N d'ordre : 18/2023-D/MT

REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
Université des Sciences et de la Technologie Houari Boumediène

Faculté de Mathématiques

Thèse de Doctorat en Sciences

Présentée pour l'obtention du grade de Docteur

En : Mathématiques

Spécialité : Probabilités-Statistique

Par : Sarah GHETTAB

Sujet

Propriétés asymptotiques des estimateurs à noyaux non symétriques pour des données incomplètes

Soutenue publiquement le 03 juillet 2023, devant le jury composé de :

M. Tarek MEDKOUR	Professeur	à l'ENSIA, Sidi Abdellah	Président.
Mme. Zohra GUESSOUM	Professeur	à l'USTHB, Alger	Directrice de Thèse.
Mme. Nadjia EL SAADI	Professeur	à l'ENSSEA, Kolea	Examinatrice.
Mme. Samira TALEB	Maître de Conférence/A	à l'USTHB, Alger	Examinatrice.
Mme. Ourida SADKI	Professeur	à l'USTHB, Alger	Invitée.

Remerciments

*Louanges à **Dieu**, de m'avoir accordé santé, courage et patience pour mener à bien ce travail.*

*C'est tout naturellement que je souhaite adresser mes premiers remerciements à ma directrice de thèse, Professeur **Zohra Guessoum**. Elle a su accompagner mes premiers pas dans la recherche par ses explications clairvoyantes sur le monde académique. Aussi grâce à sa synthèse des différents domaines de recherche elle a pu me proposer un sujet passionnant et d'actualité. Sa générosité, sa disponibilité et ses conseils avisés ont ponctué ces années de collaboration. Aussi le moteur de mon travail a été son exigence initiale, sa rigueur et son honnêteté dans ses innombrables relectures qui étaient toujours contre balancés par son soutien et sa gentillesse dans mes moments de doute. Un grand merci à elle.*

*Dans un deuxième temps, j'adresse mes remerciements et ma profonde reconnaissance aux professeurs : **Abdelkader Tatachak**, doyen de la faculté des mathématiques à l'USTHB et Madame **Ourida Sadki** professeur à l'USTHB qui m'ont permis grâce à leurs présence et leurs considérables apports durant les séances du groupe de travail des données incomplètes (GTDI), d'orienter et de développer ma recherche convenablement, merci infiniment.*

*Également, toute ma gratitude à Monsieur le professeur **Tarek Medkour** pour avoir accepté de présider cet honorable jury.*

*Je tiens aussi à remercier les autres membres du jury : Madame **Nadjia El Saadi** professeur à l'ENSSEA Koléa, Madame **Samira Taleb** H.D.R à l'USTHB pour : le temps qu'elles ont*

consacré à la lecture et l'analyse de cette thèse ainsi que leurs présence à la soutenance.

Je voudrais aussi remercier, tous mes collègues enseignants à l'ESSA-Alger et à l'USTHB, je pense en particulier à Madame Bey Siham : chef du département de probabilité et statistique de l'USTHB, Madame Graba Samia et son mari Kessira Abderrahim, pour les innombrables services qu'ils m'ont rendu, leur aide ainsi que leurs encouragements, merci beaucoup.

Un groupe de personnes sans qui tout ceci n'aurait pas été possible sont mes meilleures amies : Ilhame, Yasmine, Amina, Mouna, Soumia, Asma. Je ne puis que vous remerciez pour votre amitié pendant toutes ces années. Toutes nos discussions, tous vos conseils et tous les moments passés ensemble ont fait qui je suis. Je vous remercie.

Il va de soi que je n'oublie pas les membres du groupe de travail GTDI : Asma, Latifa, Wafa Hassiba, Hamida, Nassima, Soumia, Salwa, Sabrina, Louiza,..., pour tous les moments inoubliables partagés durant nos années de recherche. Ce fut un plaisir de travailler avec vous et j'espère que cette collaboration continuera.

Mes plus vifs remerciements s'adressent à mes chers parents, mes sœurs et frères, spécialement : Wail, Nawel, sofia, et Nouma, qui ont toujours cru en moi, m'ont soutenu et aidé par leurs encouragements et leurs disponibilités. Mille mercis.

Je tiens également à adresser mes remerciements les plus sincères à ma belle famille.

Enfin et surtout, je réitère mes remerciements les plus vifs à mon époux. Sa compréhension, ses encouragements et la patience dont il a fait preuve au cours de ces 9 années ont constitué un soutien inestimable pour moi. Qu'il trouve ici l'expression de toute ma gratitude. Évidement je n'oublie pas mes chères enfants Razane et Mohamed El Hassan qui ont égayé ces années de travail par leur présence.
A toute personne qui m'a aidé de près ou de loin durant ce parcours, merci infiniment.

Résume

Les problématiques abordés et les résultats établis dans ce travail portent essentiellement sur l'étude des propriétés asymptotiques des estimateurs à noyaux asymétriques de certaines fonctionnelles : la densité, le taux de hasard, et la fonction de répartition et qui permettent de corriger les effets de bord bien connus dans le cas de noyaux symétriques.

Dans le cadre de l'estimation non paramétrique, nous nous plaçons dans un modèle de données censurées à droite. Les observations sont au préalable considérées indépendantes et identiquement distribuées (i.i.d.).

Nous proposons un estimateur lisse de la fonction de répartition par l'utilisation du théorème de Hille. Ainsi, nous proposons un nouveau type d'estimateurs à noyau asymétrique pour la densité et la fonction de hasard qui fonctionnent convenablement aux bornes, lorsque la variable d'intérêt est positive et censurée à droite. Les estimateurs sont construits à l'aide de noyaux asymétriques d'espérance 1. Nous établissons la convergence uniforme presque sûre avec vitesse, et nous étudions les propriétés asymptotiques et la normalité des estimateurs résultants. Une large étude de simulation est menée pour conforter les résultats théoriques. Une application aux données réelles est réalisée.

Mots clés : consistance forte uniforme ; données censurées ; effet de bord ; estimateur à noyau ; estimateur produit-limite ; normalité asymptotique ; noyaux asymétriques ; vitesse de convergence.

Abstract

The issues addressed and the results established in this work relate essentially to the study of the asymptotic properties of estimators with asymmetric kernels of certain functional : the density, the hazard rate, and the distribution functions and which behave well at the boundary.

As part of the non-parametric estimation, we place ourselves in a right-censored data frameworck. The observations are first considered independent and identically distributed (i.i.d.).

We propose a smooth estimator of the distribution function by using Hille's theorem. Thus, we propose a new type of asymmetric kernel estimators for density and hazard functions that perform well at the boundary, when the variable of interest is positive and right-censored. The estimators are constructed using asymmetric kernels of expectation 1. On a compact set, we establish the almost sure uniform convergence with rate, and we study the asymptotic properties and the normality of the resulting estimators. A large simulation study is carried out to confirm the theoretical results. An application to real data is carried out.

Keywords : asymmetric kernel ; asymptotic normality ; boundary effect ; convergence rate ; data censored ; kernel estimator ; product-limit estimator ; strong uniform consistency .

Liste des abréviations

v.a variables aléatoires

i.i.d indépendantes identiquement distribuées.

p.s presque sûrement.

f.r fonction de répartition.

f.r.e fonction de répartition empirique.

LLI Loi du Logarithme Itérée.

EKM Estimateur de Kaplain Meire.

PC Pourcentage de censure.

IG inverse Gaussian.

RIG reciprocal inverse Gaussian.

VC-classes (Vapnik-$\check{C}$ervonenkis) classes.

T.C.L Théorème centrale limite.

Articles et Communications

1. Ghettab Sarah, & Guessoum Zohra (2022). Asymptotic properties of asymmetric kernel estimators for non-negative and censored data, *Communications in Statistics - Theory and Methods*. https://doi.org/10.1080/03610926.2022.2150059.

2. Ghettab, S., Guessoum, Z. (2022). Density and hazard rate estimation for right-censored data by using Log-normal kernel. *Journée Scientifique du Laboratoire MSTD(JSL'22)*, Université des Sciences et de la Technologie Houari-Boumédiène (USTHB), Alger, Février, 2022.

3. Ghettab, S., Guessoum, Z. (2021). Non-parametric density estimation for positive and censored data : Application to Log-normal kernel, *The First Online International Conference on Pure and Applied Mathematics (IC-PAM 2021),* 26-27 Mai 2021, Ouargla University, Algeria.

4. Ghettab, S., Guessoum, Z. (2020). An Asymmetric Kernel Estimators for Density function : application to breast cancer data. *6th Conference of the Stochastic Modeling Techniques and Data Analysis. International Conference (SMTDA 2020)* Barcelona, Spain, June, 2020.

5. Ghettab, S., Guessoum, Z. (2019). An Asymmetric Kernel Estimators for censored data. 4 ^{ème} *Colloque International Modélisation Stochastique et Statistique* (MSS 2019), Université des Sciences et de la Technologie Houari-Boumédiène (USTHB), Alger, Novembre, 2019. https://hal.archives-ouvertes.fr/hal-02593238/file/MSS%202019_Proceedings.pdf

6. Ghettab, S., Guessoum, Z. (2018). Some Asymptotic properties of Asymmetric Kernel Esti-

mator., *Congrès des Mathématiciens Algériens*, (CMA 2018), Université M'hamed Bougara Boumerdès (UMBB), mai 2018.

Table des matières

Introduction

Introduction Générale

L'une des principales méthodes de la théorie moderne d'estimation fonctionnelle non paramétrique est la méthode du noyau. C'est une méthode qui semble commode, simple à mettre en œuvre, et peut être appliquée à de nombreux problèmes importants d'estimation. Dans cette thèse, nous nous intéressons principalement à l'estimation de la fonction de densité et cette dernière peut être utilisée pour estimer toute autre fonction comme la fonction de risque, la fonction de régression et autres. Les premiers résultats sur l'estimation de la densité par noyau sont dus à Rosenblatt (1956)[59] et Parzen (1962) [56]. De nombreux efforts ont été consacrés à l'étude des performances optimales des estimateurs à noyaux. Pour une bibliographie générale et des exposés de synthèse, nous renvoyons aux livres de : Silverman (1986) [63], Wand et Jones (1995)[67], Bosq(1996) [9], Artur Gramacki (2017) [40].

Soit $(\Omega, \mathcal{A}, \mathbb{P})$ un espace probabilisé et soit f une fonction de densité de probabilité associée à une variable aléatoire (r.v) X à support dans $\mathbb{R}$ définie sur $(\Omega, \mathcal{A}, \mathbb{P})$. Soient $X_1, X_2, \ldots, X_n, n$-échantillon de X. L'estimateur $\hat{f}$ de Parzen-Rosenblat de f est défini par

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

où h_n est le paramètre de lissage ($h_n \to 0 \quad \text{quand} \quad n \to \infty$), et K est une fonction de densité appelée noyau supposée être symétrique. Plusieurs auteurs ont souligné que le succès de cet estimateur est bien illustrée lorsque la distribution de la densité sous-jacente n'est pas loin de la forme gaussienne c'est à dire définie sur toute la droite réelle $\mathbb{R}$, mais ont constaté son mauvais comportement surtout aux extrémités de ce dernier si la distribution est fortement asymétrique, multimodale

et à queue lourde. De telles caractéristiques sont habituelles avec par exemple la distribution de revenu, le sinistre en assurance, qui sont définis sur un support positif $[0, +\infty[$, ou comme l'âge d'une personne qui est une variable à support compact $[0, T]$. Un exemple simple est la densité exponentielle $f(x) = e^{-x}, \quad x > 0$. La FIGURE 1 montre une estimation de cette densité par $\hat{f}$ défini dans (1) en utilisant le noyau gaussien (courbe noire pointillée), basée sur une taille d'échantillon de $n = 1000$, la densité théorique est indiquée pour comparaison (courbe rouge continue).

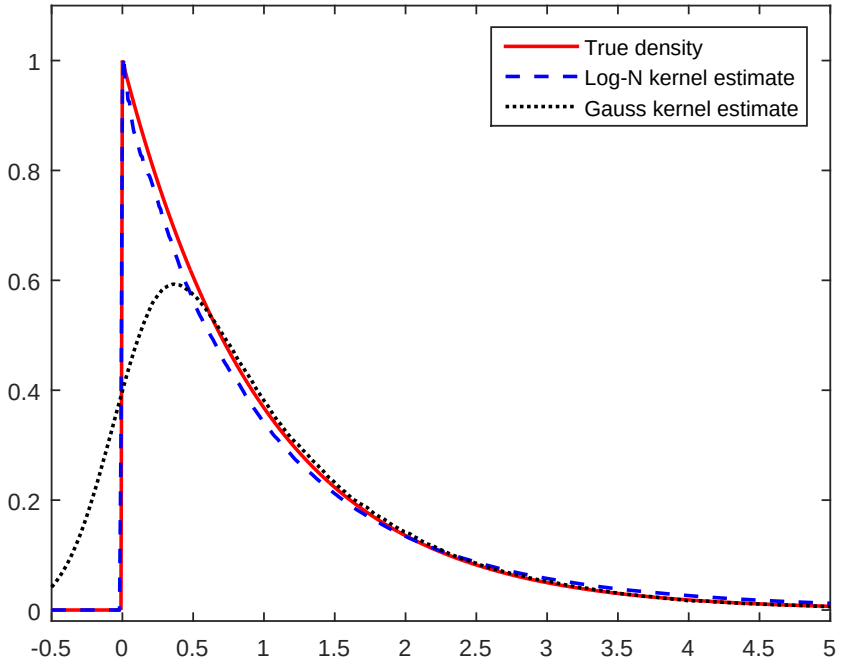


FIGURE 1 – La vrai densité exponentiel $Exp(1)$ (ligne continue), son estimation par le noyau gaussien (ligne pointillée) et le noyau Log-Normal (ligne discontinue), $n = 1000$.

Il est clair que cet estimateur souffre de deux types de problèmes aux bornes. Le premier c'est qu'il dépasse le domaine de définition et donne (incorrectement) un poids positif à une partie inaccessible $(-\infty, 0]$, le second est l'échec de l'estimation des points de discontinuité comme il est montré dans la FIGURE 1 pour des x positif et proches du point $x = 0$, l'estimateur à noyau gaussien essaie d'estimer des fréquences relativement élevées, tandis que pour des x négatifs et proches de $x = 0$ il vise à estimer des valeurs proches de zéro, d'où $\hat{f}(0)$ sous-estime la vrai valeur $f(0)$. En conséquence, l'estimateur classique $\hat{f}$ n'estime pas correctement f aux extrémités et c'est

ce qu'on appelle effet de bord ou en anglais Spill-over. Différentes approches dans la littérature ont été suggérées pour contourner ce problème, les plus célèbres sont résumées comme suit :

- Transformation des données : Cette méthode à été suggérée par Deveroye et Gyorfi (1985) [24]. Wand et *al.* (1991) [68] proposent de transformer les données, ensuite de trouver l'estimateur de la densité en utilisant les données transformées, et la rétro-transformation par un changement de variable pour l'estimation de la densité des données d'origine. Une étude rigoureuse se trouve dans Marron et Ruppert (1994) [54]. Pour la transformation logarithmique on peut se référer à Silverman (1986) [63] et Charpentier et Flachaire (2015) [13].

- Utilisation d'estimateur polynomial local est intéressante pour rectifier le biais aux bornes. Tibshirani et Hastie (1987) [65] ont remarqué empiriquement que la régression locale linéaire corrige automatiquement les effets de bord. Des années après, Fan et Gijbels (1992) [27] ont combiné l'idée de lissage par les polynômes locaux et l'utilisation d'un paramètre de lissage variable pour la fonction de régression et montrent que l'estimateur proposé n'a pas d'effet de bord. Cheng (1997) [17], Cheng et *al* (1997) [18] ont prouvé l'efficacité et la correction automatique du biais aux bornes des estimateurs polynomiales locaux pour la densité et la régression. Ruppert et Wand (1994)[60] ont étendu leurs résultats sur le biais et la variance asymptotiques dans le cas des variables multivariées.

- En 1985, Schuster suggère une technique de symétrisation qui consiste à refléter les données autour des bornes pour avoir la suite $\{X_1, -X_1, X_2, -X_2, \ldots\}$ et appliquer ensuite le lissage à noyau sur l'ensemble des données et leurs reflétées. Cline et Hart (1991) [21] ont montré que l'estimateur proposé par Schuster (1985) [62] convergera selon le critère de l'erreur quadratique moyenne intégrée comme si la densité n'avait pas de points de discontinuité. Cette approche est particulièrement utile si la fonction à estimer est nulle au voisinage de zéro. Pour plus de détails sur cette méthode, nous renvoyons à Silverman (1986, page 30) [63].

- Dans le cas où l'emplacement des points de discontinuité de la densité sont connus, une solution disponible est l'usage des noyaux de bord (boundary kernels) autour du point de discontinuité. Cette méthode a été largement utilisée dans la littérature, nous citons à titre d'exemples : Gasser et Muller (1979) [32], Gasser et *al* (1995) [33] , Muller (1991) [55], Jones (1993) [46], Zhang et Karunamuni (1998) [71]. Pour plus de détails, on peut se référer aux livres de Scott (2015) [64], et Gramacki (2017) [40]. Bien que les noyaux de bord puissent

être très utiles, il existe des problèmes potentiellement sérieux avec les données réelles : un premier inconvénient de cette méthode est que les points de discontinuité ne sont pas connus, le second plusieurs noyaux de bord peuvent donner des estimations négatives à proximité des bornes, surtout quand $f(0) = 0$.

De nouvelles approches plus adaptative que les méthodes précédentes ont été proposées. Nous citons une version de régression plus sophistiquée que l'approche de Schuster [62], celle de Hall et Wehrly (1991) [44], Cowling et Hall (1996) [22], basée sur la génération de pseudo-données (pseudodata en anglais) au voisinage des bornes du support de la densité, et l'estimateur produit bénéficie d'ordres optimaux du biais et de variance. Rice (1984) [58] qui combine deux estimateurs à noyau avec des paramètres de lissage différent pour obtenir une réduction du bais. Zhang et *al* (1999) [70] proposent une combinaison entre les méthodes de pseudo-données, de transformation et de réflexion. Dans le même concept combinant parfois quelques-unes de ces méthodes, nous citons Bouezmarni et al. (2005) [8] ainsi que Karunamuni et Alberts (2005) [48].

Une nouvelle approche qui a efficacement résolut le problème des effets de bord a été proposée par Chen (1999, 2000a) [14, 16]. Cette approche est basée sur l'utilisation des noyaux bêta et gamma pour estimer respectivement les densités à support égal à $[0, 1[$ et $[0, +\infty[$. Cela a donné naissance à une classe intéressante d'estimateurs qui ne souffrent pas du biais aux bornes. C'est la classe d'estimateurs à noyaux asymétriques. Ce sont généralement des noyaux dont le support coïncide avec celui de la densité à estimer et peuvent changer de forme selon la position du point d'estimation. Une procédure similaire a été étudiée par Scaillet (2004) [61] avec les noyaux inverses gaussiens (IG) et inverses gaussiens réciproques (RIG) pour les densité à support positif. L'avantage des estimateurs à noyaux asymétriques est qu'ils ouvrent la voie à une large utilisation : nous citons par exemple les travaux de Chen (2000a) [15] pour la régression, Bouezmarni et Rolin (2003) [6], Bouezmarni et Scaillet (2005) [7], Gustafsson et *al.* (2007) [41] avec la méthode de transformation, Fernandes et *al.* (2014) [29]. Néanmoins, la variance des estimateurs du type de ceux proposés par Chen (1999) [14] explose en $x = 0$, et pour éviter le problème, les auteurs donnent deux sortes de formalisation de la variance, une pour $x/b \longrightarrow \infty$ et l'autre pour $x/b \longrightarrow \kappa$ où κ une constante positive et b est le paramètre de lissage. Cela nous semble un peu arbitraire car il ne donne pas une image claire de ce qui se passe près du point $x = 0$ ou de sa proximité.

Les méthodes précédentes fournissent une correction efficace des effets de bords, mais les plus efficaces ont tendance à être assez compliquées. Cela décourage leurs utilisation et implique également

une analyse difficile. Cependant, la recherche de simplicité dans les estimateurs non paramétriques qui soient similaires aux estimateurs à noyaux classiques a conduit Bagai et Prakasa Rao (1996) [3] à remplacer le noyau $K(.)$ par un autre noyau non négatif. Ils montrent que l'estimateur résultant a les même propriétés asymptotiques que l'estimateur classique sous certaines conditions de régularité. Cela élimine certainement le problème de la probabilité positive dans la région négative. Ainsi, ils ont noté que seulement les r premières statistiques d'ordre contribuent à l'estimation avec $X_{(r)} \leq x \leq X_{(r+1)}$, où $X_{(i)}$ désigne $i^{ème}$ statistique d'ordre . Chaubey et Sen (1996) [11] ont proposé un estimateur de densité comme la dérivé d'une version lisse de la fonction de répartition (f.r) estimé par le théorème sur le lissage uniforme en analyse réelle, théorème dit de Hille (1948) donné dans Chaubey et Sen (2013, Theorem 2.1, pages 257-260) [12]. Ce même estimateur a été proposé plus tôt dans Gawronski et Stadtmüller (1980, 1981) [34, 35] sous le nom de histogramme lissé. Chaubey et *al* (2007) [10] ont également proposé un estimateur basé sur une généralisation du lissage par le théorème de Hille couplé à une nouvelle idée de perturbation autour du point d'estimation x pour tenir compte du biais aux bornes. Les propriétés asymptotiques de consistance et de normalité sont établies sous certaines hypothèses de régularité appropriées. Plusieurs exemples de noyaux asymétriques sont utilisés dans cet estimateur comme le noyau Gamma, Bêta du second type, et log-normal. Une manière simple pour expliquer la construction de cet estimateur est de voir que : puisque qu'on a estimé des variables positives, on remarque que le support de l'estimateur classique n'est pas adéquat avec $(0, \infty[$ et en général ce sont les noyaux asymétriques qui possèdent le même support. Par conséquent, on échange le noyau symétrique $K(.)$ par un autre asymétrique, et pour répartir correctement la masse du noyau autour de x, nous remplaçons l'argument $(x - X_i)$ (qui peut être négatif) dans l'équation (1) par un argument toujours positif défini par $\frac{X_i}{x}$. Alors, pour $x > 0$, on propose d'estimer $f(x)$ par

$$f_n^*(x) = \frac{1}{n} \sum_{i=1}^{n} \alpha_{i,x} \; K_n\left(\frac{X_i}{x}\right),$$

où (K_n) est une famille de fonctions de densité non symétriques et $\alpha_{i,x}$ est une fonctionnelle de x et X_i à déterminer pour que l'estimateur $f_n^*(x)$ s'intègre à 1. De là, nous pouvons prendre $\alpha_{i,x} = \frac{X_i}{x^2}$ et donc,

$$f_n^*(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{X_i}{x^2} \; K_n\left(\frac{X_i}{x}\right). \tag{2}$$

Maintenant, et pour éviter tout problème au point $x = 0$, nous introduisons une suite appropriée de nombres positifs $a_n \longrightarrow 0$, comme une perturbation autour de x, et nous définissons l'estimateur

de f par :

$$f_n(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{X_i}{(x+a_n)^2}\ K_n\left(\frac{X_i}{x+a_n}\right), \tag{3}$$

Une petite simulation de cet estimateur en choisissant le noyau log-normal (voir 3.3.2), pour une taille d'échantillon égale à $n = 1000$, est représenté avec la densité de la loi $Exp(1)$ sur la FIGURE 1 dans laquelle nous voyons le très bon ajustement de la densité.

Dans cette thèse on s'intéresse à des modèleS de survie où les données ne sont pas complètement observées, la variable aléatoire est généralement positive, et donc il est approprié d'utiliser le lissage par les noyaux asymétriques. Nous avons remarqué qu'il n'y a pas beaucoup d'utilisation des noyaux asymétriques pour les données incomplètes. Bouezmarni et *al.* (2011) [5] ont adapté la procédure de lissage du noyau Gamma proposée par Chen (2000b) [16], pour estimer les fonctions de densité et de hasard. Kuruwita et *al.* (2010) [50] ont proposé une méthode pivot pour estimer les densités supportées sur $[0, \infty)$ en utilisant des noyaux asymétriques après avoir souligné la déficience des approches précédentes.

Nous proposons donc dans cette thèse un nouveau type d'estimateurs à noyau performant aux bornes, basé sur des noyaux asymétriques. Plus précisément, nous combinons deux points : nous construisons un estimateur de densité à l'aide de noyaux d'espérance 1 en utilisant l'idée de perturbation. Cet estimateur est étudié pour des données complètement observées par Chaubey et *al.* (2007)[10] qui ont utilisé un noyau asymétrique sans quantifier le taux de convergence. Ici, nous avons adapté cet estimateur pour estimer la densité dans un cadre de censure à droite et nous donnons la vitesse de convergence uniforme presque sûre.
D'autre part, on s'intéresse à la fonction de hasard $\lambda(x) := \dfrac{f(x)}{1-F(x)}$, introduite à l'origine dans la littérature statistique par Watson et Leadbetter (1964) [69]. Cette fonctionnelle est une interprétation intuitive de l'évaluation du risque dans les études de survie. En effet c'est une mesure à court terme du risque qui tient compte du fait que votre risque de décès (échec) change au fil du temps. Cette fonction est importante dans plusieurs domaines de la statistique appliquée (médecine , fiabilité, ingénierie, science actuarielle, ...). De nombreuses études ont montré que la fonction de risque joue un rôle crucial dans le risque de base en finance. Nous renvoyons à Engle (2000) [25] pour un aperçu, et à Fernandes et Grammig (2000) [28] pour l'exploitation des noyaux asymétriques dans l'analyse de la durée financière.

Contribution de la thèse

Dans cette thèse, nous étudierons principalement l'estimation non-paramétrique des fonctions de densité et de taux de hasard pour des données incomplètes. Le manuscrit est structuré en trois chapitres en plus de l'introduction et la conclusion.

Le chapitre introductif expliquera bien le problème d'effet de bords associé à l'estimateur à noyau classique, nous montrons par un exemple de la loi exponentielle l'insuffisance de ce dernier au voisinage de zero. Ainsi, nous donnons les différentes méthodes existant dans la littérature pour corriger ce problème. Ensuite, on rappellera les principaux concepts utiles à la compréhension des parties essentielles de la thèse.

Le premier chapitre rappellera certaines notions de base qui sont utiles à la compréhension des concepts, mais pas forcement utilisées dans les parties essentielles de cette thèse.

Le deuxième chapitre aborde quelques estimateurs connus pour la correction du problème du biais aux bornes. L'accent est mis sur les estimateurs à noyaux asymétriques dans le cas de données complètement observées. Ainsi, nous présenterons brièvement les résultats existants.

