École Normale Supérieure de Rennes

Rapport de stage

Introduction à l'analyse de survie, méthodes non-paramétriques et extension au cadre de censure informative

Alice MORINIÈRE

Stage encadré par Mikael Escobar-Bach au sein du LAREMA

Stage effectué en mai 2022

Table des matières

1 Introduction à l'analyse de survie : motivations et définitions

L'analyse de survie est une branche des statistiques qui s'intéresse à l'étude du temps s'écoulant avant qu'un certain évènevement ne se produise. Cette durée étudiée est nommée temps de survie. Elle peut par exemple représenter un fait de type médical comme le temps de guérison après un certain traitement, sociologique tel que le temps écoulé avant un mariage ou un premier enfant, ou encore économique à l'instar de l'attente avant l'achat d'un certain produit ou encore avant l'obtiention d'un premier emploi après la fin d'études.

Cependant, lors de l'étude de phénomènes comme ceux-ci, la donnée souhaitée peut être inaccessible à cause d'une certaine censure. Prenons l'exemple du cas de la guérison : si le sujet décède avant la fin de l'étude, s'il décide ou est contraint de quitter l'étude pour une autre raison ou si l'étude se finit avant la fin de sa guérison, la seule information obtenue est alors que la guérison serait arrivée après cet évènement de censure.

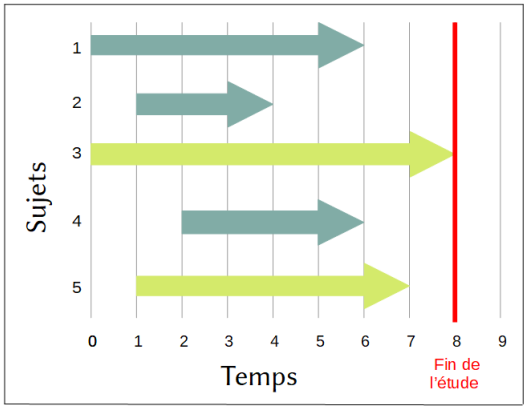

FIGURE 1 – Exemple de données censurées : chaque flèche représente un temps observé, les flèches bleues correspondent à des cas où l'on observe directement l'évènement d'intérêt, les vertes ceux où l'on observe une censure

Les données obtenues dans ce cadre sont donc composées pour chaque sujet du minimum entre le temps de survie et la censure, ainsi qu'un indicateur permettant de savoir lequel de ces deux évènements a été observé. On peut traduire cela de la façon suivante : si $Y \in \mathbb{R}_+$ est une variable aléatoire répresentant la date d'arrivée de l'évènement d'intérêt, et si $C \in \mathbb{R}_+$ est une variable aléatoire répresentant la censure, alors la donnée aléatoire est un couple aléatoire (T, δ) composé de

$$T = \min(Y, C) \quad \text{et} \quad \delta = \mathbb{1}_{\{Y \leq C\}},$$

où δ est appelé indicateur de censure.

Ce type de considérations correspond à ce que l'on appelle une censure à droite. D'autres censures existent, telles que la censure à gauche, où l'évènement peut s'être produit avant la période de l'étude, on observe alors le maximum entre l'évènement souhaité et la censure. On peut enfin s'intéresser à la censure par intervalle, où la seule information obtenue est alors un intervalle contenant le temps d'intéret. Toutes ces formes de censure ne doivent pas être confondues avec le principe de troncature : dans ce cas, si le temps de survie n'appartient pas à un certain intervalle que l'on se fixe (tel que le temps de l'étude par exemple), alors aucune information n'est disponible. Cela diffère de la censure dans la mesure où une donnée, même censurée, apporte une information partielle sur la variable à étudier.

Nous allons donc chercher dans toute la suite à trouver des stratégies pour estimer la répartition de l'évènement d'intérêt à partir de telles données, en utilisant aussi l'information contenue par les données censurées.

Pour tout ce qui va suivre, nous nous placerons dans un certain espace probabilisé complété $(\Omega, \mathscr{F}, \mathbb{P})$.

2 Intégrale de Stieltjes

Nous allons tout d'abord nous intéresser à la construction d'un outil qui nous sera utile par la suite.

2.1 Fonctions à variation bornée

Dans toute la suite, a et b désigneront deux réels tels que $a \leq b$.

Définition 2.1.1. *Soit $f : [a,b] \to \mathbb{R}$. Soit $\sigma = (a = x_0 < x_1 < ... < x_n = b)$ une subdivision de $[a,b]$. On appelle variation de f par rapport à σ :*

$$V(f,\sigma) = \sum_{i=1}^{n} |f(x_i) - f(x_{i-1})|,$$

et on appelle variation totale de f sur $[a,b]$:

$$V_a^b(f) = \sup_{\sigma} V(f,\sigma).$$

On dit que f est à variation bornée sur $[a,b]$ si $V_a^b(f) < +\infty$.

Remarque 2.1.2. *On remarque qu'une fonction monotone est une fonction à variation bornée. En effet, si l'on suppose $f : [a,b] \to \mathbb{R}$ croissante sur $[a,b]$, alors pour toute subdivision $\sigma = (a = x_0 < x_1 < ... < x_n = b)$ de $[a,b]$, on a :*

$$V(f,\sigma) = \sum_{i=1}^{n} f(x_i) - f(x_{i-1}) = f(b) - f(a).$$

Théorème 2.1.3. *Une fonction $f : [a,b] \to \mathbb{R}$ est une fonction à variation bornée sur $[a,b]$ si et seulement si elle s'écrit comme différence de deux fonctions croissantes sur $[a,b]$.*

Démonstration. Tout d'abord, soient f et g deux fonctions de $[a,b]$ dans $\mathbb{R}$. Par inégalité triangulaire, pour toute subidvision σ de $[a,b]$, on a $V(f-g,\sigma) \leq V(f,\sigma) + V(g,\sigma)$. Ainsi, $V(f-g,\sigma) \leq V_a^b(f) + V_a^b(g)$, et donc en passant à la borne supérieure dans le membre de gauche de l'inégalité, on obtient

$$V_a^b(f-g) \leq V_a^b(f) + V_a^b(g).$$

Ainsi la différence de deux fonctions à variation bornée est une fonction à variation bornée. Or on a déjà vu précédemment qu'une fonction croissante est une fonction à variation bornée. Ainsi la différence de deux fonctions croissantes est bien une fonction à variation bornée.

Réciproquement, soit f une fonction de $[a,b]$ dans $\mathbb{R}$ à variation bornée sur $[a,b]$.

Soit x et x' dans $[a,b]$, tel que $x \leq x'$. Soit σ une subdvision de $[a,x]$, σ' une subdivision de $[x,x']$, et σ'' la concaténation de ces deux subdivisions, qui est donc une subdivision de $[a,x']$. On a alors $V(f,\sigma) \leq V(f,\sigma'') \leq V_a^{x'}(f)$. Ainsi en passant à la borne supérieure dans le membre de gauche de l'inégalité, on obtient

$$V_a^x(f) \leq V_a^{x'}(f). \tag{1}$$

.

De plus, on a $V(f,\sigma'') = V(f,\sigma) + V(f,\sigma') \leq V_a^x(f) + V_x^{x'}(f)$, et donc en passant de nouveau à la borne supérieure dans le membre de gauche, on obtient

$$V_a^{x'}(f) \leq V_a^x(f) + V_x^{x'}(f).$$

De même, $V(f,\sigma) + V(f,\sigma') = V(f,\sigma'') \leq V_a^{x'}(f)$, et donc en passant à la borne supérieure dans le membre de gauche, on obtient

$$V_a^x(f) + V_x^{x'}(f) \leq V_a^{x'}(f).$$

Et donc,

$$V_a^x(f) + V_x^{x'}(f) = V_a^{x'}(f). \tag{2}$$

Ainsi, en prenant $x' = b$, on voit d'après (1) que f est à variation bornée sur $[a,x]$, et on peut donc définir la fonction $g : x \mapsto V_a^x(f)$. On écrit alors

$$\forall x \in [a,b], \quad f(x) = g(x) - (g(x) - f(x)).$$

D'une part, d'après (1), on voit que g est bien une fonction croissante. D'autre part, pour $x \leq x'$ dans $[a,b]$, on obtient, en prenant une subdivision de $[x,x']$ ne contenant que x et x'

$$f(x') - f(x) \leq |f(x') - f(x)| \leq V_x^{x'}. \tag{3}$$

Ainsi d'après (2), $f(x') - f(x) \leq g(x') - g(x)$, et donc $g - f$ est aussi une fonction croissante. Ainsi on a bien écrit f comme différence de deux fonctions croissantes. □

2.2 Définition de l'intégrale

Définition 2.2.1. *Soit α et f deux fonctions réelles définies sur un intervalle $[a,b]$. Soit $\sigma = (x_0, ...x_n)$ une subdivision de l'intervalle, telle que $a = x_0 < x_1 < .... < x_n = b$. On note*

$$\delta = \max_{k \in [\![0,n-1]\!]} (x_{k+1} - x_k).$$

Si la limite

$$\lim_{\delta \to 0} \sum_{k=0}^{n-1} f(\xi_k)(\alpha(x_{k+1}) - \alpha(x_k)),$$

où

$$\forall k \in [\![0, n-1]\!], \quad x_k \leq \xi_k \leq x_{k+1},$$

existe indépendamment du choix de la subdvision et des $(\xi_k)_{k \in [\![0,n-1]\!]}$, alors cette limite est appelée intégrale de Stieltjes de f par rapport à α entre a et b, et est notée

$$\int_a^b f(x) \mathrm{d}\alpha(x).$$

Remarque 2.2.2. *Si l'on prend pour α la fonction identité, alors on retrouve simplement l'intégrale de Riemann.*

Remarque 2.2.3. *A l'instar de l'intégrale de Riemann qui se généralise par la théorie de la mesure de Lebesgue, cette définition de l'intégrale de Stieltjes, aussi appelée Riemann-Stieltjes, se généralise grâce à l'intégrale de Lebesgue-Stieltjes. L'intégrale contre une application α correspond en fait à une intégrale contre une mesure μ telle que pour tout segment $[t,s]$, $\mu([t,s]) = \alpha(s) - \alpha(t)$.*

Donnons maintenant une critère nécessaire et suffisant pour que cette intégrale soit définie dans le cas de fonctions particulières.

Théorème 2.2.4. *Soit f et α deux fonctions réelles et bornées définies sur un intervalle $[a,b]$. On suppose de plus que α est une fonction croissante. Soit $\sigma = (x_0, ..., x_n)$ une subdivision de $[a,b]$. On pose :*

$$M_k = \sup_{x_k \leq x \leq x_{k+1}} f(x), \qquad m_k = \inf_{x_k \leq x \leq x_{k+1}} f(x),$$

$$S_\sigma = \sum_{k=0}^{n-1} M_k(\alpha(x_{k+1}) - \alpha(x_k)) \qquad et \qquad s_\sigma = \sum_{k=0}^{n-1} m_k(\alpha(x_{k+1}) - \alpha(x_k)).$$

Alors l'intégrale de Stieltjes de f par rapport à α sur $[a,b]$ existe si et seulement si

$$\lim_{\delta \to 0} (S_\sigma - s_\sigma) = 0,$$

indépendamment du choix de la subdivision.

Démonstration. Supposons tout d'abord que l'intégrale de Stieltjes de f par rapport à α sur $[a,b]$ existe. Soit $\epsilon > 0$. On peut alors trouver pour tout k dans $[\![0, n-1]\!]$ un ξ_k tel que

$$0 \leq M_k - f(\xi_k) < \epsilon.$$

On a alors

$$\sum_{k=0}^{n-1} f(\xi_k)(\alpha(x_{k+1}) - \alpha(x_k) \quad \leq \quad S_\sigma \quad \leq \quad \sum_{k=0}^{n-1} f(\xi_k)(\alpha(x_{k+1}) - \alpha(x_k)) + \epsilon(\alpha(b) - \alpha(a)),$$

et ainsi en passant à la limite on obtient alors :

$$\int_a^b f(x)\mathrm{d}\alpha(x) \quad \leq \quad \liminf_{\delta \to 0} S_\sigma \quad \leq \quad \limsup_{\delta \to 0} S_\sigma \quad \leq \quad \int_a^b f(x)\mathrm{d}\alpha(x) + \epsilon(\alpha(b) - \alpha(a)).$$

Et donc

$$\lim_{\delta \to 0} S_\sigma = \int_a^b f(x)\mathrm{d}\alpha(x).$$

On prouve de la même manière

$$\lim_{\delta \to 0} s_\sigma = \int_a^b f(x)\mathrm{d}\alpha(x),$$

et on obtient directement que

$$\lim_{\delta \to 0} (S_\sigma - s_\sigma) = 0.$$

Réciproquement, supposons que $\lim_{\delta \to 0} (S_\sigma - s_\sigma) = 0$. On pose

$$\zeta_\sigma = \sum_{k=0}^{n-1} f(\xi_k)(\alpha(x_{k+1}) - \alpha(x_k)).$$

On a alors

$$s_\sigma \leq \zeta_\sigma \leq S_\sigma.$$

D'après notre hypothèse de départ, si l'on montre que S_σ a une limite, alors cela prouvera également que ζ_σ a une limite, et donc que l'intégrale de Stieltjes est bien définie.

Soit $\epsilon > 0$. Nous souhaitons montrer qu'il existe δ_0, tel que si l'on choisit S_{σ_1} et S_{σ_2}, deux sommes liées à deux subdivisions σ_1 et σ_2 de normes respectives δ_1 et δ_2, toutes les deux inférieures à δ_0, alors

$$|S_{\sigma_1} - S_{\sigma_2}| < \epsilon.$$

Soit σ_3 la réunion de σ_1 et σ_2. Montrons

$$s_{\sigma_1} \leq S_{\sigma_3} \leq S_{\sigma_1} \quad \text{et de même} \quad s_{\sigma_2} \leq S_{\sigma_3} \leq S_{\sigma_2}. \tag{4}$$

On pose $\sigma_1 = x_0, ..., x_n$ et $\sigma_3 = \tilde{x}_0, ...\tilde{x}_m$. On note pour tout k dans $[\![0, n-1]\!]$, $\sigma(k) = \{\tilde{x}_i | \tilde{x}_i \in [x_k, x_{k+1}]\}$. D'une part, si $J \subset I$, $\sup_J f \leq \sup_I f$ et $\inf_J f \geq \inf_I f$, et d'autre part, α est supposée croissante, ainsi on obtient :

$$m_k(\alpha(x_{k+1}) - \alpha(x_k)) \leq s_{\sigma(k)} \quad \text{et } M_k(\alpha(x_{k+1}) - \alpha(x_k)) \geq S_{\sigma(k)}.$$

Ainsi, on récupère

$$s_{\sigma_1} = \sum_{i=0}^{n-1} m_k(\alpha(x_{k+1}) - \alpha(x_k)) \leq \sum_{i=0}^{n-1} s_{\sigma(k)} = s_{\sigma_3} \quad \text{et} \tag{5}$$

$$S_{\sigma_1} = \sum_{i=0}^{n-1} M_k(\alpha(x_{k+1}) - \alpha(x_k)) \geq \sum_{i=0}^{n-1} S_{\sigma(k)} = S_{\sigma_3}. \tag{6}$$

Et donc comme $s_{\sigma_3} \leq S_{\sigma_3}$, on obtient

$$s_{\sigma_1} \leq s_{\sigma_3} \leq S_{\sigma_3} \leq S_{\sigma_1},$$

et donc on a en particulier l'inégalité souhaitée pour σ_1, l'inégalité pour σ_2 se démontrant exactement de la même façon. Comme nous avons supposé $\lim_{\delta \to 0} (S_\sigma - s_\sigma) = 0$, il existe δ_0 tel que si la norme δ d'une subdivision σ est inférieur à δ_0, alors

$$S_\sigma - s_\sigma < \frac{\epsilon}{2}. \tag{7}$$

Et donc si l'on prend δ_1 et δ_2 inférieur à δ_0, on obtient d'après (4) et (7) :

$$0 \leq S_{\sigma_1} - S_{\sigma_3} < \frac{\epsilon}{2} \quad \text{et} \quad 0 \leq S_{\sigma_2} - S_{\sigma_3} < \frac{\epsilon}{2}. \tag{8}$$

Et donc

$$`|S_{\sigma_1} - S_{\sigma_2}| < \epsilon,$$

ce qui conclut la preuve. □

Nous avons donc obtenu une condition nécessaire et suffisante qui nous servira par la suite pour prouver l'existence de l'intégrale de Stieltjes dans certains cadres.

2.3 Théorèmes d'existence

Nous allons désormais prouver la bonne définition de l'intégrale de Stieltjes dans deux cadres différents.

