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

Université Frères Mentouri Constantine
Faculté des sciences exactes
Département de mathématiques

N° d'ordre : 37/DS/2023
N° de série : 04/math/2023

THÈSE

présentée pour l'obtention du diplôme
de doctorat en sciences
en mathématiques

« Estimation Non-Paramétrique de La Fonction de Répartition Bivariée : Approche de Bernstein »

Par
Dib Khaled

OPTION : Probabilités et Statistique

Devant le jury :

Président	M^me	S. Belaloui	Prof.	Université frères Mentouri, Constantine 1
Directrice de thèse	M^me	F. Messaci	Prof.	Université Salah Boubnider, Constantine 3
Co-Encadreur de thèse	M.	T. Bouezmarni	Prof.	Université de Sherbrooke, Canada
Examinateur	M.	M. Boussaboua	Prof.	Université Salah Boubnider, Constantine 3
Examinatrice	M^me	I. Laroussi	M.C.A.	Université frères Mentouri, Constantine 1

Soutenue le : 22 Juin 2023

Remerciements

Je tenais à profiter de cette occasion pour exprimer ma profonde gratitude pour chacun des membres de Jury pour l'honneur qu'ils m'ont fait de s'intéresser à ce travail et d'avoir accepté de l'évaluer.

Mes remerciements s'adressent :

A mes directeurs de thèse, madame la Professeur Fatiha MESSACI et monsieur le professeur Taoufik BOUEZMARNI qui ont accepté à Co-encadrer ce travail. La réalisation de ce projet n'aurait pas été possible sans votre expertise, vos conseils éclairés et votre dévouement constant.

A madame la professeur Soheir BELALOUI, pour avoir présidé le jury lors de ma soutenance de thèse de doctorat. Votre présence à la tête de cette assemblée a été un honneur et une source de grande confiance pour moi.

A monsieur le professeur Moussedek BOUSSABOUA et madame Docteur Ilhem LAAROUSSI, d'avoir accepté de prendre part à l'examen de ma thèse. Votre engagement en tant qu'examinateurs est une marque de votre dévouement envers la recherche académique et je me sens honoré par votre présence.

Je souhaiterais également remercier tous les professeurs, les chercheurs et le personnel qui ont contribué à mon cheminement académique.

Enfin, je tiens à exprimer ma reconnaissance envers ma famille, mes amis et mes proches pour leur soutien indéfectible tout au long de cette aventure. Leur soutien moral, leurs encouragements et leur compréhension face aux exigences de cette thèse ont été d'une importance capitale pour moi.

DIB Khaled

Constantine, Juin 2023

Table des matières

Chapitre 1

Introduction générale

Le problème de l'estimation des fonctionnelles est rencontré dans de très nombreux domaines. La démarche statistique se fonde sur le traitement des données, en général une série de réalisations d'une loi. Le but peut être d'estimer la distribution théorique qui représente la série observée.

Dans la pratique, il est parfois difficile de trouver le modèle paramétrique convenable. Par conséquent, le cadre paramétrique classique n'est pas toujours un choix judicieux. L'approche non paramétrique lève cette difficulté car elle n'exige aucune hypothèse sur la distribution du phénomène étudié et le modèle n'est pas décrit par un nombre fini de paramètres. L'avantage principal de ce modèle c'est qu'il est plus général et donc plus robuste. Son inconvénient notable est sa vitesse de convergence relativement plus lente par rapport au modèle paramétrique.

La fonction de répartition est un concept essentiel en statistique. Elle permet de décrire et de caractériser la distribution d'une loi théorique. Partant d'un ensemble d'observations $X_1, X_2, \ldots, X_n$, constituant un échantillon suivant une variable aléatoire X, souvent inconnue dans la pratique, on aimerait construire un estimateur

pouvant approcher avec une certaine précision la vraie fonction de répartition. Plusieurs estimateurs de cette fonction ont été proposés dans la littérature à travers différentes approches.

Cette thèse est consacrée à l'estimation non paramétrique de la fonction de répartition bivariée en se basant sur les polynômes de Bernstein. Ces polynômes furent introduits par Bernstein (1912) [4], dans le but de donner une preuve, basée sur le calcul des probabilités, du théorème de Weierstrass . Ce théorème porte sur l'approximation des fonctions continues sur un intervalle de forme $[a, b]$, grâce à une famille de polynômes, et portera son nom par la suite. t Motivé par ce résultat d'approximation uniforme, Vitale (1975) [49] a construit un estimateur de type Bernstein pour une fonction de densité univariée f. Un résultat de convergence au sens d'erreur en moyenne quadratique a été établi dans le même travail. Une généralisation de cet estimateur, au cas multivarié, a été introduite par Tenbusch (1994) [47].

Plus récemment, Babu et al. (2002) [1] ont étudié l'estimateur de la fonction de répartition basé sur les polynômes de Bernstein. Cet estimateur est un lissage de l'estimateur empirique F_n de la fonction de distribution F. En dérivant l'estimateur en question, on obtient une autre forme de l'estimateur de la fonction de densité étudié par [49].

Leblanc (2009 et 2012) [25] et [26] a travaillé sur les propriétés asymptotiques de l'estimateur de Bernstein. En fait il a montré que cet estimateur possède, sous certaines conditions, des propriétés similaires à celles des estimateurs classiques comme la fonction de répartition empirique et les estimateurs construits par la méthode des noyaux.

Pour le cas multivarié, Babu et Chauby (2006) [2] ont proposé un estimateur non paramétrique de la fonction de répartition bivariée. Ils ont montré sa convergence

presque sûre. Récemment Belaia (2016) [3] a étudié quelques propriétés asymptotiques du même estimateur notamment la normalité asymptotique. Tous les travaux cités dans ce qui précède sont développés sur un modèle des données complètement observées.

Dans la pratique, des perturbations peuvent empêcher le praticien d'observer des données complètes pour tout individu de l'échantillon étudié. Ce type de données incomplètes emmène les statisticiens vers les différents modèles dits "modèles de censures". Le plus fréquemment rencontré dans pratique est celui de la censure à droite. Il peut intervenir pour différentes raisons, comme la fixation du temps de l'étude. En effet, si on s'intéresse au temps de survie à une maladie grave, à la fin de l'étude certains individus peuvent être encore en vie et il est impossible d'observer leurs vrais temps de survie mais on dispose seulement de valeurs qui leur sont inférieures. Une autre raison est le cas d'individus perdus de vue dans l'étude ou qui disparaissent pour une autre cause que celle d'intérêt (accident, migration au autre). C'est ce modèle de censure qui est étudié dans cette thèse et pour lequel Kaplan et Meier (1958) [22] ont donné un estimateur de la fonction de répartition univariée, Stute (1993) [43] en a proposé une généralisation au cas multivarié. Ce fut le point de départ de cette thèse, dont L'objectif est d'explorer l'approche de Bernstein pour l'estimation de la fonction de répartition bivariée, par lissage de l'estimateur de Stute (1993). L'avantage de cette démarche est d'obtenir un estimateur dérivable dont la dérivée constitue un estimateur de la fonction de densité. Finalement, une estimation non paramétrique de la copule est déduite. Des propriétés asymptotiques des estimateurs proposées ont été établies. Il s'agit de la convergence presque sûre uniforme avec taux pour les estimateurs de la fonction de répartition bivariée et de la copule ainsi que la normalité asymptotique pour le premier.

Plan de la thèse Cette thèse est composée de cinq chapitres et est structurée de la manière suivante.

Le **chapitre 2** est consacré à la présentation de modèles de censure. Nous avons commencé par une introduction à l'analyse de survie, les modèles de censure à droite avec trois scénarios de censure à droite. Ensuite, nous avons exposé quelques fonctions utiles dans le contexte en question. A la fin du chapitre, nous avons abordé le sujet de l'estimation non paramétrique en présence de données censurées dans le cas univarié puis le cas bivarié. Des résultats du comportement asymptotique des estimateurs sont exposés.

Le **chapitre 3** est dédié à la problématique essentielle de cette thèse. Nous commençons par introduire l'approche de Bernstein qui constitue la base de la théorie de l'estimation par les polynômes de Bernstein. Nous nous intéressons par la suite aux principaux résultats de notre contribution au sujet de l'estimation non paramétrique de la fonction de répartition bivarié en présence de censure. Des propriétés asymptotiques de notre estimateur sont étudiées à la fin de ce chapitre, à savoir la convergence presque uniforme avec taux et la normalité asymptotique. Ce chapitre a constitué la partie principale de ma publication intitulée "Nonparametric bivariate distribution estimation using Bernstein polynomials under right censoring" apparue dans la revue *Communications in Statistics - Theory and Methods* en 2021.

Le **Chapitre 4** présente un deuxième volet de cette thèse. Dans un premier temps, nous commençons par exposer un peu de littérature sur les copules et expliquer le choix de la copule pour modéliser la structure de dépendance. Ensuite, nous présentons quelques familles de copules paramétriques les plus utilisées en pratique et leurs propriétés fondamentales comme les copules elliptiques et les copules Archimédiennes. Nous terminons le chapitre en proposant un estimateur non pa-

ramétrique de la copule ; cet estimateur est basé sur notre estimateur introduit au le troisième chapitre. Un résultat de convergence presque sûre avec taux de ce nouvel estimateur est établi à la fin du chapitre.

Le **Chapitre 5** présente des applications sur des données simulées afin de valider la performance de notre estimateur de la fonction de répartition bivariée. À la fin de ce chapitre, nous appliquons notre estimateur pour analyser des données réelles recueillies de la littérature du domaine de l'assurance.

A la fin de ce travail, nous présentons des conclusions et des perspectives qui représentent la suite éventuelle de ce travail.

English version

Title of the Thesis : Nonparametric bivariate distribution estimation and copula using Bernstein polynomials

Introduction Estimating multivariate distribution function is one of the important subjects in statistics. The nonparametric approaches require less assumptions and can avoid the potentially large bias caused by the misspecification of the model of the parametric methods.

The empirical distribution is the nonparametric estimator mostly used. However, the discrete form of this estimator leads to a discontinuous estimator with derivative equal to zero almost everywhere. Smoothing the empirical by kernels is investigated extensively. Another technique to smooth the empirical distribution is based on Bernstein estimator. The idea is based on the fact that any continuous function on $[0, 1]$ can be approximated by a polynomial. This result was firstly introduced by Bernstein (1912) (see also Lorentz (1986)). The Bernstein polynomial estimator was discussed for the first time by Vitale (1975) in order to estimate a density function defined on $[0, 1]$. For the bivariate dimension, Babu et al. (2002) used Bernstein polynomials for estimating the distribution function. Later, this estimator is investigated by Leblanc (2009). For the multivariate case, the Bernstein distribution estimator is studied by Babu and Chaubey (2006) and recently by Belalia (2016). All these estimators are proposed for complete data.

In lifetime data analysis, variables of interest are not usually completely observed and they can be subject of censorship. Various scenarios of censoring are possibles and many estimators, depending the censoring schemes, are proposed for the multi-

variate distribution, see for example Dabrowska (1988), van der Laan (1996), Stute (1993), Lin Ying (1993), Akritas Van Keilegom (2003) and Lopez Saint-Pierre (2012). Recently, Gribkova (2015) and in order to estimate the copula functions, introduced a class of estimators of the multivariate distribution function when censoring is present.

In this thesis, we consider the bivariate case and where only one random variable is right censored. We propose to smooth the empirical distribution version of Stute (1993) using Bernstein polynomials. The new estimator is easy to implement, free from boundary bias, and the density function, copula and copula density estimators can be derived. This manuscript is composed of six chapters.

Chapter 1 presents the general framework and the contributions of our thesis. We give a brief introduction and a short review of our results.

Chapter 2 introduces basic notions in survival analysis : censored variable, usual functions and Kaplan-Meier estimator. Then, we focus on the studying the bivariate version of the Kaplan-Meier estimator. We provide an exposition of some relevant results.

In **Chapter 3**, we expose our principal contribution in non parametric estimation. An estimator of the joint distribution function of two random variables X and Y is proposed. Especially, we consider the case when one of the two variables, says Y, is subject to right censoring. The newly proposed estimator is built using Bernstein polynomials. The asymptotic properties of this estimator such as, bias, variance and asymptotic normality are provided. Also, we prove the strong uniform convergence. This chapter was The main part of my publication entitled "Nonparametric bivariate distribution estimation using Bernstein polynomials under right censoring" appeared in the journal *Communications in Statistics - Theory and Methods* in 2021.

Chapter 4 is devoted to nonparametric copula estimation under univariate cen-

soring. We provide a short introduction to the copula theory and some parametric copulas that we need in simulation studies. A Nonparametric estimator of the copula function is constructed using our proposed estimator in Chapter 3. The consistency of this estimator is established and we give the convergence rate of this consistency. Due to the smoothness of this copula estimator, we give a nonparametric copula density estimator. This chapter is a project of a paper to be submited soon.

Chapter 5 provides an investigation of the finite sample properties of the proposed estimator through simulated examples. This numerical study showed that our estimator is Efficient, in sense of integrated bias and variance, compared to empirical estimator introduced by Stute (1993). Finally, the proposed estimator is applied to analyze the Loss-ALAE dataset from insurance.

Finaly, we present a conclusion and a discussion about our research perspectives.

Chapitre 2

Estimation de la fonction de répartition basée sur des données censurées à droite

2.1 Introduction à l'analyse de survie

Dans plusieurs domaines d'application de la statistique (médecine, sociologie, ingénierie), on s'intéresse au temps pour qu'un évènement survienne. Par exemple, le temps de survie chez des patients ayant une certaine pathologie spécifique. Aussi, en ingénierie, on cherche à modéliser la distribution de la durée écoulée avant qu'une machine tombe en panne. En économie, la durée du chômage est étudiée par plusieurs chercheurs.

Souvent on est amené dans de telles études, à analyser un échantillon d'observations incomplètes. Une des causes de cette incomplètude est due à la censure. En fait,

il ne s'agit pas de données manquantes puisqu'on observe une partie de la durée dans ce cas. Il existe plusieurs types de censure. Dans la section suivante, nous définissons cette notion censure.

2.2 Modèles de censure

La spécificité des données de survie est que l'on dispose le plus souvent de "données incomplètes". En fait, pour certains individus, on observe une valeur inférieure à la valeur réelle (il s'agit de la censure à droite) ou on observe une valeur supérieure à la valeur réelle (il s'agit de la censure à gauche). La durée est parfois censurée par intervalle lorsqu'elle est censurée à droite et à gauche. Pour plus de détails sur différents modèles de censure, le lecteur peut se référer à l'ouvrage de Lee Elisa T. [27].

Dans ces études, la variable d'intérêt, notée T, peut subir une censure par une variable dite variable de censure et notée C. Par exemple, dans le domaine des essais cliniques, la variable de censure C représente la date à laquelle un individu quitte l'étude pour une cause autre que la pathologie étudiée. Dans ce cas, l'observation est censurée à droite si $C < T$. Un exemple de la censure à gauche, où $C > T$, est l'âge auquel un individu commence à consommer la Marijuana [24] car certains individus se souviennent seulement qu'ils ont commencé cette consommation avant un certain âge. On se limitera, dans le cadre de cette thèse, à la censure à droite qui est la plus fréquente dans la pratique. La censure à droite intervient lorsqu'on n'observe pas évidemment T mais plutôt la variable $Y = T \wedge C$, où $u \wedge v = \min(u, v)$, et l'indicateur de censure $\delta = 1_{\{Y \leq C\}}$. Il existe trois scénarios différents de censure à droite.

Censure de type I : Dans ce cas, la censure est non-aléatoire, c-à-d la censure intervient dans un temps fixé. Plus concrètement, étant donné un nombre positif fixé c et un n-échantillon $T_1, \ldots, T_n$, les observations consistent en (Y_i, δ_i), où $Y_i = T_i \wedge c$ et $\delta_i = 1_{\{T_i<c\}}$ (l'indicateur de censure). Ce modèle est généralement utilisé dans les essais cliniques. En effet, il correspond à l'observation de la durée de survie de n individus (souris, machines, etc) au cours d'une expérience de durée prédéterminée c.

Exemple On applique ce modèle lors d'une étude sur des machines dans une usine ou des souris dans une urne. La censure correspond à la durée de l'étude. Par exemple, dans une étude clinique, on s'intéresse à la durée de survie de souris ayant subi un traitement dont on cherche à évaluer la performance. On observe soit la durée de vie d'une souris si le sujet décède, soit cette durée est censurée par une constante c (durée de l'étude) pour les souris encore vivantes à la fin de l'étude.

Censure de type II : Etant donné un nombre positif fixé r et un n-échantillon $T_1, \ldots, T_n$, les observations consistent en (X_i, δ_i) où $X_i = T_i \wedge T_{(r)}$ et $\delta_i = 1_{\{T_i<T_{(r)}\}}$ avec $T_{(1)} < T_{(2)} < \cdots < T_{(n)}$ est la statistiques d'ordre associée à $T_1, \ldots, T_n$. Ce modèle, souvent utilisé dans les études de fiabilité, correspond à l'observation de la durée de fonctionnement de n machines tant que r d'entre elles ne sont pas tombées en panne.

Exemple : Dans des études en fiabilité, on s'intéresse parfois à la durée de fonctionnement d'un type de machines. On prend à un moment initial un échantillon de n machines identiques. La fin de l'étude est déterminée par l'instant $T_{(r)}$ quand r machines d'entre elles tombent en panne. Evidemment, les durées de vie des $(n-r)$ mauchines restantes sont censurées par la valeur $T_{(r)}$.

La censure aléatoire : Ce type de censure est le scénario le plus fréquent en pratique. Dans ce cas, on observe un échantillon (Y_i, δ_i), où $Y_i = T_i \wedge C_i$ et $\delta_i = \mathbb{I}_{\{T_i<C_i\}}$ de taille n où $T_1, \ldots, T_n$ sont les durée réelles et $C_1, \ldots, C_n$ sont des variables aléatoires de censure.

Exemple : En médecine, ce modèle est souvent utilisé pour faire des essais thérapeutiques. Nous nous intéressons à la durée de vie des patients atteints d'une maladie. Pour tester la qualité d'un traitement, on commence notre étude par n patients. Durant l'étude, on observe la durée de vie réelle des patients décédés, mais des fois, pour d'autres raisons (déménagement, décés par une autre cause, etc) on observe une durée inférieure à la valeur réelle, c-à-d on observe la censure.

2.3 Fonctions utiles

Pour analyser la durée de vie on utilise souvent un ou plusieurs outils. Dans cette section, nous présentons des fonctions utiles pour étudier cette durée de vie : la fonction de survie, la fonction de risque et la fonction de risque cumulé.

Soit T une variable aléatoire réelle positive définie sur un espace probabilisé $(\Omega, \mathcal{A}, P)$. Cette variable aléatoire présente la durée pour qu'un "évènement d'intérêt" survienne. La fonction de répartition de T est donnée par $F(t) = P(T \leq t)$; Elle représente la probabilité que cet évènement d'intérêt se produise durant l'intervalle $[0, t]$. En analyse de survie ou en théorie de fiabilité, on s'intéresse souvent aux "chances" que l'évènement d'intérêt apparaisse après l'instant t. Elle se modélise par la fonction de fiabilité ou la fonction de survie notée par S et elle est définie sur $[0, +\infty[$

par

$$\begin{aligned} S(t) &= P(T > t) \\ &= 1 - F(t). \end{aligned}$$

Cette fonction permet de bien caractériser la loi de la variable T. En théorie de fiabilité, la fonction $S(t)$ permet d'avoir à chaque instant t, la probabilité que la durée du bon fonctionnement d'une machine dépasse le moment t ; Tandis qu'en médecine $S(t)$ n'est rien d'autre que la probabilité qu'un patient ait une durée de survie superieure à t. Concernant les propriétés analytiques de S, elles sont déduites directement de celles de F.
Si la loi de T admet une densité de probabilité f, alors la fonction de risque instantané, notée λ et définie par

$$\lambda(t) = \frac{f(t)}{S(t)} \quad , \quad si\, S(t) \neq 0.$$

Dans un contexte intuitif, $\lambda(t)$ est la probabilité que l'évènement d'intérêt se réalise à l'instant t sachant qu'il ne s'est pas réalisé avant l'instant t. Sous l'hypothèse de continuité sur f, on a :

$$\forall t > 0, \text{ si } S(t) \neq 0, \text{alors } \lambda(t) = \lim_{dt \to 0^+} \frac{P(t \leq T < t + dt | T \geq t)}{dt}. \tag{2.3.1}$$

Par conséquent, la fonction de risque instantané (ou de hasard) traduit l'évolution longitudinale du risque de survenue de l'événement. C'est ainsi que l'allure de cette fonction apporte une information immédiate sur les caractéristiques de l'objet étudié. En fait, en ingénierie, les parties décroissantes de la courbe représentative de λ correspondent à un phénomène de rodage de la machine, les parties croissantes à un phénomène de vieillissement et les parties planes à un phénomène de vie utile

A cette fonction de risque, est associée une autre fonction pour analyser est la fonction de risque cumulé définie, pour tout $t \in \mathbb{R}^+$ par :

$$\Lambda(t) = \int_0^t \lambda(u)du$$

Des résultats sont développés dans la littérature montrant des relations existantes entre les différentes quantités mathématiques précédemment définies. On peut démontrer facilement que si f existe et sous l'hypothèse $S(t) > 0$, on a les relations suivantes :

$$\lambda(t) = \frac{-d\log(S(t))}{dt} = \frac{f(t)}{S(t)},$$

$$\Lambda(t) = \int_0^t \lambda(t)du = -\log(S(t)).$$

$$\text{et } S(t) = \exp(-\Lambda(t)).$$

Par conséquent, en utilisant ces relations, on peut retrouver toutes ces fonctions grâce à une parmi elles. Ces relations sont souvent utilisées dans la théorie de l'estimation.