Le troisième chapitre allie les principaux résultats des travaux réalisés soumis à publication et parus. Il s'agit de l'étude de la fonction de densité par la méthode à noyaux asymétriques pour des données censurées aléatoirement à droite. Un nouvel estimateur lisse de la fonction de répartition pour des observations i.i.d. est proposé. L'étude d'un tel estimateur est aussi justifiée par le fait qu'il intervient explicitement dans l'expression de l'estimateur de la fonction taux de hasard que nous nous proposons d'étudier. La dérivée de la fonction de répartition donne une estimation de la densité, et en conséquence la suggestion de deux estimateurs de la fonction de hasard. La convergence uniforme presque sûre (sur un compact) avec vitesse des estimateurs sont établies en utilisant deux approches différentes, les VC-Classes pour la fonction de répartition et le recouvrement pour la densité. Une large étude numérique sur des données générées et un exemple sur un jeu de données réelles est donnés.

Dans le quatrième chapitre, nous établissons la normalité asymptotique des estimateurs de la fonction de densité et de hasard pour des données censurées en utilisant le théorème de Lyapunov.

Les intervalles de confiance sont construit et une étude numérique est établie.

Un dernier chapitre de conclusion permet de fixer les points mis en avant lors de cette thèse. Par ailleurs, en tirant avantage des résultats obtenus, nous proposons un aperçu des perspectives envisageables pour des études ultérieures liées à l'estimation par les noyaux asymétriques.

Les différentes parties de cette thèse sont reliées par une problématique commune qui est l'estimation non paramétrique des fonctions : densité, fonction de répartition, et de hasard. Une approche est proposée basée sur l'estimateur à noyaux asymétrique, une large comparaison sur une étude de données simulées et une application sur des données réelle. L'objectif principal de cette contribution est de réduire, le biais aux bornes des fonctions étudiées et la sensibilité de l'estimation au voisinage des points de discontinuité (en particulier au point zero) et à la censure.

Chapitre 1

Généralités Sur Les Modèles de Durée

Sommaire

1.1 Durée de survie

Dans de nombreux domaines de la statistique appliquée, on est amené par l'apparition d'une *variable de durée*, c'est-à-dire le temps qui passe d'un instant initial jusqu'à la survenue d'un évènement particulier qui n'est pas forcément la mort. Cette durée peut être par exemple :

- Dans le domaine médical : le temps sans tumeur, le temps entre le début du traitement et la réponse, la durée de la rémission après une intervention chirurgicale, le temps jusqu'au décès ....,

- Dans le domaine d'ingénierie et de la fiabilité : la durée de bon fonctionnement ou de défaillance des appareils, composants ou systèmes électroniques,

- Dans le domaine de l'assurance : durée de la vie humaine, durée de l'arrêt de travail, mais aussi durée d'attente entre 2 sinistres, durée avant la ruine, ...

- Dans le domaine socio-économique : durée de chômage, durée du premier mariage, durée de libération conditionnelle des criminels, durée de rechute de prix de pétrole,...,

La première particularité des données de durée est d'être générées par des variables aléatoires positives $(0, +\infty[$, la seconde c'est que *la variable de durée* peut ne pas être complètement observée du fait essentiellement de deux phénomènes : la censure lorsque, pour certains individus l'évènement étudié ne se produit pas pendant la période d'observation, et la troncature lorsque la variable aléatoire n'est observable que sur une sous partie de $(0, +\infty[$ et parfois à cause des deux phénomènes en même temps. Nous rappelons rapidement les approches classiques des différents modèles étudiés pour modéliser ce type de données néanmoins ce sont liées par le même objectif : déterminer la loi de probabilité de *la variable de durée* par l'estimation de l'une des fonctions suivantes :

1. Fonction de survie : $S(t) = P(X > t), \forall t \geq 0$. Pour t fixé c'est la probabilité de survivre jusqu'à l'instant t.

2. Fonction de répartition : $F(t) = P(X \leq t) = 1 - S(t), \forall t \geq 0$, Pour t fixé c'est la probabilité de mourir avant l'instant t.

3. Densité de probabilité : C'est une fonction $f(t) \geq 0$ telle que pour tout $t \geq 0$

$$F(t) = \int_0^t f(s) \ \ ds.$$

Si la fonction de répartition a une dérivée au point t alors

$$f(t) = \lim_{dt \to 0} \frac{P(t < X \leq t + dt)}{dt} = F^{'}(t) = -S^{'}(t).$$

Pour t fixé, la densité de probabilité caractérise la probabilité de mourir dans un petit intervalle de temps après l'instant t.

4. Taux de hasard ou la fonction de risque instantané λ est définie comme la probabilité qu'un individu fasse l'évènement considéré durant un intervalle de temps très court sachant qu'il a survécu jusqu'au début de l'intervalle

$$\lambda(t) = \lim_{h \to 0} \frac{P(X \leq t + h / X > t)}{h}$$

5. Le taux de hasard cumulé ou la fonction de risque cumulative évaluée au temps t est l'intégrale de la fonction de risque entre 0 et t

$$\Lambda(t) = \int_0^t \lambda(u) \quad du = -\ln S(t).$$

1.2 Modèle censuré

La censure est le phénomène le plus couramment rencontré lors du recueil de données de survie. Une censure se produit lorsque l'évènement étudié n'intervient pas pendant la période d'observation pour une raison ou une autre. Ainsi, pour chaque individu on considère trois durées : son temps de survie X_i, son temps de censure C_i, et la durée réellement observée Y_i.

1.2.1 Censure à Droite

La durée de vie est dite censurée à droite si l'individu n'a pas subi l'événement pendant la période d'observation. En présence de censure à droite, les durées de vie X_i ne sont pas toutes observées ; pour certaines d'entre elles, on sait seulement qu'elles sont supérieures aux durées de censure C_i. Un exemple typique est celui où l'évènement considéré est le décès du patient malade, et la durée

d'observation est une durée totale d'hospitalisation. L'expérimentateur peut également fixer une date de fin d'expérience. Ainsi, pour toute observation sans « évènement » observée avant cette date, on sait simplement que la durée de survie est supérieure à la durée de participation observée. Le modèle de censure à droite peux se dégénérer en fonction de l'un des types suivants :

Censure de type I (fixé)

Ce modèle est souvent utilisé dans les études épidémiologiques ou un essai clinique. Pour des raisons de temps ou de coût, l'investigateur mettra fin à l'étude ou rapportera les résultats avant que tous les sujets ne réalisent leurs événements. Au lieu d'observer $X_1, X_2, \ldots, X_n$ qui nous intéressent, on observe X_i que lorsque $X_i \leq C$, sinon on sait seulement que $X_i > C$. Les données observées sont représentées par des paires de variables aléatoires (Y, δ), où δ indique si la durée de vie X correspond à un événement ($\delta = 1$) ou à une censure ($\delta = 0$), et Y est égal à X, si la durée de vie est observée, et à C si elle est censurée, c'est-à-dire $Y_i = \min(X_i, C) = X_i \wedge C, \forall i = 1, \ldots, n$. Dans ce cas, le temps de censure C est souvent fixé et désigne la durée de l'étude, et le nombre d'événements observés est aléatoire.

Censure de type II (attente)

On décide d'observer la durée de fonctionnement de n appareils dans des études de fiabilité jusqu'à ce que r d'entre eux soient en pannes et on arrête l'étude à ce moment là. Soit $X_1, X_2, \ldots, X_n$ les durées de bon fonctionnement des appareils, et soit $X_{(1)}, X_{(2)}, \ldots, X_{(i)}, \ldots, X_{(n)}$ les statistiques d'ordres correspondantes. Dans ce cas la date de censure est la $r^{ème}$ statistique d'ordre et donc on observe : $Y_i = X_i \wedge X_{(r)}$. Ce cas est fréquemment utilisé en fiabilité lorsqu'on observe jusqu'à la première panne.

Censure de type III (aléatoire)

Il s'agit du type de censure le plus fréquemment utilisé. Dans de tels cas les individus ont des raisons différentes de censure à part de l'évènement d'intérêt sur laquelle porte l'étude, comme par exemple, un déménagement pour un patient, la panne d'une machine à cause d'une surtension d'électricité, ce qui mène l'effet aléatoire à la censure. Ainsi, a chaque individu i, est associé un couple de v.a (X_i, C_i) positives où X_i est son temps de survie et C_i son temps de censure, tel que seule la plus petite est observée , c'est-à-dire on observe $Y_i = X_i \wedge C_i$ et $\delta_i = \mathbb{I}_{\{Y_i = X_i\}} = \mathbb{I}_{\{X_i \leq C_i\}}$, indicateur de non censure. Généralement pour ce type, nous avons besoin d'une hypothèse d'indépendance entre *la variable de durée* et *la variable de censure* pour faire un sens à l'inférence. Dans de

nombreuses études, la censure est une combinaison de type III et de type I. Dans de telles études, certains patients sont censurés au hasard lorsque, par exemple, ils quittent le lieu de l'étude pour des raisons sans rapport avec l'événement d'intérêt, tandis que d'autres sont censurés de type I à la fin de la période d'étude fixe.

1.2.2 Censure à Gauche

Une durée de vie X_i associée à un individu i dans une étude est considérée comme censure à gauche si elle est inférieure au temps de censure C_i c'est-à-dire que l'événement d'intérêt s'est déjà produit pour l'individu avant que cette personne ne soit observée dans l'étude au temps C_i. Pour de tels individus, nous savons qu'elles ont vécu l'événement quelque temps avant le temps C_i, mais de temps exact de l'apparition de leurs événements est inconnue. La durée de vie exacte X_i sera connue si et seulement si elle est supérieure ou égale à C_i. Les données observées dans ce cas peuvent être représentées par des paires de variables aléatoires $(Y_i, \delta_i), \forall i = 1, \ldots, n$, comme dans le cas précédent, où Y_i est égal à X_i si la vraie durée de vie est observée et δ_i indique si la durée de vie exacte X_i est observée $(\delta_i = 1)$ ou $(\delta_i = 0)$ sinon. Notons que, la censure à gauche est l'opposé de la censure à droite, $Y = \max(X_i, C_i)$.

Exemple 1.1. *En recherche biomédicale, nous pouvons connaître la date d'entrée d'un patient à l'hôpital, et savoir qu'il a ensuite survécu un certain temps ; toutefois, le statisticien ne connaît pas exactement la date à laquelle les premiers symptômes de la maladie se sont produits ou ont été diagnostiqués. Il est alors fréquent de ne connaitre qu'un minorant de la date de ce premier évènement. La durée de survie, prise comme la différence entre le temps où se produit l'évènement d'intérêt et le temps où s'est produit l'évènement initial, est ici supérieure ou égale à la durée effectivement observée par le statisticien.*

1.2.3 Censure Double (ou mixte)

Souvent, si la censure à gauche se produit dans une étude, la censure à droite peut également se présenter, et les durées de vie sont considérées comme doublement censurées (voir Turnbull (1974) [66]). Encore une fois, les données peuvent être représentées par une paire de variables (Y, δ) où $Y = \max(\min(X_i, C_r), C_l)$ est la variable observée, et δ vaut 1 si Y est une vraie durée de vie, 0 si Y est la variable censurée à droite, et -1 si Y est la variable censurée à gauche. Ici, C_l est le temps avant lequel certains individus vivent l'événement et C_r est le temps après lequel certains individus vivent l'événement. X sera connu exactement s'il est inférieur ou égal à C_r et supérieur

ou égal à C_l.

Exemple 1.2. *Dans le centre d'apprentissage des petits enfants, souvent l'intérêt est de se concentrer sur le test des enfants pour déterminer le moment où un enfant apprend à accomplir certaines tâches spécifiques. L'âge auquel un enfant apprend la tâche serait considéré comme le temps de cet événement. Souvent, certains enfants peuvent déjà effectuer la tâche lorsqu'ils commencent l'étude. Dans ce cas les durées de cet événement correspondent à ces enfants sont considérées comme des censures à gauche. On associé une autre partie d'enfants qui ne peuvent pas apprendre la tâche pendant toute la période d'étude, auquel cas ces enfants seraient censurés à droite. Ainsi, cet échantillon contiendrait des données doublement censurées.*

1.2.4 Censure par Intervalle

Un type plus général de censure se présente lorsque la durée de vie X_i n'est pas observée, mais est connue qu'elle se produit dans un intervalle. Une telle censure d'intervalle se trouve lorsque les patients d'un essai clinique ou d'une étude longitudinale (comme les cohortes) bénéficient d'un suivi périodique et que l'apparition de l'événement du patient n'est connue que pour tomber dans un intervalle $(L_i, R_i]$ (L pour l'extrémité gauche et R pour l'extrémité droite de l'intervalle de censure). Ce type de censure peut également se produire dans des expériences industrielles où il y a une inspection périodique pour le contrôle du bon fonctionnement des équipements. Il apparaît que toute combinaison de censure à gauche, à droite ou par intervalle peut se produire dans une étude. Bien sûr, la censure par intervalle est une généralisation de la censure à gauche et à droite car, lorsque l'extrémité gauche est 0 et l'extrémité droite est C_l, nous avons une censure à gauche et, lorsque l'extrémité gauche est C_r et l'extrémité droite est infinie, nous avons la censure à droite.

Exemple 1.3. *Les données d'une étude rétrospective*[1] *pour comparer les effets cosmétiques de la radiothérapie seule par rapport à la radiothérapie et à la chimiothérapie adjuvante*[2] *chez les femmes atteintes d'un cancer du sein précoce sont rapportés. Les patientes ont été observées initialement tous les 4 à 6 mois mais, à mesure que leur rétablissement progressait, l'intervalle entre les visites s'allongeait. L'événement d'intérêt était la première apparition d'une rétraction du mamelon modérée ou sévère, une détérioration esthétique du sein. La durée exacte de la rétraction n'était*

1. Une étude se base sur l'acquisition de données présentes dans les dossiers médicaux des personnes ciblées ou dans un registre de données au moment de la soumission.
2. Se dit d'un traitement qui complète un traitement principal afin de prévenir un risque de récidive locale ou de métastases.

connue que dans l'intervalle entre les visites (intervalle censuré) ou après la dernière fois que la patiente a été vue (censurée à droite).

1.3 Modèle de Troncature

La troncature diffère de la censure au sens où elle concerne l'échantillonnage lui même. Ainsi, une variable X est tronquée par un sous ensemble éventuellement aléatoire A de $\mathbb{R}^+$ si au lieu d'observer X, on observe X uniquement si $X \in A$. Les éléments de l'échantillon "tronqué" appartiennent tous à A, et suivent donc la loi de X conditionnée par l'appartenance à A.

1.3.1 Troncature à Gauche

On dit qu'il y a troncature à gauche lorsque la variable d'intérêt X n'est observable que si elle est supérieure à la variable de troncature T c'est-à-dire $\{X > T\}$.

Exemple 1.4. *Durée de vie après la retraite : on étudie la durée de vie après la retraite de sujets qui entrent dans l'enquête à la suite d'un tirage au sort dans une caisse de retraite. Un sujet n'est donc observé que si sa durée de vie après la retraite excède le délai entre sa prise de retraite et l'instant de l'enquête. La durée de vie après la retraite est donc tronquée à gauche par ce délai. Elle peut aussi être censurée à droite si la fin de l'enquête a lieu alors que le sujet est toujours vivant.*

1.3.2 Troncature à Droite

On dit qu'il y a troncature à droite lorsque la variable d'intérêt X n'est observable que si elle est inférieure à la variable de troncature T c'est-à-dire $\{X < T\}$. Cette troncature apparaît, par exemple, dans le cas où on estime la distribution des étoiles à partir de la terre, et les étoiles trop éloignées ne sont pas visibles et donc elles sont tronquées à droite.

1.4 Modèle Tronqué à Gauche et Censuré à Droite (LTRC)

Soient $X_1, \ldots, X_N$, N variables aléatoires de même loi que X_i où X_i v.a à valeurs dans $\mathbb{R}^+$ représente le temps de survie du $i^{ème}$ individu, et soient $T_1, \ldots, T_N$, N variables aléatoires de même loi que $T \in \mathbb{R}^+$ représentant les variables de troncature. On introduit de même les variables aléatoires de censure $C_1, \ldots, C_N$, de même loi que $C \in \mathbb{R}^+$. Un modèle aléatoire est dit tronqué à gauche par une variable aléatoire T et censuré à droite par une variable aléatoire C noté "modèle

aléatoire LTRC", si au lieu d'observer la durée de vie X, on observe le vecteur (Z, T, δ) uniquement si $Z \geq T$, où

$$Z_i = \min(X_i, C_i), \quad \delta = \mathbb{I}_{(X_i \leq C_i)}$$

δ est la fonction indicatrice. Si $Z < T$ rien n'est observé. Formellement l'échantillon observé est noté

$$\{(Z_i, T_i, \delta_i), i = 1, \ldots, n\}, n \leq N$$

où les n triplets d'observations sont des réalisations identiquement distribuées du vecteur aléatoire (Z, T, δ) réellement observées c.à.d si $Z_i \geq T_i$.

Chapitre 2

Estimateurs à Noyaux non Symétriques : Cas des Données Complètes

Sommaire

2.1 Introduction

Dans ce chapitre, on s'intéresse aux estimateurs de la densité à support positif qui ne souffre pas du biais aux bornes, en particulier on se rapporte aux estimateurs à noyaux asymétriques proposés par Chaubey et *al.* (2007) [10]. Nous rappelons quelques propriétés asymptotiques de ces estimateurs comme la convergence uniforme presque sure et l'erreur quadratique moyenne intégrée (MISE).

2.2 Construction de l'estimateur de Chaubey

L'idée principale de la construction de l'estimateur de Chaubey et *al.* (2007) [10] est le lemme suivant basé sur une approche générale d'une technique de lissage spécialement adaptée au cas de données positives donnée par le théorème de Hille, qui est une légère variation du lemme 1 donné dans Feller (1965, XVII.1)(voir Annexe pour plus de détails sur le Lemme).

Lemme 2.1. *Soit u une fonction continue bornée. Soit $G_{x,n}$, $n = 1, 2, \ldots$ une famille de distributions d'espérance $\mu_n(x)$ et de variance $h_n^2(x)$, telle que $\mu_n(x) \longrightarrow x$ et $h_n(x) \longrightarrow 0$. Alors on a*

$$\tilde{u}(x) = \int_{-\infty}^{+\infty} u(t)\ dG_{n,x}(t) \longrightarrow u(x). \tag{2.1}$$

La convergence est uniforme dans tout semi intervalle dans lequel $h_n(x) \longrightarrow 0$ et u sont uniformément continues.

Soit $(\Omega, \mathcal{A}, \mathbb{P})$ un espace probabilisé et soit $\{X_i, i = 1, ..., n\}$ une suite de variables aléatoires positives définies sur cet espace. Ces variables aléatoires (rv's) sont supposées indépendantes et identiquement distribuées (i.i.d), de fonction de répartition commune inconnue (f.r) F, et de fonction de densité de probabilité bornée (p.d.f) f. Le lemme précédent peut être adapté pour proposer un estimateur lisse de la fonction de distribution F, en remplaçant $u(x)$ par la fonction de répartition empirique $F_n(x)$ ce qui donne

$$F_n^+(x) = \int_0^{+\infty} F_n(t)\ dG_{n,x}(t). \tag{2.2}$$

Techniquement, $G_{x,n}$ peut avoir n'importe quel support mais il est préférable de le choisir de même support que la variable aléatoire considérée, car cela éliminera le problème que l'estimateur attribue une masse positive à la région indésirable. Pour que $F_n^+(x)$ soit une fonction de distribution, $G_{x,n}$ doit être une fonction décroissante de x , ce qui peut être montré en utilisant une forme alternative

de $F_n^+(x)$:

$$\begin{aligned} F_n^+(x) &= \int_0^{+\infty} \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{\{X_i \le t\}} \, dG_{n,x}(t) \\ &= \frac{1}{n} \sum_{i=1}^{n} \int_0^{+\infty} \mathbb{I}_{\{X_i \le t\}} \, dG_{n,x}(t) \\ &= \frac{1}{n} \sum_{i=1}^{n} \int_{X_i(\omega)}^{+\infty} dG_{n,x}(t), \quad \forall \omega \in \Omega \\ &= \frac{1}{n} \sum_{i=1}^{n} (1 - G_{n,x}(X_i)) \, . \\ &= 1 - \frac{1}{n} \sum_{i=1}^{n} G_{n,x}(X_i). \end{aligned} \tag{2.3}$$

De l'equation (2.3), on remarque que $G_{n,x}(.)$ doit être une fonction décroissante en x pour que $F_n^+(.)$ soit un estimateur correct de la fonction de répartition. Ceci nous amène à proposer un estimateur lisse de la densité donné par

$$f_n^+(x) = \frac{dF_n^+(x)}{dx} = -\frac{1}{n} \sum_{i=1}^{n} \frac{d}{dx} G_{n,x}(X_i). \tag{2.4}$$

En utilisant la représentation (2.3), si on choisit l'argument de la fonction $G_{n,x}(.)$, égal à $\dfrac{X_i}{x}$, qui est décroissant par rapport x, alors pour conserver l'égalité dans (2.3) on pose $G_{n,x}(t) = L_n\left(\dfrac{t}{x}\right)$, tel que $L_n(.)$ est une famille de distribution sur $[0, \infty)$,

- de moyenne $\mu_n := \displaystyle\int_0^{+\infty} u \; K_n(u) du \longrightarrow 1$, quand $n \longrightarrow \infty$
- et de variance $\upsilon_n^2 := \displaystyle\int_0^{+\infty} u(u-1) K_n(u) \; du \longrightarrow 0$ quand $n \to \infty$,

alors un estimateur de $F(x)$ est donné par

$$F_n^+(x) = 1 - \frac{1}{n} \sum_{i=1}^{n} L_n\left(\frac{X_i}{x}\right), \tag{2.5}$$

et si on dérive par rapport x, cela conduit à l'estimateur de densité suivant

$$\frac{d}{dx}\left(F_n^+(x)\right) = \frac{1}{n \; x^2} \sum_{i=1}^{n} X_i \; K_n\left(\frac{X_i}{x}\right),$$

où $K_n(.)$ désigne la densité correspondant à la fonction de distribution $L_n(.)$.
Cependant, l'estimateur ci-dessus peut ne pas être défini en $x = 0$, sauf dans les cas où $\lim\limits_{x \to 0} \dfrac{d}{dx}\left(F_n^+(x)\right)$

existe. De plus, cette limite est typiquement nulle, ce qui n'est acceptable que lorsque nous estimons une densité f avec $f(0) = 0$. Ainsi, en vue du cas plus général où $0 \leq f(0) < \infty$, nous considérons la version perturbée suivante de l'estimateur de densité ci-dessus :

$$f_n^+(x) = \frac{1}{n(x+a_n)^2} \sum_{i=1}^{n} X_i \; K_n\left(\frac{X_i}{x+a_n}\right), \quad x \geq 0 \tag{2.6}$$

où $a_n \longrightarrow 0$ est une suite de nombres positifs qui tend vers zéro quand $n \to \infty$.

Remarque 2.1. *L'équation (2.6) montre que l'estimateur de densité est la moyenne de variables aléatoires i.i.d.,* $Y_{in} = \dfrac{X_i}{(x+a_n)^2} \; K_n\left(\dfrac{X_i}{x+a_n}\right), \quad i = 1, 2, \ldots, n$. *Dans la suite, nous illustrons notre méthode en prenant* $L_n(.)$ *comme étant la fonction de distribution Gamma de paramètres* $\left(\alpha = \dfrac{1}{\upsilon_n^2}, \beta = \upsilon_n^2\right)$.

2.3 Analogie de l'estimateur de Chaubey avec d'autres estimateurs

Dans cette section nous présentons une comparaison de l'approche de Chaubey avec certains estimateurs existants.

Estimateur de Parzen (1956)-Rosenblat (1962)

L'estimateur à noyau usuel est un cas particulier de la représentation donnée par (2.4) , en prenant $G_{n,x}(.)$ comme

$$G_{n,x}(t) = K\left(\frac{X_i - x}{h_n}\right) \tag{2.7}$$

où $K(.)$ est une fonction de distribution de moyenne nulle et de variance égale 1.

Estimateur de transformation de Wand, Marron et Ruppert (1991)

L'approche célèbre de transformation logarithmique de Wand, Marron et Rupert (1991) [68] conduit à l'estimateur de densité suivant :

$$\tilde{f}_n^L(x) = \frac{1}{n \; h_n \; x} \sum_{i=1}^{n} k\left(\frac{1}{h_n} \; \log(X_i/x)\right),$$

où $k(.)$ est une fonction de densité (noyau) avec une moyenne nulle et une variance égale à 1. Ceci est facilement vue comme un cas particulier de l'equation (2.4), en reprenant $G_{n,x}(.)$ comme dans (2.7) mais appliqué à $\log x$.

Estimateur de Chen (2000) et Scaillet (2004)

L'estimateur de Chen (2000) [16] est de la forme

$$\tilde{f}_C(t) = \frac{1}{n}\sum_{i=1}^{n} g_{x,n}(X_i), \tag{2.8}$$

où $g_{x,n}(.)$, la densité de $G_{n,x}(.)$, est une Gamma de paramètres $(\alpha = \rho_b(x), \beta = b)$ avec $b \to 0$ et $b\ \rho_b(x) \to x$. Cela peut également être motivé par le lemme de lissage précédent à partir de l'équation (2.1), en prenant $u(t) = f(t)$ et donc l'intégrale $\int f(t)\ g_{n,x}(t)dt$ peut être estimée par $\frac{1}{n}\sum_{i=1}^{n} g_{x,n}(X_i)$.

2.4 Convergence forte uniforme

La variété des écritures des estimateurs $F_n^+(x)$, et $f_n^+(x)$ a permis d'étudier les propriétés asymptotiques de ces estimateurs. Premièrement, on montre que $F_n^+(x)$ est asymptotiquement sans biais, en effet

$$\begin{aligned} E(F_n^+(x)) &= E\left(\frac{1}{n}\sum_{i=1}^{n}\left(1 - L_n\left(\frac{X_i}{x}\right)\right)\right) \\ &= \int_0^{+\infty}\left(1 - L_n\left(\frac{z}{x}\right)\right)\ dF(z) \end{aligned}$$

On fait une intégration par partie et un changement de variable, on obtient

$$E(F_n^+(x)) = \int_0^{+\infty} F(xt)\ \ K_n(t)\ \ dt,$$

on suppose que la dérivée seconde de f est bornée, alors un développement de Taylor au voisinage de x nous donne

$$F(xt) = F(x) + x(t-1)\ f(x) + \frac{1}{2}\ x^2(t-1)^2\ f^{'}(x) + o((t-1)^2).$$

En remplaçant cela dans l'équation précédente, nous avons

$$E(F_n^+(x)) \approx F(x) + \frac{1}{2}\ x^2\ f^{'}(x)\upsilon_n^2.$$

Par conséquent, en supposant $\upsilon_n^2 \to 0$ quand $n \to \infty$, nous trouvons que l'estimateur $F_n^+(x)$ est asymptotiquement sans biais. Ainsi, Chaubey et *al* (2007) ont établi la convergence uniforme presque sure de l'estimateur $F_n^+(x)$ vers la fonction de répartition empirique par un choix approprié de υ_n^2, comme donné dans le théorème 2.1.

