PEOPL'S DEMOCRATIC REPUBLIC OF ALGERIA
MINISTRY OF HICHER EDUCATION AND SCIENTIFIC RESEARCH

MOHAMED KHIDER UNIVERSITY, BISKRA
FACULTY OF EXACT SCIENCE, SCIENCE OF NATURE AND LIFE
DEPARTMENT OF MATHEMATICS

PhD THESIS

PRESENTED TO ACHIEVE

The Doctorate Degree In Mathematics

OPTION : **STATISTICS**

BY :

IDIOU Nesrine

TITLE

MULTI-PARAMETRIC COPULA ESTIMATION BASED ON MOMENTS METHOD UNDER CENSORING

Publicly defended, In 23/03/2022, front of the jury members :

Pr.	**YAHIA Djoubrane**	**University of**	**Biskra**	**President**
Pr.	**BENATIA Fatah**	**University of**	**Biskra**	**Supervisor**
Pr.	**BRAHIMI Brahim**	**University of**	**Biskra**	**Co-supervisor**
Pr.	**DJEFFAL El Amir**	**University of**	**Batna2**	**Examiner**
Dr.	**SAYAH Abdallah**	**University of**	**Biskra**	**Examiner**

Presented by :

IDIOU Nesrine

Multi-Parametric Copula Estimation Based on Moments Method Under Censoring

ESTIMATION DES COPULES MULTI-PARAMÈTRIQUES PAR LA MÉTHODE DES MOMENTS EN PRÉSENCE DE CENSURE

PhD Thesis

To Achieve the Doctoral University Degree
Doctor of Mathematics
Doctoral Degree In Mathematics
Statistics Option

Submitted to

Mohamed Khider University of Biskra
Faculty of Exact Science, Science of Nature and Life
Department of Mathematics

Supervisor

Prof. Dr. BENATIA Fatah

Algeria, in 2022

I DEDICAte *this effort to,*

My Parents. No homage could ever match the love they infuse me with. May God bless them all with good health and full lifespans. . .

Those that I love a lot. Who have been a staunch supporter of me throughout all of my endeavors. My fiance Mohamed, my sisters, and brothers especially for you. . .

My entire family, friends, colleagues, and for everyone. Who has helped me in any way, no matter how small or large their contribution. . .

I SAY, *Thank you*

Acknowledgments

Words *cannot express my gratitude to them enough !! But, I would like express my sincere thanks to:*

God, who is always there for me...

My supervisor, Pr. Benatia Fatah, for his trust, orientation, and openness to my views while accompanying me throughout the academic course. I sincerely thanks him for his unwavering moral support, for his respectful treatment, and for everything he has done for me over my PhD path...

My co-supervisor, Pr. Brahimi Brahim for his help and the precious contribution he devoted to this work...

The honorable examination committee members for accepting the assessment and examination of this thesis, and even for the time given to read the thesis carefully...

Pr. Mesbah Mounir for his warm welcome in his laboratory (LPSM), at the Sorbonne University in Paris, and for his contribution to a research work that was published in a respected journal...

Finally, I would like express my gratitude to all the professors, colleagues and friends from the mathematics faculty of Constantine University and Biskra University, including all the members of the LMA laboratory, and for everyone who has supported, directly or indirectly this work...

.

Scientific Contributions

Articles

1. *Idiou, N., Benatia, F., Mounir, M. (March 1, 2021). COPULAS AND FRAILTY MODELS IN MULTIVARIATE SURVIVAL DATA. Journal of Biostatistics and Health Sciences . Published by ISTE Ltd (2021 ISTE OpenScience). London, UKopenscience.fr. Vol. 2, No 1, 13-39.*

2. *Idiou, N., Benatia, F., Brahimi, B. (Juin 13, 2021). A SEMI-PARAMETRIC ESTIMATION OF COPULA MODELS BASED ON MOMENTS METHOD UNDER RIGHT CENSORING. TWMS, Applied and Engineering Mathematics Journal. (Accepted and to be published).*

3. *Idiou, N., Benatia, F. (Juin 21, 2021). SURVIVAL COPULA PARAMETERS ESTIMATION FOR ARCHIMEDEAN FAMILY UNDER SINGLY CENSORING. Advances in Mathematics: Scientific Journal 10 (2021), no.1, 1–4.*

4. *Idiou, N., Benatia, F. (July 16, 2021). SIMULATION TECHNIQUES OF ARCHIMEDEAN COPULA ESTIMATORS: PARAMETRIC AND SEMI-PARAMETRIC APPROACHES. EJ-MATH, European Journal of Mathematics and Statistics, Vol 2, No.3.*

5. *Idiou.N, Benatia.F, Brahimi.B. (February19,2020). BIAS AND RMSE OF ARCHIMEDEAN COPULA USING MOMENTS AND L-MOMENTS METHODS. Published in IEEE Xplore Digital Library. 978-17281-2580-0/20/31.00@2020 IEEE.*

Conference paper and communication

1. *Idiou, N. and Benatia, F. COPULA SEMI-PARAMETRIC ESTIMATOR. (9-10 July 2020). The 7th International Conference on Computer Engineering and Mathematical Sciences, ICMSCE 2020, Langkawi, Malaysia.*

2. *Idiou, N., Benatia, F. (December 20-22, 2019). MONTE-CARLO SIMULATION STUDIES OF SEMI-PARAMETRIC ESTIMATORS USING PYTHON SOFTWARE. International workshop on machine learning and python, university of Monastir, Monastir, Tunisia.*

3. *Idiou.N, Benatia.F, Brahimi.B. (February 18-19, 2020). BIAS AND RMSE OF ARCHIMEDEAN COPULA USING MOMENTS AND L-MOMENTS METHODS. International Conference on Mathematics and Information Technology, Adrar, Algeria. (Published in IEEEXplore).*

4. *Idiou, N., Benatia, F. (November 24-26, 2019). BIVARIATE COPULAS STATISTICS AND SEMI-PARAMETRIC ESTIMATION. International Conference on Stochastic and Statistical Modeling (MMS'2019). 4th edition, Houari Boumedien University USTHB, Algiers, Algeria.*

5. *Idiou, N., Benatia, F. (October 28-29, 2019). PERFORMANCE COMPARISON FOR SEMI-PARAMETRIC ESTIMATOR OF BIVARIATE COPULAS. International Conference on Financial Mathematics: Tools and Applications (MFOA'2019), University of Bedjaia, Bedjaia, Algeria.*

6. *Idiou, N., Benatia, F. (June 2, 2021). ON THE COPULAS ESTIMATION AND THEIR APPLICATIONS. 2nd National Seminaire of Mathematics, Mentouri Constantine 1 University, Algeria.*

7. *Idiou, N., Benatia, F. (June 27, 2019). COPULAS ESTIMATION AND APPLICATION OF NUMERICAL METHOD ON REAL DATA. National Conference of Applied Mathematics (JNMA'19), University of Larbi Ben M'hidi, Oum ElBouagui, Algeria.*

8. *Idiou, N. (June 25-27, 2019). DATA SCIENCE TRAINING FOR THE BENEFIT OF SOCIAL AND HEALTH SCIENCE STUDIES,University of Salah Boubnider, Constantine3, Algeria.*

9. *Idiou, N., Benatia, F. (April 28, 2019). A PARAMETRIC AND A SEMI-PARAMETRIC ESTIMATION OF MULTI PARAMETERS COPULAS. National Conference of Applied Mathematics (JMA'19), University of Abdelhafid Boussouf, Mila, Algeria.*

10. *Idiou, N., Benatia, F. (April 10, 2019). ESTIMATION OF MULTI-PARAMETERS COPULA. National Conference of Applied Mathematics (JMA'19), University of Ziane Achour, Djelfa, Algeria.*

Contents

List of Figures

List of Tables

Introduction

Measurement of the dependence between two or more random variables is a widely used statistical approach. Several more different measurements of random variable dependency have been considered, including the Pearson correlation coefficient, Kendall's tau, and Spearman's rho. Although these measures are easier to evaluate, they are not able to detect all types of independence, so another solution to this problem needed to be found. This issue was overcome by creating a copula function, which has the advantage of completely modeling dependence between variables.

The term Copula comes from the Latin word " *copũlae* ", which means a bond, link, or union, their use in statistics is a relatively new phenomenon that dates up to the end of the 1950. The concept of copula functions also had other names different writers in the 1970s, they are called "uniform representation" by Kimeldorf and Sampson (1975) [53], "dependency function" by Deheuvels (1979) [19], or even "the standard form" by Cook and Johnson (1981) [17].

In the mathematical sense, a copula is a function that serves as the primary link between the multivariate distribution function and its univariate margins. Regardless of the shape of the margins, the action of the copula is to represent the characteristics of dependence that are associated with each of the random variables. In 1959, the term copula was first used in the theory of multidimensional distributions, thanks to Sklar (1959) [84], who has shown in his theorem (2.1.3), under certain conditions that there is a unique copula function C, which is given by:

$$F\left(x_1, ..., x_n\right) = C(F_1\left(x_1\right), ..., F_n\left(x_n\right))$$

where F is the joint distribution of $X = \left(X_1, ..., X_n\right)$ and $F_1\left(x_1\right), ..., F_n\left(x_n\right)$ are the margins, these margins could be of different distributions.

For a variety of reasons, the copula was chosen to model dependence rather than the correlation coefficient, where we discovered that the latter has various limitations including as, if for example the two-order moments of random variables are not completed, the correlation coefficient is not defined. Also, for heavy-tailed distributions with infinite variances, this is not an adequate measure of dependence. More plainly and broadly correlation is a measure of dependence that does not provide us with all of the information we need about the structure of dependency.

In copula approaches viewpoint and concerning their estimate, if the margins $F_1, ..., F_n$ are known, then we bring to classic statistical inference methods. But, because the margins are generally unknown, mainly two approaches can be adopted for its estimations parametric and nonparametric. Nevertheless, in the first approach we estimate the margins parametrically, i.e., the resulting estimate of C will be entirely parametric, we

suppose that the marginal belongs to a family indexed by a parameter, so to estimate the margins, it suffices to estimate their parameter. This type of parameter estimation was established in the literature, namely the concordance approaches, also known as tau-inversion and rho-inversion, which are based on Kendall's tau and Spearman's rho rank correlation coefficients respectively.

Although in the other approach the margins will estimate non-parametrically, i.e., we do not assume that the margins belong to any family, then it will be a semi-parametric copula estimation. We quote the semi-parametric estimation method for copula based on this approach and we mention the methods of moments, L-moments and T.Lmoments, recently developed by Brahimi et al (2012) [7], which proposed an estimator of a parametric copula by the method of moments. As a logical continuation of the method of moments Benatia et al (2011) [6], have proposed also a new estimator, using the same classical procedure but for L-moments. After truncation of extreme values, using the T.L-moments introduced by Elamir and Seheult (2003), Chine and Benatia (2017) [15], have proposed a new unbiased and asymptotically normal estimator, whose advantage is to be valid even for non-existent or infinite moments distributions (Cauchy, beta ...).

In statistics, the study of the copula and its applications is a relatively modern phenomenon. Copulas are mathematical objects that fully capture the structure dependence between random variables and continue to gain popularity in several fields of mathematics, such as finance, actuarial science, hydraulics, biology, insurance, and reliability theory. Currently, they have become a necessary tool for market and credit models, risk aggregation, portfolio selection, etc. Nelsen (2006) [67], and others describe a variety of copula functions that can be used to fit a wide variety of dependence types. Several multivariate survival models taking into account, the dependence between random variables are based on the notion of copulas, because of the advantages of using this function to model a dependency structure between multivariate variables. We can claim that among these advantages that the copula allow the construction of multi-dimensional distribution models, they also model dependency structure properties, and it is able to measure the dependence for heavy-tailed distributions.

In the survival analysis area, a multivariate distribution can be constructed through the use of copulas in a survival setting. Survival analysis is a branch of statistics that attempts to model the time T before an event occurs. Since its origins in the 17^{th} century, survival modeling have progressed. This modeling can be done with data where for all individuals the survival is known. In this case, we are talking about complete data. However, due to the end of the study, withdrawal from the study, or loss of follow-up, only a part of the individuals are known to have survived. In this case, we are dealing with incomplete data and unknown survival observations are said to be censored. The probability that an individual is alive or unscathed beyond time t is given by the survival function, and when several events are involved simultaneously, we speak of multivariate survival. The modeling of bivariate or multivariate data in survival anal-

ysis has been discussed by several authors. Many approaches have been introduced for this modelisation, including Archimedean copula models ([4], [16], [46], [47], [62], [85], [95]), for reasons of dependency modeling and its approaches, even their application particularly in actuarial science and financial risk management. The following is the key reason for concentrating on this class:

1. they are easy to construct.
2. have interesting properties that further facilitate the modeling of dependency structures.
3. a large variety of copula families belonging to this class.

This family of Copula is often characterized by a generator, which is a function, thus reducing the search for a large dimensional distribution function. Archimedean copula models arise naturally from bivariate frailty models [71], in which the two failure times have given unobserved frailty W and each follows the proportional hazards model in W. However, in this aspect, an Archimedean copula is presented by:

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

where φ is a continuous, convex and decreasing function called the generator of C, defined on $I = [0,1] \to [0,\infty]$ and verifies $\varphi(1) = 0$.

The primary aim of this thesis is to extend the Copula theory results via semi-parametric estimating methods, which are presented in two interesting parts. Specifically, we propose an alternative estimation method of a survival copula $\tilde{C}$, based on a semi-parametric estimation of the classical moments method due to its simple mathematical form, given $(T_1; T_2)$ as singly or doubly right-censored. The asymptotic normality of the empirical survival copula was established for the two cases of censoring. The dependence structure between the bivariate survival times was modeled under the assumption that the underlying copula is Archimedean. A simulation study follows, which sheds light on the behavior of the process estimation method. The methodology of the proposed estimator is also illustrated by using lifetime data from the Diabetic Retinopathy Study, where its efficiency and robustness are observed.

In another part of the thesis, we are interested by Copula modeling and its applications in the analysis of multivariate survival data. We have implemented the frailty model for bivariate survival data by considering Archimedean copulas. Our main idea in this chapter focused on introducing the dependence between the survival times $T_1, ..., T_d$, using an unobserved random variable W, called frailty model with variable latent. The frailty variables considered here are latent variables that are not observed, are nevertheless one-dimensional. In the example presented, this variable characterized the effect of the individual on the recurrence time. Then we looked at Clayton-Oakes copulas in particular, and even the model with gamma-type frailty. The applications for health-related survival data were next examined.

So, this thesis is a blend of two statistical branches: the survival analysis and the Copula theory. We firstly provide a summary of the various definitions and fundamental properties of these two domains of statistical, and then we present our result that we talked about just previously. This thesis is organized as follows:

Chapter 1 : We present some basic notions, we start with foundations definitions like the distribution function, the empirical distribution function, the survival function... etc. Thus, we present the basic notion of order statistics, H-volume, and 2-increasing functions, in order that we can leverage them in the next.

Chapter 2 : The second chapter is mainly devoted to the design of copulas and their properties. We introduce the notion of the copula, a bivariate case, and also the relationship between copulas and pairs of random variables. By the way, we carried out a synthesis of the main copulas properties, the most important of which is given by Sklar's theorem, the founder of copulas. Following that, we devote a section for different types of copulas, namely the usual copulas, the family of Archimedean copulas, and the copulas of extreme values. Then, we present a multivariate generalization of the copula notion for all bivariate cases.

Chapter 3 : In this chapter, we give some relations existing between some dependency measures and copula seen in Chapter 2 as well as an order of concordance. We explore ways in which copulas can be used in the study of dependence or association between random variables.

Chapter 4 : This chapter is devoted to the basics of survival analysis, the semi-parametric and non-parametric estimation methods. In section (4.1), we start with a few reminders on the basic concepts of survival time notion. We present the following two cases of incomplete data: censored and truncated in section (4.2). In section (4.3), We present some approaches for estimating copulas models from a sample, including semi-parametric and parametric estimation methods. Whose following, we have presented two non-parametric estimation methods for a right-censored model (the most famous non-parametric estimators), known by the Kaplan-Meier estimator for the survival function [51], and the Kernel estimator for the density function [25]. At the end of this chapter, we introduce the non-parametric estimate for a mixed-censored model, known by the Patilea and Rolin estimator [75].

Chapter 5 : We consider a general framework of right-censoring, which includes all the concepts treated in the preceding chapters. In this chapter, we have introduced a new copula estimator for censored bivariate data based on the classical estimation method of moments, presented in a semi-parametric estimation framework. This chapter is divided into two parts the first focuses on the estimation of this new estimator when the data are doubly right-censoring, i.e. the two variables are right-censored at the same time. In the second part, we present this estimator and all results obtained in part one, when only one of two variables is right-censored as singly right-censoring. This chapter is structured as follows:

In part one we have presented the theoretical results of the estimator proposed, general formulas were proved with analytical forms of the obtained estimators. Taking into account Lopez and Saint Pierre's (2012) [72],

Gribkova and Lopez's (2015) [39], results, the asymptotic normality of the empirical survival copula was established. A semiparametric estimation method based on the classical moments method illustrated a conditional distribution on $\tilde{C}$. The dependence structure between the bivariate survival times was modeled under the assumption that the underlying copula is Archimedean. Accounting for various censoring patterns (singly or doubly censored), a simulation study was performed to enlighten the behavior of the procedure estimation method, showing the efficiency and robustness of the new estimator proposed.

As a logical continuation of results established by N.IDIOU et al (2021) [68], presented in part one of the chapters, a particular case of right-censoring has well detailed in part two of the chapter, as well as the empirical survival copula has also been evaluated in this case of singly-censored data. As an application, two Archimedean Copula models have been chosen to illustrate our theoretical results. A simulation study follows, which sheds light on the behavior of the process estimation method showing that the proposed estimator performs well in terms of relative bias and RMSE. The methodology of the proposed estimator is also illustrated by using lifetime data from the Diabetic Retinopathy study, where its efficiency and robustness are observed.

Chapter 6 : In this chapter, we are interested by Copula modeling and its applications in the analysis of multivariate survival data. We have used the frailty model for bivariate survival data by considering Archimedean copulas. Our main idea is this chapter focused on introducing the dependence between the survival times $T_1, ..., T_d$, using an unobserved random variable W, called frailty model with variable latent. We then focused on the particular cases of Clayton-Oakes copulas and the model with frailty gamma-type. For each of these two models, the copulas used for the bivariate survival functions are the same. However, the marginal survival functions are modeled in different ways. The variables of frailty, considered here, are latent, not observed, but one-dimensional. In the example presented, this variable characterized the effect of the individual on recovery time currency. These individuals could come from several hospitals. The differential effect, not observed, of these centers would then be a latent variable. This chapter ended with an application presented on bivariate survival data in biostatistics fields, analyzed by the Copula procedure of the SAS software.

Preliminary 1

Sommaire

In In this chapter we define some of the basic conceptions, in order that we can leverage them in the next.

1.1 Foundations definition

Definition 1.1.1 (**The distribution function**) *The distribution function of a real random variable (r.v) X, is a function that is generally noted by F_X, defined from $\mathbb{R}$ in $[0;1]$, where x in $\mathbb{R}$ associates:*

$$F_X(x) = P(X \leq x) = P(]-\infty; x]) \tag{1.1.1}$$

F_X is known as an increasing function, continuous to the right with a limit to the left at any point in the sense that:

$$\lim_{x \to -\infty} F(x) = 0 \quad and \quad \lim_{x \to \infty} F(x) = 1$$

Definition 1.1.2 (**The empirical distribution function**) *Let the sample $X_1, ..., X_n$ of a positive r.v X, for $n \geq 1$ size, with the distribution function F. The empirical distribution function F_n is defined by:*

$$F_n = \frac{1}{n} \sum_{i=1}^{n} I_{\{X_i \leq x\}}, \ \ \forall\, x \geq 0 \tag{1.1.2}$$

where $I_{\{B\}}$ is the indicator function of the set B. So we can conclude that F_n is the proportion of the n variables which are less than or equal to x.

