \\
&= u1_{\{0\leqslant u\leqslant p\}} + u1_{\{p\leqslant u\leqslant p+q\}} + u1_{\{p+q<u\leqslant 1\}} + 1_{\{u>1\}} \\
&= u1_{\{0\leqslant u\leqslant 1\}} + 1_{\{u>1\}},
\end{aligned}$$

ce qui montre que U est uniformément distribuée sur $[0,1]$. □

Démonstration du lemme 5.3. Choisissons $M_1 > 0$ dans le lemme 5.2 tel que $|\gamma_n(u)| \leqslant M_1$ pour tout u, pour n suffisamment grand. Soit g_{n_k} une sous suite de g_n. D'après le Lemme 5.2, il existe presque sûrement une sous suite (n'_k) de (n_k) et une fonction f telles que $\sup_{v\in[-M_1,M_1]}\left|f_{n'_k}(v) - f(v)\right| \to 0$, donc

$$\sup_{u\in[-M,M]}\left|f_{n'_k}\left(\gamma_{n'_k}(u)\right) - f\left(\gamma_{n'_k}(u)\right)\right| \to 0.$$

La continuité uniforme de f et l'équation (5.14) permettent d'écrire

$$\begin{aligned}
\sup_{u\in[-M,M]}\left|g_{h'_k}(u) - f(\gamma_0 u)\right| &\leqslant \sup_{u\in[-M,M]}\left|f_{n'_k}\left(\gamma_{n'_k}(u)\right) - f\left(\gamma_{n'_k}(u)\right)\right| \\
&\quad + \sup_{u\in[-M,M]}\left|f\left(\gamma_{n'_k}(u)\right) - f\left(\gamma_0 u\right)\right| \\
&\to 0,
\end{aligned}$$

et en posant $\phi_0(u) = \gamma_0\psi_0(\gamma_0 u)$ et $g(u) = f(\gamma_0 u)$ pour $-M \leqslant u \leqslant M$, on peut facilement voir que g vérifie (5.15). Ceci montre que (g_n) est presque sûrement relativement compacte, et que son ensemble limite est inclus dans l'ensemble défini par (5.15). De la même manière, on peut montrer que les deux ensembles sont égaux. □

Démonstration du lemme 5.4. Sous **(H3)** nous avons $\Lambda(dt) = F_X(dt)/S_X(t^-)$, pour $t < T_X$. Ceci implique que $1 - F_X(t) = \prod_{[0,t]}(1 - d\Lambda(s))$. D'autre

part, nous avons par définition $1 - F_n(t) = \prod_{[0,t]} (1 - d\Lambda_n(s))$, qui est bien défini pour $t < \max\{Z_j, 1 \leqslant j \leqslant n / \delta_j = 0 \text{ ou } \delta_j = 1\}$. L'équation de Duhamel (1.14) donne

$$\begin{aligned} F_X(t) - F_n(t) &= \prod_{[0,t]} (1 - d\Lambda_n(s)) - \prod_{[0,t]} (1 - d\Lambda(s)) \\ &= \int_0^t \prod_{[0,u[} (1 - d\Lambda_n(s))\, d(\Lambda - \Lambda_n)(u) \prod_{]u,t]} (1 - d\Lambda(s)) \\ &= \int_0^t 1 - F_n(u^-) \frac{1 - F_X(t)}{1 - F_X(u)} d(\Lambda - \Lambda_n)(u), \end{aligned}$$

d'où

$$\begin{aligned} \frac{F_n(t) - F_X(t)}{1 - F_X(t)} &= \int_0^t \frac{1 - F_n(u^-)}{1 - F_X(u)}\, d(\Lambda_n - \Lambda)(u) \\ &= \int_0^t \frac{1 - F_n(u-)}{1 - F_X(u)} \left(\frac{dH_n^{(0)}(u)}{F_n^{(2)}(u-) S_n^{(01)}(u-)} - d\Lambda(u) \right) \\ &= \frac{1}{n} \int_0^t \frac{1 - F_n(u-)}{1 - F_X(u)} \frac{1}{F_n^{(2)}(u-) S_n^{(01)}(u-)}\, dM_n(u), \end{aligned}$$

ce qui donne (5.18). □

Démonstration du lemme 5.5. (5.1) et (5.3) donnent $H^{(0)}(t) = \int_0^t F^{(2)}(s^-) S^{(01)}(s^-) d\Lambda(s)$. En remplaçant dans (5.19) et en utilisant le fait que $H(t) = F^{(2)}(t) F^{(01)}(t)$ (5.2) nous avons

$$\begin{aligned} M_n(t) &= n \left(H_n^{(0)}(t) - H^{(0)}(t) \right) + \int_{(0)}^t F^2(s) S^{(01)}(s^-)\, d\Lambda(s) \\ &\quad - \int_0^t F_n^2(s^-) S_n^{(01)}(s^-) d\Lambda_n(s) \\ &= n \left(H_n^{(0)}(t) - H^{(0)}(t) + \int_0^t \left(H_n(s^-) - H(s^-) \right) d\Lambda(s) \right) \\ &\quad - n \left(\int_0^t \left(F_n^{(2)}(s^-) - F^{(2)}(s-) \right) d\Lambda(s) \right). \end{aligned}$$