Théorème 2.1. *(Chaubey et al (2007), Théorème 2)*
Supposons que f possède une dérivée bornée, et supposons que $n\upsilon_n^2 = o(n^{-1})$, alors pour $\delta > 0$, on a pour tout $n \to +\infty$

$$\sup_{x\geq 0}|F_n^+(x) - F_n(x)| = O\left\{n^{-\frac{3}{4}\ (\log n)^{1+\delta}}\right\}.$$

Ainsi, pour l'estimateur asymétrique de la densité $f_n^+(x)$, on remarque que la formule donnée dans (2.6) est utile pour des objectifs calculatoires, cependant, pour l'étude des propriétés asymptotiques on utilise plutôt la représentation intégrale suivante,

$$f_n^+(x) = \int_0^{+\infty} F_n^+(t)\ \ \frac{d}{dx}\left[g_{n,x+a_n}(t)\right]\ dt \tag{2.9}$$

Les auteurs ont établi la consistance forte de l'estimateur $f_n^+(x)$, donnée par le théorème suivant.

Théorème 2.2. *(Chaubey et al (2007), Théorème 3)*
Sous les hypothèses suivantes

B0. $\lim_{n\to\infty} a_n = 0, \quad \lim_{n\to\infty} \upsilon_n = 0,$

B1. $\sup_{x\geq 0} \int_0^{+\infty} \left|\frac{d}{dx}\right| \left[g_{n,x+a_n}(t)\right] \Big| \ dt = o\left(\left(\frac{\log\log n}{n^{\frac{1}{2}}}\right)^{-1}\right)$

B2. $\sup_{u>0,\upsilon_n>0} uK_n(u) < +\infty,$

B3. *$f(.)$ est Lipschitzienne continue sur $[0, +\infty[$,*

on a quand $n \to +\infty$

$$\sup_{x\geq 0}|f_n^+(x) - f(x)| \xrightarrow{p.s} 0$$

Remarque 2.2. *On remarque que pour montrer la convergence forte les auteurs utilisent l'hypothèse **B2.** qui montre bien que la famille de noyaux $K_n(.)$ est asymétrique. Ainsi, L'hypothèse **B3.**, est une hypothèse de régularité de la fonction f, ce qui est courant dans l'estimation non paramétrique. Ils ont aussi utilisés quelques conditions supplémentaires sur les dérivées des densités $K_n(.)$. En ce qui concerne l'hypothèse **B1.**, par exemple, si on prend $g_{n,x}(t) = \frac{1}{x}K_n\left(\frac{t}{x}\right)$, avec $K_n(.)$ la densité de la loi gamma de paramètres $\Gamma(\alpha = 1/\upsilon_n^2, \beta = \upsilon_n^2)$, donné par,*

$$K_n(t) = \frac{1}{\beta^\alpha\ \Gamma(\alpha)}\ t^{\alpha-1}\ \exp\left\{-\frac{t}{\beta}\right\}, \qquad t > 0. \tag{2.10}$$

Dans ce cas

$$\begin{aligned} g_{n,x+a_n}(t) &= \frac{1}{x+a_n}\ K_n\left(\frac{t}{(x+a_n)}\right) \\ &= \frac{1}{(x+a_n)}\ \frac{1}{\beta^\alpha\ \Gamma(\alpha)}\ \left(\frac{t}{x+a_n}\right)^{\alpha-1}\ \exp\left\{-\frac{t}{\beta(x+a_n)}\right\} \\ &= \frac{t^{\alpha-1}}{\beta^\alpha\ \Gamma(\alpha)}\ \frac{1}{(x+a_n)^\alpha}\exp\left\{-\frac{t}{\beta(x+a_n)}\right\}. \end{aligned}$$

Ce qui implique

$$\begin{aligned} \frac{d}{dx}\left[g_{n,x+a_n}(t)\right] &= \frac{t^{\alpha-1}}{\beta^\alpha\ \Gamma(\alpha)}\left[\frac{-\alpha}{(x+a_n)^{\alpha+1}}\exp\left\{-\frac{t}{\beta(x+a_n)}\right\}+\frac{t}{\beta(x+a_n)^2}\exp\left\{-\frac{t}{\beta(x+a_n)}\right\}\right] \\ &= K_n\left(\frac{t}{x+a_n}\right)\left[\frac{t}{\beta(x+a_n)^3}-\frac{\alpha}{(x+a_n)^2}\right] \\ &= g_{n,x+a_n}(t)\left(\frac{t-\alpha\beta(x+a_n)}{\beta(x+a_n)^2}\right) \end{aligned}$$

Comme $\alpha = 1/\upsilon_n^2$, *et* $\beta = \upsilon_n^2$, *donc*

$$\begin{aligned} \int_0^{+\infty}\left|\frac{d}{dx}\left[g_{n,x+a_n}(t)\right]\right|dt &= \frac{1}{(x+a_n)^2\ \upsilon_n^2}\int_0^{+\infty}\left|t-(x+a_n)\right|\ g_{n,x+a_n}(t)dt \\ &= O\left(\frac{1}{(x+a_n)\upsilon_n}\right) \end{aligned}$$

de sorte que $\sup_{x\geq 0}\int_0^{+\infty}\left|\frac{d}{dx}\left[g_{n,x+a_n}(t)\right]\right|dt = O\left((a_n\upsilon_n)^{-1}\right)$, *et par un choix de* $a_n\upsilon_n = O\left(n^{-\frac{1}{2}+\delta}\right), 0 < \delta < \frac{1}{2}$., *la condition* ***B1.*** *est satisfaite.*

Preuve du Théorème 2.3 pour prouver le théorème 2.3, et pour des raisons techniques, on définit le pseudo estimateur $\tilde{f}(x)$ par

$$\tilde{f}(x) = \int_0^{+\infty} F(t)\ \ \frac{d}{dx}\left[g_{n,x+a_n}(t)\right]\ dt \tag{2.11}$$

et on considère la décomposition

$$\begin{aligned} f_n^+(x) - f(x) &= f_n^+(x) - \tilde{f}(x) + \tilde{f}(x) - f(x) \\ &= \int_0^{+\infty}\left(F_n^+(t) - F(t)\right)\frac{d}{dx}\left[g_{n,x+a_n}(t)\right]\ dt + \int_0^{+\infty} F(t)\frac{d}{dx}\left[g_{n,x+a_n}(t)\right]\ dt - f(x) \end{aligned} \tag{2.12}$$

or que $G_{n,x}(t) = L_n\left(\frac{t}{x}\right)$, ce qui implique que

$$\frac{d}{dt}\ L_n\left(\frac{t}{x}\right) = \frac{1}{x}\ K_n\left(\frac{t}{x}\right) = g_{n,x}(t).$$

D'une part par intégration par partie, on obtient

$$\tilde{f}(x) = \int_0^{+\infty} \frac{t}{(x+a_n)^2}\ K_n\left(\frac{t}{x+a_n}\right) dt = E\left(f_n^+(x)\right) \tag{2.13}$$

et d'autre part

$$\begin{aligned} E\Big(f_n^+(x)\Big) &= E\left(\frac{1}{n(x+a_n)^2}\sum_{i=1}^{n} X_i\ K_n\left(\frac{X_i}{x+a_n}\right)\right) \\ &= E\left(\frac{X_1}{(x+a_n)^2}\ K_n\left(\frac{X_1}{x+a_n}\right)\right) \\ &= \int_0^{+\infty} \frac{t}{(x+a_n)^2}\ K_n\left(\frac{t}{x+a_n}\right)\ \ f(t)\ dt \\ &= \int_0^{+\infty} z\ K_n(z)\ \ f\Big(z(x+a_n)\Big)\ \ dz. \end{aligned} \tag{2.14}$$

ainsi puisque $\int_0^{+\infty} z\ K_n(z)\ \ dz = 1$, l'equation (2.12) est

$$f_n^+(x) - f(x) = \int_0^{+\infty} \Big(F_n^+(t) - F(t)\Big)\ \ \frac{d}{dx}\left[g_{n,x+a_n}(t)\right]\ dt + \int_0^{+\infty} \Big(f\Big(t(x+a_n)\Big) - f(x)\Big)\ \ t\ K_n(t)\ \ dt.$$

et donc

$$|f_n^+(x) - f(x)| \leq \sup_{t\geq 0}|F_n^+(t) - F(t)| \int_0^{+\infty} \left|\frac{d}{dx}\right| \left[g_{n,x+a_n}(t)\right] \Big| dt + \int_0^{+\infty} |f\Big(t(x+a_n)\Big) - f(x)|\ \ t\ K_n(t)\ \ dt.$$

D'après le théorème 2.1 et l'hypothèse **B1.**, le premier terme de la dernière inégalité converge presque surement vers zero. Et le deuxième terme converge aussi vers zero comme on peut le voir comme suit. Pour toute $M > 0$

$$\begin{aligned} \sup_{x\geq 0} \int_0^{+\infty} |f\Big(t(x+a_n)\Big) - f(x)|\ \ t\ K_n(t)\ \ dt &= \max\left\{\sup_{0\leq x\leq M} \int_0^{+\infty} |f\Big(t(x+a_n)\Big) - f(x)|\ \ t\ K_n(t)\ \ dt,\right. \\ &\qquad \left.\sup_{x\geq M} \int_0^{+\infty} |f\Big(t(x+a_n)\Big) - f(x)|\ \ t\ K_n(t)\ \ dt\right\} \\ &=: \max\left\{\Delta_M, \theta_M\right\} \end{aligned}$$

par le changement de variable $z = t(x + a_n)$, et $\forall \varepsilon > 0$, on peut obtenir $M > 0$ tel que

$$\begin{aligned}\theta_M &\leq \sup_{x\geq M} \int_0^{+\infty} f(z)\ \frac{1}{x+a_n}\ \left(\frac{z}{x+a_n}\right)\ K_n\left(\frac{z}{x+a_n}\right) + \sup_{x\geq M} f(x) \int_0^{+\infty} z\ K_n(z)\ dz \\ &\leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2}\end{aligned}$$

en utilisant le théorème de la convergence dominée(par l'hypothèse **B2.**) pour le premier terme. De plus, sous l'hypothèse **B3.** nous avons par l'inégalité de Cauchy-Schwartz,

$$\begin{aligned}\Delta_M &\leq \int_0^{+\infty} |t(M-a_n) - M| t\ K_n(t)\ dt \\ &= M \int_0^{+\infty} |t-1| t\ K_n(t)\ dt + a_n \int_0^{+\infty} t^2\ K_n(t)\ dt \\ &\leq M \sqrt{\int_0^{+\infty} (t-1)^2\ K_n(t)\ dt} \sqrt{\int_0^{+\infty} t^2\ K_n(t)\ dt} + O(a_n) \\ &= MO(\upsilon_n) + O(a_n) = o(1).\end{aligned}$$

2.5 Normalité asymptotique

Le théorème suivant donne les conditions, sur K_n et f, sous lesquelles l'estimateur de la densité est asymptotiquement normal et donne la forme de sa variance asymptotique.

Théorème 2.3. *(Chaubey et al (2007), Théorème 4)*
Soient les hypothèses suivantes

B0′. *$f(.)$ est lipschitzienne continue sur $[0, +\infty[$,*

B1′. *Soit* $I_{n,l} := \int_0^{+\infty} K_n^l(w)dw = O(\upsilon_n^{-(l-1)})$, *si* $\upsilon_n \to 0$, *pour* $1 \leq l \leq 3$, *et* $\overline{M} := \lim_{n\to+\infty} \upsilon_n\ I_{n,2}$,

B2′. *Avec* $K_{n,l}^*(u) := \dfrac{K_n^l(u)}{I_{n,l}}$, *pour* $1 \leq l \leq 3$, *et quand* $\upsilon_n \to 0$, *on a*

(i) $\mu_{l,\upsilon_n} := \int_0^{+\infty} uK_{n,l}^*(u)du = 1 + O(\upsilon_n^2)$,

(ii) $\sigma_{l,\upsilon_n}^2 := \int_0^{+\infty} (u - \mu_{l,\upsilon_n})^2 K_{n,l}^*(u)du = O(\upsilon_n^2)$,

(iii) $\sup_{0<\upsilon_n<\epsilon} \int_0^{+\infty} u^{4+\delta} K_{n,l}^*(u)du < +\infty$, *pour* $\delta > 0, \epsilon > 0$,

B3′. *(i)* $\lim_{n\to\infty} n\upsilon_n = \infty, \quad \lim_{n\to\infty} n\upsilon_n a_n = \infty,$

(ii) $\lim_{n\to\infty} n\upsilon_n^3 = 0, \quad \lim_{n\to\infty} n\upsilon_n a_n^2 = 0,$

Sous B0′-B3′, on a

$$a)\sqrt{n\upsilon_n}\left(f_n^+(x) - f(x)\right) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \overline{M}\frac{f(x)}{x}\right), \text{pour } x > 0,$$

$$b)\sqrt{n\upsilon_n a_n}\left(f_n^+(0) - f(0)\right) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \overline{M}f(0)\right).$$

Remarque 2.3. *Il est clair que les conditions* ***B2', (i)-(iii)****, signifient ce qui suit : soit $T_{n,l}^*$ une suite de variables aléatoires de densité $K_{n,l}^*$, alors que $T_{n,l}^* \longrightarrow 1$, dans l'espace L_p, pour $1 \leq p \leq 4$, quand $\upsilon_n \longrightarrow 0$, pour tout $l = 1, 2, 3$.*

Remarque 2.4. *Nous illustrons les conditions* ***B1′*** *et* ***B2′*** *avec $K_n(t)$, la densité Gamma comme donnée dans l'equation 2.10. Pour $l \geq 1$, en utilisant les propriétés de la fonction gamma, on aura*

$$\begin{aligned} I_{n,l} &:= \int_0^{+\infty} K_n^l(w)dw \\ &= \frac{1}{\left[\Gamma\left(\frac{1}{\upsilon_n^2}\right)\right]^l (\upsilon_n^2)^{\frac{l}{\upsilon_n^2}}} \int_0^{+\infty} w^{\frac{l}{\upsilon_n^2}-l} \exp\left\{-\frac{l\ w}{\upsilon_n^2}\right\} dw \\ &= \frac{\Gamma\left(\frac{l}{\upsilon_n^2} - l + 1\right)}{\left[\Gamma\left(\frac{1}{\upsilon_n^2}\right)\right]^l (\upsilon_n^2)^{\frac{l}{\upsilon_n^2}} \left(\frac{l}{\upsilon_n^2}\right)^{\frac{l}{\upsilon_n^2}-l+1}} \end{aligned}$$

et

$$K_{n,l}^*(w) = \frac{1}{\left(\frac{\upsilon_n^2}{l}\right)^{\frac{l}{\upsilon_n^2}-l+1} \Gamma\left(\frac{l}{\upsilon_n^2} - l + 1\right)} w^{\frac{l}{\upsilon_n^2}-l+1-1} \exp\left\{-\frac{l\ w}{\upsilon_n^2}\right\}, \qquad w > 0.$$

qui est une densité de Gamma de paramètres $\alpha = \frac{l}{\upsilon_n^2} - l + 1$, $\beta = \frac{\upsilon_n^2}{l}$. En utilisant l'approximation de Stirling pour la fonction Gamma, on obtient facilement pour tout $l \geq 1$

B1′.
- $I_{n,l} := \int_0^{+\infty} K_n^l(w)dw \approx \frac{1}{\sqrt{l(2\pi)^{l-1}}} . \frac{1}{\upsilon_n^{l-1}\sqrt{1-\upsilon_n^2}},$
- $\overline{M} := \lim_{n\to+\infty} \upsilon_n I_{n,2} = \frac{1}{\sqrt{4\pi}} < \infty,$

B2′. (i) $\mu_{l,\upsilon_n} = 1 - \left(\frac{l-1}{l}\right)\upsilon_n^2,$

B2′. (ii) $\sigma_{l,\upsilon_n}^2 = \left(1 - \left(\frac{l-1}{l}\right)\upsilon_n^2\right)\frac{\upsilon_n^2}{l},$

B2′. (iii) $\forall k \geq 1$, *et* $\forall \epsilon > 0$,

$$\sup_{0<\upsilon_n<\epsilon} \int_0^{+\infty} u^k K^*_{n,l}(u)du = \sup_{0<\upsilon_n<\epsilon} \frac{\Gamma\left(k+\frac{l}{\upsilon_n^2}-l+1\right)}{\frac{l}{\upsilon_n^2}\Gamma\left(\frac{l}{\upsilon_n^2}-l+1\right)} = O\left(1+\left(\frac{k-l+1}{l}\right)\upsilon_n^2\right) < \infty$$

2.6 Erreur quadratique moyenne intégrée MISE

Dans cette section, nous étudions le critère d'erreur quadratique moyenne intégrée (MISE) de $f_n^+(x)$, qui se défini comme

$$\mathbf{MISE}(f_n^+(x)) := \int \mathbf{MSE}(f_n^+(x))dx. \tag{2.15}$$

Où **MSE** désigne l'erreur quadratique moyenne défini par :

$$\mathbf{MSE} = \Big(Biais\Big(f_n^+(x)\Big)\Big)^2 + Var\Big(f_n^+(x)\Big). \tag{2.16}$$

2.6.1 Calcul du biais

Le calcul du biais repose essentiellement sur des développements de Taylor, ce qui nous conduit à poser certaines conditions de régularité sur la fonction f et ses dérivées qui détermineront l'ordre du biais asymptotique en fonction des paramètres υ_n^2 et a_n. D'après l'équation 2.14, nous avons

$$E\Big(f_n^+(x)\Big) = \int_0^{+\infty} z\ K_n(z)\ \ f\Big(z(x+a_n)\Big)\ \ dz. \tag{2.17}$$

Un développement de Taylor d'ordre 1 autour de x, permet d'obtenir

$$f\Big(z(x+a_n)\Big) = f(x) + \Big(x(z-1)+z\ a_n\Big)f'(x) + o\Big(x(z-1)+za_n\Big),$$

et

$$\begin{aligned}
E\Big(f_n^+(x)\Big) &= \int_0^{+\infty} z\ K_n(z)\left\{f(x) + \Big(x(z-1)+z\ a_n\Big)f'(x) + o\Big(x(z-1)+za_n\Big)\right\}dz \\
&= f(x) + xf'(x)\int_0^{+\infty} z(z-1)\ K_n(z)dz + a_nf'(x)\int_0^{+\infty} z^2\ K_n(z)dz \\
&\quad + o\left(\int_0^{+\infty} z\left[x(z-1)+z\ a_n\right]\ K_n(z)dz\right) \\
&= f(x) + \Big((x+a_n)\upsilon_n^2 + a_n\Big)f'(x) + o\Big(\upsilon_n^2+a_n\Big).
\end{aligned} \tag{2.18}$$

et donc le biais de $f_n^+(x)$ est tel que

$$Biais\Big(f_n^+(x)\Big) = \Big((x+a_n)\upsilon_n^2 + a_n\Big)f'(x) + o\Big(\upsilon_n^2+a_n\Big). \tag{2.19}$$

2.6.2 Calcul de la variance

Maintenant, pour la variance, nous avons

$$Var\Big(f_n^+(x)\Big) = \frac{1}{n(x+a_n)^4}\ Var\left(X_1\ K_n\left(\frac{X_1}{x+a_n}\right)\right),$$

et

$$\begin{aligned} Var\left(X_1 K_n\left(\frac{X_1}{x+a_n}\right)\right) &= E\left(X_1^2 K_n^2\left(\frac{X_1}{x+a_n}\right)\right) - E^2\left(X_1 K_n\left(\frac{X_1}{x+a_n}\right)\right) \\ &=: V_1 - V_2. \end{aligned}$$

À partir de (2.18) et (2.6), on obtient

$$\begin{aligned} V_2 &= \Big(n(x+a_n)^2 E\left(f_n^+(x)\right)\Big)^2 \\ &= (x+a_n)^4\left[f(x) + \Big((x+a_n)\upsilon_n^2 + a_n\Big)f'(x) + o\Big(a_n^2+a_n\Big)\right]^2 \\ = n^2 &\quad (x+a_n)^4\left(f^2(x) + o\Big(a_n^2+a_n\Big)\right). \end{aligned}$$

Pour V_1 nous avons

$$\begin{aligned} V_1 &= E\left(\ X_1^2\ K_n^2\left(\frac{X_1}{x+a_n}\right)\right) \\ &= \int_0^{+\infty} t^2 K_n^2\left(\frac{t}{x+a_n}\right) f(t)\ dt. \end{aligned}$$

$$\begin{aligned} V_1 &= E\left(\ X_1^2\ K_n^2\left(\frac{X_1}{x+a_n}\right)\right) \\ &= \int_0^{+\infty} t^2 K_n^2\left(\frac{t}{x+a_n}\right) f(t)\ dt. \end{aligned}$$

Un changement de variable et un développement de Taylor impliquent

$$\begin{aligned} V_1 &= (x+a_n)^3 \int_0^{+\infty} z^2 K_n^2(z) f\Big(z(x+a_n)\Big)dz = (x+a_n)^3 I_{n,2} \int_0^{+\infty} z^2 K_{n,2}^*(z) f\Big(z(x+a_n)\Big)dz \\ &= (x+a_n)^3 I_{n,2} \int_0^{+\infty} z^2 K_{n,2}^*(z)\left\{f(x) + \Big(z(x+a_n)-x\Big)f'(x) + o\Big(z(x+a_n)-x\Big)\right\}dz \\ &= (x+a_n)^3 I_{n,2}\left\{f(x)\mu_{2,\upsilon_n}\left(K_n^*\right) + f'(x)\Big((x+a_n)\ \mu_{3,\upsilon_n}\left(K_n^*\right) - x\ \mu_2\left(K_n^*\right)\Big)\right. \\ &\quad \left. + o\Big((x+a_n)\ \mu_{3,\upsilon_n}\left(K_n^*\right) - x\ \mu_{2,\upsilon_n}\left(K_n^*\right)\Big)\right\} \\ &= I_{n,2}(x+a_n)^3\Big(f(x) + o\left(\upsilon_n^2 + a_n\right)\Big). \end{aligned}$$

D'où

$$Var\Big(f_n^+(x)\Big) = \frac{\overline{M}f(x)}{n(x+a_n)\upsilon_n} + o\bigg(\frac{1}{n\ \upsilon_n}\bigg),$$

Par conséquent, en combinant les formules ci-dessus, nous obtenons l'erreur quadratique moyenne de $f_n^+(x)$

$$\mathbf{MSE}(f_n^+(x)) = \frac{\overline{M}f(x)}{n(x+a_n)\upsilon_n} + \bigg[\Big((x+a_n)\upsilon_n^2 + a_n\Big)f'(x)\bigg]^2 + o\bigg(\frac{1}{n\ \upsilon_n}\bigg) + o\Big(\upsilon_n^2 + a_n\Big) \qquad (2.20)$$

est l'erreur quadratique moyenne intégrée et donnée par

$$\mathbf{MISE}(f_n^+(x)) = \frac{\overline{M}}{n\upsilon_n}\int_0^\infty \frac{f(x)}{(x+a_n)}dx + \int_0^\infty \Big[(x\upsilon_n^2 + a_n)f^{'}(x)\Big]^2 dx + o(\upsilon_n^2 + a_n) + o((n\upsilon_n)^{-1}). \quad (2.21)$$

Chapitre 3

Consistance des estimateurs à noyaux non symétriques : Cas des données Censurée

Sommaire

3.1 Introduction

On se place dans ce chapitre dans le contexte de données censurées à droite. Les résultats énoncés ici font partie d'un article qui a été publié dans la revue Communications in Statistics : Theory and Methods Ghettab et Guessoum (2022) [37]. Nous proposons de nouveaux estimateurs en utilisant un noyau non symétrique pour diminuer le biais aux bornes. Dans un premier temps, on définit un estimateur lisse de la fonction de répartition, qui est une extension des travaux du Chaubey et *al.* (2007) dans le cas censuré à droite. La dérivée de ce dernier nous donne un nouvel estimateur de la densité, et comme conséquence nous proposons deux versions d'estimation de la fonction taux de hasard : la première en utilisant l'estimateur de Kaplan-Meier pour estimer la fonction de survie, et la deuxième en utilisant l'estimateur à noyau non symétrique. La vitesse de convergence uniforme presque sûre sur un intervalle fermé et borné est établie pour ces estimateurs. Une large étude de simulation est menée pour conforter les résultats théoriques suivie d'une application sur un jeu de données réelles.

3.2 Estimateurs à noyaux non symétriques

Soit $(\Omega, \mathcal{B}, \mathbb{P})$ un espace probabilisé et soit $\{X_i, i = 1, ..., n\}$ une suite de variables aléatoires représentant des durée de vie définies sur $(\Omega, \mathcal{B}, \mathbb{P})$. Ces variables aléatoires (rv's) sont supposées indépendantes identiquement distribuées (i.i.d), de fonction de distribution commune inconnue (f.r) F, et de fonction de densité de probabilité bornée (p.d.f) f. Nous supposons que les durées de survie ne peuvent pas être observées complètement, et sont censurées à droite par une séquence de variables i.i.d $\{C_i, i = 1, ..., n\}$ de f.r continue G, qui sont indépendantes des durées de survie X_i. Les données observées seront les couples $\{(Y_i, \delta_i), i = 1, ..., n\}$ avec

$$Y_i = \min(X_i, C_i) \quad \text{et} \quad \delta_i = \mathbb{I}_{\{X_i \leq C_i\}}.$$

Ainsi les Y_i sont des variables aléatoires i.i.d et δ_i est l'indicateur de non censure. Pour définir convenablement un estimateur de la densité $f(.)$ pour les v.a. X_i dans le cas censuré quand $x \in [0, \infty[$, rappelons que d'après le théorème 2.1 de Chaubey et *al.* (2013), on peut définir un estimateur lisse de la fonction de répartition F par

$$\widehat{F}_n(x) = \int_0^{\infty} F_n(t) d\widetilde{L}_{n,x}(t), \tag{3.1}$$

où $F_n(t)$ est l'estimateur produit limite (P.L) introduit par Kaplan et Meier (1958), défini par

$$F_n(x) = \begin{cases} 1 - \prod\limits_{i:Y_{(i)} \leq x} \left(\dfrac{n-i}{n-i+1} \right)^{\delta_{(i)}} & \text{si} \quad x < Y_{(n)} \\ 1 & \text{si} \quad x \geq Y_{(n)}, \end{cases} . \tag{3.2}$$

avec $Y_{(1)}, Y_{(2)}, \ldots, Y_{(n)}$, désignant les statistiques d'ordre de $Y_1, Y_2, \ldots, Y_n$ et $\delta_{(i)}$ est l'indicateur de non censure correspondant à $Y_{(i)}$. La suite $\left\{ \widetilde{L}_{n,x}, n = 1, 2, \ldots \right\}$ est une famille de distributions de

moyenne $\mu_n(x)$ et de variance υ_n^2 tendant vers x et 0 respectivement quand n tend vers $+\infty$. On peut aussi définir l'estimateur de Kaplan-Meier (K.M) de G, par :

$$G_n(x) = 1 - \prod_{i:Y_{(i)}\leq x} \left(\frac{n-i}{n-i+1}\right)^{1-\delta_{(i)}}. \tag{3.3}$$

On remarque que l'estimateur (P.L) $F_n(x)$ peut être réécrit sous la forme :

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{\delta_i}{\left(1 - G_n\left(Y_i^-\right)\right)}\, \mathbb{I}_{\{Y_i\leq x\}}, \tag{3.4}$$

où la somme est prise sur les i tels que $1 - G_n\left(Y_i^-\right) =: \bar{G}_n(Y_i^-) \neq 0$ (voir l'annexe). Ainsi, (3.1) et (3.4) impliquent que

$$\widehat{F}_n(x) = 1 - \frac{1}{n}\sum_{i=1}^{n} \frac{\delta_i}{\left(1 - G_n\left(Y_i^-\right)\right)}\, \widetilde{L}_{n,x}(Y_i).$$

Pour que $\widehat{F}_n(x)$ soit une fonction de répartition, $\widetilde{L}_{n,x}$ doit être une fonction décroissante sur $x \in \mathbb{R}^+$, donc nous proposons d'estimer F sur $[0, +\infty[$ par

$$\widehat{F}_n(x) = 1 - \frac{1}{n}\sum_{i=1}^{n} \frac{\delta_i}{\left(1 - G_n\left(Y_i^-\right)\right)}\, L_n\left(\frac{Y_i}{x+a_n}\right), \tag{3.5}$$

où (L_n) est une famille de fonctions de distribution sur $[0, +\infty[$, de moyenne 1 et de variance υ_n^2, telle que $L_n\left(\dfrac{t}{x+a_n}\right) = \widetilde{L}_{n,x+a_n}(t)$. Si l'on suppose que $L_n(.)$ admet une fonction de densité $K_n(.)$, nous obtenons de (3.5) un nouvel estimateur pour la densité pour f, donné par :

$$\widehat{f}_n(x) = \frac{1}{n(x+a_n)^2}\sum_{i=1}^{n} \frac{\delta_i Y_i}{\bar{G}_n(Y_i^-)}\, K_n\left(\frac{Y_i}{x+a_n}\right). \tag{3.6}$$

À notre connaissance l'estimateur $\widehat{f}_n(x)$ n'a pas été considéré et étudié auparavant. De plus, à partir de (3.4), (3.5) et (3.6), nous proposons d'estimer la fonction de hasard $\lambda(x)$ en utilisant d'abord l'estimateur de Kaplan-Meier $F_n(x)$ par,

$$\widehat{\lambda}(x) = \frac{\widehat{f}_n(x)}{1 - F_n(x)}, \tag{3.7}$$

puis l'estimateur lisse $\widehat{F}_n(x)$,

$$\widetilde{\lambda}(x) = \frac{\widehat{f}_n(x)}{1 - \widehat{F}_n(x)}. \tag{3.8}$$

et nous étudions leurs propriétés asymptotiques.