Definition 1.1.3 (**The density function**) *A probability distribution has a density f, if f is a function defined on $\mathbb{R}$, positive and Lebesgue-integrable, such that the probability of the interval $[a, b]$, is given by:*

$$\int_a^b f(x)\, dx,$$

A probability density is also allowed to represent a distribution function F_X in the form of integrals as:

$$F(x) = \int_{-\infty}^{x} f(t)\, dt$$

Definition 1.1.4 (**The survival function**) *The survival function called also the survival time, often noted S or $\bar{F}$, of a positive and continuous r.v X, is given generally by:*

$$\begin{aligned} S(x) = \bar{F}(x) &= P(X > x) \\ &= 1 - F(x), \ \ x \geq 0 \end{aligned} \tag{1.1.3}$$

In terms of distribution, we have:

$$S(x) = \int_{x}^{+\infty} f(t)\, dt$$

we can also write $S'(x) = -F'(x) = -f(x)$, where S' is the derivative of S. We are also aware of:

- *$S(x)$ is a non-increasing function.*
- $\lim_{t \to 0} S(x) = 1$ *and* $\lim_{t \to \infty} S(x) = 0$.

Definition 1.1.5 (**The empirical survival function**) *Let the sample $X_1, ..., X_n$ of a positive r.v X and of $n \geq 1$ size, where S its a survival function. The empirical survival function noted by S_n, is given by:*

$$S_n = 1 - F_n = \frac{1}{n} \sum_{i=1}^{n} I_{\{X_i > x\}}, \quad \forall\, x \geq 0 \tag{1.1.4}$$

So S_n is the proportion of observations that exceeds x.

Definition 1.1.6 (**The quantile function**) *Let X be a r.v defined in $\mathbb{R}$, and F_X its distribution function. We call the quantile function of X the function, denoted by Q_X, from $]0;1[$ in $\mathbb{R}$, who has $0 < u < 1$ associate:*

$$Q_X(u) = F_X^{-1}(x) = \inf\{x : F_X(x) \geq u\},$$

where F_X is a continuous and monotonous function and F_X^{-1} represents its generalized inverse.

As is typical, we can say that a quantile function of a r.v is the inverse of its distribution function F_X. When this distribution function is strictly increasing, its inverse is defined without ambiguity. By convention, $Q_X(0)$ is the smallest of the possible values for X and $Q_X(1)$ is the largest.

Definition 1.1.7 (**The empirical quantile function**) *Let the sample $X_1, ..., X_n$, of an independent and identically distributed r.v. The empirical quantile function $Q_n(u)$ of a sample $X_1, ..., X_n$ are defined for all $u \in]0;1[$ by:*

$$\begin{aligned} Q_n(u) &= \inf\{x : F_n(x) \geq u\} \\ &= \inf\left\{x : \frac{1}{n} \sum_{i=1}^{n} I_{\{X_i \leq x\}} \geq u\right\}, \end{aligned}$$

where F_n is the empirical distribution function defined by (1.1.2).

1.2 Basic notions

1.2.1 Order statistics

Definition 1.2.1 (**Order statistics**) *Let the sample $X_1, ..., X_n$ of an independent and identically distributed r.v of the same distribution function F. The order statistics of $X_1, ..., X_n$ is the increasing rearrangement of the previous sample, noted:*

$$X_{(1,n)} \leq ... \leq X_{(n,n)}.$$

In particular, the random variable $X_{(i,n)}$ is the i^{th} order statistics for $1 \leq i \leq n$.

Definition 1.2.2 (**Extreme order statistics**) *The extreme order statistics noted by $X_{(1,n)}$ and $X_{(n,n)}$ are defined respectively by:*

$$X_{(1,n)} = \min X_{(i)} \quad \text{and} \quad X_{(n,n)} = \max X_{(i)}$$

Definition 1.2.3 (**Extreme order statistics distributions**) *The distributions $F_{X_{(1,n)}}$ and $F_{X_{(n,n)}}$ of the extreme order statistics $X_{(1,n)}$ and $X_{(n,n)}$ are respectively defined by:*

$$\begin{cases} F_{X_{(1,n)}}(x) = 1 - [1 - F(x)]^n \\ F_{X_{(n,n)}}(x) = [F(x)]^n \end{cases}$$

1.2.2 H-volume notion, 2-increasing functions

Consider $\overline{\mathbb{R}}$ the extension of $\mathbb{R}$ in $[-\infty;+\infty]$. A rectangle in $\overline{\mathbb{R}}^2$ is a Cartesian product B of two closed intervals $B = [x_1, x_2] \times [y_1, y_2]$. The vertices of rectangle B are the points $(x_1, y_1), (x_1, y_2), (x_2, y_1)$, and (x_2, y_2). A real function H with two variables is a function of domain $DomH$, a subset of $\overline{\mathbb{R}}^2$ including all images of rank $RanH$ is a subset of $\mathbb{R}$.

Definition 1.2.4 *Let S_1 and S_2 two non-empty subsets of $\mathbb{R}$ and consider H a two-dimensional function such that $DomH = S_1 \times S_2$. The H-volume of B is given by:*

$$V_H(B) = H(x_2, y_2) - H(x_2, y_1) - H(x_1, y_2) + H(x_1, y_1)$$

Notice that if we define the first order difference of H on the rectangle B by:

$$\begin{aligned} \Delta_{x_1}^{x_2} H(x, y) &= H(x_2, y) - H(x_2, y) \\ \text{and } \Delta_{y_1}^{y_2} H(x, y) &= H(x, y_2) - H(x, y_1) \end{aligned}$$

Then the H-volume of rectangle B is the second-order difference of H in B:

$$V_H(B) = \Delta_{y_1}^{y_2} \Delta_{x_1}^{x_2} H(x, y)$$

Definition 1.2.5 *A two-dimensional real function H is said to be 2-increasing if $V_H(B) \geq 0$ for any rectangle B whose vertices are included in $DomH$. When H is 2-increasing, the H-volume of the rectangle B is sometimes presented as a measure of B.*

The next lemmas will be very useful for establishing the continuity of sub-copulas and copulas in the subsequent. The first comes as a direct result of the definitions (1.2.4) and (1.2.5).

Lemma 1.2.1 *Let A and B two non-empty subsets of $\overline{\mathbb{R}}$ and H a 2-increasing function of domain $DomH = A \times B$.*
Let x_1, x_2 of S_1 where $x_1 \leq x_2$ and y_1, y_2 of S_2 where: $y_1 \leq y_2$, then the function $t \to H(t, y_2) - H(t, y_1)$ is non-decreasing on A and the function $t \to H(x_2, t) - H(x_1, t)$ is non-decreasing on B.

Definition 1.2.6 *Assume that A and B be 2-non-empty subsets of $\overline{\mathbb{R}}$ and H be a real function of 2-variable 2-increasing such that $DomH = A \times B$. Let a be the smallest element of A and b be the smallest element of B. We say that H is "grounded" if:*

$$H(x, b) = 0 = H(a, y) \; \forall \, (x, y) \;\; in \; A \times B$$

Lemma 1.2.2 *Let A and B two non-empty subsets of $\overline{\mathbb{R}}$ and H a 2-increasing function grounded of a domain $A \times B$ then H is non-decreasing with respect to each of his arguments.*

Proof. Let a and b two minimums of A and B respectively, taken $x_1 = a$ and $y_1 = b$. Assuming that A admits a maximum b_1 and a maximum b_2 for S_2. The function H of $S_1 \times B$ is said to admit the two marginal functions F and G in $\mathbb{R}$, given by:

$$\begin{aligned} DomF &= A \text{ and } F(x) = H(x, b_2) \text{ for all } x \text{ in } A \\ DomG &= B \text{ and } G(y) = H(b_1, y) \text{ for all } y \text{ in } B \end{aligned}$$

□

We ended this chapter with the following lemma concerning 2-increasing functions with marginals.

Lemma 1.2.3 *Assuming that A and B are two non-empty subsets of $\overline{\mathbb{R}}$ and H a function 2-inccreasing, with marginals domain $A \times B$. Let $(x_1, y_1)\,;(x_2, y_2)$ any two points of $A \times B$ then:*

$$|H(x_2, y_2) - H(x_1, y_1)| \leq |F(x_2) - F(x_1)| + |G(y_2) - G(y_1)|$$

Proof. From the triangular inequality we have:

$$|H(x_2, y_2) - H(x_1, y_1)| \leq |H(x_2, y_2) - H(x_1, y_2)| + |H(x_1, y_2) - H(x_1, y_1)|\,.$$

Assume that $x_1 \leq x_2$, because H is a 2-increasing function admits marginals then the lemmas (1.2.1) and (1.2.2) imply:

$$0 \leq H(x_2, y_2) - H(x_1, y_2) \leq F(x_2) - F(x_1)\,,$$

a similar inequality applies when $x_2 \leq x_1$, it follows therefore for all x_1, x_2 of A

$$|H(x_2, y_2) - H(x_1, y_2)| \leq |F(x_2) - F(x_1)|$$

The same for all y_1, y_2 of B and we have:

$$|H(x_1, y_2) - H(x_1, y_1)| \leq |G(y_2) - G(y_1)|$$

□

Copula conseptions 2

Sommaire

The most important aspects of copula theory, will indeed be discussed in this chapter. This is primarily a presumption for those unfamiliar with the copula function to read this thesis. The main definition and characteristics of this concept are presented, followed by a description of the fundamental basic properties and an outline of some demonstrations to help understanding. To make things easier, we start with a brief reminder of bivariate copulas. Let's say $(X_1, ..., X_d)$ is a random vector with $\mathbb{R}^d$ values, of joint distribution function F and marginals F_i for $i = 1, ..., d$.

Sklar in (1959) [84], shows that, under certain conditions, there exists a unique function denoted C (for copula), such that:

$$H(x_1, ..., x_d) = C(F_1(x_1), ..., F_d(x_d))$$

This representation translates the intrinsic relationships between the components by reducing normalizing all variables to uniform variables. This chapter introduces some concepts that will be necessary when discussing Copula. Then, focusing on the most important definitions, we provide a brief summary of the main properties of copulas.

2.1 Bivariate Copula

Consider a couple of random variables (X, Y) of marginal distribution functions $F_1(x) = P(X \leq x)$ and $F_2(y) = P(Y \leq y)$ respectively and a joint distribution $F(x, y) = P(X \leq x, Y \leq y)$. For each couple (x, y) we can associate three numbers $F_1(x)$, $F_2(y)$ and $F(x, y)$ belonging to the interval $[0; 1]$. In other words, for each pair of real numbers (x, y) corresponds a point $(F_1(x), F_2(y))$ in the unit square $[0; 1] \times [0; 1]$ and in turn, this point corresponds to a real number $F(x, y)$ belonging to the interval $[0; 1]$. We will then show that this correspondence, which assigns the distribution's value to each pair of marginal distribution functions, is in reality, a function (called a copula). In all that follows $\mathbb{I}$ will denote the unit interval $\mathbb{I} = [0; 1]$.

2.1.1 Copula and Sub-Copula

Before moving on to the copulas, let's define the sub-copulas first.

Definition 2.1.1 *A two-dimensional sub-copula, already identified as a 2-sub-copula or obviously a sub-copula C', is a function with the following properties:*

1. *$DomC' = S_1 \times S_2$, where S_1 and S_2 are subsets of $\mathbb{I}$ containing 0 and 1,*
2. *C' is grounded and 2-increasing,*
3. *For every u in S_1 and every v in S_2 we have :*

$$C'(u, 1) = u \text{ and } C'(1, v) = v.$$

Noted that for every (u, v) in $DomC'$, $0 \leq C'(u, v) \leq 1$, so that $RanC'$ is also a subset of $\mathbb{I}$.

Definition 2.1.2 *A two-dimensional copula, also known as a 2-copula or simply a bivariate copula C, is a function of $\mathbb{I}^2$ in $\mathbb{I}$ having the following properties:*

1. *For all $(u, v) \in \mathbb{I}^2$, the Copula C is grounded, i.e.*

$$C(u, 0) = C(0, v) = 0$$

2. *The margins are uniform, i.e.*

$$C(u, 1) = u \text{ and } C(1, v) = v, \quad \forall u, v \in \mathbb{I}. \tag{2.1.1}$$

3. *C is 2-increasing, in other terms: $\forall (u_1, u_2, v_1, v_2) \in \mathbb{I}^4$ such that $u_1 \leq u_2$ and $v_1 \leq v_2$, we have:*

$$C(u_2, v_2) - C(u_2, v_1) - C(u_1, v_2) + C(u_1, v_1) \geq 0 \tag{2.1.2}$$

Thus, (2.1.2) gives an "inclusion-exclusion" type formula for the number assigned by C to each rectangle $[u_1, u_2] \times [v_1, v_2]$ in $\mathbb{I}^2$ and states that the number so assigned must be nonnegative, where $C(u, v) = V_C([u_1, u_2] \times [v_1, v_2])$. By the way, any copula is a distribution on $\mathbb{I}^2$ with uniform marginal distributions on $\mathbb{I}$. The definition of copulas is a special case of sub-copulas, and they can be stated as follows:

Definition 2.1.3 *(**Nelsen 2006**[67]) A two-dimensional copula (or a a bivariate copula) is a two-dimensional sub-copula whose support is* $\mathbb{I}^2$.

The following theorem discusses one of the most important properties of sub-copulas.

Theorem 2.1.1 *Let* C' *a sub-copula. Then for every* (u, v) *in* $DomC'$ *we have:*

$$\max(u + v - 1, 0) \leq C'(u, v) \leq \min(u, v).$$

Proof. Let (u, v) an arbitrary point of $DomC'$.
Because $C'(u, v) \leq C'(u, 1) = u$ and $C'(u, v) \leq C'(1, v) = v$, then

$$C'(u, v) \leq \min(u, v)$$

In addition $V_{C'}([u, 1] \times [v, 1]) \geq 0$ indicates that $C'(u, v) \geq u + v - 1$, as well as $C'(u, v) \geq 0$ given:

$$C'(u, v) \geq \max(u + v - 1, 0).$$

Because each copula is a sub-copula, the previous theorem's inequality holds for the copulas as well. The bounds of this inequality are also copulas, noted the Fréchet-Hoeffding bounds as shown in Nelsen (2006) [67], which are usually mentioned by:

$$M(u, v) = \min(u, v),$$

and

$$W(u, v) = \max(u + v - 1, 0),$$

thus for each copula C and all (u, v) in $\mathbb{I}^2$ we have:

$$W(u, v) \leq C(u, v) \leq M(u, v). \tag{2.1.3}$$

This inequality is the copula version of the Fréchet-Hoeffding bounding inequality. We call M and W the upper and lower bounds of Fréchet-Hoeffding, respectively. Another important copula is the product copula $\Pi(u, v) = u.v$, which characterizes the case of independence. The following theorem establishes the continuity of sub-copulas. □

Theorem 2.1.2 ***[Uniform continuity]*** *Let a sub-copula* C', *for all* $(u_1, u_2), (v_1, v_2)$ *in* $DomC'$ *we have:*

$$\left|C'(u_2, v_2) - C'(u_1, v_1)\right| \leq |u_2 - u_1| + |v_2 - v_1|$$

As result, C' *is uniformly continuous in its domain.*

Definition 2.1.4 *Let a copula* C *and consider* a *as an arbitrary number* $\mathbb{I}$. *The horizontal section of* C *at* a *is the function from* $\mathbb{I}$ *in* $\mathbb{I}$ *given by* $t \mapsto C(t, a)$, *the vertical section of* C *at* a *is the function from* $\mathbb{I}$ *in* $\mathbb{I}$ *given by* $t \mapsto C(a, t)$, *and the diagonal section of* C *is the function* δ_C *from* $\mathbb{I}$ *in* $\mathbb{I}$ *define by* $\delta_C(t) = C(t, t)$.

The horizontal, vertical, and diagonal sections of copula C are all non-decreasing and uniformly continuous over $\mathbb{I}$.

2.1.2 Sklar's Theorem

Easily, a copula is a multivariate distribution function in support $\mathbb{I}$, as stated previously, where the marginal are uniforms. This immediately implies that a convex sum of copulas is also a copula. The following theorem known as Sklar's theorem is the central theorem of copula theory. It establishes the link between the above formal definition of copulas and distributions of random variables, thus allowing the application of copulas in statistical modeling.

Definition 2.1.5 *A distribution function F defined in $\overline{\mathbb{R}}$ is given as follows:*

1. *F is non-decreasing.*
2. *$F(-\infty) = 0$ and $F(\infty) = 1$.*

Definition 2.1.6 *A joint distribution function F defined in $\overline{\mathbb{R}}^2$ is given as follows:*

1. *F is 2-increasing.*
2. *$F(x, -\infty) = F(-\infty, y) = 0$ and $F(\infty, \infty) = 1$.*
3. *Since $DomF = \overline{\mathbb{R}}^2$, then F has marginals F_1 and F_2 defined by:*

$$F_1(x) = F(x, \infty) \quad \text{and} \quad F_2(y) = F(\infty, y).$$

Theorem 2.1.3 (Sklar's Theorem) *Let F be the joint distribution function of two random variables X_1 and X_2, and $F_1(x_1)$, $F_2(x_2)$ represent their marginal distribution functions. So there is a copula C*

$$F(x_1, x_2) = C(F_1(x_1), F_2(x_2)), \ \forall (x_1, x_2) \in \mathbb{I}^2 \tag{2.1.4}$$

- If the marginal distribution functions F_1 and F_2 are continuous, then C is unique in $RanF_1 \times RanF_2$.

- Conversely, if C is a copula and F_1 and F_2 are univariate distributions, then the function F defined by (2.1.4) is a joint distribution whose margins are F_1 and F_2, while C is unique.

Thus, the copula combines the marginals to form the multivariate distribution. This theorem provides both a parameterization of multivariate distributions and a construction scheme for the copulas (for the proof of this theorem see [67]). The next proposition is required for the proof of Sklar's theorem.

Proposition 2.1.1 *Let X a random variable of distribution function F, then:*

1. *If U is a random variable with a uniform distribution on $\mathbb{I}$, then $F^{-1}(U) \xrightarrow{d} F$.*
2. *If F is continuous, then $F(X) \xrightarrow{d} U_{\mathbb{I}}$.*

Proof. (Sklar's Theorem) For $(X, Y) = \left(F_1^{-1}(U_1), F_2^{-1}(U_2)\right)$, we have

$$\begin{aligned} F(x,y) &= P[X \leq x, Y \leq y] \\ &= P\left[F_1^{-1}(U_1) \leq x, F_2^{-1}(U_2) \leq y\right] \\ &= P[U_1 \leq F_1(x), U_2 \leq F_2(y)] \\ &= C(F_1(x), F_2(y)). \end{aligned}$$

□

Using Sklar's theorem, we can define the copulas from a couple of random variables as follows:

Definition 2.1.7 *Let H be a distribution function. Then a quasi-inverse of H is a function $H^{(-1)}$ defined on $\mathbb{I}$ such that:*
- If $t \in RanH$, then $H^{(-1)}(t)$ is all $x \in \bar{\mathbb{R}}$ such that $H(x) = t$.
- If $t \notin RanH$, then $H^{(-1)}(t) = \inf\{x | H(x) \geq t\} = \sup\{x | H(x) \leq t\}$.
If H is strictly increasing $H^{(-1)} = H^{-1}$ where H^{-1} is the ordinary inverse of H.

Lemma 2.1.1 *Let S_1 and S_2 be two non-empty subsets of $\overline{\mathbb{R}}$ and F be a 2-increasing function, with marginals of domain $S_1 \times S_2$. Assuming that $(x_1, y_1), (x_2, y_2)$ be any two points of $S_1 \times S_2$ then we have:*

$$|H(x_2, y_2) - H(x_1, y_1)| \leq |F_1(x_2) - F_1(x_1)| + |F_2(y_2) - F_2(y_1)|$$

Lemma 2.1.2 *Let' s F a joint distribution function of marginals F_1 and F_2. Then, there is an unique sub-copule C' such that:*

1. *$DomC' = RanF_1 \times RanF_2$,*
2. *For all x, y in $\overline{\mathbb{R}}^2$, $F(x,y) = C'(F_1(x), F_2(y))$.*

Proof. The joint distribution F verifies the assumptions of lemma (2.1.1) where $S_1 = S_2 = \overline{\mathbb{R}}$. So for all points (x_1, y_1) and (x_2, y_2) in $\overline{\mathbb{R}}^2$ we have :

$$|F(x_2, y_2) - F(x_1, y_1)| \leq |F_1(x_2) - F_1(x_1)| + |F_2(y_2) - F_2(y_1)|,$$

which implies that if

$$F_1(x_2) = F_1(x_1) \text{ and } F_2(y_2) = F_2(y_1),$$

then

$$F(x_2, y_2) = F(x_1, y_1).$$

Therefore, all the couples $\{(F_1(x), F_2(y), F(x,y)) : x, y \in \overline{\mathbb{R}}\}$ defined a two-dimensional real function C' of domain $RanF_1 \times RanF_2$, this function is a sub-copule and follows directly from the properties of F. □

Lemma 2.1.3 *Let C' a sub-copula, then there is a copula C such that $C(u,v) = C'(u,v)$ for all (u,v) in $DomC'$, ie any subcopula can be extended into a copula.*

Proof. see [67] □

Example 1 *Let (a,b) be any point of R^2 and consider the following distribution function:*

$$H(x,y) = \begin{cases} 0, & \text{if } x < a \text{ or } y < b \\ 1, & \text{if } x \geq a \text{ or } y \geq b \end{cases}$$

The marginals of H are the unit scale functions ε_a and ε_b. By applying the lemma (2.1.2), we get the sub-copule C′ of domain $\{0,1\} \times \{0,1\}$, such that $C'(0,0) = C'(0,1) = C'(1,0)$ and $C'(1,1) = 1$.