2.4 Estimation de la fonction de répartition avec données censurées

En pratique, les fonctions décrites dans la section précédente sont inconnues. Dans cette section nous allons présenter des estimateurs de la fonction de répartition en présence des données censurées. Nous allons distinguer le cas univarié et le cas bivarié.

2.4.1 L'estimateur de Kaplan-Meier

Soit $T_1, T_2, \ldots, T_n$ n variables aléatoires indépendantes et identiquement distribuées (i.i.d) de distribution F. L'estimateur empiriques de F pour le cas des données complètes est donné par :

$$F_n(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{\{T_i \leq t\}}. \tag{2.4.1}$$

Cet estimateur possède des propriétés intéressantes comme la convergence presque sûre uniforme, en plus d'être un estimateur sans biais. Cependant, ces propriétés sont valides pour les données complètes.

Maintenant, en présence des données censurées, on observe $Y_i = T_i \wedge C_i$ et $\delta_i = 1_{\{T_i < C_i\}}$. Utiliser F_n, basée sur les Y_i ou les valeurs réelles observées (les Y_i telles que $\delta_i = 1$) n'est pas un estimateur convergent de la distribution F.
Dans le cas de la censure, Kaplan-Meier a proposé un estimateur (EKM) de F, appelé aussi estimateur produit limite. Cet estimateur est donné par :

$$\widehat{F_n}(t) = 1 - \prod_{Y_{(j)} \leq t} \Big[\frac{n-j}{n-j+1}\Big]^{\delta_{(j)}} \tag{2.4.2}$$

où $Y_{(1)}, Y_{(2)}, \ldots, Y_{(n)}$ sont les statistiques d'ordre associées à (Y_n) et $\delta_{(i)}$ l'indice de censure associée à $Y_{(i)}$. Si la dernière observation est censurée, alors $\widehat{F_n}(t) \nrightarrow 1$, quand $t \nrightarrow 0$ prend des grandes valeurs. Pour éviter ce problème, on peut supposer que la dernière observation est complète, c-à-d $\delta_{(n)} = 1$.

Beaucoup de propriétés ont été établies pour l'estimateur de Kaplan-Meier. La première propriété qui a été démontrée est qu 'il un estimateur self-consistent (Efron (1967))[11]. Foldes et Rejto (1981) [16] ont démontré la convergence unifome presque sûre de cet estimateur avec un taux de convergence $\big(\frac{\log n}{n}\big)^{1/2}$. Sa normalité asymptotique a été établie par Breslow et Crowly (1974) [6]. Pour plus de détails sur les

propriétés de l'estimateur de Kaplan-Meier on peut diriger le lecteur vers le livre de Klein et Moeschberger [24].

Dans le graphique (2.1), sont tracés l'estimateur empirique et l'estimateur de Kaplan-Meier, de la fonction de distribution de la durée de survie de 80 malades atteints du cancer de la langue avec 34% de censure. Pour définir l'estimateur empirique, nous avons considéré toutes les données de l'échantillon comme étant des observations complètes. Ce graphe montre bien que l'estimateur empirique tend à surestimer la probabilité de décès comparée à l'estimateur de Kaplan-Meier. On voit clairement que le fait d'ignorer ou tenir compte de la nature des données amènent à deux conclusions différentes.

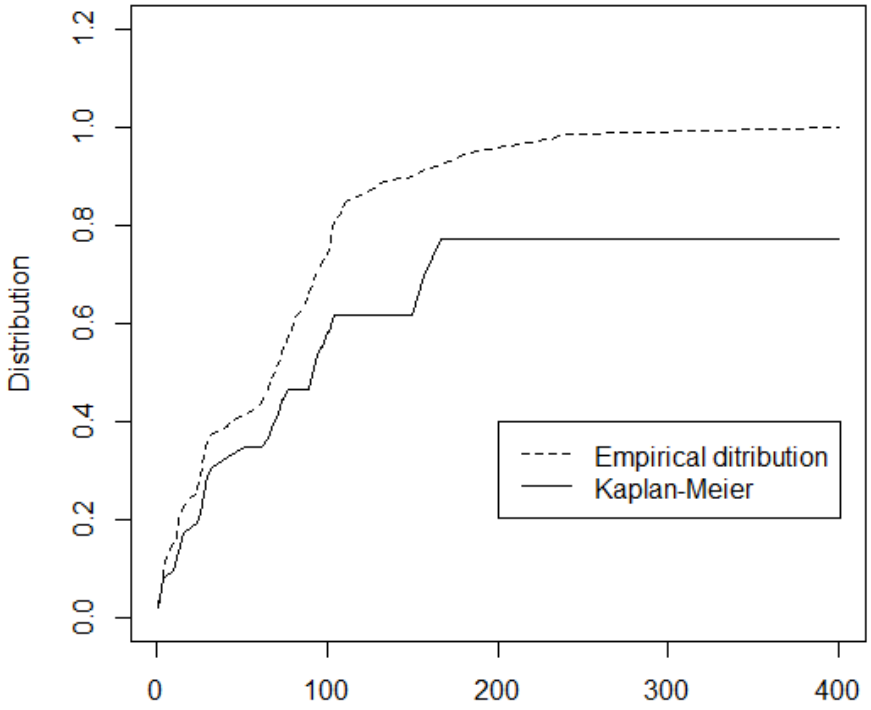


FIGURE 2.1 – Estimateurs de la fonction de distribution

Satten et Datta [40] ont proposé une réécriture de l'estimateur de Kaplan-Meier sous une forme semblable à l'expression de la fonction de répartition empirique F_n. En effet, il ont redéfini l'estimateur de Kaplan-Meier sous la forme suivante :

$$\widehat{F_n}(t) = \frac{1}{n}\sum_{i=1}^{n} w_{in}\mathbb{I}_{\{Y_i \leq t\}} \qquad avec \qquad w_{in} = \frac{\delta_i}{1-\widehat{G_n}(Y_i-)} \tag{2.4.3}$$

où $\widehat{G_n}(u)$ est l'estimateur de Kaplan-Meier de la fonction de répartition associée à la variable de censure C. $G(u-)$ est la limite à gauche de G à l'instant u.

2.4.2 L'estimation à noyau pour la fonction de répartition

Comme l'estimateur $\widehat{F_n}(t)$ est une fonction en escalier, des versions de lissage pour cet estimateur ont été largement étudiées. Le but c'est d'avoir des estimateurs qui jouissent des propriétés analytiques d'une fonction continue. L'estimateur à noyau est très abordé dans la littérature. Notons que l'estimateur à noyau de F pour les données complètes, basé sur l'échantillon $\{T_i\}_{i=1}^n$, est donné par :

$$\begin{aligned} F_n^K(t) &= \int \frac{1}{h} K\left(\frac{t-y}{h}\right) F_n(y)dy \\ &= \int H\left(\frac{t-y}{h}\right) dF_n(y) \\ &= \frac{1}{n}\sum_{i=1}^{n} H\left(\frac{t-T_i}{h}\right), \end{aligned}$$

où F_n désigne la fonction de répartition empirique définie dans (2.4.1), $H(u) = \int_{-\infty}^{u} K(v)dv$ avec K est une densité symétrique, appelée fonction noyau, et h est un réel positif appelé le paramètre ou la fenêtre de lissage. Le choix du noyau n'affecte pas beaucoup la performance de l'estimateur. Cependant, le choix de la fenêtre

est crucial. Différentes méthodes ont été proposées pour un choix dans la pratique du paramètre de lissage : méthodes de validation croisées, la règle de la normale, plug-in etc. Pour plus de détails sur cet estimateur, voir [46], [51] et [24]

Maintenant pour le cas de censure à droite, il suffit de remplacer la fonction empirique F_n par l'EKM $\widehat{F_n}(t)$. On obtient alors, pour un échantillon $(Y_i, \delta_i) = (T_i \wedge C_i, \delta_i)_{i=1}^n$, l'estimateur suivant :

$$\begin{aligned}
\widehat{F_n}^K(t) &= \int \frac{1}{h} K\left(\frac{t-y}{h}\right) \widehat{F_n}(y) dy \\
&= \int H\left(\frac{t-y}{h}\right) d\widehat{F_n}(y) \\
&= \frac{1}{n} \sum_{i=1}^{n} w_i H\left(\frac{t-Y_i}{h}\right)
\end{aligned}$$

Avec les $\frac{1}{n} w_i$ sont les sauts de Kaplan-Meier.
Les propriétés asymptotiques de cet estimateur sont étudiées par plusieurs auteurs : [5], [17] et [33].

2.4.3 L'estimateur de Kaplan-Meier dans le cas bivarié

Pour le cas bivarié, on considère le couple (X, Y) de variables aléatoires. La fonction de répartition de (X, Y) est définie par $F(x, y) = P(X \leq x, Y \leq y)$. Dans le cas de données complètes, on observe $(X_1, Y_1), \ldots, (X_n, Y_n)$ n réalisations i.i.d du couple (X, Y). L'estimateur empirique de $F(x, y)$ est donnée par :

$$F_n(x, y) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{\{X_i \leq x, Y_i \leq y\}}.$$

Stute (1993) s'est intéressé au cas bivarié de l'estimateur de Kaplan-Meier en considérant un certain cadre de censure. En fait, il a proposé un modèle bivarié de

la fonction de distribution où la variable Y subit une censure aléatoire à droite et la variable X est complètement observée. Dans ce cas, au lieu d'observer le vecteur (X, Y), on observe un vecteur aléatoire (X, T, δ) où $T = Y \wedge C$, C est la variable aléatoire de censure, et $\delta = 1_{\{Y<C\}}$ est l'indicateur de la censure. On suppose que C est indépendante de Y et que

$$P(Y \leq C|X, Y) = P(Y \leq C|Y).$$

La dernière condition peut être vue comme une propriété de Markov pour le vecteur (X, Y, C).

Etant donné un échantillon de taille n, $(X_1, T_1, \delta_1), ..., (X_n, T_n, \delta_n)$ d'observations i.i.d de même loi que (X, T, δ), Stute (1993) [43] a proposé un estimateur de $F(x, y)$, noté $\widehat{F}_n^{ST}$ défini par :

$$\widehat{F}_n^{ST}(x, y) = \frac{1}{n} \sum_{i=1}^{n} W_{in} \mathbb{I}_{\{X_i \leq x, T_i \leq y\}}, \tag{2.4.4}$$

où $W_{in} = \frac{\delta_i}{1-G_n(T_i^-)}$, $i = 1, \cdots, n$, et G_n est l'estimateur de Kaplan-Meier de la fonction de répartition de C, notée G. les poids $\frac{1}{n}W_{in}$ représentent les sauts de l'estimateur de Kaplan-Meier pour la fonction de répartition de Y. Sous l'hypothèse de l'indépendance entre C et Y, Stute [43] a montré que cet estimateur est presque sûrement convergent. Sous les mêmes conditions que [43], Stute [44] a établi la normalité asymptotique de l'estimateur $\widehat{F}_n^{ST}$.

Chapitre 3

Estimation de la fonction de répartition bivariée par les polynômes de Bernstein

3.1 Introduction

Dans ce chapitre, nous allons détailler la partie principale de notre contribution sur le sujet de cette thèse. Nous commençons préalablement par définir les polynômes de Bernstein. Nous citons après quelques propriétés qui vont être utiles par la suite. Pour plus de détails sur ces polynômes, nous recommandons le livre de Lorentz (1986)[30].

3.2 Approche de Bernstein

Le théorème de *Weierstrass* est un résultat très connu en analyse. Ce théorème affirme que pour toute fonction, $\phi(x)$, continue sur un intervalle $[a, b]$, il existe un

polynôme $P(x)$ qui converge, uniformément sur $[a, b]$, vers $\phi(x)$. Ce résultat est un corollaire du théorème de Bernstein qui constitue la base de la théorie de l'estimation par les polynômes de Bernstein.

Définition 3.2.1. Soit la fonction ϕ définie sur l'intervalle fermé $[0, 1]$. On appelle le polynôme de Bernstein d'ordre m associé à ϕ, la quantité suivante :

$$B_m(\phi)(x) = \sum_{k=0}^{m} \phi\left(\frac{k}{m}\right) P_{k,m}(x),$$

où les $P_{k,m}(x)$ sont les probabilités binomiales : $P_{k,m}(x) = \binom{k}{m} x^k (1-x)^{m-k}$, pour $x \in [0, 1]$ et $k = 0, \ldots, m$.

Le lemme suivant sera utile pour démontrer le théorème de Bernstein.

Lemme 3.2.1. *Soit la fonction suivante*

$$Q_{s,m}(z) = \sum_{k=0}^{m} (k - mz)^s \, P_{k,m}(z), \quad s \in \mathbb{N}$$

On a les résultats suivants :

1. $$Q_{0,m}(z) = 1, \tag{3.2.1}$$
$$Q_{1,m}(z) = 0, \tag{3.2.2}$$
$$Q_{2,m}(z) = mz(1-z). \tag{3.2.3}$$

2. *Pour une valeur fixée de m, et pour tout $m \in \mathbb{N}^*$, $Q_{2s,m}(z)$ est un polynôme d'indéterminé $z(1-z)$ et $Q_{2s+1,m}(z)$ est un polynôme en $z(1-z)$ multiplié par $(1-2z)$.*

3. *$Q_{2s,m}(z)$ est un polynôme d'ordre $\lfloor s/2 \rfloor$ en m, pour une valeur donnée de z, où $\lfloor x \rfloor$ est la partie entière de x.*

4. *Pour tout réel $\alpha > 0$ et $s \geq 0$, il existe une constante, $C > 0$, qui vérifie :*

$$\begin{aligned} \sum_{\substack{k=0 \\ |x-\frac{k}{m}|\geq\alpha}}^{m} P_{k,m}(z) &\leq \alpha^{-2s} m^{-2s} Q_{2s,m}(z) && (3.2.4) \\ &\leq C\alpha^{-2s} m^{-s}. && (3.2.5) \end{aligned}$$

Démonstration. Les propriétés de la loi binomiale peuvent être utilisées pour démontrer les trois premiers résultats.

1. Commençons par les trois premières égalités,
   - Pour (3.2.1), un calcul immédiat donne le résultat.
   - Concernant l'égalité (3.2.2), soit Y une variable aléatoire réelle de loi binomiale de paramètres m et z, un calcul simple donne $E(Y) = \sum_{k=0}^{m} kP_{k,m}(z) = mz$ et par conséquent
   $$\begin{aligned} Q_{1,m}(z) &= \sum_{k=0}^{m} kP_{k,m}(z) - mzQ_{0,m}(z) \\ &= 0. \end{aligned}$$
   - Finalement, pour (3.2.3), on remarque que $Q_{2,m}(z) = Var(Y) = mz(1-z)$.
2. Maintenant pour les résultats (2) et (3) du lemme, une dérivation nous donne
$$\begin{aligned} \frac{d}{dz}Q_{s,m}(z) &= -msQ_{s-1,m}(z) + \sum_{k=0}^{m}(k-mz)^s \frac{d}{dz}P_{k,m}(z) \\ &= -msQ_{s-1,m}(z) + \frac{1}{z(1-z)}Q_{s+1,m}(z). \end{aligned}$$

d'où la relation

$$Q_{s+1,m}(z) = z(1-z)\left[\frac{d}{dz}Q_{s,m}(z) + mzQ_{s-1,m}(z)\right].$$

Un raisonnement par récurrence sur cette formule donne alors les deux résultats.

3. Pour la dernière propriété du lemme, on utilise le fait que $Q_{2s,m}(z)$ est un polynôme de degré s en m et un polynôme en $z(1-z)$.

□

Le théorème suivant donne une preuve au théorème de *Weierstrass* en utilisant les polynômes de Bernstein. L'approche de Bernstein permet d'annoncer le théorème de *Weierstrass* comme suit.

Théorème 3.2.2. *[Le théorème de Bernstein]*
Soit ϕ une fonction continue sur le compact $[0,1]$, alors $B_m(\phi)(x) \to \phi(x)$, quand $m \to +\infty$ uniformément sur $[0,1]$.

Démonstration. Prenons $m \in \mathbb{N}^*$, on a pour tout $x \in [0,1]$:

$$\begin{aligned}
|\phi(x) - B_m(\phi)(x)| &= \left|\sum_{k=0}^{m}\phi(x)P_{k,m}(x) - \sum_{k=0}^{m}\phi\left(\frac{k}{m}\right)P_{k,m}(x)\right| \\
&= \left|\sum_{k=0}^{m}\left(\phi(x) - \phi\left(\frac{k}{m}\right)\right)P_{k,m}(x)\right| \\
&\leq \sum_{k=0}^{m}\left|\phi(x) - \phi\left(\frac{k}{m}\right)\right|P_{k,m}(x).
\end{aligned}$$

Comme ϕ est continue sur $[0,1]$ alors : $\exists M > 0, \forall x \in [0,1] : |\phi(x)| \leq M$. Maintenant de la continuité uniforme de ϕ, il vient
pour $\epsilon > 0, \exists \alpha > 0$ qui vérifie pour tout $x, \frac{k}{m} \in [0,1]$:

$$\left|x - \frac{k}{m}\right| < \alpha \Longrightarrow \left|\phi(x) - \phi\left(\frac{k}{m}\right)\right| \leq \frac{\epsilon}{2}$$

On obtient alors :

$$\begin{aligned}
|\phi(x) - B_m(\phi)(x)| &\leq \sum_{\substack{k=0 \\ |x-\frac{k}{m}|\leq\alpha}} \left|\phi(x) - \phi\left(\frac{k}{m}\right)\right| P_{k,m}(x) \\
&+ \sum_{\substack{k=0 \\ |x-\frac{k}{m}|\geq\alpha}} \left|\phi(x) - \phi\left(\frac{k}{m}\right)\right| P_{k,m}(x) \\
&\leq \frac{\epsilon}{2} Q_{0,m}(x) + 2M \sum_{\substack{k=0 \\ |x-\frac{k}{m}|\geq\alpha}} P_{k,m}(x) \\
&\leq \frac{\epsilon}{2} + \frac{2M}{m^2\alpha^2} Q_{2,m}(x) \\
&\leq \frac{\epsilon}{2} + \frac{M}{2m\alpha^2}.
\end{aligned}$$

La dernière inégalité vient du fait que : $Q_{2,m}(x) = mx(1-x) \leq \frac{1}{4}m$. Or $\frac{M}{2m\alpha^2} \to 0$ quand $m \to \infty$, alors il existe un ordre m_0 à partir duquel $\frac{M}{2m\alpha^2} \leq \frac{\epsilon}{2}$. Le résultat du théorème en découle. □

3.2.1 Estimation de la fonction de répartition univariée par les polynômes de Bernstein

On reprend le théorème (3.2.2) afin d'obtenir un estimateur d'une fonction de répartition dans la cas univarié F. Babu et al. (2002) [1] ont proposé l'estimateur suivant :

$$\widehat{F}^u_{m,n}(x) = \sum_{k=0}^{m} F_n\left(\frac{k}{m}\right) P_{k,m}(x), \quad x \in [0,1]. \tag{3.2.6}$$

L'idée de cet estimateur est de remplacer, dans l'approximation définie dans (3.2.1), les $F\left(\frac{k}{m}\right)$ par $F_n\left(\frac{k}{m}\right)$. L'estimateur de Babu et al. (2002) est une forme de lissage de la distribution empirique. Il est facile de voir que $\widehat{F}^u_{m,n}$ est une fonction de répartition.

Le théorème suivant montre la convergence presque sûre de cet estimateur vers la fonction de répartition. Pour toute la suite, on utilise cette notation :

$$||\zeta|| = \sup_{u\in A} |\zeta(u)|$$

où A est le support de la fonction ζ, quand $\sup_{u\in A} |\zeta(u)|$ existe.