3.3 Vitesse de convergence uniforme presque sûr des estimateurs

On définit $\tau_F = \sup\{x : \bar{F}(x) > 0\}$ et $\tau_G = \sup\{x : \bar{G}(x) > 0\}$ les bornes supérieures des supports de $\bar{F} := 1 - F$ et $\bar{G} := 1 - G$, respectivement. Dans la pratique, τ_F désigne par exemple la date limite des observations de X. On supposera ici que F et G sont des f.r continues, et on considérera la convergence sur un intervalle compact $J = [0, \tau], \tau < \tau_F$. Pour établir la convergence uniforme presque sûr p.s. de l'estimateur de la fonction de distribution lisse $\widehat{F}_n(.)$ (Théorème 3.1), l'estimateur de la densité $\hat{f}_n(.)$ (Théorème 3.2) et des estimateurs de la fonction de hasard $\widehat{\lambda}_n(.)$ et $\widetilde{\lambda}_n(.)$ (corollaire 3.1), nous avons besoins des hypothèses suivantes :

3.3.1 Hypothèses

H0. La censure C est indépendante de la vraie durée de survie X, et $\tau_F < \infty, \bar{G}(\tau_F) > 0$.

H1. K_n est une famille de fonctions définies sur $\mathbb{R}^+$, telle que pour $n \geq 1$, $K_n(.)$ est bornée et lipschitzienne continue sur J, et satisfait

(i) $\int_0^{+\infty} u\ K_n(u)du = 1,$

(ii) $\int_0^{+\infty} u(u-1)K_n(u)\ du =: \upsilon_n^2 \longrightarrow 0 \quad \text{quand} \quad n \to \infty.$

H2. Soit $I_n := \int_0^{+\infty} K_n^2(w)dw$, $\lim_{n\to+\infty} M_n := \lim_{n\to+\infty} \upsilon_n I_n < +\infty$, et $K_n^*(u) := \dfrac{K_n^2(u)}{I_n}$, de moyenne

$\mu_j(K_n^*) := \int_0^{+\infty} u^j K_n^*(u)du = 1 + O(\upsilon_n^2), \quad 1 \leq j \leq 3.$

H3. f et $\bar{f}(t) := \dfrac{f(t)}{\bar{G}(t)}$ sont deux fois continûment dérivables sur J tel que f' est bornée et $\|\Phi''\|_\infty := \sup_{x\in J}|\Phi''(x)| < +\infty,$ où Φ représente soit la fonction f soit la fonction $\bar{f}$.

H4. (i) $\lim_{n\to\infty} a_n = 0, \quad \lim_{n\to\infty} \upsilon_n = 0,$

(ii) $\lim_{n\to\infty} \dfrac{a_n}{\upsilon_n^2} = 0,$

(iii) $\lim_{n\to\infty} \sqrt{\dfrac{\log n}{n a_n^3}} = 0.$

Commentaires sur les Hypothèses

1. **H0** est généralement utilisée dans le cadre de la censure. Noter que $\bar{G}(\tau_F) > 0$ implique que $\tau_F \leq \tau_G$.

2. L'hypothèse **H1 (i)** signifie que la famille des noyaux $K_n(.)$ a une espérance égale à 1, et **H1 (ii)** signifie que $K_n(.)$ admet une variance finie υ_n^2 tend vers 0, ce qui permet de calculer le terme de biais.
3. **H2** assure l'existence de termes de variance asymptotique et **H3** est une hypothèse de régularité des fonctions f et $\bar{f}$, ce qui est courant dans l'estimation non paramétrique.
4. L'hypothèse **H4** est nécessaire pour déterminer le taux de convergence forte uniforme. Remarquons que (i), (ii) et (iii) impliquent $\lim_{n\to\infty}\sqrt{\frac{\log n}{n a_n \upsilon_n}} = 0$.

3.3.2 Quelques exemples de noyaux asymétriques

Dans cette section, nous donnons quelques exemples explicites de noyaux asymétriques K_n satisfaisant **H1** et **H2**.

Noyau Gamma

Soit $K_n(.)$ la densité de la loi gamma $\Gamma(1/\upsilon_n^2, \upsilon_n^2)$. Pour toute suite υ_n^2, **H1** est vérifiée

$$\begin{aligned} I_n = \int_0^{+\infty} K_n^2(w)dw &= \frac{1}{\left[\Gamma(1/\upsilon_n^2)\right]^2 (\upsilon_n^2)^{\frac{2}{\upsilon_n^2}}} \int_0^{+\infty} w^{2/\upsilon_n^2 - 2} e^{-\frac{2}{\upsilon_n^2}w} dw \\ &= \frac{\Gamma(\frac{2}{\upsilon_n^2} - 1)}{(2/\upsilon_n^2)^{\frac{2}{\upsilon_n^2}-1} \left[\Gamma(1/\upsilon_n^2)\right]^2 (\upsilon_n^2)^{\frac{2}{\upsilon_n^2}}} \\ &= \frac{\Gamma(\frac{2}{\upsilon_n^2} - 1)}{\upsilon_n^2 2^{\frac{2}{\upsilon_n^2}-1} \left[\Gamma(1/\upsilon_n^2)\right]^2}. \end{aligned}$$

Ce qui implique que $K_n^*(u)$ est la densité de la distribution Gamma $\Gamma(1/\upsilon_n^2 - 1, \upsilon_n^2/2)$. Par conséquent, en utilisant l'approximation de Stirling, nous obtenons

- $\lim_{n\to+\infty} \upsilon_n I_n = \frac{1}{\sqrt{4\pi}} < \infty$,
- $\mu_j(K_n^*) = \begin{cases} 1 - \frac{\upsilon_n^2}{2} & \text{si} \quad j = 1, 2 \\ 1 - \frac{\upsilon_n^4}{4} & \text{si} \quad j = 3 \end{cases} = 1 + O(\upsilon_n^2).$

Cela signifie que **H2** est vérifiée.

Noyau Bêta

Soit $K_n(.)$ la densité de la loi Bêta sur $]0, +\infty[$ $\beta(2/\upsilon_n^2 + 1, 2/\upsilon_n^2 + 2)$. Pour toute suite υ_n^2, **H1** est vérifiée et

$$I_n = \int_0^{+\infty} K_n^2(w)dw = \frac{\beta\left(\frac{4}{\upsilon_n^2} + 1, \frac{4}{\upsilon_n^2} + 5\right)}{\beta^2\left(\frac{2}{\upsilon_n^2} + 1, \frac{2}{\upsilon_n^2} + 3\right)}.$$

avec $\beta(x,y) = \dfrac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}$ est la fonction Bêta, ce qui implique que $K_n^*(u)$ est la densité d'une distribution Bêta de paramètres $\dfrac{4}{\upsilon_n^2}+1$, et $\dfrac{4}{\upsilon_n^2}+5$ respectivement, et on obtient

- $\lim\limits_{n\to+\infty} \upsilon_n I_n = \dfrac{2^{-1}e^{15}}{\sqrt{4\pi}} < \infty,$
- $\mu_j(K_n^*) = \begin{cases} 1-\dfrac{3\upsilon_n^2}{4+4\upsilon_n^2} \underset{\upsilon_n^2\to 0}{\sim} 1-\dfrac{3\upsilon_n^2}{4} & \text{si} \quad j=1,2 \\ 1-\dfrac{5\upsilon_n^4+8\upsilon_n^2}{6\upsilon_n^4+14\upsilon_n^2+8} \underset{\upsilon_n^2\to 0}{\sim} 1-\dfrac{8\upsilon_n^2}{8} & \text{si} \quad j=3 \end{cases} = 1+O(\upsilon_n^2)$

H2 est aussi satisfaite.

Noyau Log-Normal

Soit $K_n(.)$ la densité de la loi Log-Normal *Log-N*$(-\upsilon_n^2/2, \upsilon_n)$, **H1** est vérifiée et pour **H2** on a

- $\lim\limits_{n\to+\infty} \upsilon_n I_n = \dfrac{1}{\sqrt{4\pi}} < \infty,$
- $\mu_j(K_n^*) = \begin{cases} e^{-\frac{3}{4}\upsilon_n^2} \underset{\upsilon_n^2\to 0}{\sim} 1-\dfrac{3}{4}\upsilon_n^2 & \text{si} \quad j=1,3 \\ e^{-\upsilon_n^2} \underset{\upsilon_n^2\to 0}{\sim} 1-\upsilon_n^2 & \text{si} \quad j=2 \end{cases} = 1+O(\upsilon_n^2)$

3.3.3 Résultats principaux

Nous avons maintenant les résultats principaux suivantes.

Théorème 3.1. *Sous les hypothèses* ***H0, H1, H2, H3*** *et* ***H4*** *, on a, quand* $n \longrightarrow \infty$

$$\sup_{x\in J}|\widehat{F}_n(x) - F(x)| = O_{a.s}\left\{\sqrt{\frac{\log n}{n}} + \upsilon_n^2\right\}.$$

Théorème 3.2. *Sous les hypothèses* ***H0, H1, H2, H3*** *et* ***H4*** *, on a, quand* $n \longrightarrow \infty$

$$\sup_{x\in J}|\widehat{f}_n(x) - f(x)| = O_{a.s}\left\{\sqrt{\frac{\log n}{na_n\upsilon_n}} + \upsilon_n^2\right\}.$$

Corollaire 3.1. *Sous les hypothèses* ***H0, H1, H2, H3*** *et* ***H4*** *, on a, quand* $n \longrightarrow \infty$

$$a)\ \sup_{x\in J}|\widehat{\lambda}_n(x) - \lambda(x)| = O_{a.s}\left\{\sqrt{\frac{\log n}{na_n\upsilon_n}} + \upsilon_n^2\right\}.$$

$$b)\ \sup_{x\in J}|\widetilde{\lambda}_n(x) - \lambda(x)| = O_{a.s}\left\{\sqrt{\frac{\log n}{na_n\upsilon_n}} + \upsilon_n^2\right\}.$$

3.4 Simulations et étude sur données réelles

Pour étudier les performances des estimateurs proposés dans (3.5), (3.6), (3.7) et (3.8) sur un échantillon fini , et faire des comparaisons avec celui basé sur un noyau symétrique (noyau gaussien) défini dans (3.9) ci-dessous, dans la première partie nous avons réalisé une étude de simulation par la Méthode de Monte-Carlo pour examiner le comportement des estimateurs des deux fonctions de la densité et du taux de hasard dans diverses situations, et la deuxième partie illustre les méthodes avec une application à des données réelles.

3.4.1 Simulation par la Méthode de Monte-Carlo

L'algorithme suivant résume les étapes les plus importantes :
Étape 1 : Trois distributions sont considérées pour générer la variable temps de survie X_i, avec une taille d'échantillon $n = 50, 100$ et 500 :

- La loi exponentielle de paramètre $\lambda_1 = 1$ ($\mathcal{E}(1)$).
- La loi log-normale avec un paramètre d'échelle $a = 0$ et un paramètre de forme $b = 1,5$ ($Log\text{-}N(0, 1.5)$).
- La loi de Weibull avec paramètre d'échelle $a = 4.3$ et un paramètre de forme $b = 1.3$ ($W(4.3, 1.3)$).

Étape 2 : La variable temps de censure est générée à partir d'une loi exponentielle de paramètre λ_2, où λ_2 est ajusté selon le pourcentage de censure souhaité.
Étape 3 : Nous calculons les données observées $Y_i = \min(X_i, C_i)$, $\delta_i = \mathbb{I}_{\{X_i \leq C_i\}}$, et l'estimateur de Kaplan-Meier du temps de censure $\bar{G}_n(.)$ défini dans (3.3), avec une légère modification pour éviter de prendre la valeur zéro (voir Marron et Padgett (1987)),

$$\bar{G}_n(t) = \begin{cases} 1 \quad 0 \leq t \leq Y_{(1)} \\ \prod_{i=1}^{k-1} \left(\frac{n-i+1}{n-i+2}\right)^{1-\delta_{(i)}}, \quad Y_{(k-1)} < t \leq Y_{(k)}, k = 2, \ldots, n \\ \prod_{i=1}^{n} \left(\frac{n-i+1}{n-i+2}\right)^{1-\delta_{(i)}}, t > Y_{(n)}. \end{cases}$$

Étape 4 : Nous calculons l'estimateur à noyau asymétrique défini dans (3.6), en choisissant un noyau Log-normal $Log\text{-}N(-\upsilon_n^2/2, \upsilon_n)$. Afin de comparer les effets aux bornes de l'estimateur à noyau asymétrique avec celui basé sur le noyau symétrique , nous considérons l'estimateur de densité de Blum et Susarla (1980) $\hat{f}_{BS}(t)$donné par

$$\hat{f}_{BS}(t) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{t - Y_i}{h}\right) \frac{\delta_i}{\bar{G}_n(Y_i)}, \tag{3.9}$$

ainsi en l'absence de censure, (3.9) se réduit à l'estimateur de densité de noyau habituel dans (1). On calcule $\hat{f}_{BS}(.)$ en choisissant le noyau gaussien.
Étape 5 : Noter que la principale difficulté ici est le choix des paramètres υ_n^2 et a_n. Il existe plusieurs méthodes optimales pour les sélectionner. Les plus populaires sont l'erreur quadratique moyenne globale (en anglais global mean square error(GMSE)) et les critères de validation croisée. Le but ici n'est pas de donner une expression théorique des paramètres optimaux, nous allons plutôt

les calculez en sélectionnant la paire (υ_n^{2*}, a_n^*) qui minimise l'erreur quadratique moyenne globale (GMSE), dans une grille arbitraire de valeurs $\mathcal{L} = \{\upsilon_n^2 = 10^{-2} + 16.5(k-1)10^{-3},\ k = 1, 2, ..., 31\} \times \{a_n = 10^{-3} + 14.95(k-1)10^{-3},\ k = 1, 2, ..., 21\}$, en respectant l'hypothèse H4.
Pour l'estimateur à noyau symétrique, nous choisissons $h_n = O\left((1/n)^{1/5}\right)$ (voir Kuhnt et *al.* 1997).
Étape 6 : Pour les estimateurs de la fonction de hasard (3.7) et (3.8), nous suivons les étapes 1 à 5 et nous calculons l'estimateur de Kaplan-Meier de la variable temps de survie $F_n(.)$ dans (3.2).
Étape 7 : Pour faire des comparaisons numériques, nous choisissons le critère (GMSE) : Nous répétons B fois les simulations décrites dans les étapes 1 à 6 pour chaque combinaison fixé de taille d'échantillon n et de pourcentage de censure PC.
Pour une fonctionnel donnée g et son estimation $\hat{g}_n$, le GMSE d'une paire de paramètres $(\upsilon_n^2, a_n) \in [0.01, 0.5] \times [0.001, 0.3]$, est défini par

$$GMSE = \frac{1}{BH} \sum_{k=1}^{B} \sum_{j=1}^{H} \left(\hat{g}_{n,k}(x_j) - g(x_j)\right)^2,$$

où H est le nombre de points équidistants x_j appartenant à l'intervalle $[0, 3]$ pour les distributions exponentielles et log-normales et à $[0, 10]$ pour la distribution de Weibull. Ici nous prenons $H = 101$ valeurs et $\hat{g}_{n,k}(x_j)$ est la valeur de $\hat{g}_n$ calculée à l'itération k. La valeur optimale du $GMSE$, avec les paramètres globaux correspondants (υ_n^{2*}, a_n^*), pour les estimations de la fonction de densité et de la fonction taux de hasard, est indiquée dans Tableaux 3.1 - 3.4.
Étape 8 : Pour afficher la qualité d'ajustement des estimateurs, nous avons tracé la médiane des estimateurs dans les cas asymétriques et symétriques pour les paramètres optimaux globaux (υ_n^{2*}, a_n^*) donné dans le tableau 3.2 ci-dessous, obtenu à partir de 200 répétitions, avec la vraie courbe de densité pour référence, également pour la fonction taux de hasard, nous utilisons le tableau 3.4 pour choisir les paramètres optimaux.
Les résultats basés sur plusieurs tailles d'échantillons et plusieurs pourcentages de censure sont présentés dans les figures 3.1-3.3 et les figures 3.4-3.6 pour les fonctions de densité et taux de hasard respectivement.

TABLE 3.1 – Le $GMSE$ de $\widehat{f}_n$ et $\hat{f}_{BS}$ en utilisant le noyau *Log-N* et le noyau gaussien respectivement.

Distribution	PC	$n = 50$		$n = 100$		$n = 500$	
		Log-N	Gauss	*Log-N*	Gauss	*Log-N*	Gauss
$Exp(1)$	20%	0.0080	0.0132	0.0055	0.0110	0.0021	0.0075
	40%	0.0101	0.0143	0.0066	0.0124	0.0023	0.0081
	60%	0.0159	0.0182	0.0099	0.0140	0.0032	0.0089
Log-N$(0, 1.5)$	20%	0.0054	0.0099	0.0035	0.0078	0.0011	0.0052
	40%	0.0056	0.0106	0.0035	0.0088	0.0011	0.0053
	60%	0.0076	0.0124	0.0047	0.0100	0.0014	0.0063
$W(4.3, 1.3)$	20%	0.00064	0.00066	0.00043	0.00045	0.00014	0.00023
	40%	0.00077	0.00078	0.00045	0.00055	0.00014	0.00025
	60%	0.0011	0.0012	0.00064	0.00076	0.0002	0.00032

TABLE 3.2 – Les paramètres optimaux (v_n^{2*}, a_n^*) pour l'estimateur de la fonction de densité asymétrique $\widehat{f}_n$.

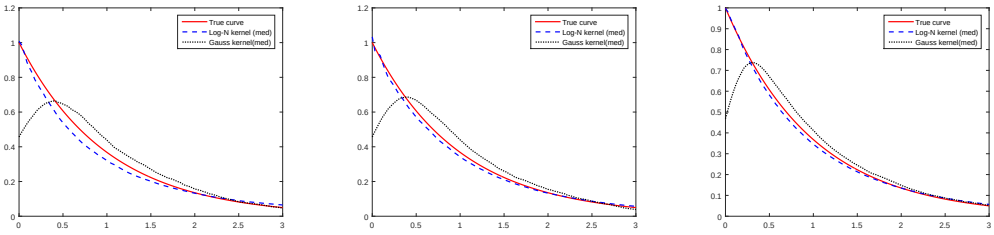
Distribution	PC	$n = 50$	$n = 100$	$n = 500$
$Exp(1)$	20%	$(0.3635, 0.0510)$	$(0.2209, 0.0617)$	$(0.1403, 0.0442)$
	40%	$(0.3511, 0.0372)$	$(0.2767, 0.0485)$	$(0.1403, 0.0298)$
	60%	$(0.4628, 0.0488)$	$(0.3759, 0.0269)$	$(0.1837, 0.0199)$
Log-N$(0, 1.5)$	20%	$(0.4566, 0.0010)$	$(0.3325, 0.0010)$	$(0.1961, 0.0010)$
	40%	$(0.4318, 0.0010)$	$(0.3449, 0.0010)$	$(0.2271, 0.0010)$
	60%	$(0.4876, 0.0010)$	$(0.3263, 0.0010)$	$(0.2271, 0.0010)$
$W(4.3, 1.3)$	20%	$(0.2395, 0.1490)$	$(0.1961, 0.0077)$	$(0.1341, 0.0010)$
	40%	$(0.1775, 0.1410)$	$(0.2209, 0.0010)$	$(0.0906, 0.0010)$
	60%	$(0.3697, 0.0010)$	$(0.1961, 0.0010)$	$(0.1216, 0.0010)$

TABLE 3.3 – Le $GMSE$ des estimateurs de la fonction de taux de hasard en utilisant le noyau *Log-N* et le noyau gaussien.

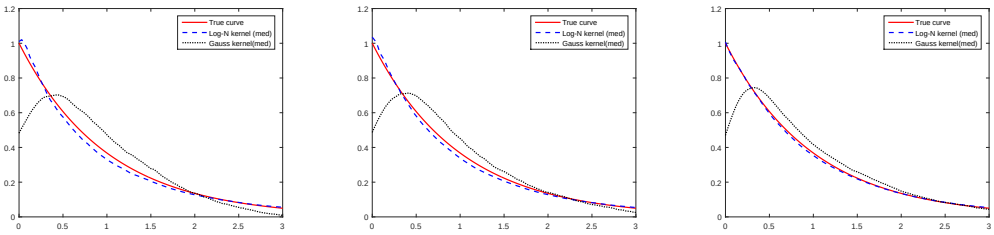
Distribution	Estim	$n = 50$			$n = 100$			$n = 500$		
		20%	40%	60%	20%	40%	60%	20%	40%	60%
$Exp(1)$	$\widehat{\lambda}_{Log\text{-}N}$	0.1112	0.1591	0.3187	0.0698	0.1137	0.2396	0.0185	0.0373	0.1145
	$\widetilde{\lambda}_{Log\text{-}N}$	0.0641	0.0613	0.0875	0.0453	0.0412	0.0562	0.0167	0.0177	0.0217
	$\widehat{\lambda}_{Gauss}$	0.3101	0.3680	0.5207	0.1629	0.2652	0.4728	0.0622	0.1134	0.3382
$Log\text{-}N(0, 1.5)$	$\widehat{\lambda}_{Log\text{-}N}$	0.0152	0.0170	0.0324	0.0090	0.0095	0.0198	0.0025	0.0027	0.0057
	$\widetilde{\lambda}_{Log\text{-}N}$	0.0133	0.0181	0.0190	0.0073	0.0141	0.0157	0.0022	0.0027	0.0127
	$\widehat{\lambda}_{Gauss}$	0.0566	0.0898	0.1555	0.0324	0.0546	0.1260	0.0137	0.0215	0.0582
$W(4.3, 1.3)$	$\widehat{\lambda}_{Log\text{-}N}$	0.0113	0.0139	0.0257	0.0072	0.0103	0.0184	0.0017	0.0030	0.0080
	$\widetilde{\lambda}_{Log\text{-}N}$	0.0063	0.0068	0.0081	0.0041	0.0044	0.0055	0.0015	0.0017	0.0022
	$\widehat{\lambda}_{Gauss}$	0.0211	0.0255	0.0397	0.0138	0.0157	0.0296	0.0033	0.0062	0.0189

TABLE 3.4 – Les paramètres optimaux (v_n^{2*}, a_n^*) des estimateurs de la fonction de taux de hasard asymétrique $\widehat{\lambda}_{Log\text{-}N}$ et $\widetilde{\lambda}_{Log\text{-}N}$.