Théorème 2.3.1. *Soit f et α deux fonctions réelles bornées définies sur un certain intervalle $[a, b]$. Si f est continue sur $[a, b]$, et si α est à variation bornée sur $[a, b]$, alors l'intégrale de Stieltjes de f par rapport à α de a à b existe.*

Démonstration. On voit grâce au théorème 2.1.3 que l'on peut supposer sans restriction que α est une fonction croissante et bornée sur $[a, b]$. De plus, comme f est une fonction continue sur $[a, b]$ qui est un compact, on sait d'après le théorème de Heine que f est uniformément continue sur $[a, b]$. Soit $\epsilon > 0$. Il existe donc, par uniforme continuité de f, δ_0 tel que pour toute subdivision σ de norme plus petite de δ_0, on ait (avec les notations précédentes)

$$M_k - m_k < \epsilon, \quad \forall k \in [\![0, n-1]\!].$$

Ainsi pour une telle subdivsion, on a (toujours avec les notations précédentes)

$$0 \leq S_\sigma - s_\sigma \leq \epsilon(\alpha(b) - \alpha(a)).$$

On obtient alors

$$\lim_{\delta \to 0} (S_\sigma - s_\sigma) = 0,$$

et donc d'après le théorème 2.2.4, on obtient bien que l'intégrale de f par rapport à α est bien définie. □

Théorème 2.3.2. *Soit f et α deux fonctions réelles bornées définies sur un certain intervalle $[a,b]$. Si f est à variation bornée sur $[a,b]$, et si α est continue sur $[a,b]$, alors l'intégrale de Stieltjes de f par rapport à α de a à b existe, et*

$$\int_a^b f(x)\mathrm{d}\alpha(\mathrm{x}) = f(b)\alpha(b) - f(a)\alpha(a) - \int_a^b \alpha(x)\mathrm{df}(\mathrm{x}).$$

Cette formule est appelée intégration par parties de l'intégrale de Stieltjes.

Démonstration. Soit σ une subdivision quelconque de $[a,b]$. On pose de nouveau

$$\zeta_\sigma = \sum_{k=0}^{n-1} f(\xi_k)(\alpha(x_{k+1}) - \alpha(x_k)),$$

où pour k dans $[\![0, n-1]\!]$, ξ_k est toujours un réel de l'intervalle $[x_k, x_{k+1}]$. En réarrangeant la somme, on obtient

$$\begin{aligned}\zeta_\sigma &= \sum_{k=1}^{n} f(\xi_{k-1})\alpha(x_k) - \sum_{k=0}^{n-1} f(\xi_k)\alpha(x_k) = -\sum_{k=1}^{n-1} \alpha(x_k)(f(\xi_k) - f(\xi_{k-1})) + f(\xi_{n-1})\alpha(b) - f(\xi_0)\alpha(a) \\ &= -\Big(\sum_{k=1}^{n-1} \alpha(x_k)(f(\xi_k) - f(\xi_{k-1}))\Big) - \alpha(a)\big(f(\xi_0) - f(a)\big) - \alpha(b)\big(f(b) - f(\xi_{n-1})\big) + f(b)\alpha(b) - f(a)\alpha(a).\end{aligned}$$

Lorsque la norme $\delta = \max_{k\in[\![0,n-1]\!]}(x_{k+1} - x_k)$ de σ tend vers 0, alors le maximum de $\xi_0 - a$, $\xi_1 - \xi_0$, ..., $\xi_{n-1} - \xi_{n-2}$, $b - \xi_{n-1}$ tend également vers 0. Ainsi, par continuité de f, on a

$$-\alpha(a)\big(f(\xi_0) - f(a)\big) - \alpha(b)\big(f(b) - f(\xi_{n-1})\big) \underset{\delta\to 0}{\longrightarrow} 0.$$

Et d'après le théorème précédent, on obtient

$$\zeta_\sigma \underset{\delta\to 0}{\longrightarrow} -\int_a^b \alpha(x)\mathrm{df}(\mathrm{x}) - f(b)\alpha(b) - f(a)\alpha(a),$$

ce qui conclut la preuve. □

La bonne définition des intégrales sortant des deux cadres précédents seront admis pour la suite.

2.4 Applications utiles à notre étude

Si l'on présente ici l'intégrale de Stieltjes, c'est parce qu'elle est couramment utilisée dans le cadre des statistiques.

On remarque tout d'abord qu'elle permet d'exprimer l'espérance d'une variable aléatoire en fonction de sa fonction de répartition. Soit X une variable aléatoire réelle de fonction de répartition F. Comme dit précédemment, on peut associer à F (fonction bornée et croissante) une mesure de Stieltjes, et alors pour tout intervalle $]s,t]$ de $\mathbb{R}$, on a

$$\int_s^t \mathrm{dF}(\mathrm{u}) = F(t) - F(s) = \mathbb{P}(]s,t]).$$

Comme l'ensemble $\{]s,t]|(s,t) \in \mathbb{R}^2\}$ est stable par intersection finie et engendre la tribu borélienne, on en déduit par le lemme des classes monotones que les mesures dF et P sont égales sur toute la tribu. Alors pour une fonction f quelconque, si $\mathbb{E}(f(X))$ est bien définie, elle vérifie

$$\mathbb{E}(f(X)) = \int f(x)\mathrm{dF}(\mathrm{x}).$$

Une fois cette constatation faite, nous allons désormais prouver un lemme sans rapport apparent à notre sujet mais qui nous sera utile pour la suite

Proposition 2.4.1. *Soit A et f deux fonctions réelles définies sur $[0,t]$, avec $t \in \mathbb{R}$. On suppose que A est cadlag et à variation bornée sur $[0,t]$, et que f est de classe C^1 sur ce même intervalle. On a alors*

$$f(A(t)) - f(A(0)) = \int_0^t f'(A(s^-))\mathrm{dA(s)} + \sum_{s \leq t} \Delta f(A(s)) - f'(A(s^-))\Delta A(s),$$

où $\Delta A(s) = A(s) - A(s^-)$.

Démonstration. Soit $S = \{t | \Delta A(t) \neq 0\}$. Comme A est à variation bornée, elle s'écrit d'après le théorème 2.1.3 comme une différence de deux fonctions croissantes, ainsi, S est un ensemble au plus dénombrable, on note donc $\#S = m$, avec $m \in \bar{\mathbb{N}}$. On pose donc $S = \{t^1, ..., t^m\}$. Soit $0 = t_1 < ... < t_n = t$ une partition de $[0,t]$ incluant S. On écrit :

$$f(A(t)) - f(A(0)) = \sum_{i=1}^n f(A(t_i)) - f(A(t_{i-1})) = \sum_{i=1}^n f(A(t_i^-)) - f(A(t_{i-1})) + \sum_{i=1}^n \Delta f(A(t_i)).$$

Soit i dans $[\![1,n]\!]$. D'après le caractère C^1 de f, on sait que f est dérivable sur $]A(t_{i-1}), A(t_i^-)[$ et continue sur $[A(t_{i-1}), A(t_i^-)]$, ainsi le théorème des accroissements finis nous donne l'existence de c_i dans $]A(t_{i-1}), A(t_i)[$ tel que

$$f(A(t_i^-)) - f(A(t_{i-1})) = f'(c_i)(A(t_i^-) - A(t_{i-1})).$$

Or A étant cadlag, on sait qu'elle est continue sur $[t_{i-1}, t_i^-]$, ainsi $A([t_{i-1}, t_i^-])$ est un intervalle, qui contient $A(t_i^-)$ et $A(t_{i-1})$, donc qui contient $[A(t_{i-1}), A(t_i^-)]$, et donc qui contient c_i. Ainsi, il existe s_i dans $[t_{i-1}, t_i^-]$ tel que $c_i = A(s_i)$.

On obtient alors :

$$\begin{aligned} f(A(t)) - f(A(0)) &= \sum_{i=1}^n f'(A(s_i))(A(t_i^-) - A(t_{i-1})) + \sum_{i=1}^n \Delta f(A(t_i)) \\ &= \sum_{i=1}^n f'(A(s_i))(A(t_i) - A(t_{i-1})) + \sum_{i=1}^n \Delta f(A(t_i)) - f'(A(s_i))\Delta A(t_i), \end{aligned}$$

et on remarque que la deuxième somme ne contient en réalité que les termes pour lesquels t_i est dans S.

On fait tendre la norme de la subdivision $(t_0, ..., t_n)$ vers 0. On a tout d'abord $s_i \longrightarrow t_i^-$ et donc $f'(A(s_i)) \longrightarrow f'(A(t_i^-))$ par continuité de $f' \circ A$, et

$$\sum_{i=1}^n f'(A(s_i))(A(t_i) - A(t_{i-1})) \longrightarrow \int_0^t f'(A(s^-))\mathrm{dA(s)}.$$

Et on obtient donc

$$f(A(t)) - f(A(0)) = \int_0^t f'(A(s^-))\mathrm{dA(s)} + \sum_{s \leq t} \Delta f(A(s)) - f'(A(s^-))\Delta A(s).$$

□

Nous allons désormais pouvoir utiliser les outils mis en place afin de parvenir à la construction de deux estimateurs classiques de la fonction de survie, celui de Kaplan-Meier et celui de Beran.

3 Estimateur de Kaplan-Meier

Dans toute cette section, nous supposerons que la variable à étudier est représentée par une variable aléatoire $Y \in \mathbb{R}_+$, que la variable de censure est répresentée par une autre variable aléatoire $C \in \mathbb{R}_+$, et que ces deux variables

sont indépendantes. La variable observée est alors le couple (T, δ) avec $T = \min(Y, C)$ et $\delta = \mathbb{1}_{\{Y \leq C\}}$. On note F la fonction de répartition de Y, G celle de C et H celle de T. On note de plus S la fonction de survie de $Y : S = 1 - F$. Enfin, on note $\tau_F = \inf\{t \in \mathbb{R}_+ | F(t) = 1\}$ et $\tau_H = \inf\{t \in \mathbb{R}_+ | H(t) = 1\}$.

On suppose alors que l'on dispose d'un échantillon aléatoire $\{(T_i, \delta_i)\}_{1 \leq i \leq n}$, où les $\{(T_i, \delta_i)\}$ sont indépendants et identiquement distribués selon la loi de (T, δ). Lorsque les $\{T_i\}_{1 \leq i \leq n}$ sont rangés par ordre croissant, on les notera $\{T_{(i)}\}_{1 \leq i \leq n}$, et on notera alors de même $\{(T_{(i)}, \delta_{(i)})\}_{1 \leq i \leq n}$.

3.1 Construction de l'estimateur

Pour construire l'estimateur de Kaplan-Meier, nous allons essayer de généraliser certaines formules valides dans le cas où la variable d'intérêt Y est à densité.

Définition 3.1.1. *Si Y est une variable aléatoire à densité, on définit le risque instantané en $t \in \mathbb{R}_+$ comme la probabilité que Y appartienne à un petit intervalle autour de t, conditionnellement au fait que Y soit supérieur à t :*

$$\lambda(t) = \lim_{h \to 0} \frac{\mathbb{P}(t \leq Y < t + h | Y \geq t)}{h}.$$

Proposition 3.1.2. *Si Y est à densité, de fonction de densité f, alors λ, f et S sont reliés de la manière suivante : pour tout t dans $\mathbb{R}_+$ tel que $t < \tau_F$,*

$$S(t) = \exp(-\int_0^t \lambda(u)\mathrm{d}u) = \exp(-\int_0^t \frac{f(u)}{S(u)}\mathrm{d}u).$$

Démonstration. Soit t dans $\mathbb{R}_+$. On a

$$\begin{aligned}\lambda(t) &= \lim_{h \to 0} \Big(\frac{\mathbb{P}((t \leq Y < t + h) \cap (Y \geq t))}{\mathbb{P}(Y \geq t) \times h} \Big) = \lim_{h \to 0} \Big(\frac{\mathbb{P}(t \leq Y < t + h)}{h} \Big) \times \frac{1}{\mathbb{P}(Y \geq t)} \\ &= \lim_{h \to 0} \Big(\frac{S(t) - S(t + h)}{h} \Big) \times \frac{1}{S(t)} = \frac{-S'(t)}{S(t)} = \frac{f(t)}{S(t)}.\end{aligned}$$

Et donc

$$\lambda(t) = [-log(S(t))]'.$$

Ainsi en intégrant on obtient bien le résultat souhaité. □

Nous allons alors essayer de généraliser ce type d'écriture dans le cas où Y n'est pas nécessairement à densité.

Définition 3.1.3. *On définit le risque cumulé de Y en $t < \tau_F$ comme l'intégrale suivante :*

$$\Lambda(t) = \int_0^t \frac{\mathrm{d}F(u)}{S(u^-)}.$$

On a vu que dans le cas où Y est à densité, on a la relation $S = \exp(-\Lambda)$. Généralisons désormais cette formule dans le cas général.

Proposition 3.1.4. *Si l'on écrit Λ de la façon suivante :*

$$\forall t \in \mathbb{R}_+, \quad \Lambda(t) = \Lambda_c(t) + \sum_{s \leq t} \Delta\Lambda(s),$$

où Λ_c est la partie continue de Λ, et où $\Delta\Lambda(s) = \Lambda(s) - \Lambda(s^-)$, alors on a la relation suivante entre S et Λ :

$$\forall t \in \mathbb{R}_+, \quad S(t) = \exp\Big(- \Lambda_c(t) + \sum_{s \leq t} \log(1 - \Delta\Lambda(s)) \Big). \tag{9}$$

Démonstration. Soit t dans $\mathbb{R}_+$. On utilise le lemme 2.4.1 en posant $f = \log$ et $A = 1 - F$, et on obtient alors :

$$\begin{aligned}\log(S(t)) &= \log(1-F(t)) - \log(1-F(0)) \\ &= \int_0^t \frac{\mathrm{d}(1-F(s))}{1-F(s^-)} + \sum_{s\le t} \log\Big(\frac{1-F(s)}{1-F(s^-)}\Big) - \frac{\Delta F(s)}{1-F(s^-)} \\ &= -\int_0^t \frac{\mathrm{d}F(s)}{S(s^-)} + \sum_{s\le t} \log\Big(\frac{1-F(s)}{1-F(s^-)}\Big) - \frac{\Delta F(s)}{1-F(s^-)} \\ &= -\Lambda_c(t) + \sum_{s\le t} \Delta\Lambda(s) + \sum_{s\le t} \log\Big(\frac{1-F(s)}{1-F(s^-)}\Big) - \frac{\Delta F(s)}{1-F(s^-)}.\end{aligned}$$

Or, pour tout s dans $\mathbb{R}_+$

$$\Delta\Lambda(s) = \Delta\Big(\int_0^t \frac{\mathrm{d}F(u)}{S(u^-)}\Big) = \frac{\Delta F(s)}{S(s^-)}.$$

Ainsi

$$\log(S(t)) = -\Lambda_c(t) + \sum_{s\le t} \log\Big(\frac{1-F(s)}{1-F(s^-)}\Big) = -\Lambda_c(t) + \sum_{s\le t} \log(1-\Delta\Lambda(s)).$$

□

Grâce à ce lien entre la fonction de survie et le risque cumulé, on peut imaginer que si l'on parvient à obtenir une estimation de Λ, il sera possible de remonter à une estimation de S. Cependant, l'expression de Λ dépend elle même de S, nous devons donc pour sortir de l'impasse trouver une autre expression du risque cumulé.

Proposition 3.1.5. *Pour tout t dans $\mathbb{R}_+$ tel que $t < \tau_H$, on a*

$$\Lambda(t) = \int_0^t \frac{\mathrm{d}H^u(s)}{1-H(s^-)},$$

où H^u est défini comme la sous-distribution suivante :

$$\forall s \in \mathbb{R}_+, \quad H^u(s) = \mathbb{P}(T \le s, \delta = 1).$$

Démonstration. Soit t dans $\mathbb{R}_+$. On écrit alors :

$$H^u(t) = \mathbb{P}(Y \le t, Y \le C) = \mathbb{E}(\mathbb{1}_{\{Y\le t\}}\mathbb{1}_{\{Y\le C\}}) = \int_{\mathbb{R}_+^2} \mathbb{1}_{\{Y\le t\}}\mathbb{1}_{\{Y\le C\}}\,\mathrm{d}(\mathbb{P}_Y \otimes \mathbb{P}_C).$$

Alors par indépendance de Y et C, on obtient :

$$\begin{aligned}H^u(t) &= \int_{\mathbb{R}_+^2} \mathbb{1}_{\{u\le t\}}\mathbb{1}_{\{u\le c\}}\,\mathrm{d}\mathbb{P}_Y(u)\mathrm{d}\mathbb{P}_C(c) = \int_{u\in\mathbb{R}_+} \mathbb{1}_{\{u\le t\}} \int_{c\in\mathbb{R}_+} \mathbb{1}_{\{u\le c\}}\,\mathrm{d}\mathbb{P}_C(c)\,\mathrm{d}\mathbb{P}_Y(u) \\ &= \int_0^t \mathbb{E}(\mathbb{1}_{u\le C})\,\mathrm{d}F(u) = \int_0^t (1-G(u^-))\mathrm{d}F(u).\end{aligned}$$

Or l'ensemble $\{[0,t] \mid t \in \mathbb{R}_+\}$ est stable par intersection finie et engendre la tribu borélienne. Donc le lemme des classes monotones nous indique que, comme les mesures $\mathrm{d}H^u$ et $(1-G^-)\mathrm{d}F$ coincident sur cet ensemble, elles sont égales.