D'après (5.16) et (5.20), nous en déduisons que

$$\begin{aligned}\mu_n(u) &= \frac{\sqrt{n}}{b_n}\left(\mathrm{H}_n^{(0)}(z+h_nu) - \mathrm{H}^{(0)}(z+h_nu) - \mathrm{H}_n^{(0)}(z) + \mathrm{H}^{(0)}(z)\right)\\ &\quad + \frac{\sqrt{n}}{b_n}\int_{(z)}^{z+h_nu}\left(\mathrm{H}_n(s^-) - \mathrm{H}(s^-)\right)\,d\Lambda(s)\\ &\quad + \frac{\sqrt{n}}{b_n}\int_{z}^{z+h_nu}\left(\mathrm{F}_n^2(s^-) - \mathrm{F}^{(2)}(s^-)\right)\,d\Lambda(s)\\ &= k_n(u) + \frac{\sqrt{n}}{b_n}\int_z^{z+h_nu}(\mathrm{H}_n(t^-) - \mathrm{H}(t^-))\,d\Lambda(t)\\ &\quad - \frac{\sqrt{n}}{b_n}\int_z^{z+h_nu}\left(\mathrm{F}_n^{(2)}(t^-) - \mathrm{F}^{(2)}(t^-)\right)\,d\Lambda(t).\end{aligned}$$

En appliquant la loi du logarithme itéré de Chung (1949) au processus empirique $\sqrt{n}\,(\mathrm{H}_n(t^-) - \mathrm{H}(t^-))$, nous obtenons

$$\sup_{t\in[z-h_n\mathrm{M},z+h_n\mathrm{M}]}\sqrt{n}\,|\mathrm{H}_n(t^-) - \mathrm{H}(t^-)| = \mathrm{O}(\sqrt{\log\log n}).$$

En inversant le temps, nous pouvons adapter la loi du logarithme itéré de Földes et Rejtő (1981a) au cas de la censure à gauche pour obtenir

$$\sup_{t\in[z-h_n\mathrm{M},z+h_n\mathrm{M}]}\sqrt{n}\,\left|\mathrm{F}_n^{(2)}(t^-) - \mathrm{F}^{(2)}(t^-)\right| = \mathrm{O}(\sqrt{\log\log n}),$$

ce qui montre que pour un certain $\mathrm{A}_n = \mathrm{O}(1/\sqrt{h_n})$,

$$\sup_{u\in[-\mathrm{M},\mathrm{M}]}|\mu_n(u) - k_n(u)| \leqslant (\Lambda(z+h_n\mathrm{M}) - \Lambda(z-h_n\mathrm{M}))\,\mathrm{A}_n$$

qui est $\mathrm{O}(\sqrt{h_n})$ d'après **(H1)**, **(H3)** et les conditions de régularité sur $\mathrm{F_X}$, $\mathrm{F_L}$ and $\mathrm{F_R}$. □

Démonstration du lemme 5.6. Posons $\mathrm{T}_n = \max\{\mathrm{Z}_j, 1 \leqslant j \leqslant n/\delta_j = 0 \text{ ou } \delta_j = 1\}$. Comme $z < \mathrm{T_X}$, $h_n \to 0$ et $\mathrm{T}_n \to \mathrm{T_X}$ p.s. quand $n \to \infty$, nous pouvons supposer que $z + h_nu < \mathrm{T}_n$ pour n suffisamment grand (car $u \in [-\mathrm{M},\mathrm{M}]$). Alors (5.18) est vérifiée pour $t = z+h_nu$. Nous pouvons alors écrire $\Pi_n(t) = \Pi_{n,1}(t) + \Pi_{n,2}(t)$ avec

$$\begin{aligned}\Pi_{n,1}(t) &= \frac{1}{\sqrt{n}}\int_0^t \frac{d\mathrm{M}_n(u)}{\mathrm{F}^{(2)}(u^-)\mathrm{S}^{(01)}(u^-)};\\ \Pi_{n,2}(t) &= \frac{1}{\sqrt{n}}\int_0^t\left(\frac{1-\mathrm{F}_n(u^-)}{1-\mathrm{F_X}(u)}\right)\frac{1}{\mathrm{F}_n^{(2)}(u^-)\mathrm{S}_n^{(01)}(u^-)}d\mathrm{M}_n(u)\\ &\quad - \frac{1}{\sqrt{n}}\int_0^t\frac{1}{\mathrm{F}^{(2)}(u^-)\mathrm{S}^{(01)}(u^-)}d\mathrm{M}_n(u).\end{aligned}$$