The extension of C' into a copula C by lemma (2.1.3) is the copula $C = \Pi$, ie $C(u,v) = uv$. Noted that each copula which coincides with C' on its domain is therefore an extension of it.

We are now able to prove Sklar's theorem that we re-learn here for convenience.

Corollary 2.1.1 *Let F, F_1, F_2 be defined in (2.1.3), where C' a sub-copula and let F_1^{-1} and F_2^{-1} be the quasi-inverses of F_1 and F_2 respectively. Then*

$$\forall (u,v) \in DomC\prime,\ C'(u,v) = F(F_1^{-1}(u), F_2^{-1}(v)).$$

Proof. see [67] □

Since a copula is a sub-copula, this corollary is also valid if C' is a copula.

Example 2 (Gumbel 1960a) *Let F be the joint distribution function given by:*

$$F(x,y) = \begin{cases} 1 - e^{-x} - e^{-y} + e^{-(x+y+\alpha xy)}, & x \geq 0,\ y \geq 0 \\ 0, & \text{otherwise} \end{cases}$$

where α is a parameter in $\mathbb{I}$. Then the marginal distribution functions are exponentials, with quasi-inverses

$$F_1^{(-1)}(u) = -\ln(1-u) \text{ and } F_2^{(-1)}(v) = -\ln(1-v) \text{ for } (u,v) \in \mathbb{I}.$$

Hence the corresponding copula is

$$C(u,v) = u + v - 1 + (1-u)(1-v)\, e^{-\alpha \ln(1-u)\ln(1-v)}$$

2.1.3 Copulas and random variables

It is essential to link a copula that describes a set of random variables (or vector Random). This is true especially for the continuity of random variables.

Theorem 2.1.4 *Let X and Y be two random variables whose distribution functions are F_1 and F_2 respectively, where the joint distribution function is F. Then, there exists a copula C satisfying $F(x,y) = C(F_1(x), F_2(y))$. If X and Y are continuous, C is unique on $RanF_1 \times RanF_2$.*

Theorem 2.1.5 (**Independence**) *Let X and Y be two continuous random variables. Then, X and Y are independent if and only if $C_{XY} = \Pi$, where Π is the produced copula.*

Copulas have one of the most important properties characterizing a dependency measure, which is the invariance by strictly increasing transformation.

Theorem 2.1.6 (**Invariance by strictly increasing transformation**) *Let X and Y be a continuous random variables having C_{XY} as a copula. If α and β are functions strictly increasing on $RanX$ and $RanY$ respectively, $C_{\alpha(X)\beta(Y)} = C_{XY}$.*

Thus C_{XY} is invariant under strictly increasing transformations of X and Y.

Proof. Let F be the joint distribution of continuous random variables X and Y, with respective marginal distributions F_1 and F_2. Let F_t be the joint distribution of the transform $(f_1(X), f_2(Y))$, of respective marginal distributions F_{1t} and F_{2t}. We have

$$F_{1t}(x) = P(f_1((X) \leq x) = P(X \leq f_1^{-1}(x)) = F_1\left(f_1^{-1}(x)\right), \forall x \in \mathbb{R} \tag{2.1.5}$$

the same

$$F_{2t}(y) = P(f_2((Y) \leq y) = P(Y \leq f_2^{-1}(y)) = F_2\left(f_2^{-1}(y)\right), \forall y \in \mathbb{R} \tag{2.1.6}$$

We conclude from (2.1.5) and (2.1.6) that for all u and v in $\mathbb{I}$

$$F_{1t}^{-1}(u) = f_1\left(F_{1t}^{-1}(u)\right) \text{ and } F_{2t}^{-1}(v) = f_2\left(F_{2t}^{-1}(v)\right)$$

Then, using Sklar's theorem we have:

$$\begin{aligned} C_{f_1(X)f_2(Y)}(u,v) &= F_t\left(F_{1t}^{-1}(u), F_{2t}^{-1}(v)\right) \\ &= P(f_1(X) \leq F_{1t}^{-1}(u), f_2(Y) \leq F_{2t}^{-1}(v)), \end{aligned}$$

because f_1 and f_2 are bijective, then

$$\begin{aligned} C_{f_1(X)f_2(Y)}(u,v) &= P(X \leq F_1^{-1}(u), Y \leq F_2^{-1}(v)) \\ &= F\left(F_1^{-1}(u), F_2^{-1}(v)\right) \\ &= C_{XY}(u,v) \end{aligned}$$

□

According to [60], this property ensures that the copula completely provides the dependency structure, regardless of the size of the marginal distributions. Thus, any measure of dependence expressed by the copula and marginal distribution functions are too. When the transformations are not necessarily strictly increasing, the following theorem is accurate.

Theorem 2.1.7 *Let X and Y be continuous random variables having C_{XY} as a copula. Let α and β be strictly monotonic functions on $RanX$ and $RanY$ respectively.*
- If α is strictly increasing and β is strictly decreasing, then

$$C_{\alpha(X)\beta(Y)}(u,v) = u - C_{XY}(u, 1-v)$$

- If α is strictly decreasing and β is strictly increasing, then

$$C_{\alpha(X)\beta(Y)}(u,v) = v - C_{XY}(1-u,v)$$

- If α and β are strictly decreasing, then

$$C_{\alpha(X)\beta(Y)}(u,v) = u + v - 1 + C_{XY}(1-u,1-v)$$

2.1.4 Fréchet-Hoeffding Boundaries

We have seen the Fréchet-Hoeffding bounds $M(u,v)$ and $W(u,v)$ for copulas like previously. Hence, its graphic representation is the continuous surface in $\mathbb{I}^3$ whose vertices are $(0,0,0)$, $(0,0,1)$, $(0,1,0)$ and $(1,1,1)$. This graph is located between the graphs of Fréchet-Hoeffding bounds, i.e. the surfaces of M and W. The contour diagram can also represent the graph of a copula.

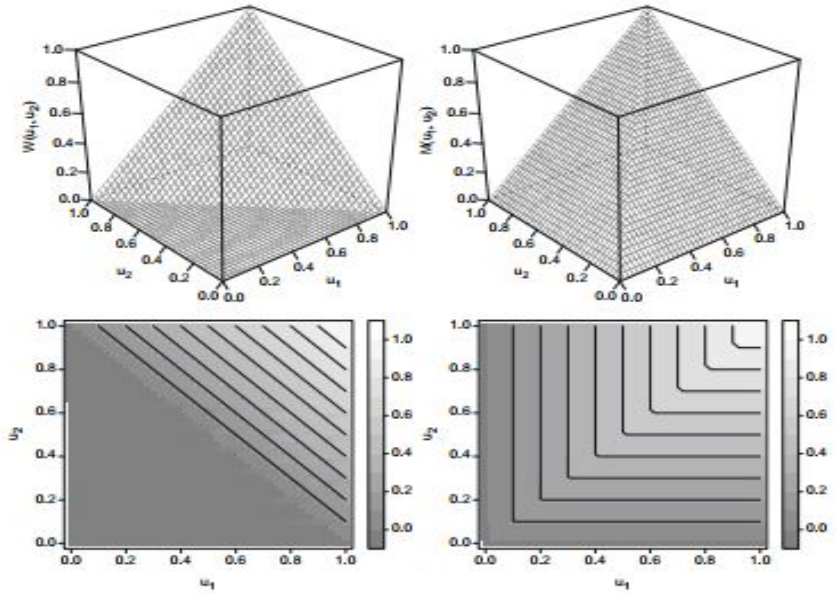

Figure 2.1 – *The Copula (top) and contour plots (bottom) of W and M respectively.*

According to the following theorem, we can say that every copula C has a lower bound and an upper bound.

Theorem 2.1.8 (Frechet-Hoeffding) *Let C a copula, for $\forall u, v \in \mathbb{I}$, then:*

$$W(u,v) = \max(u+v-1,0) \leq C(u,v) \leq M(u,v) = \min(u,v),$$

where M and W represent the lower and upper Fréchet-Hoeffding bounds (respectively).

Proof. see [67] □

Consider the two random variables X, Y of joint distribution function F and of marginals F_1 and F_2 respectively. As a consequence of Sklar's theorem and of the Frechet-Hoeffding theorem we have for x and $y \in \overline{\mathbb{R}}$

$$\max(F_1(x) + F_2(y) - 1, 0) \leq F(x,y) \leq \min(F_1(x), F_2(y)).$$

This is true since M and W are copulas, the above bounds are distribution functions and are called Fréchet-Hoeffding bounds for distribution functions F having F_1 and F_2 as marginals. When these limits are reached, we would like to notice what can be said about the random variables X and Y in this case. The following lemmas and theorem give the conditions of equality between the joint distribution function $F(x,y)$ and one of the two Frechet-Hoeffding bounds.

Lemma 2.1.4 *If F is the distribution function of the random couple (X,Y), then F is equal to the upper Frechet-Hoeffding bound if and only if for all $(x,y) \in \overline{\mathbb{R}}^2$, $P(X > x, Y \leq y) = 0$ or $P(X \leq x, Y > y) = 0$.*

Proof. We have

$$\begin{aligned} F_1(X) &= P(X \leq x) \\ &= P(X \leq x, Y \leq y) + P(X \leq x, Y > y) \\ &= F(x,y) + P(X \leq x, Y > y), \end{aligned}$$

and

$$\begin{aligned} F_2(y) &= P(Y \leq y) \\ &= P(X \leq x, Y \leq y) + P(X > x, Y \leq y) \\ &= F(x,y) + P(X > x, Y \leq y), \end{aligned}$$

then

$$M(F_1(x), F_2(y)) = F(x,y) + \min(P(X \leq x, Y > y), P(X > x, Y \leq y)),$$

so

$$F(x,y) = M(F_1(x), F_2(y)),$$

only if

$$\min(P(X \leq x, Y > y), P(X > x, Y \leq y)) = 0.$$

□

Lemma 2.1.5 *Let S a subset of $\overline{\mathbb{R}}^2$. S is decreasing if and only if for all $(x,y) \in \overline{\mathbb{R}}^2$,*

1. *$\forall (u,v) \in S,\ u \leq x \Longrightarrow v > y$. Or*
2. *$\forall (u,v) \in S,\ v > y \Longrightarrow u \leq x$.*

Theorem 2.1.9 *Let X, Y two random variables of joint distribution function F. We say that H is equal to the lower Fréchet-Hoeffding bound if and only if the support of F is a decreasing subset of $\overline{\mathbb{R}}^2$.*

2.1.5 Survival and semi-survival copulas

In this sub-section, we define the functions associated with copulas, which will be used in Chapter 5 and 6.

- **Survival copula**

  For a couple of random variables (X, Y) having F as the joint distribution function, the joint survival function is given by:

  $$S(x, y) = P(X > x, Y > y),$$

  where $S_1(x) = 1 - F_1(x)$ and $S_2(x) = 1 - F_2(y)$ are the marginal survival functions of X and Y respectively, presented in (1.1.3).

  The question arises! is there a link between the marginal survival functions and their joint survival function ! If we assume that C is the copula of X and Y. Then we have:

  $$\begin{aligned} S(x, y) &= 1 - F_1(x) - F_2(y) + F(x, y) \\ &= S_1(x) + S_2(x) - 1 + C(F_1(x), F_2(y)) \\ &= S_1(x) + S_2(x) - 1 + C(1 - S_1(x), 1 - S_2(x)). \end{aligned}$$

  Then, if we define a function $\tilde{C}$ from $I^2 \to I$ we obtain:

  $$\tilde{C}(u, v) = u + v - 1 + C(1 - u, 1 - v), \tag{2.1.7}$$

  We can notice that $\tilde{C}$ is a copula and we call it the survival copula of X and Y. This copula relates the joint survival function to its univariate marginals in a completely analogous way to that in which the copula relates the joint distribution function to its marginals.

- **Semi-Survival copula**

  For a couple of random variables (X, Y), the function $S(x, y)$ can be written as:

$(x), F_2(y))$
$= \tilde{C}(S_1(x), F_2(y))$.

  The function $\tilde{C}$ is called the semi-survival copula associated with the copula C and the expression on I^2 is given by:

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

2.1.6 Copula properties

Let X, Y be two continuous random variables with a copula C and a joint distribution function F with F_1 and F_2 as marginals.

- **Symmetry**

  If X is a random variable and a is a real number. We say that X is symmetric with respect to a if $X - a$ and $a - X$ have the same distribution i.e.

$$\forall x \in \mathbb{R}, P\left(X - a \leq x\right) = P\left(a - X \leq x\right).$$

  When X is continuous and has a distribution function H, it is equivalent to

$$H\left(a + x\right) = \bar{H}\left(a - x\right) \tag{2.1.8}$$

  When H is discontinuous, (2.1.8) holds only at the points of continuity of H. Now consider a couple of random variables (X, Y), we can define symmetry in terms of couple $(a, b) \in \mathbb{R}^2$ in a different way. One of the most important ways of symmetry is exchangeability.

Definition 2.1.8 (**Exchangeability** [67]) *Two random variables X and Y are exchangeable if the random vectors (X, Y) and (Y, X) are identically distributed. The exchangeability between two random variables X and Y, where F is the joint distribution function, can be expressed as:*

$$F\left(x, y\right) = F\left(y, x\right), \forall \left(x, y\right) \in \mathbb{R}^2$$

For identically distributed random variables, the exchangeability is equivalent to the symmetry of their copula as expressed in the following theorem.

Theorem 2.1.10 *Let X and Y be two continuous random variables having the joint distribution function F, the marginal distribution functions F_1 and F_2, respectively, and the copula C as an associated copula. We say that X, Y are exchangeable if and only if:*

$$F_1 = F_2 \text{ and } C(u, v) = C(v, u), \forall \left(u, v\right) \in \mathbb{I}^2$$

If $C(u, v) = C(v, u)$, for all $(u, v) \in \mathbb{I}^2$, we say that C is symmetrical.

- **Order relation**

The inequality of the Fréchet-Hoeffding bounds suggests the existence of a partial order on the set copulas as follows:

Definition 2.1.9 *Let C_1, C_2 be two copulas. We say that C_1 is smaller than C_2 (or C_2 is greater than C_1) and we denote by $C_1 \prec C_2$ (or $C_1 \succ C_2$) if:*

$$C_1(u, v) \leq C_2(u, v), \quad \forall \left(u, v\right) \in \mathbb{I}^2. \tag{2.1.9}$$

Remark 2.1.1 *We can notice that according to this order, the Fréchet-Hoeffding upper bound* $M = \min(u, v)$ *is greater than any other copula. And that the Fréchet-Hoeffding lower bound* $W = \max(u + v - 1, 0)$ *is smaller than any other copula.*

This punctual partial order of the copulas set is called order of concordance which is a tool for discussing the relationship between copulas and the dependence properties between random variables. Because not all copula pairs are comparable, it is clear that this order is only partial.

Example 3 *The copula produces* Π *and the copula obtained by the average of the two Fréchet Hoeffding bounds are not comparable. Beceause, if we suppose that:*

$$C(u, v) = \frac{[W(u, v) + M(u, v)]}{2},$$

then, we have $C\left(\frac{1}{4}, \frac{1}{4}\right) > \Pi\left(\frac{1}{4}, \frac{1}{4}\right)$ *and* $C\left(\frac{1}{4}, \frac{3}{4}\right) < \Pi\left(\frac{1}{4}, \frac{3}{4}\right)$
so neither $C \prec \Pi$ *nor* $\Pi \prec C$.

- **Convexity and concavity**

Definition 2.1.10 *A copula is said to be concave if we have:*

$$C(\alpha a + (1 - \alpha) c, \alpha b + (1 - \alpha) d) \geq \alpha C(a, b) + (1 - \alpha) C(c, d),$$

for all α, a, b, c, d *in* $\mathbb{I}$. *And it is said to be convex if we have:*

$$C(\alpha a + (1 - \alpha) c, \alpha b + (1 - \alpha) d) \leq \alpha C(a, b) + (1 - \alpha) C(c, d),$$

for all α, a, b, c, d *in* $\mathbb{I}$.

- **Partial derivatives**

Definition 2.1.11 *The partial derivatives of* $C(u, v)$ *almost surely exist for all* $(u, v) \in \mathbb{I}^2$,

$$0 \leq \frac{\partial C(u, v)}{\partial u} \leq 1 \qquad \text{and} \qquad 0 \leq \frac{\partial C(u, v)}{\partial v} \leq 1.$$

- **Copula's density**

Let X, Y two continuous random variables, we denote by f the joint density function associated with F, and by f_1 and f_2, the marginal density functions of X, Y respectively.

Definition 2.1.12 *The density* $c(F_1(x), F_2(y))$ *associated with* $C(F_1(x), F_2(y))$ *is defined by:*

$$c\left(F_1(x), F_2(y)\right) = \frac{\partial^2 C\left(F_1(x), F_2(y)\right)}{\partial F_1(x)\partial F_2(y)} = \frac{f\left(x,y\right)}{f_1(x)f_2(y)}$$

We thus have according to Sklar's theorem in particular:

$$f\left(x,y\right) = f_1(x)f_2(y).c\left(F_1(x), F_2(y)\right)$$

- **Harmonic copula**

  Let C a copula whose second-order partial derivatives are continuous in $\mathbb{I}^2$.

Definition 2.1.13 *The copula C is harmonic in $\mathbb{I}^2$, if C satisfies the equation:*

$$\nabla^2 C\left(u,v\right) = \frac{\partial^2}{\partial u^2} C\left(u,v\right) + \frac{\partial^2}{\partial v^2} C\left(u,v\right) = 0.$$

Example 4 *The copula $\Pi(u,v) = uv$, is a harmonic copula:*

$$\frac{\partial^2}{\partial u^2}\Pi\left(u,v\right) = \frac{\partial^2}{\partial v^2} C\left(u,v\right) = 0.$$

- **Homogeneous copula**

Definition 2.1.14 *A copula C is homogeneous of degree k if $\exists k \in \mathbb{R}$, $\forall u, v, \lambda \in I$*

$$C\left(\lambda u, \lambda v\right) = \lambda^k C\left(u,v\right). \tag{2.1.10}$$

a. The function $\Pi = uv$, is homogeneous of degree 2, because

$$\left(\lambda u\right)\left(\lambda v\right) = \lambda^2 uv.$$

b. The function $M = \min(u,v)$, is homogeneous of degree 1, because

$$\min(\lambda u, \lambda v) = \lambda \min\left(u,v\right).$$

2.2 Bivariate Copula families

Several studies have focused on the construction of different copula families. In this section, we represent the most common families in the bivariate case.

2.2.1 Usual Copulas

- **Independency copula**

Let X, Y two continuous random variables and F the joint distribution function, F_1 and F_2 the marginal distributions of X, Y respectively.

Definition 2.2.1 *If X, Y are two independent random variables, then the associated copula is a product of its marginals.*

$$C_{X,Y}(x, y) = F_1(x).F_2(y)$$

The copula thus defined is harmonic and homogeneous of degree 2.

