Contributions en inférence statistique en présence de censure multivariée

Svetlana Gribkova

Université Pierre et Marie Curie

École Doctorale de Sciences Mathématiques de Paris Centre

Thèse de doctorat

Discipline : Statistique

présentée par

Svetlana Gribkova

Contributions à l'inférence statistique en présence de censure multivariée

---

dirigée par Gérard Biau et Olivier Lopez

Soutenue le 29 septembre 2014 devant le jury composé de :

M. Patrice Bertail	Rapporteur	Université Paris Ouest Nanterre
M. Gérard Biau	Directeur	UPMC
M. Antoine Chambaz	Examinateur	Université Paris Ouest Nanterre
M. Wenceslao González Manteiga	Rapporteur	Université Santiago de Compostela
M^{me} Agathe Guilloux	Examinateur	UPMC
M. Olivier Lopez	Co-directeur	ENSAE
M. Valentin Patilea	Examinateur	ENSAI

Institut de Mathématiques de Jussieu
175, rue du chevaleret
75 013 Paris

UPMC
Ecole Doctorale de Sciences
Mathématiques de Paris Centre
4 place Jussieu
75252 Paris Cedex 05
Boite courrier 290

Remerciements

Je tiens à remercier de tout mon cœur Gérard Biau et Olivier Lopez qui ont guidé et dirigé mes recherches durant ces trois années de thèse et qui m'ont fait découvrir des sujets scientifiques fascinants. Je leur suis reconnaissante de leur présence, de leurs conseils fructueux, de leurs soutiens sans faille et de leurs encouragements amicaux.

Je n'aurais probablement pas eu ce parcours si Gérard ne m'avait pas encouragé à suivre le M2 de Statistique lorsque j'étais en deuxième année de l'École des Mines. Je le remercie donc tout d'abord pour avoir été à l'origine de cette thèse bien avant son commencement. Je le remercie ensuite pour la confiance qu'il m'a accordée en acceptant de la diriger. Enfin, je le remercie pour avoir toujours été à mon écoute en cas de difficultés, pour m'avoir fait profité de son expertise scientifique, pour ses suggestions et ses commentaires pertinents lors de la relecture de mes articles qui m'ont beaucoup aidé à les améliorer.

Je tiens à exprimer ma profonde gratitude à Olivier pour tout ce qu'il m'a appris, aussi bien sur le plan scientifique que sur le plan humain, pour toutes les heures de travail passionnantes que j'ai passées en "donskerisant" dans son bureau, dont les portes ont toujours été ouvertes pour moi. Je le remercie également pour ses conseils magiques qui transformaient d'un coup des problèmes insurmontables en simples questions techniques, pour la chance que j'ai eu de participer à des colloques, pour sa disponibilité, pour son implication, pour son amitié, son humour et son optimisme, qui m'ont aidé à surmonter les difficultés que j'ai rencontrées.

J'exprime toute ma reconnaissance aux rapporteurs de ma thèse, Patrice Bertail et Wenceslao González Manteiga, pour avoir lu et évalué ce manuscrit. Je remercie également Antoine Chambaz, Agathe Guilloux et Valentin Patilea pour avoir accepté de faire partie de mon jury.

Je remercie tout particulièrement Philippe Saint Pierre pour son cours de l'analyse de survie qui m'a donné envie de m'orienter vers ce domaine de recherche, pour le co-encadrement avec Olivier de mon stage de M2, ainsi que pour les moments que l'on partagés et pour ses conseils amicaux durant mes années de thèse.

Il m'est impossible d'oublier tous les maîtres de conférences et les professeurs du LSTA ainsi que les membres de l'équipe de statistique du LPMA avec qui j'ai eu la chance de discuter et d'échanger. Je remercie Michel Broniatowski pour sa bienveillance, pour ses conseils et pour les discussions intéressantes que nous avons eu. Je remercie Agathe, Fanny, Etienne, Nathalie, Tabea, Frederic, Jean-Patrick et Bertrand pour tous les moments conviviaux que l'on a partagés ensemble.

Je remercie également tous les amis et collègues doctorants que j'ai eu le plaisir de rencontrer. Un merci très particulier à Sarah, pour les conversations intéressantes sur les sujets mathématiques et d'autres, ainsi que pour son amitié, à Patricia pour son amitié, son soutien et son écoute, à Jean-Paul pour m'avoir aidé à corriger la majorité des fautes d'orthographe dans mon manuscrit, et pour m'avoir soutenu lors de la rédaction. Les mots de remerciement pour l'ambiance amicale et les échanges de toutes sortes vont à mes collègues du bureau Sumeya, Alexis, Salim, Nedjmaddin, à ma grande sœur de thèse Aurélie, à Cécile, Erwan, Benjamin, Baptiste, Mathieu, Assia, Roxanne, Tarn, Mokhtar, Boris.

Enfin, merci à toute ma famille pour son soutien, et tout particulièrement à ma mère qui a fait le même parcours du combattant, qui me comprend si bien et qui trouve toujours les mots justes pour m'encourager.

C'est une étape dans ma vie qui se termine avec la soutenance de thèse, ces trois années sont déjà couvertes pour moi d'un léger voile de nostalgie. C'est en même temps une nouvelle étape qui commence avec ses défis et ses aventures. Grâce à tout ce que j'ai appris et tous ceux que j'ai connus, je me sens à présent prête à les affronter.

Résumé

Résumé

L'objectif de cette thèse est d'explorer plusieurs approches pour l'étude des données censurées multivariées, à savoir l'estimation non paramétrique de la fonction de répartition jointe, la modélisation de dépendance par les modèles de copules et l'étude exploratoire par des méthodes de clustering.

Plan du manuscrit

Ce manuscrit est composé de quatre chapitres. Les Chapitres 2 à 4 sont indépendants et utilisent leurs propres notations.

Le **Chapitre 1** introduit le contexte général de cette thèse ainsi que ses contributions. Nous abordons d'abord quelques concepts liés à l'étude de données censurées : la variable de censure, les observations incomplètes et l'estimateur de Kaplan-Meier. Par la suite, nous nous concentrons sur la problématique de l'étude de relations entre les variables en présence de censure multivariée, et nous présentons les méthodes utilisées dans les chapitres suivants : l'estimation de la fonction de répartition jointe, la modélisation par les copules et le clustering. Une synthèse des résultats principaux de la thèse clôt cette première partie du manuscrit.

Le **Chapitre 2** est consacré à l'estimation de la distribution jointe des deux variables censurées dans le cadre d'un modèle de durée simplifié où la différence entre deux variables de censure est observée. Dans un premier temps, nous construisons un estimateur non paramétrique de la fonction de répartition jointe. Dans un deuxième temps, nous établissons la normalité asymptotique pour les intégrales par rapport à la mesure définie par cet estimateur. Notre approche consiste à établir une représentation de ces intégrales sous la forme d'une somme de termes indépendants et identiquement distribués et d'un reste asymptotiquement négligeable.

Nous considérons ensuite des applications, en l'occurrence l'estimation du tau de Kendall et la construction d'un test d'adéquation pour les modèles de copules, basé sur le processus de Kendall. Une étude numérique est effectuée sur des données simulées et sur des données réelles d'assurance.

Le **Chapitre 3** est dédié à la problématique de l'estimation non paramétrique de la copule bivariée, à partir d'un échantillon de données censurées. La copule est d'abord estimée par une fonction discrète qui peut être interprétée comme une extension de la copule empirique en présence de censure. La construction est basée sur

un estimateur de la fonction de répartition jointe d'une certaine forme générique, dont l'estimateur du Chapitre 2 est un cas particulier. Nous nous intéressons par la suite aux estimateurs lisses de la copule et de sa densité. Les deux approches proposées sont des adaptations respectives de deux méthodes de l'estimation par noyau proposées par [Fermanian et al., 2004] et par [Omelka et al., 2009]. La deuxième approche, basée sur la transformation des variables, permet de rendre la performance de l'estimateur de [Fermanian et al., 2004] moins influencée par les lois marginales.

Les propriétés asymptotiques sont étudiées en termes de théorème central limite fonctionnel pour les estimateurs de la fonction copule, et en termes de convergence uniforme sur les compacts strictement inclus dans $[0,1]^2$ pour les estimateurs de sa densité.

Nous présentons ensuite une procédure de test d'adéquation, basée sur nos estimateurs, pour les modèles de copules en présence de censure. Le chapitre se termine par une étude de données réelles et simulées.

Le **Chapitre 4** présente une approche exploratoire pour l'étude de données censurées. Plus précisement, nous considérons une configuration multivariée où une variable est une durée sujette à la censure, et toutes les autres variables sont observées. Nous étendons la méthode de quantification de la loi d'un vecteur aléatoire en présence de censure. Après avoir défini la distortion empirique dans le cadre censuré, nous définissons le quantificateur empirique optimal et étudions sa consistance. En particulier, nous fournissons une borne exponentielle non asymptotique pour la vitesse de convergence.

Nous proposons ensuite un algorithme de clustering pour les données censurées. Il est composé de deux étapes : la première évalue numériquement les centres de clusters, puis la seconde attribue des labels aux observations censurées.

Mots-clefs

Analyse de survie, estimation non paramétrique, censure, copule, processus empiriques, clustering, quantification.

Contributions to statistical inference in presence of multivariate censoring

Abstract

The main purpose of this thesis is to explore several approaches for studying multivariate censored data: nonparametric estimation of the joint distribution function, modeling dependence with copulas and k-clustering for the exploratory analysis.

Outline

This manuscript is composed of four chapters. Chapters 2 to 4 are independent and use their own notations.

Chapter 1 presents the general framework and the contributions of this thesis. We give a brief exposition of basic notions in survival analysis: censored variable, incomplete observations and Kaplan-Meier estimator. Then, we focus on the studying the dependence between random variables in presence of multivariate censoring and we provide an exposition of some relevant tools: estimation of the joint distribution function, modeling dependence with copulas and clustering. Eventually, we give a summary of our contributions.

Chapter 2 deals with the estimation of the joint distribution function of two censored variables in a simplified survival model in which the difference between two censoring variables is observed. We provide a new nonparametric estimator of the joint distribution function and we establish the asymptotic normality of the integrals with respect to its associated measure. The basic idea of our approach is to derive an asymptotic representation of these integrals as a sum of i.i.d. terms and of an asymptotically negligible reminder.

Then, we consider some applications: estimating Kendall's tau and constructing a new goodness-of-fit test for copula models, based on Kendall's process. The estimators are studied numerically both on simulated and real data sets.

Chapter 3 is devoted to nonparametric copula estimation under bivariate censoring. We provide a discrete and two smooth copula estimators along with two estimators of its density. The discrete estimator can be seen as an extension of the empirical copula under censoring. Its construction is based on a distribution function estimator having some generic form. In particular, it agrees with the estimator considered in Chapter 2. The two smooth copula estimators are the censored generalizations of two kernel estimators proposed respectively by [Fermanian et al., 2004] and [Omelka et al., 2009]. The second approach is based on a certain transformation of the initial variables allowing to make the performance of the estimator of [Fermanian et al., 2004] be less affected by the marginal distributions.

We show the weak convergence in $l^\infty[0,1]^2$ of empirical processes associated with the copula estimators and we derive the uniform consistency of copula density estimators on the compact sets strictly included in $[0,1]^2$.

The practical behavior of our estimators is investigated through a simulation study and two real data applications, corresponding to different censoring settings. We use our estimators to define a goodness-of-fit procedure for parametric copula models. A new bootstrap scheme is proposed to compute the critical values.

Chapter 4 provides a new exploratory approach for censored data analysis. We consider a multivariate configuration with one variable subjected to censoring and the others completely observed. We extend the probabilistic k-quantization method in the case of random vector with one censored component. The definitions of the empirical distortion and of empirically optimal quantizer are generalized in presence of one-dimensional censoring. We study the asymptotic properties of the distortion of the empirically optimal quantizer and we provide a non-asymptotic exponential bound for the rate of convergence.

Our results are then applied to construct a new two-step clustering algorithm for censored data.

Keywords

Survival analysis, nonparametric estimation, censoring, copula, empirical processes, clustering, quantization.

Table des matières

Chapitre 1

État de l'art et contributions

1.1 Introduction générale

Cette thèse est consacrée à l'étude statistique de la distribution de plusieurs variables aléatoires en présence de censure. Le phénomène de censure est lié aux événements perturbateurs qui peuvent se produire dans le laps de temps nécessaire au recueil d'une donnée. Il intervient donc fréquemment lors de mesures qui portent sur les variables modélisant le temps écoulé entre deux événements : durée de vie d'un individu, durée entre le début d'une maladie et la guérison, durée d'un épisode de chômage, etc.

Ces perturbations empêchent l'observateur d'accéder à la totalité de l'information concernant le phénomène qu'il étudie et conduisent à l'apparition d'observations incomplètes dites censurées. Du point de vue de l'inférence statistique, le traitement des données censurées nécessite de revoir les méthodes classiques dont le point de depart commun est un échantillon des réplications indépendantes identiquement distribuées (i.i.d.) des variables d'intérêt. En effet, en présence de censure, une variable d'intérêt n'est plus observée directement, et un tel échantillon sur lequel travailler est indisponible.

Dans de nombreuses applications, la durée n'est pas étudiée de façon isolée. Dans certaines situations, on cherche à comprendre sa distribution à l'aide d'un modèle statistique avec des variables explicatives (le modèle de régression). Par exemple, en médecine, on cherche à expliquer la durée de vie des patients à partir de leurs caractéristiques. Dans d'autres cas, on s'intéresse à la structure des interactions entre la durée et d'autres variables décrivant le phénomène étudié, ces variables pouvant elles-mêmes être des durées sujettes à censure. On peut évoquer les études de dépendance entre les durées de vie de jumeaux en génétique, ou entre celles de conjoints en assurance vie.

Dans les situations mentionnées ci-dessus (et dans de nombreuses autres), le statisticien est confronté au problème de l'étude de la distribution de plusieurs variables aléatoires, l'une au moins d'entre elles étant perturbée par la censure. Ce problème est inévitable car négliger les observations censurées revient à exclure de l'échantillon les durées les plus grandes, qui sont les plus concernées par la censure. Par ailleurs, il est crucial dans un grand nombre d'applications pour lesquelles le pourcentage des observations censurées peut atteindre les valeurs importantes.

Dans cette thèse, nous explorons, en présence de censure, trois approches pour l'étude de la loi jointe de variables aléatoires, à savoir l'estimation de leur fonction de répartition, la modélisation de leur structure de dépendance à l'aide des copules et le clustering.

Le reste de ce chapitre est organisé de la façon suivante. Dans les Sections 1.2 à 1.5, nous introduisons les trois domaines de la statistique correspondant à nos trois approches. La Section 1.6 présente une synthèse des principaux résultats.

1.2 Concepts de base en analyse de survie

Dans cette section, nous introduisons la notion de variable censurée et la méthode non paramétrique de Kaplan-Meier de l'estimation de sa distribution. Par la suite, nous utiliserons le terme conventionnel "durée" pour désigner la variable sujette à la censure. Néanmoins, il faut noter que l'ensemble des techniques que nous allons considérer dans cette thèse peuvent être appliquées aux variables censurées n'ayant pas de caractère temporel (comme, par exemple, les montants de sinistres).

1.2.1 Les observations censurées

La censure aléatoire est l'un des phénomènes les plus fréquents à l'origine de données incomplètes en statistique. Un cas typique de son apparition est le suivant. Lorsque l'on étudie la durée de temps écoulé jusqu'à l'apparition d'un certain événement, pour une population d'objets où des individus, il peut arriver que certains d'entre eux cessent d'être observé, alors que l'événement ne s'est pas encore produit. Pour ces individus, on disposera seulement d'une partie de l'information : la durée qui nous intéresse n'est pas observée mais on sait sûrement qu'elle est supérieure à la période de temps délimitée par le début de l'observation et l'instant auquel l'individu a cessé d'être observé. La définition formelle pour la variable T sujette à ce type de perturbations est le suivant.

La variable aléatoire T (la durée) est dite **censurée à droite** par la censure aléatoire C si au lieu d'observer T, on observe

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

Un échantillon résultant de mesures portant sur T est dit de données censurées. Il est composé de réplications i.i.d. $(Y_i, \delta_i)_{1 \leq i \leq n}$ du vecteur aléatoire (Y, δ).

Avant de s'intéresser au problème de l'inférence statistique à partir d'un échantillon de données censurées, nous allons d'abord évoquer quelques exemples classiques de son apparition.

- **En essais cliniques ou en assurance vie,** on étudie la durée de survie T des individus. La variable de censure C représente dans ce cas la date à laquelle un individu quitte l'étude pour une cause autre que le décès.
- **En fiabilité,** l'événement d'intérêt est la défaillance d'un système. Ainsi, T est la durée de son fonctionnement correct. Si au bout d'un certain temps C, le système cesse d'être utilisé, alors qu'il est encore en état de marche, T est censurée par C.

- **En assurance dommages,** on peut s'intéresser au temps T entre deux accidents subis par un conducteur. S'il annule son contrat d'assurance au bout d'un temps C, alors que le deuxième accident ne s'est pas encore produit, C est la variable de censure pour T.

Comme on l'a déjà mentionné, l'information incomplète apportée par un échantillon de données censurées ne peut pas être traitée par les procédures statistiques classiques. En particulier, la fonction de répartition empirique est inadaptée pour l'estimation de la distribution d'une durée censurée. En effet, cet estimateur s'écrit sous la forme

$$F_n^{emp}(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{T_i \leq t},$$

où $(T_i)_{1\leq i\leq n}$ sont les réplications i.i.d. de T. En présence de censure, T n'est plus la variables observée mais la durée sous-jacente. Par conséquent, $F_n^{emp}(t)$ fait intervenir des quantités non observées.

Sa généralisation pour les données censurées a été proposé par Kaplan et Meier (voir [Kaplan and Meier, 1958]) qui ont défini un estimateur consistant de F à partir de l'échantillon $(Y_i, \delta_i)_{1\leq i\leq n}$.

1.2.2 Estimateur de Kaplan-Meier

Dans cette section, nous allons considérer l'estimateur de Kaplan-Meier et ses propriétés. Dans toute la suite, on va supposer que :

- La durée T et sa censure C sont indépendantes.
- $\mathbb{P}(T = C) = 0$.

La première hypothèse rend le modèle identifiable, et la deuxième garantie la symétrie entre les variables T et C. Elles sont essentielles pour la consistance de l'estimateur que l'on va définir.

Construction de l'estimateur

Dans cette section, on cherche à construire un estimateur non paramétrique de la fonction de répartition $F(t)$ d'une durée T à partir d'un échantillon $(Y_i, \delta_i)_{1\leq i\leq n}$.

Soient $Y_{(1)}, \ldots, Y_{(n)}$ les statistiques d'ordre de l'échantillon $(Y_i)_{1\leq i\leq n}$, que l'on va supposer distinctes (dans le cas contraire, il est convenu de briser les coïncidences de manière aléatoire). On peut écrire :

$$\mathbb{P}(T > Y_{(i)}) = \prod_{j=1}^{i} \mathbb{P}(T > Y_{(j)} | T > Y_{(j-1)}) = \prod_{j=1}^{i} \left(1 - \mathbb{P}(T \leq Y_{(j)} | T > Y_{(j-1)})\right),$$

avec $Y_{(0)} = 0$. Lorsqu'on interprète T comme une durée, $\mathbb{P}(T \leq Y_{(j)} | T > Y_{(j-1)})$ est la probabilité conditionnelle que l'évènement survienne dans l'intervalle $]Y_{(j-1)}, Y_{(j)}]$, sachant qu'il ne s'est pas encore produit au moment $Y_{(j-1)}$. Cette probabilité peut être estimée par zéro si $\delta_j = 0$, et par $1/\sum_{k=1}^{n} \mathbb{1}_{Y_k \geq Y_{(j)}}$ sinon.

Pour tout $t \in [Y_{(i-1)}, Y_{(i)}[$, il n'y a pas de nouvelle information dans l'intervalle $[Y_{(i-1)}, t]$, par rapport à celle disponible à l'instant $Y_{(i-1)}$. Il est donc naturel d'estimer $\mathbb{P}(T > t)$ par la même valeur que $\mathbb{P}(T > Y_{(i-1)})$. En rassemblant ces arguments, on

arrive à la définition suivante de l'estimateur, dit de Kaplan-Meier, de la fonction de répartition F :

$$1 - F_n^{KM}(t) = \prod_{Y_{(i)} \leq t} \left(1 - \frac{1}{\sum_{j=1}^{n} \mathbb{1}_{Y_{(j)} \geq Y_{(i)}}}\right)^{\delta_i}. \tag{1.1}$$

L'expression (1.1) définit une fonction monotone, continue à droite avec limite à gauche. Les points de discontinuité de cette fonction correspondent aux observations non censurées.

Une autre façon d'écrire l'estimateur de Kaplan-Meier, utilisée par [Satten and Datta, 2001], permet de l'exprimer sous la forme similaire à celle de la fonction de répartition empirique dans le cadre non censuré :

$$F_n^{KM}(t) = \frac{1}{n} \sum_{i=1}^{n} W_{in} \mathbb{1}_{Y_i \leq t}, \quad \text{avec} \quad W_{in} = \frac{\delta_i}{1 - G_n^{KM}(Y_i-)}, \tag{1.2}$$

où $G_n^{KM}(y)$ est l'estimateur de Kaplan-Meier de la fonction de répartition $G(y)$ de la censure C, et $G(y-)$ désigne la limite à gauche de G au point y.

Propriétés de l'estimateur de Kaplan-Meier

La mesure sous-jacente. La mesure définie par $F_n^{KM}(t)$ attribue des poids uniquement aux observations non censurées. De plus, d'après (1.2), les grandes observations ont les poids plus importants, ce qui permet de compenser leur déficit dans l'échantillon dû à la présence de censure.

Contrairement à la mesure empirique, la masse totale de la mesure associée à $F_n^{KM}(t)$ peut être strictement inférieure à 1. Cela se produit précisément dans le cas où la plus grande observation est censurée.

Intégrales Kaplan-Meier. De nombreuses quantités à estimer, liées à la durée étudiée, se présentent sous la forme

$$\mathbb{E}[\phi(T)] = \int \phi(t) dF(t),$$

où ϕ est une fonction d'espérance finie. En absence de censure, leurs estimateurs sont les intégrales de ϕ par rapport à la mesure empirique. Les propriétés asymptotiques de ces intégrales découlent directement du théorème central limite et de la loi forte de grands nombres. En présence de censure, les quantités $\mathbb{E}[\phi(T)]$ sont estimées par les intégrales par rapport à la mesure sous-jacente associée à l'estimateur de Kaplan-Meier :

$$\int \phi(y) dF_n^{KM}(y) = \frac{1}{n} \sum_{i=1}^{n} \frac{\delta_i \phi(Y_i)}{1 - G_n^{KM}(Y_i-)}. \tag{1.3}$$

L'étude du comportement asymptotique des intégrales (1.3) est sensiblement plus complexe que dans le cadre non censuré, puisque les termes de la somme ne sont pas i.i.d.

Propriétés asymptotiques. Dans la littérature de l'analyse de survie, il existe deux approches à l'analyse asymptotique de l'estimateur de Kaplan-Meier. La première, basée sur les techniques de martingales, a été employée dans [Gill, 1980] et

[Gill, 1983]. Elle permet de démontrer la normalité asymptotique de l'estimateur de Kaplan-Meier et de certains de ses fonctionnelles, sans pour autant fournir le théorème central limite pour ses intégrales.

Une approche alternative, à laquelle on aura recours à plusieurs reprises, est fondée sur les représentations des intégrales (1.3) sous la forme d'une somme de termes i.i.d. d'espérance nulle et d'un reste asymptotiquement négligeable :

$$\int \phi(y) dF_n^{KM}(y) = \frac{1}{n} \sum_{i=1}^{n} \psi_\phi(Y_i, \delta_i) + R_n(\phi),$$

où $\mathbb{E}[\psi_\phi(Y, \delta)] = 0$ et $\sup_\phi R_n(\phi) = o(n^{-1/2})$. En utilisant cette technique, [Stute and Wang, 1993] démontrent la loi forte pour les intégrales de Kaplan-Meier, et [Stute, 1995] fournit le théorème central limite. Pour d'autres utilisations des représentations i.i.d. dans la littérature, le lecteur pourra consulter [Stute, 1999], [Stute et al., 2000], [Sánchez Sellero et al., 2005] et [Gannoun et al., 2005].

1.3 Analyse de survie multivariée

Dans cette section, nous entrons au cœur du sujet de la thèse. Nous présentons d'abord deux cadres importants où la censure intervient dans un contexte multivarié : le modèle de régression et la dépendance entre deux durées. Dans la première situation, la censure affecte une seule variable, alors que dans la deuxième, les deux variables peuvent y être sujettes. Nous discuterons ensuite d'une façon plus détaillée le deuxième cadre, en s'arrêtant sur les difficultés principales d'un point de vue de l'inférence statistique et sur plusieurs modèles existant dans la littérature.

1.3.1 Deux contextes multivariés

Comme nous l'avons déjà mentionné, tous les individus ne sont pas identiques et deux personnes avec des caractéristiques différentes n'aurons pas la même espérance de vie. Une façon de tenir compte de cette hétérogénéité consiste à caractériser les individus par des variables explicatives. Généralement, elles n'ont pas de caractère temporel et sont entièrement observées.

- **Contexte 1. Modèle multivarié où une seule variable est censurée.** Soit T une variable aléatoire (la durée) censurée à droite par la censure C. De plus, on considère un vecteur aléatoire X (de variables explicatives) dont aucune composante n'est affectée par la censure. Le vecteur observé prend alors la forme (Y, δ, X), où

$$\begin{cases} Y = \min(T, C), \\ \delta = \mathbb{1}_{T \leq C}. \end{cases} \tag{1.4}$$

  Un échantillon est composé des répliques i.i.d. $(Y_i, \delta_i, X_i)_{1 \leq i \leq n}$.

Pour étudier les relations entre les variables dans ce contexte, on peut envisager plusieurs approches. Par exemple, [Stute, 1993] a considéré l'estimation de la fonction de répartition jointe de T et X. Il a généralisé l'estimateur de Kaplan-Meier en

présence de covariables par un estimateur qui s'écrit sous la forme

$$F_n^S(t,x) = \frac{1}{n}\sum_{i=1}^{n} W_{in}\mathbb{1}_{Y_i\leq t, X_i\leq x}, \tag{1.5}$$

où W_{in} désigne le poids attribué à l'observation Y_i par l'estimateur de Kaplan-Meier de la fonction de répartition de T, basé sur l'échantillon $(Y_i,\delta_i)_{1\leq i\leq n}$.

Dans le cas de l'estimateur de Kaplan-Meier, l'hypothèse d'identifiabilité est l'indépendance entre T et C. Une façon simple de la généraliser en présence de covariables est de supposer l'indépendance entre (T,X) et C. En fait, [Stute, 1993] montre que l'estimateur (1.5) est consistant sous les hypothèses plus générales :

- T et C sont indépendants,
- $\mathbb{P}(T\leq C|T,X) = \mathbb{P}(T\leq C|T)$.

Une autre approche consiste à utiliser un modèle de régression (paramétrique tel que "Accelerated Failure Time", semi-paramétrique tel que le modèle de Cox où entièrement non paramétrique).

Lorsque la population hétérogène, que l'on étudie, est composée de plusieurs groupes homogènes, la modélisation est plus précise si elle est faite séparément pour chacun de ces groupes. Se pose ici une question naturelle de leur détection en présence de censure. Ce problème sera considéré dans le Chapitre 4.

Nous allons maintenant considérer un autre cas important d'apparition de censure dans des données multivariées. Il s'agit de la dépendance entre deux durées censurées. Cette situation correspond à de nombreuses applications réelles traitées dans la littérature. Par exemple, [Hougaard et al., 1992] ont considéré la dépendance entre les durées de vie des jumeaux pour étudier l'influence d'un facteur génétique, [Luciano et al., 2008] se sont penchés sur le problème de dépendance de durées de vie des conjoints en assurance, [Denuit and Van Keilegom, 2006] ont étudié la corrélation positive entre les coûts de sinistres, etc. Nous allons d'abord décrire le cadre général.

- **Contexte 2. Modèle à deux variables censurées.**
  Soit T_1 (resp. T_2) une variable aléatoire de fonction de répartition $F_1(t_1)$ (resp. $F_2(t_2)$), censurée par une censure aléatoire C_1 (resp. C_2). Dans la suite, on notera $(Y_1,Y_2,\delta_1,\delta_2)$ le vecteur observé, où

$$\begin{cases} Y_j = \min(T_j,C_j) \\ \delta_j = \mathbb{1}_{T_j\leq C_j} \end{cases} \qquad \text{pour } j=1,2.$$

  Un échantillon des observations se compose de $(Y_{1i},Y_{2i},\delta_{1i},\delta_{2i})_{1\leq i\leq n}$.

D'un point de vue de l'inférence statistique, l'estimation de la répartition jointe de deux variables censurées est significativement plus complexe par rapport à la situation univariée. Contrairement à la fonction de répartition empirique, l'estimateur de Kaplan-Meier ne se généralise pas directement au cadre multivarié. Les propriétés désirables d'un estimateur de distribution telles que les masses positives attribuées à

toutes les observations et la convergence à la vitesse $n^{-1/2}$, perçues comme naturelles en absence de censure et vérifiées également par l'estimateur de Kaplan-Meier, sont difficiles à satisfaire sous censure bivariée.

Dans ce contexte, il peut y avoir deux chemins à suivre. Le premier consiste à travailler sous les hypothèses les plus générales possibles assurant l'identifiabilité du modèle, à savoir (T_1, T_2) indépendant de (C_1, C_2). Les estimateurs de ce type existants dans la littérature ne parviennent pas pour autant à satisfaire les deux propriétés citées auparavant. Par exemple, l'un des estimateurs les plus utilisés, dû à [Dabrowska, 1988], est $n^{-1/2}$-consistant mais assigne des masses négatives à certains points du plan et ne vérifie donc pas la définition de la fonction de répartition bivariée (voir [Pruitt, 1991b]). Le même défaut caractérise l'estimateur défini par [Campbell and Földes, 1982].

L'approche de [Tsai et al., 1986], basée sur l'estimateur de la fonction de répartition conditionnelle introduite dans [Beran, 1981], permet de construire un estimateur consistant avec des poids positifs, mais sa vitesse de convergence est inférieure à $n^{-1/2}$. Le même inconvenient concerne d'autres estimateurs tels que [Pruitt, 1991a], [van der Laan, 1996] et [Akritas and Van Keilegom, 2003].

Le deuxième chemin envisageable consiste à introduire des hypothèses sur le mécanisme de censure en modélisant le lien entre C_1 et C_2. Ces modèles sont certainement moins généraux, mais permettent en revanche de définir des estimateurs convergents à la vitesse $n^{-1/2}$ et n'attribuant à aucun point du plan de poids négatif.

1.3.2 Forme générique de l'estimateur bivarié

Les modèles évoqués à la fin de la section précédente peuvent généralement être écrits sous la forme générique suivante :

$$F_n(t_1, t_2) = \frac{1}{n} \sum_{i=1}^{n} \delta_{1i}\delta_{2i} W_n(Y_{1i}, Y_{2i}) \mathbb{1}_{Y_{1i} \leq t_1, Y_{2i} \leq t_2}, \tag{1.6}$$

où $W_n(y_1, y_2)$ est une certaine fonction aléatoire à valeurs positives, construite à partir de l'échantillon. Sa forme exacte se détermine selon l'hypothèse faite sur le mécanisme de censure. Un tel estimateur attribue des poids uniquement aux observations doublement non censurées qui sont les seules qui contiennent une information complète sur le phénomène.

Pour déterminer la fonction de poids W_n, il convient de raisonner de la façon suivante. En absence de censure, toute quantité du type $\mathbb{E}(\phi(T_1, T_2))$ peut être estimée par une intégrale de fonction $\phi(t_1, t_2)$ par rapport à la mesure empirique. Pour conserver la même propriété en présence de censure, on souhaiterait que les quantités du type

$$\int \phi(t_1, t_2) dF_n(t_1, t_2) = \frac{1}{n} \sum_{i=1}^{n} \delta_{1i}\delta_{2i} W_n(Y_{1i}, Y_{2i}) \phi(Y_{1i}, Y_{2i}). \tag{1.7}$$

convergent vers $\mathbb{E}(\phi(T_1, T_2))$. Pour définir une fonction W_n qui assure cette propriété, une piste simple consiste à chercher une fonction déterministe $W^*(y_1, y_2)$ qui satisfait, pour toute fonction ϕ d'espérance finie,

$$\mathbb{E}(\phi(T_1, T_2)) = \mathbb{E}(\delta_1\delta_2 W^*(Y_1, Y_2)\phi(Y_1, Y_2)).$$

Une telle fonction W^* dépend généralement des distributions inconnues des variables du modèle. Elle ne peut donc pas être utilisée directement pour la construction de $F_n(t_1, t_2)$. En revanche, elle peut être remplacée par une version empirique. La fonction aléatoire W_n dans la forme générique est alors construite comme un estimateur de W^*. Nous allons maintenant montrer des réalisations de cette construction sur plusieurs exemples.

Avant de considérer le cadre où plusieurs variables sont censurées, revenons sur le cas de notre premier contexte, où une seule variable est censurée. Il s'agit d'un cadre très fréquemment utilisé dans les problèmes de régression sous censure (voir, par exemple, [Cao and González-Manteiga, 2008]) :

Modèle 1. La censure affecte uniquement une variable.
On considère une variable censurée T_1 et une variable observée T_2 ($\delta_2 = 1$ p.s.), sous les hypothèses

- T_1 est indépendant de C_1
- $\mathbb{P}(T_1 \leq C_1 | T_1, T_2) = \mathbb{P}(T_1 \leq C_1 | T_1)$

Les deux hypothèses d'identifiabilité permettent de montrer que, pour tout $\phi(t_1, t_2)$,

$$\mathbb{E}(\phi(T_1, T_2)) = \mathbb{E}\left[\frac{\delta_1 \delta_1}{1 - G_1(Y_1-)}\phi(Y_1, Y_2)\right],$$

où G_1 représente la fontion de répartition de la censure C_1. Cela nous amène à envisager une fontion $W^*(y_1) = (1 - G_1(y_1-))^{-1}$ qui ne dépend dans ce cas que de y_1. Son estimateur naturel est

$$W_n(y_1) = \frac{1}{1 - G_{1n}(y_1)},$$

où G_{1n} est l'estimateur de Kaplan-Meier de la fonction de répartition de la censure. On retrouve sans difficulté les poids de l'estimateur (1.5) de Stute de notre contexte 1, dont ce modèle est un cas particulier.

"Loss adjustment expenses" en assurance. La survenue d'un sinistre engendre deux types des frais pour la compagnie d'assurance : le montant à rembourser à l'assuré et le coût d'installation du sinistre. Le premier est appelé *loss* et il est connu de suite. Le deuxième est dit *ALAE* et il se détermine seulement à la cloture du sinistre. Généralement, les sinistres coûteux sont traités pendant plus longtemps et nécessitent d'être étudiés davantage, ils engendrent donc les frais plus élevés. Par conséquent, il existe une dépendance positive entre *loss* et *ALAE*. A l'arrivée d'un sinistre, la compagnie d'assurance doit faire une provision pour son *ALAE*. L'estimation de la structure de dépendance entre *loss* et *ALAE* lui permet d'optimiser le montant de la provision. Cette estimation peut être faite à partir des données sur les coûts des sinistres passés et leurs *ALAE* (les données *loss-ALAE*). Puisque le montant de remboursement est généralement limité par un plafond spécifique à chaque police, de nombreuses bases de données de ce type sont incomplètes. En effet,

la pratique assez courante consiste à enregistrer uniquement la somme remboursée et non pas le coût réel d'un sinistre. Par conséquent, la variable *loss* est censurée à droite. Quant au montant de *ALAE*, il est entièrement à la charge de la compagnie d'assurance et est donc toujours enregistré. Puisque les sinistres censurés sont les plus coûteux, ils ne peuvent pas être négligés sans créer un risque de biais de l'estimation. Il est donc nécessaire d'utiliser les techniques adaptées développées dans le cadre du Modèle 2.

Dans le cas d'une censure multivariée (i.e. quand plusieurs variables sont censurées), plusieurs configurations existent suivant les hypothèses faites sur le mécanisme de censure. La situation la plus simple se produit lorsque le phénomène se réduit à une censure commune à T_1 et T_2, ce qui est le cas du Modèle 2 ci-dessous, qui généralise le modèle de [Lin and Ying, 1993].

Modèle 2. Les censures diffèrent par une variable auxiliaire.
On considère un vecteur aléatoire (T_1, T_2) avec les composantes censurées par (C_1, C_2), et une variable $\varepsilon = C_2 - C_1$. On suppose que :

- La variable ε est observée
- (T_1, T_2) est indépendant de ε et de C_1
- C_1 est indépendant de ε

Puisque T_1 est censurée par C_1 et T_2 est censurée par $C_1+\varepsilon$, les variables T_1 et $T_2-\varepsilon$ sont censurées par la même censure C_1. La configuration de [Lin and Ying, 1993] est donc un cas particulier du Modèle 2 qui se réalise quand $\varepsilon = 0$ p.s.

Les observations dans le Modèle 2 se composent des réplications i.i.d.

$$(Y_{1i}, Y_{2i}, \delta_{1i}, \delta_{2i}, \varepsilon_i)_{1\leq i\leq n}$$

du vecteur $(Y_1, Y_2, \delta_1, \delta_2, \varepsilon)$. La construction et l'étude de l'estimateur bivarié dans ce modèle feront l'objet du Chapitre 2. En particulier, on montrera que la fonction de poids W_n peut être définie comme

$$W_n(y_1, y_2, \epsilon) = \frac{1}{1 - G_n^{C_1}(\max(y_1, y_2 - \epsilon)-)},$$

où $G_n^{C_1}$ est l'estimateur de Kaplan-Meier basé sur l'échantillon

$$(\min(C_{1i}, A_i), \eta_i)_{1\leq i\leq n}, \quad \text{avec} \quad \begin{cases} A_i = \max(T_{1i}, T_{2i} - \varepsilon_i) \\ \eta_i = 1 - \delta_{1i}\delta_{2i} \end{cases}$$

Une application en assurance. Le modèle de l'encadré correspond à une application précise à l'étude des contrats de retraite complémentaire avec une clause de réversion, souscrits par les couples mariés. Ces assurances prévoient une indemnité versée aux veuves ou aux veufs après la mort de leur conjoint. Pour étudier les risques liés aux contrats de ce type, il est crucial d'estimer et de modéliser la dépendance entre les durées de vie de l'homme et de sa femme.

Lorsqu'on modélise les durées de vie des conjoints par les variables aléatoires T_1 et T_2, les censures C_1 et C_2 désignent leurs âges à la sortie de l'étude pour une cause autre que le décès. Puisque les deux membres du même couple sont liés par un contrat unique dans de nombreux jeux de données ils sortent de l'étude simultanément. Ainsi, la différence entre les deux censures représente la différence d'âge entre deux époux qui est généralement connue. La variable $C_2 - C_1$ est donc observable, et on retrouve le Modèle 2.

L'hypothèse du Modèle 2 correspond à des situations particulières. En toute généralité, on ne peut pas déduire une variable de censure de l'autre. Le Modèle 3 proposé par [Lopez and Saint-Pierre, 2012] permet de tenir compte de la structure de dépendance entre C_1 et C_2.

Modèle 3. Le lien entre les deux censures est modélisé par une copule connue.

On considère un vecteur aléatoire (T_1, T_2) avec les composantes censurées par (C_1, C_2). Soit S_1 (resp. S_2) la fonction de survie de C_1 (resp. C_2). On suppose que

- (T_1, T_2) est indépendant de (C_1, C_2).
- Il existe une fonction (copule) connue $\tilde{\mathfrak{C}}$ telle que

$$\mathbb{P}(C_1 > y_1, C_2 > y_2) = \tilde{\mathfrak{C}}(S_1(y_1), S_2(y_2)),$$

En utilisant la même idée que dans les cas précédents, [Lopez and Saint-Pierre, 2012] définissent un estimateur de la fonction de répartition de (T_1, T_2) à partir d'un échantillon $(Y_{1i}, Y_{2i}, \delta_{1i}, \delta_{2i})_{1\leq i\leq n}$. Leur estimateur est présenté sous la forme (1.6), avec

$$W_n(y_1, y_2) = \frac{1}{\tilde{\mathfrak{C}}(S_{1n}(y_1), S_{2n}(y_2))},$$

où S_{1n} et S_{2n} sont les estimateurs de Kaplan-Meier de S_1 et de S_2.

L'estimateur défini par cette approche, converge à la vitesse $n^{-1/2}$ et n'attribue à aucun point du plan un poids négatif. Néanmoins, ces propriétés satisfaisantes découlent de l'hypothèse forte sur la structure de dépendance entre les censures. A titre d'exemple, l'estimateur de Dabrowska, n'ayant pas le deuxième avantage de l'estimateur de [Lopez and Saint-Pierre, 2012], suppose également l'indépendance entre (T_1, T_2) et (C_1, C_2), mais ne fait aucune hypothèse sur le lien entre C_1 et C_2.

Les estimateurs définis dans le cadre des Modèles 1 à 3 présentent les avantages d'attribuer les poids positifs à toutes les observations et de converger à la vitesse $n^{-1/2}$. Plus encore, sous certaines conditions sur les queues des distributions, il est possible d'obtenir le théorème central limite fonctionnel (TCLF) pour leurs intégrales sur tout $\mathbb{R}^2$. Cela a été démontré pour les modèles 1 et 3, respectivement, dans [Stute, 1995] et [Lopez and Saint-Pierre, 2012]. Quant au modèle 2, ces propriétés seront établies dans le Chapitre 2.

Si l'on veut s'affranchir de cette hypothèse sur la structure de dépendance entre C_1 et C_2, on peut se placer dans le Modèle 4, qui correspond à la situation envisagée par Dabrowska. Un estimateur du type (1.6) qui respecte les contraintes de positivité des poids a été proposé par [Lopez, 2012].

Modèle 4. Cas général.
Soit (T_1, T_2) un vecteur aléatoire avec les composantes censurées par (C_1, C_2). On suppose que (T_1, T_2) est indépendant de (C_1, C_2).

Suivant les situations modélisées, on aura intérêt à utiliser le maximum d'information sur le phénomène de censure (c'est à dire se placer dans les Modèles 1 à 3 quand c'est possible).

1.3.3 Modélisation de dépendance entre deux durées

Jusqu'à maintenant on a étudié les interactions entre deux durées à travers l'estimation de leur fonction de répartition jointe. Cette méthode considère la distribution dans son ensemble englobant les lois marginales et la structure de dépendance. Lorsque l'on souhaite se focaliser plus précisément sur la comprehension des relations entre les variables, une approche par la modélisation de dépendance peut apparaitre plus avantageuse. Néanmoins, les lois paramétriques multivariées couvrent un nombre restreint de dépendances et imposent les restrictions strictes sur les marginales. Cette difficulté peut être évitée à l'aide de la théorie de copules. Ces fonctions permettent de décomposer l'analyse de la distribution multivarié en deux étapes indépendantes : la modélisation de dépendance et l'étude des lois marginales. De plus, les modèles de copules permettent de décrire une vaste gamme des dépendances.

Cette approche s'est avéré utile dans de nombreuses études de l'analyse de survie, notamment dans l'analyse de la dépendance entre deux durées de vie. Par exemple, [Frees et al., 1996], [Youn and Shemyakin, 1999], [Youn and Shemyakin, 2001] et [Luciano et al., 2008] ont appliqué des modèles de copules dans le cadre de l'étude de la dépendance entre les durées de vie de conjoints. Cette application constitue un apport important en assurance où l'on supposait auparavant l'indépendance entre deux durées, ce qui n'était absolument pas réaliste. Les copules ont également été utilisées par [Hougaard et al., 1992] pour relier les durées de vie des jumeaux dans le cadre d'une étude de l'influence d'un facteur génétique.

Dans un autre contexte, [Frees and Valdez, 1998], [Klugman and Parsa, 1999], [Carriere, 2000], et [Denuit and Van Keilegom, 2006] ont modélisé par des copules la dépendance entres les coûts de sinistres *loss* et $ALAE$ de la section précédente.

Lorsque l'on modélise la dépendance à l'aide d'une fonction copule, on suppose qu'elle appartient à une certaine classe paramétrique. Se posent ensuite les questions de l'estimation et de l'adéquation de modèle. En présence de censure, elles sont plus complexes et certaines procédures existantes dans le cadre non censuré ne sont pas disponibles. En particulier, pour tester l'adéquation en absence de censure, un estimateur paramétrique de copule peut être comparé avec son estimateur non paramétrique. On peut alors écarter un modèle si ces estimateurs sont "trop différents". Cette logique sera formalisée dans les sections suivantes où l'on présentera

des bases de la théorie des copules et des techniques associées. En présence de censure, jusqu'au maintenant, l'estimation non paramétrique de la copule n'a pas été étudiée. Cette démarche représente l'un des objectifs de cette thèse et sera présentée en détails dans le Chapitre 3.

1.4 Éléments de la théorie des copules

Cette section a pour objectif d'introduire la théorie des copules et des techniques associées. Par souci de simplicité, toutes les notions seront présentées en dimension égale à deux. Leurs généralisations à une dimension supérieure sont directes. Pour plus d'information, le lecteur pourra consulter deux monographies très complètes par [Nelsen, 2006] et [Joe, 1997]. Nous avons également fait le choix d'expliquer les procédures d'estimation et d'adéquation à partir d'un échantillon non censuré, en discutant, lorsque cela s'avère nécessaire, leur généralisation en présence de censure.

1.4.1 Théorème de Sklar et généralités

• **Définition de copule et le théorème de Sklar.** Par définition, une copule $\mathfrak{C}(u, v) : [0, 1]^2 \to [0, 1]$ est une fonction de répartition bivariée aux lois marginales uniformes, i.e.

1. Pour tous $(u, v) \in [0, 1]^2$, $\mathfrak{C}(u, 0) = \mathfrak{C}(0, v) = 0$, $\mathfrak{C}(u, 1) = u$ et $\mathfrak{C}(1, v) = v$;
2. Pour tout quadruplet (u_1, u_2, v_1, v_2) tel que $u_1 \leq u_2$ et $v_1 \leq v_2$,

$$\mathfrak{C}(u_2, v_2) - \mathfrak{C}(u_2, v_1) - \mathfrak{C}(u_1, v_2) + \mathfrak{C}(u_1, v_1) \geq 0.$$

Les copules ont d'abord été introduites en probabilités, puis ont trouvé leurs applications en statistique grâce au résultat important démontré par [Sklar, 1959] :

Théorème 1. Soit $F(t_1, t_2)$ une fonction de répartition aux marginales F_1 et F_2. Il existe une copule $\mathfrak{C}(u_1, u_2)$, telle que

$$F(t_1, t_2) = \mathfrak{C}(F_1(t_1), F_2(t_2)). \tag{1.8}$$

De plus, si les fonctions marginales F_1 et F_2 sont continues, alors la copule $\mathfrak{C}$ est unique.