De plus, par indépendance de Y et C, on a

$$\begin{aligned}(1-H(t)) &= \mathbb{P}(T > t) = \mathbb{P}(\min(Y,C) > t) = \mathbb{P}((Y>t)\cap(C>t)) \\ &= \mathbb{P}(Y>t)\,\mathbb{P}(C>t) = (1-F(t))(1-G(t)).\end{aligned}$$

Ainsi, on obtient

$$\Lambda(t) = \int_0^t \frac{\mathrm{d}F(s)}{1-F(s^-)} = \int_0^t \frac{\mathrm{d}H^u(s)}{(1-F(s^-))(1-G(s^-))} = \int_0^t \frac{\mathrm{d}H^u(s)}{1-H(s^-)}.$$

□

Maintenant que l'on a obtenu une expression de Λ, nous allons pouvoir essayer d'approximer les différentes fonctions qui la composent.

Définition 3.1.6. *On définit pour approximer les fonctions $1-H^-$ et H^u les fonctions empiriques données par :*

$$\forall s \in \mathbb{R}_+, \quad 1-H_n(s^-) = \frac{1}{n}\sum_{j=1}^n \mathbb{1}_{\{T_j \geq s\}} \quad et \quad H_n^u(s) = \frac{1}{n}\sum_{j=1}^n \mathbb{1}_{\{T_j \leq s, \delta_j = 1\}}.$$

On définit alors un estimateur de Λ donné par :

$$\forall t \in \mathbb{R}_+, \quad \Lambda_n(t) = \int_0^t \frac{\mathrm{d}H_n^u(s)}{1-H_n(s^-)}.$$

Remarque 3.1.7. *On sait d'après la loi des grands nombres que les fonctions de répartition empiriques H_n et H_n^u convergent presque sûrement vers H et H^u respectivement .*

Il est alors possible de calculer cette intégrale. Soit t dans $\mathbb{R}_+$. On a en effet

$$\Lambda_n(t) = \int_0^t \frac{\mathrm{d}H_n^u(s)}{1-H_n(s^-)} = \frac{1}{n}\sum_{i=1}^n \int_0^t \frac{\mathrm{d}\mathbb{1}_{\{T_i \leq s, \delta_i = 1\}}}{1-H_n(s^-)} = \frac{1}{n}\sum_{i=1}^n \frac{\delta_i}{1-H_n(T_i^-)}\mathbb{1}_{\{T_i \leq t\}}.$$

Et donc :

$$\Lambda_n(t) = \frac{1}{n}\sum_{T_{(i)} \leq t} \frac{\delta_{(i)}}{1-H_n(T_{(i)}^-)}. \tag{10}$$

Or pour tout i dans $[\![1, n]\!]$, on a

$$1-H_n(T_{(i)}^-) = \frac{1}{n}\sum_{j=1}^n \mathbb{1}_{\{T_{(j)} \geq T_{(i)}\}} = \frac{1}{n} \times (n-i+1).$$

Ainsi

$$\Lambda_n(t) = \sum_{T_{(i)} \leq t} \frac{\delta_{(i)}}{n-i+1}.$$

En réinjectant cette expression dans la formule (9), et en remarquant que Λ_n est une fonction en escalier (et donc que sa partie continue est nulle), on obtient un estimateur de la fonction de survie S de la variable Y :

$$S_n(t) = \exp\Big(\sum_{T_{(i)} \leq t} \log(1-\Delta\Lambda_n(T_{(i)}))\Big) = \prod_{T_{(i)} \leq t} \big(1-\Delta\Lambda_n(T_{(i)})\big) = \prod_{T_{(i)} \leq t} \big(1-\frac{\delta_{(i)}}{n-i+1}\big). \tag{11}$$

Définition 3.1.8. *On définit l'estimateur de Kaplan-Meier, aussi appelé estimateur produit-limite, comme la fonction en escalier suivante :*

$$\forall t \in \mathbb{R}_+, \quad S_n(t) = \prod_{T_{(i)} \leq t} \Big(1-\frac{1}{n-i+1}\Big)^{\delta_{(i)}}.$$

Remarque 3.1.9. *Cet estimateur permet de prendre en considération l'information apportée par les données censurées, à l'inverse de l'estimateur naïf donné par la formule*

$$\tilde{S_n}(t) = \frac{1}{n_\delta} \sum_{\delta_i=1} \mathbb{1}_{\{T_i>t\}},$$

où $n_\delta = \#\{i \in [\![1,n]\!] | \delta_i = 1\}$. *En effet, celui-ci consiste à ne considérer que les données non censurées, mais cela peut fausser le résultat asymptotique donné par la loi des grands nombres :*

$$\tilde{S_n}(t) = \frac{n}{n_\delta} \times \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{T_i>t,\delta_i=1\}} \underset{n\to\infty}{\longrightarrow} \frac{\mathbb{P}(T>t, \delta=1)}{\mathbb{P}(\delta=1)}.$$

La limite obtenue n'est pas $S(t)$ *si* T *et* δ *ne sont pas indépendants.*

Une fois l'estimateur de Kaplan-Meier construit, nous allons désormais prouver qu'il converge bien en effet vers la fonction de survie de Y.

3.2 Théorème de Glivenko-Cantelli

Pour prouver la consistance de l'estimateur, nous allons avoir besoin du théorème de Glivenko-Cantelli. Avant d'en voir l'énoncé ainsi que la preuve, intéressons nous à un lemme qui nous sera utile pour cette dernière.

Théorème 3.2.1. ***Théorèmes de Dini*** *Soit* $a < b$ *deux réels et* $(f_n)_{n\in\mathbb{N}}$ *une suite de fonctions de* $[a,b]$ *dans* $\mathbb{R}$ *qui converge simplement vers* $f : [a,b] \to \mathbb{R}$.

1. *Si la suite des* $(f_n)_{n\in\mathbb{N}}$ *est croissante et si* f *et les* f_n *sont des fonctions continues alors la suite* $(f_n)_{n\in\mathbb{N}}$ *converge uniformément vers* f.
2. *Si les* f_n *sont des fonctions croissantes et si* f *est une fonction continue, alors la suite* $(f_n)_{n\in\mathbb{N}}$ *converge uniformément vers* f.

Démonstration. Prouvons tout d'abord le premier théorème. On suppose donc que la suite des $(f_n)_{n\in\mathbb{N}}$ est croissante, et que f et les f_n sont continues. On pose pour tout $n \in \mathbb{N}$, $g_n = f - f_n$, qui est donc une fonction continue, positive et qui converge simplement vers 0. Pour montrer la convergence uniforme de la suite, on doit montrer $\lim_{n\to\infty} \sup_{x\in[a,b]} g_n(x) = 0$, ie

$$\forall \epsilon > 0, \quad \exists N \in \mathbb{N}, \quad \forall n \geq N, \quad \sup_{x\in[a,b]} g_n(x) \leq \epsilon.$$

On raisonne par l'absurde et on suppose :

$$\exists \epsilon > 0, \quad \forall k \in \mathbb{N}, \quad \exists n_k \geq k, \quad \sup_{x\in[a,b]} g_{n_k}(x) > \epsilon.$$

Comme pour tout $k \in \mathbb{N}$, la fonction g_{n_k} est une fonction continue sur le compact $[a,b]$, on en déduit qu'il existe x_{n_k} tel que $\sup_{x\in[a,b]} g_{n_k}(x) = x_{n_k}$. On obtient ainsi une suite (x_{n_k}) dans $[a,b]$. Toujours parce que $[a,b]$ est compact, on sait qu'il existe une extractrice $\phi : \mathbb{N} \to \mathbb{N}$ tel que la suite $(x_{\phi(n_k)})_{k\in\mathbb{N}}$ converge vers x_∞ dans $[a,b]$.

Soit p dans $\mathbb{N}$. On possède $k \in \mathbb{N}$ tel que $\phi(n_k) \geq p$, et donc par croissance de la suite des $(f_n)_{n\in\mathbb{N}}$ et par définition des $(x_{n_k})_{k\in\mathbb{N}}$, on a alors

$$f(x_{\phi(n_k)}) - f_p(x_{\phi(n_k)}) \geq f(x_{\phi(n_k)}) - f_{\phi(n_k)}(x_{\phi(n_k)}) > \epsilon.$$

Et donc, en faisant tendre k vers l'infini, on obtient $f(x_\infty) - f_p(x_\infty) > \epsilon$. Or comme p est arbitrairement grand, cela contredit la convergence des f_n. On obtient donc bien la convergence uniforme des $(f_n)_{n\in\mathbb{N}}$.

Passons désormais au second théorème. On suppose donc maintenant que les f_n sont des fonctions croissantes et que f est continue. Comme $[a,b]$ est un compact, d'après le théorème de Heine, on sait que f est uniformément continue sur $[a,b]$, et ainsi

$$\forall \epsilon > 0, \quad \exists \eta > 0, \quad \forall (x,y) \in [a,b]^2, \quad |x-y| \leq \eta \Longrightarrow |f(x)-f(y)| \leq \epsilon.$$

Ainsi, il existe une subdivision $a = x_0 < x_1 < ... < x_m = b$ de $[a,b]$ telle que pour tout i dans $[\![0, m-1]\!]$, $|f(a_{i+1}) - f(a_i)| \leq \epsilon$, et donc par croissance de f (comme limite simple de fonctions croissantes), on a

$$0 \leq f(a_{i+1}) - f(a_i) \leq \epsilon. \tag{12}$$

Soit x dans $[a,b]$, et soit i dans $[\![0, m-1]\!]$ tel que $a_i \leq x \leq a_{i+1}$. Grâce à la monotonie des f_n et de f, et grâce à (12), on obtient pour tout n dans $\mathbb{N}$:

$$\begin{cases} f_n(x) - f(x) \leq f_n(a_{i+1}) - f(a_i) \leq f_n(a_{i+1}) - f(a_{i+1}) + \epsilon, \\ f_n(x) - f(x) \geq f_n(a_i) - f(a_{i+1}) \geq f_n(a_i) - f(a_i) - \epsilon. \end{cases}$$

Or par convergence simple des f_n vers f, on sait qu'il existe N dans $\mathbb{N}$ tel que

$$\forall n \geq N, \quad \forall i \in [\![0, m]\!], \quad |f_n(a_i) - f(a_i)| \leq \epsilon.$$

Ainsi, on obtient $|f_n(x) - f(x)| \leq 2\epsilon$. On a obtenu ce résultat pour un x quelconque, et donc on a bien

$$\sup_{x \in [a,b]} |f_n(x) - f(x)| \leq 2\epsilon.$$

Ainsi, on obtient bien la convergence uniforme des $(f_n)_{n \in \mathbb{N}}$. □

Nous allons désormais pouvoir énoncer et démontrer le théorème de Glivenko Cantelli.

Théorème 3.2.2. ***Théorème de Glivenko-Cantelli*** *Soit $(X_i)_{1 \leq i \leq n}$ un échantillon de variables indépendantes et identiquement distribuées, de fonction de répartition commune F. Alors presque sûrement la fonction de répartition empirique converge uniformément vers F, autrement dit*

$$\mathbb{P}\Big(\lim_{n \to \infty} \sup_{t \in \mathbb{R}} \Big| \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{X_i \leq t\}} - F(t) \Big| = 0 \Big) = 1.$$

Démonstration. Sans perdre en généralité, nous allons considérer une suite $(U_i)_{1 \leq i \leq n}$ de variables idépendantes et identiquement distribuées suivant toutes une loi uniforme sur $[0,1]$. En effet, si l'on prend une autre suite $(X_i)_{1 \leq i \leq n}$ de variables idépendantes et identiquement distribuées de fonction de répartition F. Alors pour i dans $[\![1, n]\!]$, on sait que $X_i \sim F^{-1}(U_i)$ (où F^{-1} désigne l'inverse généralisée de F donnée par $F^{-1}(p) = \inf\{x \in \mathbb{R} | F(x) \geq p\}$ pour $p \in]0,1]$) et donc

$$\begin{aligned} \sup_{t \in \mathbb{R}} \Big| \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{X_i \leq t\}} - F(t) \Big| &\sim \sup_{t \in \mathbb{R}} \Big| \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{F^{-1}(U_i) \leq t\}} - F(t) \Big| = \sup_{s \in F(\mathbb{R})} \Big| \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{U_i \leq s\}} - s \Big| \\ &\leq \sup_{s \in [0,1]} \Big| \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{U_i \leq s\}} - s \Big|. \end{aligned}$$

De plus, en se ramenant à une loi uniforme, on obtient une fonction de répartition continue, à support compact, ce qui nous permet de nous rapprocher du cadre du second théorème de Dini.

Parmi les hypothèses nécessaires à l'application du second théorème de Dini, on a donc déjà le fait que la suite de fonctions de répartition empiriques est une suite de fonctions croissantes, de $[0,1]$ dans $\mathbb{R}$ et que la fonction

limite est une fonction continue. Ainsi, pour montrer que presque sûrement les fonctions de répartition empiriques convergent uniformément vers la fonction identité, il faut montrer que presque sûrement, elles convergent simplement vers cette même fonction. Autrement dit, pour montrer le résultat souhaité, il ne reste plus qu'à prouver qu'il existe un ensemble A de mesure pleine tel que :

$$\forall s \in [0,1], \quad \forall \omega \in A, \quad \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{U_i(\omega)\leq s\}} \underset{n\to\infty}{\longrightarrow} s.$$

Or on sait d'après la loi des grands nombres :

$$\forall s \in [0,1], \quad \exists A_s \subset \Omega, \quad \text{tel que} \quad \mathbb{P}(A_s) = 1 \quad \text{et} \quad \forall \omega \in A_s, \quad \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{U_i(\omega)\leq s\}} \underset{n\to\infty}{\longrightarrow} s.$$

On pose alors

$$A = \bigcap_{s\in[0,1]\cap\mathbb{Q}} A_s.$$

Alors A est de mesure pleine comme intersection dénombrable d'ensembles de mesures pleines, et

$$\forall s \in [0,1]\cap\mathbb{Q}, \quad \forall \omega \in A, \quad \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{U_i(\omega)\leq s\}} \underset{n\to\infty}{\longrightarrow} s.$$

Soit s quelconque dans $[0,1]$. Alors il existe $(x_n)_{n\in\mathbb{N}}$ croissante et $(y_n)_{n\in\mathbb{N}}$ décroissante, toutes les deux dans $[0,1]\cap\mathbb{Q}$ et convergeant toutes les deux vers s. Alors pour ω dans A, pour n dans $\mathbb{N}$ et k dans $\mathbb{N}$, on a :

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{U_i(\omega)\leq x_k\}} \leq \quad \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{U_i(\omega)\leq s\}} \leq \quad \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{U_i(\omega)\leq y_k\}}.$$

En faisant tendre n vers l'infini dans cette inégalité, on obtient

$$x_k \leq \lim\inf_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{U_i(\omega)\leq s\}} \leq \lim\sup_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{U_i(\omega)\leq s\}} \leq y_k.$$

Et ainsi, en faisant tendre k vers l'infini, on obtient par encadrement,

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{U_i(\omega)\leq s\}} \underset{n\to\infty}{\longrightarrow} s.$$

Et on obtient bien la conclusion souhaitée.

□

Une fois ces outils démontrés, nous allons pouvoir prouver la consistance de l'estimateur de Kaplan Meier.

3.3 Consistance de l'estimateur

Pour montrer la consistance de l'estimateur, nous allons commencer par prouver un lemme utile pour la suite.