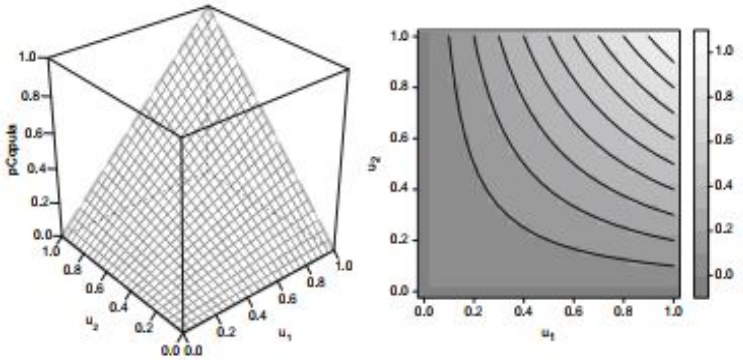

Figure 2.2 – *Copula (Left) and contour plot (right) of the independence Copula.*

- **Gaussian copula**

If we denote ϕ_ρ the joint distribution function of the bivariate normal distribution with linear correlation coefficient $\rho \in \mathbb{I}$, where ϕ the distribution function of the standard normal distribution. The Gaussian copula of the couple random X, Y is defined by:

$$C_\rho^\phi(u, v) = \phi_\rho\left(\phi^{-1}(u), \phi^{-1}(v)\right),$$

where

$$\phi_\rho\left(\phi^{-1}(u), \phi^{-1}(v)\right) = \int_{-\infty}^{\phi^{-1}(u)} \int_{-\infty}^{\phi^{-1}(v)} \frac{1}{2\pi\sqrt{1-\rho_{X,Y}^2}} \exp\left(\frac{2st\rho_{X,Y} - s^2 - t^2}{1-\rho_{X,Y}^2}\right) dsdt.$$

This copula is parameterized by the linear correlation coefficient ρ.

- $C_\rho^\phi(u, v) \longrightarrow \Pi$, when $\rho \longrightarrow 0$.
- $C_\rho^\phi(u, v) \longrightarrow W(u, v)$, when $\rho \longrightarrow -1$.
- $C_\rho^\phi(u, v) \longrightarrow M(u, v)$, when $\rho \longrightarrow +1$.

- **Student's copula**

In the univariate case, the distribution function of a Student random variable is defined by:

$$t_v(x) = \int_{-\infty}^{x} \frac{\Gamma\left(\frac{v+1}{2}\right)}{\sqrt{\pi v}\Gamma(v+2)} \left(1 + \frac{s^2}{v}\right)^{-\frac{v+1}{2}} ds,$$

where Γ is the Euler function defined by:

$$\Gamma(x) = \int_0^{\infty} t^{x-1} e^{-t} dt.$$

In the bivariate case, let $\rho \in [-1,1]$, then the bivariate distribution function is:

$$t_{\rho,v}(x,y) = \int_{-\infty}^{x}\int_{-\infty}^{y} \frac{1}{2\pi\sqrt{1-\rho^2}} \left(1 + \frac{s^2+t^2-2\rho st}{v(1-\rho^2)}\right)^{-\frac{v+2}{2}} dsdt.$$

Definition 2.2.2 *A Student's copula is a parametric copula parameterized by the linear correlation coefficient ρ and freedom degree v. This copula is defined by:*

$$\begin{aligned} C^t_{\rho,v}(u,v) &= t_{\rho,v}\left(t^{-1}_{\rho,v}(u), t^{-1}_{\rho,v}(v)\right) \\ &= \int_{-\infty}^{t_v^{-1}(u)} \int_{-\infty}^{t_v^{-1}(v)} \frac{1}{2\pi\sqrt{1-\rho^2}} \left(1 + \frac{s^2+t^2-2\rho st}{v(1-\rho^2)}\right)^{-\frac{v+2}{2}} dsdt. \end{aligned}$$

The corresponding density is then defined by:

$$c^t_{\rho,v}(u,v) = \rho^{-\frac{1}{2}} \frac{\Gamma\left(\frac{v+2}{2}\right)\Gamma\left(\frac{v}{2}\right)}{\Gamma\left(\frac{v+1}{2}\right)^2} \frac{\left(\frac{1+\left(t_v^{-1}(u)\right)^2+\left(t_v^{-1}(v)\right)^2-2\rho\left(t_v^{-1}(u)\right)\left(t_v^{-1}(v)\right)}{v(1-\rho^2)}\right)^{-\left(\frac{v+2}{2}\right)}}{\left(1+\left(t_v^{-1}(u)\right)^2\right)^{-\left(\frac{v+2}{2}\right)}\left(1+\left(t_v^{-1}(v)\right)^2\right)^{-\left(\frac{v+2}{2}\right)}}.$$

Remark 2.2.1

a) *The Gaussian and the Student copulas are both members of the elliptical copula family.*

b) *If the freedom degree $v \to \infty$, then the Student copula converges to the Gaussian copula and it is very difficult to differentiate between these two copulas.*

2.2.2 Archimedean Copulas

Before its introduction in many fields such as finance, this family of copulas was first recognized by Schweizer and Sklar (1961)[80], during the study of the t-norm, and its name is due to Ling (1965)[58].

- **Generator and Archimedien copula**

Before presenting this family of copulas, it is necessary to present the following definition and remarks:

Definition 2.2.3 *Let $\varphi : \mathbb{I} \to \mathbb{R}_+$, continuous, decreasing and convex function, such that $\varphi(1) = 0$, then φ is said to be generator.*

The pseudo-inverse of φ is defined as follows:

$$\varphi^{-1}(u) = \begin{cases} \varphi^{-1}(u) & si \quad 0 \le u \le \varphi(0) \\ 0 & si \quad \varphi(0) \le u \le +\infty \end{cases}$$

If $\varphi(0) = \infty$, then φ is strictly decreasing.

Definition 2.2.4 *The Archimedian copula is defined by:*

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

where $\varphi : \mathbb{I} \to \mathbb{R}_+$, *is a generator. This copula has the following properties:*

a. ***Symmetry***

$$C(u,v) = C(v,u), \forall (u,v) \in \mathbb{I}^2$$

b. ***Associativity***

$$C(C(u,v),z) = C(u,C(v,z)), \forall (u,v,z) \in \mathbb{I}^3$$

c. ***Contour convexity***

$$\{(u,v) \in \mathbb{I}^2 : \varphi(u) + \varphi(v) = \varphi(k)\},\ k > 0$$

d. ***Density***

$$c(u,v) = -\frac{\varphi''(C(u,v)\varphi'(u)\varphi'(v))}{(\varphi'(C(u,v)))^3}$$

- **Archimedean Copula of a Single Parameter**

The Gumbel (1960), Clayton (1978), and Frank (1978) copulas are the most well-known and widely used among this family of copulas.

- Copula of Gumbel (1960):

Gumbel's copula is a symmetrical copula defined by:

$$C_\alpha(u,v) = \exp -\left((-\ln u)^\alpha + (-\ln v)^\alpha\right)^{\frac{1}{\alpha}},$$

whose generator is defined by:

$$\varphi_\alpha(t) = (-\ln t)^\alpha,$$

where the dependency parameter $\alpha \in [1; +\infty[$, also we have:

a. $C_\alpha \to \Pi$, when $\alpha \to 1$.

b. $C_\alpha \to M$, when $\alpha \to \infty$

- Clayton's Copula (1978):

This copula is also called the Cook and Johnson copula (1981)[17], but the first to have studied it are Kimeldorf and Sampson (1975)[53], this copula is defined by:

$$C_\alpha(u,v) = \left(u^{-\alpha} + v^{-\alpha} - 1\right)^{-\frac{1}{\alpha}},$$

where the generator and the pseudo inverse are defined respectively by:

$$\varphi_\alpha(t) = \frac{1}{\alpha}\left(t^{-\alpha} - 1\right) \text{ and } \varphi_\alpha^{-1}(t) = (t+1)^{-\frac{1}{\alpha}}.$$

Where the dependency parameter $\alpha \in [-1; 0[\cup]0; +\infty.[$.

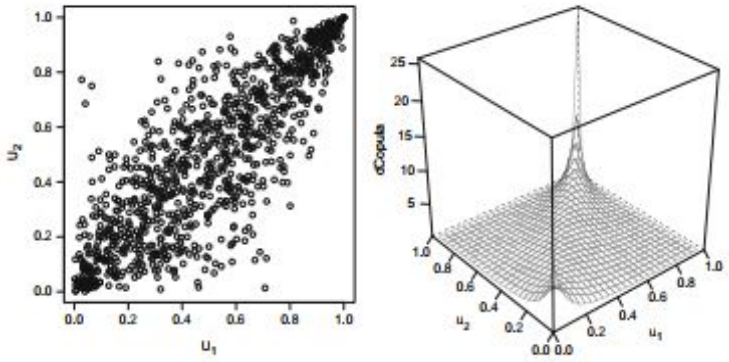

Figure 2.3 – *Scatter plot of $n = 1000$ independent observations from Gumbel Copula for θ (left) and wireframe plot of the corresponding density (right).*

a. If the parameter $\alpha \to 0$, then the marginal are independent.

b. If the parameter $\alpha \to \infty$, then we get the Fréchet-Hoffding upper bound copula M.

- Frank's Copula (1978):

This copula is a symmetric copula of a dependency parameter $\alpha \in [-\infty; 0[\cup]0; +\infty.[$. It is defined by:

$$C_\alpha(u, v) = -\frac{1}{\alpha} \ln\left(1 + \frac{(e^{-u\alpha} - 1)(e^{-v\alpha} - 1)}{e^{-\alpha} - 1}\right),$$

of generator:

$$\varphi_\alpha(t) = -\ln\left(\frac{\exp(-\alpha t) - 1}{\exp(-\alpha) - 1}\right),$$

and density

$$c_\alpha(u, v) = \frac{(\alpha - 1)\ln\alpha^{u+v}}{((\alpha - 1) + (\alpha^u - 1)(\alpha^v - 1))^2}.$$

a. $C_\alpha \to \Pi$, when $\alpha \to 0$.

b. $C_\alpha \to M$, when $\alpha \to +\infty$.

c. $C_\alpha \to W$, when $\alpha \to -\infty$.

The densities of Frank, Clayton, and Gumbel copulas are depicted in Figure (2.2.2).

2.2.3 Extreme values Copulas

Another family of copulas which is widely used is that of the extreme values. As the names indicate, it is a class of copulas related by the notion of extreme values. The distribution of extreme values in the univariate case will be presented first, followed by the copula of extreme value.
Let $X_1, ..., X_n$ a sequence of random variables i.i.d. That is M_n is given by: $M_n = \max(X_1, ..., X_n)$.

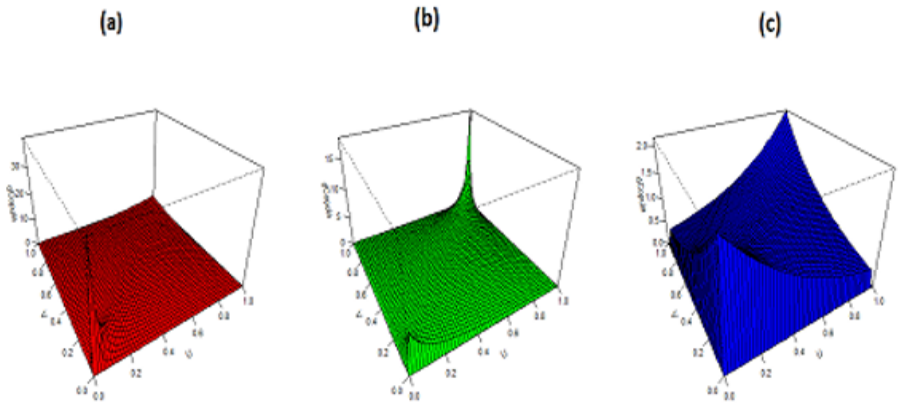

Figure 2.4 – *Copula densities (a): Clayton for $\alpha = 3$, (b): Gumbel for $\alpha = 2$ and (c): Frank for $\alpha = 2$.*

Theorem 2.2.1 ***(Fisher-Tippet, 1928)***
If there are two sequences $c_n > 0$ and $d_n \in \mathbb{R}$, such that $\frac{M_n - d_n}{c_n}$ converges in a non-degenerate distribution, then:

$$\lim_{n\to\infty} P(\frac{M_n - d_n}{c_n} \leq x) = G_\alpha(x),$$

where $G_\alpha(x)$ takes one three of the distributions lisetd below:

- **Fréchet** $\quad \Phi_\alpha(x) = \begin{cases} 0, & \text{if} \quad x \leq 0 \\ \exp\{-x^{-\alpha}\}, & \text{if} \quad x > 0 \end{cases} \qquad \alpha > 0$
- **Weibull** $\quad \Psi_\alpha(x) = \begin{cases} \exp\{-(-x)^{\alpha}\}, & \text{if} \quad x \leq 0 \\ 1, & \text{if} \quad x > 0 \end{cases} \qquad \alpha > 0$
- **Gumbel** $\quad \Lambda(x) = \exp\{-e^{-x}\}, \qquad x \in \mathbb{R}$

Lemma 2.2.1 *The statistics $\left(n\left(\frac{M_n}{\theta} - 1\right)\right)_{n\geq 1}$ converges in distribution to the Weibull random variable, when $\alpha = 1$.*

Proof. We assume that $G(x) = P(n\left(\frac{M_n}{\theta} - 1\right) \leq x)$, and

$$F_{X_n} = P(\max(X_i) \leq x) = (F(x))^n = (x/\theta)^n,$$

because the distribution function of the uniform distribution $F(x) = x/n$, then,

$$\begin{aligned} G(x) &= P\left(n\left(\frac{M_n}{\theta} - 1\right) \leq x\right) \\ &= P\left(\frac{M_n}{\theta} \leq \frac{x}{n} + 1\right) \\ &= P\left(M_n \leq \theta(\frac{x}{n} + 1)\right) \\ &= F_{M_n}(\theta(\frac{x}{n} + 1)) \\ &= (1 + \frac{x}{n})^n \\ &= \exp(x) \end{aligned}$$

so $G(x) = \exp(x)$ has a Weibull distribution. □

Lemma 2.2.2 *Let $(X_1, Y_1), ..., (X_n, Y_n)$ pairs of random variables i.i.d of a common copula C and $C_{(n)}$ the copula associated where: $X_{(n)} = \max(X_i)$, and $Y_{(n)} = \max(y_i)$, for all $i = 1, ..., n$. Then*

$$C_{(n)}(u, v) = C^n(u^{\frac{1}{n}}, v^{\frac{1}{n}}), \quad 0 \leq u, v \leq 1.$$

Proof.

$$\begin{aligned} F_{(n)}(x, y) &= P\left(\max(X_i) \leq x, \max(Y_j) \leq y\right) \\ &= P\left(X_1 \leq x, X_2 \leq x, ...X_n \leq x \text{ and } Y_1 \leq y, Y_2 \leq Y, ...Y_n \leq y\right) \\ &= (F(x, y))^n \\ &= C(F_1(x), F_2(y))^n \\ &= C^n((F_{1(n)}(x))^{\frac{1}{n}}, (F_{2(n)}(y))^{\frac{1}{n}}). \end{aligned}$$

□

The limit of the sequence $\left\{C_{(n)}\right\}$, gives us the following definition.

Definition 2.2.5 *We say that C^*, is a Copula of bivariate extreme values if there is a copula C, such that:*

$$C^*(u, v) = \lim_{n \to \infty} C^n(u^{\frac{1}{n}}, v^{\frac{1}{n}})$$

2.2.4 Bivariate extreme values distributions

Let $(X_1, Y_1), ..., (X_n, Y_n)$ pairs of random variables i.i.d of common distribution F. So there are $a_n, c_n > 0$ and $b_n, d_n \in \mathbb{R}$, such that:

$$\lim_{n \to \infty} P(\frac{X_{(n,n)} - b_n}{a_n} \leq x \leq \frac{Y_{(n,n)} - d_n}{c_n}) = G_\theta(x, y),$$

so G is a non-degenerate distribution if and only if the marginal distributions of G are univariate extreme value distributions.

- **Parametric family of bivariate extreme value copulas**

There are essentially two large families of the usual parametric model of bivariate extreme value copulas: the mixed model and the logistic model or Gumbel (1960) [40]. The other models usually come from a symmetrical or asymmetrical extension of these models. We present these distributions in table (2.1). Where $\tilde{u} = -\ln u$, $\tilde{v} = -\ln v$, and Φ is the distribution function of the reduced centered normal distribution.

Theorem 2.2.2 *For any copula of bivariate extreme values C^*, there is a convex function A defined from $\mathbb{I}$ in $\left[\frac{1}{2}, 1\right]$, such that:*

$$C^*(u, v) = \exp\left[\left(-\ln(u) + \ln(v) A\left(\frac{\ln u}{\ln(u) + \ln(v)}\right)\right)\right]$$

In addition, A checks $\max(t, 1-t) < A(t) < t, \quad \forall t \in \mathbb{I}$. A is called a generator or Pickands dependency function (see [36], [45]).

family	$C^*_\theta(u,v)$	$A_\theta(t)$
Independence	uv	$A(t)=1$
Gumbel_1	$\exp\left\{-(\tilde{u}^\theta+\ln\tilde{v}^\theta)^{\frac{1}{\theta}}\right\}, \theta\geq 1$	$\left[t^\theta+(1-t^\theta)\right]^{\frac{1}{\theta}}$
Gumbel_2	$uv\exp\left\{\theta\frac{\tilde{u}\tilde{v}}{\tilde{u}+\tilde{v}}\right\}, \theta\geq 0$	$t^2-\theta t+1$
Galambos	$uv\exp\left\{-\left(\tilde{u}^{-\theta}+\tilde{v}^{-\theta}\right)^{\frac{1}{\theta}}\right\}$	$1-\left[t^{-\theta}+(1-t^{-\theta})\right]^{-\frac{1}{\theta}}$
Husler-Reiss	$\exp\left\{\begin{array}{l}-\tilde{v}\Phi[\frac{1}{\theta}+\frac{1}{2}\theta\log(\frac{\tilde{v}}{\tilde{u}})]-\\ \tilde{u}\Phi\left[\frac{1}{\theta}\right]+\frac{1}{2}\theta\log(\frac{\tilde{u}}{\tilde{v}})\end{array}\right\}, \theta\geq 0$	$t\Phi\left\{\begin{array}{l}\left[\frac{1}{\theta}+\frac{1}{2}\theta\log(\frac{t}{1-t})\right]+\\ t\Phi\left[\frac{1}{\theta}+\frac{1}{2}\theta\log(\frac{t}{1-t})\right]\end{array}\right.$
Marchal-Olkin	$\left\{\begin{array}{l}uv^{1-\beta} \text{ si } u^\alpha<v^\beta\\ u^{1-\alpha}v \text{ si } u^\alpha>v^\beta\end{array}\right.$	$\left\{\begin{array}{l}\max\{1-\alpha t, 1-\beta(1-t)\},\\ \alpha\leq 1, \beta\geq 0\end{array}\right.$
Tawn	$uv\exp\left\{\begin{array}{l}-(1-\delta)+(\theta-\delta)\tilde{u}+\\ \left[(\theta\tilde{u})^\lambda+(\delta\tilde{v})^\lambda\right]^{\frac{1}{\lambda}}\end{array}\right\}$	$\left[\begin{array}{l}(1-\delta)+(\delta-\theta)t+\\ \left[(\theta t)^\lambda+(\delta(1-t)^\lambda\right]^{\frac{1}{\lambda}}\end{array}\right]$

Table 2.1 – *Extreme value copulas.*

2.3 Multivariate Copula

The properties of bivariate copulas in the multivariate case are investigated in this section. While some of the results are similar, others are not. Let $A_1,...,A_n$ non-empty subsets of $\mathbb{R}^n$ and G a function defined on $A_1\times...\times A_n\to\mathbb{R}$.

Definition 2.3.1 *Let a_i the smallest elements of A_i, where $i=1,...,n$. The function G is said to be grounded if it is equale to zero for all $v\in A_1\times...\times A_n$, and for at least one index k such that $v_k=a_k$*

$$G(v_1,...,v_{k-1},a_k,a_{k+1},...,a_n)=0. \tag{2.3.1}$$

Definition 2.3.2 *Let $S_1,...,S_n$ non-empty measurable parts of $\bar{\mathbb{R}}$. Let $B=[a;b]$ a n-pavement whose vertices are in Dom G. The volume G of B is then defined by:*

$$V_G(B)=\sum sgn(c)G(c),$$

where the sum is carried out on all the vertices c of B and the $sgn(c)$ is given by:

$$sgn(c)=\left\{\begin{array}{ll}1 & \text{if } c_k=a_k \text{ for an even number of } k\\ -1 & \text{if } c_k=a_k \text{ for odd number of } k\end{array}\right.$$

Definition 2.3.3 *The function G is said to be n-increasing if $V_G(B)\geq 0$ for all n-pavement B whose vertices are in Dom G.*

Definition 2.3.4 *A d-dimensionally sub-copula (or n-sub-copula) $\tilde{C}$ is a real function defined on $A_1\times...\times A_n$, where for $i=1,...,n$, the A_i are non-empty subsets of $\mathbb{I}$ containing both 0 and 1 satisfying the conditions $\tilde{C}$ is grounded (2.3.1) .*

1. $\tilde{C}$ has one-dimensional marginals, $\tilde{C}_i$ for $i=1,...,n$, such that $\forall u_i\in A_i, C_i(u_i)=u_i$.
2. $\tilde{C}$ is n-increasing .