Dans la littérature de l'analyse de survie, certains auteurs préfèrent faire appel à la fonction de survie $S(t_1, t_2) = \mathbb{P}(T_1 > t_1, T_2 > t_2)$ plutôt qu'à la fonction de répartition. Ceci n'est qu'une convention et le théorème de Sklar peut être exprimé de façon équivalente pour des fonctions de survie, i.e. il existe une copule $\tilde{\mathfrak{C}}$ dite de survie telle que

$$S(t_1, t_2) = \tilde{\mathfrak{C}}(S_1(t_1), S_2(t_2)).$$

La copule $\mathfrak{C}$ et la copule de survie $\tilde{\mathfrak{C}}$ du même vecteur aléatoire sont liées par la relation

$$\tilde{\mathfrak{C}}(u, v) = u + v - 1 + \mathfrak{C}(1 - u, 1 - v).$$

• **Copules et les mesures de dépendance.** Puisque les lois marginales n'apportent pas d'information sur la structure de dépendance, la copule apparait comme une quantité capturant toute l'information disponible sur cette structure. Par conséquent, les mesures de dépendance classiques peuvent être exprimées uniquement en fonction de la copule.

Tau de Kendall. Soient $(T_1^{(1)}, T_2^{(1)})$ et $(T_1^{(2)}, T_2^{(2)})$ deux copies i.i.d. d'un vecteur aléatoire (T_1, T_2). Le coefficient de Kendall de concordance entre les variables aléatoires T_1 et T_2 est défini comme

$$\tau_{T_1T_2} = P[(T_1^{(1)} - T_1^{(2)})(T_2^{(1)} - T_2^{(2)}) > 0] - P[(T_1^{(1)} - T_1^{(2)})(T_2^{(1)} - T_2^{(2)}) < 0].$$

De façon informelle, cet indicateur montre si les "grandes" valeurs de T_1 ont tendance à être associées aux "grandes" valeurs de T_2. Si $\mathfrak{C}$ est la copule qui lie T_1 et T_2, alors

$$\tau_{T_1T_2} = 4 \int \mathfrak{C}(u, v) d\mathfrak{C}(u, v) - 1.$$

Rho de Spearman. Une autre mesure d'association entre les variables aléatoires T_1 et T_2 est le rho de Spearman qui est défini comme un coefficient de corrélation linéaire entre les rangs :

$$\rho^S_{T_1T_2} = \rho(F_{T_1}(T_1), F_{T_2}(T_2)),$$

où F_{T_1} et F_{T_2} sont les fonctions de répartition respectives de T_1 et de T_2. Le rho de Spearman s'exprime en fonction de la copule par

$$\rho^S_{T_1T_2} = 12 \int_0^1 \int_0^1 (\mathfrak{C}(u, v) - uv) du dv.$$

• **Modèles de copules.** Le théorème de Sklar montre que la modélisation de la structure de dépendance, sous-jacente à une fonction de répartition $F(t_1, t_2)$, peut être décomposée en deux étapes indépendantes : le choix d'un modèle pour la copule $\mathfrak{C}(u, v)$ et la spécification des lois marginales.

On considère une classe paramétrique de copules $\mathcal{C}_\Theta = \{\mathfrak{C}_\theta : \theta \in \Theta\}$, indexée par un ensemble $\Theta \subseteq \mathbb{R}^k$. En suivant le théorème de Sklar, le modèle de copule pour la fonction de répartition s'écrit comme suit :

$$F(t_1, t_2) = \mathfrak{C}_\theta(F_1(t_1), F_2(t_2)). \tag{1.9}$$

Il est dit *paramétrique* lorsque les lois marginales sont également modélisées par des lois paramétriques, et **semi-paramétrique** lorsqu'elles ne sont pas spécifiées.

1.4.2 Copules archimédiennes

A l'heure actuelle, il existe un grand choix de classes paramétriques de copules. Pour une revue des familles usuelles, on renvoie le lecteur vers les monographies de [Nelsen, 2006] et [Joe, 1997]. Dans cette section, nous allons considérer une classe particulièrement importante de copules archimédiennes souvent utilisées en pratique.

Soit $\varphi : [0,1] \to [0,+\infty]$ une fonction arbitraire convexe, continue, strictement décroissante telle que $\varphi(0) = +\infty$ et $\varphi(1) = 0$, alors la fonction

$$C(u,v) = \varphi^{[-1]}(\varphi(u) + \varphi(v)), \tag{1.10}$$

définit une copule, dite archimédienne de générateur φ. Ici $\varphi^{[-1]}(t)$ désigne l'inverse généralisée de φ définie par

$$\varphi^{[-1]}(t) = \begin{cases} \varphi^{-1}(t), & \text{si } t \in [0,+\infty], \\ 0, & \text{sinon.} \end{cases}$$

La popularité de cette classe est due notamment à la facilité de construction des familles archimédiennes. En effet, pour en construire une, il suffit de prendre une famille paramétrique $\{\varphi_\theta, \theta \in \Theta\}$ de générateurs et appliquer la transformation (1.10). La diversité des familles de copules archimédiennes permet à cette classe de décrire une vaste gamme de dépendances.

D'autre part, les mesures de dépendance entre des variables aléatoires liées par une copule archimédienne peuvent s'exprimer de façon très simple à partir de son générateur φ_θ. Par exemple, pour le tau de Kendall, on a

$$\tau = 1 + 4 \int_0^1 \frac{\varphi_\theta(u)}{\varphi'_\theta(u)} du.$$

Un autre objet important qui peut facilement être exprimé à partir du générateur, est la fonction de répartition du vecteur aléatoire (T_1, T_2) transformé par sa loi $F(t_1, t_2)$. Elle est définie de la façon suivante :

$$K(z) = P(F(T_1, T_2) \leq z). \tag{1.11}$$

[Genest and Rivest, 1993] ont constaté que

$$K_\theta(z) = z - \frac{\varphi_\theta(z)}{\varphi'_\theta(z)}. \tag{1.12}$$

La fonction $K(z)$ intervient dans le processus empirique dit de Kendall. Dans la Section 1.4.4, on montrera que l'expression (1.12) permet de construire des tests d'adéquation, basé sur ce processus et valables pour les copules archimédiennes. L'existence de ces tests qui sont spécifiques à la classe archimédienne est un avantage supplémentaire de cette classe.

1.4.3 Inférence statistique

Dans cette section, nous considérons l'estimation des modèles de copules pour la fonction de répartition $F(t_1, t_2)$ à partir d'un échantillon $(T_{1i}, T_{2i})_{1 \leq i \leq n}$ de loi F.

Méthode par l'inversion d'une mesure de dépendance

Cette méthode consiste à exprimer une mesure k de dépendance (par exemple, le tau de Kendall) comme une fonction $k(\theta)$ de paramètre θ de la copule. Cette

mesure est ensuite estimée à partir des données par une valeur $\hat{k}$. Un estimateur de θ s'obtient par l'inversion de la fonction $k(\theta)$, i.e. $\hat{\theta} = k^{-1}(\hat{k})$. L'avantage de cette méthode est sa simplicité. Ses inconvénients sont une faible performance et la nécessité d'utiliser plusieurs mesures de dépendance, lorsque le paramètre de la copule est multidimensionnel.

Estimation paramétrique et semi-paramétrique

Soient $f(t_1, t_2)$ la densité de $F(t_1, t_2)$ et f_1, f_2 ses densités marginales. Par le théorème de Sklar, on a

$$f(t_1, t_2) = c(F_1(t_1), F_2(t_2)) f_1(t_1) f_2(t_2),$$

où l'on a introduit la densité de copule

$$c(u, v) = \frac{\partial^2 \mathfrak{C}(u, v)}{\partial u \partial v}.$$

Estimation paramétrique. Cette méthode utilise des modèles paramétriques pour la copule et pour les lois marginales. La fonction L de log-vraisemblance du modèle s'écrit sous la forme :

$$L = \sum_{i=1}^{n} \log c(F_1(T_{1i}, \theta_1), F_2(T_{1i}, \theta_2), \theta) + \sum_{i=1}^{n} \log f_1(T_{1i}, \theta_1) + \sum_{i=1}^{n} \log f_2(T_{2i}, \theta_2), \quad (1.13)$$

où θ désigne le paramètre de copule et θ_1, θ_2 ceux des lois marginales. Leurs estimateurs par le maximum de vraisemblance sont

$$(\hat{\theta}, \hat{\theta}_1, \hat{\theta}_2) = \arg \max_{\theta, \theta_1, \theta_2} L(\mathcal{T}_n, \theta, \theta_1, \theta_2),$$

où l'on a noté $\mathcal{T}_n = (T_{1i}, T_{2i})_{1 \leq i \leq n}$. La maximisation simultanée par rapport à tous les paramètres est coûteuse d'un point de vue numérique, ainsi elle est rarement utilisée en pratique. La procédure habituelle, proposée par [Shih and Louis, 1995], consiste à décomposer le problème selon les deux étapes suivantes :

- Estimation des paramètres θ_1, θ_2 des marginales par les maximums des vraisemblances unidimensionnelles $\hat{\theta}_1$, $\hat{\theta}_2$.
- Injection des estimateurs $\hat{\theta}_1$ et $\hat{\theta}_2$ dans la fonction de vraisemblance (1.13) et sa maximisation par rapport à θ.

Estimation semi-paramétrique. Dans un modèle semi-paramétrique de dépendance, la copule est modélisée par une famille paramétrique et les lois marginales ne sont pas spécifiées. Dans ce cas, la vraisemblance s'écrit comme

$$L = \sum_{i=1}^{n} \log c(F_1(T_{1i}), F_2(T_{1i}), \theta) + \sum_{i=1}^{n} \log\{f_1(T_{1i}) f_2(T_{2i})\},$$

où le deuxième terme ne contient pas de paramètre inconnu. La partie de la vraisemblance à maximiser est donc son premier terme qui contient pour autant les marginales inconnues. Pour ce modèle, la méthode d'estimation dite *omnibus* a été proposé dans [Genest et al., 1995] et [Shih and Louis, 1995]. Elle se décompose également en deux étapes, mais cette fois les marginales sont estimées d'une façon non paramétrique :

- Estimation des lois marginales F_1, F_2 par les estimateurs non paramétriques $\hat{F}_1$, $\hat{F}_2$.
- Injection de $\hat{F}_1$ et $\hat{F}_2$ à la place de F_1 et F_2 dans le premier terme de vraisemblance et sa maximisation par rapport à θ :

$$\hat{\theta} = \arg\max_{\theta} \sum_{i=1}^{n} \log c_\theta(\hat{F}_1(T_{1i}), \hat{F}_2(T_{1i})).$$

Un des avantages de cette méthode est l'absence de risque d'une mauvaise spécification des marginales.

En présence de censure, les deux procédures fonctionnent de manière analogue et sont également traitées par [Shih and Louis, 1995]. L'estimation paramétrique utilise les deux mêmes étapes, mais avec une fonction vraisemblance modifiée pour prendre en compte la censure. Quant à l'estimation semi-paramétrique, en plus de la modification de la vraisemblance, les estimateurs non paramétriques des marginales sont calculés par la méthode de Kaplan-Meier.

Estimation non-paramétrique

Un estimateur discret. Le premier estimateur non-paramétrique de copule, dite *copule empirique*, a été introduit par [Deheuvels, 1979]. Pour le construire, on considère tout d'abord la fonction de répartition empirique :

$$\hat{F}(t_1, t_2) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{T_{1i} \leq t_1, T_{2i} \leq t_2}.$$

Ensuite, par le théorème de Sklar, on remarque que

$$\mathfrak{C}(u, v) = F(F_1^{-1}(u), F_2^{-1}(v)).$$

Pour estimer la copule, il convient alors de remplacer dans la formule précédente la fonction de répartition F inconnue par son estimateur non paramétrique $\hat{F}$. Cela conduit à l'estimateur suivant de la fonction copule :

$$\hat{\mathfrak{C}}(u, v) = \hat{F}(\hat{F}_1^{-1}(u), \hat{F}_2^{-1}(v)). \tag{1.14}$$

Ses propriétés asymptotiques ont été étudiées en termes de processus empirique associé,

$$n^{1/2}(\hat{\mathfrak{C}}(u, v) - \mathfrak{C}(u, v)), \qquad \text{pour } u, v \in [0, 1].$$

Ce processus converge en distribution vers un pont brownien, dans l'espace $l^\infty([0,1]^2)$ des fonctions uniformément bornées sur $[0,1]^2$. Pour plus de détails, on renvoie le lecteur à [Deheuvels, 1979], [Gaenssler and Stute, 1987], [Fermanian et al., 2004] (qui ont également considéré un processus empirique associé à l'estimateur par noyau de la copule), [Tsukahara, 2005] et [Segers, 2012].

Un estimateur lisse. La copule empirique est une fonction discrète, alors que les copules usuelles sont le plus souvent continues et possèdent une densité. Il serait

donc naturel d'envisager une généralisation lisse de $\hat{\mathfrak{C}}$. Un tel estimateur, que l'on présente par la suite, a été proposé par [Fermanian et al., 2004].

Soit $k : \mathbb{R} \mapsto \mathbb{R}$ un noyau (i.e. une fonction positive, intégrable et à valeurs réelles) et $K(x) = \int_{-\infty}^{x} k(u)du$ sa primitive. Un estimateur par noyau de la fonction de répartition s'écrit sous la forme

$$\tilde{F}(t_1, t_2) = \frac{1}{n} \sum_{i=1}^{n} K_h \left(t_1 - T_{1i}\right) K_h(t_2 - T_{2i}) \,, \tag{1.15}$$

où h est appelé la **fenêtre** et $K_h(t) = K(t/h)$. L'estimateur lisse de copule, considéré par [Fermanian et al., 2004], est donné par :

$$\tilde{\mathfrak{C}}(u, v) = \tilde{F}(\tilde{F}_1^{-1}(u), \tilde{F}_2^{-1}(v)). \tag{1.16}$$

L'estimateur (1.16) permet de définir un estimateur de la densité de copule par la dérivation. Ce dernier peut servir comme un outil graphique de la sélection des modèles de copules.

De façon générale, l'intérêt des estimateurs non paramétriques est qu'ils ne nécessitent aucun a priori sur la copule sous-jacente et peuvent donc servir pour la validation des modèles. Dans la section suivante, nous présenterons, entres d'autres, des tests d'adéquation basés sur la copule empirique.

En présence de censure, ces estimateurs n'ont pas été généralisés, jusqu'à maintenant, en raison de la complexité de l'estimation de dépendance en présence de censure bivariée. Dans le Chapitre 3, nous montrerons une possibilité d'une telle généralisation à partir des estimateurs de la forme générique (1.6).

1.4.4 Tests d'adéquation pour les modèles de copules

Diverses familles de copules peuvent représenter des structures de dépendance très différentes. Lorsqu'il s'agit de modéliser un jeu de données réel, certaines classes paramétriques peuvent se révéler inadaptées, i.e. incapables de capter la spécificité de corrélation entre les variables. Les tests d'adéquation permettent de valider un modèle de copule adéquat et d'éliminer ceux qui ne correspondent pas à la dépendance étudiée.

Soient T_1 et T_2 les variables liées par une copule $\mathfrak{C}$, et soient $(T_{1i}, T_{2i})_{1 \leq i \leq n}$ les réplications i.i.d. de (T_1, T_2). Pour une classe paramétrique de copules

$$\mathcal{C}_\Theta = \{\mathfrak{C}_\theta : \theta \in \Theta\},$$

on considère le problème de test d'hypothèse suivant (test d'adéquation) :

$$H_0 : \mathfrak{C} \in \mathcal{C}_\Theta \qquad \text{contre} \qquad H_1 : \mathfrak{C} \notin \mathcal{C}_\Theta$$

L'idée commune à toutes les procédures que l'on va considérer est de comparer un estimateur paramétrique de la copule ou de sa fonctionnelle sous l'hypothèse H_0 à son estimateur non paramétrique. Les trois approches principales sont basées respectivement sur le processus de Kendall, les estimateurs de la densité de copule et la copule empirique.

Approche basée sur le processus de Kendall

La fonction $K(z)$ de Kendall définie dans (1.11) peut être écrite sous la forme

$$K(z) = \mathbb{P}(\mathfrak{C}(U_1, U_2) \leq z),$$

où $U_j = F_j(T_j)$, pour $j = 1, 2$. Un estimateur non paramétrique de $K(z)$ est alors donné par

$$K_n(z) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\hat{F}(T_{1i}, T_{2i}) \leq z}. \tag{1.17}$$

Le processus empirique associé à cet estimateur (processus de Kendall) se définit, pour $z \in [0, 1]$, par

$$n^{1/2}(K_n(z) - K(z)).$$

Ce processus a été introduit par [Genest and Rivest, 1993] et étudié en détail par [Barbe et al., 1996]. Il a été utilisé pour la construction de tests d'adéquation à des familles archimédiennes pour la première fois par [Wang and Wells, 2000b] (dans le cadre de données censurées). Leur statistique de test est basée sur la distance L_2 entre l'estimateur non paramétrique K_n et la fonction K_{θ_n}, calculée à partir de (1.12), où θ_n est un estimateur du paramètre de la copule sous H_0 :

$$d(K_n, K_{\theta_n}) = n \int_{\xi}^{1} \{K_n(z) - K_{\theta_n}(z)\}^2 \, dz,$$

où ξ est un paramètre de troncation. Par la suite, [Genest et al., 2006] ont suggéré deux autres tests utilisant les statistiques

$$d_{1n} = n \int_{0}^{1} \{K_n(z) - K_{\theta_n}(z)\}^2 \, dK_{\theta_n}(z) \quad \text{et} \quad d_{2n} = \sup_{0 \leq z \leq 1} \left| n^{1/2}(K_n(z) - K_{\theta_n}(z)) \right|.$$

Les auteurs ont également proposé une procédure de bootstrap pour l'évaluation des valeurs critiques.

Un avantage important de l'approche basée sur le processus de Kendall est sa facilité d'implementation. Son inconvénient est d'être valable uniquement pour les familles de copules archimédiennes.

Approche basée sur la densité de copule

Cette approche a été développée dans [Fermanian, 2005]. Elle se base sur la distance entre c_{θ_n}, l'estimateur paramétrique de la densité c de copule sous H_0, et son estimateur non paramétrique défini par

$$c_n(u, v) = \frac{1}{nh^2} \sum_{i=1}^{n} k\left(\frac{u - \hat{F}_1(T_{1i})}{h}\right) k\left(\frac{u - \hat{F}_2(T_{2i})}{h}\right),$$

où $\hat{F}_1, \hat{F}_2$ sont les fonctions de répartition empiriques marginales et k est un noyau. La statistique de test est basée sur la distance

$$J_n = \int_{[0,1]^2} \{c_n(u, v) - k_h * c_{\theta_n}(u, v)\}^2 \omega(u, v) du dv,$$

où $k_h(u) = k(u/h)$, $\omega(u, v)$ est une certaine fonction de pondération et "$*$" désigne la loi de convolution. La normalité asymptotique de J_n a été démontré dans [Fermanian, 2005].

Approche basée sur la copule empirique

Cette approche naturelle consiste à comparer un estimateur paramétrique de la copule sous H_0 et son estimateur non paramétrique, à savoir la copule empirique définie par (1.14). L'avantage de cette approche est de ne pas se restreindre aux copules archimédiennes et de pouvoir s'appliquer à toute famille paramétrique de copules.

Les tests d'adéquation basés sur le processus empirique de la copule ont été considérés, par exemple, dans les articles [Fermanian, 2005], [Genest and Rémillard, 2008] et [Genest et al., 2009]. Les statistiques de test correspondantes sont basées sur les distances de Cramér-von Mises et Kolmogorov-Smirnov :

$$d_n^{CvM} = n \int \left\{ (\hat{\mathfrak{C}} - \mathfrak{C}_{\theta_n})(u,v) \right\}^2 d\hat{\mathfrak{C}}(u,v)$$

et

$$d_n^{KS} = \sup_{(u,v)\in[0,1]^2} \left| n^{1/2} (\hat{\mathfrak{C}} - \mathfrak{C}_{\theta_n})(u,v) \right| .$$

La convergence faible du processus empirique de copule permet de déduire le comportement asymptotique de ces statistiques. [Genest and Rémillard, 2008] mentionnent que la façon convenable de calculer les p-valeurs pour ces tests consiste à utiliser la méthode de bootstrap.

En présence de censure, [Wang and Wells, 2000b] ont proposé d'utiliser l'approche par le processus de Kendall, en remplaçant dans (1.17) la fonction de répartition empirique par l'estimateur de Dabrowska. L'inconvénient de cet estimateur, comme on l'a déjà vu, est l'attribution de poids négatifs à certains points du plan. Il est donc inconfortable pour les procédures bootstrap qui doivent resimuler les données à partir d'une loi estimée. Dans le cas particulier du Modèle 2, cette difficulté peut être contournée en utilisant notre estimateur avec des poids positifs, compatible avec la procédure bootstrap (voir le Chapitre 2).

Pour les approches via la densité de copule ou la copule empirique, elles n'existent pas en présence de censure, faute d'estimateurs non paramétriques associés. Ces estimateurs, déjà évoqués à la fin de la Section 1.4.3, seront introduits dans le Chapitre 3. Ils nous permettront de proposer une extension de l'approche par la copule empirique en présence de censure.

1.5 Quantification et clustering

Dans les deux sections précédentes, nous avons discuté des problèmes de l'estimation et de la modélisation de dépendance pour des données censurées. Nous avons vu que le problème de la prise en compte de cette dépendance joue un rôle considérable dans les applications. Un autre problème présentant des enjeux importants dans l'analyse de durée multivariée est celui de l'hétérogénéité de la population. Par exemple, en assurance, un portefeuille d'assurés est généralement composé de classes de risque homogènes (liées à des caractéristiques socio-professionnelles, géographiques, etc.). Savoir les distinguer peut permettre à la compagnie d'assurance

de cibler mieux ses risques. En médecine, séparer la population de patients en clusters d'individus avec des caractéristiques similaires peut permettre de différencier les méthodes de traitement de façon plus efficace.

En absence de censure, le problème de partitionnement de la population en groupes homogènes peut être résolu par des méthodes de clustering, basées sur les distances entre les observations. Dans le cas où l'échantillon multivarié contient des observations d'une variable de type "durée de vie" ou "coût de sinistre", sa composante correspondante peut être sujette à la censure. Cela rend inobservables les distances entre observations et n'autorise pas l'application des méthodes standards. L'objectif du Chapitre 4 sera de proposer une extension de la méthode de clustering par quantification en présence de censure. Les sections suivantes ont pour rôle de présenter cette méthode dans son cadre standard, i.e. non censuré.

1.5.1 Principe de quantification

Le clustering (ou le partitionnement de données) est une méthode statistique de classification non supervisée. Son objectif consiste à séparer l'ensemble des individus selon un petit nombre des classes homogènes, ou clusters, en minimisant l'inertie intra-cluster et en maximisant l'inertie inter-cluster.

Nous allons nous intéresser plus particulièrement au clustering dit k-means qui cherche à créer une partition de n observations en k clusters, dans laquelle chaque observation appartient au cluster dont le centre est le plus proche. L'origine théorique de cette méthode est le principe probabiliste de la quantification dont l'objectif est de compresser l'information contenue dans une probabilité. Un opérateur mathématique permettant d'effectuer cette compression est appelée **quantificateur**.

Pour formaliser le problème, nous considérons un vecteur aléatoire Z à valeurs dans $\mathbb{R}^d$ de loi P, avec $\mathbb{E}\,\|Z\|^2 < \infty$. Soit

$$\mathcal{C} = \{c_1, \ldots, c_k\}$$

un sous-ensemble de $\mathbb{R}^d$. On appellera un k-quantificateur toute fonction mesurable

$$q : \mathbb{R}^d \to \mathcal{C} = \{c_1, \ldots, c_k\}, \text{ avec } c_i \in \mathbb{R}^d, \text{ pour } i = 1, \ldots, k.$$

L'ensemble $\mathcal{C}$ est dit un alphabet associé au quantificateur. Un $k-$quantificateur est entièrement caractérisé par son alphabet $\mathcal{C}$ et les cellules

$$S_i = \{z : q(z) = c_i\}, \quad i = 1, \ldots, k,$$

selon la règle $\{q(z) = c_i \Leftrightarrow z \in S_i\}$.

Le problème de quantification consiste à représenter Z par $q(Z)$ en choisissant le quantificateur q de sorte que la représentation est la plus précise possible. Il est à noter que $q(Z)$ peut prendre uniquement k valeurs distinctes. Ainsi, en remplaçant Z par $q(Z)$, on résume la loi P par une distribution discrète de support $\mathcal{C}$.

La précision de cette approximation se mesure par une erreur de représentation, appelée **distorsion**, donnée par

$$D(P, q) = \mathbb{E}||Z - q(Z)||^2 = \int ||z - q(z)||^2 dP(z). \tag{1.18}$$

On définit la performance optimale dans la classe Q_k de tous les k-quantificateurs par

$$D_k^*(P) = \inf_{q \in Q_k} D(P, q).$$

Un quantificateur q^* est dit optimal s'il atteint cette performance, c'est à dire

$$D(P, q^*) = D_k^*(P).$$

Un tel q^* existe toujours (voir [Pollard, 1982b]) et appartient à la classe des quantificateurs des plus proches voisins, i.e. il vérifie

$$||z - q^*(z)||^2 = \min_{c_i \in \mathcal{C}} ||z - c_i||,$$

où $\mathcal{C}$ est son alphabet. Cela signifie qu'il suffit de rechercher le quantificateur optimal parmi les quantificateurs des plus proches voisins. Toute partition de l'espace associée à un quantificateur des plus proches voisins est appelée **partition de Voronoï**.

1.5.2 Quantificateur empirique

En pratique, la loi du vecteur aléatoire Z est inconnue. Par conséquent, il est impossible de déterminer q^* en minimisant (1.18). Néanmoins, lorsque l'on dispose d'un échantillon $(Z_1, \ldots, Z_n)$ de loi P, celle-ci peut être approchée par la loi empirique

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

où δ_x désigne la mesure de Dirac en x. Au lieu de minimiser la distorsion inconnue (1.18), l'idée consiste alors à minimiser la distorsion empirique, définie par

$$D(P_n, q) = \int ||z - q(z)||^2 dP_n(z) = \frac{1}{n} \sum_{i=1}^{n} ||Z_i - q(Z_i)||^2.$$

Le quantificateur q_n^* est dit empirique optimal si

$$D(P_n, q_n^*) = \inf_{q \in Q_k} D(P_n, q).$$

Lorsque l'on a construit un tel q_n^*, il est naturel de se demander si sa performance approche asymptotiquement la distorsion minimale d'un quantificateur optimal q^*. La réponse est affirmative et est donnée par le théorème suivant (voir [Pollard, 1982b, Pollard, 1982a, Abaya and Wise, 1984, Graf and Luschgy, 1994]).

Théorème 2. (Consistance). Soit q_n^* un k-quantificateur empirique optimal construit à partir de l'échantillon $(Z_1, \ldots, Z_n)$ de loi P. On a

$$\lim_{n \to \infty} D(P, q_n^*) \to D_k^*(P) \quad \text{p.s.}$$

Dans les applications, la taille de l'échantillon est toujours finie. Il est donc important d'étudier la vitesse de convergence établie par le Théorème 2. La borne supérieure non asymptotique que nous présentons au Théorème 3 est due à [Linder et al., 1994].

Théorème 3. (Vitesse). Supposons que $P(||Z|| \leq R) = 1$ pour un certain $R \in]0, \infty[$. Il existe une constante C dépendant uniquement de d, k et R telle que

$$|\mathbb{E}D(P, q_n^*) - D_k^*(P)| \leq \frac{C}{\sqrt{n}}.$$

Pour une revue détaillée des concepts présentés dans les deux dernières sections et pour plus de résultats sur la vitesse de convergence on pourra consulter [Linder, 2002].

Dans la littérature statistique, le problème de la construction de quantificateur empirique optimal est souvent désigné par k-means clustering (voir [MacQueen, 1967]).

1.5.3 Algoritme des $k-$means

Comme on l'a vu, le k-quantificateur optimal appartient à la classe des quantificateurs des plus proches voisins qui sont entièrement caractérisés par leurs k centres. La distorsion empirique d'un quantificateur des plus proches voisins s'écrit sous la forme :

$$D(P_n, q) = \frac{1}{n} \sum_{i=1}^{n} \min_{c_j \in \mathcal{C}} ||Z_i - c_j||^2$$

Sa minimisation s'effectue donc sur l'ensemble de tous les alphabets $\mathcal{C}$ de la taille k. D'un point de vue numérique, ce problème est "NP-complet" (ne peut pas être résolue dans un temps polynomial). Par conséquent, l'évaluation des centres d'un quantificateur empirique optimal nécessite d'avoir recours à des algorithmes heuristiques qui en fournissent des valeurs approchées.

On considère les observations $(z_1, \ldots, z_n)$ de vecteur aléatoire Z de loi P. L'algorithme des k-means est une procédure itérative de partition des observations due à [Lloyd, 2006] et [Steinhaus, 1956]. Elle repose sur le théorème suivant, démontré par les mêmes auteurs.

Théorème 4. Soit q un k-quantificateur arbitraire à l'alphabet $\mathcal{C} = \{c_i\}_{i=1}^k$ et les cellules de partition $\{S_i\}_{i=1}^k$. On a :

- Si q' est un k-quantificateur des plus proches voisins ayant le même alphabet, alors $D(P, q') \leq D(P, q)$.
- Si q' est un k-quantificateur ayant les mêmes cellules de partition que q et les centres définis par $c_i' = \mathbb{E}[Z|Z \in S_i], \quad i = 1, \ldots, k$, alors $D(P, q') \leq D(P, q)$.

Dans l'algorithme des k-means, chaque itération est composée de deux étapes d'actualisation successives de l'alphabet et des cellules de partition. Nous allons maintenant présenter la procédure exacte :

- Initialiser les k centres
- Répéter jusqu'à ce qu'il n'y ait plus de changement :
  - A partir des centres $\{c_{l,i}\}_{i=1}^{k}$ obtenus à l'itération l précédente, calculer la partition de Voronoi $\{S_{l+1,i}\}_{i=1}^{k}$ associée, i.e.

    $$S_{l+1,i} = \{z_j : ||z_j - c_{l,i}|| \leq ||z_j - c_{l,m}||, \text{ pour tout } m = 1, \ldots, k\}.$$

  - Actualiser les centres par $\{c_{l+1,i}\}_{i=1}^{k}$ selon la formule

    $$c_{l+1,i} = \frac{1}{|S_{l+1,i}|} \sum_{j=1}^{n} \mathbb{1}_{z_j \in S_{l+1,i}},$$

    où $|S_{l+1,i}|$ désigne le cardinal de l'ensemble $|S_{l+1,i}|$.

Grâce au Théorème 4, la distorsion décroît à chaque itération et l'algorithme ainsi défini converge toujours en temps fini vers un minimum local.

1.6 Contributions

Les contributions de cette thèse sont regroupées dans les Chapitres 2 à 4. Le Chapitre 2 reprend un article publié dans *Journal of Multivariate Analysis* (voir [Gribkova et al., 2013]) et traite de l'estimation de la fonction de répartition dans le cadre du Modèle 2 de la Section 1.3.2. Dans le Chapitre 3, nous proposons une procédure de l'estimation non paramétrique de copule en présence de censure. Elle est basée sur les estimateurs de la fonction de répartition de la forme générique (1.6), dont l'estimateur du Chapitre 2 est un cas particulier. Enfin, le dernier chapitre aborde la problématique de l'exploration de données hétérogènes en présence de censure.

1.6.1 Estimation pour le modèle de durée où la différence entre les censures est observée

Dans cette section, nous allons d'abord définir un estimateur de la fonction de répartition jointe pour le Modèle 2. Ensuite, nous allons considérer ses applications à l'estimation du tau de Kendall et aux tests d'adéquation pour les modèles de copules, basés sur le processus de Kendall.

Estimateur de la fonction de répartition

Nous allons nous placer maintenant dans le cadre du Modèle 2 de la Section 1.3.2. Notre premier objectif consiste à définir un estimateur de la fonction de répartition jointe des deux durées T_1 et T_2 à partir d'un échantillon

$$(Y_{1i}, Y_{2i}, \delta_{1i}, \delta_{2i}, \varepsilon_i)_{1 \leq i \leq n} = (\min(T_{1i}, C_{1i}), \min(T_{2i}, C_{2i}), \delta_{1i}, \delta_{2i}, \varepsilon_i)_{1 \leq i \leq n}.$$

Suivant la logique évoquée precédemment, nous allons chercher un estimateur sous la forme :

$$F_n(t_1,t_2) = \frac{1}{n}\sum_{i=1}^{n} \delta_{1i}\delta_{2i} W_n(Y_{1i},Y_{2i},\varepsilon_i)\mathbb{1}_{Y_{1i}\leq t_1, Y_{2i}\leq t_2}, \tag{1.19}$$

où W_n sera déterminé comme un estimateur d'une certaine fonction $W^*(y_1,y_2,\epsilon)$, qui satisfait, pour tout $\phi(y_1,y_2)$ d'espérance finie, l'identité suivante :

$$\mathbb{E}[\delta_1\delta_2 W^*(Y_1,Y_2,\varepsilon)\phi(Y_1,Y_2)] = \mathbb{E}[\phi(T_1,T_2)]. \tag{1.20}$$

Soit $G^{C_1}(y_1)$ la fonction de répartition de C_1 et $S^{C_1}(y_1)$ sa fonction de survie. Sous les hypothèses du Modèle 2, on a

$$\begin{aligned}\mathbb{E}[\delta_1\delta_2\phi(Y_1,Y_2)] &= \mathbb{E}\left[\mathbb{E}\left[\mathbb{1}_{\max(T_1,T_2-\varepsilon)\leq C_1}|T_1,T_2,\varepsilon\right]\phi(T_1,T_2)\right]\\ &= \mathbb{E}\left[S^{C_1}(\max(T_1,T_2-\varepsilon)-)\phi(T_1,T_2)\right],\end{aligned}$$

ce qui suggère de choisir la fonction W^* suivante :

$$W^*(y_1,y_2,\epsilon) = \frac{1}{S^{C_1}(\max(y_1,y_2-\epsilon)-)}, \tag{1.21}$$

Cette fonction fait intervenir la fonction de survie $S^{C_1}(y_1)$ inconnue. Comme nous l'avons décrit dans la Section 1.3.2, l'idée consiste à remplacer $S^{C_1}(y_1)$ par son estimateur $S_n^{C_1}(y_1)$. L'injection de $S_n^{C_1}(y_1)$ dans (1.21) permet d'obtenir une fonction aléatoire $W_n(y_1,y_2,\epsilon)$, estimateur de $W^*(y_1,y_2,\epsilon)$.

Pour estimer $S^{C_1}(y_1)$, on remarque que la censure C_1 est observée lorsqu'elle est inférieure à une variable $A = \max(T_1,T_2-\varepsilon)$. Par conséquent, C_1 peut être considérée elle-même comme une variable censurée par A, avec un indicateur de censure $\eta = 1-\delta_1\delta_2$. Sa distribution peut donc être estimée par un estimateur de Kaplan-Meier $S_n^{C_1}$ basé sur un échantillon $(\min(C_{1i},A_i),\eta_i)_{1\leq i\leq n}$.

On définit la fonction aléatoire W_n, estimateur de W^*, en remplaçant dans (1.21) la fonction S^{C_1} par $S_n^{C_1}$:

$$W_n(y_1,y_2,\epsilon) = \frac{1}{S_n^{C_1}(\max(y_1,y_2-\epsilon)-)}. \tag{1.22}$$

Cela nous conduit à la forme finale suivante de l'estimateur de la fonction de répartition de (T_1,T_2) :

$$F_n(t_1,t_2) = \frac{1}{n}\sum_{i=1}^{n} \frac{\delta_{1i}\delta_{2i}}{S_n^{C_1}(\max(Y_{1i},Y_{2i}-\varepsilon_i)-)}\mathbb{1}_{Y_{1i}\leq t_1, Y_{2i}\leq t_2}. \tag{1.23}$$

La difficulté majeure de l'analyse asymptotique de l'estimateur (1.23) réside dans les dépendances mutuelles des termes de la somme. Pour résoudre ce problème, nous avons établi une représentation i.i.d. des intégrales par rapport à la mesure définie par $F_n(t_1,t_2)$. Elle est énoncée dans le Théorème 5 et prouvée dans le Théorème 1 de la Section 2.3.1. Afin de l'expliciter, nous introduisons d'abord la fonction suivante :

$$F^*(t_1,t_2) = \frac{1}{n}\sum_{i=1}^{n} \delta_{1i}\delta_{2i} W^*(Y_{1i},Y_{2i},\varepsilon_i)\mathbb{1}_{Y_{1i}\leq t_1, Y_{2i}\leq t_2}.$$

On remarque que, par le choix de W^*, cette fonction vérifie l'identité

$$\mathbb{E}\left[\int \phi(t_1,t_2)dF^*(t_1,t_2)\right] = \mathbb{E}[\phi(T_1,T_2)],$$

pour toute fonction ϕ d'espérance finie. Nous avons établi le résultat suivant.

Théorème 5. Sous certaines conditions (voir les Hypothèses 2 et 3 de la Section 2.3.1), il existe une classe de fonctions $\mathcal{F}$ telle que, pour $\phi \in \mathcal{F}$, on a

$$\int \phi(t_1,t_2)d(F_n - F^*)(t_1,t_2) = \frac{1}{n}\sum_{i=1}^{n} \psi_\phi(Y_{1i}, Y_{2i}, \varepsilon_i) + R_n(\phi), \tag{1.24}$$

avec

$$\sup_{\phi\in\mathcal{F}} |R_n(\phi)| = o_P(n^{-1/2}) \text{ et } E[\psi_\phi(Y_1,Y_2,\varepsilon)] = 0.$$

Par conséquent, pour toute fonction ϕ vérifiant l'Hypothèse 2 de la Section 2.3.1,

$$n^{1/2}\left(\int \phi(t_1,t_2)dF_n(t_1,t_2) - \mathbb{E}[\phi(T_1,T_2)]\right) \rightsquigarrow \mathcal{N}(0,\sigma_\phi^2), \tag{1.25}$$

où σ_ϕ^2 est donné par

$$\sigma_\phi^2 = \mathbb{E}\left[\left\{\frac{\delta_1\delta_2\phi(Y_1,Y_2)}{S^{C_1}(\min(C_1,A)-)} - E[\phi(T_1,T_2)] + \psi_\phi(Y_1,Y_2,\varepsilon)\right\}^2\right].$$

et $\rightsquigarrow$ désigne la convergence faible.

La forme exacte des termes de la somme (1.24) sera explicitée dans la section correspondante. Il est important que ces termes sont i.i.d. de l'espérance nulle, ce qui permet de déduire la normalité asymptotique des intégrales par rapport à la mesure définie par l'estimateur $F_n(t_1,t_2)$.

Ce résultat est prouvé sous les Hypothèses 2 et 3 de la Section 2.3.1. La première détermine la classe fonctionnelle pour la fonction ϕ. Elle est nécessaire pour appliquer des outils de la théorie des processus empiriques indexés par les classes de fonctions. Plus précisement, l'Hypothèse 2 exige que la fonction ϕ soit associée à une certaine classe fonctionnelle possédant la propriété de Donsker (voir [van der Vaart and Wellner, 1996] pour la définition des classes de Donsker). Cette propriété permet d'appliquer le théorème central limite uniforme sur la classe de fonctions et assure que le reste $R_n(\phi)$ soit asymptotiquement négligeable uniformément sur la classe fonctionnelle de ϕ.

L'Hypothèse 3 est nécessaire uniquement dans le cas où l'on souhaite obtenir la convergence sur tout le support de la distribution. Elle représente une condition de moments et peut être interprétée de façon informelle comme une condition empêchant une censure trop importante dans les queues des distributions.

Application à l'estimation du tau de Kendall et de la fonction $K(z)$

Le tau de Kendall. Le tau de Kendall peut être exprimé à partir de la fonction de répartition (voir, par exemple, [Nelsen, 2006]) sous la forme

$$\tau = 4 \int F(t_1, t_2) dF(t_1, t_2) - 1.$$

A partir de $F_n(t_1, t_2)$, il est naturel de définir un estimateur suivant de τ :

$$\hat{\tau} = 4 \int F_n(t_1, t_2) dF_n(t_1, t_2) - 1.$$

Sa normalité asymptotique est un corollaire du théorème de la représentation i.i.d. (voir le Corollaire 1 de la Section 2.3.1).

La fonction $K(z)$. Si $F(t_1, t_2)$ et $S_F(t_1, t_2)$ sont, respectivement, la fonction de répartition et la fonction de survie de (T_1, T_2), la fonction $K(z)$ de Kendall s'exprime comme

$$K(v) = \int \mathbf{1}_{F(t_1,t_2) \leq v} dF(t_1, t_2).$$

Nous en définissons un estimateur non paramétrique suivant :

$$K_n(v) = \int \mathbf{1}_{F_n(t_1,t_2) \leq v} dF_n(t_1, t_2), \tag{1.26}$$

A la Proposition 1 de la Section 2.3.3 , on montre que le processus empirique de Kendall, i.e. $n^{1/2} (K_n(v) - K(v))$ converge en distribution vers un processus gaussien de moyenne nulle.

Application aux tests d'adéquation

On rappelle que, par le théorème de Sklar,

$$S_F(t_1, t_2) = \mathfrak{C}(S_{T_1}(t_1), S_{T_2}(t_2)),$$

avec $S_{T_1}(t_1)$ et $S_{T_2}(t_2)$ les fonctions de survie marginales et $\mathfrak{C}(u, v)$ la copule de survie.

Soit $\mathcal{C}_\Theta = \{\mathfrak{C}_\theta : \theta \in \Theta\}$ une classe de copules archimédiennes. On note $\varphi_\theta(t)$ le générateur de la copule $\mathfrak{C}_\theta$. Nous souhaitons confronter les deux hypothèses suivantes :

$$H_0 : \ \mathfrak{C} \in \mathcal{C}_\Theta \qquad \text{contre} \qquad H_1 : \ \mathfrak{C} \notin \mathcal{C}_\Theta.$$

Nous rappelons également que, pour les copules archimédiennes,

$$K(v) = v - \frac{\varphi_\theta(v)}{\varphi'_\theta(v)}. \tag{1.27}$$

Le test que nous proposons est une extension, en présence de censure, de la méthode de [Wang and Wells, 2000b]. On considère la statistique de test suivante :

$$d(K_n, K_{\theta_n}) = \left[n \int_0^1 \left(K_n(v) - K_{\theta_n}(v) \right)^2 dv \right]^{1/2},$$

où $K_n(v)$ et K_{θ_n} sont respectivement définis par (1.26) et

$$K_{\theta_n} = v - \frac{\varphi_{\theta_n}(v)}{\varphi'_{\theta_n}(v)},$$

où θ_n est un estimateur paramétrique de la copule obtenu, par exemple, par la méthode de [Shih and Louis, 1995].

Dans la Section 2.4.2, nous considérons une application de ce test d'adéquation sur des données d'un assureur canadien portant sur les durées de vie des conjoints. Cette base de données contient plus de 90% d'observations censurées. Par conséquent, l'utilisation des méthodes adaptées à la présence de censure est cruciale pour son traitement.

La dernière remarque consiste à noter que la forme de la variance asymptotique pour la statistique de test est complexe et, pour calculer la valeur critique du test, il est nécessaire d'utiliser une procédure de bootstrap. Nous proposons donc dans la Section 2.3.2 une telle procédure, qui généralise la méthode de [Genest et al., 2006] en présence de censure.

1.6.2 Estimation non paramétrique de copule en présence de censure

Nous considérons les variables aléatoires T_1 et T_2 de fonction de répartition jointe $F(t_1, t_2)$, liées par une copule $\mathfrak{C}$ et sujettes aux censures aléatoires C_1 et C_2. Les lois marginales de F seront notées F_1 et F_2. L'objectif de cette section est de construire un estimateur non paramétrique de $\mathfrak{C}$ à partir de l'échantillon

$$(Y_{1i}, Y_{2i}, \delta_{1i}, \delta_{2i})_{1 \leq i \leq n},$$

éventuellement complété par les observations d'une variable auxiliaire (on fait notamment allusion au Modèle 2 où l'on observe la différence entre les deux censures).

Estimateur discret de la copule

Dans toute la suite nous supposons que $F_n(t_1, t_2)$ est un estimateur de F de la forme (1.6) décrite à la Section 1.3. On définit un estimateur non paramétrique de la copule par

$$\mathfrak{C}_n(u, v) = F_n(F_{1n}^{-1}(u), F_{2n}^{-1}(v)). \tag{1.28}$$

Cet estimateur peut être interprété comme une extension de la copule empirique au cadre des données censurées. Dans les Sections 3.3.2 et 3.3.3 du Chapitre 3, nous montrons les deux résultats de convergence suivants pour cet estimateur.

Théorème 6. Soit $\mathcal{T}_1 = [-\infty, A_1]$, et $\mathcal{T}_2 = [-\infty, A_2]$, tels que

$$\sup_{t_1 \in \mathcal{T}_1, t_2 \in \mathcal{T}_2} |F_n(t_1, t_2) - F(t_1, t_2)| = O_P(n^{-1/2}),$$

et tels que, pour $j = 1, 2$,

$$\sup_{t \in \mathcal{T}_j} |F_{jn}(t) - F_j(t)| = O_P(n^{-1/2}).$$

Alors, pour u_1 et u_2 définis par $F_1(\mathcal{T}_1) = [0, u_1]$, and $F_2(\mathcal{T}_2) = [0, u_2]$, et pour tout $\eta > 0$, on a

$$\sup_{u \leq u_1 - \eta, v \leq u_2 - \eta} |\mathfrak{C}_n(u, v) - \mathfrak{C}(u, v)| = O_P(n^{-1/2}). \tag{1.29}$$

Il est à noter, que la présence de censure peut rendre les estimateurs de la distribution non consistants au voisinage de la borne supérieure du support. C'est pour cette raison que la convergence dans (1.29) ne peut pas être démontrée sur $[0, 1]^2$ tout entier sans hypothèse supplémentaire sur les queues de distribution. L'utilisation de telles hypothèses, spécifiques au modèle considéré, permet en général d'étendre la convergence de l'estimateur de la distribution sur tout le support. Comme le montre le théorème suivant, l'estimateur de la copule devient alors consistant sur $[0, 1]^2$.

Théorème 7. Lorsque le processus $n^{1/2}(F_n(t_1, t_2) - F(t_1, t_2))$ converge faiblement vers un processus gaussien $\mathbb{G}_F(t_1, t_2)$ dans $l^\infty(\mathbb{R}^2)$, le processus $n^{1/2}(\mathfrak{C}_n(u, v) - C(u, v))$ converge dans $l^\infty([0, 1]^2)$ vers

$$\mathbb{Z}_{\mathfrak{C}}(u, v) = \mathbb{Z}^*_{\mathfrak{C}}(u, v) - \partial_1 \mathfrak{C}(u, v)\mathbb{Z}^*_{\mathfrak{C}}(u, 1) - \partial_2 \mathfrak{C}(u, v)\mathbb{Z}^*_{\mathfrak{C}}(1, v),$$

où $\mathbb{Z}^*_{\mathfrak{C}}(u, v) = \mathbb{G}_F(F_1^{-1}(u), F_2^{-1}(v))$.

Le Théorème 7 montre également que, si l'estimateur bivarié $F_n(t_1, t_2)$ vérifie le théorème central limite fonctionnel, alors l'estimateur de copule le vérifie aussi, et ceci indépendamment de la méthode de construction de F_n.

Enfin, on remarque que les modèles particuliers 1 à 3 de la Section 1.3 vérifient les conditions du Théorème 7. De plus, le Modèle 4, qui est le plus général, satisfait les hypothèses du Théorème 6.