Théorème 3.2.3. *Soit F une fonction de répartition continue sur $[0,1]$. Si $m,n \to \infty$, alors on a $\left|\left|\widehat{F}^u_{m,n} - F\right|\right| \to 0 \quad p.s.$*

Démonstration. La preuve de ce théorème se base sur cette inégalité :

$$\left|\left|\widehat{F}^u_{m,n} - F\right|\right| \leq \left|\left|\widehat{F}^u_{m,n} - B_m(F)\right|\right| + ||B_m(F) - F||\,.$$

On a d'une part :

$$\begin{aligned}\left|\widehat{F}^u_{m,n}(x) - B_m(F)(x)\right| &= \left|\sum_{k=0}^{m}\left[F_n\left(\frac{k}{m}\right) - F\left(\frac{k}{m}\right)\right]P_{k,m}(x)\right| \\ &\leq \max_k \left|F_n\left(\frac{k}{m}\right) - F\left(\frac{k}{m}\right)\right| \\ &\leq ||F_n - F||\,.\end{aligned}$$

D'après le théorème de Glivenko-Cantelli, on a la convergence presque sûre de la première quantité. D'autre part le théorème (3.2.2) assure la convergence de $||B_m(F) - F||$ vers 0. Le résultat du théorème est alors établi. □

Définition 3.2.2. Soit ϕ une fonction définie sur un intervalle I. Soit un réel k positif. On dit que ϕ est k-lipschitzienne sur I s'il existe une constante k qui vérifie :

$$\forall x, y \in I, \quad |f(x) - f(y)| \leq k\,|x - y|\,.$$

Un autre résultat de l'article, [1], mesure la distance entre l'estimateur $\widehat{F}^u_{m,n}(x)$ et l'estimateur empirique. En fait, sous l'hypothèse que la densité, $f := F'$, existe et lipschitzienne, on a le théorème suivant :

Théorème 3.2.4. *Soit F une fonction continue et dérivable sur $[0,1]$, de densité de probabilité f. Si f est lipschitzienne d'ordre 1 et pour $n^{2/3} \leq m \leq \left(\frac{n}{\log n}\right)^2$, on a*

$$\left|\left|\widehat{F}^u_{m,n} - F_n\right|\right| = O_{p.s.}\left(\left(\frac{\log n}{n}\right)^{1/2}\left(\frac{\log m}{m}\right)^{1/4}\right) \quad quand \quad n \to \infty.$$

Si de plus on prend : $m = n$ on obtient :

$$\left|\left|\widehat{F}^u_{m,n} - F_n\right|\right| = O_{p.s.}\left(\left(\frac{\log n}{n}\right)^{3/4}\right).$$

Pour démontrer ce théorème on aura besoin des deux lemmes suivants.

Lemme 3.2.5. *Soient $Z_1, \ldots, Z_n$ des variables aléatoires réelles indépendantes, centrées et bornées par une constante positive b. Si $V \geq \sum_{i=1}^n E(Z_i^2)$, alors $\forall 0 < s < 1$ et $0 \leq a \leq V/sb$, on a*

$$P\left(\sum_{i=1}^n Z_i > a\right) \leq \exp\left(-a^2 s(1-s)/V\right). \tag{3.2.7}$$

Lemme 3.2.6. *Si F est lipschitzienne et sur l'ensemble*

$$N_{x,m} = \left\{0 \leq k \leq m : |k - mx| \leq (m \log m)^{1/2}\right\}$$

on a, pour $2 \leq m \leq (n \log n)^2$, le résultat suivant

$$H_{n,m} = \sup_{0<x<1} \max_{k \in N_{x,m}} |F_n(k/m) - F(k/m) - F_n(x) + F(x)| = O(\beta_{n,m}) \quad p.s.$$

où $\beta_{n,m} = (n^{-1} \log n)^{1/2}(m^{-1} \log m)^{1/4}$.

Pour les preuves des deux lemmes précédents, nous dirigeons le lecteur vers l'article [1].

Le théorème (3.2.4) est démontré dans le paragraphe suivant.

Démonstration. Commençons par réecrire la différence

$$\widehat{F}^{u}_{m,n}(x)-F_n(x)=\sum_{k=m}^{m} P_{k,m}(x)\left(F_n(k/m)-F(k/m)-(F_n(x)+F(x))+(B_m(F)(x)-F(x)\right).$$

Comme f est lipschitzienne d'ordre 1 alors on a :

$$F(k/m)-F(x)=((k/m)-x)\,f(x)+O\left(((k/m)-x)^2\right).$$

En appliquant les résultats(3.2.2) et (3.2.3) du lemme (3.2.1) on obtient

$$||B_m(F)-F||=O\left(\frac{1}{m}\right). \tag{3.2.8}$$

D'autre part et en vertu le lemme (3.2.5) en prenant les variable aléatoire Z_i supposées $i.i.d.$ et qui vérifient $P(Z_i=1-x)=x=1-P(Z_i=-x)$, $a=(m\log m)^{1/2}$, $V=m/4$ et $s=1/2$, on obtient alors pour $m\geq 16$,

$$\sum_{k=0,k\notin N_{x,m}} P_{k,m}(x)\leq \frac{2}{m}. \tag{3.2.9}$$

Finalement on déduit le résultat du théorème d'après (3.2.8), (3.2.9) et le lemme (3.2.6).

□

Bien qu'on dispose de la convergence presque sûre de $\widehat{F}^{u}_{m,n}$, cet estimateur n'est pas un estimateur sans biais. On vérifie facilement le fait que $E\left(\widehat{F}^{u}_{m,n}(x)\right)=B_m(F)(x)$, et puisqu'on a $B_m(F)(x)\to F(x)$ quand $m\to\infty$ on affirme que cet estimateur est asymptotiquement sans biais. La loi du logarithme itéré pour cet estimateur a été établie par Leblanc(2009) [25], Les expressions du biais de cet estimateur et de sa variance sont développées aussi dans [25].

3.2.2 Généralisation au cas bivarié

Dans cette partie, nous allons aborder la généralisation au cas bivarié de quelques résultats qui seront fondamentaux pour notre contribution.

Pour toute fonction continue sur $[0,1]^2$, notée Ψ, on définit l'opérateur suivant :

$$B_m(\Psi)(x,y) = \sum_{k=0}^{m}\sum_{\ell=0}^{m} \Psi\left(\frac{k}{m}, \frac{\ell}{m}\right) P_{k,m}(x) P_{\ell,m}(y), \tag{3.2.10}$$

Où m est un entier qui joue le rôle du paramètre de lissage. Une version multivariée du théorème (3.2.2) est définie dans la littérature. Le théorème suivant donne la généralisation du théorème (3.2.2) au cas bivarié.

Théorème 3.2.7. *Soit Ψ une fonction continue sur $[0,1]^2$, alors on a :*

$$B_m(\Psi)(x,y) \to \Psi(x,y),$$

quand $m \to \infty$

Démonstration. Soient $(x,y) \in [0,1]^2$ et m le paramètre du lissage. On a :

$$\begin{aligned}
|B_m(\Psi)(x,y) - \Psi(x,y)| &= \left| \sum_{k=0}^{m}\sum_{\ell=0}^{m} \Psi\left(\frac{k}{m}, \frac{\ell}{m}\right) P_{k,m}(x) P_{\ell,m}(y) - \Psi(x,y) \right| \\
&= \left| \sum_{k=0}^{m}\sum_{\ell=0}^{m} \Psi\left(\frac{k}{m}, \frac{\ell}{m}\right) P_{k,m}(x) P_{\ell,m}(y) - \sum_{k=0}^{m}\sum_{\ell=0}^{m} \Psi(x,y) P_{k,m}(x) P_{\ell,m}(y) \right| \\
&= \left| \sum_{k=0}^{m}\sum_{\ell=0}^{m} \left[\Psi\left(\frac{k}{m}, \frac{\ell}{m}\right) - \Psi(x,y) \right] P_{k,m}(x) P_{\ell,m}(y) \right| \\
&\leq \sum_{k=0}^{m}\sum_{\ell=0}^{m} \left| \Psi\left(\frac{k}{m}, \frac{\ell}{m}\right) - \Psi(x,y) \right| P_{k,m}(x) P_{\ell,m}(y).
\end{aligned}$$

La fonction Ψ étant continue sur le compact $[0,1]^2$, elle est uniformément continue et bornée par une certaine constante $A > 0$ pour tout $(x,y) \in [0,1]^2$. Maintenant, pour la

dernière partie, nous appliquons des même techniques utilisées dans la démonstration du théorème (3.2.2). Finalement la convergence est vérifiée.

□

Motivés par cette approximation, Babu et Chaubey (2006) [2] ont proposé l'estimateur suivant :

$$\widehat{\Psi}_{m,n}(x,y) = \sum_{k=0}^{m}\sum_{\ell=0}^{m} F_n\left(\frac{k}{m},\frac{\ell}{m}\right) P_{k,m}(x)P_{\ell,m}(y), \quad (x,y)\in[0,1]^2.$$

On vérifie facilement que $\widehat{\Psi}_{m,n}$ est une fonction de répartition, ce qui justifie son emploi comme un estimateur naturel de la fonction de répartition. La convergence presque sûre de cet estimateur est donnée par le théorème suivant :

Théorème 3.2.8. *Soit F une fonction continue de répartition sur $[0,1]^2$. Alors,*

$$\left|\left|\widehat{\Psi}_{m,n} - F\right|\right| \to 0 \quad p.s.$$

quand $m,n\to\infty$.

Démonstration. Moyennant les mêmes techniques utilisées pour démontrer le théorème (3.2.3), on a :

$$\left|\left|\widehat{\Psi}_{m,n} - F\right|\right| \leq ||F_n - F|| + ||B_m(\Psi) - F||.$$

La deuxième partie est déduite directement du théorème (3.2.7). Pour la première partie, un résultat dans [7] donne sa convergence presque sûre vers 0. □

Dans la section suivante, nous allons énoncer les principaux résultats de notre contribution.

3.3 Estimation en présence de données censurées

Au Chapitre 2, nous avons exposé l'estimateur de la fonction de répartition bivariée proposé par Stute (1993) dans le cas de données censurées à droite. Malheureusement, cet estimateur est en escalier et sa dérivée est nulle presque partout. Nous allons considérer, pour le même modèle de données, un lissage par les polynômes de Bernstein de l'estimateur de Stute afin de traiter son problème de discontinuité. Nous commençons par présenter notre estimateur puis aborder les résultats obtenus.

Soit le vecteur aléatoire (X, Y) de fonction de répartition F défini sur $[0, 1]^2$. Il est facile de généraliser cette considération pour des données définies sur d'autres domaines dans $\mathbb{R}^2$ en utilisant les transformations adéquates. Nous allons supposer que X soit complètement observée et que Y soit le sujet d'une censure aléatoire à droite. Dans ce cas, au lieu d'observer le vecteur (X, Y), on observe un vecteur aléatoire (X, T, δ) où $T = Y \wedge C$, C est la variable aléatoire de censure de fonction de répartition G, et $\delta = 1_{\{Y<C\}}$ est l'indicateur de la censure. On suppose que C est indépendante de Y. Étant donné un échantillon de taille n, $(X_1, T_1, \delta_1), ..., (X_n, T_n, \delta_n)$ d'observations i.i.d. de même loi que (X, T, δ), le but est de proposer un estimateur par lissage par les polynômes de Bernstein, de l'estimateur de Stute (1993), $\widehat{F}_n^{ST}$, défini dans (2.4.4) et d'étudier ses propriétés asymptotiques. A cet effet, nous allons noter par L la fonctions de répartition de la variable T, $\tau = \sup_t\{t, L(t) < 1\}$, i.e, le temps terminal de T et $P_{k,m}^{\tau}(y) = P_{k,m}(y/\tau)$. L'estimateur proposé, $\widehat{F}_{m,n}$ de F est donné par :

$$\widehat{F}_{m,n}(x, y) = \sum_{k=0}^{m}\sum_{\ell=0}^{m} \widehat{F}_n^{ST}\left(\frac{k}{m}, \frac{\tau\ell}{m}\right) P_{k,m}(x)P_{\ell,m}^{\tau}(y). \tag{3.3.1}$$

Dans la pratique τ est généralement inconnu, nous le remplaçons par $T_{(n)}$, c-à-d la valeur maximale des observations $\{T_i\}_{i=1}^n$. Notons aussi que pour tout estimateur

convergent $\hat{\tau}$ tel que $\hat{\tau} - \tau = O_{p.s.}(n^{-1})$, tous les résultats ci-dessous sont vérifiés.

Remarque 3.3.1. *Il existe d'autres formes de censure à droite ; par exemple si les deux variables subissent une censure à droite. L'estimateur proposé par* (3.3.1) *est adaptable en remplaçant* $\widehat{F}_n^{ST}$ *par la fonction empirique qui prend en considération la forme de censure. Dans le but de construire des estimateurs pour la copule, plusieurs exemples de fonctions empiriques ont été discutés dans [19] et [31]*

Remarque 3.3.2. *Le fait que* $\widehat{F}_{m,n}$ *soit dérivable, un estimateur de la densité bivariée du couple* (X, Y) *en est déduit. Il est donné comme suit :*

$$\begin{aligned}\widehat{f}_{m,n}(x,y) &= \sum_{k=0}^{m}\sum_{\ell=0}^{m} \widehat{F}_n^{ST}\left(\frac{k}{m}, \frac{\tau\ell}{m}\right) P'_{k,m}(x)\left(P^{\tau}\right)'_{\ell,m}(y) \\ &= \frac{m^2}{\tau}\sum_{k=0}^{m-1}\sum_{\ell=0}^{m-1} \widehat{F}_n^{ST}\left(A_{k,l}\right) P_{k,m-1}(x) P^{\tau}_{\ell,m-1}(y)\end{aligned}$$

où $\widehat{F}_n^{ST}\left(A_{k,\ell}\right) = \widehat{F}_n^{ST}\left(\frac{k}{m}, \frac{\tau\ell}{m}\right) - \widehat{F}_n^{ST}\left(\frac{k-1}{m}, \frac{\tau\ell}{m}\right) - \widehat{F}_n^{ST}\left(\frac{k}{m}, \frac{\tau(\ell-1)}{m}\right) + \widehat{F}_n^{ST}\left(\frac{k-1}{m}, \frac{\tau(\ell-1)}{m}\right)$.

Remarque 3.3.3. *Un estimateur non paramétrique de la copule, la fonction qui modélise la structure de dépendance entre* Y *et* Y*, est développable à partir de notre estimateur pour la fonction de répartition. Pour* $(u, v) \in [0, 1]^2$*, un estimateur de la copule est donné comme suit :*

$$\begin{aligned}\widehat{C}_{m,n}(u,v) &= \widehat{F}_{m,n}\left(\widehat{F}^{-1}_{m,n,1}(u), \widehat{F}^{-1}_{m,n,2}(v)\right) \\ &= \sum_{k=0}^{m}\sum_{\ell=0}^{m} \widehat{F}_n^{ST}\left(\frac{k}{m}, \frac{\tau\ell}{m}\right) P_{k,m}\left(\widehat{F}^{-1}_{m,n,1}(u)\right) P^{\tau}_{\ell,m}\left(\widehat{F}^{-1}_{m,n,2}(v)\right),\end{aligned}$$

où $\widehat{F}_{m,n,1}(x) = \widehat{F}_{m,n}(x, +\infty)$ *(resp.* $\widehat{F}_{m,n,2}(y) = \widehat{F}_{m,n}(+\infty, y)$ *) est un estimateur pour la fonction de répartition de* X *(resp. de* Y*) et* K^{-1} *désigne l'inverse généralisé de la fonction* K*.*

Remarque 3.3.4. *Pour le cas d'un vecteur aléatoire (X, Y) à valeurs positives, on propose un autre estimateur qui ne nécessite pas des transformations des données. Cet estimateur se construit à l'aide des probabilités de Poisson et est donné par :*

$$\widehat{F}(x,y) = \sum_{k=0}^{\infty}\sum_{\ell=0}^{\infty} \widehat{F}_n^{ST}\left(\frac{k}{m}, \frac{\ell}{m}\right) \mathrm{Pois}_{k,m}(x)\,\mathrm{Pois}_{\ell,m}(y)$$

où $\mathrm{Pois}_{k,m}(z)k \in \mathbb{N}$ *sont les probabilités de Poisson de paramètre* mz *:*

$$\mathrm{Pois}_{k,m}(z) = \exp(-mz)\frac{(mz)^k}{k!} \qquad k \in \mathbb{N}.$$

Aussi, on définit un estimateur associé pour la densité :

$$\widehat{f}(x,y) = m^2 \sum_{k=0}^{\infty}\sum_{\ell=0}^{\infty} \widehat{F}_n\left(A_{k,\ell}\right) \mathrm{Pois}_{k,m}(x)\,\mathrm{Pois}_{\ell,m}(y).$$

3.4 Propriétés asymptotiques

Dans cette section, nous allons aborder des propriétés asymptotiques de l'estimateur proposé $\widehat{F}_{m,n}$. Nous commençons par la convergence uniforme presque sûre, en donnant la vitesse de convergence de cet estimateur. Ensuite, la représentation i.i.d. est établie pour ce nouvel estimateur. Ce résultat va nous aider pour calculer le biais et la variance asymptotique. Finalement, la normalité asymptotique est déduite à partir de la représentation i.i.d. pour notre estimateur.

3.4.1 La convergence presque sûre

La proposition suivante montre la convergence uniforme presque sûre de $\widehat{F}_{m,n}$ vers F sur le domaine $[0,1]^2$ et fournit la vitesse de convergence de notre estimateur. Cette vitesse est égale à la vitesse de convergence de l'estimateur de Stute (1993).

Théorème 3.4.1. *Supposons que F soit une fonction lipschitienne. Sous l'hypothèse $m^{-1/2} = O(n^{-1/2}(\log\log n)^{1/2})$, on a :*

$$\sup_{0\leq x\leq 1, 0\leq y\leq \tau} \left|\widehat{F}_{m,n}(x,y) - F(x,y)\right| = O_{a.s.}(n^{-1/2}(\log\log n)^{1/2}).$$

Preuve du théorème 3.4.1.

$$\begin{aligned}
\left|\widehat{F}_{m,n}(x,y) - F(x,y)\right| &= \left|\sum_{k=0}^{m}\sum_{\ell=0}^{m} \widehat{F}_n^{ST}\left(\frac{k}{m},\frac{\tau\ell}{m}\right) P_{k,m}(x)P^{\tau}_{\ell,m}(y) - F(x,y)\right| \\
&= \left|\sum_{k=0}^{m}\sum_{\ell=0}^{m} \left[\widehat{F}_n^{ST}\left(\frac{k}{m},\frac{\tau\ell}{m}\right) - F(x,y)\right] P_{k,m}(x)P^{\tau}_{\ell,m}(y)\right| \\
&\leq \sum_{k=0}^{m}\sum_{\ell=0}^{m} \left|\widehat{F}_n^{ST}\left(\frac{k}{m},\frac{\tau\ell}{m}\right) - F\left(\frac{k}{m},\frac{\tau\ell}{m}\right)\right| P_{k,m}(x)P^{\tau}_{\ell,m}(y) \\
&+ \sum_{k=0}^{m}\sum_{\ell=0}^{m} \left|F\left(\frac{k}{m},\frac{\tau\ell}{m}\right) - F(x,y)\right| P^{\tau}_{k,m}(x)P^{\tau}_{\ell,m}(y) \\
&= (I) + (II).
\end{aligned}$$

La partie (I) est traitée dans [43] [voir le corollaire (1.5)] et montre que :

$$\sup_{0\leq x\leq 1, 0\leq y\leq \tau} \left|\widehat{F}_n^{ST}(x,y) - F(x,y)\right| = O_{p.s.}\left(n^{-1/2}(\log\log n)^{1/2}\right).$$

Alors on obtient :

$$(I) = O_{p.s.}\left(n^{-1/2}(\log\log n)^{1/2}\right). \tag{3.4.1}$$

Maintenant pour la partie (II), nous appliquons des techniques similaires à celles utilisées dans [20]. Pour tout $x \in [0,1]$ et $y \in [0,\tau]$, notons par $\mathcal{B}_z$ la variable aléatoire

suivant la loi binomiale de paramètre (z, m), nous avons :

$$\begin{aligned} II &\leq C \sup_{0\leq x\leq 1, 0\leq y\leq \tau} \sum_{k=0}^{m}\sum_{\ell=0}^{m} \left[\left|\frac{k}{m}-x\right| + \left|\frac{\tau\ell}{m}-y\right|\right] P_{k,m}(x) P^{\tau}_{\ell,m}(y) \\ &\leq C\left[\sup_{0\leq x\leq 1}\left(E\left(\frac{\mathcal{B}_x}{m}-x\right)^2\right)^{1/2} + \sup_{0\leq y\leq \tau}\left(E\left(\frac{\tau\mathcal{B}_{y/\tau}}{m}-y\right)^2\right)^{1/2}\right] \\ &= C\left[\sup_{0\leq x\leq 1}\left(\frac{x(1-x)}{m}\right)^{1/2} + \sup_{0\leq y=\tau}\left(\frac{y(\tau-y)}{m}\right)^{1/2}\right] \\ &= O(m^{-1/2}) \end{aligned} \tag{3.4.2}$$

où C vérifie l'inégalité suivante $|F(x_1, y_1) - F(x_2, y_2)| \leq C\left(|x_1 - x_2| + |y_1 - y_2|\right)$. Finalement le résultat de la proposition découle d'après (3.4.1), (3.4.2) et sous l'hypothèse technique $m^{-1/2} = O(n^{-1/2}(\log\log n)^{1/2})$.

3.4.2 La représentation i.i.d.

Dans cette section, nous allons représenter notre estimateur sous une forme linéaire des quantités aléatoires, indépendantes et de même distribution. La proposition suivante établit la représentation asymptotique i.i.d. de l'estimateur de Bernstein $\widehat{F}_{m,n}$ avec une erreur d'ordre $O_{p.s.}\left(\frac{\log n}{n}\right)$.