Soit $\pi_{n,j}(u) = b_n^{-1}\left(\Pi_{n,j}(z+h_nu) - \Pi_{n,j}(z)\right)$ $(1 \leqslant j \leqslant 2)$. Nous avons d'une part

$$\pi_{n,1}(u) = \frac{1}{b_n\sqrt{n}} \int_z^{z+h_nu} \frac{d\mathrm{M}_n(t)}{\mathrm{F}^{(2)}(t-)\mathrm{S}^{(01)}(t-)}.$$

En posant $\mathrm{U}(t) = \mathrm{M}_n(t) - \mathrm{M}_n(z)$ et $\mathrm{V}(t) = \mathrm{F}^{(2)}(t)\mathrm{S}^{(01)}$, l'intégration par parties ($\int \mathrm{V}^- d\mathrm{U} = \mathrm{UV} - \int \mathrm{U}\, d\mathrm{V}$) donne

$$\pi_{n,1}(u) = \frac{\mu_n(u)}{\mathrm{F}^{(2)}(z+h_nu)\mathrm{S}^{(01)}(z+h_nu)} - \int_0^u \mu_n(t)\left(\frac{1}{\mathrm{F}^{(2)}(z+h_nt)\mathrm{S}^{(01)}(z+h_nt)}\right)$$

Or la suite (μ_n) est bornée d'après le théorème 5.2 et le Lemme 5.5. Donc

$$\begin{aligned}\left|\pi_{n,1}(u) - \frac{\mu_n(u)}{\mathrm{F}^{(2)}(z)\mathrm{S}^{(01)}(z)}\right| \leqslant |\mu_n(u)|\left|\frac{1}{\mathrm{F}^{(2)}(z+h_nu)\mathrm{S}^{(01)}(z+h_nu)} - \frac{1}{\mathrm{F}^{(2)}(z)\mathrm{S}^{(01)}(z)}\right| \\ + \sup_{t\in[-\mathrm{M},\mathrm{M}]}|\mu_n(t)|\left|\int_0^u\left(\frac{1}{\mathrm{F}^{(2)}(z+h_nt)\mathrm{S}^{(01)}(z+h_nt)}\right)\right|,\end{aligned}$$

ce qui implique qu'il existe une constante $\mathrm{C} > 0$ telle que

$$\begin{aligned}&\sup_{u\in[-\mathrm{M},\mathrm{M}]}\left|\pi_{n,1}(u) - \frac{\mu_n(u)}{\mathrm{F}^{(2)}(z)\mathrm{S}^{(01)}(z)}\right| \\ &\leqslant \mathrm{C}\sup_{u\in[-\mathrm{M},\mathrm{M}]}\left|\frac{1}{\mathrm{F}^{(2)}(z+h_nu)\mathrm{S}^{(01)}(z+h_nu)} - \frac{1}{\mathrm{F}^{(2)}(z)\mathrm{S}^{(01)}(z)}\right| \to 0. \quad (5.26)\end{aligned}$$

D'autre part, en utilisant la définition de M_n (5.19), nous avons

$$\begin{aligned}\pi_{n,2}(u) = &\frac{\sqrt{n}}{b_n}\int_z^{z+h_nu}\frac{1-\mathrm{F}_n(t^-)}{(1-\mathrm{F}_\mathrm{X}(t))\mathrm{F}_n^{(2)}(t^-)\mathrm{S}_n^{(01)}(t^-)}d\mathrm{H}_n^{(0)}(t) \\ &- \frac{\sqrt{n}}{b_n}\int_z^{z+h_nu}\frac{1}{\mathrm{F}^{(2)}(t^-)\mathrm{S}^{(01)}(t^-)}d\mathrm{H}_n^{(0)}(t) \quad (5.27) \\ &- \frac{\sqrt{n}}{b_n}\int_z^{z+h_nu}\left(\left(\frac{1-\mathrm{F}_n(t^-)}{1-\mathrm{F}_\mathrm{X}(t)}\right) - \frac{\mathrm{F}_n^{(2)}(t^-)\mathrm{S}_n^{(01)}(t^-)}{\mathrm{F}^{(2)}(t^-)\mathrm{S}^{(01)}(t^-)}\right)d\Lambda(t).\end{aligned}$$