Estimateurs lisses de copule

L'estimateur défini par (1.28) est une fonction discrète. Néanmoins, comme nous l'avons évoqué dans la section consacrée à la copule empirique, la plupart des copules usuelles possèdent la densité. Il est donc naturel de les approcher par des estimateurs lisses. Ces estimateurs permettent également d'estimer la densité de copule dont la visualisation peut éventuellement servir comme un outil graphique de la selection de modèles. Dans cette section, nous allons considérer deux approches à l'estimation par noyau de copule. La première est une adaptation, en présence de censure, de l'estimateur de [Fermanian et al., 2004] que l'on a présenté dans la Section 1.4.3. La deuxième utilise une autre technique, proposée dans le cadre non censuré par [Omelka et al., 2009].

Soit $k : \mathbb{R} \mapsto \mathbb{R}$ un noyau (i.e. une fonction positive, intégrable et à valeurs réelles) et $K(x) = \int_{-\infty}^{x} k(u)du$ son intégrale.

- **Estimateur 1.** On introduit un estimateur lisse de la fonction de répartition,

$$\hat{F}^1_n(t_1, t_2) = \frac{1}{n} \sum_{i=1}^{n} W_n K_h\left(t_1 - Y_{1i}\right) K_h(t_2 - Y_{2i}), \tag{1.30}$$

  où l'on a noté $K_h(x) := K(x/h)$, et on omet la dépendance précise de W_n de ses arguments. On définit l'estimateur par noyau de la copule donné par

$$\hat{\mathfrak{C}}^1_n(u, v) = \hat{F}^1_n((\hat{F}^1_{1n})^{-1}(u), (\hat{F}^1_{2n})^{-1}(v)), \tag{1.31}$$

où $\hat{F}^1_{1n}$ et $\hat{F}^1_{2n}$ sont les lois marginales de $\hat{F}^1_n$.

- **Estimateur 2.** Comme le mentionnent [Omelka et al., 2009], un inconvénient de l'estimateur (1.31) est la dépendance de sa performance des lois marginales. Ils proposent donc de standardiser les marginales par une transformation de variables, à l'aide d'une fonction de répartition $\Phi(x)$ choisie de sorte que les fonctions Φ' et $(\Phi')^2/\Phi$ sont bornées. On introduit le couple des variables aléatoires

$$(\tilde{T}_1, \tilde{T}_2) = (\Phi^{-1}[F_1(T_1)], \Phi^{-1}[F_2(T_2)]),$$

et l'estimateur suivant de leur fonction de répartition jointe :

$$\hat{F}^2_n(t_1, t_2) = \frac{1}{n}\sum_{i=1}^{n} W_n K_h\left(t_1 - \Phi^{-1}[F_{1n}(Y_{1i})]\right) K_h(t_2 - \Phi^{-1}[F_{2n}(Y_{2i})]\Big),$$

où F_{1n} et F_{2n} sont les lois marginales de l'estimateur (1.6). Puisque la copule est invariante par transformation monotone, les variables $\tilde{T}_1$ et $\tilde{T}_2$ sont liées par la même copule que T_1 et T_2, mais elles ont toutes les deux comme marginales la fonction $\Phi(x)$. Cela nous conduit à la définition du deuxième estimateur lisse de $\mathfrak{C}$:

$$\hat{\mathfrak{C}}^2_n(u, v) = \hat{F}^2_n(\Phi^{-1}(u), \Phi^{-1}(v)). \tag{1.32}$$

Sous certaines conditions (voir le Chapitre 3), on démontre pour les deux estimateurs les résultats de convergence faible énoncés dans le théorème suivant.

Théorème 8. Pour $i = 1, 2$, le processus empirique

$$n^{1/2}(\hat{\mathfrak{C}}^i_n(u, v) - \mathfrak{C}(u, v)), \qquad 1 \leq u, v \leq 1,$$

converge dans $l^\infty([0, 1]^2)$ vers le processus limite $\mathbb{Z}_{\mathfrak{C}}(u, v)$, défini dans l'énoncé du Théorème 7.

Application à l'estimation de la densité de copule

Une première application des estimateurs introduits dans les sections précédentes est l'estimation de la densité de copule. Les estimateurs de la densité se déduisent des estimateurs lisses de la copule par la relation :

$$\hat{c}^i(t_1, t_2) = \frac{\partial^2}{\partial_{t_1}\partial_{t_2}} \hat{\mathfrak{C}}^i_n(t_1, t_2), \tag{1.33}$$

for $i = 1, 2$.

Sous les hypothèses explicitées dans la Section 3.4.3 du Chapitre 3, nous obtenons, pour ces estimateurs, les vitesses de convergence uniforme sur les compacts strictement inclus dans $[0, 1]^2$. Plus particulièrement, nous démontrons le résultat suivant.

Théorème 9. Soit $\mathcal{C}$ un compact strictement inclu dans $[0,1]^2$, et soit

$$\eta_n = h^2 + \frac{[\log n]^{1/2}}{hn^{1/2}}.$$

On a les deux assertions suivantes.

- Si $nh^2[\log n]^{-1} \to \infty$, et, de plus, s'il existe $\alpha > 0$ tel que $hn^\alpha \to 0$, alors

$$\sup_{(u,v)\in\mathcal{C}} |\hat{c}_1(u,v) - c(u,v)| = O_P(\eta_n). \tag{1.34}$$

- Pour tout h tel que $nh^{10/3} \to \infty$, on obtient

$$\sup_{(u,v)\in\mathcal{C}} |\hat{c}_2(u,v) - c(u,v)| = O_P(\eta_n). \tag{1.35}$$

Applications aux données réelles et simulées

Dans la **première partie** de l'étude numérique détaillée dans le Chapitre 3, nous avons évalué des nouveaux estimateurs sur les données simulées, selon le schéma suivant :

- Les simulations des échantillons de données bivariées à partir d'une copule $\mathfrak{C}(u,v)$ connue (Clayton, Frank ou Gumbel).
- L'introduction de la censure bivariée dans les données, avec des pourcentages variés.
- L'évaluation des estimateurs non paramétriques de copule à partir des échantillons censurés.
- L'étude de la distance de Kolmogorov-Smirnov et de la distance quadratique intégrée entre ces estimateurs et la copule $\mathfrak{C}(u,v)$ à l'origine des données.

Pour la description détaillée des simulations et des résultats, nous renvoyons le lecteur vers la Section 3.5.1 du Chapitre 3.

Dans la **deuxième partie** de l'étude numérique, nous considérons une application des estimateurs non paramétriques de copule en présence de censure aux tests d'adéquation pour deux jeux de données réelles, à savoir les données d'un assureur canadien avec les observations de durées de vie des conjoints et les données ALAE qui portent sur les coûts dépendants de sinistres.

1.6.3 Quantification en présence de censure et une application au clustering

En absence de censure, nous avons défini un quantificateur empirique optimal comme celui qui minimise la distorsion par rapport à la loi empirique, basée sur les réplications i.i.d. de même loi que le vecteur à quantifier. Nous avons également vu que le problème numérique de la minimisation de la distorsion empirique était NP-complet. Par conséquent, l'évaluation des centres de clusters nécessite, en pratique, un recours à un algorithme itératif.

En présence de censure, des difficultés se retrouvent aux deux niveaux. D'une part, comme nous l'avons déjà vu, les réplications i.i.d. de vecteur d'intérêt sont indisponibles. Par conséquent, la définition habituelle de la distorsion empirique n'est pas utilisable, et la méthode classique de la quantification ne peut pas être appliquée. D'autre part, les distances euclidiennes entre les observations censurées et les centres de clusters ne sont pas observées, ce qui pose un problème à l'algorithme des k-means standard basé sur ces distances.

Dans cette section, nous allons étendre les procédures classiques de la quantification et du clustering au cas d'un vecteur aléatoire dont une composante est censurée et toutes les autres sont observées (voir le contexte 1 de la Section 1.3). Cela correspond aux situations pratiques où une des variables porte un caractère temporel et est sujette à la censure.

Notre premier objectif consistera à proposer une généralisation des définitions de la distorsion empirique et de quantificateur empirique optimal, pour pouvoir quantifier le vecteur inobservé (T, X) à partir de réplications i.i.d. $(Y_i, \delta_i, X_i)_{1\leq i\leq n}$ de vecteur observé (Y, δ, X). Cela nous permettra de construire un critère empirique dont la minimisation fournit les centres d'un quantificateur empirique optimal. De même que dans le cadre non censuré, le problème numérique associé est NP-complèt.

Notre deuxième objectif consistera à proposer un algorithme itératif permettant d'évaluer numériquement les centres de clusters et d'associer des labels à toutes les observations, à partir d'un échantillon censuré. Comme on l'a déjà mentionné, la difficulté principale de cette démarche est que toutes les distances associées aux observations censurées ne sont pas observées.

Quantification pour un vecteur à une composante censurée

Nous nous plaçons dans le contexte 1 de la Section 1.3 et nous supposons que les conditions de la consistance de l'estimateur (1.5) sont vérifiées. On note P la loi de (T, X) que l'on souhaite quantifier. Comme nous l'avons vu, la définition de la distorsion empirique fait intervenir la loi empirique, basée sur les réplications i.i.d. indisponibles de (T, X). A leur place, nous disposons de réplications i.i.d. $(Y_i, \delta_i, X_i)_{1\leq i\leq n}$ de vecteur observé (Y, δ, X). Nous rappelons que, dans le cadre non censuré, la mesure empirique servait à estimer la loi P. Dans le contexte censuré, notre approche consistera à estimer P par la mesure, engendrée par l'estimateur de Stute (1.5), et à définir par rapport à elle la distorsion empirique.

En présence de censure, le déficit de grandes observations induit les difficultés de l'estimation dans la queue de distribution et peut rendre l'estimateur inconsistant en voisinage de la borne supérieure du support. Pour palier cette difficulté, une approche qu'on retrouve fréquemment dans la littérature (voir, par exemple, [Heuchenne, 2008]) consiste à restreindre la variable étudiée à un compact $[0, \tau]$ strictement inclus dans le support de la distribution. Par ailleurs, en quantification, les résultats théoriques sur la vitesse de convergence de la distorsion d'un quantificateur empirique optimal nécessitent des variables bornées, ce qui n'est en général pas le cas pour une variable de durée. Cela est une raison supplémentaire pour utiliser la borne de troncation τ qui peut pour autant être choisie arbitrairement proche de la frontière supérieure du support. Cela revient donc à travailler avec une loi tronquée

$P^\tau := P_{(T,X)\|T\leq\tau}$. On introduit la mesure suivante :

$$\mathcal{P}_n^\tau = \frac{1}{n}\sum_{i=1}^n W_{in}^\tau \delta_{(Y_i,X_i)}, \text{ avec } \quad W_{in}^\tau = \frac{W_{in}\mathbb{1}_{Y_i\leq\tau}}{\sum_{i=1}^n W_{in}\mathbb{1}_{Y_i\leq\tau}}, \tag{1.36}$$

où W_{in} sont les poids de l'estimateur (1.5). La mesure (1.36) présente deux modifications par rapport à la mesure engendrée par l'estimateur de Stute. La première est liée à l'introduction de la troncature, alors que la deuxième rend la somme des poids égale à 1. On définit alors la distorsion empirique par

$$\begin{aligned}\mathcal{D}(\mathcal{P}_n^\tau, q) &= \frac{\sum_{i=1}^n W_{in}||(Y_i,X_i)-q(Y_i,X_i)||^2\mathbb{1}_{Y_i\leq\tau}}{n\sum_{i=1}^n W_{in}\mathbb{1}_{Y_i\leq\tau}} \\ &= \frac{1}{n}\sum_{i=1}^n W_{in}^\tau||(Y_i,X_i)-q(Y_i,X_i)||^2. \end{aligned}\tag{1.37}$$

Dans ce contexte, on appellera q_n^* un quantificateur empirique optimal s'il minimise (1.37). Pour les mêmes raisons que dans le cadre non censuré, q_n^* existe et est un quantificateur des plus proches voisins.

Les propriétés asymptotiques d'un quantificateur empirique optimal ainsi défini sont étudiées dans la Section 3 du Chapitre 4. Nous établissons d'abord la convergence presque sûre,

$$D(P^\tau, q_n^*) \xrightarrow[n\to\infty]{p.s.} D_k^*(P^\tau). \tag{1.38}$$

Ensuite, nous étudions la vitesse de convergence. On montre que, sous la condition

$$P(||(T,X)||\leq R)=1, \quad R\in]0,\infty[$$

il existent des constantes positives universelles K, K_1, K_2, L_1, L_2 telles que, pour tout $z > 4K/F_\tau^T$, avec $F_\tau^T = P(T\leq\tau)$,

$$\begin{aligned}P(\sqrt{n}|D(P^\tau,q_n^*)-D(P^\tau,q^*)|>z) &\leq 5\exp\left(-L_1z^2+L_2z\right) \\ &+ 2\left[\exp(-K_1z^2)+\exp(-\sqrt{n}K_2z)\right] \\ &+ O\left(e^{-\sqrt{n}}\right),\end{aligned}$$

où le reste $O\left(e^{-\sqrt{n}}\right)$ ne dépend pas de z.

Idée de la preuve. L'idée principal de la preuve de ce théorème est de borner la différence entre deux distorsions par une déviation maximale d'un certain processus indexé par une classe de fonctions avec la propriété de Donsker. L'inégalité exponentielle découle alors d'une inégalité de concentration de [Talagrand, 1994]. Une des difficultés principales est que q_n^* est optimal par rapport à la mesure aléatoire $\mathcal{P}_n^\tau$, dont les poids dépendent de l'estimateur de Kaplan-Meier $\hat{G}(y)$ de la variable de censure. Techniquement, pour gérer cette mesure, on remplace $\hat{G}(y)$ par sa limite $G(y)$. Pour ce faire, il est nécessaire de contrôler $\sup_y|\sqrt{n}(\hat{G}(y)-G(y))|$, où le supremum est pris sur des ensembles qui ne dépendent pas de z. Cela est possible grace à une inégalité exponentielle pour l'estimateur de Kaplan-Meier, démontrée dans [Bitouzé et al., 1999].

Algorithme de clustering

Soient $(T_i, X_i)_{1\leq i\leq n}$ les résultats de mesures de (T, X) sur n individus. En présence de censure, nous observons uniquement $(Y_i, \delta_i, X_i)_{1\leq i\leq n}$, i.e. pour certains individus à la place de la "vrai" réalisation (T_i, X_i) nous disposons uniquement de (Y_i, δ_i, X_i). On s'interroge alors sur la manière de séparer les individus en clusters par rapport aux réalisations d'intérêt $(T_i, X_i)_{1\leq i\leq n}$ en disposant uniquement de leurs versions censurées.

Nous rappelons que le problème de clustering est naturellement lié à la quantification. En effet, la quantification fournit une façon optimale de résumer la loi d'un vecteur aléatoire par k points (centres) et la règle de quantification (i.e. la règle des plus proches voisins). Par conséquent, pour effectuer le clustering des observations, il suffit d'évaluer d'abord les centres optimaux, en minimisant la distorsion empirique, et attribuer ensuite à chaque observation le label du cluster dont le centre est le plus proche. D'un point de vue algorithmique, le problème de la minimisation est NP-complet et nécessite donc un algorithm itératif permettant d'approcher numériquement les centres.

Dans notre situation, on a défini la distorsion empirique pour un vecteur à une composante censurée. Sa minimisation nous permettra de trouver les centres de clusters, en proposant une procédure itérative adaptée à un échantillon censuré. Néanmoins, en présence de censure, cette étape ne suffit pas. En effet, pour les observations censurées, leur distances aux centres de clusters ne sont pas observées. Par conséquent, il est impossible de leur attribuer un label. Nous proposons donc un algorithme en deux étapes :

- L'Étape 1 est une procédure itérative qui sert à approcher les centres minimisant la distorsion (1.37) et à attribuer les labels aux observations non censurées.
- L'Étape 2 a pour l'objectif d'estimer les distances aux centres des clusters pour chaque observation censurée, et de lui attribuer un label de cluster dont le centre est le plus proche par rapport aux distances estimées.

Idée de l'Étape 1. Dans l'algorithme des k-means standard, à chaque iteration, le centre de chaque cellule S est actualisé par la moyenne des observations associées à cette cellule. Cette moyenne est en fait un estimateur de

$$\mathbb{E}\left[(T, X)|(T, X) \in S\right]. \tag{1.39}$$

Dans le cadre non censuré, chaque observation y contribue avec le poids $1/n$. En présence de censure, l'idée même de l'estimateur de Kaplan-Meier consiste à attribuer les poids uniquement aux observations non censurées et de distribuer la masse totale entre elles de sorte à compenser le biais causé par la censure. Par la même logique, notre procédure actualisera le centre de la cellule S par

$$c = \frac{\sum_{i=1}^{n}(Y_i, X_i)^T W_{in} \mathbb{1}_{\{(Y_i,X_i)\in S, \delta_i=1\}}}{\sum_{i=1}^{n} W_{in} \mathbb{1}_{\{(Y_i,X_i)\in S, \delta_i=1\}}}, \tag{1.40}$$

où c est un estimateur de (1.39) qui se base uniquement sur les observations non censurées appartenant à la cellule, mais qui tient compte de leurs poids compensant

la censure. Le reste de la procédure sera effectué de la même façon que dans l'algorithme des k-means. A la fin de l'étape, les centres de clusters sont déterminés et les observations non censurées sont associées à leurs centres les plus proches.

Étape 1 (Évaluation des k centres).

- Initialiser les coordonnées des centres $c_1^{(0)}, \ldots, c_k^{(0)}$
- Calculer les poids W_{in} de l'estimateur de Kaplan-Meier pour l'échantillon $(Y_i, X_i, \delta_i)_{1 \leq i \leq n}$
- **Répeter jusqu'à ce que rien ne change :** pour une iteration ℓ
  - Calculer les cellules de Voronoï $S_1^\ell, \ldots, S_k^\ell$ qui correspondent aux centres $c_1^{(\ell)}, \ldots, c_k^{(\ell)}$ pour l'ensemble des observations non censurées $\{(Y_i, X_i) : \delta_i = 1, i = 1, \ldots, n\}$
  - pour $j = 1, \ldots, k$ calculer les nouveaux centres $(c_j^{(\ell+1)})_{1 \leq j \leq k}$:

$$c_j^{(\ell+1)} = \frac{\sum_{i=1}^n (Y_i, X_i)^T W_{in} \mathbb{1}_{\{(Y_i,X_i) \in S_j^l, \delta_i = 1\}}}{\sum_{i=1}^n W_{in} \mathbb{1}_{\{(Y_i,X_i) \in S_j^l, \delta_i = 1\}}}.$$

  - L'algorithme s'arrête en nombre fini ℓ^* des iterations. Pour $j = 1, \ldots, k$ attribuer à l'observation (Y_i, X_i) avec $\delta_i = 1$ le label j si $(T_i, X_i) \in S_j^{\ell^*}$.

Remarque. Les résultats théoriques nécessitent l'introduction de la borne de troncation τ. Néanmoins, elle peut être choisie arbitrairement proche de la frontière supérieure du support. Nous considérons donc comme admissible en pratique de la prendre égale à cette dernière.

Idée de l'Étape 2. Lorsqu'une observation i est censurée, on n'observe pas (T_i, X_i), et on sait seulement que $\delta_i = 0$ et que $T_i > Y_i$. La meilleure approximation de la distance inobservée $d((T_i, X_i); c)$ de cette observation au centre c est alors donnée par :

$$\mathbb{E}\left[d((T_i, X_i); c) \,|X_i, T_i > Y_i, \delta_i = 0\right]$$

Ainsi, pour chaque $i = 1, \ldots, n$, tel que $\delta_i = 0$ nous estimons la distance de la réalisation non observée (T_i, X_i) au centre $c_j^{(\ell*)}$ par un estimateur suivant de ce cette espérance conditionnelle :

$$\hat{d}_{ij} = \frac{\int_{Y_i}^\infty ||(t, X_i) - c_j^{(\ell*)}||^2 d\hat{F}(t|X_i)}{\int_{Y_i}^\infty d\hat{F}(t|X_i)}, \tag{1.41}$$

où $\hat{F}(t|x)$ est un estimateur de $F(t|X = x) = P(T \leq t|X = x)$ donné par

$$\hat{F}(t|x) = \frac{1}{n} \sum_{i=1}^n W_{in} \frac{k(\frac{x - X_i}{h})}{\sum_{j=1}^n k\left(\frac{x - X_j}{h}\right)} \mathbb{1}_{Y_i \leq t}, \tag{1.42}$$

avec un noyau $k(x)$, i.e. une fonction positive intégrable telle que $\int_{\mathbb{R}^d} k(x)dx = 1$.

En combinant (1.41) et (1.42), on obtient

$$\hat{d}_{ij} = \frac{\sum_{m=1}^{n} W_{mn} ||(Y_m, X_i) - c^{(l*)}||^2 k\left(\frac{X_i - X_m}{h}\right) \mathbb{1}_{Y_m \geq Y_i}}{\sum_{m=1}^{n} W_{mn} k(\frac{X_i - X_m}{h}) \mathbb{1}_{Y_m \geq Y_i}}. \tag{1.43}$$

Nous allons maintenant présenter la deuxième étape de notre algorithme :

Étape 2 (Attribution des labels aux observations censurées). Pour attribuer les labels aux observations censurées :

- Pour chaque observation censurée (Y_i, X_i) calculer les distances estimées $\hat{d}_{ij}$ en utilisant (1.43).
- Attribuer à (Y_i, X_i) le label $j^* = \arg\min_j \hat{d}_{ij}$.

La méthode exposée dans cette section est décrite de façon détaillée dans le Chapitre 4, où nous donnons également quelques résultats de son applications aux données.

Chapitre 2

A simplified model for studying bivariate mortality under right-censoring

Abstract In this chapter, we provide a nonparametric estimator of the distribution of bivariate censored lifetimes, in a model where the two censoring variables differ only through an additional observed variable. This situation is motivated by a particular application to insurance, where the supplementary variable corresponds to the age difference between two individuals. Asymptotic results for our estimator are provided. The new tools that we develop are used to perform goodness-of-fit tests for survival copula models. The practical performance is illustrated through simulations and a real data analysis.

This chapter corresponds to the article [Gribkova et al., 2013] published in *Journal of Multivariate Analysis*.

2.1 Introduction

In last survivor insurance, an important issue is to infer on the joint distribution of the lifetimes of two individuals linked through an insurance contract, say (T, U). One of the difficulties in studying such variables comes from the presence of bivariate censoring, with a proportion of censored observations which may be quite high. Therefore, most of the approaches used in this field are parametric (typically parametric survival copula models, see e.g. [Shih and Louis, 1995]), while nonparametric tools are rarely used, although they would be required at least to assess the validity of the proposed models. Our aim is to provide a new nonparametric estimator of the joint distribution of two lifetimes under bivariate random censoring, in a framework which is adapted to the study of problems coming from the insurance field. A specificity of such problems is that an additional variable, which carries information on the model, is generally present, this variable being the age difference between the two individuals under study. Using this information that is often neglected, one can define a quite simple nonparametric estimator of the distribution of the two

lifetimes, which is close to the Kaplan-Meier estimator ([Kaplan and Meier, 1958]) and to the estimator of [Lin and Ying, 1993].

Various approaches have been used to perform nonparametric estimation of multivariate lifetimes. Most of them focus on estimating the survival function, without focusing on the joint distribution itself. Therefore, many of them provide consistent estimators of this function, but fail to define a true distribution. For example, the estimator of the survival function proposed by [Campbell and Földes, 1982] is not monotonic. The nonparametric maximum likelihood (NPMLE) procedure of [Hanley and Parnes, 1983] leads to an estimator which is sometimes inconsistent for continuous data [Tsai et al., 1986], while the rate of convergence of a modification of this estimator suggested by [Tsai et al., 1986] achieves a slow convergence rate (slower that $n^{1/2}$ where n denotes the sample size). Another NPMLE approach is proposed by [van der Laan, 1996], introducing some modification of the data and using an interval censoring methodology. Although this estimator is shown to be asymptotically efficient for these modified data, the convergence rate is also slower than $n^{1/2}$. On the other hand, the product-limit type estimator proposed by [Dabrowska, 1988], which is often used in practice (see e.g. [Luciano et al., 2008], [Fan et al., 2000], [Wang and Wells, 2000b], [Gill et al., 1995], [Prentice and Cai, 1992]), assigns negative mass to some points on the plane [Pruitt, 1991a]. Nonparametric smoothing techniques have been used by [Pruitt, 1991b] (but the implicit definition of the estimator leads to difficulties and a weak performance according to [van der Laan, 1996]), and by [Akritas and Van Keilegom, 2003] (this last estimator presenting the advantage of defining a true distribution, but, again, with a slower convergence rate, and the necessity of an absolutely continuous censoring variables).

Among the approaches that we mention, each of them suffers either from a too slow convergence rate, or from the fact that the corresponding estimators do not provide true probability distributions. The estimator proposed by [Lopez and Saint-Pierre, 2012] does not present these drawbacks, but relies on an assumption on the joint distribution of the censoring variables which may not be reasonable for the particular application we have in mind. Indeed, a specificity of data-sets coming from last-survivor insurance, is that individuals usually quit the study (for some cause other than death) at the same time. This induces a specific dependence between the two censoring variables involved in the problem. The main idea of the new estimator that we propose consists of using this additional information.

The rest of the chapter is organized as follows. In section 2.2, we present the general censoring framework that we consider. We define a non parametric estimator of the distribution of (T, U). In section 2.3, we provide asymptotic results for estimating quantities of the type $E[\phi(T, U)]$ for a large class of functions ϕ (the survival function being only a particular case). A bootstrap procedure is proposed to compute the variance of the error in such estimation problems. Application of our nonparametric estimator to goodness-of-fit for copula models is considered. Section 2.4 illustrates our result through simulation studies and a real data analysis.

2.2 A simplified model for bivariate censoring

We first present in section 2.2.1 our bivariate right-censoring model. Estimation of the joint distribution of (T, U) is introduced in section 2.2.2.

2.2.1 Bivariate right-censoring

In the following, we consider two lifetimes (T, U), and i.i.d. replications $(T_i, U_i)_{1\leq i\leq n}$ of these random variables. In a bivariate right-censoring model, $(T_i, U_i)_{1\leq i\leq n}$ are not directly observed. Instead, one observes

$$\begin{cases} Y_i &= \inf(T_i, C_i), \text{ and } \delta_i = \mathbf{1}_{T_i\leq C_i} \\ Z_i &= \inf(U_i, D_i), \text{ and } \gamma_i = \mathbf{1}_{U_i\leq D_i}, \end{cases}$$

where $(C_i, D_i)_{1\leq i\leq n}$ consist of i.i.d. replications of a random bivariate censoring vector (C, D), and $(\delta_i, \gamma_i)_{1\leq i\leq n}$ are indicator functions allowing the distinction between censored and uncensored observations.

In many applications, such as last survivor insurance, there exists some clear relationship between the two censoring variables (C, D). In this particular case, T (resp. U) denotes the total lifetime of the husband (resp. his wife) and C (resp. D) denotes the age at which the husband (resp. the wife) stops being under observation for any cause other than death.

Usually, censoring causes are twofold : the end of the observation period (if the person is not dead at this time, his/her lifetime is not observed), or the surrender of the contract. In both situations, one can observe that, for many cases, this event which stops observation occurs at the same time for both members of the couple. Taking the example of a pension contract with a reversion clause, one can see that surrendering the contract will automatically remove the two members of the couple from the database (unless one of them possesses additional contracts that could allow the company to keep some track on him/her, which is usually not the case due to the complexity of such a tracking process). If ε denotes the age difference between the two members of the couple, then $D = C + \varepsilon$. Moreover, the variables $(\varepsilon_i)_{1\leq i\leq n}$ are observed for all couples.

To summarize, in such a framework, observations are made of

$$(Y_i, Z_i, \varepsilon_i, \delta_i, \gamma_i)_{1\leq i\leq n},$$

where the random variable ε represents the age difference between the two observed persons. We now state some identifiability assumptions.

Assumption 1. Assume that

1. $D = C + \varepsilon$.
2. (T, U) is independent from ε, and from C, and $\mathbb{P}(T = C) = \mathbb{P}(U = C+\varepsilon) = 0$.
3. C is independent from ε.

In Assumption 1, points 2 and 3 are a direct multivariate extension of the classical identifiability assumption required to ensure the consistency of the Kaplan-Meier estimator in the univariate case (see [Stute and Wang, 1993]). In section 2.2.2 below, we show how one can estimate nonparametrically the distribution of (T, U) under Assumption 1.

2.2.2 Nonparametric estimation of the distribution of (T, U)

In this section, we are interested in estimating quantities of the type $E[\phi(T,U)]$ for some function ϕ. A particular case is the joint survival function $S_F(t,u) = \mathbb{P}(T > t, U > u)$, which corresponds to $\phi(T,U) = \mathbf{1}_{T>t,U>u}$. In absence of censoring, the answer to this estimation problem consists of using the empirical means. Defining $F(t,u) = \mathbb{P}(T \leq t, U \leq u)$, one can rewrite $E[\phi(T,U)] = \int \phi(t,u)dF(t,u)$, which can be consistently estimated by $\int \phi(t,u)dF_{emp}(t,u)$, where $F_{emp}(t,u) = n^{-1}\sum_{i=1}^{n} \mathbf{1}_{T_i \leq t, U_i \leq u}$ denotes the empirical distribution function. In our framework, the empirical distribution in unfortunately unavailable, since $(T_i, U_i)_{1 \leq i \leq n}$ are not directly observed.

We propose to rely on an estimator of the type

$$\hat{F}(t,u) = \sum_{i=1}^{n} \delta_i \gamma_i W_n(Y_i, Z_i, \varepsilon_i) \mathbf{1}_{Y_i \leq t, Z_i \leq u}, \tag{2.1}$$

to generalize the empirical distribution function to our framework. Using such type of estimators, one can straightforwardly define an estimator of

$$E[\phi(T,U)] = \int \phi(t,u)dF(t,u)$$

by

$$\int \phi(t,u)d\hat{F}(t,u) = \sum_{i=1}^{n} \delta_i \gamma_i W_n(Y_i, Z_i, \varepsilon_i)\phi(Y_i, Z_i). \tag{2.2}$$

The idea is similar to [Lopez, 2012] and [Lopez and Saint-Pierre, 2012] : instead of assigning the same $n^{-1}-$mass to each observation (as it is the case when considering the empirical distribution function), one assigns mass to doubly uncensored observations (since only these observations are completely relevant to understand the dependence structure between T and U), while the mass $W_n(Y_i, Z_i, \varepsilon_i)$ is designed to compensate for censoring.

Let $G(t)$ be the cumulative distribution function of C and $S_G(t) = \mathbb{P}(C > t)$ the survival function, we define

$$F^*(t,u) = \sum_{i=1}^{n} \delta_i \gamma_i W^*(Y_i, Z_i, \varepsilon_i) \mathbf{1}_{Y_i \leq t, Z_i \leq u}, \tag{2.3}$$

with $W^*(y,z,e) = n^{-1} S_G(\max(y, z-e)-)^{-1}$ (where $S_G(u-)$ denotes the left limit of S_G at point u), one can observe that $\int \phi(t,u)dF^*(t,u)$ is an unbiased estimator of $E[\phi(T,U)]$ under Assumption 1. Indeed, for any function ψ with finite expectation, we have, under Assumption 1,

$$\begin{aligned} E[\delta\gamma\psi(Y,Z)] &= E\left[E\left[\mathbf{1}_{\max(T,U-\varepsilon)\leq C}|T,U,\varepsilon\right]\psi(T,U)\right] \\ &= E\left[S_G(\max(T,U-\varepsilon)-)\psi(T,U)\right]. \end{aligned}$$

Unfortunately, this ideal estimator F^* can not be computed in practice, since it relies on the unknown survival function S_G. Nevertheless, it is possible to estimate this function S_G. Define $\eta_i = 1 - \delta_i\gamma_i$, and $A_i = \max(T_i, U_i - \varepsilon_i)$. The variable C_i

is observed as long as $C_i < A_i$ (that is $\eta_i = 1$). Hence, C can also be considered as a right-censored variable (censored by the variable A), provided that the event $\{C_i = A_i\}$ has probability 0. This is actually the case, from point 2 in Assumption 1, where we assumed that $\mathbb{P}(T = C) = \mathbb{P}(U - \varepsilon = C) = 0$. Therefore, the distribution of C can be estimated by the Kaplan-Meier estimator based on the censored sample $(B_i, \eta_i)_{1 \leq i \leq n}$, where $B_i = \inf(C_i, A_i)$. Moreover, it is important to notice that Assumption 1 ensures consistency of the Kaplan-Meier estimator. Let us recall that following [Stute and Wang, 1993], $\mathbb{P}(A = C) = 0$ and A independent from C are the assumptions required to ensure this consistency.

Therefore, defining the Kaplan-Meier estimator $\hat{S}_G$ of S_G,

$$\hat{S}_G(t) = \prod_{k:B_k \leq t} \left(1 - \frac{d\hat{H}_0(B_k)}{\hat{H}(B_k)}\right),$$

where $\hat{H}_0(t) = n^{-1} \sum_{i=1}^n \eta_i \mathbf{1}_{B_i \leq t}$, and $\hat{H}(t) = n^{-1} \sum_{i=1}^n \mathbf{1}_{B_i \geq t}$, a natural choice of a function W_n in (2.1) is

$$W_n(y, z, e) = \frac{1}{n\hat{S}_G(\max(y, z - e)-)}. \tag{2.4}$$

This estimator is close to the estimator proposed by [Lin and Ying, 1993]. The difference, in our approach, is the presence of the additional random variable ε_i corresponding to the age difference.

2.3 Asymptotic theory

The present section is devoted to the asymptotic results on the nonparametric estimator defined in section 2.2.2. A Central Limit Theorem for (2.2) is provided in section 2.3.1. As a corollary of this result, we deduce asymptotic convergence properties when estimating Kendall's τ coefficient, which is a classical dependence measure. Section 2.3.2 provides a bootstrap procedure in order to investigate the estimation error. In section 2.3.3, we derive theoretical results that may be used to perform goodness-of-fit tests for survival copula models.

2.3.1 An asymptotic representation for estimator (2.1)

We aim to obtain an asymptotic representation for quantities of the type (2.1). Instead of considering a single function ϕ, we focus on obtaining results that hold uniformly for functions $\phi \in \mathcal{F}$, $\mathcal{F}$ denoting a class of functions. This uniformity result is required if we wish to obtain, for example, uniform consistency results for the estimation of the distribution function. The natural class of functions to be considered in this problem is $\mathcal{F}_1 = \{(t, u) \to \mathbf{1}_{t \leq x, u \leq y} : x \in \mathcal{T}, y \in \mathcal{U}\}$, where $\mathcal{T}$ and $\mathcal{U}$ denote the support of the distribution of each marginal.

In the following, we consider a class of functions $\mathcal{F}$, with envelope Φ, satisfying Assumptions 2 and 3 below. Since our proof will rely on empirical processes theory, Assumption 2 consists of assuming that a class of functions related to $\mathcal{F}$ is Donsker, that is a class with an uniform central limit theorem property (see [van der Vaart and Wellner, 1996] for a precise definition of Donsker classes).

Assumption 2. Let $\mathcal{G}$ denote the class of positive, monotonic functions bounded by 1, and $\chi(T,U,C,\varepsilon) = \delta\gamma S_G(\max(T,U-\varepsilon)-)^{-2}$. For any (t_0,u_0) in $\mathbb{R}^2$ such that $S_F(t_0,u_0) > 0$, define

$$\begin{aligned}\mathcal{H}_{t_0,u_0} &= \{(T,U,C,\varepsilon) \to \\ &\quad \to \chi(T,U,C,\varepsilon) f(T,U) g(\max(T,U-\varepsilon)-)\mathbf{1}_{T\le t_0, U\le u_0},\ f\in\mathcal{F},\ g\in\mathcal{G}\},\end{aligned}$$

and assume that $\mathcal{H}_{t_0,u_0}$ is a Donsker class of functions.

Assumption 3 is required only if we wish to obtain the consistency on the whole support of (T,U). It automatically holds if one considers bounded functions with compact support strictly included in the support of the distribution. It can be understood as an assumption on the tail of the distributions of T and U. We remark that similar assumptions have been used in [Gill, 1983], [Stute, 1996], or [Lopez and Saint-Pierre, 2012].

Assumption 3. Assume that $E\left[\Phi(T,U)^2 S_G(\max(T,U-\varepsilon)-)^{-1}\right] < \infty$. Moreover, let $F_A(t) = \mathbb{P}(A \le t)$ and define

$$\mathcal{C}(y) = \int_{-\infty}^{y} \frac{dG(t)}{[1-F_A(t)][1-G(t-)]^2}.$$

Assume that $E[\Phi(T,U)\mathcal{C}^{1/2+\upsilon}(\max(T,U-\varepsilon)-)S_G(\max(T,U-\varepsilon)-)^{-1}] < \infty$, for some $\upsilon > 0$ (arbitrary small).

If we consider the particular case of $\mathcal{F}_1$, Assumption 2 automatically holds, provided that the moment conditions of Assumption 3 hold. Generally, one can show this is also the case for parametric classes of functions or sufficiently smooth classes of functions (using permanence properties of Donsker classes).

We now state the main theoretical result of this section.

Theorem 1. Recall that $B_i = \inf(A_i, C_i)$, where $A_i = \max(T_i, U_i - \varepsilon_i)$, and that $\eta_i = \mathbf{1}_{C_i \le A_i}$, and let $H(t) = \mathbb{P}(B > t)$. Under Assumptions 1 to 3,

$$\int \phi(t,u) d(\hat{F} - F^*)(t,u) = \frac{1}{n}\sum_{i=1}^{n} \psi_\phi(Y_{1i}, Y_{2i}, \varepsilon_i) + R_n(\phi), \tag{2.5}$$

where $\sup_{\phi\in\mathcal{F}} |R_n(\phi)| = o_P(n^{-1/2})$, and

$$\begin{aligned}\psi_\phi(Y_i,Z_i,\varepsilon_i) &= \int \left\{ S_G(a) - \frac{(1-\eta_i)S_G(B_i\vee a)}{1-H(B_i)} + \int \frac{\mathbf{1}_{B_i\ge u} S_G(u\vee a) dF_A(u)}{(1-H(u))(1-F_A(u))} \right. \\ &\quad \left. - \frac{\eta_i \mathbf{1}_{B_i > a}}{1-F_A(B_i)} \right\} \frac{\phi(t,u) d\mathbb{P}_{(T,U,C,\varepsilon)}(t,u,c,e)}{S_G(a-)},\end{aligned}$$

where we used $a = \max(t, u-e)$ to shorten the notation, and where $\mathbb{P}_{(T,U,C,\varepsilon)}$ denotes the true law of (T,U,C,ε). As a consequence, since $E[\psi_\phi(Y,Z,\varepsilon)] = 0$, we have, for all $\phi \in \mathcal{F}$,

$$n^{1/2}\left(\int \phi(t,u) d\hat{F}(t,u) - E[\phi(T,U)]\right) \Longrightarrow \mathcal{N}(0,\sigma_\phi^2), \tag{2.6}$$

with

$$\sigma_\phi^2 = E\left[\left\{\frac{\delta_i\gamma_i\phi(Y_i,Z_i)}{S_G(B_i-)} - E[\phi(T,U)] + \psi_\phi(Y_i,Z_i,\varepsilon_i)\right\}^2\right],$$

and $\Longrightarrow$ denotes the weak convergence.

The proof of this result is postponed to the Appendix section. Equation (2.6) can be used to compute asymptotic confidence intervals, provided that one is able to consistently estimate the asymptotic variance σ_ϕ^2. This can be done by replacing all the unknown distribution functions involved in the expression of σ_ϕ^2 by their empirical counterparts. However, this approximation may be too rough in practice, and bootstrap procedures seem to be more appropriate if one wishes to compute confidence interval. This bootstrap method is shown in section 2.3.2.

In addition to the application of Theorem 1 to the class $\mathcal{F}_1$ (corresponding to the estimation of the joint survival function), we show how this result may be used to provide asymptotic results for the estimation of Kendall's τ coefficient. Kendall's τ coefficient is a classical dependence measure which can be defined in the following way. For two random variables (T,U), $\tau = \mathbb{P}((T_1-U_1)(T_2-U_2)>0) - \mathbb{P}((T_1-U_1)(T_2-U_2)<0)$, where (T_1,U_1) and (T_2,U_2) are independent replications of (T,U). There exists a relationship between τ and the distribution function F, that is $\tau = 4\int F(x,y)dF(x,y) - 1$, see e.g. [Nelsen, 2006]. Therefore, a natural estimator of τ is

$$\hat{\tau} = 4\int \hat{F}(x,y)d\hat{F}(x,y) - 1. \tag{2.7}$$

As it is shown in [Wang and Wells, 2000a], censoring may cause this estimator not to be consistent in some particular situations. Indeed, defining $S_H(y,z) = \mathbb{P}(Y>y, Z>z)$ the survival function of the observed times, $\mathcal{S}_1 = \{(t,u) : S_H(t,u)>0\}$, and $\mathcal{S}_2 = \{(t,u) : S_F(t,u)>0\}$, we can see that some part of the distribution, namely $\mathcal{S}_2 - \mathcal{S}_1$ is never observed, since the corresponding observations are always censored. If this difference of sets is empty, this does not introduce bias in the estimation of τ. In other situations, some bias will arise and can be evaluated according to the method of [Wang and Wells, 2000a]. Corollary 1 below shows that this estimator admits an asymptotic representation.

Corollary 1. Let ψ_F denote function ψ from the Theorem 1 applied to function $\phi = F$. To shorten the notation, we will denote $\psi_{t,u}$ the function corresponding to $\phi(Y,Z) = \mathbf{1}_{Y\le t, Z\le u}$. Assume that $\mathcal{S}_2 - \mathcal{S}_1 = \emptyset$. Then,

$$\begin{aligned}\hat{\tau} - \tau &= 4\left\{\int F(t,u)d[F^*-F](t,u) + \int [F^*(t,u)-F(t,u)]dF(t,u)\right.\\ &\quad \left. + \frac{1}{n}\sum_{i=1}^n\left\{\psi_F(Y_i,Z_i,\varepsilon_i) + \int \psi_{t,u}(Y_i,Z_i,\varepsilon_i)dF(t,u)\right\}\right\}\\ &\quad + o_P(n^{-1/2}).\end{aligned} \tag{2.8}$$

In the representation of Corollary (1), each term is a sum of i.i.d. quantities with zero expectation and finite variance. Therefore, Corollary 1 shows that $\hat{\tau}$ is asymptotically Gaussian. Its asymptotic variance (which has a complex form) can be

deduced from this representation. Nevertheless, we do not emphasize this variance, since we recommend using bootstrap procedures to investigate the law of $\hat{\tau}$ (see section 2.4.1).

Let us also mention that, if the assumption $\mathcal{S}_2 - \mathcal{S}_1 = \emptyset$ does not hold, the result is still true, but with τ replaced by $4\int_{\mathcal{S}_1} F(x,y)dF(x,y) - 1$.

Proof.[Proof of Corollary 1] Write

$$\begin{aligned}\hat{\tau} - \tau &= 4\Big\{\int F(t,u)d[\hat{F}-F](t,u) + \int[\hat{F}(t,u)-F(t,u)]dF(t,u) \\ &\quad + \int[\hat{F}(t,u)-F(t,u)]d[\hat{F}-F](t,u)\Big\}. \end{aligned} \tag{2.9}$$

Applying Theorem 1 to function F, the first term of (2.9) can be expanded as

$$\int F(t,u)d[F^*-F](t,u) + \frac{1}{n}\sum_{i=1}^{n}\psi_F(Y_i,Z_i,\varepsilon_i) + o_P(n^{-1/2}).$$

Moreover, again from Theorem 1,

$$\begin{aligned}\int[\hat{F}(t,u)-F(t,u)]dF(t,u) &= \int[F^*(t,u)-F(t,u)]dF(t,u) \\ &\quad + \frac{1}{n}\sum_{i=1}^{n}\int\psi_{t,u}(Y_i,Z_i,\varepsilon_i)dF(t,u) + o_P(n^{-1/2}).\end{aligned}$$

The third term of (2.9) can be rewritten as

$$\begin{aligned}&\int[F^*(t,u)-F(t,u)]d[F^*-F](t,u) + \int[\hat{F}(t,u)-F^*(t,u)]d[F^*-F](t,u) \\ &+ \int[\hat{F}(t,u)-F^*(t,u)]d[\hat{F}-F^*](t,u) + \int[F^*(t,u)-F(t,u)]d[\hat{F}-F^*](t,u) \\ &:= \mathcal{T}_1 + \mathcal{T}_2 + \mathcal{T}_3 + \mathcal{T}_4.\end{aligned}$$

The term $\mathcal{T}_1$ is a second order degenerate $U-$statistics and is therefore of order $O_P(n^{-1})$. To study $\mathcal{T}_2$, apply Theorem 1 to the class of indicator functions $\mathbf{1}_{T\le t,U\le u}$ to obtain

$$\mathcal{T}_2 = \frac{1}{n}\sum_{i=1}^{n}\int\psi(T_i,U_i,\varepsilon_i)d[F^*-F](t,u) + o_P(n^{-1/2}).$$

The integral in this decomposition is zero, since $E[\psi_\phi(T,U,\varepsilon)] = 0$, and since $\int\phi d[F^*-F](t,u) = 0$. Finally, observe that $\mathcal{T}_3$ can be rewritten as

$$\mathcal{T}_3 = \frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i\gamma_i[\hat{F}(Y_i,Z_i)-F^*(Y_i,Z_i)][\hat{S}_G(B_i-)-S_G(B_i-)]}{S_G(B_i-)\hat{S}_G(B_i-)},$$

which is bounded by

$$|\mathcal{T}_3| \le \sup_{t,u}|\hat{F}(t,u)-F^*(t,u)|\sup_b|\hat{S}_G(b)-S_G(b)|\sup_b\left|\frac{S_G(b)}{\hat{S}_G(b)}\right| \times \frac{1}{n}\sum_{i=1}^{n}\frac{\eta_i}{S_G(B_i-)^2}.$$

The first supremum is $O_P(n^{-1/2})$ from Theorem 1, the second one is $o_P(1)$ from the uniform consistency of Kaplan-Meier estimator (see [Stute and Wang, 1993]), while the third one is $O_P(1)$ (see [Gill, 1983]). Moreover, the empirical mean on the right-hand side is $O_P(1)$ provided that each term of the sum has finite expectation, which is the case from Assumption 3. Combining these facts leads to $\mathcal{T}_3 = o_P(n^{-1/2})$. Term $\mathcal{T}_4 = o_P(n^{-1/2})$ by exactly the same arguments and simpler. This concludes the proof.

2.3.2 Bootstrap procedure

As we already mentioned, the asymptotic results of Theorem 1 may be difficult to use when it comes to approximating the law of $\int \phi(t,u)d\hat{F}(t,u)$. The problem comes form the complex from of the asymptotic variance, and therefore from the difficulty to estimate it in an accurate way. Therefore, the aim of the present section is to propose a bootstrap procedure that allows to circumvent this problem.