Théorème 3.4.2. *(Représentation i.i.d.)*
Sous l'hypothèse que F et G soient continues, la représentation i.i.d. de l'estimateur $\widehat{F}_{m,n}$ est donnée par :

$$\widehat{F}_{m,n}(x, y) = \frac{1}{n}\sum_{i=1}^{n}\frac{\eta_i(x, y)}{1 - G(T_i)} + O_{a.s}\left(\frac{\log n}{n}\right) \tag{3.4.3}$$

où

$$\eta_i(x,y)=\sum_{k=0}^{m}\sum_{\ell=0}^{m}\mathbb{I}_{\left\{X_i\leq\frac{k}{m},T_i\leq\frac{\tau\ell}{m}\right\}}P_{k,m}(x)P^{\tau}_{\ell,m}(y).$$

Preuve du théorème 3.4.2. Nous commençons par remarquer le fait que

$$\widehat{F}_n^{ST}(u,v)=\frac{1}{n}\sum_{i=1}^{n}W_{in}\mathbb{I}_{\left\{X_i\leq\frac{k}{m},T_i\leq\frac{\tau\ell}{m}\right\}},\quad\text{avec}\quad W_{in}=n\left(F_n(+\infty,T_i)-F_n(+\infty,T_i^-)\right). \tag{3.4.4}$$

D'après [32], il est possible de montrer que W_{in} s'écrit sous la forme suivante : $W_{in}=W_i+R_{i,n}$, où $W_i=\dfrac{\delta_i}{1-G(T_i)}$ et

$$R_{i,n}=O_{p.s}(n^{-1}\log n)\quad\text{and}\quad\mathbb{E}(R_{i,n})=O(n^{-1}). \tag{3.4.5}$$

Alors, on a :

$$\begin{aligned}\widehat{F}_{m,n}(x,y)&=\frac{1}{n}\sum_{i=1}^{n}W_{in}\sum_{k=0}^{m}\sum_{\ell=0}^{m}\mathbb{I}_{\left\{X_i\leq\frac{k}{m},T_i\leq\frac{\tau\ell}{m}\right\}}P^{\tau}_{k,m}(x)P_{\ell,m}(y)\\&=\frac{1}{n}\sum_{i=1}^{n}\left[W_i\left(\sum_{k=0}^{m}\sum_{\ell=0}^{m}\mathbb{I}_{\left\{X_i\leq\frac{k}{m},T_i\leq\frac{\tau\ell}{m}\right\}}P^{\tau}_{k,m}(x)P_{\ell,m}(y)\right)\right]+O_{p.s}(n^{-1}\log n).\end{aligned}$$

et le résultat découle à partir de (3.4.5).

Hypothèse 1. *Les dérivées partielles de F, du deuxième ordre sont continues sur $[0,1]^2$. Elles sont notées par : $F_x=\frac{\partial F(x,y)}{\partial x}$, $F_y=\frac{\partial F(x,y)}{\partial y}$, $F_{xx}=\frac{\partial^2 F(x,y)}{\partial x\partial x}$, $F_{yy}=\frac{\partial^2 f(x,y)}{\partial y\partial y}$ et $F_{xy}=\frac{\partial^2 F(x,y)}{\partial x\partial y}$.*

Théorème 3.4.3. *Sous l'hypothèse (1) et les même condition du théorème (3.4.2), si $m\to+\infty$ et $nm^{1/2}\to+\infty$, pour n suffisamment grand, alors pour tout $0<x<1$ et $0<y<\tau$, nous avons*

(1)

$$\mathbb{E}\left[\widehat{F}_{m,n}(x,y)\right]=F(x,y)+\frac{1}{2m}F_{xx}(x,y)x(1-x)+\frac{1}{2m}F_{yy}(x,y)y(\tau-y)+o(m^{-1})+O(1/n). \tag{3.4.6}$$

(2)

$$\mathbb{V}ar\left[\widehat{F}_{m,n}(x,y)\right] = n^{-1}\sigma^2(x,y) - n^{-1}m^{-1/2}V(x,y) + o(n^{-1}m^{-1/2}), \quad (3.4.7)$$

où

$$\sigma^2(x,y) = H^2(x,y) - F(x,y) \text{ et } V(x,y) = H_x(x,y)\left[2x(1-x)/\pi\right]^{1/2} + H_y(x,y)\left[2y(\tau-y)/\pi\right]^{1/2}$$

avec

$$H(x,y) = \int_0^x \int_0^y \frac{dF(t,s)}{1-G(t)}$$

une fonction dérivable sur $[0,1]^2$.

La proposition indique que notre estimateur est assymptotiquement sans biais quand m tend vers l'infini. Aussi, la première partie de (3.4.7), c-à-d $n^{-1}\sigma^2(x,y)$, est la même variance asymptotique de l'estimateur de Stute (1993) $\widehat{F}_n^{ST}$. Alors, notre estimateur réduit asymptotiquement la variance de $\widehat{F}_n^{ST}$ puisque le deuxième terme dans la variance asymptotique est négatif. Finalement, dans le cas des données dont on n'observe aucne censure, c-à-d $\tau = 1$ et $G = 0$, le bias asymptotique et la variance dans l'expression (3.4.6) et (3.4.7) se réduisent aux expressions établies pour les données complètement observées pour l'estimateur empirique bivarié.

Preuve du théorème 3.4.3. Pour la partie (1) de la proposition, On déduit rapidement le fait que :

$$\begin{aligned}
\mathbb{E}\left[\eta_i(x,y)\right] &= \sum_{k=0}^{m}\sum_{\ell=0}^{m} \mathbb{E}\left(W_i \mathbb{I}_{\left\{X_i \le \frac{k}{m}, T_i \le \frac{\tau\ell}{m}\right\}}\right) P_{k,m}(x)\, P^{\tau}_{\ell,m}(y) \\
&= \sum_{k=0}^{m}\sum_{\ell=0}^{m} F\left(\frac{k}{m}, \frac{\tau\ell}{m}\right) P_{k,m}(x) P^{\tau}_{\ell,m}(y) \\
&\equiv F^{\tau}_m(x,y).
\end{aligned}$$

Maintenant, si F est une fonction deux fois dérivable sur $[0,1]^2$, un développement de Taylor bivarié au voisinage de (x,y) nous donne :

$$\begin{aligned} F(k/m,\tau\ell/m) = F(x,y) &+ F_x(x,y)(k/m-x) + F_y(x,y)(\tau\ell/m-y) \\ &+ \frac{1}{2}F_{xx}(x,y)(k/m-x)^2 + \frac{1}{2}F_{yy}(x,y)(\tau\ell/m-y)^2 \\ &+ F_{xy}(x,y)(k/m-x)(\tau\ell/m-y) + o\left((k/m-x)^2+(\tau\ell/m-y)^2\right) \end{aligned}$$

Nous réécrivons $F_m^\tau(x,y)$ comme suit :

$$\begin{aligned} F_m^\tau(x,y) &= F(x,y) + \frac{1}{2}F_{xx}(x,y)\sum_{k=0}^{m}(k/m-x)^2P_{k,m}(x) + \frac{1}{2}F_{yy}(x,y)\sum_{\ell=0}^{m}(\tau\ell/m-y)^2P_{\ell,m}^\tau(y) \\ &+ o\left((k/m-x)^2+(\tau\ell/m-y)^2\right) \\ &= F(x,y) + \frac{1}{2}m^{-1}F_{xx}(x,y)x(1-x) + \frac{1}{2}m^{-1}F_{yy}(x,y)y(\tau-y) + o(m^{-1}). \qquad (3.4.8) \end{aligned}$$

Donc, à partir de (3.4.5)[$\mathbb{E}(R_{i,n}) = O(n^{-1})$] et (3.4.8), nous obtenons (3.4.6).

Pour la deuxième partie de la proposition, c-à-d. la variance asymptotique, nous avons

$$\begin{aligned} \mathbb{E}\left[\eta_i(x,y)^2\right] &= \sum_{k=0}^{m}\sum_{\ell=0}^{m}\sum_{k'=0}^{m}\sum_{\ell'=0}^{m}\mathbb{E}\left(W_i^2\mathbb{I}_{\left\{X_i\le\frac{k}{m},T_i\le\frac{\tau\ell}{m}\right\}}\mathbb{I}_{\left\{X_i\le\frac{k'}{m},T_i\le\frac{\tau\ell'}{m}\right\}}\right) \\ &\times P_{k,m}(x)\,P_{k',m}(x)\,P_{\ell,m}\left(\frac{y}{\tau}\right)P_{\ell',m}\left(\frac{y}{\tau}\right). \end{aligned}$$

Nous commençons par calculer $\mathbb{E}\left[W_i^2\mathbb{I}_{\left\{X_i\leq\frac{k}{m},T_i\leq\frac{\tau\ell}{m}\right\}}\mathbb{I}_{\left\{X_i\leq\frac{k'}{m},T_i\leq\frac{\tau\ell'}{m}\right\}}\right]$. En fait,

$$\begin{aligned}
\mathbb{E}\left[W_i^2\mathbb{I}_{\left\{X_i\leq\frac{k}{m},T_i\leq\frac{\tau\ell}{m}\right\}}\mathbb{I}_{\left\{X_i\leq\frac{k'}{m},T_i\leq\frac{\tau\ell'}{m}\right\}}\right] &= \mathbb{E}\left[\frac{\delta_i}{(1-G(T_i))^2}\mathbb{I}_{\left\{X_i\leq\frac{k}{m},T_i\leq\frac{\tau\ell}{m}\right\}}\mathbb{I}_{\left\{X_i\leq\frac{k'}{m},T_i\leq\frac{\tau\ell'}{m}\right\}}\right] \\
&= \mathbb{E}\left[\frac{\delta_i}{(1-G(Y_i))^2}\mathbb{I}_{\left\{X_i\leq\frac{k\wedge k'}{m},T_i\leq\tau\frac{\ell\wedge\ell'}{m}\right\}}\right] \\
&= \mathbb{E}\left[\frac{\mathbb{I}_{\{Y_i<C_i\}}}{(1-G(Y_i))^2}\mathbb{I}_{\left\{X_i\leq\frac{k\wedge k'}{m},Y_i\leq\tau\frac{\ell\wedge\ell'}{m}\right\}}\right] \\
&= \mathbb{E}\left[\frac{1}{(1-G(Y_i))^2}\mathbb{I}_{\left\{X_i\leq\frac{k\wedge k'}{m},Y_i\leq\tau\frac{\ell\wedge\ell'}{m}\right\}}\mathbb{E}\left(\mathbb{I}_{\{Y_i<C_i\}}|X_i,Y_i\right)\right] \\
&= \mathbb{E}\left[\frac{\mathbb{I}_{\left\{X_i\leq\frac{k\wedge k'}{m},Y_i\leq\tau\frac{\ell\wedge\ell'}{m}\right\}}}{1-G(T_i)}\right] \\
&= \int_0^{\frac{k\wedge k'}{m}}\int_0^{\tau\frac{\ell\wedge\ell'}{m}}\frac{dF(t,s)}{1-G(s)} \\
&\equiv H\left(\frac{k\wedge k'}{m},\tau\frac{\ell\wedge\ell'}{m}\right).
\end{aligned}$$

Alors,

$$\begin{aligned}
\mathbb{E}\left[\eta_i(x,y)^2\right] &= \sum_{k=0}^{m}\sum_{\ell=0}^{m}\sum_{k'=0}^{m}\sum_{\ell'=0}^{m}H\left(\frac{k\wedge k'}{m},\tau\frac{\ell\wedge\ell'}{m}\right)P_{k,m}(x)P_{k',m}(x)\,P_{\ell,m}\left(\frac{y}{\tau}\right)P_{\ell',m}\left(\frac{y}{\tau}\right) \\
&= \sum_{k=0}^{m}\sum_{\ell=0}^{m}H\left(\frac{k}{m},\frac{\tau\ell}{m}\right)P_{k,m}^2(x)P_{\ell,m}^2\left(\frac{y}{\tau}\right) \\
&+ \sum_{k=0}^{m}\sum_{\ell=0}^{m}\sum_{k'=0,k'\neq k}^{m}\sum_{\ell'=0,\ell'\neq\ell}^{m}H\left(\frac{k\wedge k'}{m},\tau\frac{\ell\wedge\ell'}{m}\right)P_{k,m}(x)P_{k',m}(x)\,P_{\ell,m}\left(\frac{y}{\tau}\right)P_{\ell',m}\left(\frac{y}{\tau}\right) \\
&+ \sum_{k=0}^{m}\sum_{\ell=0}^{m}\sum_{\ell'=0,\ell'\neq\ell}^{m}H\left(\frac{k}{m},\tau\frac{\ell\wedge\ell'}{m}\right)P_{k,m}^2(x)P_{\tau\ell,m}\left(\frac{y}{\tau}\right)P_{\tau\ell',m}\left(\frac{y}{\tau}\right) \\
&+ \sum_{k=0}^{m}\sum_{\ell=0}^{m}\sum_{k'=0,k'\neq k}^{m}H\left(\frac{k\wedge k'}{m},\frac{\tau\ell}{m}\right)P_{k,m}(x)P_{k',m}(x)\,P_{\ell,m}^2\left(\frac{y}{\tau}\right) \\
&\equiv \theta_1+\theta_2+\theta_3+\theta_4.
\end{aligned}$$

En appliquant un développement de Taylor de $H\left(\frac{k}{m},\frac{\tau\ell}{m}\right)$ au voisinage (x,y), nous

obtenons, pour θ_1

$$\theta_1 = H(x,y)S_m(x)S_m(y/\tau) + O\left(S_m(x)I_m(y/\tau) + S_m(y/\tau)I_m(x)\right),$$

où $S_m(z) = \sum_{k=0}^{m} P_{k,m}^2(z)$ and $I_m(z) = \sum_{k=0}^{m} \left|\frac{k}{m} - z\right| P_{k,m}^2(z)$. De même un autre développement de Taylor de H au voisinage de (x,y), nous donne l'écriture suivante pour θ_2,

$$\begin{aligned}\theta_2 = {} & H(x,y)(1-S_m(x))(1-S_m(y/\tau)) \\ & + 2H_x(x,y)R_{1m}(x)[1-S_m(y/\tau)] + 2\tau H_y(x,y)[1-S_m(x)]R_{1m}(y/\tau) \\ & + O\left(R_{2m}(x)[1-S_m(y/\tau)] + [1-S_m(x)]R_{2m}(y/\tau)\right).\end{aligned}$$

Pour θ_3,

$$\begin{aligned}\theta_3 = {} & H(x,y)S_m(x)[1-S_m(y/\tau)] \\ & + H_x(x,y)[1-S_m(y/\tau)]O(I_m(x)) + 2H_y(x,y)S_m(x)R_{1m}(y/\tau) \\ & + O\left(I_m(x)[1-S_m(y/\tau)] + 2S_m(x)R_{2m}(y/\tau)\right).\end{aligned}$$

Finalement pour θ_4

$$\begin{aligned}\theta_4 = {} & H(x,y)(1-S_m(x/\tau))S_m(y) \\ & + 2H_x(x,y)R_{1m}(x)S_m(y/\tau) + H_y(x,y)[1-S_m(x)]O(I_m(y/\tau)) \\ & + O\left(2R_{2m}(x)S_m(y/\tau) + (1-S_m(x))I_m(y/\tau)\right)\end{aligned}$$

Donc, en collectant les termes et en appliquant des résultats dans les lemmes (3.2.5) et (3.2.6), on obtient :

$$\mathbb{E}\left[\eta_i^2(x,y)\right] = H(x,y) - 2m^{-1/2}\left(H_x(x,y)[2\pi x(1-x)]^{1/2} + H_y(x,y)\tau\left[2\pi\frac{y}{\tau}\left(1-\frac{y}{\tau}\right)\right]^{1/2}\right) + o(m^{-1/2}).$$

Maintenant, nous déduisons à partir des expressions précédentes que :

$$\begin{aligned}\mathbb{V}ar\left(\eta_i(x,y)\right) &= H(x,y) - 2m^{-1/2}\left(H_x(x,y)[2\pi x(1-x)]^{1/2} + H_y(x,y)[2\pi y(\tau-y)]^{1/2}\right) + o(m^{-1/2}) \\ &= \sigma^2(x,y) - m^{-1/2}V(x,y) + o(m^{-1/2}).\end{aligned}$$

Le théorème suivant donne un résultat théorique intéressant de notre estimateur puisqu'il est utile pour les tests et la construction des intervalles de confiance. C'est la normalité asymptotique et elle est énoncée dans le théorème suivant.

Théorème 3.4.4. *Sous l'hypothèse (1) et les même condition du théorème (3.4.2). Nous avons,*

1. *Pour m et n tendant vers l'infini, alors*

$$n^{1/2}\left(\widehat{F}_{m,n}(x,y) - F_m^{\tau}(x,y)\right) \xrightarrow{D} N\left(0, \sigma^2(x,y)\right).$$

2. *De plus, si $mn^{-1/2}$ tend vers l'infini :*

$$n^{1/2}\left(\widehat{F}_{m,n}(x,y) - F(x,y)\right) \xrightarrow{D} N\left(0, \sigma^2(x,y)\right).$$

Preuve du Théorème 3.4.4. Le théorème (3.4.2) nous donne la représentation *i.i.d.*, donc la normalité asymptotique se déduit en vérifiant la condition de Lindeberg du théorème de la limite centrale.
Nous commençons par définir les variables suivantes :

$$Z_i = \eta_i(x,y) - F_m^{\tau}(x,y)$$

et

$$s_n^2 = nVar\left(\eta_i(x,y)\right) = n\sigma^2(x,y) - nm^{-1/2}V(x,y) + o(nm^{-1/2}).$$

Le théorème de la limite centrale s'applique si la condition de lindeberg suivante est vérifiée :

$$s_n^{-2} \sum_{i=1}^{n} E\left(Z_i^2 1_{\{|Z_i| \geq \epsilon s_n\}}\right) \longrightarrow 0.$$

C'est facile de remarquer que : $\frac{s_n}{\sqrt{n}} \xrightarrow{D} \sigma(x,y)$ quand $n \to \infty$, nous avons aussi $|Z_i| \leq 2$ pour tout i alors, d'après le théorème de la limite centrale, nous avons le résultat :

$$\frac{\sqrt{n}\left(\sum_{i=1}^{n} \eta_i(x,y) - F_m^{\tau}(x,y)\right)}{\frac{s_n}{\sqrt{n}}} \xrightarrow{D} N(0,1)$$

et par conséquent :

$$n^{1/2}\left(\widehat{F}_{m,n}(x,y) - F_m^{\tau}(x,y)\right) \xrightarrow{D} N\left(0, \sigma^2(x,y)\right).$$

Pour la deuxième partie du théorème, les même techniques de démonstration sont utilisées.

Chapitre 4

Estimation de la copule bivariée par les polynômes de Bernstein

4.1 Généralités sur les copules bivariées

L'étude de dépendance entre deux variables aléatoires est un sujet d'étude qui a attiré l'attention des statisticiens depuis très longtemps. Les mesures de dépendance ponctuelles comme le coefficient de corrélation, le tau de Kendall et le rho de Spearman sont souvent utilisées pour décrire la dépendance entre deux variables statistiques. La copule est un outil relativement robuste pour modéliser toute la structure de dépendance entre plusieurs variables statistiques et semble être mieux représentative qu'un simple coefficient . L'application dans de nombreux domaines, où la dépendance entre variables est utile, fait appel à ce concept. Dans ce chapitre nous allons donner des généralités sur la théorie des copules. Ensuite, nous rappellerons quelques exemples des copules classiques. Finalement, nous proposons un estimateur non-paramétrique de la copule et nous étudions la convergence presque sûre de cet esti-

mateur. Afin de mieux présenter les calculs, nous nous restreignons dans ce chapitre, à l'étude des copules bidimensionnelles.

Définition 4.1.1. Une copule bivariée est une fonction $C : [0,1]^2 \longrightarrow [0,1]$ qui vérifie les conditions suivantes :

(i) $C(0.u) = C(u,0) = 0. \quad \forall u \in [0,1]$.

(ii) $C(1.u) = C(u,1) = u. \quad \forall u \in [0,1]$. les lois marginales sont uniformes dans $[0,1]$.

(iii) $C(u_2,v_2) - C(u_2,v_1) - C(u_1,v_2) + C(u_1,v_1) \geq 0. \quad \forall (u_1,v_1),(u_2,v_2) \in [0,1]^2$ tel que $u_2 \geq u_1$ et $v_2 \geq v_1$. On dit que C est 2-non-décroissante.

Une copule C nous permet d'écrire une fonction de répartition F d'un couple (X_1, X_2) à l'aide des fonctions de répartition marginales de chacune des deux variables. Nous énonçons dans la suite, le fameux théorème de Sklar (1959) qui constitue le fondement théorique de la théorie des copules.

Théorème 4.1.1. *[Sklar(1959)]*
Soit H une fonction de répartition d'un couple de variables aléatoires, de lois marginales H_1 et H_2, alors il existe une copule C telle que pour tout $x = (x_1, x_2) \in \mathbb{R}^2$

$$H(x_1, x_2) = C(H_1(x_1), H_2(x_2)).$$

Démonstration. Pour $(x_1, x_2) \in \mathbb{R}^2$, on a :

$$\begin{aligned} H(x_1,x_2) &= P(X_1 \leq x_1, X_2 \leq x_2) \\ &= P(H_1(X_1) \leq H_1(x_1), H_2(X_2) \leq H_2(x_2)) \\ &= C\left(H_1(x_1), H_2(x_2)\right). \end{aligned}$$

□

Le théorème de Sklar permet donc de relier la notion de copule à celle de fonction de répartition bivariée.

Le corollaire suivant est fondamental pour l'estimation de la copule.

Corollaire 4.1.2. *Soit H une fonction de répartition conjointe, de marginales H_1 et H_2 continues. Soient H_1^{-1} et H_2^{-1} les fonctions inverses généralisées de H_1 et H_2 respectivement. Alors pour tout $u = (u_1, u_2) \in [0,1]^2$ on a :*

$$C(u_1, u_2) = H\left(H_1^{-1}(u_1), H_2^{-1}(u_2)\right). \tag{4.1.1}$$

C est donc unique.