Definition 2.3.5 (**n-Sub-Copula**) *An n-dimensional sub-copula (or n-sub-copula) is a function C' having the following properties:*

1. *$DomC' = S_1 \times S_2 \times S_n$, where any S_k is a subset of $\mathbb{I}$ containing 0 and 1.*
2. *C' is grounded and n-increasing.*
3. *C' has marginals C'_k, $k = 1, ..., n$ who satisfy*
$$C'_k(u) = u \text{ for all } u \in S_k.$$

Note that for every u in $DomC'$, $0 \leq C'(u) \leq 1$, $RanC'$ is also a subset of $\mathbb{I}$. Then, in this sense, an n-copula is stated as follows:

Definition 2.3.6 (n-Copula) *An n-dimensional copula (or n-copula) is an n-sub-copula whose domain is $\mathbb{I}^n$.*

Definition 2.3.7 (n-Copula) *An n-dimensional copula C, is defined from $\mathbb{I}^d$ in $\mathbb{I}$ having the following properties:*

- *For at least one of $u = 0$ coordinates, $\forall u \in \mathbb{I}^n$, then $C(u) = 0$.*
- *For all coordinates equals 1 except u_i, then $C(u) = u_i$.*
- *$\forall u, v \in \mathbb{I}^d$, such that $u \leq v$, we have $V_C([u, v]) \geq 0$.*

2.3.1 Sklar's theorem

The significance of this theorem is the same as that of the bivariate case. The following is the n-dimensional form of Sklar's theorem.

Theorem 2.3.1 (n-dimensional Sklar's theorem) *Let H an n-dimensional distribution function of marginal distribution functions $F_1, ..., F_n$. So there is a n-copula C such that for all $x \in \bar{\mathbb{R}}^n$,*
$$H(x_1, ..., x_d) = C(F_1(x_1), ..., F_n(x_n)) \tag{2.3.2}$$

If the functions $F_1, ..., F_n$ are continuous, then C is unique.

As in the bivariate case, the continuity of marginals is a sufficient condition for the uniqueness of the copula. The following corollary is very useful in estimating the copula.

Corollary 2.3.1 (Inverse Sklar's theorem) *Let H be an n-dimensional distribution function whose marginals are $F_1, F_2, ..., F_n$, where C the associated n-copula and let $F_1^{-1}, F_2^{-1}, ..., F_n^{-1}$ be the quasi-inverses of $F_1, F_2, ..., F_n$, respectively. Then for all $u \in \mathbb{I}^n$*
$$C'(u_1, u_2, ...u_n) = F(F_1^{-1}(u_1), F_2^{-1}(u_2), ..., F_n^{-1}(u_n))$$

Example 5
$$\Pi^n(u) = u_1...u_n.$$
$$M^n(u) = \min(u_1, ..., u_n).$$

Remark 2.3.1 *The function $W^n(u) = \max(u_1 + ...u_n + -n + 1, 0)$ is not a n-copula for $n > 2$, it is not n-increasing for n-volume $\left[\frac{1}{2}, 1\right]^n \subset \mathbb{I}^n$ (see[14]). But in a Sklar's theorem (1998), he verifies that the copula $C(u) = \max(u_1 + ...u_n - 10)$ is a copula $\forall u \in \mathbb{I}^d$ and for $n > 2$ (see [67]).*

Theorem 2.3.2 (Multivariate Fréchet-Hoeffding bounds) *Every multivariate copula satisfies the following inequality:*
$$\max(u_1 + ...u_n - 1; 0) \leq C(u) \leq \min(u_1, ..., u_n).$$

2.3.2 A multivariate copula's properties

Theorem 2.3.3 (**Existence and Uniqueness**) *Let $X_1, ..., X_n$ random variables, including marginal distribution functions $F_1, F_2, ..., F_n$ respectively, and the joint distribution function H. Then there exists an n-copula C such that (2.3.2) is verified. If $F_1, F_2, ..., F_n$ are all continuous, C is unique.*

Theorem 2.3.4 (**Uniform continuity**) *A copula C is uniformly continuous across its entire domain, especially for all u, v in $\mathbb{I}^d$, we have*

$$|C(u) - C(v)| \leq \sum_{k=1}^{n} |v_k - u_k| .$$

Theorem 2.3.5 (**Invariance**) *Let $(X_1, ..., X_n)$ a vector of continuous random variables, of distribution function F associated with a copula C and $(\alpha_1, ..., \alpha_n)$ a series of strictly increasing functions. Then, the joint distribution function of the random vector $(\alpha_1(X_1), ..., \alpha_d(X_d))$is also associated with the same copula C.*

$$C_{\alpha_1(X_1),...,\alpha_2(X_d)} = C_{X_1,...,X_d}(u)$$

Theorem 2.3.6 (**Partial derivatives**) *Let the copula C. The partial derivatives of C almost certainly exist, for all $i = 1, ..., d$ and for all $u \in \mathbb{I}^d$, we have*

$$0 \leq \frac{\partial C(u)}{\partial u_i} \leq 1.$$

In addition, the functions

$$u \to \frac{\partial C(u)}{\partial u_i} \quad \textit{are non-decreasing.}$$

Definition 2.3.8 (**Copula's density**) *The density c associated with the copula C is defined by:*

$$c(F_1(x_1), ..., F_d(x_d)) = \frac{\partial^d C(F_1(x_1), ..., F_d(x_d)))}{\partial F_1(x_1) ... \partial F_d(x_d)} = \frac{h(x_1, ...x_d)}{f_1(x_1), ...f_d(x_d)}$$

$$c(F_1(x_1), ..., F_d(x_d)) = \frac{h(F_1^{-1}(u_1), ..., F_d^{-1}(u_d)}{\prod_{i=1}^{d} f_i(F_i^{-1}(u_i))},$$

such that h is the density of H and f_i is the density of F_i.

If the multivariate distribution function H is absolutely continuous, then by using Sklar's theorem we can present the density function h as a function of the density c and the marginal densities $f_1, ..., f_d$ by:

$$h(x_1, ...x_d) = c(F_1(x_1), ..., F_d(x_d)) \prod_{i=1}^{d} f_i(x_i)$$

2.3.3 Multivariate parametric copula

Here we generalize all cases of bivariate copulas, we give the multivariate form to each case presented previously.

1. **Independency copula**

   The random variables $X_1, ..., X_d$ are independent if and only if

   $$H(x_1, ...x_d) = F_1(x_1) \times ... \times F_d(x_d).$$

   We therefore define the multivariate independence copula by

   $$\Pi^d(u) = u_1 \times ... \times u_d,$$

   such that $u_i = F_i(x_i)$ for $i = 1, ..., d$.

2. **Gaussian copula**

   Let Φ_R^d a normal multivariate standard distribution with a correlation matrix R, so the Gaussian copula is defined by

   $$C_R^{d,Ga}(u) = \Phi_R^d(\Phi^{-1}(u_1), ..., \Phi^{-1}(u_d)),$$

   where Φ^{-1} is the generalized inverse of the univariate normal standard distribution function Φ. According to (1.15) we have:

   $$\frac{1}{(2\pi)^{\frac{d}{2}} |R|^{\frac{1}{2}}} \exp(-\frac{1}{2} X^t R^{-1} X) = c_R^{d,Ga}(\Phi(x_1), ..., \Phi(x_d)) \times \prod_{j=1}^{d} (\frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2} x_j^2)).$$

3. **Student Copula**

   Let $T_{R,v}^d$ a multivariate distribution function of the student distribution, and a correlation matrix R, $T_{R,v}^d$ is defined by:

   $$T_{R,v}^d = \int_{-\infty}^{x_1} ... \int_{-\infty}^{x_d} \frac{\Gamma(\frac{v+d}{2})}{\Gamma(\frac{v}{2})(2\pi)^{\frac{d}{2}} |R|^{\frac{1}{2}}} (1 + \frac{(X-\mu)^t R^{-1} (X-\mu)}{v})^{-\frac{v+d}{2}} dx_1 ... dx_d,$$

   so the Student copula is defined by

   $$C_{v,R}^{d,T}(u) = \int_{-\infty}^{t_v^{-1}(u_1)} ... \int_{-\infty}^{t_v^{-1}(u_d)} \frac{\Gamma(\frac{v+d}{2})}{\Gamma(\frac{v}{2})(2\pi)^{\frac{d}{2}} |R|^{\frac{1}{2}}} \left(1 + \frac{(X-\mu)^t R^{-1} (X-\mu)}{v}\right)^{-\frac{v+d}{2}} dx_1 ... dx_d,$$

   Such that $t_v^{-1}$is the generalized inverse of the distribution function of the univariate Student distribution of v freedom degree. The density of $C_{v,R}^{d,T}(u)$ is given by:

   $$c_{v,R}^{d,T}(u) = |R|^{-\frac{1}{2}} \frac{\Gamma(\frac{v+d}{2})}{\Gamma(\frac{v}{2})} \left(\frac{\Gamma(\frac{v}{2})}{\Gamma(\frac{v+1}{2})}\right)^d \left(\frac{\left(1 + \frac{1}{v} \varsigma^t R^{-1} \varsigma\right)^{-\frac{v+d}{2}}}{\prod_{j=1}^{d} \left(1 + \frac{\varsigma_j^2}{v}\right)^{-\frac{v+1}{2}}} \right),$$

   where $\varsigma_j^2 = t_v^{-1}(u_j)$.

4. **Archimedean Copula**

   As the bivariate case the multivariate Archimidean copulas is defined by a generator $\varphi(t)$. The general form of this family of copulas is defined by:

$$C^{A,d}(u) = \varphi^{-1}(\varphi(u_1) + ... + \varphi(u_d)),$$

   and their density is given by

$$c^{A,d}(u) = \varphi^{-1}(\varphi(u_1) + ... + \varphi(u_d))\dot{\varphi}(u_1)..\dot{\varphi}(u_d)),$$

   The three most important and useful families of multivariate Archimedean copulas are now shown.

   (a) **Clayton Copula:**

$$C_\alpha^{d,Cla}(u) = \left(1 - d + \sum_{i=1}^{d} u_i^{-\alpha}\right)^{\frac{-1}{\alpha}},\ \alpha > 0.$$

   $C_\alpha^{Cla} = \Pi(u)$ when $\alpha = 0$.

   (b) **Frank Copula:**

$$C_\alpha^{d,Fr}(u) = -\frac{1}{\alpha}\log\left\{1 + \frac{\prod_{i=1}^{d}(\exp(-\alpha u_i) - 1)}{(\exp(-\alpha - 1))^{d-1}}\right\}, \quad \alpha > 0.$$

   (c) **Gumbel Copula:**

$$C_\alpha^{d,Gum}(u) = \exp\left\{-\left[\sum_{i=1}^{d}(-\log(u_i))^\alpha\right]^{\frac{1}{\alpha}}\right\},$$

   such that $\alpha \in [1, \infty[$ and we have

- $C_\alpha^{d,Gum}(u) = \Pi^d(u)$ when $\alpha \to 1$.
- $C_\alpha^{d,Gum}(u) = M^d(u)$ when $\alpha \to \infty$.

5. **Copula of extreme values**

   Let $X_i = (X_{i1}, ..., X_{id})$, $i \in \{1, ...n\}$ a sample of random vectors i.i.d with joint distribution function H and whose marginals are $F_1, ..., F_d$. The Copula $C_{(n)}$ of $M_{j,n}$ is defined by:

$$C_{(n)}(u) = C^n(u_1^{\frac{1}{n}}, ..., u_d^{\frac{1}{n}})^n, (u_1, ..., u_d) \in [0,1]^d,$$

   such that

$$M_{j,n} = \max(X_{1,j}, ...X_{n,j}),\ j = 1, ..., d.$$

Definition 2.3.9 *A Copula C^* is a copula of multivariate extreme values, if there exists a copula C such that:*

$$C^*(u) = \lim_{n\to\infty} C^n(u_1^{\frac{1}{n}}, ..., u_d^{\frac{1}{n}})^n.$$

- **Multivariate extreme value distribution**

The multidimensional theory of extreme values is concerned with the limit distribution

$$\lim_{n\to\infty} P(\frac{M_{1,n}-b_{1,n}}{a_{1,n}} \leq x_1, ..., \frac{M_{d,n}-b_{d,n}}{a_{d,n}} \leq x_d) = G_\theta(x_1, ..., x_d),$$

where G is a non-degenerate distribution if and only if the marginal distributions of G are univariate extreme value distributions.

Theorem 2.3.7 *For any copula of multivariate extreme values C^*, there is a convex function A defined from $\triangle_{d-1} = \left\{ (w_1, ..., w_d) \in [0, \infty[^d : \sum_{j=1}^{d} = 1 \right\}$ in $\left[\frac{1}{d}; 1\right]$, such that*

$$C^*(u) = \exp\left[\left(\sum_{j=1}^{d} \ln(u_j)\right) A\left(\frac{\ln u_1}{\sum_{j=1}^{d} \ln(u_j)}, ..., \frac{\ln u_d}{\sum_{j=1}^{d} \ln(u_j)}\right)\right].$$

In addition, A checks

$$\max(w_1, ..., w_d) < A(w_1, ..., w_d) < 1.$$

There is a new family of copulas defined by Capéraà et al. (2000) [12], the archimax family of copulas, this family is made up of both Archimidean and of extreme values at the same time.

Copula and dependence 3

Sommaire

The relationship between dependency measures and copulas seen in Chapter 2, is the subject of this chapter. We look at how copulas can be employed in the study of random variable dependency or association.

3.1 Association measures

To couple two or several multivariate distribution functions which are composed of identically distributed marginal distributions, it is necessary to measure the dependence between the margins. This is done from the measurement of a partial order between the pairs of data making up the observations.

3.1.1 Concordance measures

Let (x_i, y_i) and (x_j, y_j) two observations of a couple of random variables (X, Y), where $i \in \{1, ..., n\}$ and $j \in \{1, ..., n\}$.

Definition 3.1.1 *We say that (x_i, y_i) and (x_j, y_j) are concordant if and only if*

$$(x_i - x_j)(y_i - y_j) > 0 \Longleftrightarrow (x_i < x_j \text{ and } y_i < y_j) \text{ or } (x_i > x_j \text{ and } y_i > y_j).$$

We say that (x_i, y_i) and (x_j, y_j) are discordant if and only if

$$(x_i - x_j)(y_i - y_j) < 0 \Longleftrightarrow (x_i < x_j \text{ and } y_i > y_j) \text{ or } (x_i > x_j \text{ and } y_i < y_j).$$

Definition 3.1.2 *A numeric measure κ association between two random variables (X, Y) whose copula is C said to be a measure of concordance if and only if it satisfies the following properties:*

1. $\kappa_{X,Y}$ is defined for each continuous random variables couple (X, Y).
2. $-1 \leq \kappa_{X,Y} \leq 1$, where $\kappa_{X,X} = 1$ and $\kappa_{X,-X} = -1$.
3. $\kappa_{X,Y} = \kappa_{Y,X}$.
4. If X and Y are independent, then $\kappa_{X,Y} = 0$.
5. If the respective copulas of (X_1, Y_1) and (X_2, Y_2) are such that $C_1 \prec C_2$, then $\kappa_{X_1,Y_2} \leq \kappa_{X_2,Y_2}$.
6. $\kappa_{-X,Y} = \kappa_{X,-Y} = -\kappa_{X,Y}$.
7. If (X_n, Y_n) is a sequence of continuous random variables, where the copula $\{C_n\}$ converges pointwise to C, then $\lim_{n\to\infty} \kappa_{X_n,Y_n} = \kappa_{X,Y}$.
8. If $\alpha(X)$ and $\beta(Y)$ are strictly increasing functions, then $\kappa_{\alpha(X),\beta(Y)} = \kappa_{X,Y}$.

Theorem 3.1.1 *Let (X_1, Y_1) and (X_2, Y_2) two independent random vectors of continuous random variables whose joint distribution functions are H_1 and H_2 respectively. Such that X_1 and X_2 have the same distribution function F_1, by the way Y_1 and Y_2 have the same distribution function F_2. Let C_1 and C_2 the copulas associated to (X_1, Y_1) and (X_2, Y_2), respectively. Suth that:*

$$H_1(x, y) = C_1\left(F_1(x), F_2(y)\right)$$
$$H_2(x, y) = C_2\left(F_1(x), F_2(y)\right).$$

Consider Q as the difference between the concordance probability and that of the discordant of (X_1, Y_1) and (X_2, Y_2), i.e.

$$Q = P\{(X_1 - X_2)(Y_1 - Y_2) > 0\} - P\{(X_1 - X_2)(Y_1 - Y_2) < 0\}, \quad (3.1.1)$$

then

$$Q = Q(C_1, C_2) = 4\int\int_{[0,1]^2} C_2(u,v)dC_1(u,v) - 1. \quad (3.1.2)$$

Proof. Since all random variables in this case are continuous, then

$$P\{(X_1 - X_2)(Y_1 - Y_2) < 0\} = 1 - P\{(X_1 - X_2)(Y_1 - Y_2) \geq 0\},$$

then $Q = 2P\{(X_1 - X_2)(Y_1 - Y_2) > 0\} - 1$.