Under univariate censoring, two main methodologies have been proposed in the literature to perform bootstrap, see [Reid, 1981] and [Efron, 1981]. Reid's approach consists of simulating i.i.d. samples under the Kaplan-Meier distribution of the lifetimes estimated from observed data. On the other hand, Efron's methodology consists of using the nonparametric estimators of the distribution of the lifetime and of the censoring to resimulate samples. [Akritas, 1986] showed that only Efron's methodology was consistent. We therefore propose to adopt this strategy.

The basic idea consists of simulating variables (T,U) according to the estimated distribution defined by $\hat{F}$ (and renormalized in order to ensure that the total mass is equal to one). The censoring can be simulated similarly using $\hat{G}$, while ε is simulated according to its empirical distribution $\hat{F}_\varepsilon(t) = n^{-1}\sum_{i=1}^n \mathbf{1}_{\varepsilon_i \leq t}$. The procedure is summarized below.

To compute B bootstrap $n-$samples, repeat for $b = 1, ..., B$ the following simulation scheme,

1. Simulate independent variables $(T_i^b, U_i^b)_{1\leq i\leq n}$ under the probability distribution
   $\hat{F}/\hat{F}(+\infty,+\infty)$, where we recall that $\hat{F}(+\infty,+\infty) = \lim_{t\to+\infty,u\to+\infty}\hat{F}(t,u)$ is not necessarily equal to one.
2. Simulate independent variables $(\varepsilon_i^b)_{1\leq i\leq n}$ under the probability distribution $\hat{F}_\varepsilon$.
3. Simulate independent variables $(C_i^b)_{1\leq i\leq n}$ under the probability distribution $\hat{G}/\hat{G}(+\infty)$.
4. The $b-$th bootstrap sample is composed of $(Y_i^b, Z_i^b, \delta_i^b, \gamma_i^b, \varepsilon_i)_{1\leq i\leq n}$, where $Y_i^b = \inf(T_i^b, C_i^b)$, $Z_i^b = \inf(U_i^b, C_i^b + \varepsilon_i^b)$, $\delta_i^b = \mathbf{1}_{T_i^b \leq C_i^b}$, $\gamma_i^b = \mathbf{1}_{U_i^b \leq C_i^b + \varepsilon_i^b}$.

2.3.3 Application to survival copula inference

Survival copula models are a common tool to model dependence between two lifetimes (T,U). Indeed, the bivariate survival function $S_F(t,u) = \mathbb{P}(T > t, U > u)$ of the random vector (T,U) admits, by Sklar's Theorem ([Sklar, 1959]), a copula representation, that is

$$S_F(t,u) = \mathfrak{C}(S_T(t), S_U(u)),$$

where $S_T(t) = \mathbb{P}(T > t)$ and $S_U(u) = \mathbb{P}(U > u)$, and where $\mathfrak{C}$ is a survival copula function (see e.g. [Nelsen, 2006]). To understand the dependence between T and U, which is represented by the copula function $\mathfrak{C}$, it is natural to search for an estimator of $\mathfrak{C}$, usually based on a parametric or semiparametric model, see e.g. [Shih and Louis, 1995]. Nonparametric inference is then required to assess the validity of the model.

[Wang and Wells, 2000b] proposed to extend the methodology of [Genest and Rivest, 1993] in presence of censoring. This approach relies on the estimation of the function $v \to K(v) = \mathbb{P}(S_F(T,U) \leq v)$. If we consider the particular case of Archimedean copula families (that is copulas defined as $\mathfrak{C}(u,v) = \phi^{-1}(\phi(u)+\phi(v))$ where the generator ϕ is a convex function satisfying the conditions of Theorem 4.3.4 in [Nelsen, 2006]), there exists a one-to-one correspondance between the generator ϕ and the function K, through the relationship

$$K(v) = v - \frac{\phi(v)}{\phi'(v)}. \tag{2.10}$$

The basic idea of goodness-of-fit procedures based on K consists of comparing a parametric estimator, (based on an estimator $\phi_{\hat{\theta}}$ depending on the parametric model and on $\hat{\theta}$, the association parameter estimated from the data) to a nonparametric one. [Wang and Wells, 2000b] used an estimator based on the nonparametric estimator of [Dabrowska, 1988]. The nonparametric estimator that we propose to use is defined as

$$\hat{K}(v) = \int \mathbf{1}_{\hat{S}_F(t,u)\leq v} d\hat{F}(t,u). \tag{2.11}$$

In Proposition 1, we show that the process $n^{1/2}(\hat{K}(v) - K(v))$ converges towards a Gaussian process. This kind of result is essential to legitimate goodness-of-fit techniques that will be fully discussed in section 2.4.2. Nevertheless, we do not focus on the estimation of the asymptotic covariance process. In practice, since its computation seems rather delicate, it is preferable to rely on bootstrap procedure (see section 2.4.2).

Proposition 1. Assume that :

1. The distribution function $K(v)$ admits a continuous bounded derivative $k(v)$.
2. Given $S_F(t,u) = v$, there exists a version of the conditional distribution of (Y,Z) and a countable family $\mathcal{P}$ of partitions $\mathcal{E}$ on $\mathcal{I}$ (where $\mathcal{I}$ denotes the support of (T,U)) into a finite number of Borel sets satisfying

$$\inf_{\mathcal{E}\in\mathcal{P}} \max_{E\in\mathcal{E}} diam(E) = 0,$$

such that, for all $E \in \mathcal{E}$, the mapping

$$v \to \mu_v(E) = k(v)\mathbb{P}\left((T,U) \in E | S(T,U) = v\right)$$

is continuous.

Then, there exists a zero-mean Gaussian process W such that,

$$\left\{n^{1/2}\left(\hat{K}(v) - K(v)\right),\ v \in \mathbb{R}\right\} \Longrightarrow \quad \left\{-\int\int \mathbf{1}_{S_F(t,u)>v} dW(t,u) - \int\int W(t,u) d\mu_\cdot(t,u),\ v \in \mathbb{R}\right\},$$

where $\Longrightarrow$ denotes the weak convergence of the stochastic process.

Proof. From Theorem 1, one can deduce $n^{1/2}(\hat{S}_F(t,u) - S_F(t,u)) \Longrightarrow W(t,u)$, where W is a Gaussian process with mean zero. Consequently, Theorem 1 in [Wang and Wells, 2000b] applies.

2.4 Simulations and real data example

In this section, we investigate the finite sample size behaviour of our procedure. This investigation is done through simulation studies and a real data example. The data that we consider have been initially studied by [Frees et al., 1996], and was studied by [Carriere, 2000], [Youn and Shemyakin, 1999], [Youn and Shemyakin, 2001] and [Luciano et al., 2008]. We refer to [Frees et al., 1996] for a more detailed description of this dataset, containing lifetimes of two members of a couple who subscribed an insurance contract. The dataset concerns 14947 contracts from a large Canadian insurer, observed between Decembre 29th, 1988 and Decembre 31th, 1993[1]. After elimination of same-sex contracts and of couples with more than one policies (for which we only keep one policy), 11454 contracts remain. In addition to bivariate censoring, observations are subject to left truncation, which was not considered by our approach. Neglecting left-truncation will lead to a slight over-estimation of the lifetimes, which, from the prospective of an insurer who wishes to evaluate his liabilities in the case of a pension contract, represents a cautious approach.

In section 2.4.1, we discuss the problem of estimating Kendall's τ coefficient, illustrating the theoretical results of Corollary 1. In section 2.4.2, we study the practical implementation of the goodness-of-fit procedure for copula models, based on the process $\hat{K}$ defined in (2.11).

2.4.1 Nonparametric estimation of Kendall's τ coefficient

Real data example. Using $\hat{\tau}$ defined in equation (2.7), we find an estimated value of Kendall's τ coefficient which is $\hat{\tau} = 0.6696$, which is roughly of the same order as the values obtained by other authors on the same data-set (for example, for a specific generation, [Luciano et al., 2008] obtained an estimation which is 0.6039). We used the nonparametric bootstrap procedure described in section 2.3.2 to approximate the law of $\hat{\tau}$. Through $B = 1000$ bootstrap replications, we obtain an estimation of the distribution of $\hat{\tau}$ which is represented in Figure 2.1 below. We observe that the distribution of $\hat{\tau}$ obtained using the bootstrap procedure does not seem to be Gaussian. Therefore, it legitimates to rely rather on this bootstrap procedure than on normal approximation to investigate uncertainty in estimating τ.

Simulation study. To illustrate the convergence of $\hat{\tau}$, we present some results of a simulation study. The random lifetimes (T, U) are simulated from a Clayton copula model (see Table 2.2 in section 2.4.2 for a precise definition) with association parameter $\theta = 2$ (which corresponds to a value $\tau = 0.5$), with marginals following a Weibull distribution. Weibull distribution is parametrized through a shape parameter α and a scale parameter β, and admits a density

$$f(t) = \frac{\alpha}{\beta}\left(\frac{t}{\beta}\right)^{\alpha-1} \exp\left(-\frac{t^\alpha}{\beta^\alpha}\right),$$

for $t \geq 0$. We consider the case $\alpha = 10$ and $\beta = 1.7$. The censoring variable C is simulated according to an exponential distribution with parameter λ (with mean

1. The author wishes to thank the Society of Actuaries, through the courtesy of Edward J. Frees and Emiliano Valdez, for allowing use of the data in our study

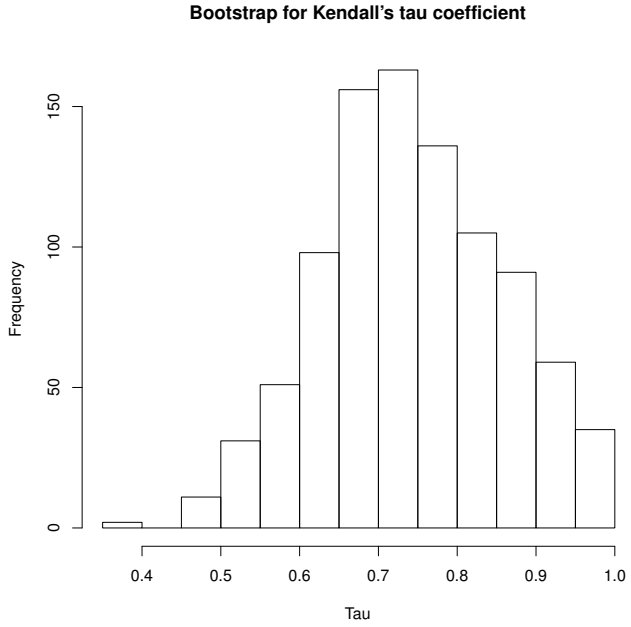


FIGURE 2.1 – Histogram of the distribution of $\hat{\tau}$ using the bootstrap procedure.

λ^{-1}). Different values of λ are considered in order to change the average proportion of doubly uncensored observations. Random variables ε_i are simulated according to an exponential distribution with parameter $\mu = 50$.

For each considered value of the parameter λ, we generate $n-$samples for different values of n. We repeat $N = 1000$ times the simulation scheme in order to estimate the bias $E[\hat{\tau} - \tau]$, the variance $Var(\hat{\tau})$, and the mean-squared error $E[(\hat{\tau} - \tau)^2]$. Results are presented in Table 2.1 below.

Model	Criterion	n=1000	n=2000
$\alpha = 10$	MSE =	0.004537	0.002548
$\beta = 1.1$	Bias =	0.06686	0.05020
(35% of uncensored)	Variance =	6.6984e-5	2.8133e-5
$\alpha = 10$	MSE =	0.006949	0.004482
$\beta = 1.7$	Bias =	0.08275	0.06650
(20% of uncensored)	Variance =	1.020e-4	5.9425e-5

TABLE 2.1 – Estimation of the mean-squared error and related quantities for the estimation of Kendall's τ coefficient.

2.4.2 Goodness-of-fit for semiparametric copula models

A goodness-of-fit procedure based on $\hat{K}$. Consider a parametric family of Archimedean survival copulas $\mathcal{F}_C = \{\mathfrak{C}_\theta : \theta \in \Theta\}$. We will denote ϕ_θ the Archime-

dean generator of copula $\mathfrak{C}_\theta$. We describe how to extend the procedure proposed by [Genest and Rivest, 1993] to test

$$H_0 : \mathfrak{C} \in \mathcal{F}_C,$$

against

$$H_1 : \mathfrak{C} \notin \mathcal{F}_C.$$

The principle of the test consists of computing an estimator $\hat{\theta}$ (assuming that H_0 holds) from the data, then using $\phi_{\hat{\theta}}$ and (2.10) computing a parametric estimator $K_{\hat{\theta}}$ of function K. Next, considering some distance d between curves, the test statistic is $\mathcal{T}_n = d(\hat{K}, \hat{K}_{\hat{\theta}})$, where $\hat{K}$ is defined in (2.11). H_0 is rejected when $\mathcal{T}_n > s_\alpha$, where s_α is a critical value that ensures that the procedure achieves level α. In the following, we will consider the particular case $d(\hat{K}, K_\theta) = [\int_0^1 \left(\hat{K}(v) - K_\theta(v)\right)^2 dv]^{1/2}$.

To estimate $\hat{\theta}$, one can either rely on a semiparametric maximum likelihood procedure, as it is done in [Shih and Louis, 1995], or take $\hat{\theta} = \arg\min_{\theta\in\Theta} d(\hat{K}, K_\theta)$, which has been done in [Luciano et al., 2008] and seems more natural in our framework. Therefore, we will use this second approach, and our test statistic may be rewritten as $\mathcal{T}_n = \min_{\theta\in\Theta} d(\hat{K}, K_\theta)$. To compute the critical values, a bootstrap procedure is required. In our framework, [Wang and Wells, 2000b] proposed a bootstrap methodology, which has been shown to fail to be consistent by [Genest et al., 2006]. Therefore, we prefer to adapt the consistent resampling plan which was defined in [Genest et al., 2006] to the presence of censoring. This results on using the bootstrap procedure defined in section 2.3.2, but replacing Step 1 by

1'. Simulate independent variables $(T_i^b, U_i^b)_{1\leq i\leq n}$ under the distribution defined by $\mathfrak{C}_{\hat{\theta}}$ and with marginal distributions defined by the Kaplan-Meier estimators (univariate) of T and U,

which corresponds to an approximation of the law of (T, U) under H_0. Alternatively, in a full parametric modelling of the distribution of (T, U), parametric distributions may be used instead of the nonparametric Kaplan-Meier estimators. Based on B bootstrap replications of $\mathcal{T}_n$, the critical value s_α can be determined in order to ensure a level α of the procedure.

Real data example. We consider three copula models that have been used by [Luciano et al., 2008] to study the mortality of a particular generation in the data-set (without distinguishing between generations). These models are Clayton, Frank copula models, and a copula called Nelsen 4.2.20 (corresponding to the copula defined in formula 4.2.20 in [Nelsen, 2006]). Definition of these three Archimedean families in recalled in Table 2.2 below. Estimators $\hat{\theta}$ are computed by minimization of the distance $d(\hat{K}, K_\theta)$.

The graphical comparison between $\hat{K}$ and $K_{\hat{\theta}}$ is presented in Figure 2.2 below. Table 2.3 presents the results of the test procedure described above, comparing the value of the test statistic $\mathcal{T}_n$ to the quantiles of the distribution of $\mathcal{T}_n$ under the null hypothesis (computed using our bootstrap procedure).

The model with the smallest value of the test-statistic is Nelsen's 4.2.20 copula model. Frank's copula model achieves a value of the test-statistic which is quite close from Nelsen's 4.2.20 case. However, in this last model, the corresponding $p-$value

Model	$\phi_\theta(t)$	$\mathfrak{C}_\theta(u,v)$	$\hat{\theta}$
Clayton	$\theta^{-1}(t^{-\theta}-1)$	$(u^{-\theta}+v^{-\theta}+1)^{-1/\theta}$	4.8991
Frank	$-\log\left(\frac{\exp(-\theta t)-1}{\exp(-\theta)-1}\right)$	$-\theta^{-1}\log\left(1+\frac{(\exp(-\theta u)-1)(\exp(-\theta v)-1)}{(\exp(-\theta)-1)}\right)$	11.4115
4.2.20	$\exp(t^{-\theta})-e$	$\log\left(\exp(u^{-\theta})+\exp(v^{-\theta})-e\right)^{-1/\theta}$	1.338

TABLE 2.2 – Expression of the considered copula families. The column $\hat{\theta}$ presents the association parameter estimated on the data-set, minimizing distance d.

Model	Test statistic	95% quant.	97.5 % quant.	99 % quant.	p-value
Clayton	0.06229	0.16638	0.17600	0.19362	0.533
Frank	0.05434	0.04330	0.04667	0.05203	0.008
Nelsen 4.2.20	0.05181	0.12243	0.12245	0.12246	0.492

TABLE 2.3 – Goodness-of-fit procedure for a the three survival copula models considered (Clayton, Frank, Nelsen 4.2.20).

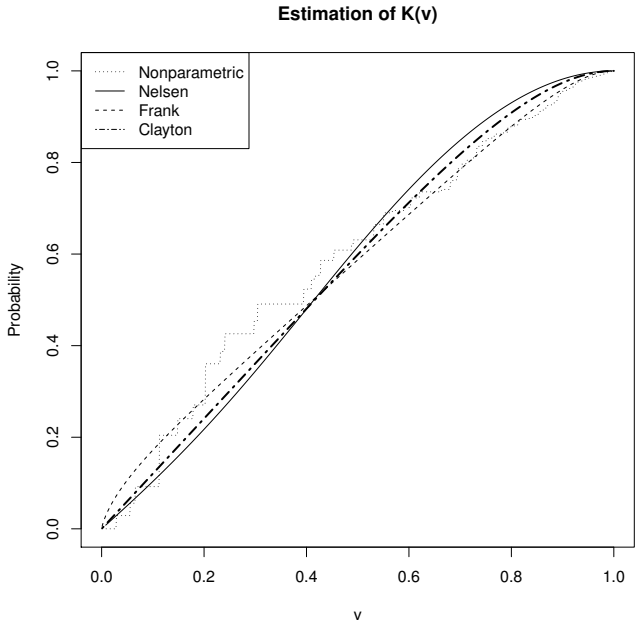


FIGURE 2.2 – Graphical comparison between $\hat{K}$ and $K_{\hat{\theta}}$ for the three copula models considered.

is small, while it is not the case for the two other models. Graphically, it seems that all the models that we consider have difficulties to capture the behaviour of function $\hat{K}$ for values of v between 0.2 and 0.4.

2.5 Conclusion

The estimator that we considered in this chapter is designed for applications where the censoring times for both individuals differ only through an observed random variable. In the example that we consider, this observed variable represents the age difference. We only considered the case of two random lifetimes (T, U), but the procedure can easily be generalized to more lifetimes. The main difficulty of this extension comes from the fact that the procedure requires to put mass only at fully observed observations. To obtain a sufficient number of such observations when the number of lifetimes is high, one would need to have a large value of the sample size n. We mainly focused on applying our results to the study of copula models. Other applications of this technique could be considered, as regression models (see [Lopez and Saint-Pierre, 2012] for related problems), or the evaluation of various dependence measures, see [Fan et al., 2000] or [Hougaard, 2000].

2.6 Appendix

2.6.1 Proof of Theorem 1

Let (t_0, u_0) denote some point in $\mathbb{R}^2$ such that $S_F(t_0, u_0) > 0$.
First case : $\phi(t, u) = 0$ **for** $t \geq t_0$ **or** $u \geq u_0$. One can write

$$\int \phi(t,u) d(\hat{F} - F^*)(t,u) = \frac{1}{n} \sum_{i=1}^{n} \frac{\delta_i \gamma_i [S_G(B_i-) - \hat{S}_G(B_i-)] \phi(Y_i, Z_i)}{S_G(B_i-)^2} + R_{1n}(\phi),$$

with

$$|R_{1n}(\phi)| \leq \sup_{a \leq a_0} \frac{|\hat{S}_G(a) - S_G(a)|^2}{S_G^2(a) \hat{S}_G(a)} \times \left(\frac{1}{n} \sum_{i=1}^{n} \delta_i \gamma_i \Phi(Y_i, Z_i) \right), \tag{2.12}$$

where a_0 is some point in $\mathbb{R}$ such that $S_G(a_0) > 0$. It follows from the uniform $n^{1/2}-$consistency of $\hat{S}_G$ (see [Gill, 1983]) that the right-hand side in (2.12), which does not depend on ϕ, is $o_P(n^{-1/2})$. Moreover, let us observe that the functions

$$\begin{aligned} f_n(y, z, e, \delta, \gamma) &= \delta \gamma \hat{S}_G(b-) \phi(y, z) [S_G(b-)]^{-2}, \\ f(y, z, e, \delta, \gamma) &= \delta \gamma S_G(b-) \phi(y, z) [S_G(b-)]^{-2}, \end{aligned}$$

are two elements of the Donsker class $\mathcal{H}_{t_0,u_0}$ defined in Assumption 2. Moreover, using the uniform convergence rate of $\hat{S}_G$, one obtains that $\|f_n - f\|_\infty \to 0$. Therefore, the asymptotic equicontinuity of Donsker classes (see Lemma 19.24 in [van der Vaart, 1998]) ensures that

$$\int \phi(t,u) d(\hat{F} - F^*)(t,u) = \int \frac{[S_G(a-) - \hat{S}_G(a-)] \phi(t,u) d\mathbb{P}_{(T,U,C,\varepsilon)}(t,u,c,e)}{S_G(a-)} + R_{2n}(\phi), \tag{2.13}$$

where $\sup_{\phi\in\mathcal{F}}|R_{2n}(\phi)| = o_P(n^{-1/2})$. Next, the representation (2.5) follows from [Stute, 1996] or [Gijbels and Veraverbeke, 1991], since

$$\begin{aligned} S_G(a) - \hat{S}_G(a) &= \frac{1}{n}\sum_{i=1}^{n}\left\{\int \frac{\mathbf{1}_{B_i\geq u}S_G(u\vee a)dF_A(u)}{(1-H(u))(1-F_A(u))} - \frac{(1-\eta_i)S_G(B_i\vee a)}{1-H(B_i)} - \right\} \\ &\quad + \left\{S_G(a) - \frac{\eta_i\mathbf{1}_{B_i>a}}{1-F_A(B_i)}\right\} + R(a), \end{aligned}$$

where $\sup_{a\leq a_0}|R(a)| = o_P(n^{-1/2})$, where a_0 is some point such that $\mathbb{P}(B > a_0) > 0$ (by assuming that ϕ is zero for large values of t or u, we ensure that, in (2.13), only terms with b smaller than such a a_0 appear). To deduce equation (2.6), it suffices to observe that

$$\begin{aligned} \int \phi(t,u)d[\hat{F}-F](t,u) &= \frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i\gamma_i\phi(Y_i,Z_i)}{S_G(B_i-)} - E\left[\phi(T,U)\right] \\ &\quad + \frac{1}{n}\sum_{i=1}^{n}\psi_\phi(Y_i,Z_i,\varepsilon_i) + o_P(n^{-1/2}). \end{aligned}$$

Each of these two i.i.d. sums have zero mean, and the Central Limit Theorem applies.

General case : the general case follows from a combination of the first case and of Lemma 1. Indeed, for a general function $\phi(t,u)$ and an arbitrary point (t',u'), let us consider a decomposition

$$n^{1/2}\int \phi(t,u)d(\hat{F}-F^*)(t,u) = P_n(t',u',\phi) + R_n(t',u',\phi),$$

with

$$P_n(t',u',\phi) = n^{1/2}\int \phi(t,u)\mathbb{1}_{t\leq t',u\leq u'}d(\hat{F}-F^*)(t,u)$$

and

$$R_n(t',u',\phi) = n^{1/2}\int \phi(t,u)\mathbb{1}_{(t,u)\notin[0,t']\times[0,u']}d(\hat{F}-F^*)(t,u).$$

Define $\mathcal{I}_{(t',u')} = [0,t']\times[0,u']$, and let $\mathcal{I}$ denote the support of (Y,Z), and (t_1,u_1) its upper bound. Remark that,

$$|R_n(t',u',\phi)| = n^{1/2}\left|\int \phi(t,u)\mathbf{1}_{(t,u)\notin\mathcal{I}_{(t',u')}}d[F-F^*](t,u)\right| \leq M_n\Gamma_n(t',u'), \quad (2.14)$$

where

$$M_n = \sup_a n^{1/2}\left|\frac{(\hat{S}_G-S_G)(a)}{\mathcal{C}(a)^{1/2+\upsilon}}\right| \times \sup_a\left|\frac{S_G(a)}{\hat{S}_G(a)}\right|,$$

and

$$\Gamma_n(t',u') = \frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i\gamma_i\mathcal{C}^{1/2+\upsilon}(B_i-)\mathbf{1}_{(Y_i,Z_i)\notin\mathcal{I}_{(t',u')}}\Phi(Y_i,Z_i)}{S_G(B_i-)^2}.$$

The result now follows from Lemma 1. To check its conditions, let us first observe that, by the first case of the proof, the process $P_n(t',u',\phi)$ converges on $\mathcal{I}_{(t_0,u_0)}\times\mathcal{F}$ for any (t_0,u_0) such that $t_0 < t_1$ and $u_0 < u_1$. Condition 1 is satisfied due to the finitness of the variance of the terms in the i.i.d. sum in the right-hand side of

decomposition (2.5). Condition 2 follows from (2.14). Indeed, $M_n = O_P(1)$ from Theorem 2.1 in [Gill, 1983], since $\int \mathcal{C}(a)^{-1-2v} d\mathcal{C}(a) < \infty$ (see condition (2.1) in [Gill, 1983]). Moreover, we have $\Gamma_n(t', u') \leq \Gamma_n(t_0, u_0)$ for any $(t', u') \notin \mathcal{I}_{(t_0,u_0)}$ and $\Gamma_n(t_0, u_0) \to \Gamma(t_0, u_0)$ in probability due to Assumption 3. Condition 3 is obviously satisfied.

2.6.2 A technical Lemma

Lemma 1. Assume that, for any (t_0, u_0) such that $\mathcal{I}_{(t_0,u_0)}$ is strictly included in the interior part of $\mathcal{I}$,

$$(P_n(t', u', \phi))_{(t',u') \in \mathcal{I}_{(t_0,u_0)}} \Longrightarrow (W_{V_\phi}(t', u'))_{(t',u') \in \mathcal{I}_{(t_0,u_0)}},$$

where $W_{V_\phi}(t', u')$ is a Gaussian process with the covariance function $V_\phi(t'_1, u'_1, t'_2, u'_2)$ and $\Longrightarrow$ denotes the weak convergence. Moreover, let $R_n(t', u', \phi) = P_n(t_1, u_1, \phi) - P_n(t', u', \phi)$ and assume that the following conditions hold,

1. V_ϕ is continuous in the point (t_1, u_1, t_1, u_1) and $\sup_{\phi \in \mathcal{F}} |V_\phi| < \infty$,
2. $|R_n(t', u', \phi)| \leq M_n \times \Gamma_n(t_0, u_0)$, for all $(t', u') \in \mathcal{I} - \mathcal{I}_{(t_0,u_0)}$, with $M_n = O_P(1)$, and $\Gamma_n(t_0, u_0) \to \Gamma(t_0, u_0)$ in probability,
3. $\lim_{(t_0,u_0)\to(t_1,u_1)} \Gamma(t_0, u_0) = 0$.

Then $P_n(t_1, u_1, \phi) \Longrightarrow \mathcal{N}(0, V_\phi(t_1, u_1, t_1, u_1))$.

Proof. From Theorem 13.5 in [Billingsley, 1999] and from condition 1, it suffices to show that, for all $\varepsilon > 0$,

$$\lim_{(t_0,u_0)\to(t_1,u_1)} \limsup_{n\to\infty} \mathbb{P}\left(\sup_{(t',u') \in \mathcal{I} - \mathcal{I}_{(t_0,u_0)}, \phi \in \mathcal{F}} |R_n(t', u', \phi)| > \varepsilon \right) = 0. \tag{2.15}$$

Using condition 2 in the Lemma, the probability in equation (2.15) is bounded, for all $M > 0$, by

$$\mathbb{P}(|\Gamma_n(t', u') - \Gamma(t', u')| > \varepsilon/M - \Gamma(t', u')) + \mathbb{P}(M_n > M). \tag{2.16}$$

Moreover, from condition 4,

$$\limsup_{n\to\infty} \mathbb{P}(|\Gamma_n(t', u') - \Gamma(t', u')| > \varepsilon/M - \Gamma(t', u')) = \mathbf{1}_{\varepsilon/M - \Gamma(t',u') \leq 0}.$$

Since $\Gamma(t', u') \to 0$ when $(t', u') \to (t_1, u_1)$ (condition 5), we can deduce that

$$\lim_{(t',u')\to\tau_H} \limsup_{n\to\infty} \mathbb{P}(|\Gamma_n(t', u') - \Gamma(t', u')| > \varepsilon/M - \Gamma(t', u')) = 0.$$

Hence,

$$\lim_{(t_0,u_0)\to(t_1,u_1)} \limsup_{n\to\infty} \mathbb{P}\left(\sup_{(t',u') \in \mathcal{I} - \mathcal{I}_{(t_0,u_0)}, \phi \in \mathcal{F}} |R_n(t', u', \phi)| > \varepsilon \right) \leq \limsup_{n\to\infty} \mathbb{P}(M_n > M).$$

To conclude the proof, let M tend to infinity.

Chapitre 3

Nonparametric copula estimation under bivariate censoring

Abstract. In this chapter, we consider nonparametric copula inference under bivariate censoring. Based on an estimator of the joint cumulative distribution function, we define a discrete and two smooth estimators of the copula. The construction that we propose is valid for a large range of estimators of the distribution function, and therefore for a large range of bivariate censoring frameworks. Under some conditions on the tails of the distributions, the weak convergence of the corresponding copula processes is obtained in $l^\infty([0,1]^2)$. We derive the uniform convergence rates of the copula density estimators deduced from our smooth copula estimators. Investigation of the practical behavior of these estimators is done through a simulation study and two real data applications, corresponding to different censoring settings. We use our nonparametric estimators to define a goodness-of-fit procedure for parametric copula models. A new bootstrap scheme is proposed to compute the critical values.

This chapter corresponds to the preprint [Gribkova and Lopez, 2013] on *HAL*.

3.1 Introduction

When it comes to analyzing dependence between random variables, copula models have appeared as a common and flexible tool. According to Sklar's Theorem ([Sklar, 1959]), the multivariate distribution function $F(t_1,\dots,t_d) = \mathbb{P}(T_1 \leq t_1,\dots,T_d \leq t_d)$ of a random vector $(T_1,\dots,T_d)$ can be coupled to its univariate marginal distributions $F_j(t_j) = \mathbb{P}(T_j \leq t_j)$, for $j = 1,...,d$, by the relation

$$F(t_1,\dots,t_d) = \mathfrak{C}(F_1(t_1),\dots,F_d(t_d)), \tag{3.1}$$

where $\mathfrak{C}$ is a copula function, that is, by definition, a multivariate distribution on $[0,1]^d$ with uniform marginal distributions. The copula function appears as a quantity that contains all the information about the dependence structure of the random vector, and is unique in the case where the marginal distributions are continuous. In numerous situations in lifetime data analysis, estimation of $\mathfrak{C}$ must be performed from indirect observations of the variables $(T_1,...,T_d)$, due to the presence of

censoring (see e.g. [Fleming and Harrington, 1991] for a description of censoring mechanism). In this chapter, we define a new class of nonparametric estimators of the copula function $\mathfrak{C}$ adapted to various schemes of multivariate random censoring. We derive asymptotic properties of these estimators, and investigate their practical behavior through a simulation study and some applications to real data.

Copula models represent a flexible alternative to fully parametric models of multivariate distribution function, allowing to study the dependence structure separately from the marginal distributions. This property of a copula becomes of prime importance for applications in economics and insurance. A detailed introduction to copula theory and dependence modeling can be found in [Joe, 1997] or [Nelsen, 2006]. For a recent survey on copula models in econometrics we refer to [Patton, 2012]. Various applications in finance and/or insurance are considered by [Frees and Valdez, 1998], [Embrechts et al., 2003], [Cherubini et al., 2004] and [Bouyé et al., 2007].

Copula estimation for uncensored data has been extensively studied in the literature. Several statistical procedures are available (see e.g. [Genest and Rivest, 1993], [Tsukahara, 2005] for parametric or semiparametric modeling, and [Choros et al., 2010] for a global review of the existing methods). A nonparametric estimation approach was introduced by [Deheuvels, 1979], who defined the empirical copula function. We refer to [Fermanian et al., 2004] and [Segers, 2012] for some recent studies of asymptotic properties of this estimator. Unlike parametric or semiparametric models, the empirical copula function is not affected by misspecification (see [Fermanian and Scaillet, 2005] for more details). Therefore, nonparametric approaches are required to construct goodness-of-fit procedures, such as those studied in [Fermanian, 2005]. The empirical copula is an irregular discrete estimator and cannot be used for a copula density estimation. Its smooth versions were considered by several authors, see e.g. [Fermanian et al., 2004], and [Omelka et al., 2009] who proposed techniques designed to reduce boundary bias.

Such nonparametric estimators all rely on the empirical distribution function of $(T_1, ..., T_d)$. Under random censoring, this empirical estimator is unavailable, since the variables T_j subject to censoring are not directly observed. Parametric or semiparametric approaches can be adapted, by performing maximum likelihood estimation with a form of the likelihood criterion which takes the incompleteness of the observations into account, see [Shih and Louis, 1995]. On the other hand, the extension of nonparametric procedures to the censored framework requires replacing the unavailable empirical distribution function by a suitable nonparametric estimator of F. Many estimators of F have been proposed in the literature, see e.g. [Dabrowska, 1988], [Akritas and Van Keilegom, 2003], [van der Laan, 1996], see also a review of most existing procedures in [Lopez and Saint-Pierre, 2012]. Depending on the identifiability assumptions on the censoring mechanism, different procedures can be introduced to take specific forms of censoring into account. [Wang and Wells, 1997] and [Lopez and Saint-Pierre, 2012] considered estimators which are consistent under some restrictions on the dependence structure of the censored variables, [Gribkova et al., 2013] proposed an estimator which is adapted to a simplified censoring framework, corresponding to some specificity of insurance datasets. Goodness-of-fit procedures for censored copula models have been studied by [Wang and Wells, 2000b], [Luciano et al., 2008], [Gribkova et al., 2013] (who used

an extension of the procedure of [Genest and Rivest, 1993]), but are only valid in the particular case of Archimedean copula models.

In this chapter, we propose a general procedure of nonparametric copula estimation under multivariate censoring, based on the availability of a nonparametric estimator of the distribution function. This estimator is required to be of some generic form, which is compatible with many multivariate censoring schemes (that is under various types of identifiability assumptions). We construct three classes of copula estimators. The first one is non-smooth and can be considered as an extension of [Deheuvels, 1979]. The two others are smooth estimators, based on kernel estimation of the distribution of either the variables $(T_j)_{1\leq j\leq d}$, or a transformed version of them (such as the one proposed by [Omelka et al., 2009] to reduce boundary effects in copula smoothing). The weak convergence of the corresponding copula processes is obtained. Moreover, we derive uniform asymptotic convergence rates of the copula density estimators obtained through differentiation of our smooth copula estimators. The practical behavior of these new estimators is investigated through a simulation study, and two applications to real datasets. As a by-product of our estimators, we propose a goodness-of-fit procedure (with computation of the critical values through a new bootstrap scheme) which is consistent for a large number of copula models, even in a non-Archimedean framework.

The rest of the chapter is organized as follows. In section 3.2, we introduce the general multiple censoring model that we consider. We describe some examples of specific censoring schemes that will be covered by our general framework. Empirical copula estimators are defined and studied in section 3.3. Section 3.4 is devoted to the construction and theoretical study of smooth copula estimators. Simulation studies and real data applications are considered in section 3.5. Technical arguments are presented in the Appendix section.

3.2 Model description and examples

In this section, we describe our framework. Section 3.2.1 presents the general setup along with some notations and first model assumptions. Several classical examples are recalled in section 3.2.2.

3.2.1 General setup

For the sake of simplicity, we focus on the two-dimensional case. The extension of our results to higher dimensions is straightforward.

Consider a random vector (T_1, T_2) with the cumulative distribution function $F(t_1, t_2) = \mathbb{P}(T_1 \leq t_1, T_2 \leq t_2)$ and marginal distribution functions F_1, F_2. We will denote by $\mathfrak{C}$ the associated copula function, that is

$$F(t_1, t_2) = \mathfrak{C}(F_1(t_1), F_2(t_2)).$$

To ensure the unicity of the copula function $\mathfrak{C}$, we will assume that the variables T_1 and T_2 are continuous.

In some cases, in particular in lifetime data analysis, one of these two variables (or even both of them) may be subject to censoring, and thus may not be directly

observed. Instead of observing the variable T_j $(j = 1, 2)$, one observes a minimum between it and another (censoring) random variable, which will be denoted by C_j. The available data is then composed of i.i.d. replications $(Y_{1i}, Y_{2i}, \delta_{1i}, \delta_{2i})_{1\leq i\leq n}$ of random vector $(Y_1, Y_2, \delta_1, \delta_2)$, where $Y_j = \min(T_j, C_j)$ and $\delta_j = \mathbb{I}_{T_j\leq C_j}$ for $j = 1, 2$. Sometimes, in addition to this information, one may observe some auxiliary variables. Such situation is illustrated by our Example 3 below. Throughout the sequel, we will assume that the support of the distribution of T_j is included in the support of the corresponding censoring variable C_j for $j = 1, 2$. This assumption is classical in lifetime data analysis. If it is not verified, a part of the distribution is not observed and thus the distribution function cannot be estimated consistently on the whole support of (T_1, T_2).

As we have already mentioned in the introduction, different bivariate distribution function estimators can be considered, depending on the censoring scheme. However, most of them can be written in some generic form. From now on, we will assume that F can be estimated consistently by an estimator $\mathbb{F}_n$ of the following form,

$$\mathbb{F}_n(t_1, t_2) = \frac{1}{n}\sum_{i=1}^{n} W_{in}\mathbb{I}_{Y_{1i}\leq t_1, Y_{2i}\leq t_2}, \tag{3.2}$$

where W_{in} are random weights, designed to compensate asymptotically the bias caused by the particular structure of the data. The appropriate weight W_{in} to be used depends strongly on the identifiability conditions that are required to infer on F. Basically, these assumptions describe a dependence structure between (C_1, C_2) and (T_1, T_2), and may differ from one application to another. In all the examples that we consider, we will assume that $W_{in} = \delta_{1i}\delta_{2i}\hat{g}(Y_{1i}, Y_{2i})$, where $\hat{g}$ is a function estimated from the data, converging towards a limit function g, where g satisfies the following condition,

$$\forall\phi \in L^1,\ E\left[\delta_1\delta_2 g(Y_1, Y_2)\phi(Y_1, Y_2)\right] = E[\phi(T_1, T_2)]. \tag{3.3}$$

Classical situations where the indicated assumptions hold, are described in the following subsection.

3.2.2 Examples

In the examples that we propose, we only consider estimators of the distribution function that correspond to a positive measure. The results that we derive in the following can be adapted to distributions with eventual negative masses at some observations, such as the estimator considered by [Dabrowska, 1988] (see also [Pruitt, 1991a]). However, the resulting copula estimators are not true copula functions due to this negative mass, therefore we do not focus on such cases.

Example 1 : censoring acts only on one of the two variables. In this situation, $C_2 = \infty$ a.s., and consequently $Y_2 = T_2$ and $\delta_2 = 1$ a.s. In such a setting, the estimator $\mathbb{F}_n^{(1)}$ defined by [Stute, 1993] is of the form (3.2), that is

$$\mathbb{F}_n^{(1)}(t_1, t_2) = \frac{1}{n}\sum_{i=1}^{n} W_{in}^{(1)}\ \mathbb{I}_{Y_{1i}\leq t_1, Y_{2i}\leq t_2}, \tag{3.4}$$

where the random weights $W_{in}^{(1)}$ are the jumps of the Kaplan-Meier estimator of the distribution function of T_1. This estimator is consistent provided that C_1 is independent from T_1, and $\mathbb{P}(T_1 \leq C_1|T_2, T_1) = \mathbb{P}(T_1 \leq C_1|T_1)$. A practical way of rewriting $W_{in}^{(1)}$ consists of linking this jump to the Kaplan-Meier estimator of the censoring variable C_1. Indeed, defining a Kaplan-Meier estimator of the censoring variable,

$$\hat{G}(t) = 1 - \prod_{i:Y_{1i} \leq t} \left(1 - \frac{1}{\sum_{j=1}^{n} \mathbb{I}_{Y_{1j} \geq Y_{1i}}}\right)^{1-\delta_{1i}},$$

according to [Satten and Datta, 2001],

$$W_{in}^{(1)} = \frac{\delta_{1i}}{1 - \hat{G}(Y_{1i}-)}.$$

Introducing $G(t) = \mathbb{P}(C_1 \leq t)$, this weight can be seen as an approximation of

$$W_{i}^{(1)} = \frac{\delta_{1i}}{1 - G(Y_{1i}-)}.$$

Example 2 : censoring variables linked through a copula function. This situation is described in [Lopez and Saint-Pierre, 2012]. In this framework, (C_1, C_2) is supposed to be independent from (T_1, T_2). Another assumption is made on the joint distribution of the random vector (C_1, C_2). It is assumed that the joint survival function S_G can be expressed as

$$S_G(c_1, c_2) = \mathbb{P}(C_1 \geq c_1, C_2 \geq c_2) = \mathcal{C}_G(S_1(c_1), S_2(c_2)),$$

where $\mathcal{C}_G$ is a known survival copula, and S_1 and S_2 are the marginal survival functions of C_1 and C_2. Note that the last assumption can be relaxed by estimating $\mathcal{C}_G$ from a parametric model. The only impact of this additional modeling is a modification of the asymptotic distribution of the distribution function estimator, without any modification of the convergence rate provided that the parametric model is sufficiently regular. Denoting by $\hat{S}_1$ and $\hat{S}_2$ the Kaplan-Meier estimators of S_1 and S_2, the estimator of [Lopez and Saint-Pierre, 2012] corresponds to

$$W_{in}^{(2)} = \frac{\delta_{1i}\delta_{2i}}{\mathcal{C}_G(\hat{S}_1(Y_{1i}), \hat{S}_2(Y_{2i}))},$$

with its limit equal to $W_{i}^{(2)} = \delta_{1i}\delta_{2i}[\mathcal{C}_G(S_1(Y_{1i}), S_2(Y_{2i}))]^{-1}$.

Example 3 : censoring variables differ only through an additional observed variable. Here we consider a model, studied in [Gribkova et al., 2013]. It corresponds to a classical situation, appearing when it comes to study the insurance contracts, related to two individuals (generally two members of a same couple). If, for example, T_1 (resp. T_2) is the total lifetime of the husband (resp. his wife), the censoring variables C_1 and C_2 are their ages at a moment of the exit from the study, for a reason different from death. The observed variables are then $Y_1 = T_1 \wedge C_1$ and $Y_2 = T_2 \wedge C_2$. Besides these variables, the age difference ε between two members of the couple is generally observed. In most cases, the two members of the couple,

if both alive, exit the study at the same time. This leads to a link between two censoring variables through the relationship $C_2 - C_1 = \varepsilon$. The observations are then formed of i.i.d. replications $(Y_{1i}, Y_{2i}, \varepsilon_i, \delta_{1i}, \delta_{2i})_{1\leq i\leq n}$. In such setting, a consistent estimator of the distribution function F is of the form (3.2), with the weights given by

$$W_{in}^{(3)} = \frac{\delta_{1i}\delta_{2i}}{1 - \tilde{G}(\max(Y_{1i}, Y_{2i} - \varepsilon_i)-)}, \tag{3.5}$$

where $\tilde{G}(y)$ is a Kaplan-Meier estimator of $G(y) = \mathbb{P}(C_1 \leq y)$ (see [Gribkova et al., 2013]). The censoring variable is observed since one of the two lifetimes is censored, so the estimator $\tilde{G}(y)$ is based on the observations $(\inf(A_i, C_i), \mathbb{1}_{C_i\leq A_i})_{1\leq i\leq n}$, where $A_i = \inf(T_{1i}, T_{2i} - \varepsilon_i)$.

An additional example. An estimator of the form (3.2) has also been considered by [Lopez, 2012]. It can be applied to a large set of situations, since it requires only that (T_1, T_2) is independent from (C_1, C_2), without making any assumption on the dependence structure of two censoring variables. Moreover, this estimator can be used in presence of bivariate left-truncation. Since, for now, there does not exist any weak convergence result for this estimator, we are unable to prove a weak convergence of the corresponding empirical copula process. However, the uniform consistency property of this estimator permits to establish a uniform convergence of the copula estimator, see Theorem 2 below.

3.3 Discrete nonparametric copula estimator

In this section, we first define a nonparametric copula estimator, which extends the empirical copula of [Deheuvels, 1979] in our framework. The asymptotic properties of this estimator are considered in section 3.3.2 (uniform $n^{1/2}-$consistency) and in section 3.3.3 (weak convergence of the corresponding empirical process).

3.3.1 Definition of the estimator

By the definition of the copula function $\mathfrak{C}$, we have

$$\mathfrak{C}(u, v) = F(F_1^{-1}(u), F_2^{-1}(v)), \quad 0 \leq u, v \leq 1, \tag{3.6}$$

where L^{-1} denotes the generalized inverse of a monotone function L. Therefore, the copula function can be estimated nonparametrically by considering an empirical version of (3.6), that is

$$\mathfrak{C}_n(u, v) = \mathbb{F}_n(\mathbb{F}_{1n}^{-1}(u), \mathbb{F}_{2n}^{-1}(v)), \quad 0 \leq u, v \leq 1, \tag{3.7}$$

where $\mathbb{F}_n$ is defined in (3.2) and $\mathbb{F}_{1n}(t_1) = \mathbb{F}_n(t_1, \infty)$, and $\mathbb{F}_{2n}(t_2) = \mathbb{F}_n(\infty, t_2)$. In the uncensored case, this definition reduces to that of the empirical copula estimator introduced in [Deheuvels, 1979]. If $\mathbb{F}_n$ is a true distribution function (that is, a monotonic function with $\mathbb{F}_n(+\infty, +\infty) = 1$), then $\mathfrak{C}_n$ is a true copula function. For the examples considered in the previous sections, in some situations, the total mass of $\mathbb{F}_n$ may be strictly less than one (this is a classical issue for Kaplan-Meier

estimator in the univariate case), leading to estimators $\mathfrak{C}_n$ with total mass less than one. In this case, the residual mass may be affected to the point $(1,1)$ in order to retrieve a true copula function. In the case where $\mathbb{F}_n$ allocates negative mass to some observations (for example the estimator of [Dabrowska, 1988]), this definition can not be used since $\mathbb{F}_{jn}$ (for $j=1,2$) may not be monotonic and $\mathbb{F}_{jn}^{-1}$ may not be defined. Nevertheless, (3.7) is asymptotically equivalent, up to the terms of order $O_P(1/n)$ uniformly on u and v, to

$$\tilde{\mathfrak{C}}_n(u,v) = \frac{1}{n}\sum_{i=1}^{n} W_{in} \mathbb{1}_{\mathbb{F}_{1n}(Y_{1i})\leq u, \mathbb{F}_{2n}(Y_{2i})\leq v}, \tag{3.8}$$

which is still valid for non-monotonic functions $\mathbb{F}_{jn}$ (although, in this case, $\mathfrak{C}_n$ is not a copula).