Proposition 3.3.1. *En reprenant les notations précédentes, on a pour tout $\tau < \tau_H$:*

$$\forall t \in [0,\tau], \quad |-\log(S_n(t)) - \Lambda_n(t)| \leq O\Big(\frac{1}{n}\Big) \quad \textit{presque sûrement.}$$

Démonstration. Soit $\tau < \tau_H$ et t dans $[0, \tau]$. Montrons

$$0 < -\log(S_n(t)) - \Lambda_n(t) < \frac{1}{n}\int_0^t \frac{\mathrm{d}H_n^u(s)}{(1-H_n(s))(1-H_n(s^-))} \leq \frac{1}{n} \times \frac{H_n(t)}{(1-H_n(t))^2}.$$

On a d'une part

$$\begin{aligned}\frac{1}{n}\int_0^t \frac{\mathrm{d}H_n^u(s)}{(1-H_n(s))(1-H_n(s^-))} &= \frac{1}{n^2}\sum_{i=1}^n \int_0^t \frac{\mathrm{d}\mathbb{1}_{\{T_{(i)}\leq s, \delta_{(i)}=1\}}}{(1-H_n(s))(1-H_n(s^-))} \\ &= \sum_{i=1}^n \frac{\delta_{(i)}\mathbb{1}_{\{T_{(i)}\leq t\}}}{n^2(1-H_n(T_{(i)}))(1-H_n(T_{(i)}^-))}.\end{aligned}$$

D'autre part, on a d'après (10) et (11)

$$\begin{aligned}-\log(S_n(t)) - \Lambda_n(t) &= -\sum_{T_{(i)}\leq t} \log(1-\Delta\Lambda_n(T_{(i)}) - \frac{1}{n}\sum_{T_{(i)}\leq t}\frac{\delta_{(i)}}{1-H_n(T_{(i)}^-)} \\ &= \sum_{i=1}^n \mathbb{1}_{\{T_{(i)}\leq t\}}\Big(-\log\big(1-\frac{\delta_{(i)}}{n(1-H_n(T_{(i)}^-))}\big) - \frac{\delta_{(i)}}{n(1-H_n(T_{(i)}^-))}\Big) \\ &= \sum_{i=1}^n \mathbb{1}_{\{T_{(i)}\leq t\}}\delta_{(i)}\Big(-\log\big(1-\frac{1}{n(1-H_n(T_{(i)}^-))}\big) - \frac{1}{n(1-H_n(T_{(i)}^-))}\Big).\end{aligned}$$

Or

$$\forall x \in \mathbb{R}_+^*, \quad 0 < -\log(1-\frac{1}{x+1}) - \frac{1}{x+1} < \frac{1}{x(x+1)}. \tag{13}$$

Donc en utilisant ici cette inégalité, avec $x+1 = n(1-H_n(T_{(i)}^-))$, on obtient

$$0 < -\log(S_n(t)) - \Lambda_n(t)) < \sum_{i=1}^n \frac{\mathbb{1}_{\{T_{(i)}\leq t\}}\delta_{(i)}}{n(1-H_n(T_{(i)}^-))(n(1-H_n(T_{(i)}^-))-1)}.$$

Or

$$n(1-H_n(T_{(i)}^-))-1) = n - \sum_{j=1}^{i-1} 1 - 1 = n - \sum_{j=1}^{i} 1 = n(1-H_n(T_{(i)})).$$

Et donc

$$0 < -\log(S_n(t)) - \Lambda_n(t) < \frac{1}{n}\int_0^t \frac{\mathrm{d}H_n^u(s)}{(1-H_n(s))(1-H_n(s^-))}.$$

Enfin

$$\begin{aligned}\frac{1}{n}\int_0^t \frac{\mathrm{d}H_n^u(s)}{(1-H_n(s))(1-H_n(s^-))} &= \sum_{i=1}^n \frac{\delta_{(i)}\mathbb{1}_{\{T_{(i)}\leq t\}}}{n^2(1-H_n(T_{(i)}))(1-H_n(T_{(i)}^-))} \\ &\leq \frac{1}{n(1-H_n(t))^2} \times \Big(\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{T_{(i)}\leq t\}}\delta_{(i)}\Big) \\ &\leq \frac{H_n(t)}{n(1-H_n(t))^2}.\end{aligned}$$

Comme $H_n(t)$ converge presque sûrement vers $H(t)$, et que $t \leq \tau < \tau_H$, on sait que $\frac{H_n(t)}{(1-H_n(t))^2}$ est presque sûrement borné. Alors on obtient bien

$$|-\log(S_n(t)) - \Lambda_n(t)| \leq O\Big(\frac{1}{n}\Big) \quad \text{presque sûrement.}$$

□

Théorème 3.3.2. *Si F est continue, alors pour tout $\tau < \tau_H$,*

$$\forall t \in [0,\tau], \quad S_n(t) \underset{n\to\infty}{\longrightarrow} S(t) \quad \textit{presque sûrement.}$$

Démonstration. Soit $\tau < \tau_H$ et t dans $[0,\tau]$.

Montrons tout d'abord que $\Lambda_n(t) \longrightarrow \Lambda(t)$ presque sûrement. On a obtenu en (10)

$$\Lambda_n(t) = \frac{1}{n}\sum_{i=1}^{n} \frac{\delta_i \mathbb{1}_{\{T_i \le t\}}}{1 - H_n(T_i^-)}.$$

On peut alors découper différemment cette somme afin de faire apparaître une somme de variables indépendantes :

$$\Lambda_n(t) = \frac{1}{n}\sum_{i=1}^{n} \frac{\delta_i \mathbb{1}_{\{T_i \le t\}}}{1 - H(T_i^-)} + \frac{1}{n}\sum_{i=1}^{n} \delta_i \mathbb{1}_{\{T_i \le t\}} \Big(\frac{1}{1 - H_n(T_i^-)} - \frac{1}{1 - H(T_i^-)} \Big).$$

D'une part, on a

$$\begin{aligned} \Big| \frac{1}{n}\sum_{i=1}^{n} \delta_i \mathbb{1}_{\{T_i \le t\}} \Big(\frac{1}{1 - H_n(T_i^-)} - \frac{1}{1 - H(T_i^-)} \Big) \Big| &= \Big| \frac{1}{n}\sum_{i=1}^{n} \delta_i \mathbb{1}_{\{T_i \le t\}} \Big(\frac{H_n(T_i^-) - H(T_i^-)}{(1 - H_n(T_i^-))(1 - H(T_i^-))} \Big) \Big| \\ &\le \frac{||H - H_n||_\infty}{(1 - H_n(t))(1 - H(t))} \times \frac{1}{n}\sum_{i=1}^{n} \delta_i \mathbb{1}_{\{T_i \le t\}}. \end{aligned}$$

Or $\frac{1}{n}\sum_{i=1}^{n} \delta_i \mathbb{1}_{\{T_i \le t\}}$ est majoré par 1 et $\frac{1}{(1-H_n(t))(1-H(t))}$ est borné presque sûrement (car $H_n(t)$ converge presque sûrement vers $H(t)$ et car $t \le \tau < \tau_H$), et d'après le théorème de Glivenko-Cantelli, $||H - H_n||_\infty \longrightarrow 0$ presque sûrement. Ainsi, on obtient

$$\frac{1}{n}\sum_{i=1}^{n} \delta_i \mathbb{1}_{\{T_i \le t\}} \Big(\frac{1}{1 - H_n(T_i^-)} - \frac{1}{1 - H(T_i^-)} \Big) \underset{n\to\infty}{\longrightarrow} 0 \quad \text{presque sûrement.}$$

D'autre part, les variables aléatoires $\{\frac{\delta_i \mathbb{1}_{\{T_i \le t\}}}{1 - H(T_i^-)}\}_{1 \le i \le n}$ sont indépendantes et identiquement distribuées. Ainsi, d'après la loi des grands nombres,

$$\frac{1}{n}\sum_{i=1}^{n} \frac{\delta_i \mathbb{1}_{\{T_i \le t\}}}{1 - H(T_i^-)} \underset{n\to\infty}{\longrightarrow} \mathbb{E}\Big(\frac{\delta \mathbb{1}_{\{T \le t\}}}{1 - H(T^-)} \Big) \quad \text{presque sûrement.}$$

Et donc étant donné que $\mathbb{E}\Big(\frac{\delta \mathbb{1}_{\{T \le t\}}}{1 - H(T^-)} \Big) = \Lambda(t)$, on obtient $\Lambda_n(t) \underset{n\to\infty}{\longrightarrow} \Lambda(t)$ presque sûrement.

Comme F est supposée continue, on voit dans la preuve de la proposition 3.1.4 que le lien entre S et Λ est simplifié. On a alors en effet $\log(S) = -\Lambda$. Ainsi

$$\begin{aligned} |\log(S(t)) - \log(S_n(t))| &\le |\log(S(t)) + \Lambda_n(t))| + |-\Lambda_n(t) - \log(S_n(t))| \\ &\le |\Lambda_n(t)) - \Lambda(t)| + |-\Lambda_n(t) - \log(S_n(t))|. \end{aligned}$$

D'après ce qui précède et d'après la proposition 3.3.1, on sait que $|\Lambda_n(t)) - \Lambda(t)|$ et $|-\Lambda_n(t) - \log(S_n(t))|$ convergent tous les deux presque sûrement vers 0. Ainsi, on en déduit

$$\log(S_n(t)) \underset{n\to\infty}{\longrightarrow} \log(S(t)) \quad \text{presque sûrement.}$$

Et donc comme la fonction exponentielle est une fonction continue, on obtient bien

$$S_n(t) \underset{n\to\infty}{\longrightarrow} S(t) \quad \text{presque sûrement.}$$

□

3.4 Simulations

Pour expérimenter la bonne convergence de cet estimateur, nous allons maintenant observer quelques résultats obtenus sur des échantillons simulés.

On effectue par exemple un test à l'aide d'un échantillon simulé de la façon suivante : on crée un échantillon de taille $n = 300$, avec Y suivant une loi exponentielle de paramètre 5, C suivant une loi exponentielle de paramètre 2, les deux variables étant indépendantes. On obtient alors la courbe empirique ci dessous grâce à l'estimateur de Kaplan-Meier. On voit bien que cette courbe est fidèle à la courbe théorique.

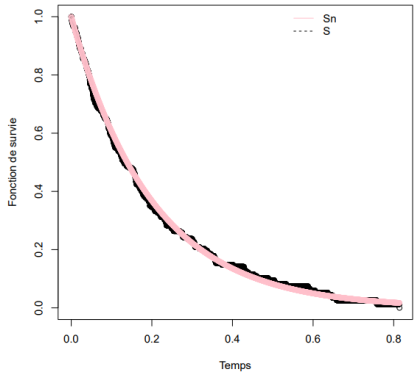

FIGURE 2 – Résultat obtenu après test de l'estimateur de Kaplan Meier sur un échantillon simulé

De plus, si l'on moyenne les résultats de 100 échantillons de taille 200 (ayant les mêmes paramètres que précédemment), on voit que la courbe empirique suit parfaitement la courbe théorique.

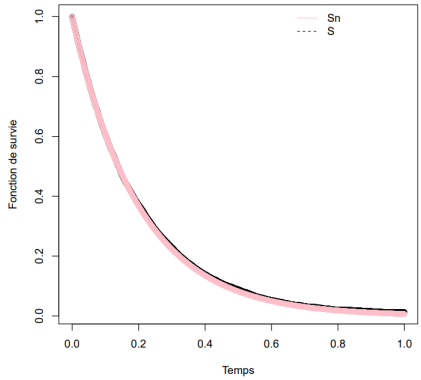

FIGURE 3 – Résultat obtenu après moyenne de tests de l'estimateur de Kaplan Meier sur plusieurs échantillons simulés

En revanche, si l'on simule des échantillons où l'on prend en compte l'influence d'une certaine covariable, l'estimateur ne fonctionne plus. Ici par exemple, les variables Y et C sont toujours indépendantes, mais chaque couple est influencé par une certaine covariable x dans $[0, 1]$ dans le sens où Y suit une loi exponentielle de paramètre $5x$ et C une loi exponentielle de paramètre $2x$. Alors le résultat donné par l'estimateur ne suit plus du tout le modèle théorique attendu lorsque l'on se place en un x d'observation quelconque.

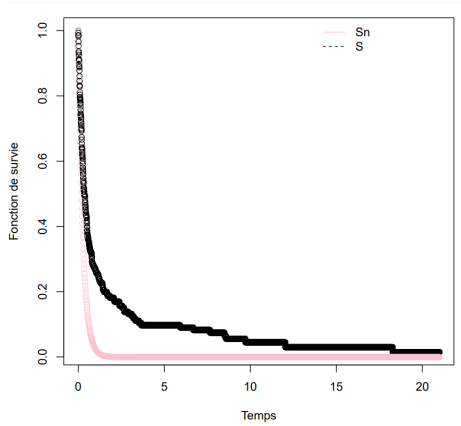

FIGURE 4 – Résultat obtenu après test de l'estimateur de Kaplan Meier sur un échantillon simulé, influencé par des covariables

C'est pour traiter ce genre de cas que nous introduisons l'estimateur suivant.

4 Estimateur de Beran

Nous allons désormais chercher un estimateur pour le cas où l'on possède ce même type de données censurées à droite, mais cette fois accompagnées de covariables. Dans le cadre médical, cela peut représenter l'âge, le sexe, la taille ou encore le poids des patients par exemple.

Dans toute cette section, nous supposerons toujours que les variables d'intérêt Y et de censure C sont indépedantes. Cependant, la variable observée est désormais un triplet (T, δ, X) où l'on a toujours $T = \min(Y, C)$ et $\delta = \mathbb{1}_{\{Y \leq C\}}$ et où $X \in \mathbb{R}_+^d$ est un vecteur aléatoire représentant les d covariables observées.

On suppose alors que l'on dispose d'un échantillon aléatoire $\{(T_i, \delta_i, X_i)\}_{1 \leq i \leq n}$, où les $\{(T_i, \delta_i, X_i)\}$ sont indépendants et identiquement distribués selon la loi de (T, δ, X) . Lorsque les $\{T_i\}_{1 \leq i \leq n}$ sont rangés par ordre croissant, on les notera $\{T_{(i)}\}_{1 \leq i \leq n}$, et on notera alors de même $\{(T_{(i)}, \delta_{(i)}, X_{(i)})\}_{1 \leq i \leq n}$.

4.1 Lois conditionnelles

Dans ce contexte, nous allons souhaiter reprendre les méthodes précédentes en conditionnant par un certain x dans $\mathbb{R}_+^d$. Cependant, si, comme nous le faisons dans la suite, nous supposons que la variable X est une variable à densité, alors il est impossible de se servir de la définition basique de probabilité conditionnelle, au risque de diviser par 0. Il est donc nécessaire de définir de façon plus générale cette loi conditionnelle.

Définition 4.1.1. *Soit $(E, \mathscr{E})$ et $(F, \mathscr{F})$ deux espaces probabilisés. On dit qu'une application $\nu : E \times \mathscr{F} \to [0, 1]$ est un noyau de transition si*

1. *Pour tout x dans E, $\nu(x, .)$ est une probabilité sur $(F, \mathscr{F})$.*
2. *Pour tout B dans $\mathscr{F}$, $\nu(., B)$ est $\mathscr{E}$-mesurable.*

On peut, à partir de cette notion, donner une nouvelle définition de loi conditionnelle.

Définition 4.1.2. *Soit X et Y deux variables aléatoires respectivement à valeurs dans $\mathbb{R}^d$ et $\mathbb{R}^k$. On appelle loi conditionnelle de Y sachant X un noyau sur $(\mathbb{R}^d \times \mathscr{B}(\mathbb{R}^k))$ tel que pour tout $A \subset E$ et $B \in \mathscr{B}(\mathbb{R}^k)$,*

$$\mathbb{P}(X \in A, Y \in B) = \int_{x \in A} \nu(x, B) \mathbb{P}_X(\mathrm{dx}).$$

On notera alors, pour x dans $\mathbb{R}^d$ et B dans $\mathscr{B}(\mathbb{R}^k)$:

$$\nu(x, B) = \mathbb{P}(Y \in B | X = x). \tag{14}$$

Le théorème suivant, que nous admettrons, nous garantit l'existence d'une telle loi conditionnelle dans le cadre précécent.

Théorème 4.1.3. ***Théorèmes de Jirina*** *Soit X et Y deux variables aléatoires respectivement à valeurs dans $\mathbb{R}^d$ et $\mathbb{R}^k$. Alors il existe une loi conditionnelle de Y sachant X.*

Nous allons désormais, à l'aide de méthodes similaires aux méthodes précédentes, essayer d'approximer cette probabilité conditionnelle.

4.2 Construction de l'estimateur

On reprend la même démarche que lors de la construction de l'estimateur de Kaplan-Meier, en conditionnant à chaque étape par rapport à $x \in \mathbb{R}_+^d$.