Sous l'hypothèse **(H3)**, la loi du logarithme itéré de Messaci and Nemouchi (2011, 2013) donne

$$\sup_{t\in[z-h_n\mathrm{M},z+h_n\mathrm{M}]}\left|\frac{1-\mathrm{F}_n(t^-)}{1-\mathrm{F}_\mathrm{X}(t^-)} - 1\right| = \mathrm{O}\left(\sqrt{\frac{\log\log n}{n}}\right); \quad (5.28)$$

de plus, nous pouvons adapter la loi du logarithme itéré de Földes et Rejtő (1981a) au cas de la censure à gauche pour avoir

$$\sup_{t\in[z-h_n\mathrm{M},z+h_n\mathrm{M}]}\left|\frac{\mathrm{S}_n^{(01)}(t^-)}{\mathrm{S}^{(01)}(t^-)}-1\right|=\mathrm{O}\left(\sqrt{\frac{\log\log n}{n}}\right), \tag{5.29}$$

$$\sup_{t\in[z-h_n\mathrm{M},z+h_n\mathrm{M}]}\left|\frac{\mathrm{F}_n^{(2)}(t^-)}{\mathrm{F}^{(2)}(t^-)}-1\right|=\mathrm{O}\left(\sqrt{\frac{\log\log n}{n}}\right). \tag{5.30}$$

En utilisant (5.28), (5.29) et (5.30), et le fait que $b_n=\sqrt{2h_n\log\log n}$, (5.27) donne que pour un certain $\mathrm{A}_n=\mathrm{O}(1/\sqrt{h_n})$

$$\begin{aligned}&\sup_{u\in[-\mathrm{M},\mathrm{M}]}\left|\pi_{n,2}(u)\right|\\&\leqslant\left(\mathrm{H}_n^{(0)}(z+h_n\mathrm{M})-\mathrm{H}_n^{(0)}(z-h_n\mathrm{M})+\Lambda(z+h_n\mathrm{M})-\Lambda(z-h_n\mathrm{M})\right)\mathrm{A}_n.\end{aligned} \tag{5.31}$$

Comme dans la démonstration du théorème 5.2, les conditions de régularité sur $\mathrm{F_X}$, $\mathrm{F_L}$ et $\mathrm{F_R}$ impliquent que $\Lambda(z+h_n\mathrm{M})-\Lambda(z-h_n\mathrm{M})=\mathrm{O}(h_n)$ et que $\mathrm{H}^{(0)}(z+h_n\mathrm{M})-\mathrm{H}^{(0)}(z-h_n\mathrm{M})=\mathrm{O}(h_n)$. De plus, en utilisant (5.16) et le théorème 5.2, nous déduisons que

$$\begin{aligned}\mathrm{H}_n^{(0)}(z+h_n\mathrm{M})-\mathrm{H}_n^{(0)}(z-h_n\mathrm{M})&=\frac{b_n}{\sqrt{n}}\left(k_n(\mathrm{M})-k_n(-\mathrm{M})\right)+\mathrm{O}(h_n)\\&=\mathrm{O}\left(b_n/\sqrt{n}\right)+\mathrm{O}(h_n)=\mathrm{O}(h_n).\end{aligned}$$

La dernière inégalité vient de l'hypothèse **(H2)**. En remplaçant dans (5.31), et en combinant avec (5.26), nous obtenons le résultat du lemme. □

Démonstration du lemme 5.7. D'après (5.6) et (5.17), nous avons $a_n(z)=(1-\mathrm{F_X}(z))\Pi_n(z)$. En remplaçant dans (5.7), et en utilisant (5.21), nous obtenons

$$\begin{aligned}\xi_n(u)&=\frac{1}{b_n}(1-\mathrm{F_X}(z+h_nu))\Pi_n(z+h_nu)-\frac{1}{b_n}(1-\mathrm{F_X}(z))\Pi_n(z)\\&=\frac{1}{b_n}(1-\mathrm{F_X}(z))(\Pi_n(z+h_nu)-\Pi_n(z))\\&\quad+\frac{1}{b_n}(\mathrm{F_X}(z)-\mathrm{F_X}(z+h_nu))\Pi_n(z+h_nu)\\&=(1-\mathrm{F_X}(z))\pi_n(u)+\frac{1}{b_n}(\mathrm{F_X}(z)-\mathrm{F_X}(z+h_nu))\Pi_n(z+h_nu).\end{aligned}$$

Nous pouvons facilement voir que (5.28) et (5.17) impliquent

$$\sup_{u\in[-M,M]} |\Pi_n(z+h_n u)| = O\left(\sqrt{\log\log n}\right).$$

En outre, l'existence de la dérivée de F_X au point z implique que $F_X(z) - F_X(z+h_n u) = O(h_n)$. Nous avons alors

$$\sup_{u\in[-M,M]} |\xi_n(u) - (1-F_X(z))\pi_n(u)| = O(\sqrt{h_n}) \to 0 \quad \text{p.s.} \qquad \square$$