In the following sections 3.3.2 and 3.3.3, we study the asymptotic behavior of the estimator $\mathfrak{C}_n$. Compared to the uncensored case, the main difficulty here is to handle the weights W_{in}. Indeed, in absence of censoring, one observes a sample composed of i.i.d. observations, each of them contributing to the estimator with the same weight equal to n^{-1}, while in our setting each weight is random and depends on the whole sample.

3.3.2 Uniform $n^{1/2}-$consistency

As in the uncensored case, asymptotic properties of the empirical copula estimator are derived from the corresponding properties of the underlying distribution function estimator $\mathbb{F}_n$. Thus, the $n^{1/2}-$consistency of $\mathfrak{C}_n$ only requires the $n^{1/2}-$convergence of $\mathbb{F}_n$, as stated in the following Theorem.

Theorem 2. Assume, without loss of generality, that (T_1, T_2) are almost surely positive, and assume that $\mathbb{F}_{jn}$ is monotonic for $j=1,2$. Let $\mathcal{T}_1 = [-\infty, A_1]$, and $\mathcal{T}_2 = [-\infty, A_2]$, such that

$$\sup_{t_1\in\mathcal{T}_1, t_2\in\mathcal{T}_2} |\mathbb{F}_n(t_1,t_2) - F(t_1,t_2)| = O_P(n^{-1/2}),$$

and, for $j=1,2$,

$$\sup_{t\in\mathcal{T}_j} |\mathbb{F}_{jn}(t) - F_j(t)| = O_P(n^{-1/2}).$$

Moreover, assume that $\sup_{i=1,\dots,n: Y_{1i}\leq A_1, Y_{2i}\leq A_2} W_{in} = O_P(1)$. Denoting $F_1(\mathcal{T}_1) = [0,u_1]$, and $F_2(\mathcal{T}_2) = [0,u_2]$, then, for any $\eta > 0$,

$$\sup_{u\leq u_1-\eta, v\leq u_2-\eta} |\mathfrak{C}_n(u,v) - \mathfrak{C}(u,v)| = O_P(n^{-1/2}).$$

Proof. Let $\varepsilon < \eta$ and consider that we are on the event $\mathcal{A}_\varepsilon = \{\sup_{t\in\mathcal{T}_j, j=1,2} |\mathbb{F}_{jn}(t) - F_j(t)| \leq \varepsilon\}$. Note that,

$$\sup_{u\leq u_1-\eta, v\leq u_2-\eta} |\mathfrak{C}_n(u,v) - \tilde{\mathfrak{C}}_n(u,v)| \leq \frac{2}{n} \sup_{i=1,\dots,n: Y_{1i}\leq A_1, Y_{2i}\leq A_2} W_{in},$$

where $\tilde{\mathfrak{C}}_n$ is defined in (3.8).Therefore, it suffices to prove the uniform consistency of $\tilde{\mathfrak{C}}_n$. Next, observe that, on $\mathcal{A}_\varepsilon$, we have for $j = 1, 2$,

$$|\mathbb{1}_{\mathbb{F}_{jn}(Y_{ji})\leq u} - \mathbb{1}_{F_j(Y_{ji})\leq u}| \leq \mathbb{1}_{|F_j(Y_{ji})-u|\leq\varepsilon},$$

for $Y_{1i} \leq A_1$ and $Y_{2i} \leq A_2$. Defining, for $u \leq u_1 - \eta$ and $v \leq u_2 - \eta$,

$$\mathbb{F}_n^*(u, v) = \frac{1}{n}\sum_{i=1}^{n} W_{in}\mathbb{1}_{F_1(Y_{1i})\leq u}\mathbb{1}_{F_2(Y_{2i})\leq v},$$

we can deduce that, on $\mathcal{A}_\varepsilon$,

$$|\tilde{\mathfrak{C}}_n(u, v) - \mathbb{F}_n^*(u, v)| \leq \frac{2}{n}\sum_{i=1}^{n} W_{in}\left\{\mathbb{1}_{|F_1(Y_{1i})-u|\leq\varepsilon} + \mathbb{1}_{|F_2(Y_{2i})-v|\leq\varepsilon}\right\}\mathbf{1}_{Y_{1i}\leq A_1, Y_{2i}\leq A_2}.$$

The presence of $\mathbf{1}_{Y_{1i}\leq A_1, Y_{2i}\leq A_2}$ in the last equation is due to the fact that only the terms with $Y_{ji} \leq A_j$ (for $j = 1, 2$) give a positive contribution to the sum that defines $\mathbb{F}_n^*$ and $\tilde{\mathfrak{C}}_n$. This is clear for $\mathbb{F}_n^*$. For $\tilde{\mathfrak{C}}_n$, observe that, due to monotonicity, $Y_{ji} > A_j$ implies $\mathbb{F}_{jn}(Y_{ji}) \geq \mathbb{F}_{jn}(A_j) \geq F_j(A_j) - \varepsilon$. Moreover,

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} W_{in}\left\{\mathbb{1}_{|F_1(Y_{1i})-u|\leq\varepsilon} + \mathbb{1}_{|F_2(Y_{2i})-v|\leq\varepsilon}\right\} &= \mathbb{F}_{1n}(F_1^{-1}(u+\varepsilon)) - \mathbb{F}_{1n}(F_1^{-1}(u-\varepsilon)) \\ &\quad + \mathbb{F}_{2n}(F_2^{-1}(v+\varepsilon)) - \mathbb{F}_{2n}(F_1^{-1}(v-\varepsilon)).\end{aligned}$$

Using the uniform convergence of $\mathbb{F}_{jn}$ for $j = 1, 2$, we can deduce that, on $\mathcal{A}_\varepsilon$,

$$|\tilde{\mathfrak{C}}_n(u, v) - \mathbb{F}_n^*(u, v)| \leq 8\varepsilon,$$

Next, we take $\varepsilon = 8^{-1}Mn^{-1/2}$. We get, for n large enough, $\mathbb{P}(n^{1/2}\sup_{u\leq u_1-\eta, u\leq u_2-\eta}|\tilde{\mathfrak{C}}_n(u, v) - \mathbb{F}_n^*(u, v)| > M) \leq \mathbb{P}(\mathcal{A}^c_{8^{-1}Mn^{-1/2}})$, where $\mathcal{A}^c$ denotes the complementary of the set $\mathcal{A}$. We then can deduce that $\lim_{M\to\infty}\bar{\lim}_{n\to\infty}\mathbb{P}(n^{1/2}\sup_{u\leq u_1-\eta, u\leq u_2-\eta}|\tilde{\mathfrak{C}}_n(u, v) - \mathbb{F}_n^*(u, v)| > M) = 0$. Moreover, $\mathbb{F}_n^*$ converges uniformly towards $\mathfrak{C}$ at rate $n^{1/2}$ for $u \leq u_1 - \eta$ and $v \leq u_2 - \eta$ from the rate of uniform convergence of $\mathbb{F}_n$.

3.3.3 Weak convergence of the censored empirical copula process

To obtain asymptotic weak convergence of $\mathfrak{C}_n$, some additional properties on the estimator $\mathbb{F}_n$ are required. First of them is the weak convergence of the empirical process $n^{1/2}(\mathbb{F}_n(t_1, t_2) - F(t_1, t_2))$, which is stated in Assumption 4. Let $l^\infty(\mathfrak{T})$ denote a space of all bounded functions $f : \mathfrak{T} \to \mathbb{R}$.

Assumption 4. Assume that $\mathbb{F}_n(t_1, t_2)$ is an $n^{1/2}-$consistent estimator of $F(t_1, t_2)$ satisfying

$$\mathbb{H}_n(t_1, t_2) := \sqrt{n}(\mathbb{F}_n(t_1, t_2) - F(t_1, t_2)) \rightsquigarrow \mathbb{G}_F(t_1, t_2) \quad \text{in} \quad l^\infty(\mathbb{R}^2), \tag{3.9}$$

where $\mathbb{G}_F(t_1, t_2)$ is a tight gaussian process and $\rightsquigarrow$ denotes the weak convergence.

The next requirement is related to the fact that the copula estimator (3.7) must be invariant under the probability integral transform, see [Fermanian et al., 2004] in the uncensored case. To express this invariance in our framework, define the pseudo-variables $(T_1^*, T_2^*, C_1^*, C_2^*)$ as

$$(T_1^*, T_2^*, C_1^*, C_2^*) = (F_1(T_1), F_2(T_2), F_1(C_1), F_2(C_2)),$$

and, for $j = 1, 2$,

$$Y_j^* = \min(T_j^*, C_j^*),\ \delta_j^* = \mathbb{I}_{T_j^* \leq C_j^*}.$$

We will denote by $(Y_{1i}^*, Y_{2i}^*, \delta_{1i}^*, \delta_{2i}^*)_{1\leq i\leq n}$ a corresponding i.i.d. sample. As F_1 and F_2 are monotonic, $\delta_j^* = \delta_j$. Joint distribution function of (T_1^*, T_2^*) is to be estimated by $\mathbb{F}_n^*$, which is the estimator calculated by the same way as (3.2), but using the pseudo-sample $(Y_{1i}^*, Y_{2i}^*, \delta_{1i}^*, \delta_{2i}^*)_{1\leq i\leq n}$.

Assumption 5. The weights attributed by the estimator (3.2) are invariant under the probability integral transform, i.e.

$$W_{in}^* = W_{in}, \quad i = 1, \ldots, n.$$

In all the examples that we consider, for each i, the weight W_{in} depends only on the indicators $(\delta_{1i}, \delta_{2i})$, and on the ranks R_{ji}, for $j = 1, 2$, of the observations Y_{ji} in samples $(Y_{j1}, ..., Y_{jn})$. Denote R_{ji}^* the rank of Y_{ji}^* in the transformed sample $(Y_{j1}^*, ..., Y_{jn}^*)$, then the monotonicity of F_1 and F_2 ensures that $R_{ji}^* = R_{ji}$. Therefore, Assumption 5 is naturally satisfied in the examples that we consider.

From Assumptions 4 and 5, the process $n^{1/2}(\mathbb{F}_n^*(u, v) - \mathfrak{C}(u, v))$ converges in $l^\infty([0, 1]^2)$ to the gaussian process $\mathbb{Z}_C^*(u, v) = \mathbb{G}_F(F_1^{-1}(u), F_2^{-1}(v))$. Defining the empirical process corresponding to the introduced copula estimator by

$$\mathbb{Z}_n(u, v) = \sqrt{n}(\mathfrak{C}_n(u, v) - \mathfrak{C}(u, v)), \qquad 0 \leq u, v \leq 1, \tag{3.10}$$

we now state the main result of this section.

Theorem 3. Suppose that F has continuous marginal distribution functions and partial derivatives of its copula function exist and are continuous. Then the censored empirical copula process $\{\mathbb{Z}_n(u, v),\ 0 \leq u, v \leq 1\}$ converges weakly in $l^\infty([0, 1]^2)$ to the tight Gaussian process,

$$\mathbb{Z}_{\mathfrak{C}}(u, v) = \mathbb{Z}_{\mathfrak{C}}^*(u, v) - \partial_1\mathfrak{C}(u, v)\mathbb{Z}_{\mathfrak{C}}^*(u, 1) - \partial_2\mathfrak{C}(u, v)\mathbb{Z}_{\mathfrak{C}}^*(1, v).$$

Theorem 3 is an extension of Theorem 3 in [Fermanian et al., 2004], which establishes, in absence of censoring, the weak convergence of the empirical copula process in $l^\infty([0, 1]^2)$. The arguments that we develop are similar to those used in the uncensored case and are based mainly on functional Delta-Method.

Proof. The first step of the proof consists of reducing the problem to the case where the marginals T_1 and T_2 are uniformly distributed. This is done through the following Lemma (proved in the Appendix section), which is a consequence of Assumption 5.

Lemma 2. Assume that F_1 and F_2 are continuous distribution functions and let $\mathfrak{C}_n(u,v)$ and $\mathfrak{C}_n^*(u,v)$ be the estimators of $\mathfrak{C}$ based respectively on the observations $(Y_{1i}, Y_{2i}, \delta_{1i}, \delta_{2i})_{1\leq i\leq n}$ and $(Y_{1i}^*, Y_{2i}^*, \delta_{1i}^*, \delta_{2i}^*)_{1\leq i\leq n}$. The equation $\mathfrak{C}_n(u,v) = \mathfrak{C}_n^*(u,v)$ holds.

From Lemma 2, it follows that $\mathbb{Z}_n(u,v) = \mathbb{Z}_n^*(u,v) = \sqrt{n}(\mathfrak{C}_n^*(u,v) - \mathfrak{C}(u,v))$, where $\mathbb{C}_n^*(u,v) = \mathbb{F}_n^*(\mathbb{F}_{1n}^{*-1}(u), \mathbb{F}_{2n}^{*-1}(v))$, where $\mathbb{F}_{1n}^*, \mathbb{F}_{2n}^*$ are the marginals of $\mathbb{F}_n^*$. Then Lemma 2 of [Fermanian et al., 2004] can be applied with $H^*(u,v) = \mathfrak{C}(u,v)$. The limiting process is then obtained by applying the Delta-Method and Theorem 3.9.28 in [van der Vaart and Wellner, 1996].

Examples 1 to 3 (continued). As we have already mentioned, Assumption 5 is quite natural and satisfied for all examples that we consider. Let us discuss now Assumption 4. In each example that we give, the difference between the distribution function and its estimator of the form (3.2) can be represented as

$$\mathbb{F}_n^{(j)}(t_1,t_2) - F^{(j)}(t_1,t_2) = \frac{1}{n}\sum_{i=1}^{n} \eta_i(t_1,t_2) + R_n^j(t_1,t_2),$$

with $\sup_{t_1,t_2} |R_n^j(t_1,t_2)| = o_P(n^{-1/2})$, where $n^{-1/2}\sum_{i=1}^n \eta_i(t_1,t_2)$ is a sum of i.i.d. terms, which converges to a Gaussian process from the empirical process theory, and $j = 1, ..., 3$. For Example 1, such a representation has been derived by [Stute, 1996] for a fixed point (t_1,t_2). The uniform convergence of the remainder term can be seen as a particular case of the results obtained by [Lopez and Saint-Pierre, 2012], which also cover our Example 2. [Gribkova et al., 2013] provide the representation for Example 3. In each case, the uniform convergence of the remainder term is obtained by adding some assumptions on the tail of the distributions of two lifetimes, compared to the tail of the distributions of the censoring variables.

3.4 Smoothed copula estimators

The procedure described in section 3.2 introduces a discrete nonparametric estimator of the copula function. However, if the underlying copula is continuous or even smooth, it may be reasonable to estimate it by a smooth function. In section 3.4.1, we introduce two nonparametric smooth copula estimators, which are valid in presence of censored observations. By taking their derivatives, we deduce two copula density estimators. Section 3.4.2 establishes the weak convergence of the empirical processes associated with the smooth copula estimators. Section 3.4.3 deals with the uniform convergence of the resulting copula density estimators.

3.4.1 Smooth estimators of the copula and its density

Let $k : \mathbb{R} \mapsto \mathbb{R}$ be a kernel function (that is a smooth function with integral over $\mathbb{R}$ equal to one) and $K(x) = \int_{-\infty}^{x} k(u)du$ its cumulative integral. Introducing a smoothing parameter $h > 0$, a classical kernel estimator of the bivariate distribution function of multiplicative form is defined through a convolution of the empirical measure with the measure of density $h^{-2}k(u/h)k(v/h)$.

For the censored data, we replace the empirical measure by the measure defined by the estimator (3.2), which leads to a natural extension of the classical Parzen-Rosenblatt estimator. This leads us to a following kernel distribution function estimator of multiplicative form,

$$\hat{\mathbb{F}}_n^1(t_1,t_2) = \frac{1}{n}\sum_{i=1}^{n} W_{in} K_h\left(t_1 - Y_{1i}\right) K_h(t_2 - Y_{2i})\,, \tag{3.11}$$

where we introduced a notation $K_h(x) := K(x/h)$. Although we use, for notation convenience, the same h for both components, different bandwidths may be used. Like all smoothing techniques, this estimator is sensitive to the choice of the bandwidth parameter(s). This question will be discussed in section 3.5. Let us introduce now a first smooth copula estimator given by

$$\hat{\mathfrak{C}}_n^1(u,v) = \hat{\mathbb{F}}_n^1((\hat{\mathbb{F}}_{1n}^1)^{-1}(u), (\hat{\mathbb{F}}_{2n}^1)^{-1}(v)). \tag{3.12}$$

In the uncensored case, this estimator reduces to the kernel estimator studied by [Fermanian et al., 2004], who established a functional central limit theorem for the associated empirical process.

In order to ensure the weak convergence of the estimator $\hat{\mathfrak{C}}_n^1$, one must control its bias through an assumption on the boundedness of the second order partial derivatives of the underlying joint distribution function F. Let us notice that these regularity conditions do not impose the uniform boundedness of the second order derivatives of the corresponding copula function itself, which would have excluded from consideration several important families of copulas. An important drawback of estimator (3.12) is that its performance depends on marginal distribution functions of the variables T_1 an T_2 (this issue was extensively discussed in [Omelka et al., 2009]). To get rid of this inconvenient, [Omelka et al., 2009] introduced some transformation of the initial variables. It leads to construction of a kernel estimator of the distribution function F of (T_1, T_2) based on pseudo-observations, whose marginal distributions are asymptotically equal to some distribution function Φ, designed to avoid corner bias problems. This method can be extended to our framework, leading to a second smooth estimator $\mathfrak{C}_n^2$ of the copula function.

Indeed, for some distribution function Φ, let us consider a couple of variables $(\tilde{T_1}, \tilde{T_2}) = (\Phi^{-1}[F_1(T_1)], \Phi^{-1}[F_2(T_2)])$ and pseudo-observations $(\Phi^{-1}[\mathbb{F}_{1n}(Y_{1i})], \Phi^{-1}[\mathbb{F}_{2n}(Y_{2i})])_{1\le i\le n}$. Since the copula function is invariant under monotone transformations, the variables $(\tilde{T_1}, \tilde{T_2})$ are coupled by the same copula as (T_1, T_2). Next, we define an estimator of the joint distribution function of $(\tilde{T_1}, \tilde{T_2})$ by

$$\hat{\mathbb{F}}_n^2(t_1,t_2) = \frac{1}{n}\sum_{i=1}^{n} W_{in} K_h\left(t_1 - \Phi^{-1}[\mathbb{F}_{1n}(Y_{1i})]\right) K_h\left(t_2 - \Phi^{-1}[\mathbb{F}_{2n}(Y_{2i})]\right),$$

where $\mathbb{F}_{1n}$ (resp. $\mathbb{F}_{2n}$) are marginal distributions of estimator (3.2). Then, the corresponding copula estimator is defined as,

$$\hat{\mathfrak{C}}_n^2(u,v) = \hat{\mathbb{F}}_n^2(\Phi^{-1}(u), \Phi^{-1}(v)). \tag{3.13}$$

Since these copula estimators are smooth, two estimators of the copula density $c(u,v)$ can be deduced by considering

$$\hat{c}^i(t_1,t_2) = \frac{\partial^2}{\partial_{t_1}\partial_{t_2}} \hat{\mathfrak{C}}_n^i(t_1,t_2), \tag{3.14}$$

for $i = 1, 2$.

3.4.2 Functional CLT for the smooth copula estimators

The proof of the weak convergence of the copula processes associated with the defined smooth estimators relies on asymptotic properties of the weights W_{in} and of the estimators of type (3.2). We first state some key assumptions that will allow, in our proofs, the replacement of the weights W_{in} by their limit quantities W_i (up to some additional terms).

Assumption 6. Let us recall that $W_{in} = \delta_{1i}\delta_{2i}\hat{g}(Y_{1i}, Y_{2i})$, where $\hat{g}(y_1, y_2)$ is a random function converging to its deterministic counterpart $g(y_1, y_2)$ satisfying (3.3). For $j = 1, 2$, denote $\tau_j = \inf\{t : F_j(t) = 1\}$. Assume that, on every set $\mathcal{Y} = [-\infty, t_1] \times [-\infty, t_2]$ with $t_1 < \tau_1$ and $t_2 < \tau_2$, g is bounded and is twice continuously differentiable with respect to its arguments with uniformly bounded partial derivatives up to order two. Moreover, assume that :

1. $\sup_{t \le t_1, u \le t_2} |\hat{g}(t_1, t_2) - g(t_1, t_2)| = O_P(n^{-1/2})$, and that the restrictions of $\hat{g}$ and g to the set $\{(t, u) : t \le t_1, u \le t_2\}$ both belong to Donsker classes of functions ;
2. for all $\psi \in \mathcal{F}$, where $\mathcal{F}$ denotes a Donsker class of bounded functions such that $\psi(y_1, y_2) = 0$ for $y_1 > t_1$ or $y_2 > t_2$, we have the following representation,

$$\frac{1}{n}\sum_{i=1}^{n}[W_{in} - W_i]\psi(Y_{1i}, Y_{2i}) = \frac{1}{n}\sum_{i=1}^{n}\eta^{\psi}(Y_{1,i}, Y_{2,i}, \delta_{1i}, \delta_{2i}) + R_n(\psi), \qquad (3.15)$$

   with $\sup_{\psi \in \mathcal{F}} |R_n(\psi)| = o_P(n^{-1/2})$, and $E[\eta^{\psi}(Y_{1,i}, Y_{2,i}, \delta_{1i}, \delta_{2i})] = 0$ for all ψ;
3. $\{(t_1, t_2, d_1, d_2) \to \eta^{\psi_{h,y_1,y_2}}(t_1, t_2, d_1, d_2) : h \in [0, 1/4], (y_1, y_2) \in \mathcal{Y}\}$ is a Donsker class of functions, with $\sup_{(y_1,y_2)\in\mathcal{Y}} E[(\eta^{\psi_{h,y_1,y_2}}(Y_{1,i}, Y_{2,i}, \delta_{1i}, \delta_{2i}))^2] \to_{h\to 0} 0$, defining

$$\begin{aligned}\phi_{h,y_1,y_2}(t_1, t_2) &= K_h(y_1 - t_1)K_h(y_2 - t_2), \\ \psi_{h,y_1,y_2}(t_1, t_2) &= \phi_{h,y_1,y_2}(t_1, t_2) - \mathbb{1}_{t_1 \le y_1, t_2 \le y_2}.\end{aligned}$$

I.i.d. representations of the type (3.15) are classical tools when it comes to studying the asymptotic properties of the estimators (3.2). Moreover, the previous assumptions are valid for all the practical examples that we consider. Indeed, the desired representations were obtained in Theorem 3.3 of [Lopez and Saint-Pierre, 2012] for Examples 1 and 2 and in Theorem 3.1 of [Gribkova et al., 2013] for Example 3. The uniform convergence of $\hat{g}$ is, in all the examples, a consequence of the uniform consistency of the Kaplan-Meier type estimators on compact sets. In each case, the restrictions of $\hat{g}$ and g to compact sets can easily seen to be in Donsker class of functions from Theorem 2.7.5 in [van der Vaart and Wellner, 1996]. Point 3 is more technical, but is reasonable, having in mind the particular shape of function η^{ψ} coming from the Examples. We refer to section 3.6.5 to see how this Assumption can be checked in the Examples.

Assumption 6 would be sufficient if we restrain ourselves to proving the convergence of $\hat{\mathfrak{C}}_n^i$ on $[0, a] \times [0, b]$ for a and b strictly less than 1. Indeed, in this case, the

convergence is not affected by the observations with Y_{1i} close to τ_1 or Y_{2i} close to τ_2, which give no contribution to the value of $\hat{\mathfrak{C}}_n^i(u,v)$ for $u < a$ and $v < b$. These large observations (close to the tail of at least one of the marginal distributions) give rise to particular difficulties, which are similar to those encountered in the univariate case. For Kaplan-Meier estimator, i.i.d. representations of the same type as in Assumption 6 can be obtained under standard conditions if one avoids investigating the right tail of the distribution (see e.g. [Gijbels and Veraverbeke, 1991]). To obtain representations valid on the whole real line, some assumptions on the distribution of the lifetime and of the censoring are required, as in [Stute, 1995]. This is the purpose of our Assumption 7 below.

Assumption 7. Assume that $E[\delta_1\delta_2 g(T_1,T_2)^2] < \infty$, and assume that there exist i.i.d. random variables (Z_i) such that

$$\sup_i |W_{in} - W_i| \leq A_n Z_i,$$

where $A_n = O_P(n^{-1/2})$ and $E[Z_i] < \infty$.

Examples 1 to 3 (continued). In each of the examples that we consider, a decomposition of the weights into $A_n Z_i$, as in Assumption 7, can be obtained. The rate of convergence of A_n can be deduced from Theorem 2.1 in [Gill, 1983]. Indeed, if we denote by $\hat{G}_j$ a Kaplan-Meier estimator of the cdf G_j of censoring variable C_j, and by $Y_{[j,n]}$ the largest observation for the variable j, the indicated theorem implies

$$\sup_{t \leq Y_{[j,n]}} \left| \frac{h(t)\{\hat{G}_j(t) - G_j(t)\}}{1 - \hat{G}_j(t)} \right| = O_P(n^{-1/2}), \tag{3.16}$$

for any function h such that $\int h(t)^2 dL_{F_j,G_j}(t) < \infty$, where $L_{F_j,G_j}(t) = \int_0^t [(1 - F_j(u))(1-G_j(u))^2]^{-1} dG_j(u)$, with F_j denoting the cdf of the corresponding lifetime T_j (here, the roles of T_j and C_j are reversed compared to Theorem 2.1 in [Gill, 1983]).

In each example that we considered, the difference between the weights W_{in} and W_i involves the difference of Kaplan-Meier estimators of the distribution of the censoring and their limits. This difference converges at rate $n^{-1/2}$ uniformly on the whole distribution support only if, according to (3.16), it is multiplied by an appropriate function h decreasing fast enough. Hence, in each case that we consider, the term A_n is of the form (3.16) for some appropriate h. This function h also naturally appears in the Z_i that we obtain.

More precisely, in Example 1, write, for some $\varepsilon > 0$ arbitrary small,

$$W_{in} - W_i = \left\{ \frac{\hat{G}(Y_i-) - G(Y_i-)}{[1 - \hat{G}(Y_i-)] L_{F_1,G}^{1/2+\varepsilon}(Y_i)} \right\} \frac{\delta_i L_{F_1,G}^{1/2+\varepsilon}(Y_i)}{1 - G(Y_i-)}.$$

Assumption 7 is valid for

$$A_n^1 = \sup_{t \leq Y_{[j,n]}} \left\{ \frac{|\hat{G}(t) - G(t)|}{L_{F_1,G}^{1/2+\varepsilon}(t)[1 - \hat{G}(t)]} \right\},$$

where $\varepsilon > 0$, and $Z_i^1 = \delta_{1i} L_{F_1,G}^{1/2+\varepsilon}(Y_i)[1 - G(Y_{1i})]^{-1}$. The condition $E[Z_i^1] < \infty$ can be expressed as

$$E[Z_i^1] = \int L_{F_1,G}^{1/2+\varepsilon}(t) dF_1(t) < \infty.$$

This condition is similar to the one in [Stute, 1995]. Similarly, for Example 2, assuming that $\mathcal{C}_G(u,v) \geq u^{\alpha_1} v^{\alpha_2}$ and that its first order partial derivatives are bounded, let us take

$$A_n^2 = \sup_{t \leq Y_{[j,n]}} \left\{ \sum_{j=1}^{2} \frac{|\hat{G}_j(t) - G_j(t)|}{L_{F_j,G_j}^{1/2+\varepsilon}(t)[1 - \hat{G}_j(t)]} \right\},$$

and

$$Z_i^2 = \frac{\delta_{1i}\delta_{2i}}{\mathcal{C}_G(1 - G_1(Y_{1i}), 1 - G_2(Y_{2i}))} \left\{ \frac{(1 - G_1(Y_{1i}))^{1-\alpha_1} L_{F_1,G_1}^{1/2+\varepsilon}(Y_{1i})}{(1 - G_2(Y_{2i}))^{\alpha_2}} \right\}$$
$$+ \frac{(1 - G_2(Y_{2i}))^{1-\alpha_2} L_{F_2,G_2}^{1/2+\varepsilon}(Y_{2i})}{(1 - G_1(Y_{1i}))^{\alpha_1}} \Bigg\}.$$

The integrability condition can be written as

$$\int\int \left\{ \frac{(1 - G_1(t_1-))^{1-\alpha_1} L_{F_1,G_1}^{1/2+\varepsilon}(t)}{(1 - G_2(t_2-))^{\alpha_2}} + \frac{(1 - G_2(t_2))^{1-\alpha_2} L_{F_2,G_2}^{1/2+\varepsilon}(t_2)}{(1 - G_1(t_1))^{\alpha_1}} \right\} < \infty.$$

For Example 3, one can take A_n^3 similar to A_n^1, but replacing F_1 by $F^* = \mathbb{P}(A \leq t)$ in the definition of $L_{F^*,G}$, and $\hat{G}$ by $\tilde{G}$. Defining $B_i = \max(Y_{1i}, Y_{2i} - \varepsilon_i)$, we have $Z_i^3 = \delta_{i1}\delta_{i2} L_{F^*,G}^{1/2+\varepsilon}(B_i)[1 - G(B_i-)]^{-1}$, with the condition $E[Z_i^3] < \infty$ which can be rewritten as $\int L_{F^*,G}^{1/2+\varepsilon}(a) dF^*(a) < \infty$.

The main tool to obtain the weak convergence of the empirical process associated with $\mathfrak{C}_n^1$ is Theorem 4, establishing the asymptotic equivalence of the smoothed estimator $\hat{\mathbb{F}}_n^1(t_1,t_2)$ to the estimator $\mathbb{F}_n(t_1,t_2)$, up to some negligible remainder term. The proof of Theorem 4 is given in the Appendix section.

Theorem 4. Consider a symmetric kernel function k with compact support such that $k \geq 0$ and $\int u^2 k(u) du < \infty$. Let $F(t_1,t_2)$ be twice differentiable distribution function with the second order derivatives uniformly bounded on $\mathbb{R}^2$. We suppose that $h^2\sqrt{n} \to 0$. Under Assumptions 6 and 7, we have,

$$\sqrt{n} \sup_{t_1,t_2 \in \mathbb{R}} |\hat{\mathbb{F}}_n^1(t_1,t_2) - \mathbb{F}_n(t_1,t_2)| \xrightarrow[\mathbb{P}]{} 0. \tag{3.17}$$

Remark 1. The condition $h^2\sqrt{n} \to 0$ of Theorem 4 is verified for $h = O(n^{-1/3})$, which is the usual optimal rate for distribution function estimator in absence of censoring.

Corollary 2. Theorem 4 implies the weak convergence of the process $\sqrt{n}(\hat{\mathbb{F}}_n^1(t_1,t_2) - F(t_1,t_2))$ to the tight gaussian limit process $\mathbb{G}_F(t_1,t_2)$ from the Assumption 4.

Corollary 3. If the condition $h^2\sqrt{n} \to 0$ is not satisfied, it can be seen from the proof of Theorem 4 (see Appendix section), that

$$\sup_{t_1,t_2\in\mathbb{R}} |\hat{\mathbb{F}}_n^1(t_1,t_2) - \mathbb{F}_n(t_1,t_2)| = O_P(h^2).$$

Proving the convergence of $\mathfrak{C}_n^2$ follows a different path, and requires some assumptions on the function Φ involved in the transformation of the observations, and on the behavior of the partial derivatives of the copula function near the boundaries of $[0,1]^2$. These requirements (mainly same as those of [Omelka et al., 2009]) are listed in Assumption 8. The restrictions on the copula function are quite reasonable, since most of classical copula families follow this property, see Appendix D in [Omelka et al., 2009].

Assumption 8. Assume that $\mathfrak{C}$ is twice continuously differentiable on $]0,1[^2$, and that

$$\begin{aligned}
\frac{\partial^2\mathfrak{C}(u,v)}{\partial u^2} &= O\left(\frac{1}{u(1-u)}\right),\\
\frac{\partial^2\mathfrak{C}(u,v)}{\partial v^2} &= O\left(\frac{1}{v(1-v)}\right),\\
\frac{\partial^2\mathfrak{C}(u,v)}{\partial u\partial v} &= O\left(\frac{1}{\sqrt{uv(1-u)(1-v)}}\right).
\end{aligned}$$

Moreover, assume that Φ is strictly increasing distribution function with Φ' and Φ'^2/Φ bounded.

We also add an invariance properties of the weights W_{in} after transformation of the variables by Φ. Like Assumption 5, this assumption automatically holds if the weights only depend on the ranks of the lifetimes (since, for $j = 1, 2$, the transformation $\Phi^{-1}(F_j(\cdot))$ is increasing).

Assumption 9. Recall that $W_{in} = \delta_{1i}\delta_{2i}\hat{g}(Y_{1i}, Y_{2i})$, where $\hat{g}$ is some random function obtained from the sample $(Y_{1i}, Y_{2i}, \delta_{1i}, \delta_{2i})_{1\le i\le n}$. Let us consider the sample of transformed observations

$$(M_{1i}, M_{2i}, \delta_{1i}, \delta_{2i}) := (\Phi^{-1}(F_1(Y_{1i})), \Phi^{-1}(F_1(Y_{2i})), \delta_{1i}, \delta_{2i})_{1\le i\le n},$$

and the corresponding weights $W_{in}^{\Phi} := \delta_{1i}\delta_{2i}\hat{g}^{\Phi}(Y_{1i}, Y_{2i})$, where $\hat{g}^{\Phi}$ is computed by the same method as $\hat{g}$, but based on the sample of transformed observations. Assume that $W_{in}^{\Phi} = W_{in}$.

We now state the main result of this section.

Theorem 5. For $i = 1, 2$, we consider smoothed empirical copula process $\hat{\mathbb{Z}}_n^i(u,v)$ associated with the estimator $\mathfrak{C}_n^i(u,v)$,

$$\hat{\mathbb{Z}}_n^i(u,v) = \sqrt{n}(\hat{\mathfrak{C}}_n^i(u,v) - \mathfrak{C}(u,v)), \quad 0 \le u, v \le 1. \tag{3.18}$$

- Under the conditions of Theorem 4, the empirical process $\{\hat{\mathbb{Z}}_n^1(u,v), 0 \leq u,v \leq 1\}$ converges weakly in $l^\infty([0,1]^2)$ to the tight Gaussian process $\{\mathbb{Z}_{\mathfrak{C}}(u,v), 0 \leq u,v \leq 1\}$, defined in Theorem 3.
- Under Assumptions 6 to 9, for kernel function $k(x)$ verifying the conditions of Theorem 4, the empirical process $\{\hat{\mathbb{Z}}_n^2(u,v), 0 \leq u,v \leq 1\}$ converges weakly to the same gaussian process $\{\mathbb{Z}_{\mathfrak{C}}(u,v), 0 \leq u,v \leq 1\}$, provided that $h^2\sqrt{n} \to 0$.

Proof. The result for the case $i = 1$ follows directly from our Theorem 4 and technics used by [Fermanian et al., 2004] in the uncensored case. Indeed, the result obtained by [Fermanian et al., 2004] is a consequence of their Lemma 7, the stochastic equicontinuity of the process (3.18) and its finite-dimensional convergence. Arguments, used in Lemma 7 can not be applied in the presence of censoring. Thus establishing that the supremum of the difference between the discrete estimator of the distribution function and its smoothed version is of the order of $o_P(n^{-1/2})$ constitutes the main difficulty, resolved by Theorem 4. The rest of the arguments are applicable directly to our case. The proof for the case $i = 2$ is postponed to the Appendix section.

The results of Theorem 5 and Theorem 3 can be applied to construct a nonparametric goodness-of-fit test of the hypothesis $H_0 : \mathfrak{C} \in \mathcal{C}_\Theta$ against $H_1 : \mathfrak{C} \notin \mathcal{C}_\Theta$, where $\mathcal{C}_\Theta = \{\mathfrak{C}_\theta, \theta \in \Theta\}$ is some parametric class of copulas. The test procedure, which will be explained in details in section 3.5.2.1, is analogous to the test, based on Deheuvels copula in the uncensored case and uses the limit distribution of the empirical process associated with the copula estimator. Other approach, which we will not develop here, is a goodness-of-fit test based on copula densities. In the incensored case, it was studied by [Fermanian, 2005]. This method can also be adapted to our framework, using density estimators (3.14).

3.4.3 Uniform consistency of copula density estimators

In this section, we derive the uniform consistency of the copula density estimators (3.14) on a compact subset of $[0,1]^2$, which we denote by $\mathcal{C} = [\theta_1, \theta_2] \times [\theta_1', \theta_2']$, where $\theta_1 > 0$, $\theta_1' > 0$, and $\theta_2 < 1$, $\theta_2' < 1$. The maximum size of the compact $\mathcal{C}$ is restricted by the following assumption.

Assumption 10. Assume that, for all $\mathcal{Y} = [-\infty, t_1] \times [-\infty, t_2]$ where $t_1 < \tau_1$ and $t_2 < \tau_2$,

$$\sup_{i:(Y_{1i},Y_{2i})\in\mathcal{Y}} |W_{in} - W_i| = o_P(\eta_n), \quad \eta_n = h^2 + \frac{[\log n]^{1/2}}{h\sqrt{n}}.$$

Examples 1 to 3 (continued). This assumption holds for all examples due to the uniform convergence of the Kaplan-Meier estimators (of the censoring variables) involved in the definition of W_{in}, on sets that do not contain the tail of the distribution. A regularity assumption on $\mathcal{C}_G$ (first derivatives uniformly bounded) must be added in Example 2. In this case, it is easy to check that $\sup_{i:(Y_{1i},Y_{2i})\in\mathcal{Y}} |W_{in} - W_i| = O_P(n^{-1/2})$.

The next assumption serve for controlling the denominator, appearing in the derivatives of kernel copula estimator $\hat{\mathfrak{C}}_n^1$.

Assumption 11. Assume that there exist a constant c such that

$$\inf_{x\in\mathcal{C}} f_i(F_i^{-1}(x)) > c, \quad i = 1, 2.$$

Moreover, assume that the density f of (T_1, T_2) is twice continuously differentiable with partial derivatives up to order two uniformly bounded.

We now state the main result of this section, which is proved in the Appendix section.

Theorem 6. Recall that

$$\eta_n = h^2 + \frac{[\log n]^{1/2}}{h\sqrt{n}}.$$

- Under Assumptions 10 and 11, for a kernel function k satisfying the assumptions of Theorem 4, and for h such that $hn^\alpha \to 0$ for some $\alpha > 0$, and $nh^2[\log n]^{-1} \to \infty$, we have

$$\sup_{(u,v)\in\mathcal{C}} |\hat{c}_1(u, v) - c(u, v)| = O_P(\eta_n). \tag{3.19}$$

- Assume that Assumption 8 holds and
  1. The kernel function k is four times continuously differentiable,
  2. The function $(x, y) \to c(\Phi(x), \Phi(y))\Phi'(x)\Phi'(y)$ is C^2 on every compact set.

  If $nh^{10/3} \to \infty$, then we have

$$\sup_{(u,v)\in\mathcal{C}} |\hat{c}_2(u, v) - c(u, v)| = O_P(\eta_n). \tag{3.20}$$

3.5 Simulation study and real data analysis

This section is divided into two main parts. In the first one (section 3.5.1) we present the results of a simulation study of our estimators. Their performance is evaluated on censored datasets, simulated using several parametric copula models. The second part (section 3.5.2) is devoted to real data applications. A goodness-of-fit test based on our estimators is defined in section 3.5.2.1. In section 3.5.2.2, we study a dataset where only one variable is censored. In section 3.5.2.3, our nonparametric copula estimation techniques are applied to a joint survival dataset from a Canadian insurer.

3.5.1 Simulation study

To investigate the finite sample behavior of our estimators we carried out simulation studies in different settings. We consider Model 1, illustrating our Example 1 (where only one variable is censored) and Model 2, corresponding to Example 2 with two censored variables and an assumption on the joint distribution of the censoring. We do not present results from Example 3 (censoring variables are linked through an additional variable), since they are quite similar to those of Example 1 and 2.

3.5.1.1 Simulation scheme

In each setting, we simulate 1000 bivariate samples of size $n = 200$ according to the simulation schemes described below.

Distribution of lifetimes.

- The marginal distribution of T_1 and T_2 are simulated according to Weibull distributions with the shape parameters $k_1 = 2, k_2 = 2.2$ and the scale parameters $\lambda_1 = 3.1, \lambda_2 = 4.1$.
- To model the dependence structure, we consider three Archimedean copula families : Clayton, Frank and Gumbel (see Table 3.1 for the expressions of corresponding copulas). For each family, we consider two values of the dependence parameter, corresponding to Kendall's τ coefficients equal to $\tau_1 = 0.25$ and $\tau_2 = 0.75$. These values are summarized in Table 3.2.

Model	$\mathfrak{C}_\theta(u, v)$
Clayton	$\max(u^{-\theta} + v^{-\theta} - 1, 0)^{-1/\theta}$
Frank	$-\theta^{-1} \log\left(1 + \frac{(\exp(-\theta u)-1)(\exp(-\theta v)-1)}{(\exp(-\theta)-1)}\right)$
Nelsen 4.2.20	$[\log\left(\exp(u^{-\theta}) + \exp(v^{-\theta}) - e\right)]^{-1/\theta}$
Joe	$1 - [(1-u)^\theta + (1-v)^\theta - (1-u)^\theta(1-v)^\theta]^{1/\theta}$
Gumbel	$\exp[-\{(-\log u)^\theta + (-\log v)^\theta\}^{1/\theta}]$

TABLE 3.1 – Archimedean copulas

Copula	$\tau = 0.25$	$\tau = 0.75$
Clayton	0.66	6.00
Frank	2.30	14.00
Gumbel	1.33	4.00

TABLE 3.2 – Values of copula parameters.

Distribution of censoring variables.

- Censoring variables are modelized by Pareto distributions, that is, their survival function is equal to

$$\mathbb{P}(C > t) = \begin{cases} \frac{1}{(t+1)^\lambda} & \text{if } t \geq 0 \\ 1 & \text{if } t < 0. \end{cases}$$

  The values of the Pareto distribution parameters are chosen in order to achieve approximatively 25% of censored observations (for each censored marginal) in a first case, and 50% of censored observations in a second case.
- In case of Model 1, only one lifetime is censored. In Model 2, two lifetimes are censored by two independent censoring variables.

For each of the obtained censored datasets, we perform nonparametric copula estimation, using the three copula estimators which we introduced. To reduce the computational time, a fixed bandwidth is used to assess the performance of the

smooth estimators ($h = 0.2$). A data-driven choice of bandwidth is discussed in the real data applications.

3.5.1.2 Results.

In order to evaluate the performance of the estimators, we compute two distances between the estimated copula function and the true underlying copula. We first consider a Kolmogorov-Smirnov distance (KS in the following), that is

$$d_{KS}(\hat{\mathfrak{C}}, \mathfrak{C}) = \sup_{u,v} |\hat{\mathfrak{C}}(u, v) - \mathfrak{C}(u, v)|,$$

and a square-root of a quadratic integrated distance (RMSE in the following),

$$d_{RMSE}(\hat{\mathfrak{C}}, \mathfrak{C}) = \left[\int (\hat{\mathfrak{C}}(u, v) - \mathfrak{C}(u, v))^2 dudv\right]^{1/2},$$

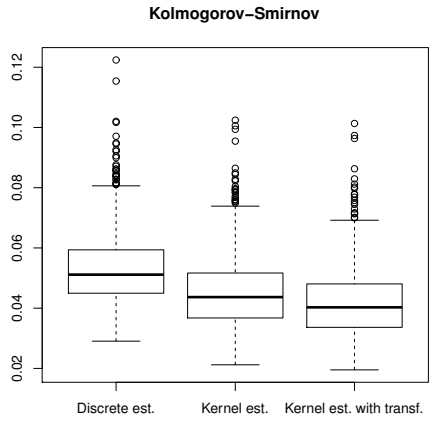
where $\hat{\mathfrak{C}}$ denotes one of the defined nonparametric copula estimators.

These distances calculated for 1000 replications are presented through boxplots on Figures 3.1 to 3.3. As the results are quite similar in several situations, we present here only the selected cases. Figure 3.1 for Frank copula permits to compare the errors of the estimation for different levels of censoring and values of Kendall's τ coefficient. As expected, estimation becomes less precise if we increase the percentage of censoring (compare (d) and (f) of Figure 3.1). The second observation is that, at fixed level of censoring, error decreases when the correlation between variables becomes stronger (compare (a),(b) of Figure 3.1 with (e),(f)). Figure 3.2 presents some results for Gumbel and Clayton copulas. Figure 3.3 illustrates Example 3, where two variables are censored. Here the situation is quite similar to the previous case.

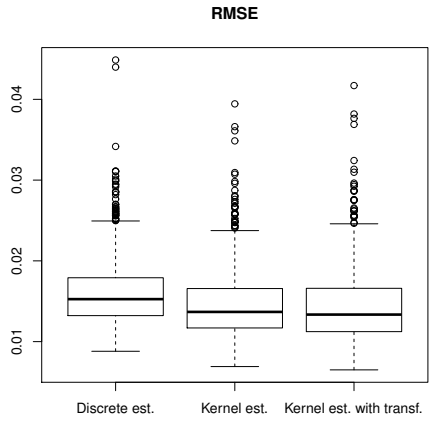
Most of the presented figures show that the performance of two kernel estimators is superior to that of the discrete estimator. These results are in accord with the results of [Omelka et al., 2009] in the uncensored case, and are natural due to the fact that the data was simulated using copulas which are smooth on $]0, 1[^2$. While the performances of two kernel estimators (3.12) and (3.13) are quite close, the interest of considering the estimator (3.13) is that the transformation of initial variables makes it less sensitive to the marginal distributions of variables, than the estimator (3.12).

3.5.2 Real data applications

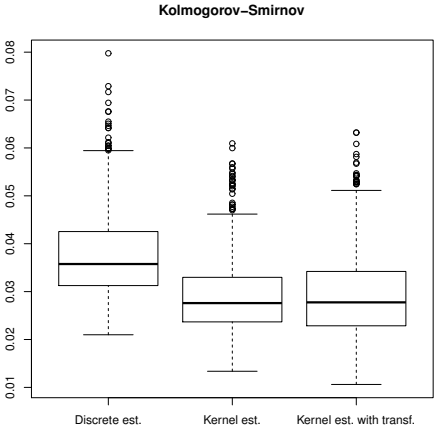
In this section we consider some applications of our results to two real data sets. We first present in section 3.5.2.1 a general method of goodness-of-fit testing of a parametric model, based on the nonparametric estimators defined in the previous sections, and a bootstrap procedure permitting to compute the $p-$values. Section 3.5.2.2 is devoted to a bivariate non life insurance data set, where one variable represents the indemnity to be paid for a claim and the other variable is the associated allocated loss adjustment expense. The second example is a life insurance data representing joint lifetimes of couples subscribed an insurance contract, which is studied in section 3.5.2.3.