On notera pour s dans $\mathbb{R}_+$, $F_x(s) = \mathbb{P}(Y \leq s | X = x)$, $S_x = 1 - F_x$, $H_x(s) = \mathbb{P}(T \leq s | X = x)$ et $H_x^u(s) = \mathbb{P}(T \leq s, \delta = 1 | X = x)$. Enfin, on notera $\tau_{F_x} = \inf\{t \in \mathbb{R}_+ | F_x(t) = 1\}$ et $\tau_{H_x} = \inf\{t \in \mathbb{R}_+ | H_x(t) = 1\}$.

Définition 4.2.1. *On définit le risque cumulé de Y en $t < \tau_{F_x}$, conditionnellement au fait que $X = x$ de la façon suivante :*

$$\Lambda_x(t) = \int_0^t \frac{\mathrm{d}F_x(u)}{S_x(u^-)}.$$

On prouve exactement de la même façon que précédemment que les formules suivantes sont vérifiées : pour $t < \tau_{H_x}$, on a

$$\log(S_x(t)) = -\Lambda_{x,c}(t) + \sum_{s \leq t} \log(1 - \Delta\Lambda_x(s)), \tag{15}$$

et

$$\Lambda_x(t) = \int_0^t \frac{\mathrm{d}H_x^u(s)}{1 - H_x(s^-)}.$$

Cependant, nous devons désormais tenir compte des covariables : pour estimer H_x et H_x^u, il parait intuitif de tenir plus compte des données dont les covariables sont "proches" de x. C'est donc pour cela que l'on introduit des poids.

Définition 4.2.2. *On définit les poids $\{w_i(x)\}_{1 \leq i \leq n}$ par rapport à x de la manière suivante :*

$$\forall i \in [\![1, n]\!], \quad w_i(x) = \frac{\frac{1}{h_n^d} K(\frac{x - X_i}{h_n})}{\sum_{j=1}^n \frac{1}{h_n^d} K(\frac{x - X_j}{h_n})},$$

où h_n est une constante dépendant de n, telle que $lim_{n \to \infty}(h_n) = 0$ et où K, appelé noyau, est une densité sur $\mathbb{R}_+^d$.

On peut à partir de ces poids essayer d'estimer de nouveaux les fonctions composant Λ_x.

Définition 4.2.3. *On définit pour approximer les fonctions H_x^u et $1 - H_x$ les fonctions empiriques données par :*

$$\forall s \in \mathbb{R}_+, \quad 1 - H_{n,x}(s^-) = \sum_{j=1}^n w_i(x) \mathbb{1}_{\{T_j \geq s\}} \quad et \quad H_{n,x}^u(s) = \sum_{j=1}^n w_i(x) \mathbb{1}_{\{T_j \leq s, \delta_j = 1\}}.$$

On définit alors un estimateur de Λ_x donné par :

$$\forall t \in \mathbb{R}_+, \quad \Lambda_{n,x}(t) = \int_0^t \frac{\mathrm{d}H_{n,x}^u(s)}{1 - H_{n,x}(s^-)}.$$

Remarque 4.2.4. *On remarque que cette fois, contrairement au cas sans covariable, la convergence de $H_{n,x}$ et $H^u_{n,x}$ vers H_x et H^u_x n'est pas assurée par la loi des grands nombres à cause de la dépendance en n portée par h_n dans l'expression des poids.*

On peut de nouveau calculer l'intégrale obtenue. Soit t dans $\mathbb{R}_+$. On a alors

$$\Lambda_{n,x}(t) = \int_0^t \frac{\mathrm{d}H^u_{n,x}(s)}{1 - H_{n,x}(s^-)} = \sum_{i=1}^n w_i(x) \int_0^t \frac{\mathrm{d}\mathbb{1}_{\{T_i \leq s, \delta_i = 1\}}}{1 - H_{n,x}(s^-)} = \sum_{i=1}^n \frac{w_i(x)\delta_i}{1 - H_{n,x}(T_i^-)} \mathbb{1}_{\{T_i \leq t\}}.$$

Et donc

$$\Lambda_{n,x}(t) = \sum_{T_{(i)} \leq t} \frac{w_{(i)}(x)\delta_{(i)}}{1 - H_{n,x}(T_{(i)}^-)}. \tag{16}$$

Or pour tout i dans $[\![1, n]\!]$, on a

$$1 - H_{n,x}(T_{(i)}^-) = \sum_{j=i}^n w_{(j)}(x) = 1 - \sum_{j=1}^{i-1} w_{(j)}(x).$$

Ainsi,

$$\Lambda_{n,x} = \sum_{T_{(i)} \leq t} \frac{w_{(i)}(x)\delta_{(i)}}{1 - \sum_{j=1}^{i-1} w_{(j)}(x)}.$$

En réinjectant de nouveau l'expression obtenue dans la formule (15), et en remarquant que $\Lambda_{n,x}$ est une fonction en escalier (et donc que sa partie continue est nulle), on obtient un estimateur de la fonction de survie S_x :

$$S_{n,x}(t) = \exp\Big(\sum_{T_{(i)} \leq t} \log(1 - \Delta\Lambda_{n,x}(T_{(i)})\Big) = \prod_{T_{(i)} \leq t} \big(1 - \Delta\Lambda_{n,x}(T_{(i)})\big) = \prod_{T_{(i)} \leq t} \Big(1 - \frac{\delta_{(i)} w_{(i)}(x)}{1 - \sum_{j=1}^{i-1} w_{(j)}(x)}\Big). \tag{17}$$

Définition 4.2.5. *On définit l'estimateur de Beran comme la fonction en escalier suivante :*

$$\forall t \in \mathbb{R}_+, \quad S_{n,x}(t) = \prod_{T_{(i)} \leq t} \Big(1 - \frac{w_{(i)}(x)}{1 - \sum_{j=1}^{i-1} w_{(j)}(x)}\Big)^{\delta_{(i)}}.$$

Une fois cet estimateur construit, il va de nouveau s'agir de prouver sa consistance.

4.3 Consistance de l'estimateur

Passons tout d'abord en revue quelques lemmes utiles à la preuve de la consistance. Nous allons par la suite supposer que X est une variable aléatoire à densité, de fonction de densité f.

Tout d'abord, le théorème suivant sert d'équivalent au théorème de Glivenko-Cantelli dans notre contexte.

Théorème 4.3.1. *On suppose*

$$\lim_{n\to\infty} h_n = 0 \quad \textit{et} \quad \lim_{n\to\infty} n h_n^d = +\infty.$$

On suppose de plus

$$\forall \rho > 0, \quad \sum_{n \geq 1} \exp(-\rho n h_n^d) < +\infty.$$

Alors

$$\mathbb{P}\Big(\lim_{n\to\infty} \sup_{t \in \mathbb{R}} \Big| H_{n,x}(t) - H_x(t) \Big| = 0 \Big) = 1.$$

Nous admettrons ce théorème par manque de notions et d'outils. Une preuve est cependant disponible dans la référence [5].

Proposition 4.3.2. *On définit f_n comme la fonction de densité empirique suivante :*

$$f_n(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h_n^d} K\Big(\frac{x - X_i}{h_n}\Big).$$

On suppose que K est de carré intégrable, et que

$$\lim_{n\to\infty} nh_n^d = +\infty.$$

Alors

$$f_n(x) \underset{n\to\infty}{\longrightarrow} f(x) \quad \textit{en probabilité.}$$

Démonstration. Tout d'abord, on obtient en effectuant un changement de variable :

$$\mathbb{E}\Big(\frac{1}{h_n^d} K\Big(\frac{x - X}{h_n}\Big)\Big) = \int_{\mathbb{R}_+^d} \frac{1}{h_n^d} K\Big(\frac{x-u}{h_n}\Big) f(u)\mathrm{d}u = \int_{\mathbb{R}_+^d} K(v) f(x + vh_n)\mathrm{d}v.$$

Et donc, étant donnée que $lim_{n\to\infty} h_n = 0$, et que K est une densité sur $\mathbb{R}_+^d$ on a :

$$\mathbb{E}\Big(\frac{1}{h_n^d} K\Big(\frac{x - X}{h_n}\Big)\Big) \underset{n\to\infty}{\longrightarrow} f(x) \int_{\mathbb{R}_+^d} K(v)\mathrm{d}v = f(x).$$

Soit $\epsilon > 0$. D'après ce qui précède, on sait qu'il existe $N \in \mathbb{N}$ tel que

$$\forall n \geq N, \quad |\mathbb{E}\Big(\frac{1}{h_n^d} K\Big(\frac{x - X}{h_n}\Big)\Big) - f(x)| < \frac{\epsilon}{2}.$$

Ainsi

$$\begin{aligned}\mathbb{P}(|f_n(x) - f(x)| > \epsilon) &\leq \mathbb{P}(|f_n(x) - \mathbb{E}\Big(\frac{1}{h_n^d} K\Big(\frac{x - X}{h_n}\Big)\Big)| + |\mathbb{E}\Big(\frac{1}{h_n^d} K\Big(\frac{x - X}{h_n}\Big)\Big) - f(x)| > \epsilon) \\ &\leq \mathbb{P}(|f_n(x) - \mathbb{E}\Big(\frac{1}{h_n^d} K\Big(\frac{x - X}{h_n}\Big)\Big)| > \frac{\epsilon}{2}).\end{aligned}$$

Alors d'après l'inégalité de Tchebychev et grâce au fait que les $(X_i)_{1\leq i\leq n}$ sont indépendantes et identiquement distribuées, on obtient

$$\mathbb{P}(|f_n(x) - f(x)| > \epsilon) \leq \frac{\mathrm{Var}\Big(\frac{1}{n}\sum_{i=1}^n \frac{1}{h_n^d} K\big(\frac{x-X_j}{h_n}\big)\Big)}{(\frac{\epsilon}{2})^2} = \frac{4\sum_{i=1}^n \mathrm{Var}\Big(K\big(\frac{x-X_j}{h_n}\big)\Big)}{(nh_n^d\epsilon)^2} = \frac{4\mathrm{Var}\Big(K\big(\frac{x-X}{h_n}\big)\Big)}{n(h_n^d)^2\epsilon^2}.$$

Or

$$\begin{aligned}\mathrm{Var}\Big(K\big(\frac{x-X}{h_n}\big)\Big) &= \mathbb{E}\Big(\big|K\big(\frac{x-X}{h_n}\big) - \mathbb{E}\big(K\big(\frac{x-X}{h_n}\big)\big)\big|^2\Big) \leq \mathbb{E}\Big(\big|K\big(\frac{x-X}{h_n}\big)\big|^2\Big) \\ &\leq \int_{\mathbb{R}_+^d} K\Big(\frac{x-u}{h_n}\Big)^2 f(u)\mathrm{d}u = \int_{\mathbb{R}_+^d} K(v)^2 f(x - vh_n) h_n^d \mathrm{d}v.\end{aligned}$$

Donc

$$\frac{1}{h_n^d}\mathrm{Var}\Big(K\big(\frac{x-X}{h_n}\big)\Big) = I := \int_{\mathbb{R}_+^d} K(v)^2 f(x - vh_n)\mathrm{d}v \underset{n\to\infty}{\longrightarrow} ||K||_2^2 f(x).$$

Ainsi

$$\mathbb{P}(|f_n(x) - f(x)| > \epsilon) \leq \frac{4I}{nh_n^d\epsilon^2} \underset{n\to\infty}{\longrightarrow} 0.$$

□

En fait, il est possible de montrer un théorème bien plus fort en supposant quelques hypothèses supplémentaires.

Théorème 4.3.3. *On suppose que f est uniformément continue, que K est bornée et intégrable, et qu'elle vérifie*

$$\int_0^\infty x^{d-1}L(x)\mathrm{dx} < \infty.$$

On suppose de plus que la suite $(h_n)_{n\in\mathbb{N}}$ vérifie

$$\lim_{n\to\infty} h_n = 0 \quad et \quad \lim_{n\to\infty} \frac{nh_n^{2d}}{\log(n)} = +\infty.$$

Alors

$$\mathbb{P}\Big(\lim_{n\to\infty} \sup_{x\in\mathbb{R}^d} \Big|f_n(x) - f(x)\Big| = 0\Big) = 1.$$

Nous admettrons de nouveau ce théorème qui nécessite trop d'outils sortant des limites de ce rapport. Cependant une preuve est disponible dans la référence [6].

Proposition 4.3.4. *On suppose que x est tel que $f(x) \neq 0$. En reprenant les notations précédentes, et en supposant toujours que K, f et $(h_n)_{n\in\mathbb{N}}$ vérifient les hypothèses des deux théorèmes précédents, on a pour tout $\tau < \tau_{H_x}$:*

$$\forall t \in [0,\tau], \quad |-\log(S_{n,x}(t)) - \Lambda_{n,x}(t)| \leq O\Big(\frac{1}{nh_n^d}\Big) \quad \textit{presque sûrement.}$$

Démonstration. Soit $\tau < \tau_H$ et t dans $[0,\tau]$. On a d'après (16) et (17) :

$$\begin{aligned}
-\log(S_{n,x}(t)) - \Lambda_{n,x}(t) &= -\sum_{T_{(i)}\leq t} \log(1-\Delta\Lambda_{n,x}(T_{(i)}) - \sum_{T_{(i)}\leq t} \frac{w_{(i)}(x)\delta_{(i)}}{1-H_{n,x}(T_{(i)}^-)} \\
&= \sum_{i=1}^n \mathbb{1}_{\{T_{(i)}\leq t\}}\Big(-\log\big(\frac{w_{(i)}(x)\delta_{(i)}}{1-H_{n,x}(T_{(i)}^-)}\big) - \frac{w_{(i)}(x)\delta_{(i)}}{1-H_{n,x}(T_{(i)}^-)}\Big) \\
&= \sum_{i=1}^n \mathbb{1}_{\{T_{(i)}\leq t\}}\delta_{(i)}\Big(-\log\big(\frac{w_{(i)}(x)}{1-H_{n,x}(T_{(i)}^-)}\big) - \frac{w_{(i)}(x)}{1-H_{n,x}(T_{(i)}^-)}\Big).
\end{aligned}$$

Donc en reprenant l'inégalité (13), cette fois ci en prenant $x+1 = \frac{1-H_{n,x}(T_{(i)}^-)}{w_{(i)}(x)}$, on obtient :

$$\begin{aligned}
0 < -\log(S_{n,x}(t)) - \Lambda_{n,x}(t) &< \sum_{i=1}^n \mathbb{1}_{\{T_{(i)}\leq t\}}\delta_{(i)} \frac{1}{\Big(\frac{1-H_{n,x}(T_{(i)}^-)}{w_{(i)}(x)}\Big)\Big(\frac{1-H_{n,x}(T_{(i)}^-)}{w_{(i)}(x)} - 1\Big)} \\
&< \sum_{i=1}^n \mathbb{1}_{\{T_{(i)}\leq t\}}\delta_{(i)} \frac{w_{(i)}(x)^2}{(1-H_{n,x}(T_{(i)}^-))(1-H_{n,x}(T_{(i)}^-) - w_{(i)}(x))} \\
&< \sum_{i=1}^n \mathbb{1}_{\{T_{(i)}\leq t\}}\delta_{(i)} \frac{w_{(i)}(x)^2}{(1-H_{n,x}(T_{(i)}^-))(1-H_{n,x}(T_{(i)}))} \\
&< \frac{1}{(1-H_{n,x}(t))^2} \sum_{i=1}^n \mathbb{1}_{\{T_{(i)}\leq t\}}\delta_{(i)} w_{(i)}(x)^2.
\end{aligned}$$

Or pour tout i dans $[\![1,n]\!]$, on a

$$w_i(x) = \frac{\frac{1}{h_n^d}K(\frac{x-X_i}{h_n})}{\sum_{j=1}^n \frac{1}{h_n^d}K(\frac{x-X_j}{h_n})} = \frac{K(\frac{x-X_i}{h_n})}{nh_n^d f_n(x)} \leq \frac{||K||_\infty}{nh_n^d f_n(x)}.$$

Et donc

$$0 < -\log(S_{n,x}(t)) - \Lambda_{n,x}(t) < \frac{||K||_\infty H_{n,x}(t)}{(1-H_{n,x}(t))^2 f_n(x)} \times \frac{1}{nh_n^d}.$$

Or K est borné, $\frac{H_{n,x}(t)}{(1-H_{n,x}(t))^2}$ est borné presque sûrement (car $H_{n,x}(t)$ converge presque sûrement vers $H_x(t)$ et car $t \leq \tau < \tau_{H_x}$), et $\frac{1}{f_n(x)}$ est borné presque sûrement car $f_n(x)$ converge presque sûrement vers $f(x)$ qui est non nul. Ainsi, on obtient bien

$$| -\log(S_{n,x}(t)) - \Lambda_{n,x}(t)| \leq O\Big(\frac{1}{nh_n^d}\Big) \quad \text{presque sûrement.}$$

□

Théorème 4.3.5. *On suppose de nouveau que x est dans le support de f, et on effectue de nouveau toutes les hypothèses effectuées au cours des théorèmes précédents sur f, K et sur la suite $(h_n)_{n\in\mathbb{N}}$. On suppose de nouveau que K est de carré intégrable. Enfin, on suppose que F_x est une fonction continue. Alors pour tout $\tau < \tau_{H_x}$,*

$$\forall t \in [0,\tau], \quad S_{n,x}(t) \underset{n\to\infty}{\longrightarrow} S_x(t) \quad \textit{en probabilité.}$$

Démonstration. Soit $\tau < \tau_{H_x}$ et t dans $[0,\tau]$.