Chapitre 6

Simulation

Pour donner une idée de la performance des estimateurs à noyau de la densité et du taux de hasard pour des tailles d'échantillons finies, nous les calculons et nous traçons leurs graphes pour les comparer à la vraie densité et au vrai taux de hasard respectivement pour deux tailles d'échantillons et pour deux modèles. Rappelons qu'en présence de censure mixte, au lieu d'observer un échantillon de la variable d'intérêt X, nous observons un échantillon du couple (Z, A) où $Z = \max(\min(X, R), L)$ et

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

Modèles et échantillons Nous commençons par simuler les temps de survie d'intérêt $\{X_i,\ 1 \leqslant i \leqslant n\}$, les temps de censure à droite $\{R_i,\ 1 \leqslant i \leqslant n\}$ et les temps de censure à gauche $\{L_i,\ 1 \leqslant i \leqslant n\}$ selon deux modèles et pour deux tailles de l'échantillon ($n = 300$ et $n = 500$) pour chaque modèle.

1. Modèle 1
   — X_i suit une loi de Weibull : $Weibull(2, 2)$,
   — R_i suit une loi de Weibull : $Weibull(4, 2)$,
   — L_i de Weibull : $Weibull(0.6, 2)$.
2. Modèle 2
   — X_i suit une loi log-logistique : $LogLogis(3.7, 6)$,
   — R_i suit une loi log-logistique : $LogLogis(5, 6)$,
   — L_i suit une loi de Weibull : $Weibull(3, 6)$.

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

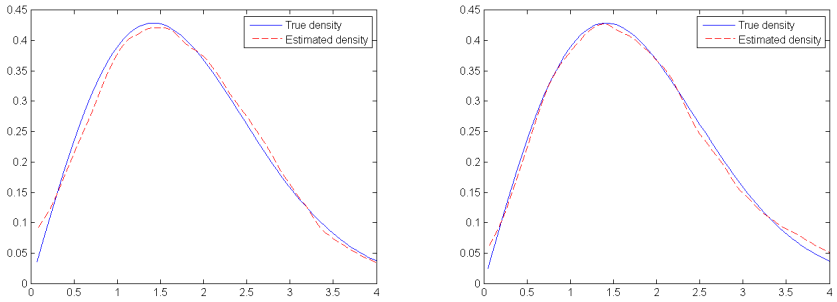

FIGURE 6.1 – Estimateur de la densité pour le modèle 1, avec $n = 300$ et $n = 500$

Taux de censure Les paramètres des lois sont choisis de façon à obtenir des taux de censure acceptables. Nous avons obtenu des taux de censure à droite d'environ 15–20% et des taux de censure à gauche d'environ 10–15% pour les deux modèles.

Calcul des estimateurs Nous calculons alors l'estimateur f_n de la densité

$$f_n(x) = \frac{1}{h_n} \int \mathrm{K}\left(\frac{x-y}{h_n}\right) d\mathrm{F}_n(y),$$

et l'estimateur $\tilde{\lambda}_n$

$$\tilde{\lambda}_n(x) = \frac{1}{h_n} \int \mathrm{K}\left(\frac{x-y}{h_n}\right) d\Lambda_n(y),$$

Ces deux estimateurs dépendent d'un noyau K et d'une fenêtre h qu'il faut choisir pour les calculer. Nous savons que le choix du noyau K n'est pas déterminant, et nous choisissons le noyau d'Epanechnikov. En revanche, le choix de h_n est crucial. Nous calculons la fenêtre optimale par minimisation de la distance L_1 entre chaque estimateur et la vraie courbe, sur un ensemble de valeurs de la fenêtre $h_n \in [0.1, 2]$.

Conclusion Les graphes des figures 6.1, 6.2, 6.3 and 6.4 confirment la bonne performance des deux estimateurs. On voit aussi que les résultats s'améliorant avec l'augmentation de la taille de l'échantillon, sans surprise. Les courbes à gauche (resp. droite) sont obtenues pour $n = 300$ (resp. $n = 500$).