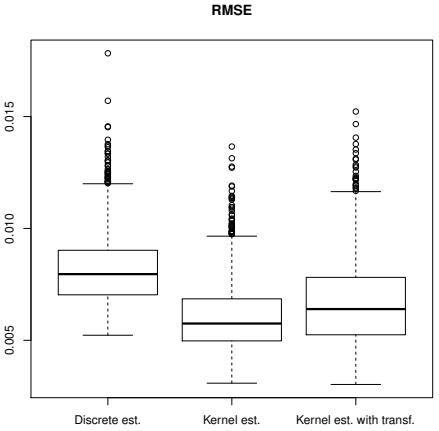
(a) Frank copula, 50% of censoring, $\tau = 0.25$



(b) Frank copula, 50% of censoring, $\tau = 0.25$

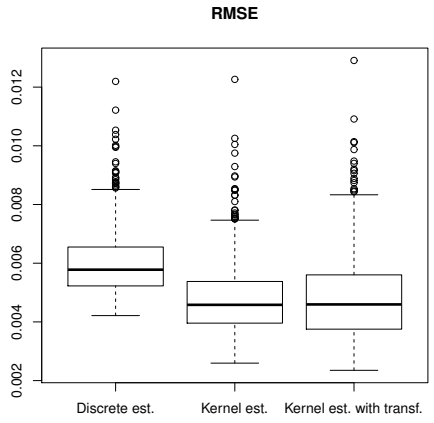


(c) Frank copula, 50% of censoring, $\tau = 0.75$



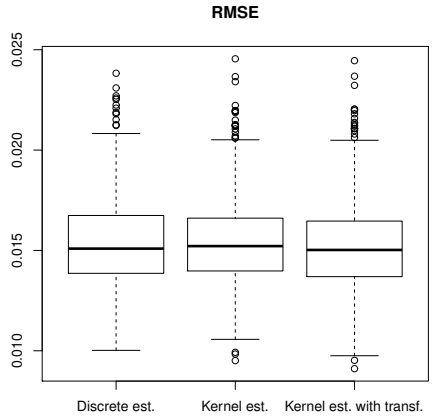
(d) Frank copula, 50% of censoring, $\tau = 0.75$

(e) Frank copula, 25% of censoring, $\tau = 0.75$

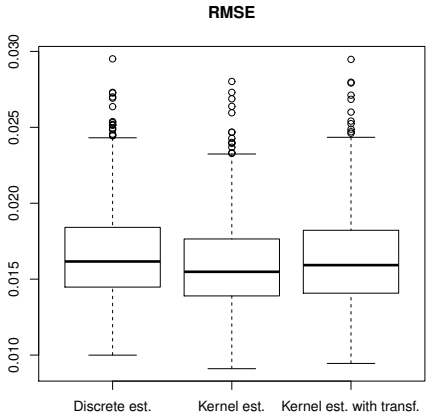


(f) Frank copula, 25% of censoring, $\tau = 0.75$

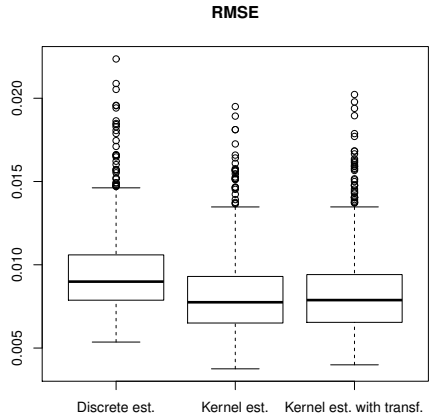
FIGURE 3.1 – Model 1 (only one variable is censored) : Frank copula



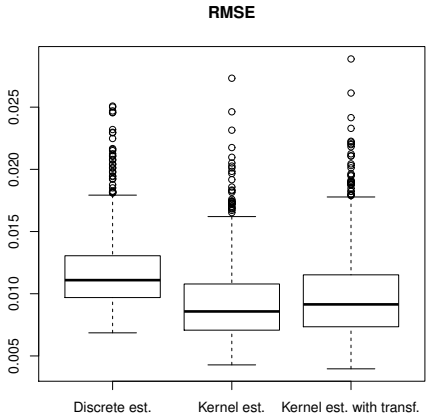
(a) Clayton copula, 25% of censoring, $\tau = 0.75$



(b) Clayton copula, 50% of censoring, $\tau = 0.75$



(c) Gumbel copula, 25% of censoring, $\tau = 0.75$



(d) Gumbel copula, 50% of censoring, $\tau = 0.75$

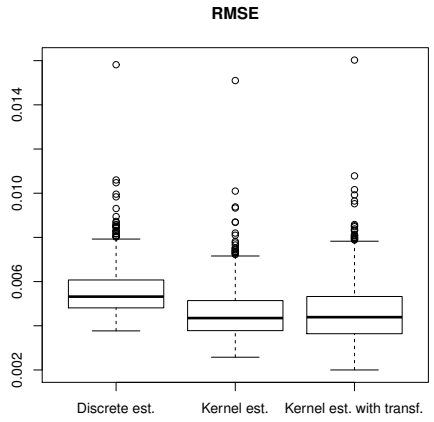
Figure 3.2 – Model 1 (only one variable is censored) : Clayton and Gumbel copulas

3.5.2.1 Goodness-of-fit procedure based on the nonparametric copula estimators

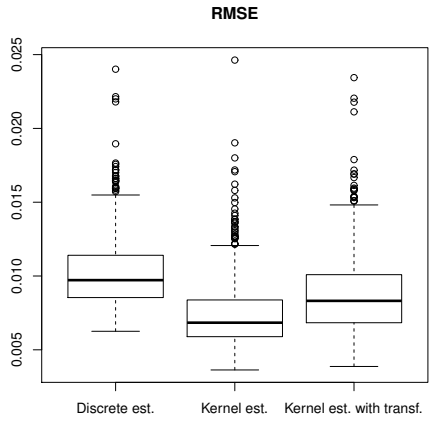
Let $\mathfrak{C}_n$ be one of three nonparametric copula estimators defined previously, and $\mathfrak{C}_{\hat{\theta}}$ its parametric estimator under H_0. Here $\hat{\theta}$ is a $\sqrt{n}$-consistent estimator of θ. In our study, we obtained $\hat{\theta}$ using the maximum likelihood method studied by [Shih and Louis, 1995] (the alternative method which can be used is based on a relationship between Kendall's τ coefficient and θ (see [Luciano et al., 2008])). Consider a Cramér-Von-Mises type of distance between the estimators $\mathfrak{C}_n$ and $\mathfrak{C}_{\hat{\theta}}$, defined by

$$d_n = n \int_0^1 \left(\mathfrak{C}_n(u,v) - \mathfrak{C}_{\hat{\theta}}(u,v)\right)^2 d\mathfrak{C}_n(u,v). \tag{3.21}$$

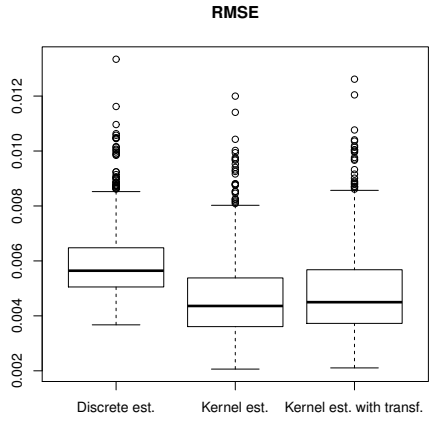
Other kind of distances between $\mathfrak{C}_n$ and $\mathfrak{C}_{\hat{\theta}}$ may be used instead. It follows from Theorem 3 that d_n admits a weak limit under H_0, while, under H_1, d_n tends to infinity with probability tending to one. Thus the critical region of the test is of the



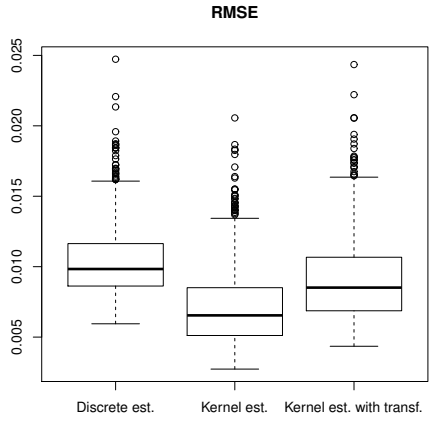
(a) Frank copula, 25% of censoring, $\tau = 0.75$



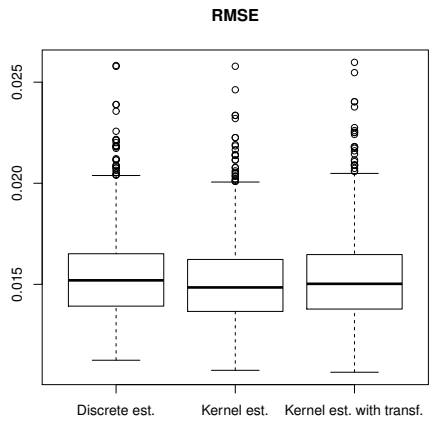
(b) Frank copula, 50% of censoring, $\tau = 0.75$



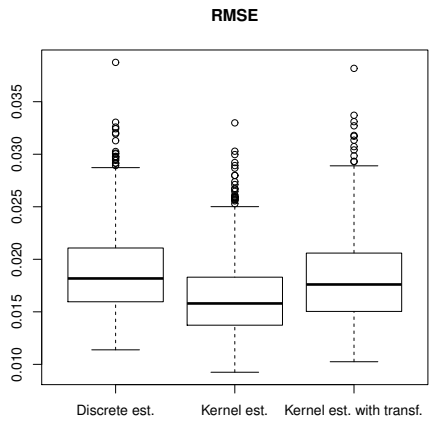
(c) Gumbel copula, 25% of censoring, $\tau = 0.75$



(d) Gumbel copula, 50% of censoring, $\tau = 0.75$



(e) Clayton copula, 25% of censoring, $\tau = 0.75$



(f) Clayton copula, 50% of censoring, $\tau = 0.75$

FIGURE 3.3 – Model 2 : two variables are censored

form $\mathcal{R} = \{d_n > d\}$. The limit law of d_n can be computed using the exact form of $\mathbb{Z}_C$ and of a limit distribution for $\hat{\theta}$.

Alternatively, one can compute critical values and $p-$values via bootstrap. This is the path that we used in our examples, relying on the bootstrap procedure described below. To this aim, suppose that $\mathbb{G}_n$ is some estimator of the joint distribution function of the censoring variables, defining a true distribution function. In all the examples presented in this chapter, such an estimator is available (up to some normalization of the Kaplan-Meier estimator of the censoring distribution, or to some affectation of the residual mass to infinity).

Bootstrap procedure.

For $b = 1, ..., B$, where B denotes the number of bootstrap replications,

1. simulate $(T_{1i}^b, T_{2i}^b)_{1\leq i\leq n}$ according to the distribution defined by $\mathfrak{C}_{\hat{\theta}}$ and the marginal distributions $\mathbb{F}_{1n}$ and $\mathbb{F}_{2n}$;
2. simulate $(C_{1i}^b, C_{2i}^b)_{1\leq i\leq n}$ according to $\mathbb{G}_n$;
3. compute the $b-$th bootstrap sample $(Y_{1i}^b, Y_{2i}^b, \delta_{1i}^b, \delta_{2i}^b)_{1\leq i\leq n}$ based on the simulated variables ;
4. using the $b-$th bootstrap sample, compute estimators $\hat{\theta}^b$ and $\mathfrak{C}_n^b$ and the corresponding distance d_n^b;

Then, use the vector $(d_n^b)_{1\leq b\leq n}$ to estimate the $p-$value.

3.5.2.2 Loss-ALAE data.

This bivariate dataset was provided by the US Insurance Services Office and studied previously by [Denuit and Van Keilegom, 2006] and [Frees and Valdez, 1998]. It contains 1500 observations, composed of losses (indemnities to be paid by insurance company) and of allocated loss adjustment expenses (ALAE's), associated with each loss. ALAE's are additional costs which are be related to lawyer's fees or claim investigation expenses. Each contract has a specific policy limit C_1 (a maximal claim amount). Denote the loss variable by T_1 and the ALAE variable by T_2. If the amount T_{1i} of the i-th claim exceeds the corresponding limit C_{1i}, only C_{1i} is registered by the insurance company and the loss variable is censored. As the ALAE variable is always observed, we are in the case described in Example 1, with the bivariate estimator of the joint distribution function given by (3.4).

As the expensive claims are usually associated with greater settlement costs, large values of the loss variable are expected to be associated with large values of ALAE's. In the reinsurance practice it is important to model correctly this association. We refer to [Denuit and Van Keilegom, 2006] for more details. This paper mention that, although the data contains only 34 censored observations, they have a much higher mean than the complete data (217.941\$ versus 37.110\$), so that the estimation can be biased if censored observations are not taken into account.

In order to identify which parametric copula family is more adapted to modeling the dependance structure of the data, we performed a goodness-of-fit test (using the estimator (3.7)) for four families of archimedean copulas : Frank, Gumbel, Clayton, Joe (see Table 3.1 for the corresponding copula expressions). The results are given in Table 3.3. As expected from the structure of the data, only the extreme value

Model	$\hat{\theta}$	Test statistic	95% quantile	97.5 % quantile	99 % quantile	p-value
Frank	3.30	$9,64 \times 10^{-5}$	1.10×10^{-5}	1.43×10^{-5}	1.88×10^{-5}	< 0.001
Gumbel	1.50	2.13×10^{-5}	4.99×10^{-5}	5.61×10^{-5}	6.11×10^{-5}	0.851
Clayton	1.00	37.8×10^{-5}	1.97×10^{-5}	2.20×10^{-5}	2.58×10^{-5}	< 0.001
Joe	1.90	6.71×10^{-5}	5.92×10^{-5}	6.49×10^{-5}	7.35×10^{-5}	0.019

TABLE 3.3 – Loss-ALAE data : goodness-of-fit for the considered copula models.

copula models give non zero p-values. Gumbel's copula outperforms the three other models. This result is in concordance with [Denuit and Van Keilegom, 2006] and [Frees and Valdez, 1998] who noticed that Gumbel's copula furnishes the best fit. Compared to their procedure, our method however is able to reject clearly Frank's and Clayton's models.

3.5.2.3 Canadian insurer's data.

This dataset belongs to a large Canadian insurer and contains joint lifetimes (T_1, T_2) of members of the couples who subscribed an insurance contract [1]. Besides the lifetimes, an additional variable ε is observed, which is the age difference between two individuals of the same couple. 11947 couples were observed between December, 29, 1988, and December 31, 1993. In our study, we eliminated the same-sex contracts and we kept only one contract for couples with more than one policy. The remaining sample concerns 11454 observations. As most couples were still alive at the end of the observation period, the dataset contains a huge proportion of censored observations (98,2% with at least one censored lifetime), the censoring variables (C_1, C_2) being the ages of individuals at the exit from the study. In the present approach, we neglected the left-truncation phenomenon as in [Gribkova et al., 2013]. For more details on this dataset we refer to [Carriere, 2000], [Frees et al., 1996], [Luciano et al., 2008] and [Youn and Shemyakin, 1999].

As both members of a couple are removed from the study at the same moment, censoring variables are related by $C_2 = C_1 + \varepsilon$, which corresponds to the situation described in our Example 3, where the bivariate distribution function estimator is of the form (3.2) with the weights given by (3.5).

We recall that, if two variables T_1 and T_2 are coupled by $C(u, v)$, their joint survival function $S(t_1, t_2) = P(T_1 > t_1, T_2 > t_2)$ can be written as $S(t_1, t_2) = \tilde{\mathcal{C}}(S_1(t_1), S_2(t_2))$, where S_1, S_2 are marginal survival functions and $\tilde{\mathcal{C}}$ is a copula function with

$$\tilde{\mathcal{C}}(u, v) = C(1 - u, 1 - v) + u + v - 1.$$

For easier comparison with former studies of this dataset, we present the results of the estimation of $\tilde{C}$ rather than C itself.

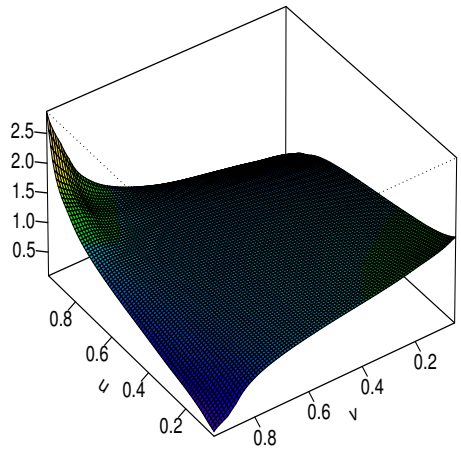
1. Copula density estimation. Denote by $\tilde{c}$ the copula density associated with $\tilde{C}$. We use $\hat{c}^1(1 - u, 1 - v)$ and $\hat{c}^2(1 - u, 1 - v)$ according to the definition (3.14)

1. The authors wish to thank the Society of Actuaries, through the courtesy of Edward J. Frees and Emiliano Valdez, for allowing use of the data in this research

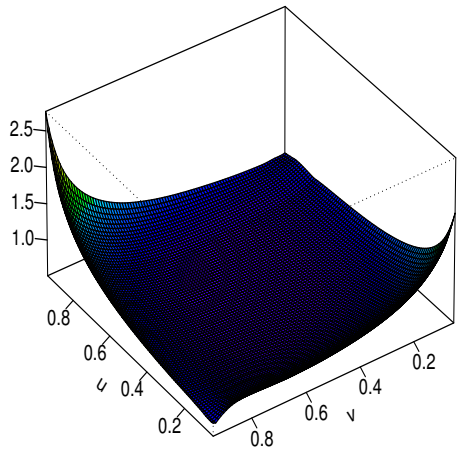
to estimate $\tilde{c}$. To select the bandwidth, we propose the following heuristic criterion based on a reference copula $\mathfrak{C}_{ref}$. We select a bandwidth in a set $\mathcal{H}$ such that

$$\hat{h}^j = \arg\min_{h\in\mathcal{H}} \int (\mathfrak{C}_n^j(u,v) - \mathfrak{C}_{ref}(u,v))^2 dudv, \quad j = 1, 2.$$

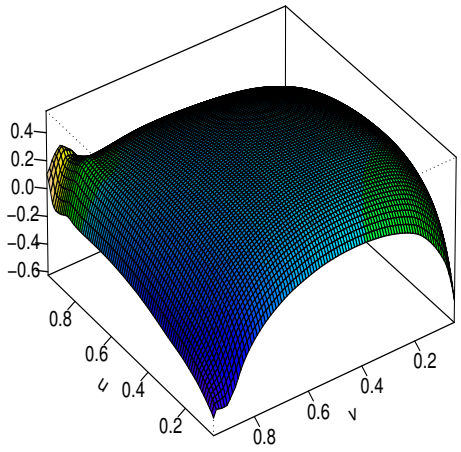
As a reference copula, we considered Frank's copula with parameter value specified in Table 3.4, due to the shape of the estimated density (which had similarities with Frank's copula). The two estimations of the survival copula density $\tilde{c}$ that we obtained are represented by Figure 3.4 (a) and (b). The difference between these two estimators is represented in Figure 3.4 (c). This difference is pronounced in the corners of the unit square. This is not surprising due to the fact that the estimator $\hat{\mathfrak{C}}_n^2$ is designed to improve estimation on the border of $[0,1]^2$.



(a) Estimator $\hat{c}^1(1-u, 1-v)$, $h = 0.14$



(b) Estimator $\hat{c}^2(1-u, 1-v)$, $h = 2$



(c) Difference between two density estimators

Figure 3.4 – Nonparametric copula density estimation. Data from a Canadian insurer.

2. Goodness-of-fit for semiparametric copula models. We now apply the methodology proposed at the end of the section 3.4.2 to perform a goodness-of-fit test. Three Archimedean copula families are considered : Clayton, Frank and

Nelsen 4.2.20 (see Table 3.1 for the corresponding expressions). The test statistics is given by (3.21), where we used the discrete estimator (3.7) for $\mathfrak{C}_n$. The estimated copula parameters and p-values of the test are presented in Table 3.4. They were calculated via the procedures described in [Gribkova et al., 2013]. It is not surprising

Model	$\hat{\theta}$	Test statistic	95% quantile	97.5 % quantile	99 % quantile	p-value
Clayton	4.89	0.00076	0.00208	0.00251	0.00310	0.391
Frank	11.41	0.00036	0.00092	0.00115	0.00143	0.416
Nelsen	1.33	0.00116	0.00137	0.00160	0.00230	0.103

Table 3.4 – Canadian data set : goodness-of-fit for the considered survival copula models.

to see that, at level 5%, none of the considered models is rejected. Indeed, the huge percentage of censoring observations induces a considerable loss of power of any goodness-of-fit procedure, even if total number of observations is high. Nevertheless, the obtained p-values can be used to give insights about the choice of the appropriate model. Let us remark also that the idea of Frank's copula as a good candidate for describing dependence structure of the present data is confirmed by the visual form of the plots of our nonparametric copula density estimators.

3.6 Appendix

3.6.1 Proof of Lemma 2

Let m be the total number of couples composed of doubly uncensored observations and let $(Y_{[j,1]}, \ldots, Y_{[j,m]})$ for $j = 1, 2$ be the order statistics, corresponding to these observations. Denote as $W_{[j,i]}$ the weight associated with $Y_{[j,i]}$. We recall that the weights attributed to the censored observations are equal to 0. Under Assumption 5, both $\mathfrak{C}_n$ and $\mathfrak{C}_n^*$ are defined on the same set of greed points, i.e.

$$(u_i, v_j) = \left(\frac{1}{n} \sum_{k=1}^{m} W_{[1,k]} \mathbb{1}_{k \leq i}, \frac{1}{n} \sum_{l=1}^{m} W_{[2,l]} \mathbb{1}_{l \leq j} \right), \quad 1 \leq i, j \leq m.$$

Then, similarily to [Fermanian et al., 2004], with probability one,

$$\begin{aligned}
\mathfrak{C}_n(u_i, v_j) &= \mathbb{F}_n(Y_{[1,i]}, Y_{[2,j]}) \\
&= \mathbb{F}_n(F_1^{-1}F_1(Y_{[1,i]})), F_2^{-1}F_2(Y_{[2,j]})) \\
&= \mathbb{F}_n^*(F_1(Y_{[1,i]}), F_2(Y_{[2,j]})) \\
&= \mathbb{F}_n^*((\mathbb{F}_{1n}^*)^{-1}\mathbb{F}_{1n}^*(F_1(Y_{[1,i]})), (\mathbb{F}_{2n}^*)^{-1}\mathbb{F}_{2n}^*(F_2(Y_{[2,j]}))) \\
&= \mathfrak{C}_n^*(\mathbb{F}_{1n}^*(F_1(Y_{[1,i]})), \mathbb{F}_{2n}^*(F_2(Y_{[2,j]}))) \\
&= \mathfrak{C}_n^*(u_i, v_j).
\end{aligned}$$

3.6.2 Proof of Theorem 4

The proof is decomposed in two parts. First, we consider $t_1 < \tau_1$ and $t_2 < \tau_2$ and study the convergence for $(y_1, y_2) \in \mathcal{Y} = [-\infty, t_1] \times [-\infty, t_2]$. In this case, since the kernel function k has compact support, one can consider that W_{in} is bounded by some finite function (for n large enough, the observations with values of $Y_{1i} \geq t_1 + \varepsilon$ or $Y_{2i} \geq t_2 + \varepsilon$ for any $\varepsilon > 0$ give a contribution zero to the sum). Next, we rely on a tightness argument to make $t_1 \to \tau_1$ and $t_2 \to \tau_2$.

1. Convergence for $(\mathbf{y_1}, \mathbf{y_2}) \in \mathcal{Y}$. As stated before, since k has compact support, for n large enough, only terms with $(Y_{1i}, Y_{2i}) \in \mathcal{C}_\varepsilon$ contribute to $\hat{\mathbb{F}}_n^1(y_1, y_2)$ for $y_1 < t_1$ and $y_2 < t_2$. Therefore, we will assume, in this first part of the proof, that $(Y_{1i}, Y_{2i}) \in \mathcal{Y}_\varepsilon = [-\infty, t_1 + \varepsilon] \times [-\infty, t_2 + \varepsilon]$ for all i.

Recall the notation $K_h(y) = \int_{-\infty}^{yh^{-1}} k(u)du$. The class of functions

$$\mathcal{F}_1 = \{(y_1, y_2) \to \phi_{h,y_1,y_2}(t_1, t_2) = K_h(y_1 - t_1) K_h(y_2 - t_2), h \in [0, 1/4]\},$$

is Donsker from Lemma A.1 in [Omelka et al., 2009]. Therefore, it follows from (3.15) in Assumption 6 that

$$\begin{aligned}\hat{\mathbb{F}}_n^1(y_1, y_2) - \mathbb{F}_n(y_1, y_2) &= \frac{1}{n}\sum_{i=1}^n W_i \left[\phi_{h,y_1,y_2}(Y_{i1}, Y_{i2}) - \mathbb{1}_{Y_{i1} \leq y_1, Y_{i2} \leq y_2}\right] \\ &\quad + \frac{1}{n}\sum_{i=1}^n \eta^{\psi_{h,y_1,y_2}}(Y_{1i}, Y_{2i}, \delta_{1i}, \delta_{2i}) + R_n(y_1, y_2),\end{aligned}$$

where $\sup_{y_1,y_2} |R_n(y_1, y_2)| = o_P(n^{-1/2})$, and where we recall that

$$\psi_{h,y_1,y_2}(t_1, t_2) = \phi_{h,y_1,y_2}(t_1, t_2) - \mathbb{1}_{t_1 \leq y_1, t_2 \leq y_2}.$$

From the Donsker assumption on the functions $\eta^{\psi_{h,y_1,y_2}}$ (see Assumption 6), we get, from the asymptotic equicontinuity of Donsker classes and the fact that the L^2-norms of $\eta_i^{\psi_{h,y_1,y_2}}$ uniformly tend to zero,

$$\sup_{h,(y_1,y_2)\in\mathcal{Y}} \left|\frac{1}{n}\sum_{i=1}^n \eta^{\psi_{h,y_1,y_2}}(Y_{1i}, Y_{2i}, \delta_{1i}, \delta_{2i})\right| = o_P(n^{-1/2}),$$

since, for all h and all (y_1, y_2), $E[\eta^{\psi_{h,y_1,y_2}}(Y_{1i}, Y_{2i}, \delta_{1i}, \delta_{2i})] = 0$.

Next, let W be a random variable having the same distribution as the variables $(W_i)_{1 \leq i \leq n}$. Observe that, again, since $y_1 < t_1$ and $y_2 < t_2$, W can be considered as almost surely bounded, and $W \times \mathcal{F}_1$ is a Donsker classes of functions, from a permanence properties of Donsker classes, see [van der Vaart and Wellner, 1996] Example 2.10.10. Therefore,

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^n W_i \left[\phi_{h,y_1,y_2}(Y_{i1}, Y_{i2}) - \mathbb{1}_{Y_{i1} \leq y_1, Y_{i2} \leq y_2}\right] \\ &\qquad = \int\int [\phi_{h,y_1,y_2}(t_1, t_2) - \mathbb{1}_{t_1 \leq y_1, t_2 \leq y_2}] dF(t_1, t_2) + o_P(n^{-1/2}),\end{aligned}$$

where the o_P-rate holds uniformly in (y_1, y_2), and where we used that, for any function ϕ with finite first moment, $E[W\phi(Y_1, Y_2)] = E[\phi(T_1, T_2)]$. Using, again, a

second order Taylor expansions and the differentiability assumptions on F, we get $\sup_{y_1<t_1,y_2<t_2} |\hat{\mathbb{F}}^1_n(y_1,y_2) - \mathbb{F}_n(y_1,y_2)| = o_P(n^{-1/2})$.

2. Convergence in the right tail of the distribution. We apply Lemma 7.1 in [Lopez and Saint-Pierre, 2012] to the process

$$R_n(t_1,t_2) = n^{1/2}\left\{\hat{\mathbb{F}}^1_n(\tau_1,\tau_2) - \mathbb{F}_n(\tau_1,\tau_2) - \hat{\mathbb{F}}^1_n(t_1,t_2) + \mathbb{F}_n(t_1,t_2)\right\}.$$

This can be done by checking that $|R_n(t_1,t_2)| \leq M_n\Gamma_n(t_1,t_2)$, with $M_n = n^{1/2}A_n$ and $\Gamma_n(t_1,t_2) = \frac{1}{n}\sum_{i=1}^n Z_i \mathbb{I}_{Y_{1i}>t_1,Y_{2i}>t_2}$. Assumption 7 ensures that M_n and Γ_n satisfy the conditions 2 to 5 of Lemma 7.1 in [Lopez and Saint-Pierre, 2012].

3.6.3 Proof of Theorem 5 (case $i = 2$)

Let $L_{1i} = \Phi^{-1}(F_1(T_{1i}))$ and $L_{2i} = \Phi^{-1}(F_2(T_{2i}))$. These random variables have marginal distribution Φ, and have joint distribution function $F^\Phi(l_1,l_2) = \mathbb{P}(L_1 \leq l_1, L_2 \leq l_2) = C(\Phi(l_1),\Phi(l_2))$. Since the transformation $\Phi^{-1}(F_1(\cdot))$ is increasing, we can observe that the censoring model is equivalent (up to this transformation), to the model based on, for $j = 1, 2$,

$$\begin{aligned} M_{ji} &= \Phi^{-1}(F_j(Y_{ji})), \\ \delta^\Phi_{ji} &= \mathbb{1}_{L_{ji}\leq D_{ji}} = \delta_{ji}, \end{aligned}$$

where D_{ji} are the transformed censoring variables, that is $D_{ji} = \Phi^{-1}(F_j(C_{ji}))$. We will denote by $\mathbb{F}^\Phi_n$ the estimator of the joint distribution of (L_1, L_2) similar to $\mathbb{F}_n$ but based on the transformed variables. From Assumption 9, the weights $W^\Phi_{in} = \delta_{1i}\delta_{2i}\hat{g}^\Phi(M_{1i}, M_{2i})$ based on the transformed model are the same as the weights W_{in}. Moreover, define $g^\Phi(l_1,l_2) = g(\Phi^{-1}(F_1(l_1)), \Phi^{-1}(F_1(l_2)))$.

Next, define $\hat{M}_{ji} = \Phi^{-1}(\mathbb{F}_{jn}(Y_{ji}))$. Let $f_{nj}(m) = \Phi^{-1}(\mathbb{F}_{jn}(F_j^{-1}(\Phi(m))))$. We have $\hat{M}_{ji} = f_{nj}(M_{ji})$. We can decompose

$$\begin{aligned} \hat{\mathfrak{C}}^2_n(\Phi(l_1),\Phi(l_2)) &= \frac{1}{n}\sum_{i=1} W_{in}\Big[K_h(l_1 - \hat{M}_{1i})K_h(l_2 - \hat{M}_{2i}) \\ &\qquad - K_h(l_1 - M_{1i})K_h(l_2 - M_{2i})\Big] + \hat{\mathbb{F}}^{1,\Phi}_n(l_1,l_2), \quad (3.22) \end{aligned}$$

where $\hat{\mathbb{F}}^{1,\Phi}_n$ denotes the estimator $\hat{\mathbb{F}}^1_n$ based on the transformed variables. From Theorem 4, we see that this second term in (3.22) satisfies

$$\hat{\mathbb{F}}^{1,\Phi}_n(l_1,l_2) = \mathbb{F}^\Phi_n(l_1,l_2) + R^\Phi_n(l_1,l_2),$$

with $\sup_{l_1,l_2} |R^\Phi_n(l_1,l_2)| = o_P(n^{-1/2})$. Indeed, the limit F^Φ of $\hat{\mathbb{F}}^{1,\Phi}_n$ is assumed to satisfy the assumptions of Theorem 4 as a consequence of Assumption 8. Moreover, the relation $\mathbb{F}^\Phi_n(l_1,l_2) = \mathbb{F}^*_n(\Phi(l_1),\Phi(l_2))$ holds.

Hence the proof consists of showing the negligibility of the first term in (3.22). The path of the proof is similar to the one of Theorem 4 : we first consider $l_1 < t_1^\Phi$ and $l_2 < t_2^\Phi$ (with obvious extension of the notation (t_1,t_2)) which allows to consider bounded weights W_i. Next, we use a tightness argument to obtain convergence on the whole plane. This last part is exactly the same as in the proof of Theorem 4, we therefore only focus on the case $l_1 < t_1^\Phi$ and $l_2 < t_2^\Phi$.

Since $\Phi^{-1}(\mathbb{F}_{jn})$ are increasing, we can see that the functions in the bracket of the first term of (3.22) belong to a Donsker class of functions, from Lemma A.1 in [Omelka et al., 2009]. Therefore, this first term can be rewritten as

$$\begin{aligned}
&\int\int \left[K_h\left(l_1 - f_{1n}(l)\right) - K_h\left(l_1 - l\right)\right] K_h\left(l_2 - l'\right) \frac{\hat{g}^{\Phi}(l,l')dF^{\Phi}(l,l')}{g^{\Phi}(l,l')} \\
&+ \int\int \left[K_h\left(l_2 - f_{2n}(l')\right) - K_h\left(l_2 - l'\right)\right] K_h\left(l_1 - l\right) \frac{\hat{g}^{\Phi}(l,l')dF^{\Phi}(l,l')}{g^{\Phi}(l,l')} \\
&+ \int\int \left[K_h\left(l_1 - f_{1n}(l)\right) - K_h\left(l_1 - l\right)\right]\left[K_h\left(l_2 - f_{2n}(l')\right) - K_h\left(l_2 - l'\right)\right] \frac{\hat{g}^{\Phi}(l,l')dF^{\Phi}(l,l')}{g^{\Phi}(l,l')} \\
&= \mathcal{T}_1 + \mathcal{T}_2 + \mathcal{T}_3,
\end{aligned}$$

up to some remainder terms that are $o_P(n^{-1/2})$ uniformly in (l_1, l_2), and where we used the fact that $E[\delta_{i1}\delta_{i2}g(Y_{1i}, Y_{2i})\phi(Y_{i1}, Y_{i2})] = E[\phi(T_{i1}, T_{i2})]$. The last term $\mathcal{T}_3$ is a second order term. A Taylor expansion and the uniform convergence of $\mathbb{F}_{jn}$ for $j = 1, 2$ show that this term is $O_P(n^{-1})$ uniformly in (l_1, l_2). The first two terms $\mathcal{T}_1$ and $\mathcal{T}_2$ can be studied in a similar way due to their symmetric definition, hence we only focus on $\mathcal{T}_1$.

To study $\mathcal{T}_1$, we first replace $\hat{g}$ by g, observing that, from a first order Taylor expansion,

$$\begin{aligned}
&\left|\int\int \left[K_h\left(l_1 - f_{1n}(l)\right) - K_h\left(l_1 - l\right)\right] K_h\left(l_2 - l'\right) \frac{\{\hat{g}^{\Phi}(l,l') - g^{\Phi}(l,l')\}dF^{\Phi}(l,l')}{g^{\Phi}(l,l')}\right| \\
&\qquad \leq C_0 \sup_{i:(y_1,y_2)\in\mathcal{C}_\varepsilon} |\hat{g}(y_1,y_2) - g(y_1,y_2)| \sup_t |\mathbb{F}_{1n}(t) - F_1(t)| h^{-1},
\end{aligned}$$

for some absolute constant C_0. Indeed, since $dF^{\Phi}(l,l') = c(\Phi(l), \Phi(l'))\Phi'(l)\Phi'(l')dldl'$, we get $|f_{1n}(l) - l|dF^{\Phi}(l,l') \leq C_0' \sup_t |\mathbb{F}_{1n}(t) - F_1(t)|$. Since $nh^2 \to \infty$ and using Assumption 6, one obtains that

$$\begin{aligned}
\mathcal{T}_1 &= \int \left[K_h\left(l_1 - f_{1n}(l)\right) - K_h\left(l_1 - l\right)\right] \int K_h\left(l_2 - l'\right) dF^{\Phi}(l,l') + o_P(n^{-1/2}) \\
&= \int \left[K_h\left(l_1 - f_{1n}(l)\right) - K_h\left(l_1 - l\right)\right] \int k(w_2)\partial_1 F^{\Phi}(l, l_2 + w_2 h)dw_2 + o_P(n^{-1/2}),
\end{aligned}$$

where the o_P-rate does not depend on (l_1, l_2).

From a second order Taylor expansion of $\partial_1 F^{\Phi}(l, l_2 + w_2 h)$ and the boundedness of its derivatives (due to the presence of Φ), we get $\int K_h\left(l_2 - l'\right) dF^{\Phi}(l,l') = \partial_1 F^{\Phi}(l, l_2)dl + O(h^2)$, where $O(h^2)-$rate does not depend on l. A Taylor expansion of K_h leads to

$$\begin{aligned}
\mathcal{T}_1 &= \frac{1}{h}\int k\left(\frac{l_1 - l}{h}\right)\left[F_1(F_1^{-1}(\Phi(l))) - \mathbb{F}_{1n}(F_1^{-1}(\Phi(l)))\right]\partial_1 C(\Phi(l), \Phi(l_2))\Phi'(l)dl \\
&+ o_P(n^{-1/2}).
\end{aligned}$$

Performing a change of variables with $v = [l_1 - l]h^{-1}$, one gets,

$$\begin{aligned}
\mathcal{T}_1 = &\int k(v)\left[\mathbb{F}_{1n}(F_1^{-1}(\Phi(l_1 + vh))) - F_1(F_1^{-1}(\Phi(l_1 + vh)))\right]\partial_1 C(\Phi(l_1 + vh), \Phi(l_2))dv \\
&+ o_P(n^{-1/2}).
\end{aligned} \tag{3.23}$$

Using the differentiability of Φ, one can replace $F_1(F_1^{-1}(\Phi(l_1 + vh)))$ by $F_1(F_1^{-1}(\Phi(l_1))) + vh\Phi'(l_1) + O(h^2)$, uniformly in l_1.

Since $\hat{g}^\Phi$ tends uniformly to g^Φ we have,

$$\begin{aligned}\mathbb{F}_{1n}(F_1^{-1}(\Phi(l_1+vh))) - \mathbb{F}_{1n}(F_1^{-1}(\Phi(l_1))) &= \frac{1}{n}\sum_{i=1}^n \delta_{i1}\delta_{i2}\hat{g}^\Phi(M_{i1}, M_{i2})\,[\mathbb{1}_{M_{1i}\le l_1+vh} \\ &\quad -\mathbb{1}_{M_{1i}\le l_1}] \\ &= \frac{1}{n}\sum_{i=1}^n f_n(\delta_{i1}, \delta_{i2}, M_{i1}, M_{i2}),\end{aligned}$$

where f_n belongs to a Donsker class $\mathcal{F}'$ (since the class of indicators function is Donsker, and that multiplication by a bounded Donsker class does not modify the Donsker property, see [van der Vaart and Wellner, 1996]), with $\|f_n\|_2 \to 0$. We then can write

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n f_n(\delta_{i1}, \delta_{i2}, M_{i1}, M_{i2}) &= \int \hat{g}^\Phi(u_1,u_2)\left[\mathbb{1}_{u_1\le\phi(l_1+vh)} - \mathbb{1}_{u_1\le\phi(l_1)}\right]\frac{dC(u_1,u_2)}{g^\Phi(u_1,u_2)} \\ &\quad +o_P(n^{-1/2}),\end{aligned}$$

with the o_P-rate holding uniformly in l_1. In this equation, we can replace $\hat{g}^\Phi$ by g^Φ up to some residual term which is $O_P(n^{-1/2}h) = o_P(n^{-1/2})$. Indeed,

$$\begin{aligned}&\int\{\hat{g}^\Phi(u_1,u_2) - g^\Phi(u_1,u_2)\}\left[\mathbb{1}_{u_1\le\phi(l_1+vh)} - \mathbb{1}_{u_1\le\phi(l_1)}\right]\frac{dC(u_1,u_2)}{g^\Phi(u_1,u_2)} \\ &\qquad = \int_{u_2}\int_{u_1=l_1}^{u_1=l_1+vh}[\hat{g}^\Phi(u_1,u_2) - g^\Phi(u_1,u_2)]\frac{dC(u_1,u_2)}{g^\Phi(u_1,u_2)} \\ &\qquad \le C_0 h \sup_{u_1,u_2}|\hat{g}^\Phi(u_1,u_2) - g^\Phi(u_1,u_2)|.\end{aligned}$$

Next, it follows from a Taylor expansion that

$$\mathbb{F}_{1n}(F_1^{-1}(\phi(l_1+vh))) - \mathbb{F}_{1n}(F_1^{-1}(\phi(l_1))) = vh\phi'(l_1)C(\phi(l_1),1) + o_P(n^{-1/2}).$$

This, combined with (3.23), the differentiability of $\partial_1 C$, and the fact that $\int vk(v)dv = 0$, shows that

$$\mathcal{T}_1 = -[\mathbb{F}_{1n}(F_1^{-1}(\phi(l_1))) - F_1(F_1^{-1}(\phi(l_1)))]\partial_1 C(\phi(l_1),\phi(l_2)) + o_P(n^{-1/2})$$

Since $\mathbb{F}_{1n}(F_1^{-1}(\phi(l_1))) = \mathbb{F}_n^*(\phi(l_1),\infty)$, we finally get

$$\hat{\mathfrak{C}}_n^2(u_1,u_2) = \mathbb{F}_n^*(u_1,u_2) - \mathbb{F}_n^*(u_1,\infty)\partial_1 C(u_1,u_2) - \mathbb{F}_n^*(\infty,u_2)\partial_2 C(u_1,u_2) + o_P(n^{-1/2}),$$

and the convergence of $\mathbb{F}_n^*(u_1,u_2)$ leads to the appropriate asymptotic distribution.

3.6.4 Proof of Theorem 6

A. Convergence of $\hat{\mathbf{c}}_1$. Let f_{1n}, f_{2n} denote the derivatives of $\hat{\mathbb{F}}_{1n}, \hat{\mathbb{F}}_{2n}$. Decompose

$$\begin{aligned}\hat{c}_1(u,v) &= \frac{1}{nh^2}\sum_{i=1}^{n}\frac{W_i k\left(\frac{\hat{\mathbb{F}}_{1n}^{-1}(u)-Y_{1i}}{h}\right)k\left(\frac{\hat{\mathbb{F}}_{2n}^{-1}(v)-Y_{2i}}{h}\right)}{f_{1n}(\hat{\mathbb{F}}_{1n}^{-1}(u))f_{2n}(\hat{\mathbb{F}}_{2n}^{-1}(v))} \\ &\quad + \frac{1}{nh^2}\sum_{i=1}^{n}(W_{in}-W_i)\frac{k\left(\frac{\hat{\mathbb{F}}_{1n}^{-1}(u)-Y_{1i}}{h}\right)k\left(\frac{\hat{\mathbb{F}}_{2n}^{-1}(v)-Y_{2i}}{h}\right)}{f_{1n}(\hat{\mathbb{F}}_{1n}^{-1}(u))f_{2n}(\hat{\mathbb{F}}_{2n}^{-1}(v))} \\ &:= \frac{J_{1n}(u,v)}{f_{1n}(\hat{\mathbb{F}}_{1n}^{-1}(u))f_{2n}(\hat{\mathbb{F}}_{2n}^{-1}(v))} + \frac{J_{2n}(u,v)}{f_{1n}(\hat{\mathbb{F}}_{1n}^{-1}(u))f_{2n}(\hat{\mathbb{F}}_{2n}^{-1}(v))}. \qquad (3.24)\end{aligned}$$

Observe that, for $(u,v)\in\mathcal{C}$, $(\hat{\mathbb{F}}_{1n}^{-1}(u),\hat{\mathbb{F}}_{2n}^{-1}(v))\in\mathcal{Y}$ with probability tending to one, where $\mathcal{Y}=[0,t_1]\times[0,t_2]$ for some $t_1<\tau_1$ and $t_2<\tau_2$. Indeed, recall that $\hat{\mathbb{F}}_{1n}^{-1}(u)=\inf\{f:\hat{\mathbb{F}}_{1n}(t)\geq u\}$. There exists some $\eta>0$ such that $t_1=F_1^{-1}(u)+\eta$ is strictly less than τ_1. Moreover, $\hat{\mathbb{F}}_{1n}(t_1)=F_1(t_1)+\varepsilon_n(t_1)$, where $\varepsilon_n(t_1)$ tends to zero, and $F_1(t_1)>u$, which shows that $\hat{\mathbb{F}}_{1n}^{-1}(u)\leq t_1$ with probability tending to one. Moreover, since k has compact support, with probability tending to one, only the points i such that $(Y_{1i},Y_{2i})\in\mathcal{Y}_\varepsilon$ give a non-zero contribution to the sum, where the definition of $\mathcal{Y}_\varepsilon$ can be found in the proof of Theorem 4.

The proof is then composed of several steps. We show first that the second term in (3.24) negligible. Indeed, following Assumption 10 and using previous observation, we require to bound the difference of the weights $(W_{in}-W_i)$ for the indexes i corresponding to observations in $\mathcal{Y}_\varepsilon$. Then, using Assumption 11, we get

$$\sup_{(u,v)\in\mathcal{C}}|J_{2n}(u,v)|\leq O_P(\eta_n)\times\sup_{(y_1,y_2)\in\mathcal{Y}_\varepsilon}\left|\frac{1}{nh^2}\sum_{i=1}^{n}k\left(\frac{y_1-Y_{1i}}{h}\right)k\left(\frac{y_2-Y_{2i}}{h}\right)\right|,$$

where the supremum on the right-hand side is $O_P(1)$, by Theorem 4 in [Einmahl and Mason, 2005]. Notice that the denominators in (3.24) are bounded away from zero. To see that, it suffices to write, for $j=1,2$,

$$f_{jn}(\hat{\mathbb{F}}_{jn}^{-1}(u))=\sum_{i=1}^{n}W_i k\left(\frac{\hat{\mathbb{F}}_{jn}^{-1}(u)-Y_{1i}}{h}\right)+O_P(\eta_n). \qquad (3.25)$$

Thus, there exists $\varepsilon>0$, such that, with probability tending to one

$$\inf_{u\in\mathcal{C}}f_{jn}(\hat{\mathbb{F}}_{jn}^{-1}(u))=\inf_{u\in\mathcal{C}}f_j(\hat{\mathbb{F}}_{jn}^{-1}(u))+O_P(\eta_n)\geq\inf_{x\leq t_j+\varepsilon}f_j(x)+O_P(\eta_n)>\frac{c}{2},$$

and the second term in (3.24) is therefore negligible.