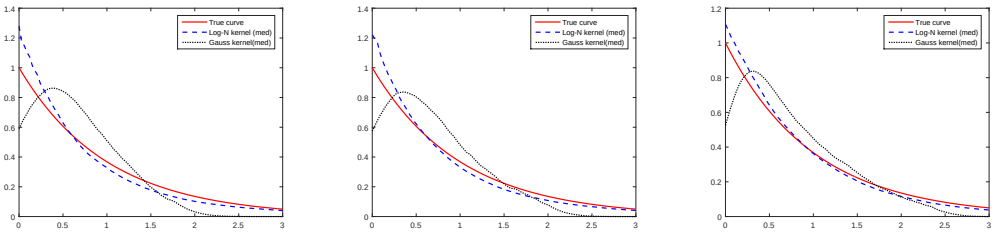
Distribution		$n = 50$		$n = 100$		$n = 500$	
		$\widehat{\lambda}_{Log\text{-}N}$	$\widetilde{\lambda}_{Log\text{-}N}$	$\widehat{\lambda}_{Log\text{-}N}$	$\widetilde{\lambda}_{Log\text{-}N}$	$\widehat{\lambda}_{Log\text{-}N}$	$\widetilde{\lambda}_{Log\text{-}N}$
$Exp(1)$	20%	(0.3387, 0.0825)	(0.1278, 0.0360)	(0.1527, 0.0638)	(0.1154, 0.0207)	(0.0720, 0.0279)	(0.0534, 0.0118)
	40%	(0.4690, 0.0333)	(0.1775, 0.0497)	(0.3325, 0.0239)	(0.1154, 0.0444)	(0.1589, 0.0173)	(0.1216, 0.0093)
	60%	(0.4566, 0.0167)	(0.3263, 0.0459)	(0.4194, 0.0154)	(0.3139, 0.0442)	(0.3139, 0.0334)	(0.1341, 0.0606)
$Log\text{-}N(0, 0.5)$	20%	(0.4566, 0.0010)	(0.500, 0.0010)	(0.3449, 0.0010)	(0.4380, 0.0010)	(0.2209, 0.0010)	(0.2209, 0.0010)
	40%	(0.4876, 0.0010)	(0.4380, 0.1517)	(0.4628, 0.0010)	(0.4938, 0.0180)	(0.2333, 0.0010)	(0.4628, 0.0010)
	60%	(0.4814, 0.0010)	(0.4380, 0.2270)	(0.4380, 0.0010)	(0.4132, 0.1858)	(0.3015, 0.0010)	(0.3822, 0.1587)
$W(4.3, 1.3)$	20%	(0.2023, 0.1190)	(0.1154, 0.0010)	(0.1465, 0.0562)	(0.0720, 0.0108)	(0.0410, 0.0383)	(0.0383, 0.0033)
	40%	(0.3015, 0.0217)	(0.1775, 0.0010)	(0.1899, 0.1508)	(0.1030, 0.0116)	(0.0782, 0.0330)	(0.0534, 0.0010)
	60%	(0.3822, 0.0404)	(0.2643, 0.0010)	(0.4690, 0.0010)	(0.1589, 0.0282)	(0.2147, 0.1631)	(0.0968, 0.0175)



(a) $n = 50, PC \backsimeq 20\%$ (b) $n = 100, PC \backsimeq 20\%$ (c) $n = 500, PC \backsimeq 20\%$

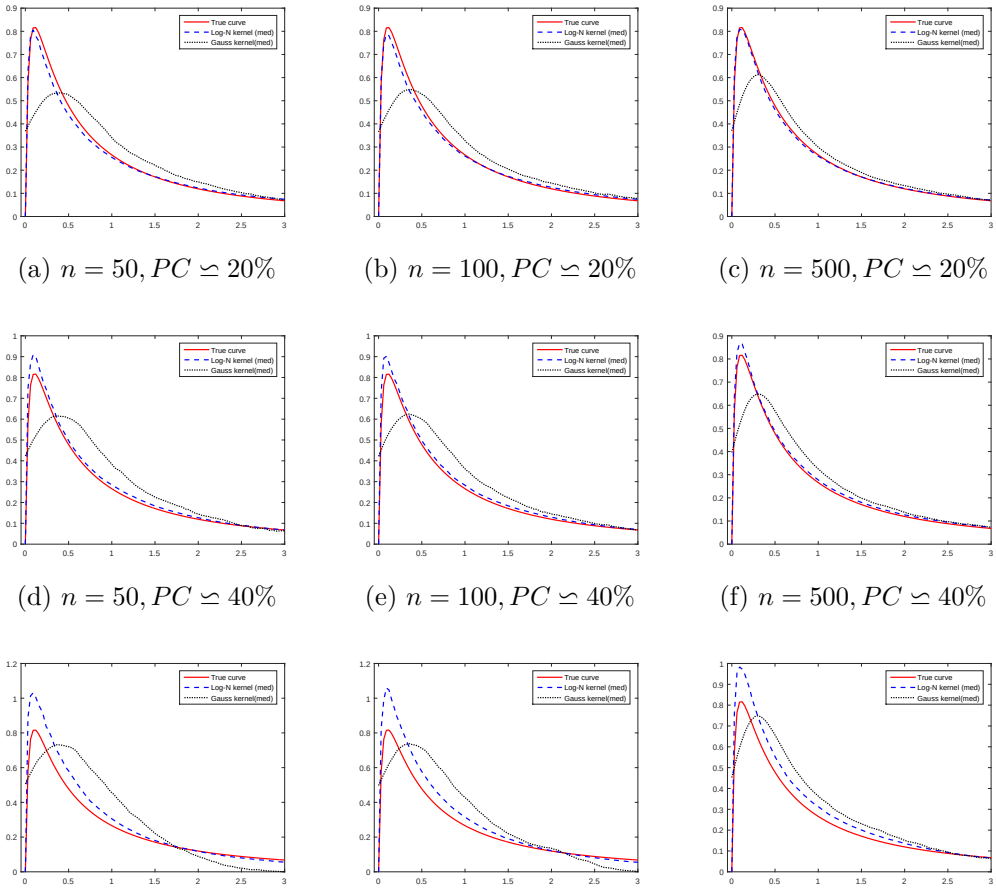


(d) $n = 50, PC \backsimeq 40\%$ (e) $n = 100, PC \backsimeq 40\%$ (f) $n = 500, PC \backsimeq 40\%$



(g) $n = 50, PC \backsimeq 60\%$ (h) $n = 100, PC \backsimeq 60\%$ (i) $n = 500, PC \backsimeq 60\%$

FIGURE 3.1 – Densité exponentielle $\mathcal{E}(1)$. Vraie fonction (ligne continue) et densités estimées par le noyau *Log-N* (ligne discontinue) et le noyau gaussien (ligne pointillée).

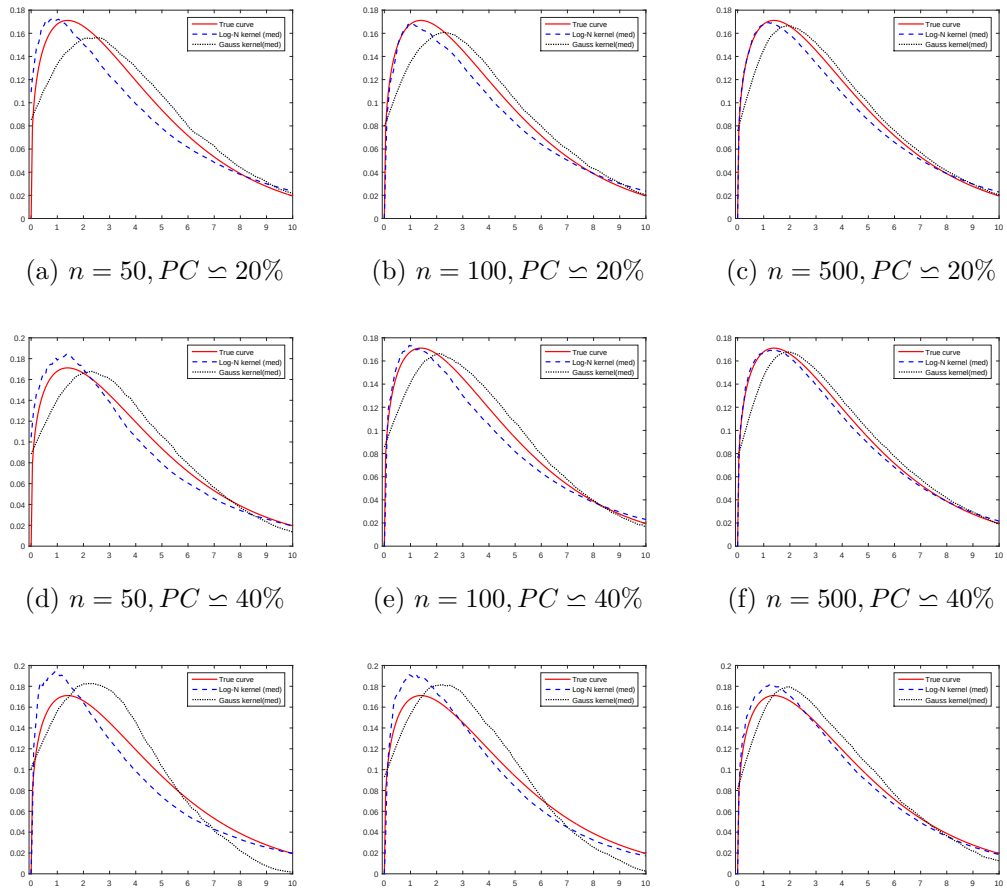


(a) $n = 50, PC \backsimeq 20\%$ (b) $n = 100, PC \backsimeq 20\%$ (c) $n = 500, PC \backsimeq 20\%$

(d) $n = 50, PC \backsimeq 40\%$ (e) $n = 100, PC \backsimeq 40\%$ (f) $n = 500, PC \backsimeq 40\%$

(g) $n = 50, PC \backsimeq 60\%$ (h) $n = 100, PC \backsimeq 60\%$ (i) $n = 500, PC \backsimeq 60\%$

FIGURE 3.2 – Densité Log-normal $Log - N(0, 1.5)$. Vraie fonction (ligne continue) et densités estimées par le noyau *Log-N* (ligne discontinue) et le noyau gaussien (ligne pointillée).

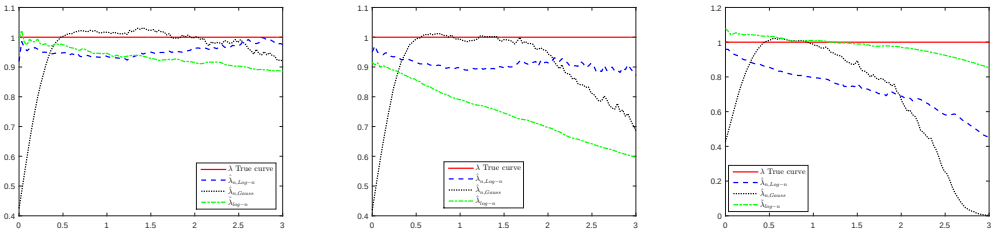


(a) $n = 50, PC \backsimeq 20\%$ (b) $n = 100, PC \backsimeq 20\%$ (c) $n = 500, PC \backsimeq 20\%$

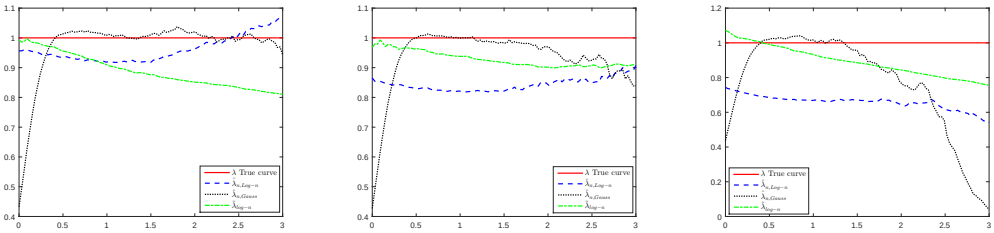
(d) $n = 50, PC \backsimeq 40\%$ (e) $n = 100, PC \backsimeq 40\%$ (f) $n = 500, PC \backsimeq 40\%$

(g) $n = 50, PC \backsimeq 60\%$ (h) $n = 100, PC \backsimeq 60\%$ (i) $n = 500, PC \backsimeq 60\%$

FIGURE 3.3 – Densité de Weibull $W(4.3, 1.3)$. Vraie fonction (ligne continue) et densités estimées par le noyau *Log-N* (ligne discontinue) et le noyau gaussien (ligne pointillée)

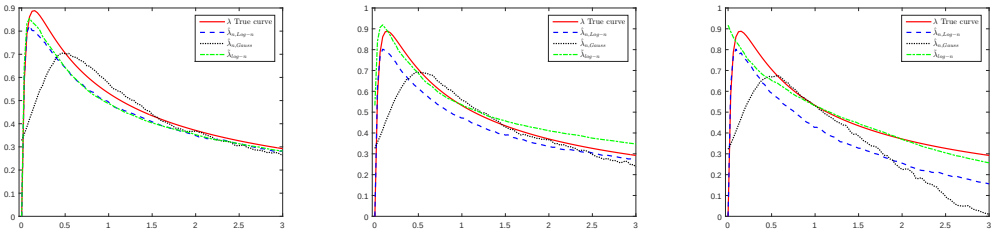


(a) $n = 500, PC \backsimeq 20\%$ (b) $n = 500, PC \backsimeq 40\%$ (c) $n = 500, PC \backsimeq 60\%$

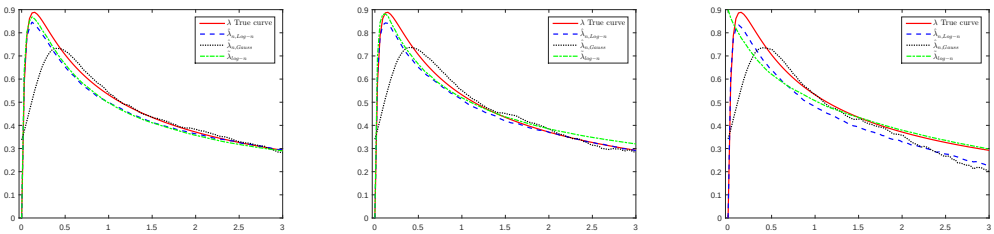


(d) $n = 1000, PC \backsimeq 20\%$ (e) $n = 1000, PC \backsimeq 40\%$ (f) $n = 1000, PC \backsimeq 60\%$

FIGURE 3.4 – Fonction de hasard pour la loi $Exp(1)$. Vraie fonction (ligne continue) et les fonctions estimée par le noyau *Log-N* $\widehat{\lambda}(.)$ (ligne continue), $\widetilde{\lambda}(.)$ (ligne discontinue), et le noyau gaussien (ligne pointillée).

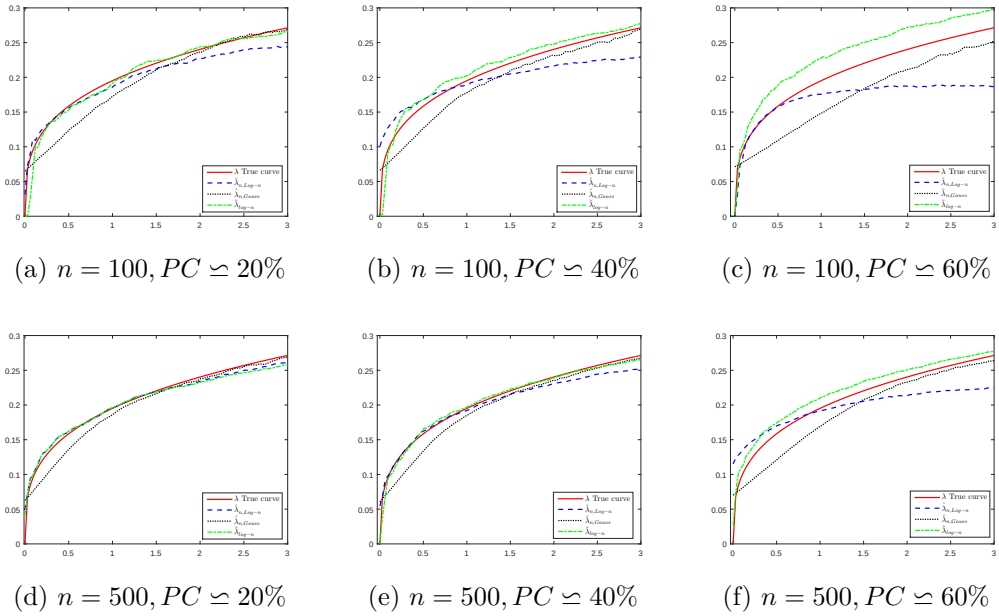


(a) $n = 100, PC \backsimeq 20\%$ (b) $n = 100, PC \backsimeq 40\%$ (c) $n = 100, PC \backsimeq 60\%$



(d) $n = 500, PC \backsimeq 20\%$ (e) $n = 500, PC \backsimeq 40\%$ (f) $n = 500, PC \backsimeq 60\%$

FIGURE 3.5 – Fonction de hasard de la loi *Log-N*$(0, 1.5)$. Vrai fonction (ligne continue) et les fonction estimée par le noyau *Log-N* $\widehat{\lambda}(.)$ (ligne discontinue), $\widetilde{\lambda}(.)$ (ligne trait pointillé), et par le noyau gaussien (ligne pointillé).



(a) $n = 100, PC \backsimeq 20\%$ (b) $n = 100, PC \backsimeq 40\%$ (c) $n = 100, PC \backsimeq 60\%$

(d) $n = 500, PC \backsimeq 20\%$ (e) $n = 500, PC \backsimeq 40\%$ (f) $n = 500, PC \backsimeq 60\%$

FIGURE 3.6 – Fonction de hasard de la loi $W(4.3, 1.3)$. Vraie fonction (ligne continue) et la fonction estimée par le noyau *Log-N* $\widehat{\lambda}(.)$ (ligne discontinue), $\widetilde{\lambda}(.)$ (ligne trait pointillé), et le noyau gaussien (ligne pointillé).

Commentaire sur les résultats de simulation

Les résultats des expériences de simulation indiquent que la GMSE diminue lorsque la taille de l'échantillon augmente et cela est vrai pour les estimateurs à noyau asymétrique et symétrique. La qualité de l'estimation est affecté par le pourcentage de censure comme on peut le voir dans les tableaux 3.1-3.4 avec cependant un net avantage pour l'estimateur à noyau asymétrique. D'après les FIGURES 3.1-3.6, on peut voir que les estimateurs proposés sont vraiment meilleurs près des bornes et moins affectés par la censure par rapport aux estimateurs à noyau symétrique, donc notre méthode est beaucoup plus efficace que la méthode classique. De plus, les deux estimateurs de la fonction taux de hasard $\widehat{\lambda}_n$ et $\widetilde{\lambda}_n$ sont très proches l'un de l'autre pour un échantillon de grande taille, comme le montrent les FIGURES 3.4-3.6 et la Table 3.3, par exemple pour la densité de la loi log-Normal avec 20% de censure et $n = 500$ les valeurs du GMSE sont $0,0025$ et $0,0022$ respectivement, par rapport au noyau gaussien ($0,0137$), la différence est vraiment moins prononcée. Sur la base concrète de tous les résultats ci-dessus, l'utilisation de l'approche du noyau asymétrique est une bonne alternative pour traiter l'effet du biais aux bornes, pour les densités à support $[0, +\infty)$.

3.4.2 Application sur un jeu de données réelles : mélanome malin (cancer de la peau)

Dans cette section, nous analysons un ensemble de données réelles pour illustrer l'efficacité de l'estimateur à noyau asymétrique en présence de données censurées. De plus, nous comparons avec l'estimateur à noyau gaussien. Les données consistent en des mesures effectuées sur des patients atteints de mélanome malin (cancer de la peau) entre 1962 et 1977. Chaque patient a subi une opération radicale à l'hôpital universitaire d'Odense, au Danemark. Cela signifie que la tumeur a été complètement retirée avec la peau à une distance d'environ 2,5 cm autour d'elle. Les 71 données étudiées ici (converties en années pour simplifier), ont été extraites d'un ensemble de 205 données de survie au mélanome (voir Andersen et al. (1993), Tableau A1 p. 709). Parmi les 71 données sélectionnées, 57 d'entre elles représentent la durée de vie réelle X, c'est-à-dire le temps écoulé entre l'opération et le décès par cancer de la peau (indiqué par 1 parmi les 205 observations du tableau A1 d'Andersen et al. (1993)). Les données restantes (14) sont censurées et correspondent à des patients décédés d'une cause indépendante du cancer de la peau (indiqué par 3 dans le tableau A1 d'Andersen et al. (1993)). La durée moyenne de survie est de 3,479 ans (allant de 0,027 à 9,474 ans) et le pourcentage de censure PC $\approx 24,5\%$.
Dans la FIGURE 3.7 nous avons tracé l'histogramme des 57 données de la variable X, en prenant 7 classes, afin d'avoir une idée sur la forme de la distribution des données observées . Ce dernier suggère une distribution de Weibull. Le test d'ajustement de Kolmogorov-Smirnov ne rejette pas l'hypothèse selon laquelle les données proviennent de $W(4.25, 1.25)$ avec une p-value égale à 0,52.

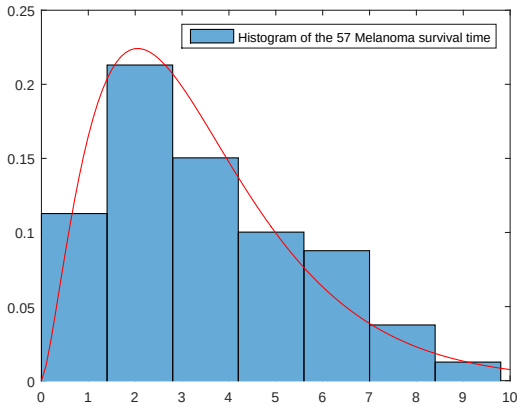
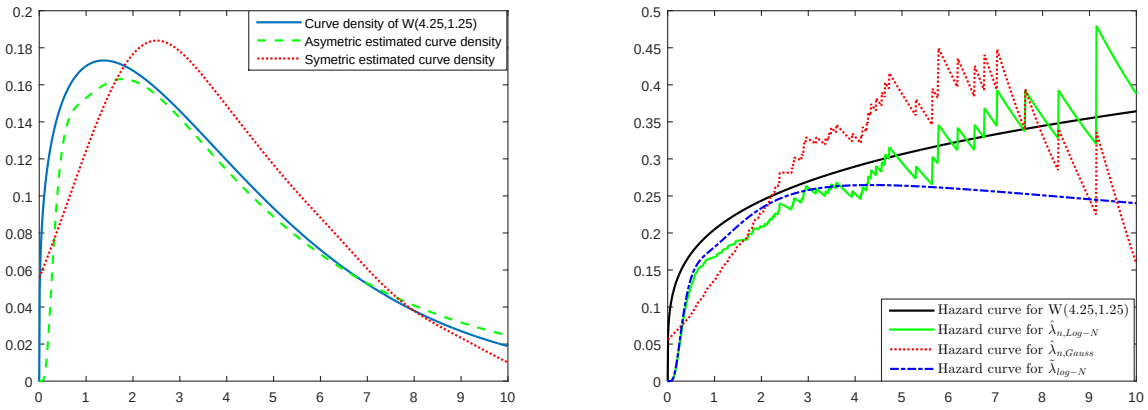


FIGURE 3.7 – Histogramme du temps de survie réel (57 données).

Les estimateurs à noyau asymétrique et symétrique de la densité et des fonctions de hasard, en utilisant les 71 données observées, sont représentés sur la FIGURE 3.8 avec la distribution $W(4.25, 1.25)$.



(a) (b)

FIGURE 3.8 – La fonction de densité W(4.25,1.25) (a) et la fonction de hasard (b) et les estimateurs à noyau asymétrique et symétrique correspondants.

Ici encore, nous voyons que l'estimateur à noyau asymétrique fonctionne bien pour la fonction de densité et la fonction taux de hasard.

3.5 Démonstrations des résultats

3.5.1 Preuve du Théorème 3.1

De (3.5) nous pouvons réécrire $\widehat{F}_n(x)$ comme suit

$$\widehat{F}_n(x) = \sum_{i=1}^{n} \frac{\delta_i}{n\bar{G}_n(Y_i^-)} \left(1 - L_n\left(\frac{Y_i}{x+a_n}\right)\right) =: \sum_{i=1}^{n} \frac{\delta_i}{n\bar{G}_n(Y_i^-)} \bar{L}_n\left(\frac{Y_i}{x+a_n}\right),$$

et définir

$$\widetilde{F}_n(x) = \sum_{i=1}^{n} \frac{\delta_i}{n\bar{G}(Y_i)} \bar{L}_n\left(\frac{Y_i}{x+a_n}\right),$$

L'inégalité triangulaire permet d'écrire

$$\left|\widehat{F}_n(x) - F(x)\right| \leq \left|\widehat{F}_n(x) - \widetilde{F}_n(x)\right| + \left|\widetilde{F}_n(x) - E(\widetilde{F}_n(x))\right| + \left|E(\widetilde{F}_n(x)) - F(x)\right|.$$

Étape 1. Nous prouvons que

$$\sup_{x\in J}\left|\widetilde{F}_n(x) - E(\widetilde{F}_n(x))\right| = O_{a.s}\left(\sqrt{\frac{\log n}{n}}\right). \tag{3.10}$$

Pour cela, considérons la suite indépendante et identiquement distribuée (i.i.d) $(Y_1,\delta_1),\ldots,(Y_n,\delta_n)$, et définissons la classe de fonctions :

$$\mathcal{F}_n = \left\{\psi_{n,x} : J\times\{0,1\} \longrightarrow \mathbb{R}^+/ \quad \psi_{n,x}(t,\delta) = \frac{\delta}{n\bar{G}(t)}\bar{L}_n\left(\frac{t}{x+a_n}\right); x\in J\right\}.$$

D'après le lemme 3*b*) de Giné et Guillou (1999), $\mathcal{F}_n$ est une VC (Vapnik-$\check{C}$ervonenkis) classe de fonctions mesurables et uniformément bornées, d'enveloppe $\psi(n,\delta) = \dfrac{\delta}{n\bar{G}(t)}$. Il est claire que,

$$\sum_{i=1}^{n} \{\psi_{n,x}(Y_i,\delta_i) - E\left(\psi_{n,x}(Y_1,\delta_1)\right)\} = \widetilde{F}_n(x) - E(\widetilde{F}_n(x)).$$

De plus

$$\sup_{\psi_{n,x}\in\mathcal{F}_n} \|\psi_{n,x}(t,\delta)\|_\infty = \sup_{\substack{(t,\delta)\in \\ \mathbb{R}^+\times\{0,1\}}} \left|\frac{2\delta}{n\bar{G}(t)}\right| \leq \frac{2}{n\bar{G}(\tau_F)} =: U_n$$

De l'hypothèse **H0**, nous avons $\bar{G}(t) \geq \bar{G}(\tau_F) > 0$, alors

$$\begin{aligned}
\sup_{x\in J} Var\left(\psi_{n,x}(Y_1,\delta_1)\right) &\leq \sup_{x\in J} E\left(\psi^2_{n,x}(Y_1,\delta_1)\right) \\
&= E\left(E\left(\frac{\mathbb{I}_{\{X_1\leq C_1\}}}{n^2\bar{G}^2(X_1)}\bar{L}^2\left(\frac{X_1}{x+a_n}\right)|X_1\right)\right) \\
&= \int_0^{+\infty} \frac{\bar{L}^2\left(\frac{t}{x+a_n}\right)}{n^2\bar{G}^2(t)} E\left(\mathbb{I}_{\{t\leq C_1\}}|X_1=t\right) f(t)dt \\
&\leq \frac{1}{n^2\bar{G}(\tau_F)}\int_0^{+\infty} f(t)dt \\
&\leq \frac{1}{n^2\bar{G}(\tau_F)} =: \sigma_n^2
\end{aligned}$$

En appliquant l'inégalité de Talagrand (Giné et Guillou (2002), Théorème (2.1) avec $t = M\sqrt{\dfrac{\log n}{n}}$, où M est une constante positive), il existe deux constantes positives M_1 et M_2 telles que

$$\begin{aligned}
P &\left\{\sup_{\psi_{n,x}\in\mathcal{F}_n}\left|\sum_{i=1}^{n}\psi_{n,x}(Y_i,\delta_i) - E(\psi_{n,x}(Y_i,\delta_i))\right| \geq M\sqrt{\frac{\log n}{n}}\right\} \\
&\leq M_1 \exp\left\{-\frac{M\sqrt{n\log n}\bar{G}(\tau_F)}{M_1}\log\left[1 + \frac{M\dfrac{1}{n\bar{G}(\tau_F)}\sqrt{\dfrac{\log n}{n}}}{M_1\left(\dfrac{1}{\sqrt{n\bar{G}(\tau_F)}} + \dfrac{1}{n\bar{G}(\tau_F)}\sqrt{\log\dfrac{M_2}{\bar{G}(\tau_F)}}\right)^2}\right]\right\} \\
&\leq M_1 \exp\left\{-\frac{M\bar{G}_{(\tau_F)}\sqrt{n\log n}}{M_1}\log\left[1 + \frac{\dfrac{M}{M_1}\sqrt{\frac{\log n}{n}}}{\left(1+\sqrt{\dfrac{1}{n\bar{G}(\tau_F)}\log\dfrac{M^2}{\bar{G}(\tau_F)}}\right)^2}\right]\right\}.
\end{aligned}$$

En faisant maintenant une approximation de la fonction logarithme et de la fonction racine carrée au voisinage de zéro, nous obtenons

$$\begin{aligned} P & \left\{ \sup_{\psi_{n,x}\in\mathcal{F}_n} \left| \sum_{i=1}^{n} \{\psi_{n,x}(Y_1,\delta_1) - E(\psi_{n,x}(Y_1,\delta_1))\} \right| \geq M\sqrt{\frac{\log n}{n}} \right\} \\ \leq & \; M_1 n^{-\frac{M^2 \bar{G}_{(\tau_F)}}{M_1^2}}. \end{aligned}$$

Pour un choix approprié de M, le membre de droite de la dernière inégalité est le terme général d'une série convergente. Par conséquent on a bien (3.10).