Montrons tout d'abord que $\Lambda_{n,x}(t) \longrightarrow \Lambda_x(t)$. On a obtenu en (16)

$$\Lambda_{n,x}(t) = \sum_{i=1}^n \frac{w_i(x)\delta_i \mathbb{1}_{\{T_i\leq t\}}}{1-H_{n,x}(T_i^-)}.$$

Ainsi, en découpant la somme on obtient :

$$\Lambda_{n,x}(t) = \sum_{i=1}^n \frac{w_i(x)\delta_i \mathbb{1}_{\{T_i\leq t\}}}{1-H_x(T_i^-)} + \sum_{i=1}^n w_i(x)\delta_i \mathbb{1}_{\{T_i\leq t\}}\Big(\frac{1}{1-H_{n,x}(T_i^-)} - \frac{1}{1-H_x(T_i^-)}\Big). \tag{18}$$

Et d'une part,

$$\begin{aligned}\Big|\sum_{i=1}^n w_i(x)\delta_i \mathbb{1}_{\{T_i\leq t\}}\Big(\frac{1}{1-H_{n,x}(T_i^-)} - \frac{1}{1-H_x(T_i^-)}\Big)\Big| &= \Big|\sum_{i=1}^n w_i(x)\delta_i \mathbb{1}_{\{T_i\leq t\}} \frac{H_{n,x}(T_i^-) - H_x(T_i^-)}{(1-H_{n,x}(T_i^-))(1-H_x(T_i^-))}\Big| \\ &\leq \frac{||H_{n,x}-H_x||_\infty}{(1-H_{n,x}(t))(1-H_x(t))} \times \sum_{i=1}^n w_i(x)\delta_i \mathbb{1}_{\{T_i\leq t\}}.\end{aligned}$$

Or $\frac{1}{(1-H_{n,x}(t))(1-H_x(t))}$ est borné (car $H_{n,x}(t)$ converge presque sûrement vers $H_x(t)$ et car $t \leq \tau < \tau_{H_x}$), et $\sum_{i=1}^n w_i(x)\delta_i \mathbb{1}_{\{T_i\leq t\}}$ est borné par 1. Ainsi, comme $||H_{n,x}-H_x||_\infty$ tend vers 0 presque sûrement d'après le théorème 4.3.1, on obtient :

$$\sum_{i=1}^n w_i(x)\delta_i \mathbb{1}_{\{T_i\leq t\}}\Big(\frac{1}{1-H_{n,x}(T_i^-)} - \frac{1}{1-H_x(T_i^-)}\Big) \underset{n\to\infty}{\longrightarrow} 0 \quad \text{presque sûrement.} \tag{19}$$

D'autre part, on a

$$\sum_{i=1}^n \frac{w_i(x)\delta_i \mathbb{1}_{\{T_i\leq t\}}}{1-H_x(T_i^-)} = \frac{1}{f_n(x)} \times \frac{1}{n}\sum_{i=1}^n \frac{\delta_i \mathbb{1}_{\{T_i\leq t\}} K(\frac{x-X_i}{h})\frac{1}{h_n^d}}{1-H_x(T_i^-)}. \tag{20}$$

Or

$$\mathbb{E}\Big(\frac{\delta \mathbb{1}_{\{T\leq t\}} K(\frac{x-X}{h})\frac{1}{h_n^d}}{1-H_x(T^-)}\Big) = \frac{1}{h_n^d}\int_{\mathbb{R}_+}\int_{\mathbb{R}_+^d} \frac{K(\frac{x-u}{h_n})\mathbb{1}_{\{v\leq t\}}}{1-H_x(v^-)} f(u)\,\mathrm{du}\,\mathrm{dH_x^u(v)} = \frac{1}{h_n^d}\int_0^t \frac{\mathrm{dH_x^u(v)}}{1-H_x(v^-)}\int_{\mathbb{R}_+^d} K(\frac{x-u}{h_n}) f(u)\mathrm{du}.$$

On a vu précédemment grâce à un changement de variables :

$$\frac{1}{h_n^d}\int_{\mathbb{R}_+^d} K\big(\frac{x-u}{h_n}\big)f(u)\mathrm{du} \underset{n\to\infty}{\longrightarrow} f(x).$$

Ainsi,

$$\mathbb{E}\Big(\frac{\delta \mathbb{1}_{\{T\le t\}}K\big(\frac{x-X}{h}\big)\frac{1}{h_n^d}}{1-H_x(T^-)}\Big) \underset{n\to\infty}{\longrightarrow} \Lambda_x(t)f(x).$$

Soit $\epsilon>0$. D'après ce qui précède, on sait qu'il existe $N\in\mathbb{N}$ tel que

$$\forall n\ge N,\quad \Big|\mathbb{E}\Big(\frac{\delta \mathbb{1}_{\{T\le t\}}K\big(\frac{x-X}{h}\big)\frac{1}{h_n^d}}{1-H_x(T^-)}\Big)-\Lambda_x(t)f(x)\Big|<\frac{\epsilon}{2}.$$

Or

$$\mathbb{P}\Big(\Big|\Lambda_x(t)f(x)-\frac{1}{n}\sum_{i=1}^n\frac{\delta_i\mathbb{1}_{\{T_i\le t\}}K(\frac{x-X_i}{h})\frac{1}{h_n^d}}{1-H_x(T_i^-)}\Big|>\epsilon\Big)\le \mathbb{P}\Big(\Big|\Lambda_x(t)f(x)-\mathbb{E}\Big(\frac{\delta \mathbb{1}_{\{T\le t\}}K\big(\frac{x-X}{h}\big)\frac{1}{h_n^d}}{1-H_x(T^-)}\Big)\Big|+ \Big|\mathbb{E}\Big(\frac{\delta \mathbb{1}_{\{T\le t\}}K\big(\frac{x-X}{h}\big)\frac{1}{h_n^d}}{1-H_x(T^-)}\Big)-\frac{1}{n}\sum_{i=1}^n\frac{\delta_i\mathbb{1}_{\{T_i\le t\}}K(\frac{x-X_i}{h})\frac{1}{h_n^d}}{1-H_x(T_i^-)}\Big|>\epsilon\Big).$$

Et donc

$$\mathbb{P}\Big(\Big|\Lambda_x(t)f(x)-\frac{1}{n}\sum_{i=1}^n\frac{\delta_i\mathbb{1}_{\{T_i\le t\}}K(\frac{x-X_i}{h})\frac{1}{h_n^d}}{1-H_x(T_i^-)}\Big|>\epsilon\Big)\le \mathbb{P}\Big(\Big|\mathbb{E}\Big(\frac{\delta \mathbb{1}_{\{T\le t\}}K\big(\frac{x-X}{h}\big)\frac{1}{h_n^d}}{1-H_x(T^-)}\Big)-\frac{1}{n}\sum_{i=1}^n\frac{\delta_i\mathbb{1}_{\{T_i\le t\}}K(\frac{x-X_i}{h})\frac{1}{h_n^d}}{1-H_x(T_i^-)}\Big|>\frac{\epsilon}{2}\Big).$$

On a d'après l'inégalité de Tchebychev et grâce au fait que les $(X_i)_{1\le i\le n}$ sont indépendantes et identiquement distribuées :

$$\mathbb{P}\Big(\Big|\mathbb{E}\Big(\frac{\delta \mathbb{1}_{\{T\le t\}}K\big(\frac{x-X}{h}\big)\frac{1}{h_n^d}}{1-H_x(T^-)}\Big)-\frac{1}{n}\sum_{i=1}^n\frac{\delta_i\mathbb{1}_{\{T_i\le t\}}K(\frac{x-X_i}{h})\frac{1}{h_n^d}}{1-H_x(T_i^-)}\Big|>\frac{\epsilon}{2}\Big)\le\frac{4}{\epsilon^2}\mathrm{Var}\Big(\frac{1}{n}\sum_{i=1}^n\frac{\delta_i\mathbb{1}_{\{T_i\le t\}}K(\frac{x-X_i}{h})\frac{1}{h_n^d}}{1-H_x(T_i^-)}\Big) \le\frac{4\mathrm{Var}\Big(\frac{\delta \mathbb{1}_{\{\mathrm{T}\le \mathrm{t}\}}\mathrm{K}\left(\frac{\mathrm{x}-\mathrm{X}}{\mathrm{h}}\right)}{1-\mathrm{H_x(T^-)}}\Big)}{n\epsilon^2h_n^{2d}}.$$

Or

$$\begin{aligned}\mathrm{Var}\Big(\frac{\delta \mathbb{1}_{\{\mathrm{T}\le \mathrm{t}\}}\mathrm{K}\left(\frac{\mathrm{x}-\mathrm{X}}{\mathrm{h}}\right)}{1-\mathrm{H_x(T^-)}}\Big)&\le\int_{\mathbb{R}_+}\int_{\mathbb{R}_+^d}\Big(\frac{K(\frac{x-u}{h_n})\mathbb{1}_{\{v\le t\}}}{1-H_x(v^-)}\Big)^2f(u)\,\mathrm{du}\,\mathrm{dH_x^u(v)}\\&\le\int_0^t\frac{\mathrm{dH_x^u(v)}}{(1-H_x(v^-))^2}\int_{\mathbb{R}_+^d}\Big(K\big(\frac{x-u}{h_n}\big)\Big)^2f(u)\mathrm{du}\\&\le h_n^d\int_0^t\frac{\mathrm{dH_x^u(v)}}{(1-H_x(v^-))^2}\int_{\mathbb{R}_+^d}K(y)^2f(x+yh_n)\mathrm{dy}.\end{aligned}$$

Et

$$I_2:=\int_{\mathbb{R}_+^d}K(y)^2f(x+yh_n)\mathrm{dy}\underset{n\to\infty}{\longrightarrow}f(x)||K||_2^2.$$

On note

$$I(t)=\int_0^t\frac{\mathrm{dH_x^u(v)}}{(1-H_x(v^-))^2},$$

et on sait que $I(t)$ est bornée comme $t \leq \tau < \tau_{H_x}$. Ainsi,

$$\mathbb{P}\Big(\Big|\frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i \mathbb{1}_{\{T_i\leq t\}}K(\frac{x-X_i}{h})\frac{1}{h_n^d}}{1-H_x(T_i^-)} - \mathbb{E}\Big(\frac{\delta \mathbb{1}_{\{T\leq t\}}K\big(\frac{x-X}{h}\big)\frac{1}{h_n^d}}{1-H_x(T^-)}\Big)\Big| > \epsilon\Big) \leq \frac{4I(t)I_2}{nh_n^d\epsilon^2} \underset{n\to\infty}{\longrightarrow} 0.$$

Et donc

$$\mathbb{P}(\Big|\Lambda_x(t)f(x) - \frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i \mathbb{1}_{\{T_i\leq t\}}K(\frac{x-X_i}{h})\frac{1}{h_n^d}}{1-H_x(T_i^-)}\Big| > \epsilon) \underset{n\to\infty}{\longrightarrow} 0. \tag{21}$$

On déduit de la découpe (18), des résultats (19),(20) et (21), et de la convergence de $f_n(x)$ vers $f(x)$ (qui est non nul) :

$$\Lambda_{n,x}(t) \underset{n\to\infty}{\longrightarrow} \Lambda_x(t) \quad \text{en probabilité.}$$

Enfin, comme F_x est supposée continue, la formule (15) est simplifiable sous la forme $\log(S_x) = -\Lambda_x$. Ainsi :

$$\begin{aligned}|\log(S_x(t)) - \log(S_{n,x}(t))| &\leq |\log(S_x(t)) + \Lambda_{n,x}(t))| + |-\Lambda_{n,x}(t) - \log(S_{n,x}(t))| \\ &\leq |\Lambda_{n,x}(t)) - \Lambda_x(t)| + |-\Lambda_{n,x}(t) - \log(S_{n,x}(t))|.\end{aligned}$$

D'après ce qui précède et d'après la proposition 4.3.4, on sait que $|\Lambda_{n,x}(t)) - \Lambda_x(t)|$ et $|-\Lambda_{n,x}(t) - \log(S_{n,x}(t))|$ convergent tous les deux vers 0 en probabilité. Ainsi, on déduit

$$\log(S_{n,x}(t)) \underset{n\to\infty}{\longrightarrow} \log(S_x(t)) \quad \text{en probabilité.}$$

Et donc comme la fonction exponentielle est une fonction continue, on obtient bien

$$S_{n,x}(t) \underset{n\to\infty}{\longrightarrow} S_x(t) \quad \text{en probabilité.}$$

□

4.4 Simulations

De même que pour l'estimateur précédent, nous allons tester le bon fonctionnement de l'estimateur sur des échantillons simulés.

Pour tester l'estimateur, nous allons simuler le même genre de variables qu'expliquées dans la partie précédente : Y et C sont toujours indépendantes, mais pour chaque couple simulé, on simule tout d'abord une covariable suivant une loi uniforme entre $[0, 1]$, qui va influencer les lois de Y et C, Y suivra une loi exponentielle de paramètre $40x$ et C une loi exponentielle de paramètre $10x$. On se place alors en un point d'observation particulier (ici $x_0 = 0.5$) et on utilise l'estimateur de Beran. Les paramètres ici utilisés sont

$$K : u \longmapsto \frac{3}{4}(1-u^2)\mathbb{1}_{\{|u|<1\}}$$

et la suite $(h_n)_{n\in\mathbb{N}}$ est prise constante. Le choix de sa valeur résulte de ce que l'on appelle une cross-validation : on calcule pour différentes valeurs les poids obtenus, l'estimateur, ainsi que des quantités en lien, et on choisit la valeur qui minimise ces dites quantités.

La taille de l'échantillon est ici de 200. On voit que l'estimation obtenue est alors fidèle à la courbe théorique attendue.

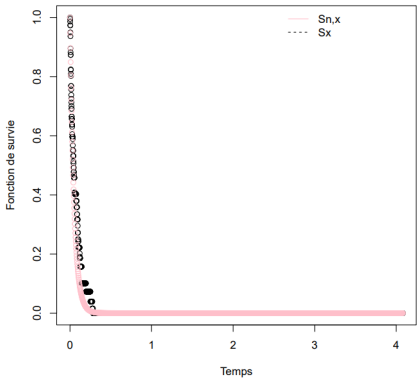

FIGURE 5 – Résultat obtenu après test de l'estimateur de Beran sur un échantillon simulé

On réalise de même une moyenne de 500 échantillons de taille 100, et on voit de nouveau que la courbe se lisse pour suivre la courbe théorique.

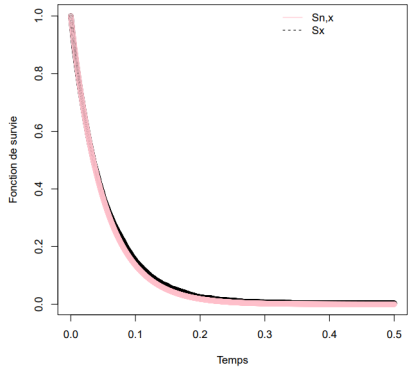

FIGURE 6 – Résultat obtenu après moyenne de tests de l'estimateur de Beran sur plusieurs échantillons simulés

Cependant, si l'on simule un échantillon pour lequel les variables Y et C ne sont plus indépendantes (par exemple à l'aide de copules, comme nous le verrons dans la partie suivante), alors l'estimateur ne fonctionne plus. Ici, les deux variables sont toujours influencées par leur covariable, mais elles sont également dépendantes, liées par une copule. Alors en utilisant les mêmes paramètres que précédemment, on se rend compte que l'estimateur n'est plus du tout performant.

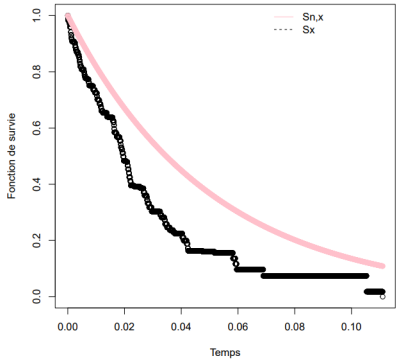

FIGURE 7 – Résultat obtenu après test de l'estimateur de Beran sur un échantillon simulé, avec variables non indépendantes

C'est pour traiter ce genre de cas que nous introduisons l'estimateur suivant.