Let us study now the first term. It follows from Theorem 4 in [Einmahl and Mason, 2005] that

$$\begin{aligned}&\sup_{(y_1,y_2)\in\mathcal{Y}_\varepsilon}\left|n^{-1}h^{-2}\sum_{i=1}^{n}W_i k\left(\frac{y_1-Y_{1i}}{h}\right)k\left(\frac{y_2-Y_{2i}}{h}\right)\right. \\ &\left.-h^{-2}\int k\left(\frac{y_1-t_1}{h}\right)k\left(\frac{y_2-t_2}{h}\right)d\mathbb{P}_{(T_1,T_2)}(t_1,t_2)\right| = O_P\left(\frac{[\log n]^{1/2}}{h\sqrt{n}}\right),\end{aligned}$$

where we used (3.3). Hence,

$$J_{1n}(u,v) = h^{-2}\int k\left(\frac{\hat{\mathbb{F}}_{1n}^{-1}(u)-t_1}{h}\right)k\left(\frac{\hat{\mathbb{F}}_{2n}^{-1}(v)-t_2}{h}\right)d\mathbb{P}_{(T_1,T_2)}(t_1,t_2)+O_P\left(\frac{[\log n]^{1/2}}{h\sqrt{n}}\right).$$

Using a second order Taylor expansion, and the fact that the derivatives of the density f up to order 2 are uniformly bounded (Assumption 11), we get

$$\sup_{u,v}|J_{1n}(u,v) - f(\hat{\mathbb{F}}_{1n}^{-1}(u),\hat{\mathbb{F}}_{2n}^{-1}(v))| = O_P(\eta_n).$$

Again by [Einmahl and Mason, 2005], there exists ε such as, for $j = 1,2$, with probability tending to one,

$$\sup_{u:(u,v)\in\mathcal{C}}|f_{jn}(\hat{\mathbb{F}}_{jn}^{-1}(u)) - f_j(\hat{\mathbb{F}}_{jn}^{-1}(u))| \leq \sup_{x<t_j+\varepsilon}|f_{jn}(x)-f_j(x)| = O_P(\eta_n).$$

Therefore, we have

$$\sup_{(u,v)\in\mathcal{C}}\left|\hat{c}_1(u,v) - \frac{f(\hat{\mathbb{F}}_{1n}^{-1}(u),\hat{\mathbb{F}}_{2n}^{-1}(v))}{f_1(\hat{\mathbb{F}}_{1n}^{-1}(u))f_2(\hat{\mathbb{F}}_{2n}^{-1}(v))}\right| = O_P(\eta_n).$$

To conclude it remains to prove that

$$J_n := \sup_{(u,v)\in\mathcal{C}}\left|\frac{f(\hat{\mathbb{F}}_{1n}^{-1}(u),\hat{\mathbb{F}}_{2n}^{-1}(v))}{f_1(\hat{\mathbb{F}}_{1n}^{-1}(u))f_2(\hat{\mathbb{F}}_{2n}^{-1}(v))} - \frac{f(F_1^{-1}(u),F_2^{-1}(v))}{f_1(F_1^{-1}(u))f_2(F_2^{-1}(v))}\right| = O_P(\eta_n).$$

By the assumption of Theorem, the density of (T_1,T_2) is twice continuously differentiable with bounded second derivatives and the marginal densities are bounded away from zero, so by using Taylor expansion, there exists a constant L, such as

$$\begin{aligned} J_n &= \sup_{u,v\in\mathcal{C}}\left|c(\hat{\mathbb{F}}_{1n}^{-1}(u),\hat{\mathbb{F}}_{2n}^{-1}(v)) - c(F_1^{-1}(u),F_2^{-1}(v)\right| \\ &\leq L\left(\sup_u|\hat{\mathbb{F}}_{1n}^{-1}(u)-F_1^{-1}(u)| + \sup_v|\hat{\mathbb{F}}_{2n}^{-1}(v)-F_2^{-1}(v)|\right) \end{aligned}$$

We will show now that the bound from the previous inequality is of the order of $O_P(\eta_n)$. For an arbitrary constant M for the first term (the second term is analogous) we have

$$\begin{aligned} & P(\sup_u|\hat{\mathbb{F}}_{1n}^{-1}(u)-F_1^{-1}(u)| > M\eta_n) \\ \leq\ & P(\exists u: u > \mathbb{F}_{1n}(F_1^{-1}(u)+M\eta_n)) + P(\exists u: u < \mathbb{F}_{1n}(F_1^{-1}(u)-M\eta_n)) \\ \leq\ & P\left\{\sup_u|\mathbb{F}_{1n}(F_1^{-1}(u)+M\eta_n) - F_1(F_1^{-1}(u)+M\eta_n)| \right. \\ & \qquad\qquad \left. > \inf_u(F_1(F_1^{-1}(u)+M\eta_n)-u)\right\} \\ & +P\left\{\sup_u|\mathbb{F}_{1n}(F_1^{-1}(u)-M\eta_n) - F_1(F_1^{-1}(u)-M\eta_n)| \right. \\ & \qquad\qquad \left. > \inf_u(u - F_1(F_1^{-1}(u)-M\eta_n))\right\}. \end{aligned}$$

By Corollary 3, $\sup_u |F_{1n}(u) - F_1(u)| = O_P(h^2)$, and, choosing M sufficiently large, we obtain $P(\sup_u |\hat{\mathbb{F}}_{1n}^{-1}(u) - F_1^{-1}(u)| > M\eta_n) \to 0$.

B. Convergence of $\hat{\mathbf{c}}_2$. We have

$$\hat{c}_2(u,v) = \frac{\sum_{i=1}^n W_{in} k\left(\frac{\Phi^{-1}(u)-\Phi^{-1}[\mathbb{F}_{1n}(Y_{1i})]}{h}\right) k\left(\frac{\Phi^{-1}(v)-\Phi^{-1}[\mathbb{F}_{2n}(Y_{2i})]}{h}\right)}{nh^2\Phi'(\Phi^{-1}(u))\Phi'(\Phi^{-1}(v))}.$$

Let us use a notation $y_{1i}^n(u) := \Phi^{-1}(u) - \Phi^{-1}[\mathbb{F}_{1n}(Y_{1i})]$ and $y_{1i}(u) := \Phi^{-1}(u) - \Phi^{-1}[F_1(Y_{1i})]$, with a similar definition for y_{2i}^n and y_{2i}. Using a 4-th order Taylor expansion, we get

$$\begin{aligned}
\hat{c}_2(u,v) &= \frac{1}{nh^2\Phi'(\Phi^{-1}(u))\Phi'(\Phi^{-1}(v))} \sum_{i=1}^n W_{in} k\left(\frac{y_{1i}^n(u)}{h}\right) k\left(\frac{y_{2i}^n(v)}{h}\right) \\
&= \frac{1}{\Phi'(\Phi^{-1}(u))\Phi'(\Phi^{-1}(v))} \left\{ \frac{1}{nh^2} \sum_{i=1}^n W_{in} k\left(\frac{y_{1i}(u)}{h}\right) k\left(\frac{y_{2i}(v)}{h}\right) + I_n(u,v) \right\},
\end{aligned}$$

where

$$\begin{aligned}
I_n(u,v) &= \frac{1}{nh^2} \sum_{i=1}^n W_{in} k\left(\frac{y_{1i}^n(u)}{h}\right) k\left(\frac{y_{2i}^n(v)}{h}\right) - \frac{1}{nh^2} \sum_{i=1}^n W_{in} k\left(\frac{y_{1i}(u)}{h}\right) k\left(\frac{y_{2i}(v)}{h}\right) \\
&= \sum_{m=1}^3 \sum_{l=0}^m \sum_{i=1}^n \frac{W_{in}}{nh^2} k^{(l)}\left(\frac{y_{1i}(u)}{h}\right) k^{(m-l)}\left(\frac{y_{2i}(v)}{h}\right) \frac{(y_{1i}^n(u) - y_{1i}(u))^l (y_{2i}^n(v) - y_{2i}(v))^{m-l}}{m!\, h^m} \\
&+ \frac{1}{nh^2} \sum_{l=0}^4 \sum_{i=1}^n W_{in} k^{(l)}\left(\frac{\tilde{y}_{1i}(u)}{h}\right) k^{(4-l)}\left(\frac{\tilde{y}_{2i}(v)}{h}\right) \frac{(y_{1i}^n(u) - y_{1i}(u))^l (y_{2i}^n(v) - y_{2i}(v))^{4-l}}{4!\, h^4} \\
&:= \sum_{m=1}^3 \sum_{l=0}^m I_{ml}^n(u,v) + \sum_{l=0}^4 I_{4l}^n(u,v),
\end{aligned}$$

using the notation $k^{(l)}$ for the $l-$th derivative of k, and $\tilde{y}_{1i}(u)$ (resp. $\tilde{y}_{2i}(v)$) for some point between $y_{1i}(u)$ and $y_{1i}^n(u)$ (resp. $y_{2i}(v)$ and $y_{2i}^n(v)$).

We get

$$\begin{aligned}
\sup_{(u,v)\in\mathcal{C}} |\hat{c}_2(u,v) - c(u,v)| &\leq \sup_{u,v\in\mathcal{C}} \left| \frac{\sum_{i=1}^n W_{in} k\left(\frac{y_{1i}(u)}{h}\right) k\left(\frac{y_{2i}(v)}{h}\right)}{nh^2\Phi'(\Phi^{-1}(u))\Phi'(\Phi^{-1}(v))} - c(u,v) \right| \\
&\quad + \sup_{(u,v)\in\mathcal{C}} |I_n(u,v)|. \qquad (3.26)
\end{aligned}$$

The proof consists of showing separately the convergence rate of each term in this decomposition.

1. Convergence of the first term in (3.26). We will show that this term converges at the rate $[\log n]^{1/2} n^{-1/2} h^{-1} + h^2$. Let us introduce the sets $\mathcal{U} = \{u : \exists v \ s.t. \ (u,v) \in \mathcal{C}\}$, and $\mathcal{V} = \{v : \exists u \ s.t. \ (u,v) \in \mathcal{C}\}$.

The first term in (3.26) can be written as

$$\sup_{(u,v)\in\mathcal{C}} \left| \frac{\left\{ \frac{1}{nh^2}\sum_{i=1}^n W_{in} k\left(\frac{y_{1i}(u)}{h}\right) k\left(\frac{y_{2i}(v)}{h}\right) - c(u,v)\Phi'(\Phi^{-1}(u))\Phi'(\Phi^{-1}(v))\right\}}{\Phi'(\Phi^{-1}(u))\Phi'(\Phi^{-1}(v))} \right|$$
$$\leq \sup_{x\in\Phi^{-1}(\mathcal{U}),y\in\Phi^{-1}(\mathcal{V})} \left| \left\{ \frac{1}{nh^2}\sum_{i=1}^n W_{in} k\left(\frac{x-\Phi^{-1}[F_1(Y_{1i})]}{h}\right) k\left(\frac{y-\Phi^{-1}[F_2(Y_{2i})]}{h}\right) \right.\right.$$
$$\left.\left. - \quad c(\Phi(x),\Phi(y))\Phi'(x)\Phi'(y)\right\} [\Phi'(x)\Phi'(y)]^{-1} \right|$$

Note that $c(\Phi(x),\Phi(y))\Phi'(x)\Phi'(y)$ is the density of the distribution function of random variables $(\Phi^{-1}(F_1(Y_1)),\Phi^{-1}(F_2(Y_2)))$, evaluated at the point (x,y). As stated at the beginning of Section 3.6.2, one can use the compactness of the support of k to deduce that only the points i corresponding to observations in $\mathcal{Y}_\varepsilon$ contribute to the sum. Therefore, one can replace (up to some negligible term) W_{in} by W_i using Assumption 10. It then follows from [Einmahl and Mason, 2005] that the resulting quantity converges towards 0 at rate $[\log n]^{1/2} n^{-1/2} h^{-1} + h^2$, the rate h^2 coming from the convergence rate of the expectation (deduced from classical arguments on kernel estimators, and the regularity of $c(\Phi(x),\Phi(y))\Phi'(x)\Phi'(y)$).

2. Convergence of $\mathbf{I_n(u,v)}$. Let us now consider the second term of (3.26),

$$\sup_{(u,v)\in\mathcal{C}} |I_n(u,v)| \leq \sum_{m=1}^{3}\sum_{l=0}^{m} \sup_{(u,v)\in\mathcal{C}} |I_{ml}^n(u,v)| + \sum_{l=0}^{4} \sup_{(u,v)\in\mathcal{C}} |I_{4l}^n(u,v)|.$$

First observe that

$$\sup_{u\in\mathcal{U}} |y_{1i}(u) - y_{1i}^n(u)| \leq \sup_{x<t_1+\varepsilon} |\Phi^{-1}(F_1(x)) - \Phi^{-1}(F_{1n}(x))|,$$

for some t_1 such that $t_1+\varepsilon<\tau_1$, since $1\notin\mathcal{U}$. Since Φ^{-1} has a continuous bounded derivative on $\mathcal{T}_1$ from Assumption 8, $\sup_{u\in\mathcal{U}} |y_{1i}(u) - y_{1i}^n(u)|$ has the same convergence rate as $\sup_{t<t_j+\varepsilon} |F_j(t) - F_{jn}(t)| = O_P(n^{-1/2})$. The same is true considering $\sup_{v\in\mathcal{V}} |y_{2i}(v) - y_{2i}^n(v)|$. Therefore, one gets

$$\begin{aligned}\sup_{(u,v)\in\mathcal{C}} |I_{4l}^n(u,v)| &\leq \frac{1}{h^6} \sup_{(u,v)\in\mathcal{C}} |(y_{1i}^n(u) - y_{1i}(u))^l (y_{2i}^n(v) - y_{2i}(v))^{4-l}| \times O_{\mathbb{P}}(1) \\ &= \frac{1}{n^2h^6} \times O_{\mathbb{P}}(1).\end{aligned}$$

For the terms up to the third order in the Taylor expansion, one obtains, for $m = 1,...,3$,

$$\begin{aligned}\sup_{(u,v)\in\mathcal{C}} |I_{ml}^n(u,v)| &\leq \frac{1}{m!h^m} \sup_{(u,v)\in\mathcal{C}} |(y_{1i}^n(u) - y_{1i}(u))^l (y_{2i}^n(v) - y_{2i}(v))^{m-l}| \\ &\quad \times \sup_{(u,v)\in\mathcal{C}} \left[\frac{1}{nh^2}\sum_{i=1}^n W_{in} k^{(l)}\left(\frac{y_{1i}(u)}{h}\right) k^{(m-l)}\left(\frac{y_{2i}(v)}{h}\right)\right] \\ &= \frac{1}{n^{m/2}h^m} \times O_{\mathbb{P}}(1).\end{aligned}$$

The condition on h implies that $\sup_{(u,v)\in\mathcal{C}} |I_n(u,v)| = O_{\mathbb{P}}(h^{-1}n^{-1/2})$.

3.6.5 Properties of the functions η^ψ is the Examples

The aim of this section is to show that part 3 of Assumption 6 holds for the functions η^ψ corresponding to the three standard examples we consider. We focus on Example 1, where η^ψ has the simplest form, which can be found in [Stute, 1996]. We then explain how these arguments may be extended to Examples 2 and 3.

In the case of Example 1, we have

$$\begin{aligned}\eta^\psi(Y_1, Y_2, \delta_1, \delta_2) &= \frac{(1-\delta_1)\int_{Y_1}^{\tau_1}\int_{-\infty}^{\infty}\psi(y_1,y_2)dF(y_1,y_2)}{1-H(Y_1)} \\ &\quad - \int \frac{\mathbf{1}_{Y_1\geq y_1}[1-F(y_1)]\{\int_{y_1}^{\infty}\int_{-\infty}^{\infty}\psi(t_1,t_2)dF(t_1,t_2)\}dG(y_1,y_2)}{[1-H(y_1)]^2[1-G(y_1)]},\end{aligned}$$

where $H(y_1) = \mathbb{P}(Y_1 \geq y_1)$. From this expression, one can see that, in the case where ψ is a nonnegative function, we can write

$$\eta^\psi(Y_1, Y_2, \delta_1, \delta_2) = \frac{(1-\delta_1)f_1(Y_1)}{1-H(Y_1)} + f_2(Y_1),$$

where f_1 and f_2 are monotone functions. Moreover, if ψ satisfies the requirements of part 2 of Assumption 6, f_1 and f_2 are bounded, and $f_1(Y_1) = 0$ for $Y_1 > t_1$. Then, one can see that η^ψ belongs to the class $\mathcal{F}_2 = (1-\delta_1)[1-H(Y_1)]^{-1}\mathbb{1}_{Y_1\leq t_1}\times\mathcal{F}_1+\mathcal{F}_1$, where $\mathcal{F}_1$ is the class of positive functions bounded by some absolute constant (this class is Donsker from Theorem 2.7.5 in [van der Vaart and Wellner, 1996]). Therefore, from Examples 2.10.7 and 2.10.10 in [van der Vaart and Wellner, 1996], $\mathcal{F}_2$ class is Donsker. The functions ψ_{h,y_1,y_2} are not non-negative, but they are the sum of two non-negative functions. Moreover, $\psi \to \eta^\psi$ is linear, therefore the family of functions $\eta^{\psi_{h,y_1,y_2}}$ is Donsker from Example 2.10.7 in [van der Vaart and Wellner, 1996]. For Examples 2 and 3, the arguments are similar, since η^ψ can always be decomposed into a sum of bounded monotonic terms.

To show that the expectation of $(\eta^{\psi_{h,y_1,y_2}})^2$ tends to zero, observe that, for $j = 1, 2$,

$$|f_j(Y_1)| \leq C \sup_{(y,z)\in\mathcal{Y}} \left|\int_y^{\tau_1}\int_{-\infty}^{\infty}\psi(t_1,t_2)dF(t_1,t_2)\right|,$$

for some constant C. Similar bounds can be found in Examples 2 and 3. Next, observe that, from Fubini's Theorem,

$$\begin{aligned}&\int_y^{\tau_1}\int_{-\infty}^{\infty}\psi_{h,y_1,y_2}(t_1,t_2)dF(t_1,t_2) = \\ &\int_{-y}^{\infty}\int_{-\infty}^{\infty}[F(y_1-hw_1, y_2-hw_2) - F(y_1,y_2)]k(w_1)k(w_2)dw_1dw_2.\end{aligned}$$

A Taylor expansion then shows that, for $(y_1, y_2) \in \mathcal{Y}$, $\|\eta^{\psi_{h,y_1,y_2}}\|_\infty \leq \tilde{C}h$, which tends to zero.

Chapitre 4

Quantization and clustering in presence of censoring

Abstract. In this chapter, we study the fixed-rate quantization problem for random vectors with a censored component and its application to clustering for censored data. We define the empirical distortion and an empirically optimal quantizer in our framework. We show that the distortion of an empirically optimal k-quantizer converges almost surely to the minimal distortion over the class of all k-quantizers. A rate of convergence is given by a non asymptotic exponential bound.

We consider then an application of the defined procedure to k-clustering. We provide a two-step algorithm for clustering of observations with one censored component. The first step deals with the numerical approximation of the centers of clusters minimizing the empirical distortion under censoring. The second step deals with assigning labels to censored observations.

4.1 Introduction

Analyzing a lifetime variable for a population of subjects is the important task of many medical and actuarial studies. In real life, any population is heterogeneous and two of its individuals with different characteristics will not have the same expected lifetime. In order to take this difference between the subjects into account, it is common to characterize them by a vector of covariates.

Despite that individuals are in general heterogeneous, they can form several homogeneous groups (clusters) inside the same population. Detecting these groups may have important concerns in applications as the statistical modeling of the lifetime with its vector of covariates is more precise when applied separately to each of the existing clusters. For example, a population of policyholders in insurance is generally composed of several risk classes according to social, professional or geographic characteristics. Detecting these classes allow the insurance company to manage more efficiently the heterogeneity of its portfolio and to optimize the provisions. Another example concerns medical applications, where modeling survival times separately for each homogeneous groups of patients may help to diversify treatment alternatives.

The problem of finding groups in a way that subjects in the same group are in a certain sense more similar to each other than those in other groups is known as

clustering. The existing literature propose various clustering algorithms. However, most of them fail to work for the described applications, for the reason that the lifetime variable is frequently subjected to censoring. Indeed, for numeric data, the similarities between observations are measured by some distance. In presence of censoring, instead of observing a lifetime T, one observes $\min(T, C)$ and $\delta = \mathbb{1}_{T \leq C}$, where C is a random variable called censoring. Therefore, the precise positions of censored observations in the space are unknown, neither the distances between them.

The purpose of this chapter is to investigate the fixed-rate quantization problem for random vectors subjected to censoring acting in one dimension and to provide, as a natural application, a new algorithm allowing for k-clustering of multivariate data with one censored component. By the following we present some basic material on the fixed-rate quantization in the standard setting (without censoring). For a fuller description of this theory, we refer the reader to [Gersho and Gray, 1991], [Graf and Luschgy, 2000] and [Linder, 2002].

Let us first set up some notations. We consider a random vector (T, X) taking values in $\mathbb{R}^{d+1}$, where T stands for lifetime and X has a meaning of a d-dimensional random vector of covariates. The aim of k-quantization consists in summarizing the distribution P of the random vector (T, X) by k points in $\mathbb{R}^{d+1}$. It can be achieved by means of the k-point quantizers. A k-point quantizer is an application

$$q : \mathbb{R}^{d+1} \to \mathcal{C},$$

where

$$\mathcal{C} = \{(c_1, \ldots, c_k), c_i \in \mathbb{R}^{d+1} \text{ for } i = 1, \ldots, n\}$$

is called a codebook.

Let Q_k be the set of all k-point quantizers. The error (distortion) committed by an arbitrary k-point quantizer $q \in Q_k$ which summarizes (T, X) by $q(T, X)$ is defined by

$$D(P, q) = \mathbb{E}_P ||(T, X) - q(T, X)||^2. \tag{4.1}$$

The optimal performance over the class of k-point quantizers is given by

$$D_k^*(P) = \inf_{q \in Q_k} D(P, q).$$

A quantizer q^* is called optimal if $D(P, q^*) = D_k^*(P)$. It may be shown (see [Linder, 2002]), that such quantizer exists. Moreover, q^* is a nearest neighbor quantizer, that is a quantizer of the form

$$q(t, x) = \arg\min_{c_i \in \mathcal{C}} ||(t, x) - c_i||^2,$$

with the distortion

$$D(P, q^*) = \inf_{\mathcal{C} \in (\mathbb{R}^{d+1})^k} \mathbb{E} \min_{c_i \in \mathcal{C}} ||(T, X) - c_i||^2. \tag{4.2}$$

The last assertion means that the task of determining the optimal quantizer is reduced to the class of the nearest neighbor quantizers.

In practice, the law P of (T, X) is unknown and the optimal quantizer q^* cannot be calculated. Therefore, instead of minimizing (4.2), one minimizes its empirical version, where P is replaced with its approximation P_n constructed by means of observations of (T, X). If there is no censoring, such approximation is given by the empirical distribution induced by a sample $(T, X)_{1\leq i\leq n}$ of i.i.d. observations of (T, X). This leads us to the definition of the empirical distortion :

$$D(P_n, q) = \frac{1}{n}\sum_{i=1}^{n} ||(T_i, X_i) - q(T_i, X_i)||^2. \tag{4.3}$$

A quantizer $q_n^* \in Q_k$, minimizing (4.3) is called empirically optimal and its distortion is equal to $D(P, q_n^*) = \inf_{q\in Q_k} D(P_n, q)$.

In presence of censoring, the observed vector is no longer (T, X) but $(Y, \delta, X) = (\min(T, C), \mathbb{1}_{T\leq C}, X)$, where C is a censoring random variable. Therefore, the classical definition of the empirical distortion uses unobserved quantities and can not be used in this setting.

The rest of the chapter is organized as follows. In Section 4.2, we propose a generalized definition of the empirical distortion adapted to the presence of censoring and we define the empirically optimal quantizer as its minimizer. Section 4.3 deals with the asymptotic results for the distortion of empirically optimal quantizer. The new clustering algorithm is presented in Section 4.4. Section 4.5 proceeds with a numerical study of our algorithm.

4.2 Quantization under censoring

In this section, we are concerned with the quantization of random vector (T, X) taking values in $\mathbb{R}^{d+1}$, in presence of censoring acting on the variable T. As we have already seen, in this setting, we dispose only of i.i.d. replications

$$(Y_i, \delta_i, X_i)_{1\leq i\leq n}$$

of the observed vector (Y, δ, X). Therefore, the unknown law P of (T, X) can not be estimated by the empirical distribution and the classical definition of the empirical distortion is not appropriate. The basic idea of our approach is to approximate P by another random measure arising from the available observations. This measure is associated with the estimator of the distribution function of (T, X) due to [Stute, 1993]. Its consistency requires the following identifiability assumptions which will supposed to be satisfied throughout the paper :

- T and C are independent
- $P(T \leq C|X, Y) = P(T \leq C|Y)$

This set of assumptions is standard in survival analysis. For more details, we refer to [Stute, 1993], [Stute, 1996], [Stute, 1999], [Gannoun et al., 2007], [Lopez, 2009] and [Sánchez Sellero et al., 2005].

Let $Y_{[i:n]}$ be the i-th order statistics of the sample $(Y_1, \ldots, Y_n)$. We will denote by $\delta_{[i:n]}$ and $X_{[i:n]}$ the corresponding realizations of the indicator and of the covariate.

With these notations, the estimator of [Stute, 1993] takes the following form :

$$\hat{F}_n(t,x) = \sum_{i=1}^{n} W_{[i:n]} \mathbb{1}_{Y_{[i:n]} \leq t, X_{[i:n]} \leq x}, \quad t \in \mathbb{R},\ x \in \mathbb{R}^d, \tag{4.4}$$

where $W_{[i:n]}$ is the weight assigned to $Y_{[i:n]}$ by the univariate Kaplan-Meier estimator (see [Kaplan and Meier, 1958]) evaluated from sample $(Y_i, \delta_i)_{1 \leq i \leq n}$. It has the following expression (see [Stute and Wang, 1993]) :

$$W_{[i:n]} = \frac{\delta_{[i:n]}}{n-i+1} \prod_{j=1}^{i-1} \left(\frac{n-j}{n-j+1} \right)^{\delta_{[j:n]}}, \quad i = 1, \ldots, n. \tag{4.5}$$

We will adopt here an alternative form of the estimator (4.4) which has previously been used by [Satten and Datta, 2001]. Let W_{in} denote the weight attributed to the i-th observation and let $\hat{G}$ be a Kaplan-Meier estimator of the distribution function G of the censoring variable C. Using this notation, the estimator (4.4) can be written as follows :

$$\hat{F}_n(t,x) = \sum_{i=1}^{n} W_{in} \mathbb{1}_{Y_i \leq t, X_i \leq x}, \quad \text{with} \quad W_{in} = \frac{\delta_i}{n(1 - \hat{G}(Y_i-))}, \tag{4.6}$$

where $G(y-)$ denotes the left-hand limit of G at y.

In presence of censoring, the lack of large observations creates some difficulties for estimating tails of distributions and can make estimators inconsistent in the neighborhood of the upper bound of support. A common approach for overcoming this difficulty (see for instance [Heuchenne, 2008]) consists in truncating the distribution by a compact set $[0, \tau]$ strictly included in its support. However, the truncation τ may be chosen arbitrarily close to the upper bound of the support. This choice seems to be the best adapted to our theory, also for one more reason. In classical theory of quantization, obtaining theoretic results on rates of convergence of the empirically optimal quantizer requires that the variables have a bounded support (see [Pollard, 1982a], [Linder et al., 1994], [Bartlett et al., 1998] and [Linder, 2002]), which is usually not the case for the lifetimes. The introduction of truncation helps to get rid of this problem.

For the described reasons, in what follows, we will deal with the truncated distribution $P^\tau := P_{(T,X)|T \leq \tau}$. Let us consider now the corresponding distribution function $F^\tau(t,x)$ and the following estimator of it :

$$F_n^\tau(t,x) = \frac{\sum_{i=1}^n W_{in} \mathbb{1}_{Y_i \leq t, X_i \leq x} \mathbb{1}_{Y_i \leq \tau}}{\sum_{i=1}^n W_{in} \mathbb{1}_{Y_i \leq \tau}}, \quad t \in \mathbb{R},\ x \in \mathbb{R}^d. \tag{4.7}$$

We point out that (4.7) is an adaptation of the estimator (4.4) by introducing truncation and by normalizing a sum of its weights by 1. This estimator induces a probability measure

$$\mathcal{P}_n^\tau = \sum_{i=1}^{n} W_{in}^\tau \delta_{(Y_i, X_i)}, \ \text{with} \quad W_{in}^\tau = \frac{W_{in} \mathbb{1}_{Y_i \leq \tau}}{\sum_{i=1}^n W_{in} \mathbb{1}_{Y_i \leq \tau}}.$$

Now, it is natural to define the empirical distortion under censoring as the distortion with respect to the empirical law $\mathcal{P}_n^\tau$:

$$\begin{aligned}\mathcal{D}(\mathcal{P}_n^\tau, q) &= \frac{\sum_{i=1}^n W_{in}||(Y_i, X_i) - q(Y_i, X_i)||^2 \mathbb{1}_{Y_i \leq \tau}}{\sum_{i=1}^n W_{in} \mathbb{1}_{Y_i \leq \tau}} \\ &= \sum_{i=1}^n W_{in}^\tau ||(Y_i, X_i) - q(Y_i, X_i)||^2. \end{aligned} \tag{4.8}$$

In this context, a quantizer $q_n^* \in Q_N$ will be called empirically optimal when it minimizes the empirical distortion (4.8). This quantizer always exist due to [Pollard, 1982b]. The next section will be concerned with the asymptotic properties of q_n^*, that is with the convergence of its distortion $\mathcal{D}(\mathcal{P}^\tau, q_n^*)$ towards the optimal distortion $D_N^*(P^\tau)$ and with its rate.

4.3 Consistency of the empirical design

This section studies the asymptotic behavior of the empirically optimal quantizer q_n^*. Theorem 7 establishes the almost sure convergence of the distortion of q_n^* towards the minimal distortion $D_N^*(P^\tau)$. Theorem 8 provides an exponential inequality for the difference between these two distortions. Corollary 1 gives the rate of almost sure convergence.

4.3.1 Almost sure convergence

The following Theorem 7 establishes the almost sure convergence of the distortion of the empirically optimal quantizer. The proof is based on the fact that the absolute value of the difference between two distortions is bounded by a Wasserstein distance between the probability measure $\mathcal{P}_n^\tau = \sum_{i=1}^n W_{in}^\tau \delta_{(T_i, X_i)}$ and the conditional distribution P^τ of (T, X), given $T \leq \tau$. We show that the indicated distance converges almost surely to zero.

Theorem 7. For all $N \geq 1$, the empirically optimal N-quantizer satisfies

$$D(P^\tau, q_n^*) \xrightarrow[n\to\infty]{a.s.} D_N^*(P^\tau). \tag{4.9}$$

Proof. We recall that the Wasserstein distance between two probability measures μ and ν is defined by

$$\rho(\mu, \nu) = \inf_{X\sim\mu, Y\sim\nu} (\mathbb{E}||X - Y||^2)^{1/2},$$

where the infinium is taken over all random vectors (X, Y) having marginal distributions μ et ν, respectively. Any nearest neighbor quantizer satisfies (see [Linder, 2002]) :

$$|D(\mu, q)^{1/2} - D(\nu, q)^{1/2}| \leq \rho(\mu, \nu),$$

and

$$|D_N^*(\mu)^{1/2} - D_N^*(\nu)^{1/2}| \leq \rho(\mu, \nu).$$

Applying these inequalities to the probability measures $\mathcal{P}_n^\tau$ and P^τ, we obtain

$$|D(\mathcal{P}_n^\tau, q_n^*)^{1/2} - D(P^\tau, q^*)^{1/2}| \leq \rho(\mathcal{P}_n^\tau, P^\tau). \tag{4.10}$$

By the following, we will show that the right-hand side of (4.10) converges to zero almost surely, which implies the assertion of the theorem. To this aim, we recall that $\rho(\mathcal{P}_n^\tau, P^\tau) \overset{a.s.}{\underset{n\to\infty}{\to}} 0$ is equivalent to

$$P\left[\mathcal{P}_n^\tau \underset{n\to\infty}{\Rightarrow} P^\tau\right] = 1 \quad \text{and} \quad P\left[\int ||(t,x)||^2 d\mathcal{P}_n^\tau(t,x) \underset{n\to\infty}{\to} \int ||(t,x)||^2 dP^\tau(t,x)\right] = 1, \tag{4.11}$$

where $\Rightarrow$ denotes the weak convergence (see, for example, [Rachev and Rüschendorf, 1998]). We will prove that both conditions of (4.11) are satisfied. At first, let us recall ([Stute, 1993]) that, for any measurable function $\phi(t,x) : \mathbb{R}^{d+1} \to \mathbb{R}$, we have

$$\int \phi(t,x) dF_n(t,x) \underset{n\to\infty}{\overset{a.s.}{\to}} \mathbb{E}_P[\phi(T,X)]. \tag{4.12}$$

The same property is true for the modified estimator. Indeed,

$$\begin{aligned} \int \phi(t,x) dF_n^\tau(t,x) &= \frac{\sum_{i=1}^n W_{in}\phi(Y_i,X_i)\mathbb{1}_{Y_i\leq\tau}}{\sum_{i=1}^n W_{in}\mathbb{1}_{Y_i\leq\tau}} \\ &= \frac{\int \phi(t,x)\mathbb{1}_{t\leq\tau} dF_n(t,x)}{\int \mathbb{1}_{t\leq\tau} dF_n(t,x)} \underset{n\to\infty}{\longrightarrow} \mathbb{E}_{P^\tau}(\phi(T,X)) \quad \text{a.s.} \end{aligned}$$

Now, take $\phi(t,x) = ||(t,x)||^2$. This leads to

$$P\left(\int ||(t,x)||^2 d\mathcal{P}_n^\tau(t,x) \underset{n\to\infty}{\longrightarrow} \int ||(t,x)||^2 dP^\tau(t,x)\right) = 1. \tag{4.13}$$

Thus, it remains to prove that $P(\mathcal{P}_n^\tau \underset{n\to\infty}{\Rightarrow} P^\tau) = 1$. To that aim, we invoke arguments similar to [Varadarajan, 1958], who showed the weak convergence of empirical measure on the set of probability one. By Lévy criterium, all we need to obtain is

$$P\left(\forall u \in \mathbb{R}^{d+1} : \psi_n(u) \to \psi(u)\right) = 1,$$

where $\psi(u) = \int \exp(i\langle(t,x),u\rangle) dP^\tau(t,x)$ and $\psi_n(u) = \int \exp(i\langle(t,x),u\rangle) dP_n^\tau(t,x)$ are the Fourrier transforms of P^τ and P_n^τ. Remark that the event

$$\Omega(u) = \{\omega : \psi_n(u) \to \psi(u)\}$$

satisfies $P^\tau(\Omega(u)) = 1$ for all u, because of property (4.12) applied to $\phi(u) = \exp(i\langle(t,x),u\rangle)$. Let T be a countable dense subset of $\mathbb{R}^{d+1}$ and consider an event

$$\Omega_0 = \bigcap_{u\in T} \Omega(u) \bigcap \{\mathcal{P}_n^\tau||(t,x)|| \to P^\tau||(t,x)||\},$$

which is of probabilty equal to one. For any $u \in \mathbb{R}^{d+1}$ and $\omega_0 \in \Omega_0$, consider a sequence $\{u_k\}_{k=1}^\infty$, such that $u_k \in T$ and $u_k \to u$. For any fixed k, we have :

$$\begin{aligned} |\psi_n(\omega_0,u) - \psi(u)| &\leq |\psi_n(\omega_0,u) - \psi_n(\omega_0,u_k)| + |\psi_n(\omega_0,u_k) - \psi(u_k)| + \\ &\quad |\psi(u_k) - \psi(u)| \\ &\leq ||u-u_k|| \left(E_{P_n^\tau}||(T,X)|| + E_{P^\tau}||(T,X)||\right) \\ &\quad + |\psi_n(\omega_0,u_k) - \psi(u_k)|, \end{aligned}$$

with $E_{P_n^\tau}||(T,X)|| \underset{n\to\infty}{\to} E_{P^\tau}||(T,X)||$, as $\omega_0 \in \Omega_0$. Moreover, $\omega_0 \in \Omega(u_k)$ implies that, for any k,

$$\lim_{n\to\infty} \sup |\psi_n(\omega_0, u) - \psi(u)| \leq 2||u - u_k||E_{P^\tau}||(T,X)||.$$

Now, let k tend to infinity. This concludes the proof.

4.3.2 Exponential inequality

From now on, we will assume that the support of the random variable (T,X) is bounded, i.e. there exists some constant $R > 0$ such that, $P(||(T,X)|| \leq R) = 1$. In the previous section, we established the almost sure consistency of the empirical design. However, in real life applications the size of data sample is always finite. Therefore, a natural question concerns the rate of convergence. The following theorem provides a non asymptotic exponential bound for the difference between the distortion of empirically optimal quantizer and the minimal distortion.

Theorem 8. There exist some positive universal constants K, K_1, K_2, L_1, L_2 such that, for any $z > 4K/F_\tau^T$, with $F_\tau^T = P(T \leq \tau)$, the following inequality holds :

$$\begin{aligned} P(\sqrt{n}|D(P^\tau, q_n^*) - D(P^\tau, q^*)| > z) &\leq 5\exp\left(-L_1 z^2 + L_2 z\right) \\ &+ 2\left[\exp(-K_1 z^2) + \exp(-\sqrt{n}K_2 z)\right] \\ &+ O\left(e^{-\sqrt{n}}\right). \end{aligned}$$

We note that the remainder term $O\left(e^{-\sqrt{n}}\right)$ does not depend on z. It arises from the control of the difference between the distribution functions of (T,X) and C and their respective estimators on sets, which are not depending on z.

For the sake of clarity, the proof of Theorem 8 is postponed to Section 4.6. It is based on the empirical process theory applied to classes of functions indexed by N-point quantizers. The main idea is to bound the difference between the distortions by a deviation of the supremum of some empirical process indexed by a Donsker functional class. The exponential inequality follows then from a concentration inequality of [Talagrand, 1994]. One of the main difficulties is that the quantizer q_n^* is optimal with respect to the empirical measure $\mathcal{P}_n^\tau$ with random weights, depending on the Kaplan-Meier estimator $\hat{G}(y)$ of the censoring variable. In order to handle such measure, we need to replace $\hat{G}(y)$ with its deterministic limit $G(y)$. To that aim, it is necessary to control $\sup_y |\sqrt{n}(\hat{G}(y) - G(y))|$, where the supremum is taken over sets which are not depending on z. This is done through an exponential inequality of [Bitouzé et al., 1999] for the Kaplan-Meier estimator.

The following corollary provides a rate of the almost sure convergence in Theorem 7.

Corollary 4. For every probability measure P and $\tau > 0$, as $n \to \infty$

$$|D(P^\tau, q_n^*) - D(P^\tau, q)| = O\left(\frac{\log n}{\sqrt{n}}\right) \quad \text{a.s.} \tag{4.14}$$

Proof. The result is a corollary of Theorem 4.14 and Borel-Cantelli Lemma applied to the sequence of events $\Omega_n = \{|D(P^\tau, q_n^*) - D(P^\tau, q)| > z_n\}$ with $z_n = \log n$.

4.4 Clustering under censoring

4.4.1 Definition of clustering algorithm

Let $(T_i, X_i)_{1\leq i\leq n}$ be i.i.d. realizations of random vector (T, X). In presence of censoring, one observes only $(Y_i, \delta_i, X_i)_{1\leq i\leq n}$. This means that for some subjects, instead of observing the "true" realization (T_i, X_i), we have at our disposal only (Y_i, δ_i, X_i). A natural question is how to separate subjects into k clusters with respect to the unobserved realizations of interest $(T_i, X_i)_{1\leq i\leq n}$, having at hand only their censored versions. That is the purpose of the present section.

We recall that clustering is naturally related to vector quantization. It gives an optimal way of summarizing a distribution of random vector by k points (centers) and the rule of quantization (i.e. the nearest neighbor rule). In standard case, in order to perform clustering, one needs to evaluate the optimal centers by minimizing the empirical distortion and to assign to each observation a label of the cluster with the closest center. We emphasize that the corresponding minimization problem is NP-hard and the optimal centers are to be approximated using an iterative algorithm.

In our framework, the basic idea is the same. However, there are several difficulties related to the presence of censoring. They lead to an algorithm in two steps. Firstly, the centers of clusters are to be evaluated by minimizing the empirical distortion (4.8). As in non censored case, the corresponding numerical problem is NP-hard. Therefore, we need to define an iterative procedure for the approximation of the unknown centers. It must be adapted to the presence of censoring. The main issue in carrying out this task is that all distances related to censored observations are unobserved. Therefore, the classical k-means algorithm breaks down. Step 1 of our algorithm provides its generalization in our framework and allows for finding the centers of clusters. In contrast to the standard setting, that is not sufficient for assigning labels to all observations. Indeed, each non censored observation still can be affected to cluster with the closest center. For censored observations, the distances to centers are not observed and one is not able to assign them labels. Hence, our algorithm needs a second stage. The aim of Step 2 is to estimate the unknown distances related to each censored observation and to assign it the label of the cluster with center which is the closest with respect to the estimated distances.

Idea of Step 1. At each iteration of standard k-means algorithm, a center of cell S is actualized by a mean of its observations. This means estimates

$$\mathbb{E}\left[(T, X)|(T, X) \in S\right]. \tag{4.15}$$

In non censored case, each observation contributes with the weight $1/n$. In presence of censoring, the basic idea of estimators is to attribute weights only to non censored observations and to compensate for censoring by the distribution of the total mass. Following the same idea, we propose to actualize the center c of cell S by

$$c = \frac{\sum_{i=1}^{n}(Y_i, X_i)^T W_{in} \mathbb{1}_{\{(Y_i,X_i)\in S, \delta_i=1\}}}{\sum_{i=1}^{n} W_{in} \mathbb{1}_{\{(Y_i,X_i)\in S, \delta_i=1\}}}. \tag{4.16}$$

Here c is an estimator of (4.15) based on non censored observations which are taken with their weights compensating for censoring. The rest of the step is analogous to

the classical k-means. At the end of Step 1, the centers of k clusters are evaluated and all non censored observations received their labels.

Step 1 (Evaluation of k centers).

- Initialize the centers by $c_1^{(0)}, \dots, c_k^{(0)}$
- Evaluate the weights W_{in} of Kaplan-Meier estimator based on the sample $(Y_i, \delta_i)_{1 \le i \le n}$
- **Repeat until nothing changes :** for the iteration ℓ
  - Calculate Voronoï cells $S_1^\ell, \dots, S_k^\ell$ corresponding to centers $c_1^{(\ell)}, \dots, c_k^{(\ell)}$ for the set of non censored observations $\{(Y_i, X_i) : \delta_i = 1, i = 1, \dots, n\}$
  - For $j = 1, \dots, k$ calculate new centers $(c_j^{(\ell+1)})_{1 \le j \le k}$ as

$$c_j^{(\ell+1)} = \frac{\sum_{i=1}^n (Y_i, X_i)^T W_{in} \mathbb{1}_{\{(Y_i,X_i) \in S_j^l, \delta_i = 1\}}}{\sum_{i=1}^n W_{in} \mathbb{1}_{\{(Y_i,X_i) \in S_j^l, \delta_i = 1\}}}.$$

  - The algorithm stoppes in a finite number ℓ^* of iterations. For $j = 1, \dots, k$ attribute to observation (Y_i, X_i) with $\delta_i = 1$ a label j if $(T_i, X_i) \in S_j^{\ell^*}$.

Remark. Our theoretical results need to introduce a troncation bound τ which can be chosen arbitrarily close to the upper bound of the support of distribution. In practice, τ can be chosen equal to this bound without significant impact on the results.

Idea of Step 2. If the i-th observation is censored, we do not observe (T_i, X_i) and we only know that $\delta_i = 0$ and that $T_i > Y_i$. The best approximation of the unobserved distance $d((T_i, X_i); c)$ from this observation to center c is given by :

$$\mathbb{E}\left[d((T_i, X_i); c) \,|X_i, T_i > Y_i, \delta_i = 0\right].$$

Therefore, for each $i = 1, \dots, n$, such that $\delta_i = 0$ we estimate the distance between (T_i, X_i) and the center $c_j^{(\ell*)}$ by the following estimator of this conditional expectation :

$$\hat{d}_{ij} = \frac{\int_{Y_i}^\infty ||(t, X_i) - c_j^{(\ell*)}||^2 d\hat{F}(t|X_i)}{\int_{Y_i}^\infty d\hat{F}(t|X_i)}, \tag{4.17}$$

with $\hat{F}(t|x)$ estimating $F(t|X = x) = P(T \le t|X = x)$ given by

$$\hat{F}(t|x) = \frac{1}{n} \sum_{i=1}^n W_{in} \frac{k(\frac{x - X_i}{h})}{\sum_{j=1}^n k\left(\frac{x - X_j}{h}\right)} \mathbb{1}_{Y_i \le t}, \tag{4.18}$$

where $x \in \mathbb{R}^d$ and $k(x)$ is a kernel, that is a positive integrable function such that $\int_{\mathbb{R}^d} k(x)dx = 1$. Combining (4.17) and (4.18), we obtain

$$\hat{d}_{ij} = \frac{\sum_{m=1}^n W_{mn} ||(Y_m, X_i) - c^{(l*)}||^2 k\left(\frac{X_i - X_m}{h}\right) \mathbb{1}_{Y_m \ge Y_i}}{\sum_{m=1}^n W_{mn} k(\frac{X_i - X_m}{h}) \mathbb{1}_{Y_m \ge Y_i}}. \tag{4.19}$$

We present now the second step of our algorithm :

Étape 2 (Assigning labels to censored observations). In order to assign labels to censored observations :

- For each censored observation (Y_i, X_i) evaluate the estimated distances $\hat{d}_{ij}$ using (1.43).
- Assign to (Y_i, X_i) a label $j^* = \arg\min_j \hat{d}_{ij}$.

4.4.2 Number of clusters

Similarly to the other k-means type algorithms, our procedure uses the number k of clusters as the input value. However, in practice k is unknown and is to be chosen adaptively. This issue was extensively studied in absence of censoring. A lot of criterions were proposed in the literature, a complete review and a comparative Monte Carlo study of most of them can be found in [Milligan and Cooper, 1985]. Several of these criterions can be adapted in presence of censoring. We propose here a rule for choosing the number of clusters, which is an adaptation of the criterion of [Krzanowski and Lai, 1988] proposed for non censored data. In absence of censoring, one have to calculate, for some range of values of k, the pooled within-cluster sum of squares S_k and a quantity

$$DIFF(k) = (k-1)^{2/(d+1)} S_{k-1} - k^{2/(d+1)} S_k,$$

where $d+1$ is the total number of the variables. [Krzanowski and Lai, 1988] proposed to chose k maximizing

$$KL(k) = \left| \frac{DIFF(k)}{DIFF(k+1)} \right|. \tag{4.20}$$

In our case, we propose to replace S_k by the weighted sum of squares involving only non censored observations :

$$\mathcal{D}_k = \sum_{i=1}^{n} W_{in} \min_{c_j \in \mathcal{C}} \|(Y_i, X_i) - c_j\|^2,$$

where $\mathcal{C} = \{c_1, \ldots, c_k\}$ are the centers of k clusters resulting from our iterative algorithm. Similarly to (4.20), the optimal number of clusters is to be chosen as the value of k maximizing

$$\left| \frac{(k-1)^{2/(d+1)} \mathcal{D}_{k-1} - k^{2/(d+1)} \mathcal{D}_k}{k^{2/(d+1)} \mathcal{D}_k - (k+1)^{2/(d+1)} \mathcal{D}_{k+1}} \right|.$$

4.5 Simulations and a real data study

4.5.1 Simulations

In this section, we evaluate the performance of our algorithm on simulated data sets. We proceed in the following way. At the first step, a complete data sample with known clusters is created and a k-means algorithm is applied, in order to obtain a

partition of the data into k clusters. The accuracy of this partition is then compared to that of the partition produced by our algorithm, having as the input the censored version of the initial sample.