Étape 2. On a

$$\begin{aligned} \sup_{x\in J}\left|\widehat{F}_n(x) - \widetilde{F}_n(x)\right| &\leq \sup_{x\in J}\left|\frac{\bar{G}(x) - \bar{G}_n(x)}{\bar{G}_n(x)}\right| \times \sup_{x\in J}\left|\frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i}{\bar{G}(Y_i)}\bar{L}_n\left(\frac{Y_i}{x+a_n}\right)\right| \\ &\leq \frac{\sup_{x\in J}|\bar{G}_n(x) - \bar{G}(x)|}{\inf_{x\in J}|\bar{G}(x)| - \sup_{x\in J}|\bar{G}_n(x) - \bar{G}(x)|} \times \sup_{x\in J}\left|\frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i}{\bar{G}(Y_i)}\bar{L}_n\left(\frac{Y_i}{x+a_n}\right)\right| \\ &=: l_1 \times l_2. \end{aligned}$$

Concernant l_1, l'utilisation de la loi du logarithme itéré (LIL) de Deheuvels et Einmahl (2000), donne

$$l_1 = O_{a.s}\left(\sqrt{\frac{\log\log n}{n}}\right).$$

Pour l_2, et sous les hypothèses **H1** et **H2**, un changement de variables et les propriétés de l'espérance conditionnelle, nous avons

$$\begin{aligned} E\left(\frac{\delta_1}{\bar{G}(Y_1)}\bar{L}_n\left(\frac{Y_1}{x+a_n}\right)\right) &= E\left(E\left(\frac{1}{\bar{G}(X_1)}\bar{L}_n\left(\frac{X_1}{x+a_n}\right)\mathbb{I}_{\{X_1\leq C_1\}}|X_1\right)\right) \\ &= \int_0^{+\infty}\frac{1}{\bar{G}(t)}\bar{L}_n\left(\frac{t}{x+a_n}\right)E\left(\mathbb{I}_{\{t\leq C_1\}}|X_1=t\right)f(t)dt \\ &= \int_0^{+\infty}\bar{L}_n\left(\frac{t}{x+a_n}\right)f(t)dt, \end{aligned}$$

et pour un $x \in J$ fixé, une intégration par partie, et les hypothèses **H1**, **H3** et **H4**(ii) donnent

$$\begin{aligned} E\left(\frac{\delta_1}{\bar{G}(Y_1)}\bar{L}_n\left(\frac{Y_1}{x+a_n}\right)\right) &= \int_0^{+\infty}F(z(x+a_n))K_n(z)dz \\ &= F(x) + O(\upsilon_n^2) = O(1). \end{aligned} \tag{3.11}$$

Ainsi par la loi forte des grands nombres (SLLN), on obtient

$$\sup_{x\in J}|\widehat{F}_n(x) - \widetilde{F}_n(x)| = O_{a.s}\left(\sqrt{\frac{\log\log n}{n}}\right)$$

Étape 3. En utilisant (3.11) on obtient facilement

$$\sup_{x\in J}\Big|E\left(\widetilde{F}_n(x)\right)-F(x)\Big|=O(\upsilon_n^2)\qquad\text{quand}\quad n\to\infty.$$

Cela achève la preuve. □

3.5.2 Preuve du Théorème 3.2

Pour prouver le théorème 3.2, et pour des raisons techniques on définit le pseudo-estimateur de $f(x)$ par

$$\widetilde{f}_n(x)=\frac{1}{n(x+a_n)^2}\sum_{i=1}^{n}\frac{\delta_i Y_i}{\bar{G}(Y_i)}\;K_n\left(\frac{Y_i}{x+a_n}\right),\tag{3.12}$$

et on considère la décomposition classique

$$\sup_{x\in J}\Big|\widehat{f}_n(x)-f(x)\Big|\le\sup_{x\in J}\Big|\widehat{f}_n(x)-\widetilde{f}_n(x)\Big|+\sup_{x\in J}\Big|\widetilde{f}_n(x)-E\left(\widetilde{f}_n(x)\right)\Big|+\sup_{x\in J}\Big|E\left(\widetilde{f}_n(x)\right)-f(x)\Big|.\tag{3.13}$$

La démonstration est donc une conséquence des lemmes suivants :

Lemme 3.1. *Sous les hypothèses* ***H1, H2, H3*** *et* ***H4*** *, et pour $x\in J$, on a*

$$Biais\Big(\widetilde{f}_n(x)\Big)=\Big((x+a_n)\upsilon_n^2+a_n\Big)f'(x)+o\Big(\upsilon_n^2+a_n\Big),\tag{3.14}$$

et

$$Var\Big(\widetilde{f}_n(x)\Big)=\frac{M_n\bar{f}(x)}{n(x+a_n)\upsilon_n}+O\left(\frac{\upsilon_n}{n\;a_n}\right),$$

où M_n et $\bar{f}(.)$ sont donnés dans les hypothèses.

Preuve. *L'utilisation de techniques standard d'espérance conditionnelle donne*

$$\begin{aligned}E\Big(\widetilde{f}_n(x)\Big)&=E\left(\frac{\delta_1\;Y_1}{(x+a_n)^2\bar{G}(Y_1)}\;K_n\left(\frac{Y_1}{x+a_n}\right)\right)\\&=\int_0^{+\infty}\frac{t}{(x+a_n)^2}\;K_n\left(\frac{t}{x+a_n}\right)\;\;f(t)\;dt\\&=\int_0^{+\infty}z\;K_n(z)\;\;f\Big(z(x+a_n)\Big)\;\;dz.\end{aligned}$$

Par l'hypothèse ***H1****, et un développement de Taylor d'ordre 1 autour de x, on obtient*

$$f\Big(z(x+a_n)\Big)=f(x)+\Big(x(z-1)+z\;a_n\Big)f'(x)+o\Big(x(z-1)+za_n\Big),$$

et

$$\begin{aligned} E\Big(\widetilde{f}_n(x)\Big) &= \int_0^{+\infty} z\ K_n(z)\left\{f(x)+\Big(x(z-1)+z\ a_n\Big)f'(x)+o\Big(x(z-1)+za_n\Big)\right\}dz \\ &= f(x)+xf'(x)\int_0^{+\infty} z(z-1)\ K_n(z)dz + a_nf'(x)\int_0^{+\infty} z^2\ K_n(z)dz \\ &\quad +o\left(\int_0^{+\infty} z\left[x(z-1)+z\ a_n)\right]\ K_n(z)dz\right) \\ &= f(x)+\Big((x+a_n)\upsilon_n^2+a_n\Big)f'(x)+o\Big(\upsilon_n^2+a_n\Big), \end{aligned} \tag{3.15}$$

Ce qui conclut le premier résultat du lemme 3.1. Maintenant, pour la variance, nous avons

$$Var\Big(\widetilde{f}_n(x)\Big)=\frac{1}{n(x+a_n)^4}\ Var\left(\frac{\delta_1Y_1}{\bar{G}(Y_1)}K_n\left(\frac{Y_1}{x+a_n}\right)\right),$$

et

$$\begin{aligned} Var\left(\frac{\delta_1Y_1}{\bar{G}(Y_1)}K_n\left(\frac{Y_1}{x+a_n}\right)\right) &= E\left(\frac{\delta_1^2Y_1^2}{\bar{G}^2(Y_1)}K_n^2\left(\frac{Y_1}{x+a_n}\right)\right)-E^2\left(\frac{\delta_1Y_1}{\bar{G}(Y_1)}K_n\left(\frac{Y_1}{x+a_n}\right)\right). \\ &=: V_1-V_2. \end{aligned}$$

À partir de (3.15) et de l'hypothèse ***H3****, on obtient*

$$\begin{aligned} V_2 &= \Big((x+a_n)^2E\left(\widetilde{f}_n(x)\right)\Big)^2 \\ &= (x+a_n)^4\left[f(x)+\Big((x+a_n)\upsilon_n^2+a_n\Big)f'(x)+o\Big(\upsilon_n^2+a_n\Big)\right]^2 \\ &= (x+a_n)^4\left(f^2(x)+O\Big(\upsilon_n^2+a_n\Big)\right). \end{aligned}$$

Pour V_1 nous avons

$$\begin{aligned} V_1 &= E\left[E\left(\frac{\mathbb{I}_{\{X_1\le C_1\}}\ X_1^2}{\bar{G}^2(X_1)}\ K_n^2\left(\frac{X_1}{x+a_n}\right)|X_1\right)\right] \\ &= \int_0^{+\infty}\frac{t^2}{\bar{G}(t)}K_n^2\left(\frac{t}{x+a_n}\right)f(t)\ dt. \end{aligned}$$

Les hypothèses ***H2*** *et* ***H3****, un changement de variable et un développement de Taylor impliquent*

$$\begin{aligned} V_1 &= (x+a_n)^3 \int_0^{+\infty} z^2 K_n^2(z) \bar{f}\Big(z(x+a_n)\Big)dz = (x+a_n)^3 I_n \int_0^{+\infty} z^2 K_n^*(z) \bar{f}\Big(z(x+a_n)\Big)dz \\ &= (x+a_n)^3 I_n \int_0^{+\infty} z^2 K_n^*(z) \left\{ \bar{f}(x) + \Big(z(x+a_n)-x\Big)\bar{f}'(x) + o\Big(z(x+a_n)-x\Big)\right\} dz \\ &= (x+a_n)^3 I_n \left\{ \bar{f}(x)\mu_2\left(K_n^*\right) + \bar{f}'(x)\Big((x+a_n)\ \mu_3\left(K_n^*\right) - x\ \mu_2\left(K_n^*\right)\Big) \right. \\ &\quad \left. + o\Big((x+a_n)\ \mu_3\left(K_n^*\right) - x\ \mu_2\left(K_n^*\right)\Big)\right\} \\ &= I_n(x+a_n)^3\Big(\bar{f}(x) + O\left(\upsilon_n^2 + a_n\right)\Big). \end{aligned}$$

D'où l'hypothèse ***H4****(ii), donne*

$$Var\Big(\widetilde{f}_n(x)\Big) = \frac{M_n \bar{f}(x)}{n(x+a_n)\upsilon_n} + O\left(\frac{\upsilon_n}{n\ a_n}\right),$$

ce qui conclut la preuve du lemme 3.1.□

Lemme 3.2. *Sous les hypothèses* ***H0, H1*** *et* ***H4*** *, et pour tout $x \in J$, nous avons*

$$\sup_{x\in J}\left|\widetilde{f}_n(x) - E\Big(\widetilde{f}_n(x)\Big)\right| = O_{a.s}\left(\sqrt{\frac{\log n}{n\ a_n^2}}\right).$$

Preuve. *On utilise un recouvrement du compact J par un nombre fini b_n d'intervalles $J_1, J_2, \ldots, J_{b_n}$ de même longueur égale à $\gamma_n = O(\frac{a_n^3}{n\upsilon_n})$ et centrés aux points $x_1, \ldots, x_{b_n}$ respectivement. Notons que comme J est borné, il existe une constante C telle que $b_n \leq C\gamma_n^{-1}$. Alors pour tout $x \in J$, (3.12) implique que*

$$\widetilde{f}_n(x) - E\Big(\widetilde{f}_n(x)\Big) = \frac{1}{n}\ \sum_{i=1}^{n} Z_i(x, a_n),$$

où

$$Z_i(x,a_n) = \frac{\delta_i Y_i}{(x+a_n)^2 \bar{G}(Y_i)}\ K_n\left(\frac{Y_i}{x+a_n}\right) - E\left[\frac{\delta_1 Y_1}{(x+a_n)^2\bar{G}(Y_1)}\ K_n\left(\frac{Y_1}{x+a_n}\right)\right].$$

Remarquons que

$$\begin{aligned} \sup_{x\in J}\left|\widetilde{f}_n(x) - E\Big(\widetilde{f}_n(x)\Big)\right| &\leq \sup_{x\in J}\frac{1}{n}\left|\sum_{i=1}^{n} Z_i(x,a_n)\right| \\ &\leq \max_{k=1,\ldots,b_n}\sup_{x\in J_k}\frac{1}{n}\left|\sum_{i=1}^{n} Z_i(x,a_n) - Z_i(x_k,a_n)\right| \\ &\quad + \max_{k=1,\ldots,b_n}\frac{1}{n}\left|\sum_{i=1}^{n} Z_i(x_k,a_n)\right|. \end{aligned} \tag{3.16}$$

*Puisque $K_n(.)$ est borné et Lipschitzien (par l'hypothèse **H1**), on a*

$$\begin{aligned}\frac{1}{n}\left|\sum_{i=1}^{n} Z_i(x,a_n) - Z_i(x_k,a_n)\right| &\leq \left|\frac{1}{n}\sum_{i=1}^{n}\frac{\tau_F}{\bar{G}(Y_i)}\frac{1}{(x+a_n)^2}K_n\left(\frac{Y_i}{x+a_n}\right) - \frac{1}{(x_k+a_n)^2}K_n\left(\frac{Y_i}{x_k+a_n}\right)\right.\\ &\quad \left. + E\left(\frac{\tau_F}{\bar{G}(Y_1)}\frac{1}{(x+a_n)^2}K_n\left(\frac{Y_1}{x+a_n}\right) - \frac{1}{(x_k+a_n)^2}K_n\left(\frac{Y_1}{x_k+a_n}\right)\right)\right|\\ &\leq \frac{2\tau_F}{\bar{G}(\tau_F)}\left[\frac{\tau_F\gamma_n}{(x+a_n)^3(x_k+a_n)} + \|K_n\|_\infty\left|\frac{1}{(x+a_n)^2} - \frac{1}{(x_k+a_n)^2}\right|\right],\end{aligned}$$

où $\|K_n\|_\infty := \sup_u |K_n(u)|$. Comme x et x_k appartiennent à J_k, cela implique qu'il existe une constante positive m_k telle que : $x \leq m_k$ et $x_k \leq m_k$, et la dernière inégalité devient,

$$\begin{aligned}\frac{1}{n}\left|\sum_{i=1}^{n} Z_i(x,a_n) - Z_i(x_k,h_n)\right| &\leq \frac{2\tau_F\ \gamma_n}{\bar{G}(\tau_F)\ a_n^2}\left\{\frac{\tau_F}{a_n^2} + \frac{2\|K_n\|_\infty|m_k+a_n|}{a_n^2}\right\}\\ &= O\left(\frac{\gamma_n}{a_n^4}\right) = O\left(\frac{1}{na_n\upsilon_n}\right).\end{aligned}$$

Ensuite, soit $U_i(x,a_n) = a_n^2 Z_i(x,a_n)$, puis pour le dernier terme de (3.16), on a pour tout $\epsilon > 0$

$$\begin{aligned}P\left(\max_{k=1,\ldots,b_n}\frac{1}{n}\left|\sum_{i=1}^{n} Z_i(x_k,a_n)\right| > \epsilon\right) &\leq \sum_{k=1}^{b_n} P\left(\frac{1}{n}\left|\sum_{i=1}^{n} Z_i(x_k,a_n)\right| > \epsilon\right)\\ &\leq P\left(\left|\sum_{i=1}^{n} U_i(x,a_n)\right| > na_n^2\epsilon\right). \qquad (3.17)\end{aligned}$$

Comme $E\left(U_i(x,a_n)\right) = 0$, $|U_i(x,a_n)| \leq \frac{2\tau_F}{\bar{G}(\tau_F)}\|K_n\|_\infty := M_1 < \infty$, et
$Var\left(U_i(x,a_n)\right. =: S^2 = O\left(\frac{a_n^3}{\upsilon_n}\right)$, on peut appliquer l'inégalité de Bernstein (Voir Ferraty et Vieu (2006) corollaire A.9, p 235) pour obtenir

$$\begin{aligned}P\left(\left|\sum_{i=1}^{n} U_i(x,a_n)\right| > na_n^2\epsilon\right) &\leq 2\exp\left\{-\frac{a_n^4\epsilon^2 n}{2S^2\left(1+\frac{a_n^2\epsilon M}{S^2}\right)}\right\}\\ &\leq 2\exp\left\{-\frac{n\epsilon^2 a_n^2}{2\left(M\frac{a_n}{\upsilon_n}+\epsilon\ M_1\right)}\right\}\\ &=: A_n,\end{aligned}$$

et donc, (3.17) devient,

$$P\left(\max_{k=1,\ldots,b_n}\frac{1}{n}\left|\sum_{i=1}^{n}Z_i(x_k,a_n)\right|>\epsilon\right) \leq b_n\ A_n \leq C\gamma_n^{-1}2\exp\left\{-\frac{n\ a_n^2\epsilon^2}{2\left(M\frac{a_n}{\upsilon_n}+\epsilon M_1\right)}\right\}$$
$$\leq C\left(\frac{n\upsilon_n}{a_n^3}\right)\times n^{-2}\exp\left\{\left(\log n\right)\left(2-\frac{n\epsilon^2 a_n\upsilon_n/\log n}{2\left(M+\epsilon M_1\ \frac{\upsilon_n}{a_n}\right)}\right)\right\},$$

si on remplace ϵ *par* $\epsilon_0\sqrt{\dfrac{\log n}{na_n\upsilon_n}}$*, on obtient*

$$P\left(\left|\sum_{i=1}^{n}Z_i(x,a_n)\right|>\epsilon\right)\leq\frac{C\upsilon}{na_n^3}n^{2-\epsilon_0^2\Big/2\left(M+M_1\epsilon_0\sqrt{\frac{\upsilon_n\log n}{na_n^3}}\right)}\approx\frac{C\upsilon}{na_n^3}n^{2-C_1\epsilon_0^2}. \tag{3.18}$$

Par l'hypothèse ***H4****, le second membre de (3.18) est le terme général d'une série convergente et Le lemme de Borel-Cantelli permet alors de conclure.*□

On revient maintenant au premier terme de la décomposition (3.13) pour lequel on a

Lemme 3.3. *Si les hypothèses* ***H1,H2,H3*** *et* ***H4*** *sont satisfaites alors pour tout* $x\in J$

$$\sup_{x\in J}\left|\widehat{f}_n(x)-\widetilde{f}_n(x)\right|=O_{p.s}\left(\sqrt{\frac{\log\log n}{n}}\right)$$

Preuve. *On a*

$$\begin{aligned}
\sup_{x\in J}\left|\widehat{f}_n(x)-\widetilde{f}_n(x)\right| &\leq \sup_{x\in J}\left|\frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{\bar{G}_n(Y_i)}-\frac{1}{\bar{G}(Y_i)}\right)\frac{\delta_i Y_i}{(x+a_n)^2}\ K_n\left(\frac{Y_i}{x+a_n}\right)\right| \\
&\leq \sup_{x\in J}\left|\frac{\bar{G}(x)-\bar{G}_n(x)}{\bar{G}_n(x)}\right|\times\sup_{x\in J}\left|\frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i Y_i}{(x+a_n)^2\bar{G}(x)}\ K_n\left(\frac{Y_i}{x+a_n}\right)\right| \\
&\leq \frac{\sup_{x\in J}|\bar{G}_n(x)-\bar{G}(x)|}{\inf_{x\in J}|\bar{G}(x)|-\sup_{x\in J}|\bar{G}_n(x)-\bar{G}(x)|}\times\sup_{x\in J}|\widetilde{f}_n(x)| \\
&\leq M\sqrt{\frac{\log\log n}{n}}\times O(1)=O_{p.s}\left(\sqrt{\frac{\log\log n}{n}}\right),
\end{aligned}$$

la dernière inégalité est donnée en utilisant la (LLI) pour l'estimateur de Kaplan Meier, et par le lemme 3.2 qui montre que $\left|\widetilde{f}_n(x)\right|$ *est uniformément presque sûrement bornée.* □

Enfin, on a le lemme 3.1 et l'hypothèse **H4**(ii) impliquent $\sup_{x\in J}\left|\widetilde{f}_n(x)-E\left(\widetilde{f}_n(x)\right)\right|=O_{a.s}\left(\upsilon_n^2+a_n\right)=O_{a.s}\left(\upsilon_n^2\right)$, ceci, avec le lemme 3.2 et le lemme 3.3 complètent la preuve du Théorème 3.2. □

3.5.3 Preuve du corollaire 3.1

Pour prouver a), considérons la décomposition suivante

$$\widehat{\lambda}_n(x) - \lambda(x) = \frac{1}{1-\widehat{F}_n(x)}\left[\left(\widehat{f}_n(x) - f(x)\right) + f(x)\frac{\widehat{F}_n(x) - F(x)}{1-F(x)}\right],$$

alors le résultat découle du (LLI) et du théorème 3.2.
La preuve de l'item b) est similaire en utilisant en plus le théorème 3.1. □

Chapitre 4

Normalité Asymptotique des Estimateurs à noyaux asymétriques

Sommaire

4.1 Introduction

On se propose, pour compléter l'étude des propriétés asymptotiques des estimateurs à noyaux asymétriques étudiés dans les chapitres précédents, d'établir la normalité asymptotique de ces estimateurs. Ce chapitre est organisé en trois sections. La première présente quelques résultats préliminaires utiles dans la preuve de nos résultats. Dans la section suivante, on est dans le cas censuré à droite, nous présentons notre deuxième résultat qui fait partie de l'article (Ghettab et Guessoum (2022)), en utilisant le théorème centrale limite (T.C.L) de Lyapounov pour montrer la normalité asymptotique de l'estimateur de la densité , et comme conséquence couplé par le lemme de Slutsky, nous avons la normalité asymptotique de l'estimateur de la fonction taux de hasard. Ensuite, comme application de cette propriété, nous avons construit les intervalles de confiance de nos estimateurs. La dernière section est dédiée à l'illustration numérique sous forme de simulation.

4.2 Résultats préliminaires

4.2.1 Théorème central limite de Lyapunov

On donne ici une version du théorème centrale limite de Lyapunov associé à une suite de variable aléatoire indépendante. Nous trouvons la preuve de ce théorème par la méthode des fonctions caractéristiques dans le livre de Chung K.L. (2001) [19].

Théorème 4.1. *(Chung K.L. (2001), Théorème 7.1.2) Soit $X_1, \ldots, X_n$ une suite de variables aléatoires indépendantes, d'espérance nulle $E(X_j) = 0$, et de variance finie $\sigma^2(X_j) = \sigma_j^2 < \infty$, et dont le moment d'ordre 3 existe $E(|X_j|^3) = \gamma_j < \infty$. Posons*

$$S_n = \sum_{j=1}^{n} X_j, \quad s_n^2 = \sum_{j=1}^{n} \sigma_j^2, \quad \Gamma_n = \sum_{j=1}^{n} \gamma_j.$$

Alors sous l'hypothèse

$$\frac{\Gamma_n}{s_n^3} \longrightarrow 0, \quad \textit{quand} \;\; n \to \infty$$

on a

$$\frac{S_n}{s_n} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1).$$

et $\xrightarrow{\mathcal{L}}$ représente la convergence en Loi.

4.2.2 Lemme de Slutsky (Slutsky, 1925)

Ce théorème, démontré par Eugen Slutsky en 1925 [26] est aussi appelé théorème de Cramèr dans la littérature. Son utilité provient du fait qu'il permet de montrer, entre autres, que la suite des images $f(X_n, Y_n)_{n\in\mathbb{N}}$ d'un couple de v.a. dont on connaît les convergences en loi, converge elle-même en loi vers l'image $f(X, C)$ pour peu que C soit une constante. Pour plus de détails on se réfère au livre Probability and Random Processes, third edition (Page 318) [36].

Lemme 4.1. *(a) Si $(X_n)_{n\in\mathbb{N}}$ est une suite de variables aléatoires qui converge en loi vers une v.a. X, et si $(Y_n)_{n\in\mathbb{N}}$ est une suite de variables aléatoires qui converge en probabilité vers une constante C, alors la suite $(X_nY_n)_{n\in\mathbb{N}}$ converge en loi vers CX, et la suite $\dfrac{X_n}{Y_n}$ converge en loi vers $\dfrac{X}{C}$ si $C \neq 0$.*
(b) Si $(X_n)_{n\in\mathbb{N}}$ est une suite de variables aléatoires qui converge en loi vers 0, et si $(Y_n)_{n\in\mathbb{N}}$ est une suite de variables aléatoires qui converge en probabilité vers une v.a Y, et si $g : \mathbb{R}^2 \longrightarrow \mathbb{R}$ est telle que $g(x, y)$ est une fonction continue en y pour tout x, et $g(x, y)$ est continue en $x = 0$ quel que soit y, alors la suite de couples $g(X_n, Y_n)_{n\in\mathbb{N}}$ converge en loi vers le couple $(0, Y)$.

4.3 Cas i.i.d / censuré à droite

Dans cette section on se place dans le cas censuré à droite. Nous présentons notre deuxième résultat relatif à l'article (Ghettab et Guessoum (2022)). Soit $(X_n)_{n\geq 1}$ une suite de variables aléatoires indépendantes identiquement distribuées (i.i.d) dite ' temps de survie' définies sur un espace probabilisé $(\Omega, \mathcal{B}, \mathbb{P})$. On suppose que ces variables ne sont pas complètement observées mais sont censurées à droite par une suite de v.a.i.i.d $(C_n)_{n\geq 1}$ appelée 'temps de censure' de f.r G, qui sont indépendantes des variables durées de survie X_i. Dans le modèle censuré à droite on observe les couples $\{(Y_i, \delta_i), i = 1, ..., n\}$ où

$$Y_i = \min(X_i, C_i) \quad \text{et} \quad \delta_i = \mathbb{I}_{\{X_i \leq C_i\}},$$

avec Y_i sont des variables aléatoires i.i.d et δ_i est l'indicateur de non censure. Ainsi, on va étudier la convergence en loi des estimateurs suivants :
Estimateur de la densité f de la v.a temps de survie :

$$\widehat{f}_n(x) = \frac{1}{n(x + a_n)^2} \sum_{i=1}^{n} \frac{\delta_i Y_i}{\bar{G}_n(Y_i^-)} \; K_n\left(\frac{Y_i}{x + a_n}\right), \tag{4.1}$$

avec :

→ $(K_n(.))$ est une famille de noyaux asymétriques d'espérance 1 et de variance finie υ_n^2 (voir l'hypothèse H1),

→ a_n est une suite de nombres positifs qui tend vers 0,

→ $\bar{G}_n(Y_i^-)$ c'est la limite à gauche de l'EKM de la f.r de la v.a temps de censure au point Y_i.