Démonstration. Sous l'hypothèse de la continuité de H_1 et H_2 et pour $(u_1, u_2) \in [0,1]^2$ on a :

$$\begin{aligned} C(u_1, u_2) &= P(H_1(X_1) \leq u_1, H_2(X_2) \leq u_2) \\ &= P(X_1 \leq H_1^{-1}(u_1), X_2 \leq H_2^{-1}(u_2)) \\ &= H\left(H_1^{-1}(u_1), H_2^{-1}(u_2)\right). \end{aligned}$$

□

La densité d'une copule :

En supposant la différentiabilité de H_1 et de H_2, la densité de la loi jointe du vecteur aléatoire (X_1, X_2) a la forme suivante :

$$f(x_1, x_2) = f_1(x_1) f_2(x_2) c\left(H_1(x_1), H_2(x_2)\right)$$

telle que c est la densité de la copule C, définie par

$$c(u_1, u_2) = \frac{\partial^2 C(u_1, u_2)}{\partial u_1 \partial u_2}.$$

On peut réécrire la densité de la copule en fonction des densités marginales, de la densité jointe et les inverses généralisées de la façon suivante :

$$c(u_1, u_2) = \frac{f\left(H_1^{-1}(u_1), H_2^{-1}(u_2)\right)}{f_1\left(H_1^{-1}(u_1)\right) f_2\left(H_2^{-1}(u_2)\right)} \tag{4.1.2}$$

À partir de cette relation, on obtient des estimateurs de la densité de la copule en utilisant des estimateurs de la densité jointe et des estimateurs des lois marginales.

Pour plus de détails concernant les fondements de la théorie des copules, nous orientons le lecteur vers les travaux de Renyi [39] et Schweizer et al. [41]

4.2 Modèles de copules :

Il existe un grand nombre de copules adaptées à différentes situations de dépendance ; toute fonction de répartition associée à un vecteur dont les marginales sont uniformes sur [0.1] définit une copule. Toutefois, quelques formes particulières sont souvent utilisées en pratique du fait de leur simplicité de mise en œuvre. Nous pouvons notamment citer les exemples ci-après :

4.2.1 Copules elliptiques

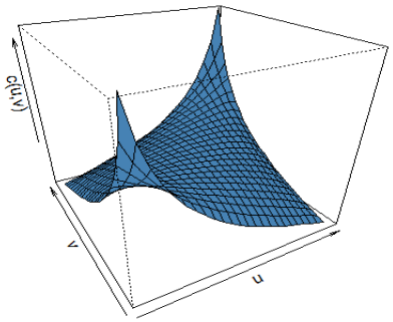
La famille des copules elliptiques est une famille de copules qui comprend deux exemples classiques : la copule Gaussienne et la copule de Student.

La copule Gaussienne

Définition 4.2.1. Soit Φ la fonction de répartition d'une loi normale centrée réduite et Φ^{-1} son inverse. La copule Gaussienne bidimensionnelle est donnée par :

$$C_\rho(u_1, u_2) = \Phi_\rho(\Phi^{-1}(u_1), \Phi^{-1}(u_2))$$

où ρ est le coefficient de corrélation, Φ_ρ représente la distribution normale centrée de matrice de corrélation égale à $\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$. Le graphe suivant montre une densité d'une copule gaussienne de paramètre $\rho = 0.5$



La copule Gaussienne modélise les formes de dépendances linéaires de la même façon que la distribution normale, mais avec des distributions marginales quelconques. Cependant, cette copule ne présente pas de dépendance de queue, en d'autres mots, elle ne présente pas de dépendance pour les petites et grandes valeurs de u_1 et u_2. Elle ne peut donc pas capter la dépendance dans les extrêmes. Cette propriété des copules gaussienne motive l'utilisant de la copule de Student.

Copule de Student

Définition 4.2.2. Soit T_ν la fonction de répartition d'une loi de Student de ν degrés de liberté, T_ν^{-1} est son inverse. La copule de Student est donnée par :

$$C_{(\rho,\nu)}(u_1, u_2) = T_{(\rho,\nu)}(T^{-1}(u_1), T^{-1}(u_2))$$

où ρ est le coefficient de corrélation $T_{(\rho,\nu)}$ est la distribution de Student bidimensionnelle standard de matrice de corrélation en fonction de ρ et de degrés de liberté égale à ν.

La figure (4.1) illustre la densité d'une de copule de Student bidimensionnelle avec $\rho = 0.5$ et $\nu = 4$.

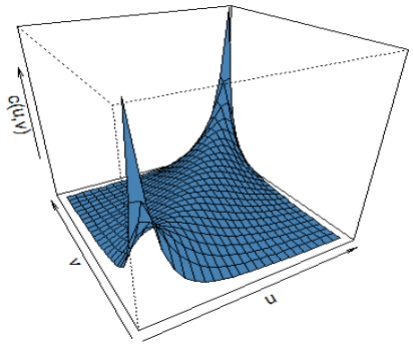


FIGURE 4.1 – Densité d'une de copule Student bidimensionnelle avec $\rho = 0.5$ et $\nu = 4$

Pour résumer, les deux copules, Gaussienne et de Student, sont utilisées pour modéliser la dépendance symétrique. Toutefois, lorsque le coefficient de corrélation est proche de 0, la copule de Student, avec des petites valeurs du degré de liberté, montre des dépendances de queue à droite (pour les petites valeurs de u_1 et u_2) et à gauche (pour les grandes valeurs de u_1 et u_2) contrairement à la copule Gaussienne.

4.2.2 Copules Archimédiennes

Une famille de copules particulièrement importante dans la pratique est celles de copules Archimédiennes. Cette classe de copule a été introduite par Genest et Mackay dans [18]. Pour une étude approfondie sur ces copules, le lecteur peut se référer à l'ouvrage de Nelsen [36].

Les copules Archimédiennes sont très populaires en statistique car elles sont faciles à construire et à simuler, en plus elles peuvent exprimer une large variété de structures de dépendance. L'idée d'une copule Archimédienne, de générateur ϕ, est que la

tranformation $\omega(u) = \exp(-\phi(u))$ appliquée aux marginales « rend les composantes indépendantes » :

$$\omega(C(u_1, u_2)) = \omega(u_1)\omega(u_2)$$

Définition 4.2.3. Soit ϕ une fonction positive, définie sur $[0, 1]$ et qui vérifie les conditions suivante :

(i) ϕ est deux fois continûment dérivable sur $[0, 1]$.

(ii) $\phi(1) = 0$, $\phi'(u) < 0$ et $\phi''(u) > 0$ sur $[0, 1]$.

La copule Archimidienne se définit alors par :

$$C(u_1, u_2) = \begin{cases} \phi^{-1}\left(\phi(u_1) + \phi(u_2)\right) & si \quad \phi(u_1) + \phi(u_2) \leq \phi(0) \\ 0 & sinon. \end{cases}$$

La fonction ϕ est dite la fonction génératrice de la copule archimidienne. En effet, chaque copule Archimidienne est définie à partir d'un choix d'une fonction ϕ vérifiant les deux conditions précédentes. Dans ce qui suit, nous présentons trois familles de copules Archimidiennes. nous commençons par introduire la copule de Gumbel.

Copule de Gumbel

La copule de Gumbel est obtenue en choisissant la fonction génératrice suivante

$$\phi_\theta(s) = (-\ln(s))^\theta$$

pour $\theta \geq 1$, et donc $\phi_\theta^{-1}(s) = \exp(-s^{1/\theta})$. Par conséquent, la Copule de Gumbel est donnée par :

$$C(u_1, u_2) = \exp\left(-\left[(-\ln u_1)^\theta + (-\ln u_2)^\theta\right]^{1/\theta}\right).$$

Cette copule est utile pour modéliser les structures de dépendance qui se caractérisent par une dépendance moyenne (à faible) dans la queue de gauche de la distribution

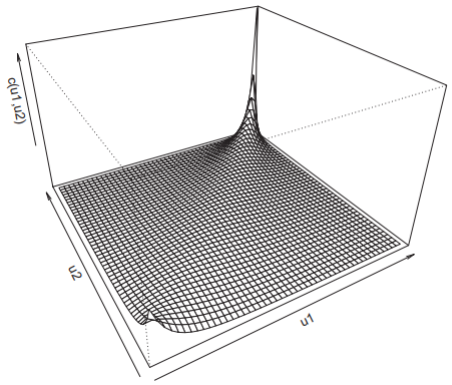


FIGURE 4.2 – Densité de la copule de Gumbel bidimensionnelle avec $\theta = 3$.

jointe et une dépendance importante dans la queue de droite de la distribution jointe. La figure (4.2) illustre la densité d'une copule de Gumbel bidimensionnelle avec $\theta = 3$.

Copule de Clayton

C'est une copule archimidienne obtenue à partir du générateur :

$$\phi_\theta(s) = \frac{1}{\theta}\left(s^{-\theta} - 1\right)$$

pour $\theta \geq 0$, et donc $\phi_\theta^{-1}(s) = (\theta s + 1)^{-\frac{1}{\theta}}$. Par conséquent, la Copule de Clayton est donnée explicitement par :

$$C(u_1, u_2) = \left(u_1^{-\theta} + u_1^{-\theta} - 1\right)^{-\frac{1}{\theta}}.$$

La copule de Clayton est adéquate dans le cas de forte dépendance pour les petites valeurs de X_1 et X_2 et une dépendance presque nulle dans la queue de droite. La figure (4.3) représente la densité d'une copule de Clayton bidimensionnelle avec $\theta = 0.7$.

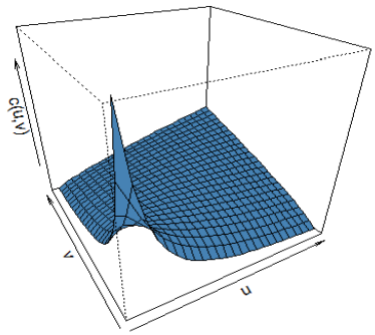


FIGURE 4.3 – Densité de la copule de Clayton bidimensionnelle avec $\theta = 0.7$.

Copule de Frank

Une autre copule Archimidienne obtenue à partir du générateur :

$$\phi_\theta(s) = -\ln\left[\frac{e^{-\theta s}-1}{e^{\theta}-1}\right]$$

pour $\theta \in \mathbb{R}^*$, et aussi, $\phi_\theta^{-1}(s) = -\frac{1}{\theta}\ln\left[e^{-s}(e^{\theta}-1)+1\right]$.

Par conséquent, la copule de Frank est donnée explicitement par :

$$C(u_1,u_2) = -\frac{1}{\theta}\ln\left[1+\frac{(e^{-\theta u_1}-1)(e^{-\theta u_2}-1)}{(e^{-\theta}-1)}\right].$$

Cette copule joue un rôle pour modéliser les structures de dépendance où le phénomène étudié n'admet pas de dépendance à la queue, ni à gauche ni à droite. La figure (4.4) représente la densité d'une copule de Frank bidimensionnelle avec $\theta = 3$.

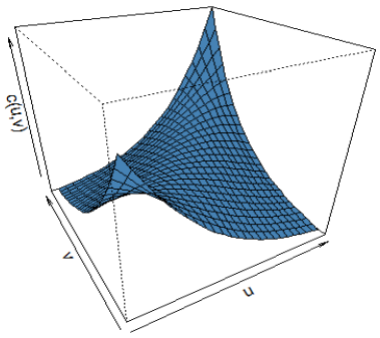


FIGURE 4.4 – Densité de la copule de Frank bidimensionnelle avec $\theta = 3$.

Remarque 4.2.1. *Les copules elliptiques permettent des corrélations positives et négatives. Cependant, les trois copules Archimedienne modélisent les dépendance positives. Pour les données avec une dépendance négative, souvent on utilise des rotations de trois copules Archimidienne. le livre de Joe (2014) [21] présente des techniques pour le choix de la copule qui modélise une formes de dépendance.*

4.3 Estimation non paramétrique de la copule :

Dans la pratique, la copule qui décrit la structure de dépendance n'est évidemment pas connue en général. Le but est donc de l'estimer à partir des observations. Plusieurs approches sont proposées pour estimer la copule dans une étude. Dans cette section, nous nous limitons à l'approche non paramétrique conformément au sujet de notre thèse car cette approche ne postule aucune hypothèse à la forme de dépendance, en plus, elle s'utilise en présence de données de types divers.

4.3.1 Copule empirique

Un estimateur naturel de la copule est motivé par la relation (4.1.1), c'est La copule empirique qui a été introduite par Deheuvels dans [8] sous le nom de "fonction empirique de dépendance". Supposons qu'on a des réalisations $(X_1, Y_1), (X_2, Y_2), ..., (X_n, X_n)$ du couple (X, Y) de fonction jointe H dont les marginales sont H_1 et H_2, la copule empirique introduite par Deheuvels(1980) pour les données completes est définie par :

$$C_n(u_1, u_2) = H_n\left(H_{1,n}^{-1}(u_1), H_{2,n}^{-1}(u_2)\right) \tag{4.3.1}$$

où $H_n(x, y) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{X_i \leq x, Y_i \leq y\}}$, $H_{1,n}(x) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{Y_i \leq x\}}$ et $H_{2,n}(x) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{Y_i \leq y\}}$.

Deheuvels [8] a montré que cette copule empirique satisfait plusieurs propriétés, on cite la convergence presque sûre vers la vraie copule C et la normalité asymptotique. Le processus de la copule empirique est présenté dans sa forme standard par :

$$\mathbb{C}_n(u_1, u_2) = \sqrt{n}\left(C_n(u_1, u_2) - C(u_1, u_2)\right), \qquad (u_1, u_2) \in [0, 1]^2 \tag{4.3.2}$$

Dans le cas i.i.d., le comportement asymptotique du processus $\mathbb{C}_n$ est établi dans Wellner (2013)[50]. Des généralisations sur des espaces plus larges, de copules, sont étudiées dans plusieurs travaux, voir comme exemples, Fermanian (2004) [13], Van der Vaart et Wellner (2007) [48] et Segers (2012) [42].

4.4 Estimateur non paramétrique de la copule par les polynômes de Bernstein :

Dans cette partie, nous proposons un estimateur de la copule bivariée C par les polynômes de Bernstein. Nous considérons un cas de les données incomplètes à cause

de la censure à droite d'une seule variable. Cet estimateur est noté par $\widehat{C}_{m,n}$ et il est construit à partir de notre estimateur $\widehat{F}_{m,n}$ pour la fonction de répartition bivariée étudiée dans la section (3.3).

Soient F_1 et F_2 les fonctions de réparition marginales de X et Y respectivement, $\widehat{F}_1$ et $\widehat{F}_2$ leurs estimateurs empiriques associés. Nous proposons d'estimer la fonction de copule par

$$\widehat{C}_{m,n}(u,v) = \widehat{F}_{m,n}(\widehat{F}_1^{-1}(u), \widehat{F}_2^{-1}(v)), \tag{4.4.1}$$

où $\widehat{F}_1^{-1}$ et $\widehat{F}_2^{-1}$ sont respectivement les inverses généralisés de $\widehat{F}_1$ et $\widehat{F}_2$.

Nous établissons, dans le théorème (4.4.1), la convergence presque sûre de $\widehat{C}_{m,n}(u,v)$ vers $C(u,v)$ sur $[0,1]^2$. La vitesse de convergence est également donnée dans ce théorème.

Théorème 4.4.1. *Sous l'hypothèse* $m^{-1/2} = O(n^{-1/2}(\log\log n)^{1/2})$ *, pour* m *et* n *tendant vers l'infini on a :*

$$\left|\left|\widehat{C}_{m,n} - C\right|\right|_\tau = O_{p.s.}\left(n^{-1/4}\log n\right),$$

Où

$$||Z||_\tau = \sup_{0\le u\le 1, 0\le v\le F_2(\tau)} |Z(u,v)|.$$

Démonstration. Nous commençons la preuve par l'identité suivante :

$$\begin{aligned}
\widehat{C}_{m,n}(u,v) - C(u,v) &= \widehat{F}_{m,n}(\widehat{F}_1^{-1}(u), \widehat{F}_2^{-1}(v)) - \widehat{F}_{m,n}(F_1^{-1}(u), F_2^{-1}(v)) \\
&+ \widehat{F}_{m,n}(F_1^{-1}(u), F_2^{-1}(v)) - F(F_1^{-1}(u), F_2^{-1}(v)) \\
&\equiv (I) + (II).
\end{aligned}$$

pour la partie (II), nous avons d'après la proposition (3.4.1), si $m = O(n^{-1/2}(\log\log n)^{1/2})$, alors :

$$\sup_{0\leq x\leq 1, 0\leq y\leq \tau} \left|\widehat{F}_{m,n}(x,y) - F(x,y)\right| = O_{p.s.}(n^{-1/2}(\log\log n)^{1/2}),$$

il en vient donc :

$$\sup_{0\leq x\leq 1, 0\leq y\leq \tau} \left|\widehat{F}_{m,n}(F_1^{-1}(u), F_2^{-1}(v)) - F(F_1^{-1}(u), F_2^{-1}(v))\right| = O_{a.s.}(n^{-1/2}(\log\log n)^{1/2}). \tag{4.4.2}$$

Maintenant pour la partie (I), nous utilisons un développement de Taylor de la quantité $\widehat{F}_{m,n}(\widehat{F}_1^{-1}(u), \widehat{F}_2^{-1}(v))$ au voisinage du point $(F_1^{-1}(u), F_2^{-1}(v)) =: (\xi_1, \xi_2)$. Ce développement nous permet d'écrire :

$$\begin{aligned}
\widehat{F}_{m,n}(\widehat{F}_1^{-1}(u), \widehat{F}_2^{-1}(v)) &= \widehat{F}_{m,n}(\xi_1, \xi_2) \\
&+ \frac{\partial \widehat{F}_{m,n}(\tilde{\xi}_1, \tilde{\xi}_2)}{\partial \xi_1}\left[\widehat{F}_1^{-1}(u) - \tilde{\xi}_1\right] \\
&+ \frac{\partial \widehat{F}_{m,n}(\tilde{\xi}_1, \tilde{\xi}_2)}{\partial \xi_2}\left[\widehat{F}_2^{-1}(u) - \tilde{\xi}_2\right],
\end{aligned}$$

où $\tilde{\xi}_1$ (respectivement $\tilde{\xi}_2$) est entre $\widehat{F}_1^{-1}(u)$ et $F_1^{-1}(u)$ (respectivement entre $\widehat{F}_2^{-1}(v)$ et $F_2^{-1}(v)$). Concernant les quantités $\widehat{F}_1^{-1}(u) - \tilde{\xi}_1$ et $\widehat{F}_2^{-1}(u) - \xi_2$, nous avons un résultat dans le lemme 3 dans Lo et Singh [29] qui indique :

$$\left|\left|\widehat{F}_i^{-1} - F_i^{-1}\right|\right| = O_{p.s.}\left(n^{-1/2}\log n\right) \quad \text{pour} \quad i = 1, 2. \tag{4.4.3}$$

Maintenant pour $\frac{\partial \widehat{F}_{m,n}(\tilde{\xi}_1, \tilde{\xi}_2)}{\partial \xi_1} =: \widehat{f}_{m,n1}(\tilde{\xi}_1, \tilde{\xi}_2)$, nous avons

$$\begin{aligned}\widehat{f}_{m,n1}(\tilde{\xi}_1,\tilde{\xi}_2) &= \sum_{k=0}^{m}\sum_{\ell=0}^{m}\widehat{F}_n^{ST}\left(\frac{k}{m},\frac{\tau\ell}{m}\right)P'_{k,m}(\tilde{\xi}_1)P^{\tau}_{\ell,m}(\tilde{\xi}_2)\\ &= \frac{m}{\tau}\sum_{k=0}^{m-1}\sum_{\ell=0}^{m}\left[\widehat{F}_n^{ST}\left(\frac{k}{m},\frac{\tau\ell}{m}\right)-\widehat{F}_n^{ST}\left(\frac{k-1}{m},\frac{\tau\ell}{m}\right)\right]P_{k,m}(\tilde{\xi}_1)P^{\tau}_{\ell,m}(\tilde{\xi}_2).\end{aligned}$$

Pour la différence $\left[\widehat{F}_n^{ST}\left(\frac{k}{m},\frac{\tau\ell}{m}\right)-\widehat{F}_n^{ST}\left(\frac{k-1}{m},\frac{\tau\ell}{m}\right)\right]=:D$ nous posons

$$\begin{aligned}D &= \left[\widehat{F}_n^{ST}\left(\frac{k}{m},\frac{\tau\ell}{m}\right)-F\left(\frac{k}{m},\frac{\tau\ell}{m}\right)\right]-\left[\widehat{F}_n^{ST}\left(\frac{k-1}{m},\frac{\tau\ell}{m}\right)-F\left(\frac{k-1}{m},\frac{\tau\ell}{m}\right)\right]\\ &+ \left[F\left(\frac{k}{m},\frac{\tau\ell}{m}\right)-F\left(\frac{k-1}{m},\frac{\tau\ell}{m}\right)\right]\\ &= D_1+D_2.\end{aligned}$$

Pour le terme D_1, la proposition 2 dans Rabhi et Bouezmarni [38] nous donne la vitesse de convergence de ce terme, en fait $D_1 = O_{p.s.}\left(n^{-3/4}\log n\right)$. Concernant le terme D_2, sous l'hypothèse que F soit lipschitzienne d'ordre 1 on obtient :

$$\begin{aligned}\left|\left[F\left(\frac{k}{m},\frac{\tau\ell}{m}\right)-F\left(\frac{k-1}{m},\frac{\tau\ell}{m}\right)\right]\right| &\leq \left|\frac{k}{m}-\frac{k-1}{m}\right|\\ &= \frac{1}{m},\end{aligned}$$

d'où $D_2 = O(m^{-1})$. Finalement le résultat du théorème en découle.