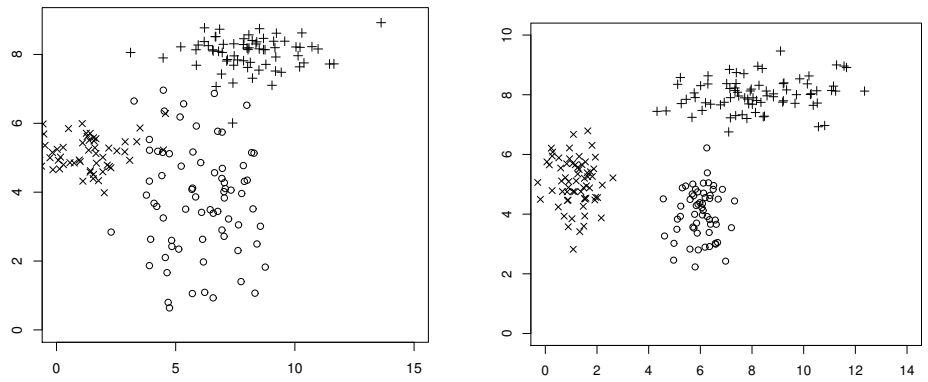
This comparison is done through the corrected Rand's statistics (see [Rand, 1971] and [Hubert and Arabie, 1985]). Rand's index permits to compare two partitions P_1 and P_2 in order to know how close they are. In the set of all possible pairs of observations let A (for "agreement") denote the number of pairs which are of one of the following types :

- Pairs of observations belonging to the same class in P_1 and P_2
- Pairs of observations belonging to a different class in P_1 and to a different class in P_2

The total number of pairs being $n(n-1)/2$, Rand's statistics is defined by $R_{P_1P_2} = 2A/(n(n-1))$. The closer $R_{P_1P_2}$ is to one, the closer are two partitions.

Plan of simulation study :

For each of three different levels of censoring $(15\%, 30\%, 45\%)$, we generated 1000 bivariate data sets of $n = 200$ observations. For each $j = 1, \ldots, 1000$, the j-th data set is composed of $k = 3$ clusters (clusters are supposed to be known), forming the partition denoted by $\mathcal{P}_j^0$. Data are simulated using a Gaussian mixture, in two following cases : groups are close (see Figure 4.1, (a)) and groups are well separated (see Figure 4.1, (b)).



(a) Case 1 : groups are close (b) Case 2 : groups are well separated

Figure 4.1 – Examples of simulated data.

The exact scheme of simulations is the following. For each level of censoring, $k = 3$ and $j = 1, \ldots, 1000$,

- Simulate a complete data sample $(T_i^{(j)}, X_i^{(j)})_{1 \leq i \leq n}$. Apply a k-means algorithm, leading to a partition of these data into k clusters. Denote this partition by $\mathcal{P}_j^c$.

- Simulate a sample $(C_i^{(j)})_{1\leq i\leq n}$ from censoring variable and get the censored data set as $(\min(T_i^{(j)}, C_i^{(j)}), \delta_i^{(j)}, X_i^{(j)})_{1\leq i\leq n}$. Apply the algorithm described in Section 4.4 and denote the resulting partition into k clusters by $\mathcal{P}_j$.
- Calculate Rand's statistics $R_{\mathcal{P}_j^c\mathcal{P}_j^0}$ and $R_{\mathcal{P}_j^c\mathcal{P}_j^0}$.

The value of $R_{\mathcal{P}_j^c\mathcal{P}_j^0}$ shows how accurate is the the partition created by k-means (applied to sample before censoring) with respect to the known "true" partition $\mathcal{P}_j^0$, and $R_{\mathcal{P}_j^c\mathcal{P}_j^0}$ have the same meaning for our algorithm.

Results. In Table 4.1 (case of close groups) and Table 4.2 (case of well separated groups), we present the mean values (over $N = 1000$ data sets) $R^N_{\mathcal{P}\mathcal{P}^0} = 1/N \sum_{j=1}^n R_{\mathcal{P}_j\mathcal{P}_j^0}$ and $R^N_{\mathcal{P}^c\mathcal{P}^0} = 1/N \sum_1^n R_{\mathcal{P}_j^c\mathcal{P}_j^0}$ of the corrected Rand's statistics. Not

Level of censoring	15%	30%	45%
$R^N_{\mathcal{P}^c\mathcal{P}^0}$	0.931	0.929	0.930
$R^N_{\mathcal{P}\mathcal{P}^0}$	0.905	0.878	0.851
$R^N_{\mathcal{P}^c\mathcal{P}^0} - R^N_{\mathcal{P}\mathcal{P}^0}$	0.026	0.051	0.079

TABLE 4.1 – Corrected Rand's statistics for simulated data, close groups

Level of censoring	15%	30%	45%
$R^N_{\mathcal{P}^c\mathcal{P}^0}$	0.993	0.994	0.994
$R^N_{\mathcal{P}\mathcal{P}^0}$	0.972	0.942	0.923
$R^N_{\mathcal{P}^c\mathcal{P}^0} - R^N_{\mathcal{P}\mathcal{P}^0}$	0.021	0.052	0.071

TABLE 4.2 – Corrected Rand's statistics for simulated data, separated groups

surprisingly, both methods perform better in the case of three well separated groups. We remark also that the agreement between the partition by our method and the true partition decreases when the proportion of censored observations increase and the difference between our method (based on available in reality information) and the procedure based on full (unavailable in reality) information becomes more important. However, it does not rise drastically and the agreement for our method remains relatively good even at 45% of censoring.

4.5.2 Real data analysis : PBC data

In this section, we illustrate our results by an application to a real data set. We consider the data from the Mayo Clinic trial in primary biliary cirrhosis (PBC) of the liver which is a rare and fatal chronic liver disease. The study had been conducted between 1974 and 1984. For a total number of 418 patients the recorded measurements are the time at risk (censored), censored indicator and 17 covariates such as a patient age, sex, clinical, biochemical and histological measurements. After excluding the participants with missing values of some covariates, we obtained a data

set of 258 observations, 111 of which are non censored. The detailed description of the data set can be found in [Fleming and Harrington, 1991].

For the easier interpretability of the results we performed clustering using only variables which were shown to be important for the survival (see the study of the same data in the regression setting conducted by [Grambsch et al., 1989]). These covariates are the patients' age, total serum bilirubin mentioned as one of the most important factors influencing the lifetime, serum albumin concentration and the prothrombin time. We excluded the severity of edema variable as our method permits to take into account only the quantitative covariates.

Before clustering, the observations of each variable were normalized by dividing by the corresponding range. Our algorithm has detected four clusters of patients. Table 4.3 presents the mean values of the survival time and of the covariates for each group and for all of the patients. The results show that the discriminative

	Survival	Age	Bilirubin	Prothrombin	Albumin
Group 1	3001.66	47.11	2.16	10.70	3.54
Group 2	2394.97	51.77	3.88	11.23	3.54
Group 3	1746.10	53.57	5.41	11.14	3.38
Group 4	1145.09	58.49	8.21	11.30	3.49
Overall means	2180.26	50.42	3.34	10.75	3.51

Table 4.3 – Clusters found in PBC data

variables seem to be the survival time, the bilirubin level and the age. Group 1 is characterized by the most important survival time associated with the low level of bilirubin and the lowest age of patients. In contrast, Group 4 is represented by the lowest survival, a very high bilirubin level and the most important age. Groups 2 and 3 are the medium cases between 1 and 4. One can see clearly that the survival is strongly associated with the bilirubin level. This fact is in concordance with the recognized importance of the factor. The mean prothrombin and albumin levels seem to be rather close for the different groups.

In conclusion, the group of the highest risk is composed of patients with the most important level of bilirubin and great ages while the lowest risk corresponds to the youngest patients with the lowest level of bilirubin.

4.6 Proof of Theorem 8

In this section, we are giving the proof of the exponential inequality announced in Section 4.3.

Proof. We have

$$P(\sqrt{n}|D(P^\tau, q_n^*) - D(P^\tau, q^*)| > z) \;\leq\; P(\sqrt{n} \sup_{q \in Q_N} |D(P^\tau, q) - D(\mathcal{P}_n^\tau, q)| > z/2).$$

Using notation $f_q(y,x) = ||(y,x) - q(y,x)||^2$, $F_\tau^T = P(T \leq \tau)$, $\hat{P}_n = \sum_{i=1}^n W_{in}\delta_{(Y_i,X_i)}$ and $F_{\tau,n}^T = \hat{P}_n(T \leq \tau)$, we obtain

$$\begin{aligned}
|D(P^\tau, q) - D(P_n^\tau, q)| &= \left| \int f_q(y,x) \mathbb{1}_{y\leq\tau} \frac{dP(y,x)}{P(T\leq\tau)} - \int f_q(y,x)\mathbb{1}_{y\leq\tau}\frac{d\hat{P}_n(y,x)}{\hat{P}_n(T\leq\tau)} \right| \\
&= \left| \frac{1}{F_\tau^T} \int f_q(y,x)\mathbb{1}_{y\leq\tau} d(P-\hat{P}_n) \right. \\
&\qquad \left. + \frac{F_{\tau,n}^T - F_\tau^T}{F_\tau^T F_{\tau,n}^T} \int f_q(y,x)\mathbb{1}_{y\leq\tau} d\hat{P}_n \right|.
\end{aligned}$$

Therefore,

$$P(\sqrt{n} \sup_{q\in Q_N} |D(P^\tau, q) - D(\mathcal{P}_n^\tau, q)| > z/2) \leq T_1 + T_2,$$

where

$$\begin{aligned}
T_1 &:= P\left(\sqrt{n} \sup_{q\in Q_N} \left| \int f_q(y,x)\mathbb{1}_{y\leq\tau} d(P-\hat{P}_n) \right| > \frac{z}{4} F_\tau^T \right), \\
T_2 &:= P\left(\sqrt{n} \sup_{q\in Q_N} \left| \frac{F_{\tau,n}^T - F_\tau^T}{F_\tau^T F_{\tau,n}^T} \int f_q(y,x)\mathbb{1}_{y\leq\tau} d\hat{P}_n \right| > \frac{z}{4} \right).
\end{aligned}$$

In the following we will consider separately the terms T_1 and T_2.

1. The first term $\mathbf{T_1}$. Let us denote by $P_n(y,x,\delta) = \frac{1}{n}\sum_{i=1}^n \delta_{(Y_i,X_i,\delta_i)}$ the empirical measure of the available observations. For any function $\phi(y,x)$, we have

$$E\left[\frac{\delta}{1-G(Y-)} \phi(Y,X) \right] = E\left[\frac{E[\mathbb{1}_{Y\leq C}|Y,X]\phi(Y,X)}{1-G(Y-)} \right] = E\left[\phi(T,X)\right].$$

Therefore, the term T_1 can be written as

$$\begin{aligned}
T_1 &= P\left(\sqrt{n} \sup_{q\in Q_N} \left| \int f_q(y,x) \frac{\delta\mathbb{1}_{y\leq\tau} dP(y,x,\delta)}{1-G(y-)} \right.\right. \\
&\qquad \left.\left. - \int f_q(y,x) \frac{\delta\mathbb{1}_{y\leq\tau} dP_n(y,x,\delta)}{1-\hat{G}_n(y-)} \right| > \frac{z}{4} F_\tau^T \right) \\
&\leq T_{11} + T_{12},
\end{aligned}$$

where

$$\begin{aligned}
T_{11} &:= P\left(\sqrt{n} \sup_{q\in Q_N} \left| \int f_q(y,x) \frac{\delta\mathbb{1}_{y\leq\tau}}{1-G(y-)} d(P-P_n) \right| > \frac{z}{4} F_\tau^T \right), \\
T_{12} &:= P\left(\sqrt{n} \sup_{q\in Q_N} \left| \int f_q(y,x)\delta\mathbb{1}_{y\leq\tau} \left[\frac{\hat{G}_n(y-) - G(y-)}{(1-\hat{G}_n(y-))(1-G(y-))} \right] dP_n \right| > \frac{z}{4} F_\tau^T \right).
\end{aligned}$$

Term $\mathbf{T_{11}}$. In order to handle the term T_{11}, let us first introduce first a class of functions

$$\mathcal{F}_1 = \left\{ g_q : g_q = \frac{\delta\mathbb{1}_{t\leq\tau}}{1-G(t-)} f_q(t,x), \quad q \in Q_N \right\}, \tag{4.21}$$

indicated by N-quantizers. For any $u > 0$, we have the following majoration :

$$P\left(\sqrt{n}\sup_{q\in Q_N}\left|\int f_q(y,x)\frac{\delta\mathbb{1}_{y\leq\tau}}{1-G(y-)}d(P-P_n)\right|>u\right)\leq P\left(\sqrt{n}\sup_{f\in\mathcal{F}_1}|(P_n-P)f|>u\right),$$

where we used the notation $Pf = \int f dP$ and $P_n f = \int f dP_n$. The exponential inequality for T_{11} follows from a concentration inequality proposed by [Talagrand, 1994] in the form, used by [Einmahl and Mason, 2005]. Let us first remark that,

$$\mathcal{F}_1 = h_\tau(t,\delta)\times\mathcal{F}_2, \tag{4.22}$$

where

$$\mathcal{F}_2 := \{f_q(t,x), \quad q\in Q_N\} \quad \text{and} \quad h_\tau(t,\delta) = \frac{\delta\mathbb{1}_{t\leq\tau}}{1-G(t-)}.$$

The class $\mathcal{F}_2$ is $P-$Donsker. Indeed, as proved e.g. in [Linder, 2002], the collection of sets $\{\{(t,x) : f_q(t,x) > u\}, u > 0, q \in Q_N\}$ forms a VC-class. Therefore, the class $\mathcal{F}_2$ is VC-major by definition given in Section 2.6.4 of [van der Vaart and Wellner, 1996], and is $P-$Donsker by Theorem 2.6.14 of Section 2.6.4.

Moreover, the function $h : (t,\delta) \to \frac{\delta\mathbb{1}_{t\leq\tau}}{1-G(t-)}$ is bounded. Consequently, $\mathcal{F}_1$ is $P-$Donsker as the pointwise product of the $P-$Donsker class $\mathcal{F}_2$ and the bounded function (see the permanence property in Example 2.10.10 of [van der Vaart and Wellner, 1996]).

We are now ready to apply the inequality of [Talagrand, 1994]. It states that, for any pointwise measurable class $\mathcal{F}$, satisfying $||f||_{\mathcal{F}} := \sup_{f\in\mathcal{F}}||f|| \leq M$ for some constant $0 < M < \infty$, we have for all $u > 0$,

$$\begin{aligned}P\left(\sqrt{n}\sup_{f\in\mathcal{F}}|(P_n-P)f|\geq\frac{A_1}{\sqrt{n}}\left(E||P_n^0 f||_{\mathcal{F}}+u\right)\right) \;\leq\; & 2\Big[\exp(-A_2u^2/n\sigma_{\mathcal{F}}^2)\\ & +\exp(-A_2u/M)\Big],\end{aligned} \tag{4.23}$$

where $P_n^0 f = \sum_{i=1}^n \varepsilon_i f(Y_i, X_i, \delta_i)$, with i.i.d. Rademacher random variables $(\varepsilon_i)_{1\leq i\leq n}$, $\sigma_{\mathcal{F}}^2 = \sup_{f\in\mathcal{F}} Var(f(Y,X,\delta))$ and A_1, A_2 are universal constants. For any function $f \in \mathcal{F}_1$, we have $||f||_{\mathcal{F}_1} \leq 4R^2(1-G(\tau-))^{-1}$. The application of the inequality (4.23) to the class of functions $\mathcal{F}_1$ gives

$$\begin{aligned}P\;\left(\sqrt{n}\sup_{f\in\mathcal{F}_1}|(P_n-P)f|\geq A_1u+\frac{A_1E||P_n^0f||_{\mathcal{F}_1}}{\sqrt{n}}\right)\\ \leq 2\left[\exp(-A_2u^2/\sigma_{\mathcal{F}_1}^2)+\exp(-\sqrt{n}A_2u/M)\right],\end{aligned} \tag{4.24}$$

where $M := 4R^2(1-G(\tau-))^{-1}$.

Using Propostion 1 of [Einmahl and Mason, 2005], we will show that the term $A_1E||P_n^0f||_{\mathcal{F}_1}/\sqrt{n}$ is uniformly bounded by some constant B_1. Indeed, as all functions of the class $\mathcal{F}_1$ are uniformly bounded, the only condition to be verified is the inequality $N(\varepsilon,\mathcal{F}_1) \leq C\varepsilon^{-\nu}$ on covering numbers, for some constants $C, \nu \geq 1$ and every $\varepsilon \in (0,1)$. This condition is satisfied. Indeed, by [Linder, 2002], the considered class of functions is a VC class. Therefore, Theorem 2.6.7 of

[van der Vaart and Wellner, 1996] applies and gives the required bound on the covering number.

According to Proposition 1 of [Einmahl and Mason, 2005], there exists some constant B_1, such that $E||P_n^0 f||_{\mathcal{F}} \leq B_1\sqrt{n}$. The inequality (4.24) takes form

$$P\left(\sqrt{n}\sup_{f\in\mathcal{F}_1}|(P_n-P)f| \geq A_1u + B_1\right) \leq 2\left[\exp(-A_2u^2/\sigma_{\mathcal{F}_1}^2) + \exp(-\sqrt{n}A_2u/M)\right], \tag{4.25}$$

for any $u > 0$ and some universal constants A_1, B_1. For any $v > K := \min(A_1+B_1, 1)$ (4.25) can be rewritten in the form

$$P\left(\sqrt{n}\sup_{f\in\mathcal{F}_1}|(P_n-P)f| \geq v\right) \leq 2\left[\exp(-K_1u^2) + \exp(-\sqrt{n}K_2u)\right], \tag{4.26}$$

where $K_1 = A_2(\sigma_{\mathcal{F}_1}(A_1+B_1))^{-2}$ and $K_2 = A_2M^{-1}(A_1+B_1)^{-1}$. Therefore, for any $z > 4K/F_\tau^T$,

$$T_{11} \leq 2\left[\exp(-K_1z^2) + \exp(-\sqrt{n}K_2z)\right].$$

Term $\mathbf{T_{12}}$. Let us use the following decomposition,

$$\begin{aligned}
T_{12} &\leq P\left(\sqrt{n}\sup_{q\in Q_N}\left|\int f_q(y,x)\mathbb{1}_{y\leq\tau}\delta dP_n\right| \times \sup_{y\leq\tau}\left|\frac{(\hat{G}_n-G)(y-)}{1-\hat{G}_n(y-)}\right| > \frac{z}{4}F_\tau^T(1-G(\tau))\right)\\
&\leq P\left(\sqrt{n}\sup_{y\leq\tau}\left|\frac{(\hat{G}_n-G)(y-)}{1-\hat{G}_n(y-)}\right| > \frac{z}{16R^2}F_\tau^T(1-G(\tau))\right)\\
&\leq P(\mathcal{A}_n) + P(\mathcal{B}_n),
\end{aligned}$$

where

$$\begin{aligned}
\mathcal{A}_n :\ = &\left\{\sqrt{n}\sup_{y\leq\tau}\left|\frac{\hat{G}_n(y-)-G(y-)}{1-\hat{G}_n(y-)}\right| > \frac{zF_\tau^T(1-G(\tau))}{16R^2}\right\}\\
&\cap\left\{\sup_{y\leq\tau}|\hat{G}_n(y-)-G(y-)| \leq \frac{1-G(\tau)}{2}\right\},
\end{aligned}$$

and

$$\mathcal{B}_n := \left\{\sup_{y\leq\tau}|\hat{G}_n(y-)-G(y-)| > \frac{1-G(\tau)}{2}\right\}.$$

For the first term we have,

$$\begin{aligned}
P(\mathcal{A}_n) &\leq P\left(\left\{\sqrt{n}\sup_{y\leq\tau}\left|\hat{G}_n(y-)-G(y-)\right| > \frac{z}{32R^2}F_\tau^T(1-G(\tau))^2\right\}\right)\\
&\leq 2.5\exp\{-2\lambda_1^2(\tau)z^2 + C\lambda_1(\tau)z\},
\end{aligned}$$

with $\lambda_1(\tau) = F_\tau^T(1-F_\tau^T)(1-G(\tau))^2/(32R^2)$, where the last inequality follows from Theorem 2 of [Bitouzé et al., 1999]. The same theorem applied to the second term gives,

$$\begin{aligned}
P(\mathcal{B}_n) &\leq P\left(\sup_{y\leq\tau}|(1-F^T(y-))(\hat{G}_n(y-)-G(y-)| > \frac{(1-G(\tau))(1-F_\tau^T)}{2}\right)\\
&\leq 2.5\exp\{-\sqrt{n}(-2\tilde{\lambda}_1^2(\tau) + C\tilde{\lambda}_1(\tau))\},
\end{aligned}$$

where $\tilde{\lambda}_1(\tau) = (1 - G(\tau))(1 - F_\tau^T))/2$.

2. The second term $\mathbf{T_2}$. The estimation of the second term T_2 is similar to that of T_{12}. Indeed,

$$\begin{aligned} T_2 &= P\left(\sqrt{n}\sup_{q\in Q_N}\left|\frac{F_{\tau,n}^T - F_\tau^T}{F_{\tau,n}^T}\int f_q(y,x)\mathbb{1}_{y\le\tau}d\hat{P}_n\right| > \frac{z}{4}F_\tau^T\right) \\ &= P\left(\sqrt{n}\sup_{q\in Q_N}\left|\frac{F_{\tau,n}^T - F_\tau^T}{F_{\tau,n}^T}\int f_q(y,x)\mathbb{1}_{y\le\tau}\frac{\delta}{1-\hat{G}_n(y-)}dP_n\right| > \frac{z}{4}F_\tau^T\right) \\ &\le P\left(\sqrt{n}\left|\frac{F_{\tau,n}^T - F_\tau^T}{F_{\tau,n}^T}\right| > \frac{z}{32R^2}F_\tau^T(1-G(\tau))\right) \\ &\quad +2.5\exp\{-\sqrt{n}(-2\tilde{\lambda}_1^2(\tau) + C\tilde{\lambda}_1(\tau))\} \\ &\quad =: T_{21} + T_{22}. \end{aligned}$$

The first term can be decomposed as $T_{21} = P(\mathcal{A}_n^{'}) + P(\mathcal{B}_n^{'})$, where

$$\mathcal{A}_n^{'} = \left\{\sqrt{n}\left|\frac{F_{\tau,n}^T - F_\tau^T}{F_{\tau,n}^T}\right| > \frac{z}{32R^2}F_\tau^T(1-G(\tau))\right\} \cap \left\{|F_{\tau,n}^T - F_\tau^T| \le F_\tau^T/2\right\},$$

and

$$\mathcal{B}_n^{'} = \left\{|F_{\tau,n}^T - F_\tau^T)| > F_\tau^T/2\right\}.$$

Using again [Bitouzé et al., 1999] we obtain,

$$\begin{aligned} P(\mathcal{A}_n^{'}) &\le P\left(\sqrt{n}\sup_{y\le\tau}\left|(1-G(y-))(\hat{F}_n(y) - F(y))\right| > \frac{z}{64R^2}(F_\tau^T)^2(1-G(\tau))^2\right) \\ &\le 2.5\exp\{-2\lambda_2^2(\tau)z^2 + C\lambda_2(\tau)z\}, \end{aligned}$$

where $\lambda_2(\tau) = (F_\tau^T)^2(1-G(\tau))^2/(64R^2)$. Moreover,

$$P(\mathcal{B}_n^{'}) \le 2.5\exp\{-\sqrt{n}(-2\tilde{\lambda}_2^2(\tau) + C\tilde{\lambda}_2(\tau))\},$$

with $\tilde{\lambda}_2(\tau) = F_\tau^T(1-G(\tau-))/2$. Bringing together all the inequalities, we obtain the assertion of the theorem.

Conclusion et perspectives

Dans cette thèse, nous avons considéré l'estimation non paramétrique, la modélisation par des copules et la quantification de la loi jointe des variables aléatoires censurées. La première approche est une façon d'étudier la loi jointe dans son ensemble, sans séparer la structure de dépendance des lois marginales. La deuxième approche permet de s'affranchir de l'étude des marginales et de se focaliser sur la modélisation de la structure de dépendance. Nous avons appliqué nos résultats théoriques sur les données réelles d'assurance portant sur les durées de vie de conjoints. En regardant plus précisément les données, on peut se rendre compte que la population sous-jacente possède une structure hétérogène et se compose de classes de risque. La prise en compte de cette structure est une façon d'améliorer la précision de l'analyse statistique. Par exemple, sur les données que l'on a considérées, on pourrait utiliser les familles de copules spécifiques pour chaque groupe homogène de conjoints. Ces considérations nous ont conduit à la troisième question de la thèse, celle de la détection sous censure de groupes homogènes au sein d'une population hétérogène. Nous avons fait un premier pas dans cette direction en traitant le cas des observations multivariées avec une seule composante censurée.

Plus que des conclusions, nous tenons à dégager un certain nombre de perspectives de recherches futures :

Troncature. Dans cette thèse, nous nous sommes focalisés sur l'étude des données affectées par la censure, sans pour autant tenir compte d'un deuxième phénomène parasite assez courant, à savoir la troncature. Elle est présente notamment dans les données canadiennes portant sur les durées de vie jointes et n'a pas pu être prise en compte par nos estimateurs. Une piste de recherche serait la généralisation de nos méthodes en présence de troncature.

Tests d'adéquation. Dans le Chapitre 3, nous avons introduit des estimateurs de la copule et de sa densité en présence de censure bivariée. Les estimateurs de la copule ont été utilisés pour construire un nouveau test d'adéquation, sans pour autant étudier ses propriétés. De manière claire, la censure entraine une perte d'information et diminue la puissance du test. Cette dernière nécessite d'être étudiée théoriquement afin de juger l'importance de l'impact de la censure, mais aussi pour comparer la performance de notre méthode à celle des autres méthodes existantes.

Dans le cadre non censuré, il existe une approche différente des tests d'adéquations pour les modèles de copules, qui est basée sur l'estimateur de sa densité (voir [Fermanian, 2005]). En utilisant les estimateurs de la densité de copule définis dans le Chapitre 3, il est possible de généraliser cette approche en présence de censure.

Évolution de la dépendance dans le temps. Si l'on revient sur le cas de deux durées dépendantes, [Spreeuw, 2006] souligne l'évolution au cours du temps du degré de dépendance entre les durées de vie de conjoints. Par exemple, [Parkes et al., 1969] et [Jagger and Sutton, 1991] montrent que le décès d'un des deux conjoints a un impact immédiat sur la durée de vie du second, mais qui disparait progressivement au cours des six mois suivants. Par conséquent, pour les compagnies d'assurance, les risques liés aux contrats à deux têtes ont également une évolution temporelle qu'il serait pertinent de modéliser. Pour faire cela, une piste envisageable serait d'introduire de manière explicite la dépendance temporelle dans la copule.

Clustering des observations censurées. Dans le Chapitre 4, nous avons envisagé une étude exploratoire de données multivariées en présence de censure. Néanmoins, nous nous sommes restreint à la situation où seulement une composante du vecteur d'observations est censurée. Notre approche pourrait être généralisée au cas de plusieurs variables censurées. Pour cela, au lieu de quantifier la mesure empirique associée à l'estimateur de [Stute, 1993], il faudrait prendre celle associée à un estimateur de la loi jointe de plusieurs variables censurées (par exemple, celle associée à l'estimateur plus général donné dans [Lopez, 2012]).

Il pourrait également être envisageable et utile d'explorer d'autres approches à la classification non supervisée en présence de censure.

Copules et grande dimension en présence de censure. Dans les études génétiques des maladies, où les données sont de grande dimension, un problème important est la sélection de gènes pertinents pour décrire la durée de survie. Une méthode très utilisée dans la littérature est basée sur la régression de Cox et est valable uniquement si la durée et sa censure sont indépendantes. Cette hypothèse, bien que garantissant l'identifiabilité des modèles statistiques, peut s'avèrer irréaliste pour ce type d'applications. Pour s'en affranchir, on peut envisager la modélisation de la dépendance entre la durée et sa censure par une copule. Un pas très récent dans cette direction a été fait par [Emura and Chen, 2014], qui ont utilisé cette approche pour mettre en évidence le biais de la procédure classique de la sélection de gènes causé par la censure dépendante, et pour proposer une procédure alternative basée sur les copules et permettant d'en tenir compte. Cette idée intéressante mérite certainement d'être explorée davantage.

Les copules peuvent également servir pour d'autres problématiques de grande dimension. A titre d'exemple, une approche explorable serait de modéliser par les copules de type hiérarchique où *vine copula* la dépendance entre la durée censurée et un vecteur de variables explicatives de grande dimension.

Une autre perspective envisageable serait la prise en compte, à l'aide des copules conditionnelles, d'un vecteur de variables explicatives dans la modélisation de la dépendance entre deux durées de vie.

Bibliographie

[Abaya and Wise, 1984] Abaya, E. F. and Wise, G. L. (1984). Convergence of vector quantizers with applications to optimal quantization. *SIAM J. Appl. Math.*, 44(1) :183–189.

[Akritas, 1986] Akritas, M. G. (1986). Bootstrapping the Kaplan-Meier estimator. *J. Amer. Statist. Assoc.*, 81(396) :1032–1038.

[Akritas and Van Keilegom, 2003] Akritas, M. G. and Van Keilegom, I. (2003). Estimation of bivariate and marginal distributions with censored data. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 65(2) :457–471.

[Barbe et al., 1996] Barbe, P., Genest, C., Ghoudi, K., and Rémillard, B. (1996). On Kendall's process. *J. Multivariate Anal.*, 58(2) :197–229.

[Bartlett et al., 1998] Bartlett, P. L., Linder, T., and Lugosi, G. (1998). The minimax distortion redundancy in empirical quantizer design. *IEEE Trans. Inform. Theory*, 44(5) :1802–1813.

[Beran, 1981] Beran, R. (1981). Nonparametric regression with randomly censored survival data. Technical report, University of California, Berkeley.

[Billingsley, 1999] Billingsley, P. (1999). *Convergence of probability measures*. Wiley Series in Probability and Statistics : Probability and Statistics. John Wiley & Sons Inc., New York, second edition. A Wiley-Interscience Publication.

[Bitouzé et al., 1999] Bitouzé, D., Laurent, B., and Massart, P. (1999). A Dvoretzky-Kiefer-Wolfowitz type inequality for the Kaplan-Meier estimator. *Ann. Inst. H. Poincaré Probab. Statist.*, 35(6) :735–763.

[Bouyé et al., 2007] Bouyé, E., Durrleman, V., Nikeghbali, A., Riboulet, G., and Roncalli, T. (2007). Copulas for Finance - A Reading Guide and Some Applications. *Social Science Research Network Working Paper Series*.

[Campbell and Földes, 1982] Campbell, G. and Földes, A. (1982). Large-sample properties of nonparametric bivariate estimators with censored data. In *Nonparametric statistical inference, Vol. I, II (Budapest, 1980)*, volume 32 of *Colloq. Math. Soc. János Bolyai*, pages 103–121. North-Holland, Amsterdam.

[Cao and González-Manteiga, 2008] Cao, R. and González-Manteiga, W. (2008). Goodness-of-fit tests for conditional models under censoring and truncation. *J. Econometrics*, 143(1) :166–190.

[Carriere, 2000] Carriere, J. F. (2000). Bivariate survival models for coupled lives. *Scand. Actuar. J.*, (1) :17–32.

[Cherubini et al., 2004] Cherubini, U., Luciano, E., and Vecchiato, W. (2004). *Copula methods in finance*. Wiley Finance Series. John Wiley & Sons Ltd., Chichester.

[Choros et al., 2010] Choros, B., Ibragimov, R., and Permiakova, E. (2010). Copula estimation. In F. Durante, W. Haerdle, P. J. and T. Rychlik, e., editors, *Workshop on Copula Theory and its Applications. Lecture Notes in Statistics- Proceedings*. Springer.

[Dabrowska, 1988] Dabrowska, D. M. (1988). Kaplan-Meier estimate on the plane. *Ann. Statist.*, 16(4) :1475–1489.

[Deheuvels, 1979] Deheuvels, P. (1979). La fonction de dépendance empirique et ses propriétés. Un test non paramétrique d'indépendance. *Acad. Roy. Belg. Bull. Cl. Sci. (5)*, 65(6) :274–292.

[Denuit and Van Keilegom, 2006] Denuit, M. P. O. and Van Keilegom, I. (2006). Bivariate archimedean copula models for censored data in non-life insurance. *Journal of Actuarial Practice*, 13 :5–32.

[Efron, 1981] Efron, B. (1981). Censored data and the bootstrap. *J. Amer. Statist. Assoc.*, 76(374) :312–319.

[Einmahl and Mason, 2005] Einmahl, U. and Mason, D. M. (2005). Uniform in bandwidth consistency of kernel-type function estimators. *Ann. Statist.*, 33(3) :1380–1403.

[Embrechts et al., 2003] Embrechts, P., Lindskog, F., and McNeil, A. J. (2003). Modelling dependence with copulas and applications to risk management. In Rachev, S., editor, *Handbook of Heavy Tailed Distributions in Finance*, pages 329–384. North-Holland.

[Emura and Chen, 2014] Emura, T. and Chen, Y.-H. (2014). Gene selection for survival data under dependent censoring : A copula-based approach. *Statistical Methods in Medical Research*.

[Fan et al., 2000] Fan, J., Hsu, L., and Prentice, R. L. (2000). Dependence estimation over a finite bivariate failure time region. *Lifetime Data Anal.*, 6(4) :343–355.

[Fermanian, 2005] Fermanian, J.-D. (2005). Goodness-of-fit tests for copulas. *J. Multivariate Anal.*, 95(1) :119–152.

[Fermanian et al., 2004] Fermanian, J.-D., Radulović, D., and Wegkamp, M. (2004). Weak convergence of empirical copula processes. *Bernoulli*, 10(5) :847–860.

[Fermanian and Scaillet, 2005] Fermanian, J.-D. and Scaillet, O. (2005). Some statistical pitfalls in copula modelling for financial applications. In Klein, E. e., editor, *Capital Formation, Governance and Banking*, pages 59–74. New York : Nova Science Publishers.

[Fleming and Harrington, 1991] Fleming, T. R. and Harrington, D. P. (1991). *Counting processes and survival analysis*. Wiley Series in Probability and Mathematical Statistics : Applied Probability and Statistics. John Wiley & Sons Inc., New York.

[Frees and Valdez, 1998] Frees, E. and Valdez, E. A. (1998). Understanding relationships using copulas. *North American Actuarial Journal*, 2 :1–25.

[Frees et al., 1996] Frees, E. W., Carriere, J. F., and Valdez, E. A. (1996). Annuity valuation with dependent mortality. *Journal of Risk and Insurance*, 63(2) :229–261.

[Gaenssler and Stute, 1987] Gaenssler, P. and Stute, W. (1987). *Seminar on empirical processes*, volume 9 of *DMV Seminar*. Birkhäuser Verlag, Basel.

[Gannoun et al., 2007] Gannoun, A., Saracco, J., and Yu, K. (2007). Comparison of kernel estimators of conditional distribution function and quantile regression under censoring. *Stat. Model.*, 7(4) :329–344.

[Gannoun et al., 2005] Gannoun, A., Saracco, J., Yuan, A., and Bonney, G. E. (2005). Non-parametric quantile regression with censored data. *Scand. J. Statist.*, 32(4) :527–550.

[Genest et al., 1995] Genest, C., Ghoudi, K., and Rivest, L.-P. (1995). A semiparametric estimation procedure of dependence parameters in multivariate families of distributions. *Biometrika*, 82(3) :543–552.

[Genest et al., 2006] Genest, C., Quessy, J.-F., and Rémillard, B. (2006). Goodness-of-fit procedures for copula models based on the probability integral transformation. *Scand. J. Statist.*, 33(2) :337–366.

[Genest and Rémillard, 2008] Genest, C. and Rémillard, B. (2008). Validity of the parametric bootstrap for goodness-of-fit testing in semiparametric models. *Ann. Inst. Henri Poincaré Probab. Stat.*, 44(6) :1096–1127.

[Genest et al., 2009] Genest, C., Rémillard, B., and Beaudoin, D. (2009). Goodness-of-fit tests for copulas : a review and a power study. *Insurance Math. Econom.*, 44(2) :199–213.

[Genest and Rivest, 1993] Genest, C. and Rivest, L.-P. (1993). Statistical inference procedures for bivariate Archimedean copulas. *J. Amer. Statist. Assoc.*, 88(423) :1034–1043.

[Gersho and Gray, 1991] Gersho, A. and Gray, R. M. (1991). *Vector Quantization and Signal Compression*. Kluwer Academic Publishers, Norwell, MA, USA.

[Gijbels and Veraverbeke, 1991] Gijbels, I. and Veraverbeke, N. (1991). Almost sure asymptotic representation for a class of functionals of the Kaplan-Meier estimator. *Ann. Statist.*, 19(3) :1457–1470.

[Gill, 1983] Gill, R. (1983). Large sample behaviour of the product-limit estimator on the whole line. *Ann. Statist.*, 11(1) :49–58.

[Gill, 1980] Gill, R. D. (1980). *Censoring and stochastic integrals*, volume 124 of *Mathematical Centre Tracts*. Mathematisch Centrum, Amsterdam.

[Gill et al., 1995] Gill, R. D., van der Laan, M. J., and Wellner, J. A. (1995). Inefficient estimators of the bivariate survival function for three models. *Ann. Inst. H. Poincare Probab. Statist.*, 31 :547–597.

[Graf and Luschgy, 1994] Graf, S. and Luschgy, H. (1994). *Consistent Estimation in the Quantization Problem for Random Vectors*. Angewandte Mathematik und Informatik. Univ.

[Graf and Luschgy, 2000] Graf, S. and Luschgy, H. (2000). *Foundations of quantization for probability distributions*. Lecture notes in mathematics. Springer-Verlag New York, Inc., Secaucus, NJ, USA.

[Grambsch et al., 1989] Grambsch, P. M., Dickson, E. R., Wiesner, R. H., and Langworthy, A. (1989). Application of the mayo primary biliary cirrhosis survival model to mayo liver transplant patients. *Mayo Clinic Proceedings*, 64(6) :699–704.

[Gribkova and Lopez, 2013] Gribkova, S. and Lopez, O. (2013). Nonparametric copula estimation under bivariate censoring.

[Gribkova et al., 2013] Gribkova, S., Lopez, O., and Saint-Pierre, P. (2013). A simplified model for studying bivariate mortality under right-censoring. *J. Multivariate Anal.*, 115 :181–192.

[Hanley and Parnes, 1983] Hanley, J. A. and Parnes, M. N. (1983). Nonparametric estimation of a multivariate distribution in the presence of censoring. *Biometrics*, 39(1) :129–139.

[Heuchenne, 2008] Heuchenne, C. (2008). Strong uniform consistency results of the weighted average of conditional artificial data points. *J. Statist. Plann. Inference*, 138(5) :1496–1515.

[Hougaard, 2000] Hougaard, P. (2000). *Analysis of multivariate survival data*. Statistics for Biology and Health. Springer-Verlag, New York.

[Hougaard et al., 1992] Hougaard, P., Harvald, B., and Holm, N. V. (1992). Measuring the similarities between the lifetimes of adult danish twins born between 1881–1930. *Journal of the American Statistical Association*, 87(417) :17–24.

[Hubert and Arabie, 1985] Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of Classification*, 2 :193–218.

[Jagger and Sutton, 1991] Jagger, C. and Sutton, C. J. (1991). Death after marital bereavement–is the risk increased ? *Stat Med*, 10(3) :395–404.

[Joe, 1997] Joe, H. (1997). *Multivariate models and dependence concepts*, volume 73 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, London.

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

[Klugman and Parsa, 1999] Klugman, S. A. and Parsa, R. (1999). Fitting bivariate loss distributions with copulas. *Insurance Math. Econom.*, 24(1-2) :139–148. 1st IME Conference (Amsterdam, 1997).

[Krzanowski and Lai, 1988] Krzanowski, W. J. and Lai, Y. T. (1988). A criterion for determining the number of groups in a data set using sum-of-squares clustering. *Biometrics*, 44(1) :23–34.

[Lin and Ying, 1993] Lin, D. Y. and Ying, Z. (1993). A simple nonparametric estimator of the bivariate survival function under univariate censoring. *Biometrika*, 80(3) :573–581.

[Linder, 2002] Linder, T. (2002). *Learning-theoretic methods in vector quantization*, volume 434 of *CISM Courses and Lectures*, pages 163–210. Springer-Verlag, Vienna.

[Linder et al., 1994] Linder, T., Lugosi, G., and Zeger, K. (1994). Rates of convergence in the source coding theorem, in empirical quantizer design, and in universal lossy source coding. *IEEE Trans. Inform. Theory*, 40(6) :1728–1740.

[Lloyd, 2006] Lloyd, S. (2006). Least squares quantization in pcm. *IEEE Trans. Inf. Theor.*, 28(2) :129–137.

[Lopez, 2009] Lopez, O. (2009). Single-index regression models with right-censored responses. *J. Statist. Plann. Inference*, 139(3) :1082–1097.

[Lopez, 2012] Lopez, O. (2012). A generalization of kaplan-meier estimator for analyzing bivariate mortality under right-censoring and left-truncation with applications to model-checking for survival copula models. *Preprint*.

[Lopez and Saint-Pierre, 2012] Lopez, O. and Saint-Pierre, P. (2012). Bivariate censored regression relying on a new estimator of the joint distribution function. *J. Statist. Plann. Inference*, 142(8) :2440–2453.

[Luciano et al., 2008] Luciano, E., Spreeuw, J., and Vigna, E. (2008). Modelling stochastic mortality for dependent lives. *Insurance Math. Econom.*, 43(2) :234–244.

[MacQueen, 1967] MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proc. Fifth Berkeley Sympos. Math. Statist. and Probability (Berkeley, Calif., 1965/66)*, pages Vol. I : Statistics, pp. 281–297. Univ. California Press, Berkeley, Calif.

[Milligan and Cooper, 1985] Milligan, G. and Cooper, M. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, 50(2) :159–179.

[Nelsen, 2006] Nelsen, R. B. (2006). *An introduction to copulas*. Springer Series in Statistics. Springer, New York, second edition.

[Omelka et al., 2009] Omelka, M., Gijbels, I., and Veraverbeke, N. (2009). Improved kernel estimation of copulas : weak convergence and goodness-of-fit testing. *Ann. Statist.*, 37(5B) :3023–3058.

[Parkes et al., 1969] Parkes, C. M., Benjamin, B., and Fitzgerald, R. G. (1969). Broken heart : a statistical study of increased mortality among widowers. *Br Med J*, 1(5646) :740–3.

[Patton, 2012] Patton, A. J. (2012). A review of copula models for economic time series. *J. Multivariate Anal.*, 110 :4–18.

[Pollard, 1982a] Pollard, D. (1982a). A central limit theorem for k-means clustering. *Ann. Probab.*, 10(4) :919–926.

[Pollard, 1982b] Pollard, D. (1982b). Quantization and the method of k-means. *IEEE Trans. Inform. Theory*, 28(2) :199–205.

[Prentice and Cai, 1992] Prentice, R. L. and Cai, J. (1992). Covariance and survivor function estimation using censored multivariate failure time data. *Biometrika*, 79(3) :495–512.

[Pruitt, 1991a] Pruitt, R. C. (1991a). On negative mass assigned by the bivariate Kaplan-Meier estimator. *Ann. Statist.*, 19(1) :443–453.

[Pruitt, 1991b] Pruitt, R. C. (1991b). Strong consistency of self-consistent estimators : general theory and an application to bivariate survival analysis. Technical Report 543, University of Minnesota, Minneapolis.

[Rachev and Rüschendorf, 1998] Rachev, S. T. and Rüschendorf, L. (1998). *Mass transportation problems. Vol. II.* Probability and its Applications (New York). Springer-Verlag, New York. Applications.

[Rand, 1971] Rand, W. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336) :846–850.

[Reid, 1981] Reid, N. (1981). Estimating the median survival time. *Biometrika*, 68(3) :601–608.

[Sánchez Sellero et al., 2005] Sánchez Sellero, C., González Manteiga, W., and Van Keilegom, I. (2005). Uniform representation of product-limit integrals with applications. *Scand. J. Statist.*, 32(4) :563–581.

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

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

[Shih and Louis, 1995] Shih, J. H. and Louis, T. A. (1995). Inferences on the association parameter in copula models for bivariate survival data. *Biometrics*, 51(4) :1384–1399.

[Sklar, 1959] Sklar, M. (1959). Fonctions de répartition à n dimensions et leurs marges. *Publ. Inst. Statist. Univ. Paris*, 8 :229–231.

[Spreeuw, 2006] Spreeuw, J. (2006). Types of dependence and time-dependent association between two lifetimes in single parameter copula models. *Scandinavian Actuarial Journal*, 2006(5) :286–309.

[Steinhaus, 1956] Steinhaus, H. (1956). Sur la division des corps matériels en parties. *Bull. Acad. Polon. Sci. Cl. III.*, 4 :801–804 (1957).

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

[Stute, 1995] Stute, W. (1995). The central limit theorem under random censorship. *Ann. Statist.*, 23(2) :422–439.

[Stute, 1996] Stute, W. (1996). Distributional convergence under random censorship when covariables are present. *Scand. J. Statist.*, 23(4) :461–471.

[Stute, 1999] Stute, W. (1999). Nonlinear censored regression. *Statist. Sinica*, 9(4) :1089–1102.

[Stute et al., 2000] Stute, W., González Manteiga, W., and Sánchez Sellero, C. (2000). Nonparametric model checks in censored regression. *Comm. Statist. Theory Methods*, 29(7) :1611–1629.

[Stute and Wang, 1993] Stute, W. and Wang, J.-L. (1993). The strong law under random censorship. *Ann. Statist.*, 21(3) :1591–1607.

[Talagrand, 1994] Talagrand, M. (1994). Sharper bounds for Gaussian and empirical processes. *Ann. Probab.*, 22(1) :28–76.

[Tsai et al., 1986] Tsai, W.-Y., Leurgans, S., and Crowley, J. (1986). Nonparametric estimation of a bivariate survival function in the presence of censoring. *Ann. Statist.*, 14(4) :1351–1365.

[Tsukahara, 2005] Tsukahara, H. (2005). Semiparametric estimation in copula models. *Canad. J. Statist.*, 33(3) :357–375.

[van der Laan, 1996] van der Laan, M. J. (1996). Efficient estimation in the bivariate censoring model and repairing NPMLE. *Ann. Statist.*, 24(2) :596–627.

[van der Vaart, 1998] van der Vaart, A. W. (1998). *Asymptotic statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge.

[van der Vaart and Wellner, 1996] van der Vaart, A. W. and Wellner, J. A. (1996). *Weak convergence and empirical processes*. Springer Series in Statistics. Springer-Verlag, New York. With applications to statistics.

[Varadarajan, 1958] Varadarajan, V. S. (1958). Weak convergence of measures on separable metric spaces. *Sankhyā*, 19 :15–22.

[Wang and Wells, 1997] Wang, W. and Wells, M. T. (1997). Nonparametric estimators of the bivariate survival function under simplified censoring conditions. *Biometrika*, 84(4) :863–880.

[Wang and Wells, 2000a] Wang, W. and Wells, M. T. (2000a). Estimation of Kendall's tau under censoring. *Statist. Sinica*, 10(4) :1199–1215.

[Wang and Wells, 2000b] Wang, W. and Wells, M. T. (2000b). Model selection and semiparametric inference for bivariate failure-time data. *J. Amer. Statist. Assoc.*, 95(449) :62–76. With a comment by Edsel A. Peña and a rejoinder by the authors.

[Youn and Shemyakin, 1999] Youn, H. and Shemyakin, A. (1999). Statistical aspects of joint life insurance pricing. *Proceedings of the business and statistics section of the American Statistical Association*, pages 34–38.

[Youn and Shemyakin, 2001] Youn, H. and Shemyakin, A. (2001). Pricing practices for joint last survivor insurance. *Actuarial Research Clearing House*.