Estimateurs de la fonction taux de hasard λ :
On propose deux estimateurs

$$\widehat{\lambda}(x) = \frac{\widehat{f}_n(x)}{1 - F_n(x)}, \tag{4.2}$$

et

$$\widetilde{\lambda}(x) = \frac{\widehat{f}_n(x)}{1 - \widehat{F}_n(x)}, \tag{4.3}$$

avec :

→ $F_n(x) = \dfrac{1}{n}\sum_{i=1}^{n} \dfrac{\delta_i}{\left(1 - G_n\left(Y_i^-\right)\right)}\ \mathbb{I}_{\{Y_i \le x\}}$, c'est l'EKM de F la f.r de la v.a temps de survie,

→ $\widehat{F}_n(x) = 1 - \dfrac{1}{n}\sum_{i=1}^{n} \dfrac{\delta_i}{\left(1 - G_n\left(Y_i^-\right)\right)}\ L_n\left(\dfrac{Y_i}{x + a_n}\right)$ un estimateur à noyau asymétrique de F,

où $L_n(x) = \int_0^x K_n(t)dt$, est une famille de f.r associe à $K_n(.)$ ou ce qu'on appelle aussi noyau intégré.

4.3.1 Hypothèses et résultats

Pour $x \in]0, +\infty[$, on définit $\tau_F = \sup\{x : \bar{F}(x) > 0\}$ et $\tau_G = \sup\{x : \bar{G}(x) > 0\}$ les bornes supérieures de $\bar{F} := 1 - F$ et $\bar{G} := 1 - G$, respectivement. Dans la pratique, τ_F désigne par exemple la date limite des observations de X. On supposera ici que F et G sont des f.r continues. Pour établir la convergence en loi de l'estimateur de la densité à noyau asymétrique $\hat{f}_n(.)$ (Théorème 4.2) et les estimateurs à noyau asymétriques du taux de risque $\widehat{\lambda}_n(.)$ et $\widetilde{\lambda}_n(.)$ (corollaire 4.1), nous avons besoins des hypothèses suivantes :

H0. La durée de censure C est indépendante de la vraie durée de survie X, et $\tau_F < \infty, \bar{G}(\tau_F) > 0$.

H1. K_n est une famille de fonctions définies sur $\mathbb{R}^+$, telle que pour $n = 1, 2, \ldots$, $K_n(.)$ est bornée et lipschitzienne sur J, et satisfait

(i) $\int_0^{+\infty} u\ K_n(u)du = 1,$

(ii) $\int_0^{+\infty} u(u-1)K_n(u)\ du =: \upsilon_n^2 \longrightarrow 0 \ \text{ quand } \ n \to \infty,$

H2'. $I_{n,l} := \int_0^{+\infty} K_n^l(w)dw = O(\upsilon_n^{-(l-1)}), \text{pour } 1 \le l \le 3, \text{ et } K_{n,l}^*(u) := \dfrac{K_n^l(u)}{I_{n,l}}$, de moyenne $\mu_j(K_{n,l}^*) := \int_0^{+\infty} u^j K_{n,l}^*(u)du = 1 + O(\upsilon_n^2), \quad 1 \le j \le 3.$

H3'. Les fonctions $f, f_1(t) := \dfrac{f(t)}{\bar{G}(t)}$ et $f_2(t) := \dfrac{f(t)}{\bar{G}^2(t)}$ sont deux fois continûment dérivables sur J telle que $\|g^{''}\|_\infty := \sup_{x \in J}|g^{''}(x)| < +\infty, \forall g = f, f_1, \text{ou} f_2.$

H4'. (i) $\lim_{n\to\infty} a_n = 0, \quad \lim_{n\to\infty} \upsilon_n = 0,$

(ii) $\lim_{n\to\infty} \dfrac{a_n}{\upsilon_n^2} = 0,$

H5. (i) $\lim_{n\to\infty} \upsilon_n \log\log n = 0.$

(ii) $\lim_{n\to\infty} n\upsilon_n^5 = 0.$

(iii) $\lim_{n\to\infty} n\upsilon_n = \infty.$

Remarque 4.1. *Notez que les hypothèses **H3'** et **H2'** impliquent les hypothèses de la consistance forte **H3** et **H2** (dans le chapitre précédent), telle que $I_{n,2} = I_n$ et $K^*_{n,2}(u) = K^*_n(u)$.*

4.3.2 Résultats

Théorème 4.2. *Sous les hypothèses **H0, H1, H2', H3', H4'** et **H5** , nous avons*

$$\sqrt{n\upsilon_n}\left(\widehat{f}_n(x) - f(x)\right) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \sigma^2(x)\right),$$

*avec $\sigma^2(x) = \overline{M}\dfrac{f_1(x)}{x}, \overline{M} := \lim_{n\to+\infty} \upsilon_n\, I_{n,2}$, où $I_{n,2}$, et f_1 ce sont des quantités définies dans **H2', H3'** et respectivement.*

Remarque 4.2. *On remarque que dans le cas de non censure c'est-à-dire $\bar{G}(.) = 1$, la variance obtenue dans le théorème précédent $\overline{M}\dfrac{f_1(x)}{x}$ devient $\overline{M}\dfrac{f(x)}{x}$ c'est ce qui a été obtenu par Chaubey et al (2007) [10].*

Du Théorème**??** on déduit

Corollaire 4.1. *Sous les hypothèses du théorème 4.2,*

$$a)\sqrt{n\upsilon_n}\left(\widehat{\lambda}_n(x) - \lambda(x)\right) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \sigma^2(x)\left[\bar{F}(x)\right]^{-2}\right),$$

$$b)\sqrt{n\upsilon_n}\left(\tilde{\lambda}(x) - \lambda(x)\right) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \sigma^2(x)\left[\bar{F}(x)\right]^{-2}\right).$$

Application à la construction d'intervalles de confiance
Pour déterminer les intervalles de confiances de nos estimateurs, on remarque que les variances asymptotiques dépendent de certaines fonctions inconnus $f, \bar{G}$ et $\bar{F}$, alors on utilise la méthode de plug-in pour obtenir des estimations des variances en remplaçant ces quantités par leurs estimateurs $\widehat{f}_n, \bar{G}_n, \bar{F}_n$ ou $\widehat{F}_n$ respectivement. Ainsi, on obtient le corollaire suivant :

Corollaire 4.2. *Sous les hypothèses du théorème 4.2, nous avons*

(a) Un intervalle de confiance de niveau $1-\eta$ avec $0 \le \eta \le 1$ pour la densité $f(x)$ est donné par

$$\left[\widehat{f}_n(x) - \frac{\widehat{\sigma}_n(x)}{\sqrt{n\upsilon_n}} \times z_{1-\frac{\eta}{2}}, \widehat{f}_n(x) + \frac{\widehat{\sigma}_n(x)}{\sqrt{n\upsilon_n}} \times z_{1-\frac{\eta}{2}}\right]$$

(b) L'intervalle de confiance de niveau $1-\eta$ *avec* $0 \leq \eta \leq 1$ *pour* $\lambda(x)$ *est donné par*

$$\left[\widehat{\lambda}_n(x) - \frac{\left(\bar{F}_n(x)\right)^{-1}\widehat{\sigma}_n(x)}{\sqrt{n\upsilon_n}} \times z_{1-\frac{\eta}{2}}, \widehat{\lambda}_n(x) + \frac{\left(\bar{F}_n(x)\right)^{-1}\widehat{\sigma}_n(x)}{\sqrt{n\upsilon_n}} \times z_{1-\frac{\eta}{2}}\right]$$

et

$$\left[\tilde{\lambda}_n(x) - \frac{\left(1-\widehat{F}_n(x)\right)^{-1}\widehat{\sigma}_n(x)}{\sqrt{n\upsilon_n}} \times z_{1-\frac{\eta}{2}}, \tilde{\lambda}_n(x) + \frac{\left(1-\widehat{F}_n(x)\right)^{-1}\widehat{\sigma}_n(x)}{\sqrt{n\upsilon_n}} \times z_{1-\frac{\eta}{2}}\right]$$

où $z_{1-\frac{\eta}{2}}$ *est le quantile d'ordre* $1-\frac{\eta}{2}$ *de la loi normale centrée réduite, et* $\widehat{\sigma}_n^2(x) = \dfrac{\overline{M}}{x}\dfrac{\widehat{f}_n(x)}{\bar{G}_n(x)}$, *est l'estimateur plug-in de* $\sigma^2(x)$.

4.3.3 Preuve du Théorème 4.2

Pour $x > 0$, nous avons

$$\begin{aligned}\sqrt{n\upsilon_n}\left(\widehat{f}_n(x) - f(x)\right) =& \sqrt{n\upsilon_n}\left(\widehat{f}_n(x) - \widetilde{f}_n(x)\right) + \sqrt{n\upsilon_n}\left(\widetilde{f}_n(x) - E\left(\widetilde{f}_n(x)\right)\right)\\ &+ \sqrt{n\upsilon_n}\left(E\left(\widetilde{f}_n(x)\right) - f(x)\right).\\ =:& \Lambda_1 + \Lambda_2 + \Lambda_3\end{aligned}$$

où $\widetilde{f}_n(x)$ est le pseudo-estimateur de $f(x)$ défini en (3.12). Nous montrons que les termes Λ_1 et Λ_3 sont négligeables. En effet, à partir du lemme 3.3 et sous l'hypothèse **H5(i)**, on obtient

$$\begin{aligned}\Lambda_1 =& \sqrt{n\upsilon_n}\left(\widehat{f}_n(x) - \widetilde{f}_n(x)\right)\\ =& O_{p.s}\left(\sqrt{\upsilon_n \log\log n}\right) = o(1).\end{aligned} \tag{4.4}$$

De même, à partir du lemme 3.1, et sous l'hypothèse **H4'(ii)** et **H5(ii)**, on a

$$\begin{aligned}\Lambda_3 =& \sqrt{n\upsilon_n}\left(E\left(\widetilde{f}_n(x)\right) - f(x)\right)\\ =& O_{p.s}\left(\sqrt{n\upsilon_n^5}\right) = o(1).\end{aligned} \tag{4.5}$$

Considérons maintenant le terme dominant Λ_2, et observons que la convergence dans le théorème 4.2 peut être déduite du fait que

$$\Lambda_2 \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \sigma^2(x)\right).$$

Ce dernier résultat sera établi grâce au lemme suivant

Lemme 4.2. *Sous les hypothèses du théorème 4.2, on a*

$$\sqrt{n\upsilon_n a_n}\left(\widetilde{f}_n(x) - E\left(\widetilde{f}_n(x)\right)\right) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \sigma^2(x)\right)$$

avec $\sigma^2(x) = \dfrac{\overline{M} f_1(x)}{x}, \overline{M} := \lim\limits_{n\to+\infty} \upsilon_n\, I_{n,2}$, *où* $I_{n,2}, f_1$ *sont des quantités définies dans* ***H2'***, *et* ***H3'***.

Preuve du lemme 4.2
Soit

$$\sqrt{n\upsilon_n}\left(\widetilde{f}_n(x)-E\left(\widetilde{f}_n(x)\right)\right)=:\sum_{j=1}^{n}W_{jn}(x). \tag{4.6}$$

où

$$\begin{aligned}W_{jn}(x):=&\sqrt{\frac{\upsilon_n}{n}}\left[\frac{\delta_j Y_j}{(x+a_n)^2\,\bar{G}_n(Y_j^-)}\;K_n\left(\frac{Y_j}{x+a_n}\right)-E\left(\frac{\delta_j Y_j}{(x+a_n)^2\,\bar{G}_n(Y_j^-)}\;K_n\left(\frac{Y_j}{x+a_n}\right)\right)\right]\\ =&\sqrt{\frac{\upsilon_n}{n}}\left(R_{jn}-E(R_{jn})\right)\end{aligned}$$

et

$$R_{jn}:=\frac{\delta_j Y_j}{(x+a_n)^2\,\bar{G}_n(Y_j^-)}\;K_n\left(\frac{Y_j}{x+a_n}\right).$$

Maintenant, pour montrer que (4.6) est asymptotiquement normalement distribué, nous pouvons vérifier la condition de Lyapunov (voir Chung (2001) [19] Theorem 7.1.2, page 222), c'est-à-dire on va vérifier que

$$\frac{\sum\limits_{j=1}^{n}E(|W_{jn}|^3)}{\left(\sum\limits_{j=1}^{n}E(W_{jn}^2)\right)^{\frac{3}{2}}}\longrightarrow 0 \text{ quand } n\to\infty,$$

ainsi

$$\frac{\sum\limits_{j=1}^{n}W_{jn}}{\left(\sum\limits_{j=1}^{n}E(W_{jn}^2)\right)^{\frac{1}{2}}}=\frac{\sqrt{n\upsilon_n}\left(\widetilde{f}_n(x)-E\left(\widetilde{f}_n(x)\right)\right)}{\left(\sum\limits_{j=1}^{n}E(W_{jn}^2)\right)^{\frac{1}{2}}}\;\xrightarrow{\mathcal{D}}\;\mathcal{N}(0,1).$$

Puisque les R_{jn} sont positifs $(1\leq j\leq n)$, on a

$$\frac{\sum\limits_{j=1}^{n}E(|W_{jn}|^3)}{\left(\sum\limits_{j=1}^{n}E(W_{jn}^2)\right)^{\frac{3}{2}}}\leq\frac{E\left[(R_{1n}+E(R_{1n}))^3\right]}{\sqrt{n}(\,Var(R_{1n}))^{\frac{3}{2}}}=\;\frac{E(R_{1n}^3)+4\;E^3(R_{1n})+3E(R_{1n}^2)E(R_{1n})}{\sqrt{n}(E(R_{1n}^2)-E^2(R_{1n}))^{\frac{3}{2}}}.$$

Ensuite, par le lemme 3.1 du chapitre précédent et l'hypothèse **H4**, on peut montrer que

$$\begin{aligned}E(R_{1n}) &= E\left(\frac{\delta_1 Y_1}{(x+a_n)^2\,\bar{G}_n(Y_1^-)}\;K_n\left(\frac{Y_1}{x+a_n}\right)\right)\\ &= f(x)+\Big((x+a_n)\upsilon_n^2+a_n\Big)f'(x)+o\Big(\upsilon_n^2+a_n\Big)\\ &= O(1)\end{aligned} \tag{4.7}$$

et

$$\begin{aligned} E(R_{1n}^2) &= E\left(\frac{\delta_1^2 Y_1^2}{(x+a_n)^4\ \bar{G}_n^2(Y_1^-)}\ K_n^2\left(\frac{Y_1}{x+a_n}\right)\right) \\ &= \frac{1}{(x+a_n)^4} E\left[E\left(\frac{\mathbb{I}_{\{X_1\le C_1\}}\ X_1^2}{\bar{G}^2(X_1)}\ K_n^2\left(\frac{X_1}{x+a_n}\right)|X_1\right)\right] \\ &= \frac{1}{(x+a_n)^4}\int_0^{+\infty}\frac{t^2}{\bar{G}(t)}K_n^2\left(\frac{t}{x+a_n}\right) f(t)\ dt. \end{aligned}$$

Les hypothèses **H2'**, **H3'**, et un changement de variable et un développement de Taylor impliquent

$$\begin{aligned} E(R_{1n}^2) &= \frac{1}{(x+a_n)}\int_0^{+\infty} z^2 K_n^2(z) f_1\Big(z(x+a_n)\Big)dz \\ &= \frac{I_{n,2}}{(x+a_n)}\int_0^{+\infty} z^2 K_{n,2}^*(z) f_1\Big(z(x+a_n)\Big)dz \\ &= \frac{I_{n,2}}{(x+a_n)}\int_0^{+\infty} z^2 K_{n,2}^*(z)\left\{f_1(x)+\Big(z(x+a_n)-x\Big)f_1'(x)+o\Big(z(x+a_n)-x\Big)\right\}dz \\ &= \frac{I_{n,2}}{(x+a_n)}\Big\{f_1(x)\mu_2\left(K_{n,2}^*\right)+f_1'(x)\Big((x+a_n)\ \mu_3\left(K_{n,2}^*\right)-x\ \mu_2\left(K_{n,2}^*\right)\Big) \\ &\quad +o\Big((x+a_n)\ \mu_3\left(K_{n,2}^*\right)-x\ \mu_2\left(K_{n,2}^*\right)\Big)\Big\} \\ &= \frac{I_{n,2}}{(x+a_n)}\Big(f_1(x)+O\left(\upsilon_n^2+a_n\right)\Big), \end{aligned}$$

d'après les hypothèses **H3'** et **H4**(i). Ce qui implique

$$\upsilon_n\ E(R_{1n}^2)\underset{n\to+\infty}{\longrightarrow}\frac{\overline{M}}{x}f_1(x), \tag{4.8}$$

ensuite

$$\begin{aligned} E(R_{1n}^3) &= E\left(\frac{\delta_1^3 Y_1^3}{(x+a_n)^6\ \bar{G}_n^3(Y_1^-)}\ K_n^3\left(\frac{Y_1}{x+a_n}\right)\right) \\ &= \int_0^{+\infty}\frac{t^3}{(x+a_n)^6\bar{G}^3(t)}\ K_n^3\left(\frac{t}{x+a_n}\right)E\left(\mathbb{I}_{\{t\le C_1\}}\Big|T_1=t\right)\quad f(t)\ \ dt \\ &= \int_0^{+\infty}\frac{t^3}{(x+a_n)^6}\ K_n^3\left(\frac{t}{x+a_n}\right)\quad f_2(t)\ dt. \end{aligned}$$

Les hypothèses **H2'** et **H3'**, un changement de variable et un développement de Taylor de $f_2(.)$

impliquent

$$\begin{aligned} E(R_{1n}^3) &= \frac{I_{n,3}}{(x+a_n)^2}\Big\{f_2(x)\mu_3(K_{n,3}^*) + f_2^{'}(x)\left[\mu_4(K_{n,3}^*) - \mu_3(K_{n,3}^*)\right] + a_n f_2^{'}(x)\mu_4(K_{n,3}^*) \\ &\quad + o\left((x+a_n)^2\mu_4(K_{n,3}^*) - x\mu_3(K_{n,3}^*)\right)\Big\} \\ &= O(\upsilon_n^{-2}). \end{aligned}$$

Par conséquent, la condition de Lyapunov est vérifiée sous l'hypothèse **H5(iii)** telle que

$$\frac{\sum_{j=1}^{n} E(|W_{jn}|^3)}{\left(\sum_{j=1}^{n} E(W_{jn}^2)\right)^{\frac{3}{2}}} = O\left(\frac{1}{\sqrt{n\upsilon_n}}\right), \tag{4.9}$$

ce qui implique que

$$\frac{\sqrt{n\upsilon_n}\left(\widetilde{f}_n(x) - E\left(\widetilde{f}_n(x)\right)\right)}{\sqrt{\upsilon_n\, Var(R_{1n})}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1),$$

ou de la même manière

$$\sqrt{n\upsilon_n}\left(\widetilde{f}_n(x) - E\left(\widetilde{f}_n(x)\right)\right) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \sigma^2(x)\right),$$

où $\sigma^2(x) = \lim_{n\to\infty} \upsilon_n\, Var(R_{1n})$, par (4.8) et (4.7) on a $\sigma^2(x) = \dfrac{\overline{M}}{x} f_1(x)$. ■
Enfin la démonstration du théorème 4.2 est complétée en combinant (4.4), (4.5) et le lemme 4.2. ■

4.3.4 Preuve du corolaire 4.1

Suivant Lo et *al* (1989) [52] soit

$$\begin{aligned} \sqrt{n\upsilon_n}\left(\widehat{\lambda}_n(x) - \lambda(x)\right) &= \sqrt{n\upsilon_n}\left\{\frac{\widehat{f}_n(x)}{1-F_n(x)} - \frac{f(x)}{1-F(x)}\right\} \\ &= \sqrt{n\upsilon_n}\left\{\frac{\widehat{f}_n(x)}{\bar{F}_n(x)} - \frac{\widehat{f}_n(x)}{\bar{F}(x)} + \frac{\widehat{f}_n(x)}{\bar{F}(x)} - \frac{f(x)}{\bar{F}(x)}\right\} \\ &= \sqrt{n\upsilon_n}\left[\widehat{f}_n(x)\left(\frac{1}{\bar{F}_n(x)} - \frac{1}{\bar{F}(x)}\right) + \left(\bar{F}(x)\right)^{-1} \times \right. \\ &\quad \left.\left(\widehat{f}_n(x) - f(x)\right)\right] \end{aligned} \tag{4.10}$$

alors, à partir du théorème 4.2 et du Lemme de Slutsky (voir 4.1 b)), on peut facilement prouver la partie a) puisque le premier terme à gauche converge vers zéro en probabilité quand $n \to \infty$. La preuve de la partie b) est similaire à a) en remplaçant $F_n(x)$ par $\widehat{F}_n(x)$ dans (4.10). ■

4.4 Illustration numérique

Dans cette section, on essaye de vérifier numériquement les résultats théoriques de la normalité asymptotique. Nous allons comparer la forme de la densité normale centrée réduite à celle de l'écart normalisé entre l'estimateur de la densité et la fonction théorique. Ainsi, nous traçons la borne supérieure et inférieure de confiance à 95%. Nous considérons l'algorithme suivant :
Étape 1 : Générer un n échantillon de la variable temps de survie $X_i \sim W(4.3, 1.3)$.
Étape 2 : Générer un n échantillon de la variable temps de censure $C_i \sim Exp(\lambda)$, où λ est ajusté selon le pourcentage de censure souhaité.
Étape 3 : Calculer les données observées $Y_i = \min(X_i, C_i)$, $\delta_i = \mathbb{I}_{\{X_i \leq C_i\}}$, et l'estimateur de Kaplan-Meier du temps de censure $\bar{G}_n(.)$ avec une légère modification pour éviter de prendre la valeur zéro (voir Marron et Padgett (1987)),

$$\bar{G}_n(t) = \begin{cases} 1 \quad 0 \leq t \leq Y_{(1)} \\ \prod_{i=1}^{k-1} \left(\frac{n-i+1}{n-i+2}\right)^{1-\delta_{(i)}}, \quad Y_{(k-1)} < t \leq Y_{(k)}, k = 2, \ldots, n \\ \prod_{i=1}^{n} \left(\frac{n-i+1}{n-i+2}\right)^{1-\delta_{(i)}}, t > Y_{(n)}. \end{cases}$$

Étape 4 : Nous fixons $x = 1$ et nous calculons $B = 200$ estimations de $\widehat{f}_{n,j}(1), 1 \leq j \leq B$, défini dans 3.6, en choisissant le noyau Log-normal $Log\text{-}N(-\upsilon_n^2/2, \upsilon_n)$, et les paramètres optimaux (υ_n^{2*}, a_n^*) à partir du tableau (3.2) pour la distribution $W(4.3, 1.3)$.
Étape 5 : Nous déduisons $\forall j = 1, \ldots, B$, l'écart normalisé entre la valeur estimée $\widehat{f}_{n,j}(1)$ et la théorique $f(1)$ qui correspond à la densité d'une Weibull $W(4.3, 1.3)$, défini au point $x = 1$ par :

$$Z_{n,j} := \sqrt{\frac{2\sqrt{\pi}\ n\ x\ \upsilon_n^*\ \bar{G}_{n,j}(1)}{\widehat{f}_{n,j}(1)}} \left(\widehat{f}_{n,j}(1) - f(1)\right) \tag{4.11}$$

où la constante $2\sqrt{\pi}$ est la valeur de $\overline{M} := \lim_{n \to +\infty} \upsilon_n\ I_{n,2}$ qui convient au noyau $Log\text{-}N(-\upsilon_n^2/2, \upsilon_n)$ (voir la section 3.3.2).
Étape 6 : Nous disposons maintenant de $B = 200$ estimations des écarts normalisés $Z_{n,j}$ au point $x = 1$, pour lesquels on estime la densité par la méthode à noyau classique en utilisant le noyau gaussien et pour le choix du paramètre de lissage on prend la fenêtre classique $h_B = C\ B^{-\frac{1}{5}}$ (voir Silverman (1986), page 40).
Étape 7 : Ainsi dans la même figure on trace le graphe de la densité estimée des écarts normalisées $Z_{n,j}$ à partir de $B = 200$ trajectoires et on compare avec la densité gaussienne, une fois on fait varier le PC, et une autre fois pour des différentes valeurs de n .
Étape 8 : Une autre manière de vérification est d'utiliser le graphe des écarts normalisées observées, ordonnées en ordre croissant notées $Z_{(n,j)}, \forall j = 1, \ldots, B$ en fonction des quartiles de la loi gaussienne, ce graphe est connu sous le nom de QQ-Plot.
Étape 9 : pour B = 300 réplications, nous encadrons la valeur estimée $\widehat{f}_n(x)$ par les bornes supérieure et inférieure données dans le Corollaire 4.2 . Cet intervalle de confiance est construit à 95%

pour $x \in [0.5, 6]$, pour des différents choix de la taille d'échantillon avec un pourcentage de censure $PC = 20\%, PC = 40\%$, respectivement. Nous traçons dans les figures 4.5 et 4.6 l'estimateur $\widehat{f}_n(x)$ avec la densité théorique encadrés par les deux bornes inférieure et supérieure.

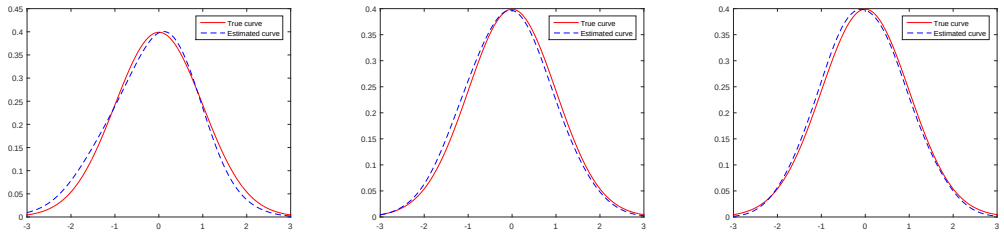


FIGURE 4.1 – densité des écarts normalisé pour $PC = 20\%$, $n = 100, n = 500$ et $n = 1000$ respectivement.