However,

$$P\{(X_1 - X_2)(Y_1 - Y_2) > 0\} = P(X_1 > X_2, Y_1 > Y_2) + P(X_1 < X_2, Y_1 < Y_2).$$

These probabilities can be evaluated by integration on one of the distribution functions of (X_1, Y_1) or (X_2, Y_2). Integrate with respect to that of (X_1, Y_1)

$$\begin{aligned} P(X_1 > X_2, Y_1 > Y_2) &= \int\int_{\mathbb{R}^2} P(X_2 < x, Y_2 < y)\, dC_1(F_1(x), F_2(y)) \\ &= \int\int_{\mathbb{R}^2} C_2(F_1(x), F_2(y))dC_1(F_1(x), F_2(y)), \end{aligned}$$

by changing variables, $u = F_1(x)$ and $v = F_2(y)$, we obtain

$$P(X_1 > X_2, Y_1 > Y_2) = \int\int_{[0,1]^2} C_2(u,v)dC_1(u,v).$$

In a similar way

$$\begin{aligned} P(X_1 < X_2, Y_1 < Y_2) &= \int\int_{\mathbb{R}^2} P(X_2 > x, Y_2 > y)\, dC_1(F_1(x), F_2(y)) \\ &= \int\int_{\mathbb{R}^2} (1 - F_1(x) - F_2(y) + C_2(F_1(x), F_2(y)))\, dC_1(F_1(x), F_2(y)) \\ &= \int\int_{[0,1]^2} (1 - u - v + C_2(u,v))dC_1(u,v), \end{aligned}$$

Now since C_1 is a joint distribution function of two random variables (U, V) uniforms. So $E(U) = E(V) = \frac{1}{2}$, therefore

$$P(X_1 < X_2, Y_1 < Y_2) = 1 - \frac{1}{2} - \frac{1}{2} + \int\int_{[0,1]^2} C_2(u,v)dC_1(u,v),$$

finally

$$P\{(X_1 - X_2)(Y_1 - Y_2) > 0\} = 2\int\int_{[0,1]^2} C_2(u,v)dC_1(u,v).$$

By grouping these results we deduce

$$Q = Q(C_1, C_2) = 4\int\int_{[0,1]^2} C_2(u,v)dC_1(u,v) - 1.$$

□

Corollary 3.1.1 *Let C_1, C_2 two copulas and either Q a measure of agreement and disagreement (3.1.2), Q has the following properties*

1. Q is symmetrical
$$Q(C_1, C_2) = Q(C_2, C_1).$$
2. Q is non-decreasing

   if $C_1 \leq \grave{C}_1$ and $C_2 \leq \grave{C}_2$ for all $(u, v) \in \mathbb{I}^2$, then $Q(C_1, C_2) \leq Q(\grave{C}_1, \grave{C}_2)$.
3. We can replace the copula C by the survival copula $\hat{C}$, because
$$Q(C_1, C_2) = Q(\hat{C}_1, \hat{C}_2).$$

Example 6 *We calculate the measure Q for the copulas M, W and Π. The support of M is the diagonal $u = v$ in $\mathbb{I}^2$, and since M has uniform margins, it follows that if g is an integrable function whose domain is $\mathbb{I}^2$, then*

$$\int \int_{I^2} g(u, v) dM(u, v) = \int_0^1 g(u, u) du.$$

Therefore, we have

$$\begin{aligned} Q(M, M) &= 4 \int \int_{I^2} \min(u, v) dM(u, v) - 1 \\ &= 4 \int_0^1 u du - 1 \\ &= 1 \end{aligned}$$

$$\begin{aligned} Q(M, \Pi) &= 4 \int \int_{I^2} uv dM(u, v) - 1 \\ &= 4 \int_0^1 u^2 du - 1 \\ &= \frac{1}{3} \end{aligned}$$

$$\begin{aligned} Q(M, W) &= 4 \int \int_{I^2} \max(u + v - 1, 0) dM(u, v) - 1 \\ &= 4 \int_0^1 (2u - 1) du - 1 \\ &= 0 \end{aligned}$$

Likewise, because the support of W is the diagonal $v = 1 - u$, we have

$$\int \int_{\mathbb{I}^2} g(u, v) dW(u, v) = \int_0^1 g(u, 1 - u) du$$

Then

$$\begin{aligned} Q(W, \Pi) &= 4 \int \int_{I^2} uv dW(u, v) - 1 \\ &= 4 \int_0^1 u(1 - u) du - 1 \\ &= -\frac{1}{3} \end{aligned}$$

$$\begin{aligned} Q(W,W) &= 4\int\int_{I^2} \max(u+v-1,0)dW(u,v) - 1 \\ &= 4\int_0^1 0du - 1 \\ &= -1 \end{aligned}$$

For the copula Π, we have $d\Pi(u,v) = dudv$, then

$$\begin{aligned} Q(\Pi,\Pi) &= 4\int\int_{I^2} uvd\Pi(u,v) - 1 \\ &= 4\int_0^1\int_0^1 uvdudv - 1 \\ &= 0 \end{aligned}$$

3.1.2 Kendall's Tau

Let (X_1, Y_1) and (X_2, Y_2) two continuous, independent and identically distributed random vectors of joint distribution functions F. Kendall's tau $\tau_{X,Y}$ of the random vector (X, Y), is defined by:

$$\tau_{X,Y} = P\left\{(X_1 - X_2)(Y_1 - Y_2) > 0\right\} - P\left\{(X_1 - X_2)(Y_1 - Y_2) < 0\right\}. \tag{3.1.3}$$

We can define the Kendall's tau as a function of a copula C, using function Q defined in (3.1.2). The following theorem represents the relationship between Kendall's tau and copulas.

Theorem 3.1.2 *Let X, Y two continuous random variables whose copula is C. The Kendall's tau of X and Y is defined by:*

$$\tau_{X,Y} = Q(C,C) = 4\int\int_{\mathbb{I}^2} C(u,v)dC(u,v) - 1 \tag{3.1.4}$$

Since the random variables $U = F(x)$ and $V = G(y)$ are uniform random variables, then equation (3.1.4) becomes:

$$\tau_{X,Y} = 4E(C(U,V)) - 1.$$

Example 7 *Let C_θ a Farlie-Gumbel-Morgenstern copula, of parameter $\theta \in [-1;1]$, since C_θ is absolutely continue, then*

$$dC_\theta(u,v) = \frac{\partial^2 C_\theta(u,v)}{\partial u \partial v}dudv = 1 + \theta(1-2u)(1-2v)\,dudv$$

So $\int\int_{\mathbb{I}^2} C_\theta(u,v)dudv = \frac{1}{4} + \frac{\theta}{18}$ and $\tau_\theta = \frac{2\theta}{9}$. We present in the following table some copulas and their corresponding kendall's tau, such that $D_k(\theta) = \int_0^\theta \frac{x}{\theta}/(e^x - 1)dx$ is a Debye function.

Copula	Kendall's tau
Normale	$2\pi^{-1}\arcsin(\rho)$
Gumbel	$(\theta-1)/\theta$
Frank	$\frac{1-4}{\theta}+\frac{4D_k(\theta)}{\theta}$
Clayton	$\theta/(\theta+2)$

Table 3.1 – *Kendall's tau for some copulas.*

3.1.3 Spearman's Rho

Let (X_1, Y_1), (X_2, Y_2) and (X_3, Y_3) three independent copies of random vector (X, Y) . The Spearman's Rho, noted by $\rho_{X,Y}$, is defined by:

$$\rho_{X,Y} = 3(P((X_1 - X_2)(Y_1 - Y_3) > 0) - P((X_1 - X_2)(Y_1 - Y_3) < 0)) \tag{2.12}$$

We can define Spearman's rho as a function of a copula C, almost as Kendall's tau.

Theorem 3.1.3 *Let X, Y two continuous random variables whose copula is C. The Spearman's Rho of X and Y is defined by:*

$$\begin{aligned}\rho_{X,Y} &= 3Q(C,\Pi),\\ &= 12\int\int_{\mathbb{I}^2} uvdC(u,v) - 3\\ &= 12\int\int_{\mathbb{I}^2} C(u,v)dudv - 3\end{aligned}$$

Because the variables U, V are uniform, where $E(U) = E(V) = \frac{1}{2}$, with variance $var(U) = var(V) = \frac{1}{12}$, then $\rho_{X,Y}$ can be written by:

$$\rho_{X,Y} = \frac{E(UV) - E(U)E(V)}{\sqrt{var(U)\sqrt{var(V)}}} \tag{2.13}$$

Proof. We have $E(UV) = \int\int_{\mathbb{I}^2} uvdC(u,v)$, then

$$\begin{aligned}\rho_{X,Y} &= 12\int\int_{I^2} uvdC(u,v) - 3\\ &= 12E(UV) - 3\\ &= \frac{E(UV) - 1/4}{1/12}\\ &= \frac{E(UV) - E(U)E(V)}{\sqrt{var(U)\sqrt{var(V)}}}\end{aligned}$$

□

Example 8 *Because Sparman's rho is defined as a function of a parametric copula C_θ we can noted $\rho_{X,Y}$ by ρ_θ.*

- Let C_θ a Farlie-Gumbel-Morgenstern copula, of parameter $\theta \in [-1;1]$, then

$$C_\theta(u,v) = uv + \theta uv(1-u)(1-v),$$

so $\int\int_{\mathbb{I}^2} C_\theta(u,v)dudv = \frac{1}{4} + \frac{\theta}{36}$. The Spearman's Rho is then $\rho_\theta = \frac{\theta}{3}$.

- Let $C_{\alpha,\beta}$ the copula of Marshall Olkin, of parameters $0 < \alpha$ and $\beta < 1$ defined by:

$$C_{\alpha,\beta}(u,v) = \begin{cases} u^{1-\alpha}v, & u^\alpha \geq v^\beta \\ uv^{1-\beta}, & u^\alpha \leq v^\beta \end{cases},$$

then

$$\int\int_{\mathbb{I}^2} C_{\alpha,\beta}(u,v)dudv = \frac{1}{2}\left(\frac{\alpha+\beta}{2\alpha-\alpha\beta+2\beta}\right),$$

the Spearman's Rho is $\rho_{\alpha,\beta} = \frac{3\alpha\beta}{2\alpha-\alpha\beta+2\beta}$.

3.2 Dependence measure

Definition 3.2.1 *A numerical measure of association δ (we note it $\delta_{X,Y}$) between two continuous random variables X, Y whose copula is C is said to be a dependency measure if and only if it satisfies the following properties:*

1. $\delta_{X,Y}$ is defined for each couple (X, Y) of continuous random variables.
2. $0 \leq \delta_{X,Y} \leq 1$.
3. $\delta_{X,Y} = \delta_{Y,X}$.
4. $\delta_{X,Y} = 0$ if and only if X and Y are independent.
5. $\delta_{X,Y} = 1$ if and only if each of X and Y is a strictly monotonic function of the other almost certainly.
6. If $\alpha(X)$ and $\beta(Y)$ are almost certainly strictly monotonic functions, so $\delta_{\alpha(X),\beta(Y)} = \delta_{X,Y}$.
7. If (X_n, Y_n) is a sequence of continuous random variables, where the copula $\{C_n\}$ converges pointwise to C, then $\lim_{n\to\infty} \delta_{X_n,Y_n} = \delta_{X,Y}$.

Example 9 (Schweizer and Wolffs σ measure) *Sperman's Rho of two continuous random variables X and Y is defined by $\rho_{X,Y} = 12\int\int_{\mathbb{I}^2}(C(u,v) - uv)dudv$. This integral represents the volume between the copula C and the copula produced Π. If we change the difference $(C(u,v) - uv)$ by $|C(u,v) - uv|$, then we get a measurement based on the distance L_1 between the graph of C and Π, this distance represents the measurement σ of Schweizer and Wolffs, and it is defined by:*

$$\sigma_C = \sigma_{X,Y} = 12\int\int_{\mathbb{I}^2} |C(u,v) - uv|\, dudv. \tag{3.2.1}$$

Theorem 3.2.1 *Let a continuous random variables X and Y and a copula C. The quantity σ_C defined in (3.2.1) is a dependency measure. Schweizer and Wolffs (1981)[81], assures that any distances between surfaces $z = C(u,v)$ and $z = uv$ represent a non-parametric measure of dependence.*

So $\forall\ 1 \leq p < \infty$, the distances L_p between C and Π is defined by:

$$L_p = \left(k_p \int\int_{\mathbb{I}^2} |C(u,v) - uv|^p\, dudv\right)^{\frac{1}{p}}, \tag{3.2.2}$$

such that k_p is a constant. From the quantity (3.2.2) we can define the following dependency measures:

- **The measure $\Phi_{X,Y}$**

$$\text{If } p = 2, \text{ then } \quad \Phi_{X,Y} = \Phi_C = \left(90 \int \int_{I^2} |C(u,v) - uv|^2 \, dudv\right)^{\frac{1}{2}} \tag{3.2.3}$$

$\Phi^2_{X,Y}$ represents the dependence index between the variables X and Y.

- **The measure $\Lambda_{X,Y}$**

$$\text{For } \ p = \infty, \ \Lambda_{X,Y} = \Lambda_C = 4 \sup_{u,v \in I} |C(u,v) - uv| \tag{3.2.4}$$

Remark 3.2.1 *After the definitions of dependency measures σ_C, Φ_C and Λ_C the latter are found to be based on Spearman's Rho coefficient. There are other measures of dependence based on another coefficient such as the Gini coefficient (see Nelsen, 2006, p.211).*

3.2.1 Tail dependency

Tail dependency is a local measure because it measures the dependence at the level of the distribution tails. There are two tail dependency coefficients defined as follows:

Definition 3.2.2 *Let X and Y two continuous random variables have respectively a distribution functions F_1 and F_2. The lower tail dependence coefficient λ_L of X and Y is defined as:*

$$\lambda_L(X,Y) = \lim_{\alpha \to 0^+} P\left(X \leq F^{-1}(\alpha) / Y \leq G^{-1}(\alpha)\right). \tag{3.2.5}$$

The upper tail dependence coefficient λ_u is by the way defined as:

$$\lambda_U(X,Y) = \lim_{\alpha \to 1^-} P\left(X > F^{-1}(\alpha) / Y > G^{-1}(\alpha)\right). \tag{3.2.6}$$

We can define these measures according to a copula C.

Definition 3.2.3 *Let X and Y two continuous random variables of copula C, then we have*

$$\lambda_L(X,Y) = \lim_{u \to 0^+} \frac{C(u,u)}{u}, \tag{3.2.7}$$

- *When $\lambda_L \in]0,1]$, then C has a lower tail dependency.*
- *When $\lambda_L = 0$, then C has no lower tail dependency.*

$$\lambda_U(X,Y) = \lim_{u \to 1^-} \frac{1 - 2u + C(u,u)}{1-u}$$

- *When $\lambda_U \in]0,1]$ then C has a upper tail dependency.*
- *When $\lambda_L = 0$ then C has no upper tail dependency.*

We now present some copulas with the dependency coefficients of tails, if they exist in the following table:

	Copula $C(u,v)$	λ_L	λ_U
Archimedean $C_\alpha(u,v)$	Clayton	$2^{-\frac{1}{\alpha}}$	0
	Gumbel	0	$2-2^{\frac{1}{\alpha}}$
	Frank	0	0
$C_\rho(u,v)$	Gaussian	0	0
$C_{\alpha,\beta}(u,v)$	Marchall-Olkin	0	$\min(\alpha,\beta)$

Table 3.2 – *Tail dependency coefficients of some copulas.*

Remark 3.2.2

- *The Clayton's copula has a lower tail dependency, but Gumbel's copula has an upper tail dependency.*
- *The Gaussian copula has no tail dependency, except for $\rho = 1$, such that*

$$\lambda_L = \lambda_U = \begin{cases} 0 & \text{if} \quad \rho < 1 \\ 1 & \text{if} \quad \rho = 1 \end{cases}$$

- *Frank's copula has no neither inferior nor superior tail dependency such as the Gaussian copula.*

Survival Analysis and Copulas

4

Sommaire

Surviving data analysis is a discipline of statistics concerned with the modeling of "lifetimes." It is generally the time elapsed between an origin date and an occurrence date events that generally correspond to the onset of illness, death, relapse, etc. The probability that an individual is alive or unscathed beyond time t is given by the survival function. When several events are involved simultaneously, we speak of multivariate survival.

The survival time and its aspects are presented in this chapter, as well as the established results. By the way, the notion of survival analysis is introduced.

4.1 Survival time notion

The term "Survival time" called also the lifetimes refers to the time elapsed until a certain event occurs. Is the time elapsed between an origin date t_0 and the occurrence date t of the event. This can represent an illness, a relapse, a cure, a machine breakdown, a claim, and so on. We evaluate the distribution of the variable of interest in survival analysis, which is an area of statistics concerned with the modeling of a lifetime. When the variable of interest T is continuous, one of the five identical functions can be used to define or describe this distribution:

- the distribution function F,
- the survival function S,
- the density function f,
- the instantaneous hazard function λ,
- the cumulative hazard function Λ.

In survival analysis, the value of mean survival and median survival may also be of relevance. In survival time theory these functions can be defined in terms of survival time as a sequence:

- **The distribution function**

  The survival variable has a distribution function as any other continuous random variable. This distribution function, is at time t, the probability that the event takes place before the date t.

- **The survival function**

  For a fixed time t, is the probability that the event occurs after t, or the probability that an individual will live beyond a date t.

- **The density function**

  Is the probability of the event occurring within a small period of time after instant t. It is defined by:

$$f(t) = \lim_{\Delta_t \to 0} \frac{P(t \leq T < t + \Delta_t)}{\Delta_t}$$

- **The instantaneous hazard function**

  In the modeling of the survival function, a fundamental concept is that of the hazard function or risk function $\lambda(t)$ for a fixed instant t. It is the probability of the event occurring in a small time interval after t, conditional on the fact that it does not take place until t. Is defined by:

$$\lambda(t) = \lim_{\Delta_t \to 0} \frac{P(t \leq T < t + \Delta_t | T \geq t)}{\Delta_t} \tag{4.1.1}$$

- **The cumulative hazard function**

  The cumulative hazard function (Andersen, Borgan, Gill et Keiding (1993)) [3], or integrated hazar function (Hougaard [1999]) [48], often noted $\Lambda(t)$ is the integral of the instantaneous risk and is given by:

$$\Lambda(t) = \int_0^t h(x)\, dx \tag{4.1.2}$$

In the coming chapters, we will explore these definitions in considerable detail.

4.2 Incomplete data

One of the criteria of survival data is the existence of incomplete observations. Because of the censoring and truncation processes data is frequently collected in partially. Censored or truncated data results from not having access to all the information. Instead of observing independent and identically distributed realizations of duration Y, we observe the realization of the variable Y subject to various disturbances, whether or not independent of the event studied. Censoring and truncation procedures can both be present simultaneously.

4.2.1 Truncated notion

The truncation prevents entirely the observation of the variable Y (in most cases, the extreme values), and leads to a loss of information (only a subsample). It is said that there is:

- **Left truncation**

  When Y is only observable if it is greater than one fixed or random positive C threshold. This is a pattern that first appeared in astronomy, where samples are composed of astral objects of a certain zoned. The absolute and apparent luminosities of an astral object are respectively defined as its brightness observed at a fixed distance and from then the earth and we only observe objects that are sufficiently shiny, that is to say those for where the luminosity $M \geq m$, m being the truncation variable. In this case, we have N objects in the sample, but we are unable to observe that the n sufficiently shiny objects.

- **Right trancattion**

  When Y is only visible, if it is smaller than C.

- **Interval truncation**

  when Y is truncated on the right and on the left. This type of truncation is encountered when studying patients in a registry: patients diagnosed before setting up the registry or listed after consultation of the registry will not be included in the study. Lyndell-Bell [21], proposed a non-parametric estimate of the distribution function of Y within the framework of the truncation model and the asymptotic properties: the strong law and asymptotic normality were studied by Woodroofe [66].

4.2.2 Censoring notion

In reality, having a sample with complete data is sometimes not available. Censorship is one of the most frequent phenomena at the origin of incomplete data in statistics. A data is said to be "censored" if the exact value is unknown, but only an estimate, lower or higher, that is to say rough information of the type $T \geq C$ or $T \leq C$. Such information is very poor, poorer than saying "T is between a and b ", since only one of the two bounds is known. In the analysis of survival times, censoring occurs when the survival T is only known for some of the individuals "the data for which survival is unknown are said censored". The variable of interest T is not observed and it is limited superially or inferiorly by a variable (of censoring, generally noted C) which has been observed.

Given that in biostatistics and epidemiology, the main focus of the studies is the explanation for the occurrence of an event of interest (death, rejection of a transplant, end of study, withdrawal from study, loss of follow-up, etc), all available information must be analyzed. However, due to the fact that the phenomenon of censoring is in itself a special case incompleteness of the data, observational studies only very rarely present complete data when within a framework of survival analysis. Thus, it is necessary, for the clinician be quick to use statistical methods that take into account the censored data.

In addition, censoring can be informative or non-informative: in the event of censoring informative, there is a dependence between the survival time and the censoring time. We take the example of a patient lost to follow-up: his voluntary withdrawal may, for example, result from the fact that the patient is near death or decides to stop treatment to die in a certain time dignity, its censoring is then dependent on the time of death. For an individual i, we consider:

- its survival time T_i.
- its censoring time C_i.
- the time actually observed X_i.

1. **Right-censoring**

   The variable of interest is said to be right-censoring if the individual concerned have no information on his last sighting. In this model of censoring, we observe the couple $(Z;\delta)$, where Z is the observed duration and δ is a binary variable which presenting the nature of this duration and takes the value 1 if the variable is observed and 0 if it is censoring. This model is the most common in practice, it is for example adapted to the case where the event of interest is the survival time to a disease and where the end date of the study is previously fixed; patients alive at the end of the study provide right-censored data. Censoring is not necessarily fixed, it can be random, this is the case for example of an individual lost to follow-up or died in an accident in study course. Or is when the event considered is the death of a sick patient and the duration of observation is a total duration of hospitalization. Or in a therapeutic trial, this censoring

can be caused by loss of sight. The censoring variable C in this case represents the date on which an individual leaves the study for a cause other than death.

2. **Left-censoring**

   The survival time is said to be left-censoring when the individual has already undergone the event before to be observed. In this case we do not know the survival time but we only know that it is less than a certain known date. A well-known example of this type of censoring looks at the time when baboons come down from trees to eat. The event of interest (descent from the tree) is observed for baboons who descended after the arrival of observers and is censored for those who went down before arrival observers. We find also this kind of phenomenon in reliability studies when the failure of an electronic device or component do not allow to continue observation for another device or component.

3. **Interval censoring**

   A date is interval-censored if instead of observing the time of the event, the only information available is that it occurred between two known dates. We find this model usually in follow-up studies where patients are monitored periodically. For example, in the case of cohort follow-up, people are often followed intermittently (not continuously), then we only know that the event has occurred produced between these two observation times.

4. **Double (or mixed) censoring**

   If data is censored on both on the right and left sides, it is said to be mixed censored. There are several non-parametric models that deal with this kind of data. For example, the models Morales et al.[65], Patilea and Rolin (2006)[75], and Turnbull (1974)[88], which is the most used, and several works are based on this model.