□

Remarque 4.4.2. *Le fait que $\widehat{C}_{m,n}$ soit dérivable, un estimateur de la densité de la copule bivariée du couple (X,Y) est déduit. Il est donné comme suit :*

$$\begin{aligned}\widehat{c}_{m,n}(u,v) &= \frac{\partial^2\widehat{C}_{m,n}(u,v)}{\partial u\partial v}\\ &= \sum_{k=0}^{m}\sum_{\ell=0}^{m}\widehat{F}_n^{ST}\left(\frac{k}{m},\frac{\tau\ell}{m}\right)\left(P_{k,m}\left(\widehat{F}^{-1}_{m,n,1}(u)\right)\right)'\left(P^{\tau}_{\ell,m}\left(\widehat{F}^{-1}_{m,n,2}(v)\right)\right)'.\end{aligned}$$

Malheureusement, la forme de l'estimateur est assez compliqué.

Remarque 4.4.3. *Pour régler le problème de l'estimateur précédent, nous proposons un autre estimateur de la copule dont on peut déduire plus facilement un estimateur de la densité de la copule qui est donné par*

$$\tilde{C}_{m,n}(u,v) = \sum_{k=0}^{m}\sum_{\ell=0}^{m} \tilde{C}_n\left(\frac{k}{m}, \frac{\tau\ell}{m}\right) P_{k,m}(u) P^{\tau}_{\ell,m}(v) \tag{4.4.4}$$

où $\tilde{C}_n(x,y) = \widehat{F}_n^{ST}\left(H_{n,1}^{-1}(x)\right), \left(H_{n,2}^{-1}(y)\right)$. $H_{n,1}^{-1}(x)$ *est l'estimateur empirique de le fonction de répartition pour les données complètes et* $H_{n,2}^{-1}(y)$ *est l'estimateur de Kaplan-Meier pour les donnée censurées à droite.*

Remarque 4.4.4. *Un autre estimateur peut se déduire des estimateurs présentés dans les équations* (4.4.3) *et* (4.4.4)*, c'est un estimateur du tau de Kendall* τ *pour le modèle de censure considéré dans le chapitre 3. Nous partons de la définition probabiliste du tau de Kendall donnée pour un vecteur de variables aléatoires continues, comme suit :*

$$\begin{aligned} \tau &= -1 + 4\int C(U,V)\, dP \\ &= -1 + 4\mathbb{E}\left(C(U,V)\right). \end{aligned}$$

L'estimateur non paramétrique du τ *de Kendall que nous proposons, est donné par l'expression :*

$$\widehat{\tau}_{m,n}(X,Y) = -1 + 4\int_0^1\int_0^1 \widetilde{C}_{m,n}(u,v)\widetilde{c}_{m,n}(u,v)\, du\, dv \tag{4.4.5}$$

où $\widetilde{C}_{m,n}(u,v)$ *et* $\widetilde{c}_{m,n}(u,v)$ *sont les estimateurs non paramétriques de la copule et de la densité de la copule, respectivement.*

Chapitre 5

Applications et simulations

Ce chapitre est consacré à une étude de simulation suivie d'une application de notre estimateur de la fonction de répartition sur des données réelles. Le but de cette partie est de comparer la performance de notre estimateur avec l'estimateur de Stute (1993).

5.1 Étude de simulation

Dans cette partie, nous allons faire une comparaison entre notre estimateur de la fonction de répartition bivariée et l'estimateur de Stute (1993) sur des données simulées. Cette étude de simulation est faite avec le Logiciel R. Nous allons commencer par expliquer le principe de notre simulation et définir les différents modèles que nous considérons dans la suite. Les résultats de simulation seront recueillis dans les tableaux de la section (5.1.3) et finalement nous discuterons les résultats.

5.1.1 Présentation des modèles

Nous considérons les mêmes hypothèses de censure définies dans la section (2.4.3). Notre variable d'intérêt est le couple (X, Y), au lieu d'observer ce vecteur, on observe un vecteur aléatoire (X, T, δ) où $T = Y \wedge C$, C est la variable aléatoire de censure, et $\delta = \mathbb{I}_{\{Y<C\}}$ est l'indicateur de la censure. On suppose que C est indépendante de Y.

Nous commençons par générer, à partir de plusieurs modèles, des réalisations d'un couple de variables aléatoires (U, V) dont la copule associée est connue. Nous obtenons un échantillon de réalisations du vecteur aléatoire $\left(F_X^{-1}(U), F_Y^{-1}(V)\right)$. Dans la pratique, on appelle ces réalisation des pseudo-observations. Pour faire entrer le phénomène de censure dans cette simulation, la variable Y est supposée censurée par une variable C. Le taux de censure dans chaque échantillon simulé, est contrôlé par le paramètre p, nous allons prendre $p = 50\%$ puis $p = 25\%$. Les modèles que nous allons considérer dans notre étude de simulation sont :

(i) Modèle 1 : Le vecteur (U, V) est choisi suivant une copule gaussienne.
(ii) Modèle 2 : Le vecteur (U, V) est choisi suivant une copule de Clayton.
(iii) Modèle 3 : Le vecteur (U, V) est choisi suivant une copule de Gumbel.

Pour chaque modèle nous varions les lois marginales pour faire la comparaison à travers plusieurs cas de figures. Nous générons X selon la loi bêta de paramètres $B(2, 2)$ et Y selon la loi bêta de paramètres $(\alpha, \beta) = (3, 3)\, et\, (0.5, 0.5)$. Ce choix pour la variable Y est proposé pour avoir deux lois marginales de formes différentes.

Pour la variable de censures C, nous la prenons suivant une loi uniforme sur $[0; a_p]$

où $a_p = \frac{\alpha}{(\alpha+\beta)p}$. La valeur de a_p assure un degré de censure égal à p. En effet,

$$\begin{aligned} p &= P(C \leq Y) \\ &= 1 - P(C \geq Y) \\ &= 1 - \int_0^1 \int_y^{a_p} \frac{1}{a_p}.f_Y(y)dc\ dy \\ &= 1 - \frac{1}{a_p}\int_0^1 (a_p - y)f_Y(y)\ dy \\ &= 1 - \frac{1}{a_p}\left(a_p - \frac{\alpha}{\alpha+\beta}\right) \\ &= \frac{\alpha}{(\alpha+\beta)a_p}. \end{aligned}$$

Nous contrôlons la corrélation des données, dans chaque modèle, par un choix de la valeur du taux de Kendall (Tau), nous proposons les valeurs suivantes ($Tau = 0.25$ puis $Tau = 0.75$).

Les réalisations simulées sont utilisées pour calculer $\widehat{F}_{m,n}$ et $\widehat{F}_n^{ST}$ pour différentes tailles d'échantillons ($n = 50$, $n = 100$, $n = 200$ et $n = 500$). Le nombre de réplications effectué est égale à 2000 réplications.

5.1.2 Les critères de comparaison :

Notre estimation dépend de l'échelle du lissage. Nous choisissons une grille, de taille dépendante de la valeur du τ, $\mathbf{s}_i = (u_i, w_i) \in \{(0.01, 0.01), (0.01, 0.03), \ldots, (\tau, \tau)\}$, qu'on va utiliser pour calculer le biais, la variance et l'erreur quadratique moyenne (EQM) pour chacun des deux estimateurs $\widehat{F}_{m,n}$ et $\widehat{F}_n^{ST}$. La valeur de τ est choisie selon les échantillon simulés, elle nous permet d'avoir l'intervalle sur lequel nous calculons les biais intégrés et les variances intégrées pour chaque cas de figure.

Biais

Le biais d'un estimateur $\hat{\theta}$ de θ est égal à

$$Biais(\hat{\theta}) := \mathbb{E}(\hat{\theta}) - \theta.$$

Un estimateur avec un bias positif est un estimateur qui sur-estime en moyenne la vraie valeur de θ. Un biais négatif indique que l'estimateur sous-estime en moyenne le vrai paramètre. Un estimateur $\hat{\theta}$ est sans biais pour θ si et seulement si son bias est nul.

Dans notre travail, on estime une fonction et donc on calcule le carré du biais intégré donné explicitement par :

$$\mathbb{IB}iais^2\left(\widehat{F}_{n,m,j}\right) = \frac{1}{I}\sum_{i=1}^{I}\left(\frac{1}{N}\sum_{j=1}^{N}\widehat{F}_{n,m,j}(\mathbf{s}_i) - F(\mathbf{s}_i)\right)^2.$$

$\widehat{F}_{n,m,j}(\boldsymbol{s})$, est notre estimateur de Bernstein obtenu dans la j^{eme} réplication et $F(\boldsymbol{s})$ étant la fonction de répartition théorique calculée à partir du modèle choisi, et ce pour $j = 1, \ldots, N$.

Par le même principe nous obtenons le carré du biais intégré de l'estimateur de Stute :

$$\mathbb{IB}iais^2\left(\widehat{F}^{ST}_{n,j}\right) = \frac{1}{I}\sum_{i=1}^{I}\left(\frac{1}{N}\sum_{j=1}^{N}\widehat{F}^{ST}_{n,j}(\mathbf{s}_i) - F(\mathbf{s}_i)\right)^2$$

Variance

la variance d'un estimateur θ est généralement définie par l'expression :

$$\mathbb{IV}ar(\theta) = \mathbb{E}(\hat{\theta^2}) - \left[\mathbb{E}(\hat{\theta})\right]^2.$$

Dans notre modèle , la variance intégrée est donnée par

$$\mathbb{IV}ar\left(\widehat{F}_{n,m,j}\right) = \frac{1}{I}\sum_{i=1}^{I}\left(\frac{1}{N}\sum_{j=1}^{N}\left(\widehat{F}_{n,m,j}(\mathbf{s}_i) - \frac{1}{N}\sum_{j}\widehat{F}_{n,m,j}(\mathbf{s}_i)\right)^2\right)^2. \qquad (5.1.1)$$

et pour l'estimateur de Stute,

$$\mathbb{V}ar\left(\widehat{F}_{n,j}^{ST}\right) = \frac{1}{I}\sum_{i=1}^{I}\left(\frac{1}{N}\sum_{j=1}^{N}\left(\widehat{F}_{n,j}^{ST}(\mathbf{s}_i) - \frac{1}{N}\sum_{j}\widehat{F}_{n,j}^{ST}(\mathbf{s}_i)\right)^2\right)^2. \tag{5.1.2}$$

Erreur quadratique moyenne

L'Erreur quadratique moyenne notée EQM est une bonne mesure du bruit aléatoire, en effet elle combine le biais et la variance à la fois. Généralement, son expression est donnée comme suit :

$$EQM_\theta(\hat{\theta}) = \mathbb{E}(\hat{\theta} - \theta)^2.$$

Un résultat qui exprime l'EQM en fonction de la variance et du biais est donné par :

$$EQM(\hat{\theta}) = Bias(\hat{\theta})^2 + Var(\theta)$$

Par conséquent, les erreurs quadratiques moyennes associée à nos estimateurs sont alors :

$$\begin{aligned} EQM\left(\widehat{F}_{n,m,j}\right) &= \mathbb{B}ias^2 + \mathbb{V}ar \\ &\approx \frac{1}{I}\sum_{i=1}^{I}\left(\left(\frac{1}{N}\sum_{j=1}^{N}\widehat{F}_{n,m,j}(\mathbf{s}_i) - F(\mathbf{s}_i)\right)^2\right) \\ &\quad + \frac{1}{I}\sum_{i=1}^{I}\left(\frac{1}{N}\sum_{j=1}^{N}\left(\widehat{F}_{n,m,j}(\mathbf{s}_i) - \frac{1}{N}\sum_{j}\widehat{F}_{n,m,j}(\mathbf{s}_i)\right)^2\right)^2, \end{aligned}$$

et pour l'estimateur de Stute, l'expression de l'EQM est :

$$\begin{aligned} EQM\left(\widehat{F}_{n,j}^{ST}\right) &= \mathbb{IB}ias^2 + \mathbb{IV}ar \\ &\approx \frac{1}{I}\sum_{i=1}^{I}\left(\left(\frac{1}{N}\sum_{j=1}^{N}\widehat{F}_{n,j}^{ST}(\mathbf{s}_i) - F(\mathbf{s}_i)\right)^2\right) \\ &\quad + \frac{1}{I}\sum_{i=1}^{I}\left(\frac{1}{N}\sum_{j=1}^{N}\left(\widehat{F}_{n,j}^{ST}(\mathbf{s}_i) - \frac{1}{N}\sum_{j}\widehat{F}_{n,j}^{ST}(\mathbf{s}_i)\right)^2\right)^2. \end{aligned}$$

5.1.3 Résultats de simulation

Les résultats de simulation (biais, variance et erreur quadratique moyenne) sont reportés dans les tables 5.1-5.9.

Biais intégré

Pour les trois modèles de copules avec deux tau de Kendall, les biais intégrés sont présentés dans les tables 5.1, 5.4 et 5.7 pour les deux scénarios des marginales et pour les deux degrés de censure. Pour tous les cas de figure (modèles de copules, lois des marginales et valeurs de taux), si on fixe une valeur du taux de censure, on remarque que le biais des deux estimateurs diminue avec la taille de l'échantillon. Par exemple, dans la table 5.1 avec $Y \sim Beta(3,3)$ et $Tau = 0.25$, pour $p = 0.25$, le biais de notre estimateur est égal à 130.2954e-06 pour $n = 50$ et 1.187839e-06 pour $n = 500$. La variation du paramètre de l'indépendance n'influe pas sur la diminution du biais en augmentant la valeur de n. En effet, pour une valeur du tau de Kendall $Tau = 0.75$; le biais diminue en augmentant la taille de l'échantillon. Aussi, notre estimateur domine l'estimateur de Stute au niveau du bias pour les modèles considérés. Nous remarquons aussi qu'avec un taux de censure élevé $p = 0.50$, le biais intégré obtenu

par notre estimateur est souvent plus bas que celui obtenu par l'estimateur de Stute. Pour chaque modèle, on constate que si on fixe la taille de l'échantillon, le biais augmente pour un taux de censure élevé dans l'échantillon simulé. Pour les grandes valeurs de n, par exemple $n = 500$, on note que parfois, l'estimateur de Stute donne une valeur du biais intégré inférieure à celle obtenue par notre estimateur, c'est une conséquence du fait que l'estimateur de Stute est l'estimateur empirique.

Variance intégrée

En analysant les tables 5.2, 5.5, et 5.8 pour les deux cas de figure des marginales et pour les deux taux de censure dans les échantillons simulés. Pour tous les scénarios (modèles de copules, lois des marginales et valeurs de taux), si on fixe une valeur de la censure, on remarque que la variance des deux estimateurs diminue avec l'augmentation de la taille de l'échantillon. Par exemple, dans la table 5.2 avec $Y \sim Beta(3, 3)$ et $Tau = 0.25$, pour $p = 0.25$, la variance de notre estimateur est égale à 1.382712e-06 pour $n = 50$ et 0.1385608e-07 pour $n = 500$. Aussi, notre estimateur domine absolument l'estimateur de Stute au niveau de la variance pour les modèles considérés. Nous remarquons aussi que même avec un taux de censure élevé $p = 0.50$, la variance intégrées obtenue par notre estimateur est toujours inférieure à celle obtenue par l'estimateur de Stute et ce quelle que soit la taille de l'échantillon. Notre estimateur est une forme de lissage de l'estimateur de Stute, par conséquent, la variance intégrée de cette forme de lissage est toujours inférieure à la variance intégrée obtenue par la forme discrète.

L'erreur quadratique moyenne

Comme dans les cas avec le biais intégré et la variance intégrée, nous remarquons d'après les tables 5.3, 5.6, et 5.9 que notre estimateur domine l'estimateur de Stute au niveau de l'erreur moyenne quadratique, pour les deux cas de figure des marginales et pour les deux taux de censure dans les échantillons simulés. Pour tous les scénarios (modèles de copules, lois des marginales et valeurs de taux), si on fixe une valeur de la censure, on remarque que l'erreur quadratique moyenne des deux estimateurs diminue avec l'augmentation de la taille de l'échantillon. Par exemple, dans la table 5.3 avec $Y \sim Beta(3,3)$ et $\tau = 0.25$, pour $p = 0.25$, l'erreur quadratique moyenne de notre estimateur est égale à 131.5361e-6 pour $n = 50$ et 1.312085e-06 pour $n = 500$. Cette diminution importante est un bon signe de performance de notre estimateur pour la fonction de répartition bivariée. Aussi, notre estimateur domine globalement l'estimateur de Stute au niveau de l'erreur quadratique moyenne pour les modèles considérés. Nous remarquons aussi que même avec un taux de censure élevé $p = 0.50$, l'erreur quadratique moyenne obtenue par notre estimateur est toujours inférieure à celle obtenue par l'estimateur de Stute indépendamment de la taille de l'échantillon.

Modèle 1 : Copule Gaussienne								
Le taux de Kendall = 0.25								
Marginales	$\boldsymbol{Y \sim beta(3,3),\quad X \sim beta(2,2)\ et\quad C \sim U([0;a_p])}$							
	$n=50$		$n=100$		$n=200$		$n=500$	
Estimation	$p=0.5$	$p=0.25$	$p=0.5$	$p=0.25$	$p=0.5$	$p=0.25$	$p=0.5$	$p=0.25$
$\widehat{F}_n^{ST}$	14.079	11.37	3.27	3.45	1.34	0.905	0.179	0.17
$\hat{F}_{n,m}$	13.03	10.34	2.81	2.94	0.58	0.682	0.119	0.086
Marginales	$\boldsymbol{Y \sim beta(0.5,0.5),\quad X \sim beta(2,2)\ et\quad C \sim U([0;a_p])}$							
	$n=50$		$n=100$		$n=200$		$n=500$	
Estimation	$p=0.5$	$p=0.25$	$p=0.5$	$p=0.25$	$p=0.5$	$p=0.25$	$p=0.5$	$p=0.25$
$\widehat{F}_n^{ST}$	12.71	6.97	2.77	2.43	1.03	0.505	0.126	0.062
$\widehat{F}_{n,m}$	12.63	6.69	2.8	2.27	1.03	0.451	0.129	0.039
Le taux de Kendall = 0.75								
Marginales	$\boldsymbol{Y \sim beta(3,3),\quad X \sim beta(2,2)\ et\quad C \sim U([0;a_p])}$							
	$n=50$		$n=100$		$n=200$		$n=500$	
Estimation	$p=0.5$	$p=0.25$	$p=0.5$	$p=0.25$	$p=0.5$	$p=0.25$	$p=0.5$	$p=0.25$
$\widehat{F}_n^{ST}$	7.42	6.4	1.43	0.59	0.59	0.248	0.125	0.033
$\hat{F}_{n,m}$	6.44	5.5	1.11	0.37	0.39	0.156	0.104	0.1
Marginales	$\boldsymbol{Y \sim beta(0.5,0.5),\quad X \sim beta(2,2)\ et\quad C \sim U([0;a_p])}$							
	$n=50$		$n=100$		$n=200$		$n=500$	
Estimation	$p=0.5$	$p=0.25$	$p=0.5$	$p=0.25$	$p=0.5$	$p=0.25$	$p=0.5$	$p=0.25$
$\widehat{F}_n^{ST}$	6.27	4.31	2.11	0.58	0.61	0.377	0.053	0.109
$\hat{F}_{n,m}$	5.76	3.86	1.85	0.47	0.507	0.299	0.069	0.096

TABLE 5.1 – Biais ($\times 10^5$) intégrés des deux estimateurs selon le modèle 1 et les différents scénarios considérés pour la dépendance, les lois marginales, les tailles des échantillons et les taux de censure.

Modèle 1 : Copule Gaussienne								
Le taux de Kendall = 0.25								
Marginales	$\boldsymbol{Y \sim beta(3,3),\quad X \sim beta(2,2)\ et\quad C \sim U([0;a_p])}$							
	$n=50$		$n=100$		$n=200$		$n=400$	
Estimation	$p=0.5$	$p=0.25$	$p=0.5$	$p=0.25$	$p=0.5$	$p=0.25$	$p=0.5$	$p=0.25$
$\widehat{F}_n^{ST}$	0.14	0.197	0.0688	0.099	0.034	0.050	0.0139	0.02
$\hat{F}_{n,m}$	0.12	0.174	0.0616	0.088	0.0304	0.044	0.0124	0.018
Marginales	$\boldsymbol{Y \sim beta(0.5,0.5),\quad X \sim beta(2,2)\ et\quad C \sim U([0;a_p])}$							
	$n=50$		$n=100$		$n=200$		$n=500$	
Estimation	$p=0.5$	$p=0.25$	$p=0.5$	$p=0.25$	$p=0.5$	$p=0.25$	$p=0.5$	$p=0.25$
$\widehat{F}_n^{ST}$	0.177	0.26	0.090	0.13	0.045	0.068	0.018	0.027
$\hat{F}_{n,m}$	0.165	0.24	0.084	0.12	0.042	0.062	0.016	0.024
Le taux de Kendall = 0.75								
Marginales	$\boldsymbol{Y \sim beta(3,3),\quad X \sim beta(2,2)\ et\quad C \sim U([0;a_p])}$							
	$n=50$		$n=100$		$n=200$		$n=500$	
Estimation	$p=0.5$	$p=0.25$	$p=0.5$	$p=0.25$	$p=0.5$	$p=0.25$	$p=0.5$	$p=0.25$
$\widehat{F}_n^{ST}$	0.15	0.2	0.073	0.1	0.036	0.047	0.015	0.02
$\hat{F}_{n,m}$	0.13	0.17	0.066	0.09	0.032	0.042	0.013	0.017
Marginales	$\boldsymbol{Y \sim beta(0.5,0.5),\quad X \sim beta(2,2)\ et\quad C \sim U([0;a_p])}$							
	$n=50$		$n=100$		$n=200$		$n=500$	
Estimation	$p=0.5$	$p=0.25$	$p=0.5$	$p=0.25$	$p=0.5$	$p=0.25$	$p=0.5$	$p=0.25$
$\widehat{F}_n^{ST}$	0.191	0.25	0.092	0.128	0.047	0.064	0.0191	0.026
$\hat{F}_{n,m}$	0.179	0.23	0.087	0.118	0.044	0.059	0.0179	0.024

TABLE 5.2 – Variances intégrées ($\times 10^5$) des deux estimateurs selon le modèle 1 et les différents scénarios considérés.