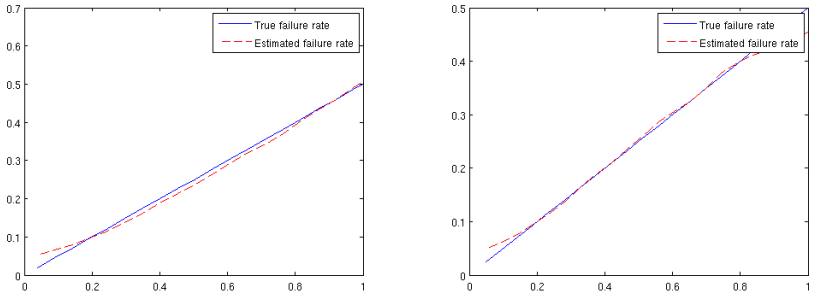

FIGURE 6.2 – Estimateur du taux de hasard pour le modèle 1, avec $n = 300$ et $n = 500$

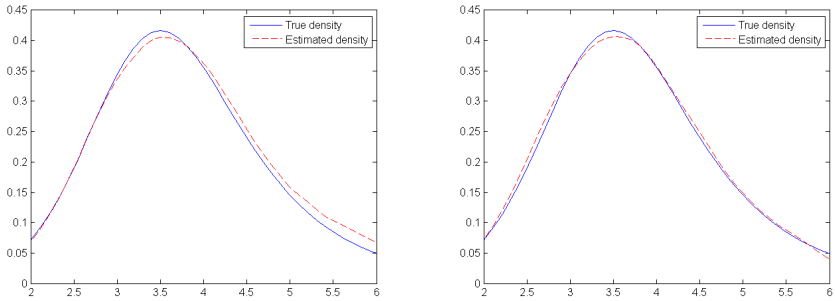

FIGURE 6.3 – Estimateur de la densité pour le modèle 2, avec $n = 300$ et $n = 500$

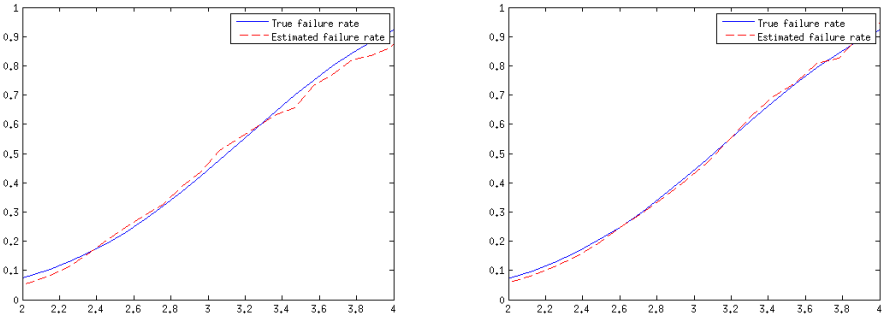

FIGURE 6.4 – Estimateur du taux de hasard pour le modèle 2, avec $n = 300$ et 500

Conclusion et Perspectives

Nous nous sommes intéressés à des estimateurs de la fonction de survie, de la densité et du taux de hasard pour lesquels nous avons montré la convergence uniforme presque complète et donné des vitesses de convergence, dans un contexte de censure mixte et pour des données indépendantes. Bien que chacune des fonctions estimées caractérise la loi de la variable aléatoire étudiée, l'un ou l'autre des estimateurs introduits peut s'avérer plus approprié selon le contexte et l'objectif poursuivi.

Nous avons aussi fourni quelques résultats relatifs au processus empirique pour ce même modèle. Il s'agit d'une loi fonctionnelle du logarithme itéré pour les accroissements du processus empirique, qui nous ont permis d'établir des lois fortes pour des estimateurs à noyau de la fonction de densité et du taux de hasard.

Ceci constitue une extension, au cas des données soumises à la censure mixte, des résultats disponibles pour des données complètes ou censurées à droite.

Une première perspective de recherche serait l'extension de la loi fonctionnelle du logarithme itéré au cas uniforme, d'une manière similaire au travail de Deheuvels et Einmahl (2000).

Une question importante concernant l'estimation par la méthode du noyau est le choix du paramètre de lissage h. Il serait donc intéressant de développer des méthodes qui permettent ce choix à partir des données. Pour justifier théoriquement ce genre de méthode, il faudrait aussi établir la convergence uniforme par rapport au paramètre de lissage h.

Une autre perspective est d'étudier ces mêmes propriétés sous divers types de dépendances : α-mélange, φ-mélange, association, *etc.*

Le modèle de censure mixte est intéressant et réaliste lorsque nous le rapportons au domaine de la fiabilité. Depuis son introduction par Patilea et Rolin (2006), un certain nombre de travaux le concernant ont vu le jour. Ci-

tons : l'estimation de la régression (Kebabi *et al.*, 2011 ; Kebabi et Messaci, 2012 ; Messaci, 2010) et des quantiles conditionnels (Volgushev et Dette, 2013), la loi du logarithme itéré Messaci et Nemouchi (2011, 2013) pour l'estimateur de Patilea-Rolin et l'estimation de la fonction de survie (Shen, 2011, 2012).