These four categories described above can arise according to the mode or mechanism of censoring. Thus, in the literature we find the following types:

- **Censoring of type I (fixed)**

  The observer fixed a value (for example a non-random end of experiment date). If we choose the right-censoring, despite of looking at the variables $T_1, T_2, ..., T_n$ which interest us, we observe T_i when it is less than a fixed duration C. Otherwise we only know that T_i is greater than C. Then, we observe a variable $Z_i = \min(T_i, C)$, $i = 1, ..., n$. For example, in the industrial domain, when observing the survival time of an electronic component during a time interval of $[0; C]$.

- **Censoring of type II (waiting, until the k^{th} dies)**

  We observe the life durations of n patients until k of them have died and we stop at this moment. If we order the variables $T_1, T_2, ...T_k, ..., T_n$, we get the order statistics $T_{(1)}, T_{(2)}, ...T_{(k)}, ..., T_{(n)}$.

The date of censoring is then T_{rk} and only the first k times are observed. In this case, we are talking about type **II** of censoring. For example, to test the reliability of a complex system we put n systems of the same type into working order and we stop when the k^{th} failure is observed.

- **Censoring of type III (randomly)**

  Censoring is generally not controlled and is also a random variable. Thus, for each individual $i = 1, 2, ..., n$ is associated a couple of random variable durations (T_i, C_i), we have what is actually observed X_i is an indicator which is equal to 1 if the event is observed and 0 if it is censored. In a therapeutic trial for example, this censoring can be caused by loss of sight, death, etc. In other words, let $T_1, T_2, ..., T_n$ a sample of a positive random variable T, we say that there is a random censoring of this sample if there is another positive random variable C of a sample $C_1, C_2, ..., C_n$, where in this case instead of observing the T_i, we observe a couple of variables (Z_i, δ_i), where

  $$Z_i = \min(T_i, C_i) \text{ and } \delta_i = I_{T_i \leq C_i} \text{ for } i = 1, 2, ..., n.$$

  and $\delta_i = I_{T_i \leq C_i}$ represents the indicator function of censored data, which specifies if our variable of interest is observed or not. For example in the right-censoring, we only observe the variable
  $Z_i = \min(T_i, C_i)$, if $T_i < C_i$ in this case $\delta_i = 1$ and the duration of interest is observed $(Z_i = T_i)$.
  Otherwise, if $T_i \geq C_i$ the variable in this case is censored $(Z_i = C_i)$ and the indicator equal to $\delta_i = 0$, i.e. we observe incomplete durations.
  By the way, we say there is a random left-censoring instead of observation $T_1, T_2, ..., T_n$ we observe the copule $Z_i = \max(T_i, C_i)$ where $\delta_i = I_{T_i \geq C_i}$ for $i = 1, 2, ..., n$.

In this thesis, we only interested by the case of right-censoring of the random type. The following are the primary estimators that play a significant role in the censored data framework.

4.3 Semi-parametric estimation for Copula models

4.3.1 Maximum Likelihood Estimation (MLE)

Assuming a multivariate parametric copula C_θ, where $\theta = (\theta_1, ..., \theta_d) \in \Theta$ be the vector of copula parameters and β be the vector of marginal parameters. Given the relatively simple functional form the self-selection likelihood function under an Archimedean copula, MLE can be employed to jointly estimate all parameters of the unknown parameters vector $(\beta_1, ..., \beta_d, \theta)$ at the same time. Assume that we observe d-independent realizations $(X_{i1}, \ldots, X_{ip}), i = 1, \ldots, d$, specified by p-margins with cumulative distribution function (CDF) F_i. However, the density of F is given by:

$$f(x_1, ..., x_d; \theta) = c_\theta \left[\left(F_{1,\beta_1}(x_1), ..., F_{d,\beta_d}(x_d) \right); \theta \right] \prod_{i=1}^{d} f_{i,\beta_i}(x_i) \tag{4.3.1}$$

That is associated with a sample $(X_{i1}, \ldots, X_{ip})_{i=1,\ldots,d}$, where c_θ is a density of a parametric copula C_θ and f_{i,β_i} is a density of F_{i,β_i}. A parametric and a semi-parametric approaches both presented seek to maximize a likelihood approximation based on (4.3.1). Consequently, the parameter vector to be estimated in the parametric approach is $\alpha = (\beta, \theta)$ and by maximizing the log-likelihood function $L(\beta_1, ..., \beta_d; \theta)$ defined by:

$$L(\beta_1, ..., \beta_d; \theta) = \sum_{i=1}^{n} \log f\left(x_1, ..., x_d; \theta\right) \tag{4.3.2}$$

$$\begin{aligned} L(\beta_1, ..., \beta_d; \theta) &= \sum_{i=1}^{n} \log c_\theta \left((F_{1,\beta_1}(x_1), ..., F_{d,\beta_d}(x_d); \theta) \prod_{j=1}^{d} f_{j,\beta_j}(x_j) \right) \\ &= \sum_{i=1}^{n} \log c_\theta((F_{1,\beta_1}(x_1), ..., F_{d,\beta_d}(x_d); \theta) + \sum_{i=1}^{n} \sum_{j=1}^{d} \log \prod_{j=1}^{d} f_{j,\beta_j}(x_{ij}), \end{aligned}$$

then the estimator of θ, noted $\hat{\theta}_n^{MV}$ is

$$\hat{\theta}_{MLE} = \arg\max L(\beta_1, ..., \beta_d; \theta)$$

See Lehmann and Casella [56], for more details. This estimator is consistent and satisfies the asymptotic normality property:

$$\sqrt{n}\left(\hat{\theta}_{MLE} - \theta\right) \to N(0, I^{-1}(\theta)),$$

such that $I(\theta)$ is the Fisher information matrix. This matrix is estimated by the inverse of the Hessian matrix of the likelihood function.

4.3.2 Margin Inference Function Method (IFM)

This method was introduced by Joe [49], in a general framework. He called the inference function method for margins (IFM) because estimation functions relate to likelihood functions (univariate or multivariate). The IFM was used primarily for multivariate models in which a multi-parameter numerical optimization for maximum likelihood estimation is unattainable or takes a too long time if we talk about a time-consuming viewpoint. Therefore, the estimation by the IFM method can be decreased the computational load potentially associated with the estimation of the maximum likelihood. Hence, this method proceeds in two stages:

- Stage 1: Estimate the margins parameters firstly.
- Stage 2: fix the marginal parameters obtained in the first stage, and then estimate the copula parameters.

See Joe [42], [49]. By analogy, for $j = 1, ..., d$ the unknown margins parameter vectors $(\beta_1, ..., \beta_d)$ are first estimated by:

$$\beta_{n,j} = \arg\sup \left(\sum_{i=1}^{n} \log f_{j,\beta_j}\left(X_{ij}\right) \right), \text{ where } \beta_j \in \mathbb{R}^{p_j}$$

By the way in the second stage, estimating the vector of the unknown copula parameter θ is performed by:

$$\hat{\theta}_{IFM} = \arg\sup \sum_{i=1}^{n} \log c_\theta((F_{1,\beta_1}(x_1), ..., F_{d,\beta_d}(x_d); \theta),$$

where $\theta \in \Theta$. The main advantage of this approach is that this method (IFM) is done in two stages. Wherever, under the MLE estimate, when models are usually multivariate, the number of parameters increases, and the numerical optimization becomes more complicated. Also, for some models multi-dimensional numerical integration is needed and this becomes increasing difficult for the theoretical properties of the IFM method (see [42]).

Again, the estimator θ satisfies the asymptotic normality property as Joe [49], has been shown:

$$\sqrt{n}\left(\hat{\theta}_{IFM} - \theta\right) \to N(0, V^{-1}(\theta)),$$

where $V(\theta)$ is the Godambe information matrix, defined by:

$$V(\theta) = D^{-1} M \left(D^{-1}\right)^t,$$

where

$$D = E\left[\frac{\partial}{\partial\theta}(g(\theta)^t)\right],\ M = E\left[g(\theta)^t g(\theta)\right] \text{ and } g(\theta) = \left(\frac{\partial}{\partial\gamma_1}L_1, ..., \frac{\partial}{\partial\gamma_d}L_d\right)$$

4.3.3 The Pseudo-maximum likelihood method (PML)

The Pseudo-maximum likelihood method (PML), was proposed in the case where the margins $F_1, ..., F_d$ associated with $X_1, ..., X_d$ are unknown, she does in two stages:

- We replace the margins $F_1, ..., F_d$ by their natural estimates (empirical estimator), they are defined by:

$$\hat{F}_{j,n}(x_{ij}) = \frac{1}{n}\sum_{i=1}^{n} 1\left(X_{ij} \leq x\right)$$

- By maximizing the pseudo log-likelihood to estimate θ, such that:

$$L(\theta) = \sum_{i=1}^{n} \ln c_\theta\left\{\hat{F}_{1,n}(x_{i1}), \hat{F}_{2,n}(x_{i2}), ..., \hat{F}_{d,n}(x_{id})\right\},$$

then the estimator $\hat{\theta}_{PML}$ is $\hat{\theta}_{PML} = \arg\sup L(\theta)$.

4.3.4 Moments Estimation method based on Kendall's Tau and Spearman's Rho

The estimation methods based on the correlation coefficients of Spearman's rho or Kendall's tau rank are called tau-inversion methods (respectively, rho-inversion) or concordance method (mentioned in part one).

Take advantage of the relationship between these coefficients of dependency and the parameter of the copula θ. Early references on the former in a copula setting are among others Oakes [70], Genest [34], Genest and Rivest [33]. This method consists of estimating the parameters sought using some measure of association such as Kendall's rate and Spearman's rho, where there is a relationship between these measures and the copula dependency parameter. Let (X, Y) a couple of random variables whose copula is C_θ, of parameter θ, such that $\theta \in \Theta \subset \mathbb{R}$.

- **Moment Estimator Based on Kendall's Tau**

  Assuming that there is a relationship between the Kendall's Tau and the parameter θ. This relation is defined by:

  $$\tau_{X,Y} = g(\theta), \tag{4.3.3}$$

  where g is a continuous and differentiable function. Then, an estimator $\hat{\theta}_n^{TK}$ of θ is defined by: $\hat{\theta}_n^{TK} = g^{-1}(\tau_n)$, such that τ_n is the empirical estimator of τ.

  This estimator is asymptotically normal

  $$\sqrt{n}(\hat{\theta}_{KT} - \theta) \to N(0, \hat{\sigma}_\tau^2),$$

  such that $\hat{\sigma}_\tau^2$ is the empirical variance of σ_τ^2

  $$\hat{\sigma}_\tau^2 = (4S_{\hat{g}}(\tau_n))^2,$$

  where

  $$\begin{cases} S^2 = \frac{1}{n}\sum_{i=1}^{n}\left(W_i + \tilde{W}_i - 2\bar{W}\right)^2 \\ W_i = \frac{1}{n}\sum_{j=1}^{n} 1\left(X_j \leq X_i, Y_j \leq Y_i\right) \\ \tilde{W}_i = \frac{1}{n}\sum_{j=1}^{n} 1\left(X_i \leq X_j, Y_i \leq Y_j\right) \end{cases}$$

- **Moment Estimator Based on Spearman's Rho**

  Similarly, if we assume that Spearman's Rho is defined as a function of θ by the following relation:

  $$\rho = h(\theta),$$

  where h is a continuous and differentiable function. Then the estimator $\hat{\theta}_n$ of θ is defined by:

  $$\theta = h^{-1}(\rho_n),$$

  such that, ρ_n is the empirical estimator of ρ. This estimator is asymptotically normal

  $$\sqrt{n}(\hat{\theta}_{RS} - \theta) \to N(0, (\sigma_n h^{-1}(\rho_n))^2),$$

  such that σ_n^2 is the estimator of σ^2

  $$\sigma_n^2 = 144(-9A_n^2 + B_n + 2C_n + 2D_n + 2\varepsilon_n),$$

where

$$A_n = \frac{1}{n}\sum_{i=1}^{n}\frac{R_i}{n+1}\frac{S_i}{n+1}$$

$$B_n = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{R_i}{n+1}\right)^2\left(\frac{S_i}{n+1}\right)^2$$

$$C_n = \frac{1}{n^3}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\frac{R_i}{n+1}\frac{S_i}{n+1}1(R_k \leq R_i, S_k \leq S_j) + \frac{1}{4} - A_n$$

$$D_n = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{S_i}{n+1}\frac{S_j}{n+1}\max\left(\frac{R_i}{n+1}, \frac{R_j}{n+1}\right) \text{ and}$$

$$\varepsilon_n = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{R_i}{n+1}\frac{R_j}{n+1}\max\left(\frac{S_i}{n+1}, \frac{S_j}{n+1}\right)$$

These all methods have been used extensively by many authors, their importance discussed in the introductory part. In this thesis, we are interested by the last one due to its simple mathematical form, where the data is right-censored.

4.4 Non-parametric estimation for right-censoring model

4.4.1 Kaplan-Meier Estimator

In the case of right-censoring, the empirical survival function of the variable T is no longer valid because since it involves unobserved quantities. In particular, estimating the distribution of a duration censored by the empirical distribution function was impossible. In order to estimate the T distribution, it was necessary to construct a survival function estimator in the presence of censored data. The non-parametric estimation problem of a right-censored random variable distribution function, was originally considered by Kaplan and Meier (1958) [51]. They provide a good estimator of the survival function $S_T(t) = 1 - F_T(t)$, having the following form:

$$\hat{S}_n^T(t) = \prod_{j/Z'_j \leq t}\left(1 - \frac{M\left(Z'_j\right)}{C\left(Z'_j\right)}\right) \tag{4.4.1}$$

where

- $\left(Z'_j\right)_{1\leq j\leq M}$ $(M \leq n)$ are the distinct values of $Z_i = \min(T_i, C_i)$ arranged in ascending order.
- $M\left(Z'_j\right) = \sum_{i=1}^{n}\delta_i I_{\left\{Z_i = Z'_j\right\}}$ is the exact number of deaths at the moment Z'_j.
- $C\left(Z'_j\right) = \sum_{i=1}^{n} I_{\left\{Z_i \geq Z'_j\right\}}$ is the number of individuals at risk just before the moment Z'_j.

The Kaplan-Meier estimator is also called the Boundary Product. It's a function in the staircase (whose value changes only at times corresponding to events observed), decreasing, continuing to the right. Noted that when there is no censoring, the Kaplan-Meier estimator reduces to the empirical survival function $S_n(t)$. It coincides with the empirical distribution function F_n when there is no censored data.

The asymptotic behavior of this estimator has aroused the interest of many authors. For independent random variables, Breslow and Crowley (1974) [8], were the first to deal with its convergence and asymptotic normality. By imposing the continuity of the distribution function of the interest variable and that of the censoring variable, Foldes and Rejto (1981b) [24], find a uniform convergence rate almost complete for $\hat{S}_n^T$ of the order $\sqrt{\frac{\ln(n)}{n}}$, the latter also introduced the law of iterated logarithm for the Kaplan-Meier estimator during the same year of 1981 (See Foldes and Rejt (1981a) [23]). Stute and Wang (1993) [85], in turn deal with its uniform convergence almost sure. The functional law of the iterated logarithm for right-censored data or truncated is deduced by Gu and Lai (1990) [41].

Under the strong dependence hypothesis of the variables of interest Cai (1998) [9] showed the consistency of this estimator, by specifying its speed of convergence. In his 2001 article (See Cai (2001) [10]), he generalizes the result of Cai and Roussas (1992) [11], to the Kaplan-Meier estimator, namely, the distribution of the iterated logarithm, under certain regularity and strong mixing conditions.

4.4.2 Kernel density estimator

Assume that the right-censored positive random variable T has a density of probability f_T. Foldes et al. (1981), proposed an extension to the estimator of Rosenblatt (1956) [76], and Parzen (1962) [74], which is expressed as follows:

$$\hat{f}_n^T(t) = \frac{1}{h_n} \int K\left(\frac{t-z}{h_n}\right) d\hat{F}_n^T(z),$$

where $\hat{F}_n^T = 1 - \hat{S}_n^T$ is the Kaplan-Meier estimator given by (4.4.1). In the same work of Foldes et al. (1981) [25], they proved its almost complete convergence. By the way, its asymptotic normality is established by Mielniczuk (1986) [64], and improved in Diehl and Stute (1988) [20]. Subsequently, Kagba (2004) [50], has given its convergence in quadratic mean. In the case of dependent data, only a few publications deal for the density function estimator. We can cite the work of Cai (1998) [9], who proposed an almost sure rate of convergence for stationary processes and α-mixers. Later, Liebscher (2002) [57], improved this result.

4.5 Non-parametric estimation for mixed censoring model

A new class of estimators is to be presented when the observations T_i are subjected to a censoring mechanism, this model is carried on the non-parametric estimate and discussed by Patilea and Rolin (2006) [75].

4.5.1 The Patilea and Rolin Estimator

Assuming that we have observed a sample $(Z_i;\delta_i)_{1\leq i\leq n}$ of the pair $(Z;\delta)$ where $Z=(T\wedge C)\vee L=\max(X,L)$, for $X=(T\wedge C)$ and T, L, C are positive and independent random variables representing respectively the variable of interest, the left-censored variable, and the right-censored variable.

Let H be the distribution function of Z and $H^{(0)}$ its sub-distribution for uncensored observations having the following expressions:

$$H(t)=P(Z\leq t)=F_L(t)\,F_X(t)=F_C(t)\,(1-S_T(t)\,S_C(t)),$$

and

$$H^{(0)}(t)=P(Z\leq t,\delta=0)=\int_0^t F_L(x)\,S_C(x)\,dF_T(x)$$

As well as their empirical versions are given respectively by:

$$H_n(t)=\frac{1}{n}\sum_{i=1}^{n}I_{\{Z_i\leq t\}},$$

and

$$H_n^{(0)}(t)=\frac{1}{n}\sum_{i=1}^{n}I_{\{Z_i\leq t,\delta_i=0\}}=\frac{1}{n}\sum_{i=1}^{n}I_{\{Z_i\leq t,\ T_i-C_i\leq 0,\ L_i-T_i\leq 0\}}$$

We noted Z'_j $(1\leq j\leq M)$ the distinct values of Z_i arranged in increasing order and for $k\in\{0,1,2\}$:

$$D_{kj}=\sum_{i=1}^{n}I_{\left\{Z_i=Z'_j,\delta_i=k\right\}}$$

The non-parametric estimator, denoted by $\tilde{S}_n$, of S_T, is the bounded product estimator given by Patilea and Rolin (2006) by the form:

$$\begin{aligned}\tilde{S}_n(t)&=1-\tilde{F}_n(t)\\&=\prod_{j/Z'_j\leq t}\left(1-\frac{D_{0j}}{n\mathring{F}_n\left(Z'_{j-1}\right)-nH_n\left(Z'_{j-1}\right)}\right)\end{aligned}$$

where $\mathring{F}_n$ is the Kaplan-Meier estimator of the distribution function F_L, defined by inverting time as:

$$\mathring{F}_n(t)=\prod_{j/Z'_j\leq t}\left(1-\frac{D_{2j}}{nH_n\left(Z'_j\right)}\right)$$

The almost sure uniform convergence of the Patilea-Rolin estimator is proved in this same article.

A semi-parametric estimation of copula models under right-censoring

5

Sommaire

In this chapter, we have introduce a new copula estimator for censored bivariate data based on the classical estimation method of moments, presented in a semi-parametric estimation framework. This chapter is divided into two parts the first focuses on the estimation of this new estimator when the data are doubly right-censoring, i.e. the two variables are right-censored at the same time. In the second part, we present this estimator and all results obtained by part one, in the case of singly right-censoring.

Part I

A semi-parametric estimation of copula models based on moments methods under right-censoring

A semi-parametric estimation of copula models based on moments methods under right-censoring

Idiou Nesrine[1], Benatia Fatah[2], Brahimi Brahim[3]

Abstract

Based on the classical estimation method of moments, a new copula estimator was proposed for censored bivariate data. As theoretical results, general formulas were proved with analytical forms of the obtained estimators. Taking into account Lopez and Saint-Pierre's(2012) [72], Gribkova and Lopez's (2015) [39], results, the asymptotic normality of the empirical survival copula was established. The dependence structure between the bivariate survival times were modeled under the assumption that the underlying copula is Archimedean. Accounting for various censoring patterns (singly or doubly censored), a simulation study was performed to enlighten the behavior of the procedure estimation method shows the efficiency and robustness of the new estimator proposed.