Modèle 1 : Copule Gaussienne								
Le taux de Kendall = 0.25								
Marginales	$\boldsymbol{Y \sim beta(3,3),\quad X \sim beta(2,2)\ et\quad C \sim U([0;a_p])}$							
	$n=50$		$n=100$		$n=200$		$n=400$	
Estimation	$p=0.5$	$p=0.25$	$p=0.5$	$p=0.25$	$p=0.5$	$p=0.25$	$p=0.5$	$p=0.25$
$\widehat{F}_n^{ST}$	14.22	11.57	1.34	3.55	0.84	0.95	0.793	0.192
$\hat{F}_{n,m}$	13.15	10.51	2.88	3.02	0.61	0.726	0.131	0.104
Marginales	$\boldsymbol{Y \sim beta(0.5,0.5),\quad X \sim beta(2,2)\ et\quad C \sim U([0;a_p])}$							
	$n=50$		$n=100$		$n=200$		$n=500$	
Estimation	$p=0.5$	$p=0.25$	$p=0.5$	$p=0.25$	$p=0.5$	$p=0.25$	$p=0.5$	$p=0.25$
$\widehat{F}_n^{ST}$	12.89	7.23	2.85	2.56	1.07	0.572	0.144	0.051
$\hat{F}_{n,m}$	12.80	6.94	2.88	2.39	1.07	0.513	0.146	0.064
Le taux de Kendall = 0.75								
Marginales	$\boldsymbol{Y \sim beta(3,3),\quad X \sim beta(2,2)\ et\quad C \sim U([0;a_p])}$							
	$n=50$		$n=100$		$n=200$		$n=500$	
Estimation	$p=0.5$	$p=0.25$	$p=0.5$	$p=0.25$	$p=0.5$	$p=0.25$	$p=0.5$	$p=0.25$
$\widehat{F}_n^{ST}$	7.57	6.66	1.5	0.685	0.625	0.295	0.139	0.053
$\hat{F}_{n,m}$	6.57	5.7	1.18	0.459	0.428	0.198	0.117	0.123
Marginales	$\boldsymbol{Y \sim beta(0.5,0.5),\quad X \sim beta(2,2)\ et\quad C \sim U([0;a_p])}$							
	$n=50$		$n=100$		$n=200$		$n=500$	
Estimation	$p=0.5$	$p=0.25$	$p=0.5$	$p=0.25$	$p=0.5$	$p=0.25$	$p=0.5$	$p=0.25$
$\widehat{F}_n^{ST}$	6.44	4.56	2.205	0.71	0.66	0.44	0.0720	0.135
$\hat{F}_{n,m}$	5.95	4.09	1.94	0.59	0.55	0.36	0.087	0.120

TABLE 5.3 – EQM ($\times 10^5$) des deux estimateurs selon le modèle 1 et les différents scénarios considérés.

Modèle 2 : Copule de Clayton								
Le taux de Kendall = 0.25								
Marginales	$\boldsymbol{Y \sim beta(3,3),\quad X \sim beta(2,2)\ et\quad C \sim U([0;a_p])}$							
	$n=50$		$n=100$		$n=200$		$n=500$	
Estimation	$p=0.5$	$p=0.25$	$p=0.5$	$p=0.25$	$p=0.5$	$p=0.25$	$p=0.5$	$p=0.25$
$\widehat{F}_n^{ST}$	16.35	9.58	4.84	3.29	0.96	0.600	0.198	0.202
$\hat{F}_{n,m}$	15.35	8.87	4.28	2.85	0.428	0.418	0.114	0.133
Marginales	$\boldsymbol{Y \sim beta(0.5,0.5),\quad X \sim beta(2,2)\ et\quad C \sim U([0;a_p])}$							
	$n=50$		$n=100$		$n=200$		$n=500$	
Estimation	$p=0.5$	$p=0.25$	$p=0.5$	$p=0.25$	$p=0.5$	$p=0.25$	$p=0.5$	$p=0.25$
$\widehat{F}_n^{ST}$	11.93	13.003	3.41	2.69	0.818	1.16	0.184	0.125
$\hat{F}_{n,m}$	12.08	12.64	3.5	2.578	0.884	1.09	0.202	0.108
Le taux de Kendall = 0.75								
Marginales	$\boldsymbol{Y \sim beta(3,3),\quad X \sim beta(2,2)\ et\quad C \sim U([0;a_p])}$							
	$N=50$		$n=100$		$n=200$		$n=500$	
Estimation	$p=0.5$	$p=0.25$	$p=0.5$	$p=0.25$	$p=0.5$	$p=0.25$	$p=0.5$	$p=0.25$
$\widehat{F}_n^{ST}$	7.94	6.03	1.757	2.16	0.35	0.38	0.132	0.15
$\hat{F}_{n,m}$	7.09	5.14	1.487	1.72	0.31	0.28	0.145	0.166
Marginales	$\boldsymbol{Y \sim beta(0.5,0.5),\quad X \sim beta(2,2)\ et\quad C \sim U([0;a_p])}$							
	$n=50$		$n=100$		$n=200$		$n=500$	
Estimation	$p=0.5$	$p=0.25$	$p=0.5$	$p=0.25$	$p=0.5$	$p=0.25$	$p=0.5$	$p=0.25$
$\widehat{F}_n^{ST}$	6.86	3.57	1.68	1.08	0.23	0.35	0.172	0.210
$\hat{F}_{n,m}$	6.59	3.26	1.62	0.94	0.26	0.299	0.165	0.178

TABLE 5.4 – Biais ($\times 10^5$) des deux estimateurs selon le modèle 2 et les différents scénarios considérés.

Modèle 2 : Copule de Clayton								
Le taux de Kendall = 0.25								
Marginales	$\boldsymbol{Y \sim beta(3,3),\quad X \sim beta(2,2)\ et\quad C \sim U([0;a_p])}$							
	$N = 50$		$N = 100$		$N = 200$		$N = 400$	
Estimation	$p = 0.5$	$p = 0.25$	$p = 0.5$	$p = 0.25$	$p = 0.5$	$p = 0.25$	$p = 0.5$	$p = 0.25$
$\widehat{F}_n^{ST}$	0.137	0.203	0.073	0.098	0.035	0.048	0.0141	0.202
$\widehat{F}_{n,m}$	0.122	0.181	0.066	0.087	0.032	0.048	0.0127	0.018
Marginales	$\boldsymbol{Y \sim beta(0.5,0.5),\quad X \sim beta(2,2)\ et\quad C \sim U([0;a_p])}$							
	$N = 50$		$N = 100$		200		$N = 400$	
Estimation	$p = 0.5$	$p = 0.25$	$p = 0.5$	$p = 0.25$	$p = 0.5$	$p = 0.25$	$p = 0.5$	$p = 0.25$
$\widehat{F}_n^{ST}$	0.186	0.251	0.090	0.121	0.044	0.065	0.018	0.0268
$\widehat{F}_{n,m}$	0.174	0.233	0.085	0.065	0.041	0.06	0.017	0.0246
Le taux de Kendall = 0.75								
Marginales	$\boldsymbol{Y \sim beta(3,3),\quad X \sim beta(2,2)\ et\quad C \sim U([0;a_p])}$							
	$n = 50$		$n = 100$		$n = 200$		$n = 500$	
Estimation	$p = 0.5$	$p = 0.25$	$p = 0.5$	$p = 0.25$	$p = 0.5$	$p = 0.25$	$p = 0.5$	$p = 0.25$
$\widehat{F}_n^{ST}$	0.144	0.202	0.074	0.096	0.037	0.049	0.0142	0.192
$\widehat{F}_{n,m}$	0.130	0.181	0.067	0.086	0.033	0.044	0.0128	0.0171
Marginales	$\boldsymbol{Y \sim beta(0.5,0.5),\quad X \sim beta(2,2)\ et\quad C \sim U([0;a_p])}$							
	$N = 50$		$N = 100$		200		$N = 400$	
Estimation	$p = 0.5$	$p = 0.25$	$p = 0.5$	$p = 0.25$	$p = 0.5$	$p = 0.25$	$p = 0.5$	$p = 0.25$
$\widehat{F}_n^{ST}$	0.190	0.253	0.096	0.128	0.047	0.063	0.0190	0.026
$\widehat{F}_{n,m}$	0.180	0.236	0.091	0.119	0.045	0.058	0.0180	0.024

TABLE 5.5 – Variances intégrées ($\times 10^5$) des deux estimateurs selon le modèle 2 et les différents scénarios considérés.

Modèle 2 : Copule de Clayton								
Le taux de Kendall = 0.25								
Marginales	$\boldsymbol{Y \sim beta(3,3),\quad X \sim beta(2,2)\ et\quad C \sim U([0;a_p])}$							
	$n=50$		$n=100$		$n=200$		$n=500$	
Estimation	$p=0.5$	$p=0.25$	$p=0.5$	$p=0.25$	$p=0.5$	$p=0.25$	$p=0.5$	$p=0.25$
$\widehat{F}_n^{ST}$	1649	9.78	4.91	3.39	0.995	0.648	0.203	0.222
$\widehat{F}_{n,m}$	15.47	9.05	4.34	2.94	0.779	0.460	0.127	0.151
Marginales	$\boldsymbol{Y \sim beta(0.5,0.5),\quad X \sim beta(2,2)\ et\quad C \sim U([0;a_p])}$							
	$n=50$		$n=100$		$n=200$		$n=500$	
Estimation	$p=0.5$	$p=0.25$	$p=0.5$	$p=0.25$	$p=0.5$	$p=0.25$	$p=0.5$	$p=0.25$
$\widehat{F}_n^{ST}$	12.12	13.25	3.50	2.819	0.863	1.22	0.202	1.52
$\widehat{F}_{n,m}$	12.25	12.87	3.61	2.699	0.925	1.154	0.219	0.133
Le taux de Kendall = 0.75								
Marginales	$\boldsymbol{Y \sim beta(3,3),\quad X \sim beta(2,2)\ et\quad C \sim U([0;a_p])}$							
	$n=50$		$n=100$		$n=200$		$n=500$	
Estimation	$p=0.5$	$p=0.25$	$p=0.5$	$p=0.25$	$p=0.5$	$p=0.25$	$p=0.5$	$p=0.25$
$\widehat{F}_n^{ST}$	8.09	6.23	1.83	2.258	0.387	0.424	0.148	0.166
$\widehat{F}_{n,m}$	7.24	5.33	1.55	1.804	0.348	0.329	6.0.158	0.183
Marginales	$\boldsymbol{Y \sim beta(0.5,0.5),\quad X \sim beta(2,2)\ et\quad C \sim U([0;a_p])}$							
	$n=50$		$n=100$		200		$n=500$	
Estimation	$p=0.5$	$p=0.25$	$p=0.5$	$p=0.25$	$p=0.5$	$p=0.25$	$p=0.5$	$p=0.25$
$\widehat{F}_n^{ST}$	7.053	3.83	1.777	1.21	0.278	0.408	1.191	0.240
$\widehat{F}_{n,m}$	6.767	3.493	1.708	1.05	0.305	0.358	0.183	0.202

TABLE 5.6 – EQM ($\times 10^5$) des deux estimateurs selon le modèle 2 et les différents scénarios considérés.

Modèle 3 : Copule de Gumbel								
Le taux de Kendall = 0.25								
Marginales	$\boldsymbol{Y \sim beta(3,3),\quad X \sim beta(2,2)\ et\quad C \sim U([0;a_p])}$							
	$n = 50$		$n = 100$		$n = 200$		$n = 500$	
Estimation	$p = 0.5$	$p = 0.25$	$p = 0.5$	$p = 0.25$	$p = 0.5$	$p = 0.25$	$p = 0.5$	$p = 0.25$
$\widehat{F}_n^{ST}$	11.15	9.962	2.90	3.056	0.884	0.825	0.162	0.102
$\widehat{F}_{n,m}$	10.26	8.916	2.42	2.584	0.649	0.633	0.101	0.066
Marginales	$\boldsymbol{Y \sim beta(0.5,0.5),\quad X \sim beta(2,2)\ et\quad C \sim U([0;a_p])}$							
	$n = 50$		$n = 100$		200		$n = 500$	
Estimation	$p = 0.5$	$p = 0.25$	$p = 0.5$	$p = 0.25$	$p = 0.5$	$p = 0.25$	$p = 0.5$	$p = 0.25$
$\widehat{F}_n^{ST}$	12.27	7.994	2.445	2.69	0.854	0.810	0.18	0.184
$\widehat{F}_{n,m}$	12.04	7.696	2.395	2.509	0.8224	0.732	0.1388	0.149
Le taux de Kendall = 0.75								
Marginales	$\boldsymbol{Y \sim beta(3,3),\quad X \sim beta(2,2)\ et\quad C \sim U([0;a_p])}$							
	$n = 50$		$n = 100$		$n = 200$		$n = 500$	
Estimation	$p = 0.5$	$p = 0.25$	$p = 0.5$	$p = 0.25$	$p = 0.5$	$p = 0.25$	$p = 0.5$	$p = 0.25$
$\widehat{F}_n^{ST}$	5.696	4.192	1.906	4.154	0.382	0.391	0.0377	0.03
$\widehat{F}_{n,m}$	4.0799	3.362	1.461	3.516	0.303	0.235	0.0774	0.09
Marginales	$\boldsymbol{Y \sim beta(0.5,0.5),\quad X \sim beta(2,2)\ et\quad C \sim U([0;a_p])}$							
	$n = 50$		$n = 100$		200		$n = 500$	
Estimation	$p = 0.5$	$p = 0.25$	$p = 0.5$	$p = 0.25$	$p = 0.5$	$p = 0.25$	$p = 0.5$	$p = 0.25$
$\widehat{F}_n^{ST}$	7.30	3.77	4.48	1.732	0.763	0.665	0.0634	0.0341
$\widehat{F}_{n,m}$	6.76	3.33	1.26	1.466	0.614	0.521	0.0673	0.091

TABLE 5.7 – Biais intégrés ($\times 10^5$) des deux estimateurs selon le modèle 3 et les différents scénarios considérés.

Modèle 1 : Copule de Gumbel								
Le taux de Kendall = 0.25								
Marginales	$\boldsymbol{Y \sim beta(3,3),\quad X \sim beta(2,2)\ et\quad C \sim U([0;a_p])}$							
	$n=50$		$n=100$		$n=200$		$n=500$	
Estimation	$p=0.5$	$p=0.25$	$p=0.5$	$p=0.25$	$p=0.5$	$p=0.25$	$p=0.5$	$p=0.25$
$\widehat{F}_n^{ST}$	0.140	0.194	0.067	0.098	0.034	0.048	0.139	0.0198
$\widehat{F}_{n,m}$	0.125	0.171	0.060	0.087	0.03	0.0428	0.0125	0.0175
Marginales	$\boldsymbol{Y \sim beta(0.5,0.5),\quad X \sim beta(2,2)\ et\quad C \sim U([0;a_p])}$							
	$n=50$		$n=100$		$n=200$		$n=500$	
Estimation	$p=0.5$	$p=0.25$	$p=0.5$	$p=0.25$	$p=0.5$	$p=0.25$	$p=0.5$	$p=0.25$
$\widehat{F}_n^{ST}$	0.178	0.255	0.086	0.129	0.0445	0.0648	0.0180	0.0271
$\widehat{F}_{n,m}$	0.166	0.235	0.0799	0.118	0.0414	0.0593	0.0168	0.0247
Le taux de Kendall = 0.75								
Marginales	$\boldsymbol{Y \sim beta(3,3),\quad X \sim beta(2,2)\ et\quad C \sim U([0;a_p])}$							
	$n=50$		$n=100$		$n=200$		$n=500$	
Estimation	$p=0.5$	$p=0.25$	$p=0.5$	$p=0.25$	$p=0.5$	$p=0.25$	$p=0.5$	$p=0.25$
$\widehat{F}_n^{ST}$	0.142	0.190	0.0723	0.0962	0.0361	0.482	0.0141	0.0194
$\widehat{F}_{n,m}$	0.128	0.168	0.0652	0.0858	0.0326	0.0430	0.0127	0.0173
Marginales	$\boldsymbol{Y \sim beta(0.5,0.5),\quad X \sim beta(2,2)\ et\quad C \sim U([0;a_p])}$							
	$n=50$		$n=100$		$n=200$		$n=400$	
Estimation	$p=0.25$	$p=0.5$	$p=0.25$	$p=0.5$	$p=0.25$	$p=0.5$	$p=0.25$	$p=0.5$
$\widehat{F}_n^{ST}$	0.185	0.254	0.0947	0.129	0.0464	0.0643	0.0189	0.0261
$\widehat{F}_{n,m}$	0.173	0.235	0.0887	0.119	0.0435	0.593	0.0177	0.0240

TABLE 5.8 – Variances intégrées ($\times 10^5$) des deux estimateurs selon le modèle 3 et les différents scénarios considérés.

Modèle 1 : Copule de Gumbel								
Le taux de Kendall = 0.25								
Marginales	$\boldsymbol{Y \sim beta(3,3),\quad X \sim beta(2,2)\ et\quad C \sim U([0;a_p])}$							
	$n=50$		$n=100$		$n=200$		$n=500$	
Estimation	$p=0.5$	$p=0.25$	$p=0.5$	$p=0.25$	$p=0.5$	$p=0.25$	$p=0.5$	$p=0.25$
$\widehat{F}_n^{ST}$	11.29	10.16	2.96	3.15	0.918	0.873	0.18	0.121
$\widehat{F}_{n,m}$	10.38	9.087	2.48	2.67	0.180	0.676	0.11	0.083
Marginales	$\boldsymbol{Y \sim beta(0.5,0.5),\quad X \sim beta(2,2)\ et\quad C \sim U([0;a_p])}$							
	$n=50$		$n=100$		$n=200$		$n=500$	
Estimation	$p=0.5$	$p=0.25$	$p=0.5$	$p=0.25$	$p=0.5$	$p=0.25$	$p=0.5$	$p=0.25$
$\widehat{F}_n^{ST}$	12.44	8.250	2.531	2.820	0.898	0.874	0.180	0.211
$\widehat{F}_{n,m}$	12.21	7.931	2.475	2.626	8.864	0.791	0.156	0.174
Le taux de Kendall = 0.75								
Marginales	$\boldsymbol{Y \sim beta(3,3),\quad X \sim beta(2,2)\ et\quad C \sim U([0;a_p])}$							
	$n=50$		$n=100$		$n=200$		$n=500$	
Estimation	$p=0.25$	$p=0.5$	$p=0.25$	$p=0.5$	$p=0.25$	$p=0.5$	$p=0.25$	$p=0.5$
$\widehat{F}_n^{ST}$	5.840	4.38	1.980	4.250	0.418	0.439	0.0518	0.0500
$\widehat{F}_{n,m}$	4.927	3.53	1.527	3.602	0.336	0.278	0.0900	0.108
Marginales	$\boldsymbol{Y \sim beta(0.5,0.5),\quad X \sim beta(2,2)\ et\quad C \sim U([0;a_p])}$							
	$n=50$		$n=100$		$n=200$		$n=500$	
Estimation	$p=0.5$	$p=0.25$	$p=0.5$	$p=0.25$	$p=0.5$	$p=0.25$	$p=0.5$	$p=0.25$
$\widehat{F}_n^{ST}$	7.484	4.021	1.574	1.862	0.810	0.730	0.0823	0.0603
$\hat{F}_{n,m}$	6.929	3.569	1.353	1.586	0.657	0.580	0.0850	0.115

TABLE 5.9 – EQM ($\times 10^5$) des deux estimateurs selon le modèle 3 et les différents scénarios de simulation.