Des résultats similaires pour des données dépendantes peuvent être recherchés. L'estimation de la densité conditionnelle et du mode conditionnel peut aussi faire l'objet d'un travail de recherche.

Le modèle de censure double est un modèle semblable au modèle de censure mixte abordé dans cette thèse dans le sens où on observe comme dans le cas de la censure mixte $\max(\min(X, R), L)$ mais où X, R et L ne sont pas indépendantes, mais $L \leqslant R$ presque sûrement. Il serait intéressant d'étudier les résultats acquis et en perspectives sous censure mixte, au cas de la censure double.

Enfin nous pensons que la réalisation de diverses applications sur des données régies par un modèle de censure mixte serait très intéressante.

Bibliographie

D. Bitouzé, B. Laurent et P. Massart : A Dvoretzky-Kiefer-Wolfowitz type inequality for the Kaplan-Meier estimator. *Annales de l'Institut Henri Poincare (B) Probability and Statistics*, 35(6) :735–763, 1999.

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

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

P. Deheuvels et J. H. J. Einmahl : On the strong limiting behavior of local functionals of empirical processes based upon censored data. *The Annals of Probability*, 24(1) :504–525, 1996.

P. Deheuvels et J. H. J. Einmahl : Functional limit laws for the increments of kaplan-meier product-limit processes and applications. *The Annals of Probability*, 28(3) :1301–1335, 2000.

P. Deheuvels et D. Mason : Functional laws of the iterated logarithm for the increments of empirical and quantile processes. *Ann. Prob.*, 20 :1248–1287, 1992.

P. Deheuvels et D. M. Mason : Functional laws of the iterated logarithm for local empirical processes indexed by sets. *The Annals of Probability*, 22 (3) :1619–1661, 1994.

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

R. L. Dobrushin : Generalization of kolmogorov's equations for markov processes with a finite number of possible states. *Matematicheskii Sbornik*, 75(3) :567–596, 1953.

J. D. Dollard et C. N. Friedman : *Product integration with application to differential equations*. Addison-Wesley, Reading, Mass., 1984.

A. Dvoretzky, J. Kiefer et J. Wolfowitz : Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *The Annals of Mathematical Statistics*, p. 642–669, 1956.

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

H. Finkelstein : The law of the iterated logarithm for empirical distribution. *The Annals of Mathematical Statistics*, 42(2) :607–615, 1971.

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

A. Földes, L. Rejtő et B. B. Winter : Strong consistency properties of nonparametric estimators for randomly censored data, II : Estimation of density and failure rate. *Periodica Mathematica Hungarica*, 12(1) :15–29, 1981.

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

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

R. D. Gill : Lectures on survival analysis. *In* P. Bernard, éd. : *Lectures on Probability Theory*, vol. 1581 de *Lecture Notes in Mathematics*, p. 115–241. Springer Berlin Heidelberg, 1994.

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

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

W. Hardle : A law of the iterated logarithm for nonparametric regression function estimators. *The Annals of Statistics*, 12(2) :624–635, 1984.

B. Helton : Integral equations and product integrals. *Pacific Journal of Mathematics*, 16(2) :297–322, 1966.

J. Helton : Mutual existence of sum and product integrals. *Pacific Journal of Mathematics*, 56(2) :495–516, 1975.

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

J. Huang : Asymptotic properties of nonparametric estimation based on partly interval-censored data. *Statistica Sinica*, 9 :501–519, 1999.

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

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

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

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

A. Kitouni, M. Boukeloua et F. Messaci : Rate of strong consistency for nonparametric estimators based on twice censored data. *Statistics & Probability Letters*, 96 :255–261, 2015.

M. R. Kosorok : *Introduction to Empirical Processes and Semiparametric Inference*. Springer Verlag, New York, 2008.

J. S. Mac Nerney : Integral equations and semigroups. *Illinois Journal of Mathematics*, 7(1) :148–173, 03 1963.

P. Masani : Multiplicative partial integration and the trotter product formula. *Advances in Mathematics*, 40(1) :1–9, 1981.

P. Massart : The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. *The Annals of Probability*, 18(3) :1269–1283, 1990.

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

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

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

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

E. Parzen : On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33 :1065–1076, 1962.

V. Patilea et J.-M. Rolin : Product limit estimators of the survival function for doubly censored data. Rap. tech., Institut de Statistique, Université Catholique de Louvain, July 2001.

V. Patilea et J.-M. Rolin : Product-limit estimators of the survival function with left or right censored data. Rap. tech., Institut de Statistique, Université Catholique de Louvain, 2004.

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

R. Peto : Experimental survival curves for interval censored data. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 22(1) :86–91, 1973.

M. Rosenblatt : Remarks on some nonparametric estimates of density function. *The Annals of Mathematical Statistics*, 27 :832–837, 1956.