5 Estimateur copulo-graphique

5.1 Fonctions de copule

Dans le cas où la variable observée et la variable de censure ne seraient pas indépendantes, il est nécessaire de trouver un outil permettant de caractériser la dépendance des deux variables. On utilise pour cela la notion de fonctions de copule.

Tout d'abord, on définit la notion de fonction de répartition pour un vecteur aléatoire. Soit $m \geq 1$. Soient $x = (x_1, ..., x_m)$ et $y = (y_1, ..., y_m)$ dans $\mathbb{R}^m$. On dit que $x \leq y$ si pour tout i dans $[\![1, m]\!]$, $x_i \leq y_i$.

Définition 5.1.1. *Soit X un vecteur aléatoire à valeurs dans $\mathbb{R}^m$. On appelle fonction de répartition de X la fonction $F : \mathbb{R}^m \to [0, 1]$ définie par $F(x) = \mathbb{P}(X \leq x)$ pour tout x dans $\mathbb{R}^m$.*

Définition 5.1.2. *On appelle fonction de copule de dimension m, ou plus simplement copule, toute fonction $C : [0, 1]^m \to [0, 1]$ étant la restriction à $[0, 1]^m$ de la fonction de répartition d'un vecteur aléatoire $U = (U_1, ...U_m)$ à valeurs dans $\mathbb{R}^m$, où pour tout i dans $[\![1, m]\!]$, U_i suit une loi uniforme dans $[0, 1]$.*

Exemple 5.1.1.

— *Si l'on considère le vecteur aléatoire $(U, ..., U)$ où U est une variable aléatoire de loi uniforme sur $[0, 1]$, alors la fonction de copule associée est définie par*

$$\forall (x_1, ..x_m) \in \mathbb{R}^m, \quad C(x_1, ...x_m) = \min_{1 \leq i \leq m} x_i.$$

— *Si l'on considére le vecteur aléatoire $(U_1, ..., U_m)$, où les $\{U_i\}_{1 \leq i \leq n}$ sont des variables aléatoires indépendantes suivant toutes une loi uniforme sur $[0, 1]$, alors la fonction de copule associée est définie par*

$$\forall (x_1, ..x_m) \in \mathbb{R}^m, \quad C(x_1, ...x_m) = \prod_{i=1}^{m} x_i.$$

Remarque 5.1.3. *Il existe en fait des critères analytiques pour reconnaître les fonctions de copule. En effet, si l'on se place par exemple en dimension 2, on peut prouver qu'une fonction $C : [0, 1]^2 \to [0, 1]$ est une copule si et seulement si*

1. $\forall u \in [0, 1], \quad C(u, 0) = C(0, u) = 0,$
2. $\forall v \in [0, 1], \quad C(v, 1) = C(1, v) = v,$
3. $\forall (u_1, u_2, v_1, v_2) \in [0, 1]^4, \quad C(u_2, v_2) - C(u_2, v_1) - C(u_1, v_2) + C(u_1, v_1) \geq 0.$

Montrons un lemme concernant le caractère lipschitzien (et donc continu) des copules qui nous sera utile pour la suite.

Proposition 5.1.4. *Soit C une copule de dimension m. Soient $u = (u_1, ..., u_m)$ et $v = (v_1, ..., v_m)$ dans $\mathbb{R}^m$. Alors*

$$|C(u) - C(v)| \leq \sum_{i=1}^{m} |u_i - v_i|.$$

Démonstration. Soit C une copule de dimension m, et $U = (U_1, ..., U_m)$ le vecteur aléatoire tel que C soit la fonction de répartition de U. On a alors par inégalité triangulaire :

$$\begin{aligned} |C(u) - C(v)| &= |\mathbb{E}(\mathbb{1}_{\{U \leq u\}}) - \mathbb{E}(\mathbb{1}_{\{U \leq v\}})| = |\mathbb{E}(\mathbb{1}_{\{U \leq u\}} - \mathbb{1}_{\{U \leq v\}})| \leq \mathbb{E}(|\mathbb{1}_{\{U \leq u\}} - \mathbb{1}_{\{U \leq v\}}|) \\ &\leq \mathbb{E}(\Big|\prod_{i=1}^{m} \mathbb{1}_{\{U_i \leq u_i\}} - \prod_{i=1}^{m} \mathbb{1}_{\{U_i \leq v_i\}}\Big|). \end{aligned}$$

Or on peut montrer par récurrence que pour toute suite $(a_n)_{n\in\mathbb{N}^*}$, $(b_n)_{n\in\mathbb{N}^*}$ de $\mathbb{C}^{\mathbb{N}^*}$ tel que pour tout $n \in \mathbb{N}^*$, $|a_n| \leq 1$ et $|b_n| \leq 1$, on a l'inégalité suivante :

$$\forall n \in \mathbb{N}^*, \quad |\prod_{k=1}^{n} a_k - \prod_{k=1}^{n} b_k| \leq \sum_{i=1}^{n} |a_k - b_k|.$$

Ainsi,

$$|C(u) - C(v)| \leq \mathbb{E}(\sum_{i=1}^{m} |\mathbb{1}_{\{U_i \leq u_i\}} - \mathbb{1}_{\{U_i \leq v_i\}}|) = \sum_{i=1}^{m} |u_i - v_i|.$$

□

Dans toute la suite, étant donnée F la fonction de répartition d'une variable aléatoire à valeurs réelles, on notera de nouveau F^{-1} son inverse généralisée définie par

$$\forall p \in]0, 1], \quad F^{-1}(p) = \inf\{x \in \mathbb{R} | F(x) \geq p\}.$$

On rappelle que pour tout x dans $\mathbb{R}$ et p dans $]0, 1]$, on a

$$F(x) \leq p \Longleftrightarrow x \leq F^{-1}(p).$$

et

$$F(F^{-1}(p)) \geq p,$$

avec égalité si $F^{-1}(p) > -\infty$ et si F est continue en $F^{-1}(p)$.

Théorème 5.1.5. ***Théorème de Sklar***

1. *Soit C une copule et $F_1, ...F_m$ des fonctions de répartition de variables aléatoires à valeurs réelles. Alors F définie par*
$$\forall (x_1, ..., x_m) \in \mathbb{R}^m, \quad F(x_1, ..., x_m) = C(F_1(x_1), ..., F_m(x_m)),$$
est la fonction de répartition d'un vecteur aléatoire à valeurs dans $\mathbb{R}^m$.
2. *Soit $X = (X_1, ..., X_m)$ un vecteur aléatoire à valeurs dans $\mathbb{R}^m$. On note $F_1, ..., F_m$ les fonctions de répartitions de $X_1, ..., X_m$ respectivement, et F la fonction de répartition du vecteur. Alors il existe C une copule telle que*
$$\forall (x_1, ..., x_m) \in \mathbb{R}^m, \quad F(x_1, ..., x_m) = C(F_1(x_1), ..., F_m(x_m)). \tag{22}$$
De plus, si les $F_1, ..., F_m$ sont continues, alors la copule est unique.

Démonstration. • Prouvons tout d'abord la première assertion. Soit C une copule et $F_1, ...F_m$ des fonctions de répartition de variables aléatoires à valeurs réelles. Soit $U = (U_1, ..., U_m)$ le vecteur aléatoire tel que les variables aléatoires $U_1, ..., U_m$ suivent toutes une loi uniforme sur $[0, 1]$ et tel que C soit restriction sur $[0, 1]^m$ de la fonction de répartition de U.

On pose $X = (F_1^{-1}(U_1), ..., F_m^{-1}(U_m))$, et on note F sa fonction de répartition. Soit $(x_1, ..., x_m)$ dans $\mathbb{R}^d$. On a alors

$$\begin{aligned} F(x_1, ..., x_m) &= \mathbb{P}(X_1 \leq x_1, ..., X_m \leq x_m) = \mathbb{P}(U_1 \leq F_1(x_1), ..., U_m \leq F_m(x_m)) \\ &= C(F_1(x_1), ..., F_m(x_m)). \end{aligned}$$

• Passons désormais à la seconde assertion. Soit $X = (X_1, ..., X_m)$ un vecteur aléatoire à valeurs dans $\mathbb{R}^m$. On note $F_1, ..., F_m$ les fonctions de répartitions de $X_1, ..., X_m$ respectivement, et F la fonction de répartition du vecteur.

Soit V une variable aléatoire indépendante de X suivant une loi uniforme sur $[0, 1]$. Soit i dans $[\![1, m]\!]$. Si F_i est continue, on pose $U_i = F_i(X_i)$, et sinon, on pose $U_i = F_i(X_i^-) + V \times \sum_{v \in \Delta_i} \mathbb{P}(X_i = v) \mathbb{1}_{\{X_i = v\}}$ où Δ_i réprésente l'ensemble des points de discontinuités de F_i.

Dans les deux cas, U_i suit une loi uniforme sur $[0, 1]$ et presque sûrement, on a $\{X_i \leq x\} = \{U_i \leq F_i(x)\}$.

En effet, dans le cas continu, pour tout u dans $]0, 1[$, on a

$$\mathbb{P}(U_i \leq u) = \mathbb{P}(F_i(X_i) \leq u) = \mathbb{P}(X_i \leq F^{-1}(u)) = F(F^{-1}(u)) = u.$$

Donc U_i suit effectivement une loi uniforme sur $[0, 1]$ et on a bien $\{X_i \leq x\} = \{U_i \leq F_i(x)\}$ presque sûrement.

Traitons désormais le cas où F_i n'est pas continue. Tout d'abord on remarque que U_i est bien à valeurs dans $[0, 1]$. Soit u dans $]0, 1[$.

Si $v = F_i^{-1}(u)$ n'est pas un point de continuité de F_i, alors u appartient à $]F_i(v^-), F_i(v)]$, et donc

$$\{U_i < u\} = \{U_i < F_i(v^-)\} \cup \{U_i \in]F_i(v^-), F_i(v)]\} = \{F_i(X_i) < F(v^-)\} \cup \{X_i = v, V < \frac{u - F_i(v^-)}{\mathbb{P}(X_i = v)}\}.$$

Et donc comme $\frac{u - F(v^-)}{\mathbb{P}(X_i = v)} = 1$,

$$\mathbb{P}(U_i < u) = P(X_i < v) + P(X_i = v) = u.$$

Si $v = F_i^{-1}(u)$ est un point de continuité de F_i, alors on a cette fois

$$\{U_i < u\} = \{U_i < F_i(v)\} = \{F_i(X_i) < F(v)\} = \{X_i < v\},$$

et donc

$$\mathbb{P}(U_i < u) = \mathbb{P}(X_i < F_i^{-1}(u)) = u.$$

Ainsi U_i suit bien une loi uniforme sur $[0, 1]$.

De plus, par définition on a $\{X_i \leq x\} \subset \{U_i \leq F_i(x)\}$, et d'autre part $\{U < F_i(x)\} \subset \{F_i(X^-) \leq F_i(x)\} \subset \{X \leq x\}$. Et comme U suit une uniforme on a presque sûrement $\{U \leq F_i(x)\} = \{U < F_i(x)\}$, et ainsi $\{X_i \leq x\} = \{U_i \leq F_i(x)\}$.

On en déduit alors que la fonction de répartition du vecteur $(U_1, ..., U_m)$ est une copule, que l'on nomme C, et on a alors, pour tout $(x_1, ..., x_m)$ dans $\mathbb{R}^m$

$$F(x_1, ..., x_m) = \mathbb{P}(X_1 \leq x_1, ..., X_m \leq x_m) = \mathbb{P}(U_1 \leq F_1(x_1), ..., U_m \leq F_m(x_m)) = C(F_1(x_1), ..., F_m(x_m)).$$

Enfin, si les $F_1, ..., F_m$ sont continues, alors l'intervalle $]0, 1[$ est inclus dans l'image de chacun des F_i. Ainsi, cela fixe l'image de C sur $]0, 1[^m$, et donc par continuité de la copule, cela implique son unicité.

□

Remarque 5.1.6. *Ce théorème implique des propriétés similaires pour les fonctions de survie. Plaçons nous par exemple en dimension 2 : on considère (X, Y) un vecteur aléatoire à valeurs dans $\mathbb{R}^2$. On note F_X la fonction de répartition de X, F_Y celle de Y, et F celle du couple, et de même on note S_X la fonction de survie de X, S_Y celle de Y et S celle du couple. D'après le théorème de Sklar, il existe une copule C telle que*

$$\forall (x, y) \in R^2, \quad F(x, y) = C(F_X(x), F_Y(y)).$$

On a alors, pour (x, y) dans R^2

$$\begin{aligned} S(x, y) &= \mathbb{P}(X > x, Y > y) = 1 - \mathbb{P}((X \leq x) \cup (Y \leq y)) = 1 - F_X(x) - F_Y(y) + F(x, y) \\ &= S_X(x) + S_Y(y) - 1 + C(1 - S_X(x), 1 - S_Y(y)). \end{aligned}$$

Ainsi, en posant

$$\begin{aligned} \tilde{C} : [0, 1]^2 &\to [0, 1] \\ (u, v) &\mapsto u + v - 1 + C(1 - u, 1 - v), \end{aligned}$$

on a

$$S(x, y) = \tilde{C}(S_X(x), S_Y(y)).$$

On remarque grâce aux propriétés analytiques données au cours de la remarque 5.1.3 que $\tilde{C}$ est bien une fonction de copule.

Enfin, on définit une catégorie spéciale de copule, qui nous servira par la suite. Si on la définit ici en dimension 2 pour coller à l'utilisation que l'on en fera ensuite, la définition peut être adaptée pour une dimension quelconque.

Définition 5.1.7. *Une fonction de copule C de dimension 2 est appelée copule archimédienne s'il existe une fonction $\phi : [0, 1] \to [0, +\infty]$ (que l'on appelle générateur de la copule) continue, convexe et strictement décroissante telle que $\phi(1) = 0$ et telle que*

$$\forall (x, y) \in \mathbb{R}^2, \quad C(x, y) = \phi^{[-1]}(\phi(x) + \phi(y)),$$

où $\phi^{[-1]}$ désigne la pseudo inverse de ϕ définie par :

$$\forall s \in \mathbb{R}_+, \quad \phi_x^{[-1]}(s) = \begin{cases} \phi_x^{-1}(s) & si \quad 0 \leq s \leq \phi_x(0), \\ 0 & si \quad \phi_x(0) \leq s \leq +\infty. \end{cases}$$

Remarque 5.1.8. *Le générateur ϕ dépend généralement d'un certain paramètre, qui peut être, comme nous le verrons dans la suite, en lien par exemple avec une covariable.*

Exemple 5.1.2. *Il existe plusieurs exemples connus de copules, dont voici les générateurs (θ représentant à chaque fois un certain paramètre donné) :*

- ***Clayton :*** $\phi_\theta(t) = \frac{1}{\theta}(t^{-\theta} - 1)$ *où* $\theta \in [-1, +\infty[\backslash\{0\}$,
- ***Frank :*** $\phi_\theta(t) = -\log\left(\frac{\exp(-\theta t)-1}{\exp(-\theta)-1}\right)$ *où* $\theta \in \mathbb{R}^*$,
- ***Gumbel :*** $\phi_\theta(t) = (-\log(t))^\theta$ *où* $\theta \in [1, +\infty[$,
- ***Joe :*** $\phi_\theta(t) = -\log(1 - (1 - t)^\theta)$ *où* $\theta \in [1, +\infty[$,
- ***Ali–Mikhail–Haq :*** $\phi_\theta(t) = \log\left(\frac{1-\theta(1-t)}{t}\right)$ *où* $\theta \in [-1, 1]$.

Enfin, on peut remarquer que la copule produit (associée au cas de deux variables indépendantes) est également une copule archimédienne, générée par $-\log$.

5.2 Construction de l'estimateur

Dans toute cette section, les variables d'intérêt Y et de censure C ne sont plus indépendantes. La variable observée est toujours un triplet (T, δ, X), où $T = \min(Y, C)$ et $\delta = \mathbb{1}_{\{Y \leq C\}}$ et où X représente comme précédemment une covariable, à valeurs dans $\mathbb{R}_+^d$.