5.2 Application sur des données réelles

L'étude de dépendance entre deux variables statistiques est largement appliquée dans le domaine des assurances et actuariats. Généralement, l'objectif est de prédire tout genre de dégâts et les dépenses couvertes par la compagne d'assurance. Pour analyser des données réelles en appliquant notre estimateur, nous avons pris de la littérature du domaine un échantillon de 1500 observations. Cette base de données a été collectée par "the Insurance Services Office (ISO) ", New Jersey, USA et publiée par Denuit et al. [9]. Les données à disposition rapportent des paiements d'indemnités (Loss, X) et des frais alloué au règlements des sinistres alloués (Allocated loss adjustment expenses, $ALAE$), associés à chaque réclamation d'indemnités. En pratique, chaque contrat d'assurance est associée à un montant maximum du sinistre couvert qui est spécifique à chaque contrat (la variable L). Les réclamations prennent généralement plus de temps à être réglées, par conséquent, la variable X est censurée à droite aux cas où le montant réclamé dépasse le montant maximum du sinistre couvert L à cause des coûts imprévus comme les charges d'investigation sur le dégât ou les frais des conflits judiciaires (frais des avocats par exemples). De telles données sont censurées à droite, car on ne sait pas le coût final causé par le dégât, par contre on sait que ce montant dépasse le coût dépensé jusqu'au temps où l'étude a été réalisée. Ces données ont été étudiées dans plusieurs travaux sur les données censurée à droite, comme par exemple [9] et [35]. Un résumé de ces données est présenté dans la table (5.10).

	X_i	Y_i	Données complètes	Données censurées
Nombre d'observations	1.500	1500	1.466	34
Minimum	10	15	10	5.000
premier quartile	4.000	2.333	3.750	50.000
Moyenne	41.208	12.588	37.110	217.941
Médiane	12.000	5.471	11.049	100.000
Troisième quartile	35.000	12.577	32.000	300.000
Maximum	2.173.595	501.863	2.173.595	1.000.000

TABLE 5.10 – Table des données en dollars Américains, recueillies par (ISO)

Statistiquement, on observe dans ce problème, les triplés (T_i, Y_i, δ_i) qui sont des réalisations du vecteur aléatoire (T, Y, δ), où $T_i = X_i \wedge L_i$ et δ_i l'indicateur de censure à droite défini comme suit :

$$\delta_i = \begin{cases} 1 \quad si \quad \mathrm{X}_i \leq L_i \\ 0 \quad si \quad \mathrm{X}_i > L_i \end{cases}$$

Le nuage de points dans la figure (5.1) illustre graphiquement les données de l'étude sur une échelle logarithmique. Les points en rouge sont les données censurées à droite (les points se situent à droite du nuage).

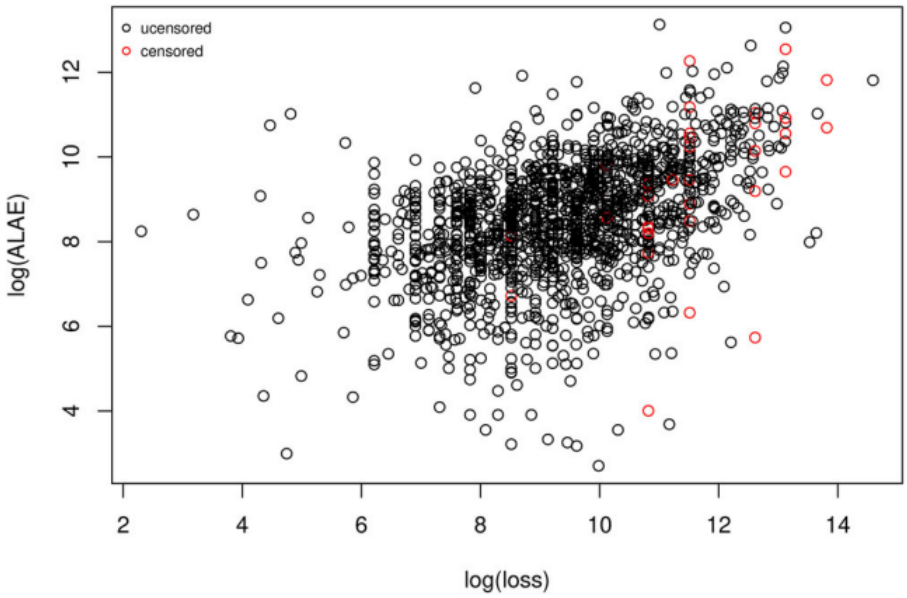


FIGURE 5.1 – Représentation graphique des variables Loss et ALAE sur une échelle logarithmique

.

Ce graphique montre qu'il y a une forte dépendance positive entre les deux variables d'intérêt. Cet aspect suggère de faire une analyse conjointe de ces deux variables. Denuit et al. [9] ont étudié la corrélation entre les variables Loss et ALAE en utilisation la modélisation par les copules. Ils ont supposé que la structure de dépendance entre les deux variables se modélise par la copule Archimidienne avec l'opérateur ϕ. Le but de leur travail c'est de proposer un estimateur non paramétrique pour le générateur ϕ, associé au modèle de copule Archimidienne. Ce modèle tient compte du fait que les pertes peuvent être censurées alors que les ALAE sont complètement observables. Cependant, nous n'avons pas d'informations sur la loi suivie par le vec-

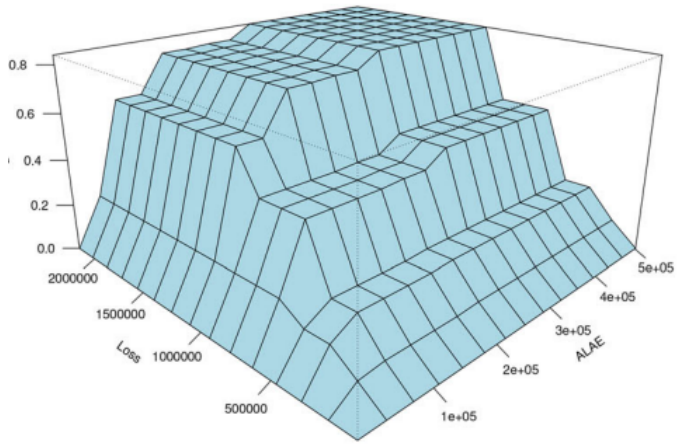
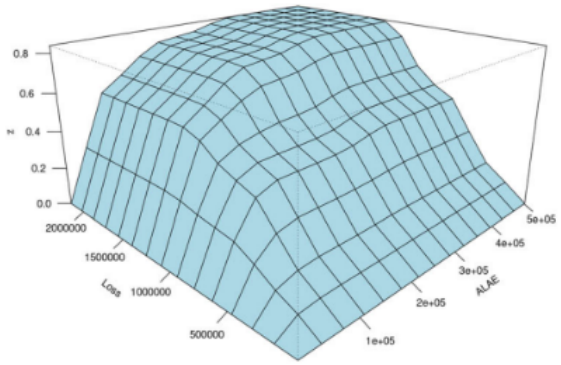


FIGURE 5.2 – Estimateur de Stute pour Loss et ALAE

teur aléatoire considéré. L'hypothèse de la copule Archimedienne n'est pas justifié. En fait, une mauvaise spécification du modèle de copule peut biaiser les conclusions même si l'opérateur ϕ est estimé avec une approche non-paramétrique. Par conséquent nous choisissons l'approche non-paramétrique pour traiter ces données. À cette fin, nous estimons la fonction de répartition jointe en utilisant l'estimateur de Stute (1993) dans (5.2)(a). Pour faire la comparaison, nous faisons l'estimation de la fonction de répartition par notre estimateur défini au chapitre 3, l'estimateur obtenu est présenté graphiquement dans (5.3)(b).



(b)

FIGURE 5.3 – Estimateur de Bernstein pour Loss et ALAE

La comparaison entre ces deux représentations nous montre que la meilleure estimation au sens de lissage, est celle construite par notre estimateur, ainsi cet aspect de lissage nous permet d'avoir la densité jointe et d'autres caractéristiques tels que les quantiles et la régression.

Conclusion et Perspectives de recherche

A la fin de ce travail, nous présentons nos conclusions et des perspectives futures pour notre recherche.

5.3 Conclusion :

Les travaux réalisés au cours de cette thèse ont apporté de nouveaux résultats dans l'estimation non-paramétrique en présence de données censurées. Notre contribution principale est l'estimation non-paramétrique de la fonction de répartition bivariée par les polynômes de Bernstein, en présence d'un certain scénario de censure. La convergence de notre estimateur ainsi que la vitesse de cette convergence ont été établies. Nous avons donné le biais et la variance associés avant d'énoncer notre résultat sur la normalité asymptotique de l'estimateur en question.

Une autre partie de notre contribution, est que nous avons extrait un estimateur non-paramétrique de la copule bivariée par les polynômes de Bernstein sous le même modèle de censure. Nos estimateurs proposés jouissent de la propriété du lissage, ce qui nous permet d'avoir rapidement, de nouveaux estimateurs en profitant du lissage des estimateurs en question. Une validation numérique de nos résultats a été réalisée à travers des simulations pour comparer notre estimateur de la fonction de répartition bivariée à celui de Stute.

5.4 Perspectives de la thèse

Nous proposons, ci dessous, quelques perspectives de recherches.

5.4.1 Extension à d'autres modèles de censure

Ce travail a été mené pour un modèle de censure à droite, très fréquent dans la pratique, mais d'autres modèles de censure peuvent exister dans la pratique (censure mixte, double...) et justifier l'extension de notre travail à ces cadres.

5.4.2 Test statistiques

Les tests statistiques sont un outil pour valider ou réfuter des hypothèses de modélisation statistique, ils sont très populaire dans l'inférence statistique. Notre estimateur peut être utilisé pour tester l'hypothèse du papier de Denuit et al. [9], c'est à dire si le modèle de dépendance est selon une copule Archimidienne.
Nous pouvons étudier les tests d'adéquation (tests of goodness of fit) en utilisant

notre estimateur pour tester l'hypothèse

$$H_0 : C = C_\theta$$

contre l'hypothèse alternative

$$H_1 : C \neq C_\theta$$

La statistique du test T sera basée sur la différence entre $\widehat{C}_{m,n}$, et $C_{\hat{\theta}}$, où $\hat{\theta}$ est estimée d'une façon consistante. Un autre travail portera sur les tests d'indépendance, l'hypothèse nulle qu'on veut tester est :

$$H_0 : C(u, v) = uv$$

contre l'hypothèse alternative :

$$H_1 : C(u, v) \neq uv$$

La statistique de ce test s'écrit comme une fonction $\mathbb{F}\left(\widehat{C}_{m,n} - uv\right)$. Pour un test de Cramer von Mises, nous prenons

$$T = \int\int \left(\widehat{C}_{m,n}(u, v) - uv\right)^2 \, du \, dv,$$

ou pour un test de Kolmogorov-Smirnov

$$T = \sup_{(u,v)\in[0,1]^2} \left|\widehat{C}_{m,n}(u, v) - uv\right|.$$

Bibliography

[1] Babu, G.J., A.J. Canty et Y.P. Chaubey. *Application of Bernstein Polynomials for smooth estimation of a distribution and density function.* journal statistical planning and inference 105 : 377−392. (2002).

[2] Babu, G. J. et Chaubey, Y. P. *Smooth estimation of a distribution function and density function on a hypercube using Bernstein polynomials for dependent random vectors.* Statistics and Probability Letters, 76 : 959−969. (2006).

[3] Belalia, M. *On the asymptotic properties of the Bernstein estimator of the multivariate distribution function.* Statistics and Probability Letters, 110 : 249−259. (2016).

[4] Bernstein, S. *Démonstration du théorème de Weierstrass fondée sur le calcul des probabilités.* Communications de la Société mathématique de Kharkow. 2-éme série, 13 : 1−2. (2012).

[5] Blum, J. R. and Susarla, V. *Maximum deviation theory of density and failure rate function estimates based on censored data.* journal of multivariate analysis, 5 : 213−222. (1980).

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

(1974).

[7] Cai, Z. and Roussas, G. G. *Uniform strong estimation under α-mixing, with rates.* Statistics and Probability Letters, 15 : 47−55. (1992).

[8] Deheuvels, P. *Non Parametric Tests for Independence.* Statistique non paramétrique asymptotique. Springer, Berlin, Heidelberg : 95−107. (1980).

[9] Denuit, M., Purcaru, O., and Van Keilegom, I. *Bivariate archimedean copula models for censored data in non-life insurance.* (2006).

[10] Dib, K., Bouezmarni, T., Belalia, M., et Kitouni, A. *Nonparametric bivariate distribution estimation using Bernstein polynomials under right censoring.* Communications in Statistics - Theory and Methods 50(23) : 5574−5584. (2021).

[11] Efron, B. *The Two Sample Problem with Censored Data.* In Proceedings of the Fifth Berkeley Symposium On Mathematical Statistics and Probability. NewYork, Prentice-Hall 4 : 831−853. (1967).

[12] Einmahl, J.H.J. *Multivariate Empirical Processes.* Technical report, CWI Tract 32, Amsterdam (1987).

[13] Fermanian, J. D., Radulovic, D., et Wegkamp, M. *Weak convergence of empirical copula processes.* Bernoulli, 10(5) : 847−860. (2004).

[14] Finkelstein D. M. et Wolfe, R. A., *A semiparametric model for regression analysis of interval-censored failure time data.* Biometrics 41 : 933−945. (1985).

[15] Fleming T. R., O'Fallon J. R. and O'Brien C. P., *Modified Kolmogorov-Smirnov Test Procedures with application to arbitrarily right censored data.* Biometrics 36 : 607−626. (1980)

[16] Földes A. et Rejtó L. : *A LIL type result for the product limit estimator.* Probability Theory and Related Fields 56 (1) : 75−86, (1981).

[17] Földes A., Rejtó L. et Winter, B. B. : *Strong consistency properties of nonparametric estimator for randomly censored data, LI : Estimation of density and failure rate.* Periodica Mathematica Hungarica 12 : 15−19, (1981).

[18] Genest, C., and MacKay, J.R. *Copules archimédiennes et families de lois bidimensionnelles dont les marges sont données.* Canadian journal of statistics 14.2 : 145−159. (1986).

[19] Gribkova, S. et Lopez, O. *Nonparametric copula estimation under bivariate censoring.* Scandinavian Journal of Statistics 42 : 925−946, (2015).

[20] Janssen, P., Swanepoel, J., and Veraverbeke, N. *Large Sample Behavior of the Bernstein Copula Estimator.* Journal of Statistical Planing and Inference, 142 : 1189−1197. (2012).

[21] Joe, H., *Dependence modeling with copulas.* CRC press (2014).

[22] Kaplan E.L. et Meier P., *Nonparametric estimation from incomplete observations.* Jasa, 457−481. (1958).

[23] Klein J. P., *Small-Sample Moments of Some Estimators of the Variance of the Kaplan-Meier and Nelson-Aalen Estimators.* Scandinavian Journal of Statistics 18 : 333−340. (1991).

[24] Klein, J. P. and M. L. Moeschberger. *Survival analysis. Techniques forcensored and truncated data.* 2nd ed. Statistics for Biology and Health. New York, NY :Springer. (2003).

[25] Leblanc, A. *Chung-Smirnov property for Bernstein estimators of distribution functions.* Journal of Nonparametric Statitistics 21 : 133−142. (2009).

[26] Leblanc, A. *On estimating distribution functions using Bernstein polynomials.* annals of the institute of statistical mathematics 64 : 919−943. (2012a).

[27] Lee, Elisa T. and J. Wang., *Statistical methods for survival data analysis.*. Vol. 476. John Wiley and Sons, (2003).

[28] Lehmann E.L., *Elements of Large-Sample Theory.* Springer, New York (1999).

[29] Lo, S.H. and Singh, K. *The product-limit estimator and the bootstrap : Some asymptotic representations.* probability theory related fields, 71 : 455−465. (1986).

[30] Lorentz, G.G. *Bernstein Polynomials.* 2nd ed., Chelsea Publishing, New York. (1986).

[31] Lopez, O. et Saint-Pierre, P. *Bivariate censored regression relying on a new estimator of the joint distribution function.*. Journal of Statistical Planning and Inference 142 (8) : 2440−2453. (2012).

[32] Major, P. et Rerjto, L. *Strong embedding of the estimator of the distribution function under random censorship.* The Annals of Statistics, 16 : 1113−1132. (1988).

[33] Marron, J. S. et Padgett, W. J., *Asymptotically optimal bandwidth selection for kernel density estimators from randomly right-censored samples.* The Annals of Statistics, 15 : 1520−1535. (1987)

[34] Mantel N., *Ranking procedures for arbitrarily restricted observations.* Biometrics 23 : 65−78. (1967).

[35] Masala, G. B., and Micocci, M. *Loss-Alae modeling through a copula dependence structure structure.* Investment Management and Financial Innovations, Volume 6, Issue 4 : 67−80 (2009).

[36] Nelson R. B., *An Introduction to Copulas.* Springer Series in Statistics. USA, (2006).

[37] Nelson W., *Theory and Applications of Hazard Plotting for Censored Failure Data.* Technometrics 14 : 945−965. (1972).

[38] Rabhi, Y. and Bouezmarni, T. *Nonparametric inference for copula density function under random censoring.* Technical Report, Department of Mathematics, University of Sherbrooke, (2016).

[39] Rényi, A., *On measures of dependence.* Acta mathematica hungarica 10.3-4 : 441−451. (1959).

[40] Satten, Glen A., and S. Datta. *The Kaplan–Meier estimator as an inverse-probability-of-censoring weighted average..* The American Statistician 55(3) : 207−210. (2001)

[41] Schweizer, B. et Wolff E. F., *On Nonparametric Measures of Dependence for Random Variables.* The Annals of Statistics, 9 : 879−885. (1981)

[42] Segers, J. *Asymptotics of empirical copula processes under non-restrictive smoothness assumptions.* Bernoulli, 18(3) : 764−782. (2012).

[43] Stute, W. *Consistent estimation under random censorship when covariables are present.* J. Multivariate Anal., 45(1) :89−103. (1993).

[44] Stute, W. *Distributional convergence under random censorship when covariables are present..*Scandinavian Journal of Statistics, 23 :461−471. (1996).

[45] Sun J., *Interval Censoring.* Encyclopedia of Biostatistics : 2090−2095. John Wiley, New York. (1998).

[46] Silverman, B. W. *Density Estimation for Statistics and Data Analysis.* Chapman and Hall, London. (1986).

[47] Tenbusch, A. *Two-dimensional Bernstein polynomial density estimation.* Metrika 41 : 233−253. (1994).

[48] Van Der Vaart, A. W. et Wellner, J. A. *Empirical processes indexed by estimated functions*. Lecture Notes-Monograph Series : 234−252. (2007)

[49] Vitale, R.A. *A Bernstein polynomial approach to density estimation.* Statistical Inference and Related Topics 2 : 87−100. (1975).

[50] Wellner, Jon. *Weak convergence and empirical processes : with applications to statistics*. Springer Science and Business Media. (2013).

[51] Wand, M. P. and Jones, M. C. *Kernel Smoothing*. Chapman and Hall, London. (1986).

Estimation Non-Paramétrique de La Fonction de Répartition Bivariée : Approche de Bernstein

Résumé *(en français)*
Dans plusieurs domaines d'applications de la statistique, l'estimation de la fonction de répartition bivariée joue un rôle essentiel. Nous proposons un tel estimateur en se basant sur l'approche de Bernstein en présence de données censurées à droite. Les résultats du comportement asymptotiques de notre estimateur ont été établis, la convergence presque sure, le biais ainsi que la normalité asymptotique. nous avons tiré un estimateur non-paramétrique de la copule a partir de notre estimateur et nous avons montré le résultat de sa convergence. Finalement nous avons illustrée la robustesse de notre contribution à travers des données simulées et données réelles.
Mots-clés
Fonction de répartition, Estimation, non-paramétrique, polynômes de Bernstein, censure, copules.

Nonparametric bivariate distribution estimation using Bernstein polynomials under right censoring

Abstract *(in english)*
In several areas of statistical application, the estimation of the bivariate distribution function plays an essential role, We propose such estimator based on Bernstein's approach in the presence of right censored data.The results of the asymptotic behavior of our estimator have been established, Almost sure convergence, bias as well as asymptotic normality. We derived a nonparametric estimator of the copula from our estimator and showed the result of its convergence. Finally, we illustrated the robustness of our contribution through simulated and real data.
Keywords
Distribution function, right censoreing, Bernstein polynomials, non-parametric, copula.

العنوان : تقدير لا وسيطي لدالة التوزيع ثنائية المتغير: مقاربة بيرنشتاين.

في العديد من الميادين التطبيقية للإحصاء نهتم بتقدير دالة التوزيع. عند دراسة متغيرين فان دالة التوزيع ثنائية البعد تلعب دورا أساسيا في نمذجة الظواهر. في هذا البحث نقدم تقديرا لدالة التوزيع ثنائية البعد باستخدام مقاربة بيرنشتاين، هذا التقدير يأخذ بعين الاعتبار نتائج الملاحظات المراقبة من اليمين.

تم تحديد نتائج السلوك المقارب لمقدرنا والتقارب شبه مؤكد. قمنا بإعطاء عبارة التحيز وبرهنا أن مقدرنا ينتهي في العينات الكبيرة الى أن يتبع متغيرا طبيعيا.

في مرحلة موالية، استخدمنا تقديرنا لدالة التوزيع في استخراج تقدير لدالة الارتباط ثنائية البعد. أعطينا برهانا على تقارب تقديرنا اللاوسيطي المستخرج.

أثبتنا النتائج التي تحصلنا عليها بواسطة دراسة محاكات بعدة نماذج على غرار دراسة بيانات حقيقية.

الكلمات الرئيسية:

دالة التوزيع ، التقدير ، غير الوسيطي، كثيرات حدود برنشتاين ، الرقابة، دالة الارتباط.