S. O. Samuelsen : Asymptotic theory for non-parametric estimators from doubly censored data. *Scandinavian Journal of Statistics*, 16 :1–21, 1989.

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

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

G. R. Shorack et J. W. Wellner : *Empirical processes with applications to statistics*. John Wiley & Sons, 1986.

A. Slavík : *Product integration, its history and applications*. Matfyzpress Prague, 2007.

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

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

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

A. W. van der Vaart et J. A. Wellner : *Weak Convergence and Empirical Processes : With Applications in Statistics*. Springer Verlag, New York, 1996.

S. Volgushev et H. Dette : Nonparametric quantile regression for twice censored data. *Bernoulli*, 19(3) :748–779, 08 2013.

V. Volterra et B. Hostinskỳ : *Opérations infinitésimales linéaires : applications aux équations différentielles et fonctionnelles*, vol. 46. Gauthier-Villars, 1938.

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

X. Xiang : Law of the logarithm for density and hazard rate estimation for censored data. *Journal of Multivariate Analysis*, 49 :278–286, 1994.

العمليّات التّجريبيّة في حالة الحجب والتقدير الدّالّي

نثبت في هذه الأطروحة التقارب المنتظم شبه الكامل، مع سرعة التقارب، لمقدّري دالّة البقاء ودالّة الخطر في حالة الحجب المزدوج. في هذا النموذج، المتغيّر محل الاهتمام X معرّض للحجب من اليمين من قبل متغير R، و $\min(X, R)$ معرّض للحجب من اليسار من طرف متغير L. ويتميز هذا النموذج باستقلال المتغيّرات الكامنة X و R و L. نستنتج نتائج تقارب مماثلة من أجل مقدّرات كثافة الاحتمال ومعدل الخطر باستعمال طريقة النّواة.

نقدّم أيضًا قانونًا دالّيًا للوغارتم المكرّر لتزايدات العملية التجريبية، المبنية على بيانات تخضع لحجب مزدوج. ونستنتج منه قوانين قويّة لمقدرات النّواة للكثافة الاحتمال ومعدل الخطر.

نخصّص إيضا فصلًا لدراسة الجداء التكامل، وهو أداة ذات أهمية أساسية في تحليل البقاء.

ننتهي مع دراسة محاكاة لتوضيح اداء المقدرات المقترحة.

العبارات الرئيسيّة: العمليات التجريبية، قانون اللّوغارتم المكرّر، الحجب المزدوج، التقارب شبه الكامل، سرعة التقارب، المحاكاة، مقدّر نهاية الجداء، كثافة الاحتمال، معدل الخطر.

Empirical processes in a censorship model and functional estimation

In this thesis we establish the almost complete uniform convergence, with rate, of estimators of the survival function and the hazard function under twice censorship. In this model, the variable of interest X is right-censored by a variable R, and $\min(X, R)$ is censored on the left by a variable L. This model is characterized by the independence of the latent variables X, R and L. We deduce similar convergence results for kernel estimators of the density and the hazard rate.

We also give a functional law of the iterated logarithm for increments of empirical process based on data subject to twice censorship. We derive strong laws for kernel estimators of the density and hazard rate.

We also reserve a chapter to the study of the product integral, a tool of primary importance in survival analysis.

We finish with a simulation study to illustrate the performances of the introduced estimators.

Keywords: Empirical processes, law of the iterated logarithm, twice censoring, almost complete convergence, convergence rates, simulation, product-limit estimator, density, failure rate.

Processus empiriques dans un modèle de censure et estimation fonctionnelle

Nous établissons dans cette thèse la convergence presque complète uniforme, en précisant la vitesse, pour les estimateurs de la fonction de survie et de la fonction de hasard sous le modèle de censure mixte. Dans le modèle considéré, la variable d'intérêt X est censurée à droite par une variable R, et $\min(X, R)$ est censuré à gauche par une variable L. Ce modèle est caractérisé par l'indépendance des variables latentes X, R, et L. Nous déduisons des résultats analogues de convergence pour des estimateurs à noyau de la densité et du taux de hasard.

Nous donnons aussi une loi fonctionnelle du logarithme itéré pour les accroissements du processus empirique basé sur des données soumise à la censure mixte. Nous déduisons des lois fortes pour des estimateurs à noyau de la fonction de densité et du taux de hasard.

Nous réservons aussi un chapitre à l'étude du produit intégral, outil de première importance en analyse de survie.

Nous terminons par une étude de simulation pour illustrer les performances des estimateurs introduits.

Mots clés : Processus empirique, loi du logarithme itéré, censure mixte, convergence presque complète, taux de convergence, simulation, estimateur produit-limite, densité, taux de hasard.