Soit x dans $\mathbb{R}_+^d$. On notera $\bar{H}_x$ la fonction de survie de T conditionnellement au fait que $X = x$:

$$\forall t \in \mathbb{R}_+, \quad \bar{H}_x(t) = 1 - H_x(t) = \mathbb{P}(T > t | X = x).$$

De plus, on pose $\mathscr{S}_x$ la fonction de survie du couple aléatoire (Y, C) conditionnellement au fait que $X = x$:

$$\forall (t_1, t_2) \in (\mathbb{R}_+)^2, \quad \mathscr{S}_x(t_1, t_2) = \mathbb{P}(Y > t_1, C > t_2 | X = x).$$

On note $\mathscr{C}_x$ (dépendant de x) la fonction de copule telle que

$$\forall (t_1, t_2) \in (\mathbb{R}_+)^2, \quad \mathscr{S}_x(t_1, t_2) = \mathscr{C}_x(\bar{F}_x(t_1), \bar{G}_x(t_2)),$$

où on note $\bar{F}_x(t_1) = 1 - F_x(t_1) = \mathbb{P}(Y > t_1 | X = x)$ et de même $\bar{G}_x(t_2) = 1 - G_x(t_2) = \mathbb{P}(C > t_2 | X = x)$.

Pour la suite, nous allons supposer que $\mathscr{C}_x$ est une copule archimédienne : il existe donc un générateur $\phi_x : [0, 1] \to [0, +\infty]$ (dépendant lui aussi de x), remplissant les conditions de la définition 5.1.7, tel que

$$\forall (t_1, t_2) \in (\mathbb{R}_+)^2, \quad \mathscr{S}_x(t_1, t_2) = \phi_x^{[-1]}[\phi_x(\bar{F}_x(t_1)) + \phi_x(\bar{G}_x(t_2))], \tag{23}$$

On remarque pour la suite que l'équation (23) implique :

$$\forall t \in \mathbb{R}_+, \quad \bar{H}_x(t) = \mathscr{S}_x(t, t) = \phi_x^{[-1]}[\phi_x(\bar{F}_x(t)) + \phi_x(\bar{G}_x(t))]. \tag{24}$$

On suppose alors de nouveau que l'on dispose d'un échantillon aléatoire $\{(T_i, \delta_i, X_i)\}_{1 \leq i \leq n}$, où les $\{(T_i, \delta_i, X_i)\}$ sont indépendants et identiquement distribués selon la loi de (T, δ, X) . Lorsque les $\{T_i\}_{1 \leq i \leq n}$ sont rangés par ordre croissant, on les notera $\{T_{(i)}\}_{1 \leq i \leq n}$, et on notera alors de même $\{(T_{(i)}, \delta_{(i)}, X_{(i)})\}_{1 \leq i \leq n}$.

On cherche à construire deux fonctions $\bar{F}_{n,x}$ et $\bar{G}_{n,x}$ approximant respectivement $\bar{F}_x$ et $\bar{G}_x$, continues à droite, en escalier ayant leurs sauts en les T_i respectivement tels que $\delta_i = 1$ et $\delta_i = 0$, et telle que $\bar{F}_{n,x}(0) = \bar{G}_{n,x}(0) = 1$.

On a vu précédemment que l'on peut approcher $\bar{H}_x$ par une fonction empirique donnée par

$$\forall s \in \mathbb{R}_+, \quad \bar{H}_{n,x} = \sum_{j=1}^{n} w_i(x) \mathbb{1}_{\{T_j > s\}},$$

où les $\{w_i\}_{1 \leq i \leq n}$ sont les poids qui ont été définis dans la définition 4.2.2.

Pour que les fonctions empiriques approchent la réalité (et donc en particulier l'équation (24)), on veut que l'équation

$$\bar{H}_{n,x}(T_i) = \phi_x^{[-1]}[\phi_x(\bar{F}_{n,x}(T_i)) + \phi_x(\bar{G}_{n,x}(T_i))]$$

soit vérfiée pour tout i dans $[\![1, n]\!]$.

On se place en T_i tel que $\delta_i = 1$. la fonction $\bar{G}_{n,x}$ n'a pas de saut en ce point, et donc $\bar{G}_{n,x}(T_i) = \bar{G}_{n,x}(T_i^-)$. La fonction $\bar{F}_{n,x}$ effectue, elle, un saut et on a alors :

$$\phi_x(\bar{F}_{n,x}(T_i^-)) - \phi_x(\bar{F}_{n,x}(T_i)) = \phi_x(\bar{H}_{n,x}(T_i^-)) - \phi_x(\bar{H}_{n,x}(T_i)).$$

Or en utilisant une somme téléscopique, on a, pour t dans $\mathbb{R}_+$

$$\phi_x(\bar{F}_{n,x}(t)) = - \sum_{T_{(i)} \leq t, \delta_{(i)} = 1} \phi_x(\bar{F}_{n,x}(T_{(i)}^-)) - \phi_x(\bar{F}_{n,x}(T_{(i)})) = - \sum_{T_{(i)} \leq t, \delta_{(i)} = 1} \phi_x(\bar{H}_{n,x}(T_{(i)}^-)) - \phi_x(\bar{H}_{n,x}(T_{(i)})).$$

Et donc

$$\bar{F}_{n,x}(t) = \phi_x^{[-1]}\Big(- \sum_{T_{(i)} \leq t, \delta_{(i)}=1} \phi_x(\bar{H}_{n,x}(T_{(i)}^-)) - \phi_x(\bar{H}_{n,x}(T_{(i)}))\Big).$$

Etant donné que l'argument de $\phi_x^{[-1]}$ est égal à $\phi_x(\bar{F}_{n,x}(t))$, que $\bar{F}_{n,x}(t) \geq 0$, et que ϕ_x est strictement décroissante, on en déduit qu'il n'est jamais supérieur à $\phi_x(0)$, et on peut donc remplacer $\phi_x^{[-1]}$ par ϕ_x^{-1}. On obtient alors l'estimateur suivant :

$$\bar{F}_{n,x}(t) = \phi_x^{-1}\Big(- \sum_{T_{(i)} \leq t, \delta_{(i)}=1} \phi_x(\bar{H}_{n,x}(T_{(i)}^-)) - \phi_x(\bar{H}_{n,x}(T_{(i)}))\Big).$$

Définition 5.2.1. *On appelle estimateur copulo-graphique la fonction en escalier suivante :*

$$\begin{aligned}\forall t \in \mathbb{R}_+, \quad \bar{F}_{n,x}(t) &= \phi_x^{-1}\Big(- \sum_{T_{(i)} \leq t, \delta_{(i)}=1} \phi_x(\bar{H}_{n,x}(T_{(i)}^-)) - \phi_x(\bar{H}_{n,x}(T_{(i)}))\Big) \\ &= \phi_x^{-1}\Big(- \sum_{T_{(i)} \leq t, \delta_{(i)}=1} \phi_x\big(\sum_{j=i}^{n} w_j(x)\big) - \phi_x\big(\sum_{j=i+1}^{n} w_j(x)\big)\Big).\end{aligned}$$

Remarque 5.2.2. *Si l'on considère la copule correpondant au cas où les variables Y et C sont indépendantes ($\mathscr{C}_x(u,v) = uv$, engendrée par $-\log$), alors on retrouve l'estimateur de Beran présenté précédemment.*

5.3 Consistance de l'estimateur

Voici un théorème qui nous donne la consistance de l'estimateur précédent dans un cadre bien précis.

Théorème 5.3.1. *On suppose que X est à valeurs dans $[0,1]$. On note $(\tilde{X}_i)_{1\leq i\leq n}$ les covariables $(X_i)_{1\leq i\leq n}$ classées par ordre croissant. On fait alors les hypothèses suivantes :*

$$\tilde{X}_n \underset{n\to\infty}{\longrightarrow} 1, \quad \max_{1\leq i\leq n-1}(\tilde{X}_{i+1} - \tilde{X}_i) = O(\frac{1}{n}) \quad et \quad \max_{1\leq i\leq n-1}(\tilde{X}_{i+1} - \tilde{X}_i) - \min_{1\leq i\leq n-1}(\tilde{X}_{i+1} - \tilde{X}_i) = o(\frac{1}{n}).$$

On suppose ensuite que K est une densité de probabilité à support fini sur $[-M, M]$ pour un certain $M > 0$, qu'elle est lipschitzienne d'ordre 1, et que son moment d'ordre 1 est nulle.

Soit $\tau < \tau_{H_x}$ (où $\tau_{H_x} = \inf\{t \in \mathbb{R}_+ | H_x(t) = 1\}$. . On suppose que sur $[0,\tau]$, H_x et H_x^u vérifie les hypothèses suivantes :

- *$\frac{\partial^2 L_x(t)}{\partial x^2}$ existe et est continue en $(x,t) \in [0,1]\times[0,T]$.*
- *$\frac{\partial^2 L_x(t)}{\partial t^2}$ existe et est continue en $(x,t) \in [0,1]\times[0,T]$.*
- *$\frac{\partial^2 L_x(t)}{\partial x \partial t}$ existe et est continue en $(x,t) \in [0,1]\times[0,T]$.*

Pour finir, on suppose que l'équation (23) est bien vérifiée et que le générateur ϕ_x de la copule archimédienne vérifie les conditions suivantes : $\frac{\partial \phi_x(v)}{\partial v}$ et $\frac{\partial^2 \phi_x(v)}{\partial v^2}$ sont lipschitziennes par rapport à x avec des constantes de Lipchitz bornées, et $\frac{\partial^3 \phi_x(v)}{\partial v^3}$ existe, est négative, et continue en $(x,v) \in [0,1]\times]0,1]$.

Alors sous toutes ces hypothèses, si $nh_n^5 \to 0$ et $\frac{(\log(n))^3}{nh_n} \to 0$, ou si $h_n = Cn^{-\frac{1}{5}}$ pour un certain $C > 0$, on a

$$\forall t \in [0,\tau], \quad (nh_n)^{-\frac{1}{2}}\big(F_{n,x}(t) - F_x(t)\big) \xrightarrow[n\to\infty]{\mathscr{L}} \mathscr{N}(0, \sigma(t)),$$

où $\sigma(t)$ est une variance dépendant de t.

Nous admettrons de nouveau ce théorème par manque d'outils, mais une preuve est disponible dans la référence [9].

5.4 Simulations

Nous allons de nouveau, à l'image des deux parties précédentes, expérimenter la bonne convergence de l'estimateur à l'aide de tests sur des échantillons simulés.

Les échantillons sont créés de la façon suivante : pour chaque couple simulé, on commence de nouveau par simuler avant tout une covariable x suivant une loi uniforme dans $[0, 1]$. Puis on simule un couple de variables tel que leur fonction de répartition suive l'équation (23), en choisissant le générateur d'une certaine famille de copule archimédienne, que l'on applique au x obtenu précédemment. Ensuite, grâce à une transformation qui utilise l'inverse généralisée de la fonction de répartition, on transforme le couple obtenu pour que Y suive une loi exponentielle de paramètre 5, et pour que C suive une loi exponentielle de paramètre 2. On se place en un point d'observation particulier, ici $x_0 = 0.5$, et on utilise l'estimateur copulo-graphique. De nouveau, la suite $(h_n)_{n\in\mathbb{N}}$ utilisée est constante et choisie grâce à la cross-validation. Quant au noyau, celui utilisé ici est le suivant :

$$K : u \longmapsto \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2}.$$

La taille de l'échantillon est 200, et la copule archimédienne utilisée est celle de Frank.

On voit que l'estimation obtenue suit la courbe théorique, mais de façon moins précise que précédemment. Cela s'explique par le fait qu'ici les paramètres sont plus compliqués à adapter, et donc tel quel le taux de censure est plus élevé que dans les parties précédentes, ce qui a évidemment comme conséquence de diminuer l'efficacité de l'estimateur.

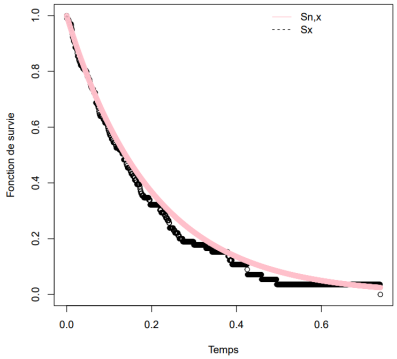

FIGURE 8 – Résultat obtenu après tests de l'estimateur copulo-graphique sur un échantillons simulé

Comme précédemment, la courbe devient bien plus précise si l'on moyenne les résultats obtenus pour 100 échantillons de taille 200.

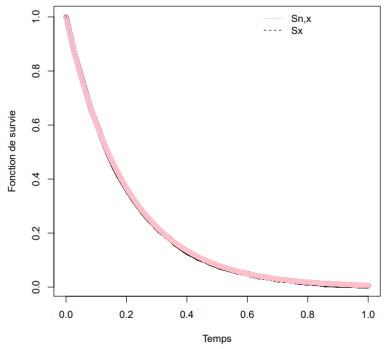

FIGURE 9 – Résultat obtenu après moyenne de tests de l'estimateur copulo-graphique sur plusieurs échantillons simulés

Cependant, pour que cette méthode fonctionne, il faut tout d'abord que Y et C soient bel et bien liées par une copule archimédienne, et il est qui plus est nécessaire de connaître la famille de cette copule, pour avoir le bon générateur.

En effet, voici un exemple où Y et C sont simulées comme précédemment, à l'exception près qu'elles sont cette fois liées par la copule de Fréchet, dite copule du minimum donnée par

$$C : (u, v) \longmapsto \min(u, v).$$

On voit alors que si l'on essaie d'utiliser l'estimateur précédent en utilisant le générateur de la copule archimédienne de Frank, on ne parvient pas du tout à approximer la fonction de survie théorique.

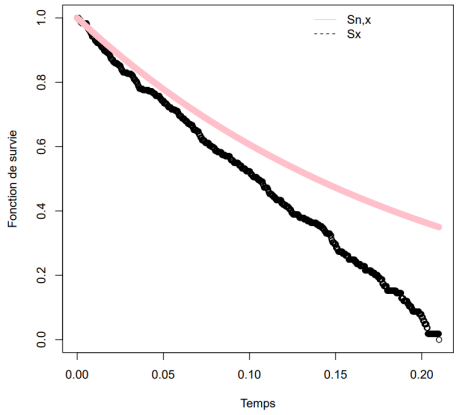

FIGURE 10 – Résultat obtenu après test de l'estimateur copulo-graphique sur un échantillon Fréchet en utilisant le générateur Frank

Voici maintenant le cas d'un échantillon où les variables sont liées par une copule archimédienne de Clayton, que l'on esssaie d'estimer en utilisant l'estimateur précédent avec le générateur de Frank.

Le graphique ci dessous résulte d'une simulation où le taux de censure était pourtant faible, mais on constate que l'estimateur a tout de même du mal à approcher la courbe théorique.

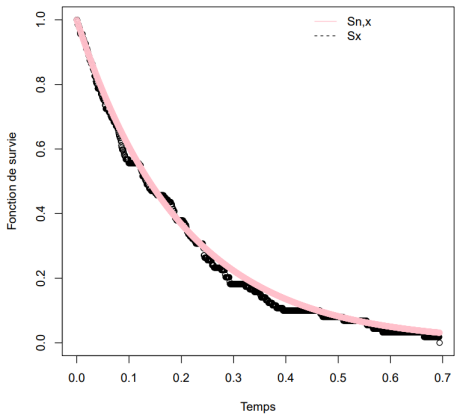

FIGURE 11 – Résultat obtenu après test de l'estimateur copulo-graphique sur un échantillon Clayton en utilisant le générateur Frank

Ainsi, il serait nécessaire de savoir si la copule reliant les deux variables est archimédienne ou non, et si oui de quel type.

6 Conclusion

Ainsi, nous avons obtenu différents estimateurs permettant d'approximer la fonction de survie d'une donnée censurée, dans plusieurs cadres. Cependant, bien les utiliser implique de savoir si la donnée et la censure sont influencées par une covariable ou non, sont indépendantes ou non, et si oui, savoir comment elles sont reliées. Ainsi, pour compléter cette étude, il s'agirait d'utiliser ou de créer des tests d'indépendance et/ou des tests pour connaître la fonction de copule associée au couple.

Références

[1] Mikael ESCOBAR-BACH, Introduction à l'analyse de survie (cours), 2019

[2] David VERNON WIDDER, The Laplace transform, 1946

[3] Li SHIU-TANG, A brief introduction to Lebesgue-Stieltjes integral, 2017

[4] Jean-Christophe BRETON, Lois conditionnelles (cours), 2008

[5] Winfried STUTE, On almost sure convergence of conditional empirical distribution functions, 1986

[6] L.P. DEVROYE et T.J. WAGNER, The strong uniform consistency of kernel density estimates, 1980

[7] Roger B. NELSON, An introduction to copulas, 2006

[8] Jean-François DEMAS, Fonctions de répartition et copules (cours), 2008

[9] Roel BRACKERS et Noel VERAVERBEKE, A copula-graphic estimator for the conditional survival function under dependent censoring, 2005