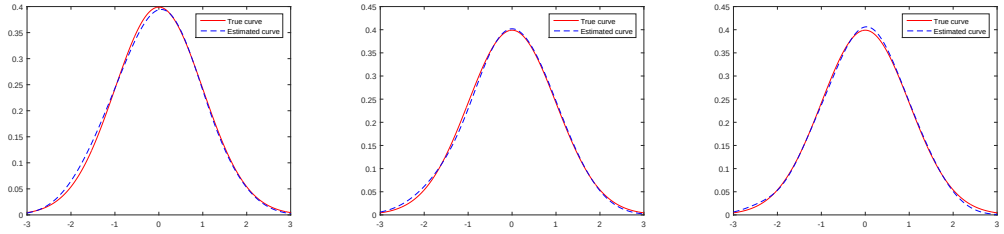


FIGURE 4.2 – densité des écarts normalisé pour $PC = 40\%$, $n = 100, n = 500$ et $n = 1000$ respectivement.

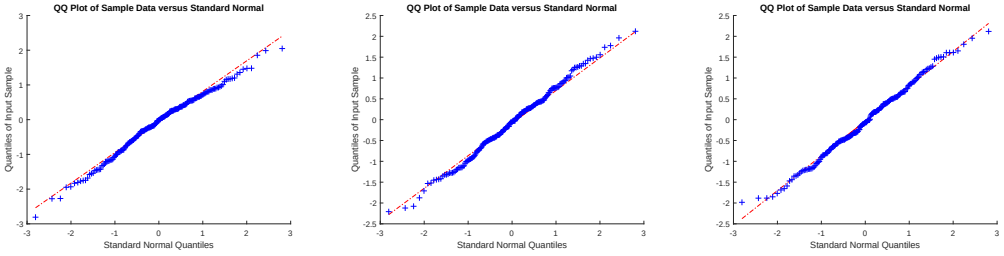


FIGURE 4.3 – QQ-plot des écarts normalisé pour $PC = 20\%$, $n = 100, n = 500$ et $n = 1000$ respectivement.

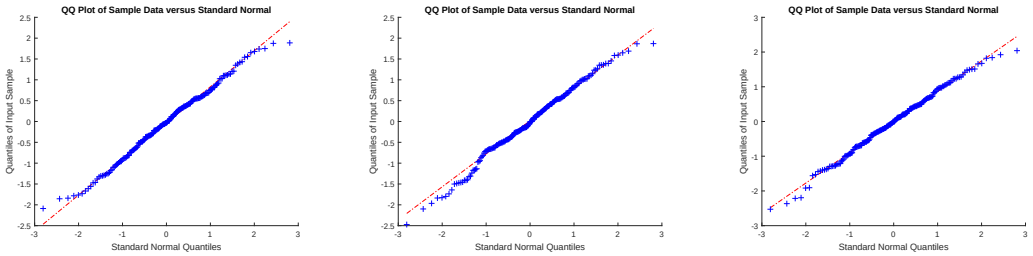


FIGURE 4.4 – QQ-plot des écarts normalisé pour $PC = 40\%$, $n = 100, n = 500$ et $n = 1000$ respectivement.

Les Figures 4.1 et 4.2 montrent que la qualité de l'ajustement s'améliore pour une grande taille d'échantillons par exemple pour $n = 1000$ peut être qu'il n'y a pas d'effet de censure mais pour $n = 100$ ou 500 on observe l'influence de la censure, ce qui est également le cas au regard du QQ-Plot correspondant dans les figures 4.3 et 4.4. Cet ajustement reste globalement satisfaisant et confirme le résultat de la normalité asymptotique énoncé dans le théorème 4.2. Comme précédemment, nous remarquons clairement l'amélioration de l'encadrement pour une taille d'échantillon n élevée, et l'influence de la censure surtout pour n petite. Ce constat confirme nos résultats théoriques.

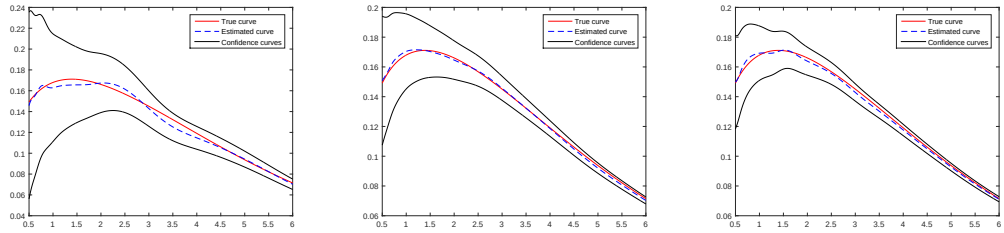


FIGURE 4.5 – Intervalles de confiances de l'estimateur de la densité $W(4.3, 1.3)$, pour $PC = 20\%$, $n = 100, n = 500$ et $n = 1000$ respectivement.

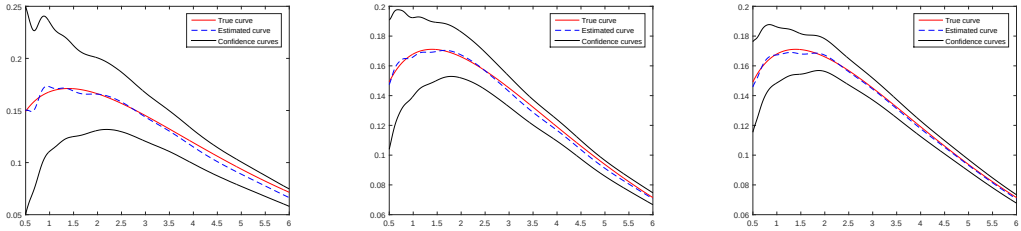


FIGURE 4.6 – Intervalles de confiances de l'estimateur de la densité $W(4.3, 1.3)$, pour $PC = 40\%$, $n = 100, n = 500$ et $n = 1000$ respectivement.

Conclusion Générale

Dans cette thèse, nous nous sommes intéressés à l'étude des modèles non paramétriques dans le but d'estimer la fonction de densité et le taux de hasard. Les résultats que nous avons établis sont liés aux propriétés asymptotiques des estimateurs à noyaux asymétriques pour un modèle de censure à droite.

Notre intérêt a dans un premier temps, porté sur l'utilisation de la technique de lissage par le théorème de Hille appliqué sur l'estimateur de Kaplan-Meier pour obtenir un estimateur lisse de la fonction de répartition dans le cas i.i.d et censuré à droite. Nous avons établi la convergence uniforme presque sûre avec vitesse de notre estimateur en employant les VC-classes.

Nous avons considéré dans un second temps le problème d'effet de bord pour la fonction de densité et taux de hasard. Nous avons proposé un nouvel estimateur à noyau asymétrique pour la fonction densité performant aux bornes lorsque la variable d'intérêt est sujette à une censure aléatoire à droite. En conséquence, deux nouveaux estimateurs de la fonction taux de hasard sont suggérés, le premier en divisant l'estimateur obtenu de la densité par notre estimateur lisse de la fonction de répartition et le deuxième en utilisant l'estimateur classique de Kaplain-Meir. Nous avons établi la convergence uniforme presque sûre sur un compact de nos estimateurs en quantifiant la vitesse, puis leurs normalités asymptotiques et nous avons donné l'expression explicite des termes asymptotiquement dominants du biais et de la variance. Une conséquence directe de ce dernier résultat est la construction d'intervalles de confiance asymptotiques ponctuels dont nous avons étudié les propriétés à travers des simulations. Une large étude numérique sur données générées a été entreprise dans le but de renforcer notre résultat théorique. Nous avons appliqué la nouvelle approche sur un exemple de données réelles liées à un cancer de la peau.

Perspectives

Comme suite de ce travail et quelques travaux envisageables, et en relation avec cette thèse, nous envisageons de :

- établir des résultats de type Berry-Esseen afin de quantifier la vitesse de convergence de la normalité.

- Étudier l'erreur quadratique moyenne intégrée (MISE), pour cherche les paramètres théorique optimaux par les différentes méthode de sélection comme la validation croisée.

- Adapter nos résultats au cas d'une autre fonctionnelle comme par exemple : la régression, la densité spectrale, quantiles, ou au cas d'un type de dépendance tel que l'association et l'α-mélange.

- Étendre nos resultats au cas de données tronquées, tronquées et censurées, ou doublement censurée.

- Aussi peut être aller vers des études de fonction de répartition, et de densité conditionnelles.

Annexe

Écriture additive de l'estimateur de Kaplan-Meier

L'estimateur de Kaplan-Meier $\widehat{F}_n$ est une fonction étagée, croissante, et continue à droite, qui ne saute qu'aux instants de varie durée de vie. Donc on peut l'écrire sous forme additive, comme une moyenne pondérée de termes identiques où le poids est lié à la censure, comme suit

$$\widehat{F}_n(x) = \sum_{i=1}^{n} \; W_i \;\; \mathbb{I}_{\{Y_i \leq x\}}$$

avec les poids W_i ce sont les sauts de $\widehat{F}_n(.)$ aux points $Y_{(i)}$. Ainsi pour $i = 1, \ldots, n$ nous avons

$$\begin{aligned}
W_{(i)} &= \widehat{F}_n(Y_{(i)}) - \widehat{F}_n(Y_{(i)}^-) \\
&= \left(1 - \prod_{j:Y_{(j)} \leq Y_{(i)}} \left(1 - \frac{1}{n-j+1}\right)^{\delta_{(j)}}\right) - \left(1 - \prod_{j:Y_{(j)} \leq Y_{(i-1)}} \left(1 - \frac{1}{n-j+1}\right)^{\delta_{(j)}}\right) \\
&= \prod_{j:Y_{(j)} \leq Y_{(i-1)}} \left(1 - \frac{1}{n-j+1}\right)^{\delta_{(j)}} \left[1 - \left(1 - \frac{\delta_{(i)}}{n-i+1}\right)\right] \\
&= \frac{\delta_{(i)} \left(1 - \hat{F}_n\left(Y_{(i)}^-\right)\right)}{n-i+1}.
\end{aligned}$$

avec $\widehat{F}_n(t^-)$ dénotes la limite à gauche de $\widehat{F}_n(.)$ at t.

Classes de Vapnik-Cervonenkis (V-C classes)

On se donne un espace métrique (T, d) et un $\epsilon > 0$. Le nombre de ϵ-recouvrement de l'espace métrique (T, d) noté $\mathcal{N}(T, d, \epsilon)$ est défini comme le nombre minimal de boules ouvertes d de centres dans T et de rayon ϵ, requis pour couvrir l'ensemble E. Une classe de fonctions mesurables $\mathcal{H}$ sur l'espace mesuré $(S, \mathcal{S})$, est une V-C classe de fonctions par rapport à l'enveloppe H s'il existe

une fonction mesurable H presque partout finie avec $|h| \leq H$ pour toute fonction $h \in \mathcal{H}$, et des nombres réels A et υ tels que :

$$\mathcal{N}(\mathcal{H}, ||.||_{L^2(\mathbb{P})}, \epsilon||H||_{L^2(\mathbb{P})}) \leq \left(\frac{A}{\epsilon}\right)^{\upsilon}$$

pour tout $\epsilon \in (0,1)$ et toute mesure de probabilité $\mathbb{P}$ sur $(S, \mathcal{S})$ pour laquelle :

$$\int M^2 \, d\mathbb{P} < \infty$$

Pour approfondir cette notion de V-C classe, vous pouvez consulter le livre de Pollard (1984) [57].

Lemme 4.3. *(Giné and Guillou (1999), Lemme 3)*

(a) *Si $\mathcal{H}$ est finie alors $\mathcal{H}$ est une V-C classe par rapport à l'enveloppe* $\max\{|h| : h \in \mathcal{H}\}$.

(b) *Si $\mathcal{H} = \{h_x : x \in J\}$ où J est une partie de $\mathbb{R}$ et $0 \leq h_x(s) \leq h_y(s)$ pour tout $x < y$ et $s \in S$, alors $\mathcal{H}$ est une V-C classe par rapport à $H = \sup\{|h| : h \in \mathcal{H}\}$.*

(c) *Si $\mathcal{H}_1$ et $\mathcal{H}_2$ sont deux V-C classe par rapport à H_1 et H_2 respectivement, alors $\{h_1 + h_2 : h_i \in \mathcal{H}_i\}$; $\{h_1 - h_2 : h_i \in \mathcal{H}_i\}$ sont des V-C classes par rapport à $(H_1 + H_2)^{\frac{1}{2}}$.*

Pour la preuve de ce lemme, veuillez consulter l'article de Giné and Guillou (1999) [38]. L'inégalité qui va suivre est celle de Talagrand.

Théorème 4.3. *(Giné and Guillou (2002), Théorème 2.1) Si $\mathcal{F}$ une V-C classe mesurable et uniformément bornée de fonctions. Soit σ^2 et U des nombres telle que $\sigma^2 \geq \sup_{f \in \mathcal{F}} Var_P f, U \geq \sup_{f \in \mathcal{F}} Var_P ||f||_\infty$ et $0 < \sigma \leq U$. Alors, il existe une constante B, et des constantes C, et L, ne dépendant que des caractéristiques A, et υ de la V-C classe, telle que*

$$P\left\{\left|\left|\sum_{i=1}^{n}(f(\xi_i) - E(f(\xi_i)))\right|\right|_{\mathcal{F}} \geq t\right\} \leq L\exp\left\{-\frac{1}{L}\frac{t}{U}\log\left[1 + \frac{tU}{L\left(\sqrt{n}\sigma + U\sqrt{\log\frac{AU}{2}}\right)^2}\right]\right\}$$

avec

$$t \geq C\left[U\log\frac{AU}{\sigma} + \sqrt{n}\sigma\sqrt{\log\frac{AU}{\sigma}}\right].$$

Bibliographie

[1] Abadir K. M., and Lawford S. (2004). Optimal Asymmetric Kernels. Economics Letters, 83 : 61-68. http://www.sciencedirect.com/science/article/pii/S0165-1765(03)00327-6

[2] Andersen P.K., Borgan O., Gill R.D. and Keiding N. (1993). *Statistical Models Based on Counting Processes*. Springer Series in Statistics .

[3] Bagai I. and Prakasa Rao B.L.S. (1996). Kernel Type Density Estimates for Positive Valued Random Variables. Sankhya. The Indian Journal of Statistics, Series A (1961-2002), vol. 57, no. 1, 1995, pp. 56–67. JSTOR, https://www.jstor.org/stable/25051030

[4] Blum J., Susarla V. (1980). Maximum deviation theory of density and failure rate function estimates based on censored data. *In : Krishnaich, P. R. (Ed.), Multivariate Analysis V. North-Holland, Amsterdam*. 213-222.

[5] Bouezmarni T., El Ghouch A., Mesfioui M. (2011). Gamma Kernel Estimators for Density and Hazard Rate of Right-Censored Data. *Journal of Probability and Statistics* 2011 :1-16 Article ID 937574.

[6] Bouezmarni T., Rolin J-M. (2003). Consistency of the beta kernel density function estimator. *Canad. J. Statist.*, 31 , No. 1, 89-98.

[7] Bouezmarni T., and Scaillet O. (2005). Consistency of Asymmetric Kernel Density Estimators and Smoothed Histograms with Application to Income Data. Econometric Theory, 21 : 390-412. https://ideas.repec.org/a/cup/etheor/v21y2005i02p390-412_05.html

[8] Bouezemarni T., Karunamuni R.J. & Alberts T. (2005). On boundary correction in kernel density estimation. Statistical Methodology 2, 191-212.

[9] Bosq D. (1985). *Nonparametric Statistics for stochastic Processes estimation and prediction*. Springer.

[10] Chaubey Y.P, Sen A, Sen P.K. (2007). A new smooth density estimator for non-negative random variables. *Technical Report No. 1/07*, Department of Mathematics and Statistics, Concordia University, Montreal, Canada.

[11] Chaubey Y.P, and Sen P.K. (1996). On smooth estimation of survival and density functions. Statistics & Risk Modeling, vol. 14, no. 1, 1996, pp. 1-22.https://doi.org/10.1524/strm.1996.14.1.1

[12] Chaubey Y.P, and Sen P.K. (2013). On nonparametric estimation of the density of a non-Negative function of observations.*Calcutta Statistical Association Bulletin.* (Special 8th Triennial Symposium Proceedings Volume). 65 :257-260.

[13] Charpentier A. et Flachaire E. (2015) Log-Transform Kernel Density Estimation of Income Distribution. L'Actualité économique. Volume 91, numéro 1-2, mars–juin 2015. DOI:https://doi.org/10.7202/1036917ar

[14] Chen S.X. (1999). Beta kernel estimators for density functions. *Comput Statist and Data Anal* 31 : 131-145. https://doi.org/10.1016/S0167-9473(99)00010-9

[15] Chen S.X. (2000a). Beta Kernel Smoothers For Regression Curves. *Statistica Sinica* 10 :73-91. https://www.jstor.org/stable/24306705

[16] Chen S.X. (2000b). Probability Density Function Estimation Using Gamma Kernels. *Ann. Inst. Statist.* Math, 52, 3 : 471-480. https://doi.org/10.1023/A:1004165218295

[17] Cheng M.Y. (1997). A Bandwidth Selector for Local Linear Density Estimators. *The Annals of Statistics* 25, 3 : 1001-1013. https://doi.org/10.1214/aos/1069362735

[18] Cheng M.Y, Fan J, Marron J.S. (1997). On automatic boundary corrections. *The Annals of Statistics* 25, 4 : 1691-1708. http://dx.doi.org/10.1214/aos/1031594737

[19] Chung K.L. (2001). A course in Probability Theory, third edition, ACADEMIC PRESS.

[20] Cleveland W., and Devlin S. (1988). Locally weighted regression : an approch to regression analysis by locally fitting. J. Amer. Statist. Assoc, vol 89,No. 403., 595-610. https://doi.org/10.2307/2289282

[21] Cline D.B., Hart J.D. (1991). Kernel estimation of densities with discontinuities or discontinuous derivatives. Statistics. 22, 69–84 . http://dx.doi.org/10.1080/02331889108802286

[22] Cowling A. and Hall H. (1996). On Pseudodata Methods for Removing Boundary Effects in Kernel Density Estimation. Journal of the Royal Statistical Society. Series B (Methodological), Vol. 58, No. 3 (1996), pp. 551-563. https://doi.org/10.1111/j.2517-6161.1996.tb02100.x

[23] Deheuvel P. and Einmahl H.J. (2000). Functional limit laws for the increments of Kaplan-Meier product limit processes and applications. *Ann. Proba* 28 : 1301-1335.

[24] Devroy L. and Gyrfi L. (1985). *Non parametric Density Estimation the L_1 View* ;Wiley Series in Probability and Mathematical Statistics.

[25] Engle R. (2000). The Econometrics of Ultra-High Frequency Data. Econometrica, 68, 1-22.

[26] Eugen Slutsky (1925). Uber stochastische Asymptoten und Grenzwerte, Metron,5.3 (1925), p. 3-89.

[27] Fan J. et Gijbels I. (1992). Variable bandwidth and local linear regression smoothers,The Annals of Statistics, p. 2008–2036. https://www.jstor.org/stable/2242378

[28] Fernandes M. and Grammig J. (2000). Nonparametric Specification Tests for Conditional Duration Models. Forthcoming in J. Econ.

[29] Fernandes M., Eduardo F. Mendes E.F. & Scaillet O. (2014). Testing for symmetry and conditional symmetry using asymmetric kernels. Ann Inst Stat Math 67, 649–671 . https://doi.org/10.1007/s10463-014-0469-6

[30] Ferraty F. and Vieu P. (2006). *Non-parametric functional data analysis : Theory and Practice*. Springer.

[31] William Feller W. (1968) . *An Introduction to Probability Theory and Its Applications*. Wiley ; 3rd edition (January 1, 1968). https://www.amazon.com/Introduction-Probability-Theory-Applications-Vol/dp/0471257087

[32] Gasser T. and Müller H.G. (1979). Kernel Estimation of Regression Functions. In Smoothing Techniques for Curve Estimation, Lecture Notes in Mathematics 757, pp. 23–68. Springer-Verlag, Berlin.https://link.springer.com/chapter/10.1007/BFb0098489

[33] Gasser T., Müller H.G. & Mammitzsch V. (1985). Kernels for nonparametric curve estimation. Journal of the Royal Statistical Society 2, 238-252.

[34] Gawronski N. and Stadtmuller U. (1980). On Density Estimation by Means of Poisson's Distribution. Scand. J. Statist., 7, 90-94.

[35] Gawronski N. and Stadtmuller U. (1981). Smoothing histograms by means of lattice and continuous distributions. Metrika, 28, 155-164.

[36] Geoffrey R. Grimmett and David R. Stirzaker (2001). Probability and Random Processes. Third edition, Published in the United States by Oxford University Press Inc., New York.

[37] Ghettab S. Gussoum Z. (2022). Asymptotic properties of asymmetric kernel estimators for non-negative and censored data, *Communications in Statistics - Theory and Methods*. https://doi.org/10.1080/03610926.2022.2150059

[38] Giné E. and Guillou A. (1999). Law of the iterated logarithm for censored data. *The Annals of Probability* 27, 4 : 2042-2067.

[39] Giné E. and Guillou A. (2002). Rates of strong uniform consistency for multivariate kernel density estimators. *Ann. I. H. Poincaré-PR* 38, 6 : 907-921.

[40] Gramacki A. Nonparametric Kernel Density Estimation and Its Computational Aspects. Studies in Big Data, vol 37. Springer International Publishing AG 2018, https://doi.org/10.1007/978-3-319-71688-6

[41] Gustafsson J., Hagmann M., Nielsen J.P. & Scaillet O. (2007). Local transformation kernel density estimation of loss distributions. Journal of Business and Economic Statistics. https://ideas.repec.org/a/bes/jnlbes/v27i2y2009p161-175.html

[42] Hagmann M., and Scaillet O. (2007). Local Multiplicative Bias Correction for Asymmetric Kernel Density Estimators. Journal of Econometrics, 141 : 213-249. https://doi.org/10.1016/j.jeconom.2007.01.018

[43] Hall P., Park B.U. (2002). New methods for bias correction at endpoints and boundaries. Ann. Stat.30(5), 1460–1479 ,http://www.jstor.org/stable/1558721

[44] Hall P., Wehrly T.E. (1991). A geometrical method for removing edge effects from kerneltype nonparametric regression estimators, J. Amer. Stat. Assoc., 86, 665-672,

[45] Hjort N., Jones M. (1996). Locally parametric nonparametric density estimation. Ann. Stat. 24(4), 1619–1647 , http://www.jstor.org/stable/2242742

[46] Jones M.C. (1993). Simple boundary correction for kernel density estimation. Stat. Comput. 3(3), 135–146, https://doi.org/10.1007/BF00147776

[47] Kaplan E.L. and Meier P. (1958). Nonparametric estimation from incomplete observations. *J. Amer. Statist* 53 : 457-481.

[48] Karunamuni R.J. , Alberts T. (2005). On boundary correction in kernel density estimation. Statistical Methodology, Volume 2, Issue 3, September 2005, Pages 191-212. https://doi.org/10.1016/j.stamet.2005.04.001

[49] Kuhnt J.W. and Padgett W.J. (1997). Local bandwidth selection for kernel density estimation from right censored data based on asymptotic mean absolute error. *Nonlinear Analysis, Theory, Methods and Application* 30, 7 : 4375-4384.

[50] Kuruwita C.N., Kulasekera K.B. and Padgett W.J. (2010). Density estimation using asymmetric kernels and Bayes bandwidths with censored data. *Journal of statistical planning and inference* 140 : 1765-1774.

[51] Linton O. and Nielsen J.P. (1994). A Multiplicative Bias Reduction Method for Nonparametric Regression. Statistics & Probability Letters, 19, 181-187. https://doi.org/10.1016/0167-7152(94)90102-3

[52] Lo S.H., Mack Y.P. and Wang J.L. (1989). Density and hazard rate estimation for censored data via a strong representation of the Kaplan-Meier estimator. *Probability Theory and Related Fields* 80, 461-473.

[53] Marron J.S. and Padgett W.J. (1987). Asymptotically optimal bandwidth selection for kernel density estimators from right-censored samples. *The annals of statistics* 15, 4 : 1520-1535.

[54] Marron J.S.,Ruppert D. (1994). Transformations to reduce boundary bias in kernel density estimation. J. R. Stat. Soc. Ser. B (Methodol.) 56, 653–671 , http://www.jstor.org/stable/2346189

[55] Müller H.-G. (1991). Smooth optimum kernel estimators near endpoints. Biometrika 78, 521–530, http://www.jstor.org/stable/2337021

[56] Parzen E. (1962). On Estimation of a Probability Density Functions and Mode. *Annals. Mathe. Statist.* 33, 3 : 1065-1076.

[57] Pollard D. (1984). Convergence of stochastic processes. Springer Verlag. Berlin.

[58] Rice J (1984). Boundary modification for kernel regression. comm, Statist, 13, 893-900, https://doi.org/10.1080/03610928408828728

[59] Rosenblatt M. (1956). Remarks on Some Nonparametric Estimates of Density Functions. *Annals. Mathe. Statist.* 27 : 832-837.

[60] Ruppert D. et Wand M. P. (1994).Multivariate Locally Weighted Least Squares Regression.Ann. Statist. 22 (3) 1346 - 1370, September, 1994. https://doi.org/10.1214/aos/1176325632

[61] Scaillet O. (2004). Density Estimation using inverse and reciprocal Gaussian kernels. *Journal of Nonparametric Statistics*. 16 : 217-226.

[62] Schuster E.F. (1985). Incorporating support constraints into nonparametric estimators of densities. Communication in Statistics Theory and Methods 5, 1123-1136. https://doi.org/10.1080/03610928508828965

[63] Silverman B.W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London. http://dx.doi.org/10.1007/978-1-4899-3324-9

[64] Scott D.W. (2015). Multivariate Density Estimation : Theory, Practice, and Visualization, 2nd Edition. WILEY Series in Probability and Statistic.https://www.wiley.com/en-be/Multivariate+Density+Estimation%3A+Theory%2C+Practice%2C+and+Visualization%2C+2nd+Edition-p-9781118575536

[65] Tibshirani R., and Hastie T. (1987). Local likelihood estimation, Journal of the American Statistical Association, vol. 82, no 398, p. 559–567. 12. https://doi.org/10.2307/2289465

[66] Turnbull B.W. (1974). Nonparametric Estimation of a Survivorship Function with Doubly Censored Data. Journal of the American Statistical Association 69 (1974) : 169–173.

[67] Wand M.P. and Jones M.C. (1995). *Kernel Smoothing*. CHAPMAN and HALL.CRC Boca Raton London New York Washington, D.C.

[68] Wand M.P., Marron J.S. and Ruppert D. (1991). Transformations in density estimation : *Journal of the American Statistical Association* 86 : 343-361. https://www.tandfonline.com/doi/abs/10.1080/01621459.1991.10475041.

[69] Watson G.S. and Leadbetter M.R. (1964). Hazard analysis I. *Biometrika* 51 : 175-184. https://doi.org/10.2307/2334205.

[70] Zhang S., Karunamuni R., Jones M.C. (1999). An improved estimator of the density function at the boundary. J. Am. Stat. Assoc. 94(448), 1231–1241 , http://www.jstor.org/stable/2669937

[71] Zhang S., Karunamuni R.J. (1998). On kernel density estimation near endpoints. J. Stat. Plan. Inference. 70, 301–316, https://doi.org/10.1016/S0378-3758(97)00187-02669937