Keywords: Archimedean copulas models, Bivariate censoring, Moment estimator, Survival copula, Right censored data.

Résumé

Sur la base de la méthode classique d'estimation des moments, un nouvel estimateur de copule a été proposé pour les données bivariées censurées. Comme résultats théoriques, des formules générales ont été prouvées avec des formes analytiques des estimateurs obtenus. En tenant compte des résultats de Lopez et Saint-Pierre (2012) [72], Gribkova et Lopez (2015) [39], la normalité asymptotique de la copule empirique de survie a été établie. La structure de dépendance entre les temps de survie bivariés a été modélisée en supposant que la copule sous-jacente est d'Archimède. Prise en compte de divers modèle de censure (uniquement ou doublement censurés), une étude de simulation a été réalisée pour éclairer le comportement de la méthode d'estimation de la procédure, a montré l'efficacité et la robustesse du nouvel estimateur proposé.

Mots clés : Modèles de copules archimédiennes, Censure bivariée, Estimateur des moments, Copule de survie, Données censurées à droite.

AMS Subject Classification: 62G05, 62G20.

5.1 Introduction

The modeling of bivariate or multivariate data in survival analysis has been discussed by several authors. Many approaches have been introduced for this modelisation, including Archimedean copula models, even their application (see [4], [16], [46], [47], [62], [85], [95]). Archimedean copula models arise naturally from bivariate frailty models ([71], [69]) in which the two failure times have given an unobserved frailty W and each follows the proportional hazards model in W. However, in this aspect, an Archimedean copula is presented by:

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

where, φ is a continuous, convex and decreasing function called the generator of C, defined on $I = [0,1] \to [0,\infty]$ and verifies $\varphi(1) = 0$. In the context of multivariate survival analysis, assume that T_1 and T_2 are two failure times conditionally independent, represented thereafter by the Archimedean copula C with the cumulative distribution function (CDF):

$$F(t_1,t_2) = P(T_1 \leq t_1, T_2 \leq t_2),$$

which can be identified according to a copula function as:

$$F(t_1,t_2) = C(F_1(t_1), F_2(t_2)),$$

where C is the associated copula function and F_1, F_2 are the margins. We noted the survival functions of T_1 and T_2 by $S_1(t_1) = P(T_1 > t_1)$ and $S_2(t_2) = P(T_2 > t_2)$ respectively and the joint survival function by:

$$S(t_1,t_2) = P(T_1 > t_1, T_2 > t_2).$$

A natural question is the following: Is there a relationship between univariate and joint survival functions !! The answer is like the following by using the copula function:

$$\begin{aligned} S(t_1,t_2) &= 1 - F_1(t_1) - F_2(t_2) + F(t_1,t_2), \\ &= S_1(t_1) + S_2(t_2) - 1 + C(1 - S_1(t_1), 1 - S_2(t_2)) \end{aligned}$$

Besides, the function $\tilde{C}$ which couples S_1 and S_2 via

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

called the survival copula of (T_1, T_2). Then, if we define $\tilde{C}$ from $I^2 \to I$ we obtain:

$$\tilde{C}(u,v) = u + v - 1 + C(1-u, 1-v), \tag{5.1.1}$$

where $(u,v) \in I^2$, see Nelsen (2006) [67]. Although this latter, can also be generated by an Archimedean copula (see [33], [32]) in the manner of the following:

$$S(t_1,t_2) = \varphi^{-1}(\varphi(S_1(t_1)) + \varphi(S_2(t_2))),$$

Hence, it was demonstrated by Genest and Rivest (1993) [33] that if (T_1, T_2) follows an Archimedean copula with the marginal survival functions $S_1(t_1)$ and $S_2(t_2)$, then

$$U = \frac{\varphi\left(S_1\left(T_1\right)\right)}{\varphi\left(S_1\left(T_1\right)\right) + \varphi\left(S_2\left(T_2\right)\right)},$$

and

$$V = \tilde{C}(S_1\left(T_1\right), S_2\left(T_2\right)) = \varphi^{-1}\left(\varphi\left(S_1\left(T_1\right)\right) + \varphi\left(S_2\left(T_2\right)\right)\right),$$

are random variables distributed independently, where U distributed uniformity on I and V follows a so-called Kendall distribution with the density function:

$$k_C\left(t\right) = \frac{\varphi\left(t\right)\varphi''\left(t\right)}{\left(\varphi'\left(t\right)\right)^2},$$

defined on $(0, 1]$, as a function of t depends on the unknown parameter θ. Assume that the two failure times T_1 and T_2 can be modeled by an Archimedean copula model and it is subject to dependence or independence right-censoring with the censoring vector (C_1, C_2), we also assume that the vector (C_1, C_2) follows an arbitrary bivariate continuous distribution. Therefore, if we denote $\delta_i = 1_{\{T_i \leq C_i\}_{i=1,2}}$ which represents the indicator function of censored data, that specifies if our variable of interest is observed or not. Then, we only observe the variable $Z_i = \min(T_i, C_i)$ if $T_i \leq C_i$ when $\delta_i = 1$, otherwise, if $T_i \geq C_i$ the variable, in this case, is censored and the indicator δ_i equal to zero $\delta_i = 0$. In this paper, we are interested by type one of censoring, where two models are presented, the first is for doubly censored variables (T_1 and T_2 both are right-censored) and the second for a singly censored when only T_1 (or T_2) is right-censored.

The issues of estimating copula parameters in literature are usually solved by maximum likelihood methods ([5], [27]). For example, if we consider the IFM method Joe (2005) presented a two-stage procedure to estimate a copula, by maximizing the copula likelihood function. Even so, this maximization generally becomes very difficult to achieve when the dimension is large and the parameter numbers are also higher. For this reason, our main aim in this paper is to propose an alternative estimation method of a survival copula $\tilde{C}$, based on the moments method due to its simple mathematical form, given (T_1, T_2) as singly or doubly right-censored. General formulas were established when the considered variable $\tilde{C}$ defined under certain conditions.

The remainder of the paper is structured as follows: in section 2, our main theorems and corollary are presented where general forms of the survival copula estimator are established. As well as, the asymptotic normality of this estimator to be verified, by considering two types of right-censored models. However, in section 3 a semi-parametric estimation based on the classical moments method illustrated for a conditional distribution on $\tilde{C}$, followed by an application presented for the Gumbel model. A simulation study evaluates the performance of our estimator presented in Section 4. Our paper ends with some discussions in Section 5.

5.2 Main results

Interesting results to be proven, related by a semi-parametric estimation based on k^{th}-moments of a variable $V = \tilde{C}(u, v)$ conditionally distributed given T_1 and T_2 as singly or doubly censored. Moreover, the following theorems and corollary illustrate our main results.

Theorem 5.2.1 (Wang and Oakes 2008) *Let (T_1, T_2) be a random pair whose distribution can be modeled by an Archimedean copula. Assuming that (T_1, T_2) is subject to dependent or independent right censoring by a censoring vector (C_1, C_2) that follows an arbitrary bivariate continuous distribution, then we have:*

1. *The distribution function of $(V | T_1 > C_1 = c_1, T_2 > C_2 = c_2)$ is*

$$F_1(v, c_1, c_2) = \frac{1}{\tilde{C}(c_1, c_2)} \left\{ v - \frac{\varphi(v) - \varphi\left(\tilde{C}(c_1, c_2)\right)}{\varphi'(v)} \right\}, \ 0 \leq v \leq \tilde{C}(c_1, c_2)$$

2. *The distribution function of $(V | T_1 > C_1 = c_1, T_2 = t_2)$ is*

$$F_2(v, c_1, t_2) = \frac{\varphi'\left(\tilde{C}(c_1, t_2)\right)}{\varphi'(v)}, \quad 0 \leq v \leq \tilde{C}(c_1, t_2)$$

3. *The distribution function of $(V | T_1 = t_1, T_2 > C_2 = c_2)$ is*

$$F_3(v, t_1, c_2) = \frac{\varphi'\left(\tilde{C}(t_1, c_2)\right)}{\varphi'(v)}, \quad 0 \leq v \leq \tilde{C}(t_1, c_2)$$

Proof. See Wang and Oakes (2008) [95]. □

Based on Theorem (5.2.1), we can show the extremely important results illustrated in Corollary (5.2.1).

Corollary 5.2.1 (IDIOU, N. et al 2021) *Under the same conditions given in Theorem (5.2.1), we have:*

1. *The k^{th} moments of $(V | T_1 > c_1, T_2 > c_2)$ for $k \geq 1$ is*

$$\begin{aligned} \mathbb{E}\left(V^k \middle| T_1 > c_1, T_2 > c_2\right) = & \frac{\left(\tilde{C}(c_1, c_2)\right)^k}{k+1} \\ & - k\left(\tilde{C}(c_1, c_2)\right)^{k-1} \varphi\left(\tilde{C}(c_1, c_2)\right) \int_0^1 \frac{v^{k-1}}{\varphi'\left(v\tilde{C}(c_1, c_2)\right)} dv \\ & + k\left(\tilde{C}(c_1, c_2)\right)^{k-1} \int_0^1 \frac{v^{k-1} \varphi\left(v\tilde{C}(c_1, c_2)\right)}{\varphi'\left(v\tilde{C}(c_1, c_2)\right)} dv. \end{aligned}$$

2. *The k^{th} moments of $(V | T_1 > c_1, T_2 = t_2)$ for $k \geq 1$ is*

$$\begin{aligned} \mathbb{E}\left(V^k \middle| T_1 > c_1, T_2 = t_2\right) = & \left(\tilde{C}(c_1, t_2)\right)^k \\ & - k\left(\tilde{C}(c_1, t_2)\right)^k \varphi'\left(\tilde{C}(c_1, t_2)\right) \int_0^1 \frac{v^{k-1}}{\varphi'\left(v\tilde{C}(c_1, t_2)\right)} dv. \end{aligned}$$

3. *The k^{th} moments of $(V|\, T_1 = t_1, T_2 > c_2)$ for $k \geq 1$ is*

$$\mathbb{E}(V^k \Big| T_1 = t_1, T_2 > c_2) = \left(\tilde{C}(t_1, c_2)\right)^k - k\left(\tilde{C}(t_1, c_2)\right)^k \varphi'\left(\tilde{C}(t_1, c_2)\right) \int_0^1 \frac{v^{k-1}}{\varphi'\left(v\tilde{C}(t_1, c_2)\right)} dv.$$

Survival empirical copula for doubly right-censoring

Initially, let us clarify that from now, we are only interested by the first model presented in corollary (5.2.1), where the two variables both are doubly censored (T_1 and T_2 both). Given the accessible observation $(Z_{1i}, Z_{2i}, \delta_{1i}, \delta_{2i})_{1\leq i\leq n}$: the independent copies of a non-negative random variable of the vector $(Z_1, Z_2, \delta_1, \delta_2)$ and the survival copula $\tilde{C}$. Assuming that $\tilde{C}$ is known and the following assumptions:

- $[H_1]$ The first and the second partial derivatives of $\tilde{C}$ are limited on I^2, where $\tilde{C}(u, v)$ is different to zero for $u \neq 0$ and $v \neq 0$.
- $[H_2]$ $\exists\ (\alpha, \beta) \in I^2$, where $\tilde{C}(u, v) \geq u^\alpha v^\beta$.
- $[H_3]$ The integral $\int \frac{dF(t_1,t_2)}{\tilde{C}(S_1(t_1),S_2(t_2))}$, is strictly less than infinity. For $\theta > 0$, where $\mathcal{F}_i(t) = \int_0^t \frac{dF_i(v)}{S_i(u)^2 S_{T_i}(u)}$, $i \in \{1, 2\}$ we have

$$\int \{\frac{S_1^{1-\alpha}(t_1)\,\mathcal{F}_1^{\frac{1}{2+\theta}}(t_1)}{S_2^{\beta}(t_2)} + \frac{S_2^{1-\beta}(t_2)\,\mathcal{F}_2^{\frac{1}{2+\theta}}(t_2)}{S_1^{\alpha}(t_1)}\} dF(t_1, t_2) < \infty.$$

- $[H_4]$ Suggesting that $\int \frac{dF(t_1,t_2)}{S_1(t_1^-)}$, is strictly less than infinity and for $\theta > 0$, we have

$$\int \left\{ \left(\int_0^{t_1} \frac{dF_1(v)}{S_1(v^-)^2 S_{T_1}(v)} \right)^{\frac{1}{2+\theta}} \right\} dF(t_1, t_2) < \infty.$$

Lopez and Saint-Pierre(2012) [72], have studied this model, noting that F can be consistently estimated by an F_n estimator in the following form:

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

that could not be used to estimate $F(t_1, t_2)$ since T_1 and T_2 are unobserved. Therefore, according to the proposition of Lopez and Saint-Pierre(2012) [72], the F estimate can be given in such form:

$$F_n(t_1, t_2) = \frac{1}{n} \sum_{i=1}^{n} \frac{\delta_{1i}\delta_{2i}}{\tilde{C}\left(\hat{S}_1(Z_{1i}), \hat{S}_2(Z_{2i})\right)} 1_{\{Z_{1i}\leq t_1, Z_{2i}\leq t_2\}}, \tag{5.2.1}$$

where $\tilde{C}$ is the survival copula given by (5.1.1) and

$$\hat{S}_1(t) = \prod_{k/Z'_{1k}<t} \left(1 - \frac{\sum_{i=1}^{n} 1_{\{Z_{1i}=Z'_{1k}, \delta_{1i}=0\}}}{\sum_{i=1}^{n} 1_{\{Z_{1i}\geq Z'_{1k}\}}}\right)$$

is the Kaplan-Meier estimate of S_1 for $((Z'_{1:k})_{1\leq k\leq m}, m \leq n)$, and $\hat{S}_2$ is the Kaplan-Meier estimate of S_2 defined by the same way. Noted Γ_{T_1} and Γ_{T_2} the support of T_1 and T_2 respectively and $l^\infty(W)$ all bounded real-valued functions space, identified on non-empty set W.

Assuming that the assumptions $[H_1] - [H_3]$ hold, Lopez and Saint Pierre's (2012), have concluded that the processes $n^{\frac{1}{2}}(F_n - F)$ converge weakly in $l^\infty(\Gamma_{T_1} * \Gamma_{T_2})$ to a centered Gaussian process (Theorem (3.4) [72]). By the way, in the event of complete data, the copula C can be estimated by:

$$\hat{C}(u,v) = F_n(F_{1n}^{-1}(u), F_{2n}^{-1}(v)),$$

where $(u,v) \in I^2$, $F_{1n}(t_1) = \lim_{t_2\to\infty} F_n(t_1,t_2)$ and $F_{2n}(t_2) = \lim_{t_1\to\infty} F_n(t_1,t_2)$, Gribkova and Lopez (2015) [39], proposed the empirical copula of C in the case of incomplete data given by:

$$C_n(u,v) = \frac{1}{n}\sum_{i=1}^{n} \frac{\delta_{1i}\delta_{2i}}{\tilde{C}\left(\hat{S}_1(Z_{1i}), \hat{S}_2(Z_{2i})\right)} 1_{\{F_{1n}(Z_{1i})\leq u, F_{2n}(Z_{2i})\leq v\}}, \tag{5.2.2}$$

when the two variables are both right-censored. By analogy, using (5.1.1) and (5.2.2), the empirical survival copula via:

$$\tilde{C}_n(u,v) = u + v - 1 + \frac{1}{n}\sum_{i=1}^{n} \frac{\delta_{1i}\delta_{2i}}{\tilde{C}\left(\hat{S}_1(Z_{1i}), \hat{S}_2(Z_{2i})\right)} 1_{\{1-F_{1n}(Z_{1i})\geq u, 1-F_{2n}(Z_{2i})\geq v\}} \tag{5.2.3}$$

Observe that for this models

$$\sup_{(u,v)\in I^2} |C_n(u,v) - \hat{C}(u,v)| = O_p\left(\frac{1}{n}\right),$$

which means that the process $n^{\frac{1}{2}}(C_n - C)$ converges weakly in $l^\infty(I^2)$ to the limiting approach L (centered Gaussian process), which either has been proven by Gribkova and Lopez (2015) [39], in theorem 2. Hence, this weak convergence allows us to prove the asymptotic normality of statistics given by the form:

$$\int_{I^2} g(u,v) dC_n(u,v),$$

noted g as a function that has a real value defined on I^2 (Van der vart and wellner (1996)).

Fermanian, Radulovic, and Wegkamp (2004), have proven this asymptotic normality in the case of complete data. By the way, thanks to Theorem 1 of M. Boukeloua (2020)[62], who proved that under some assumption when $n \to \infty$ the quantity:

$$n^{\frac{1}{2}}\left\{\int_{I^2} g(u,v) d\left(C_n(u,v) - C(u,v)\right)\right\},$$

converges in distribution to a Gaussian random variable

$$G = \int_{I^2} g(u,v) d\left(L(u,v)\right),$$

where $g \in R_2(I^2)$, the set of all real-valued functions defined on I^2, which are continuous from above and with discontinuities of the first kind.

Based on these results and if we assume the assumptions $[H_1] - [H_4]$ hold we can show the next theorem.

Theorem 5.2.2 (IDIOU, N. et al 2021) *Assuming the function $g \in R_2(I^2)$, $\tilde{C}$ and $\tilde{C}_n$ the survival copula and its empirical version respectively, then when $n \to \infty$ we have*

$$n^{\frac{1}{2}} \left\{ \int_{I^2} g(u,v) d\left(\tilde{C}_n(u,v) - \tilde{C}(u,v)\right) \right\} \xrightarrow{D} \int_{I^2} g(u,v) d\left(L(u,v)\right),$$

where the limiting is a Gaussian random variable and $(u,v) \in I^2$.

This theorem proved the asymptotic normality of the empirical survival copula, which remains valid for both models considered in the corollary (5.2.1).

5.3 Moments estimator for right-censoring

Either the following figure, T_1 and T_2 represent the survival time point and (C_1, C_2) the censoring time point. The display contains four data kinds points, including observed points (T_1, T_2), two types of singly censored points (T_1, C_2), (C_1, T_2), and doubly censored points (C_1, C_2).

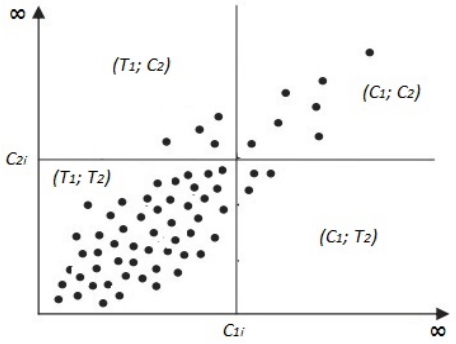

Figure 5.1 – *Censored data.*

Let (T_1, T_2) a random variables whose distribution can be modeled by an Archimedean copula and is subject to dependent or independent right censoring, $V = \tilde{C}\left(S_1\left(Z_{1i}\right), S_2\left(Z_{2i}\right)\right)$ is a conditionally distributed variable that follows a so-called Kendall distribution K_C with the density function:

$$k_C(t) = \frac{\varphi(t)\,\varphi''(t)}{\left(\varphi'(t)\right)^2},$$

defined on $(0,1]$. We define the k^{th}-moments of V for $k \geq 1$ by:

$$M_k(V|H) = E(V^k|H),$$

where $H = h_{(c_1,c_2)}$ indicate the first case of censoring (T_1 and T_2 are both right-censoring). Then, relying on the results obtained in Corollary (5.2.1), we have:

$$M_k(V|H) = \frac{\left(\tilde{C}\left(c_1, c_2\right)\right)^k}{k+1} - k\left(\tilde{C}\left(c_1, c_2\right)\right)^{k-1} \varphi\left(\tilde{C}\left(c_1, c_2\right)\right) \int_0^1 \frac{v^{k-1}}{\varphi'\left(v\tilde{C}\left(c_1, c_2\right)\right)} dv + k\left(\tilde{C}\left(c_1, c_2\right)\right)^{k-1} \int_0^1 \frac{v^{k-1}\varphi\left(v\tilde{C}\left(c_1, c_2\right)\right)}{\varphi'\left(v\tilde{C}\left(c_1, c_2\right)\right)} dv.$$

Suppose now that V belongs to a parametric family $V_\theta = \tilde{C}_\theta(u, v)$, it follows that $\varphi = \varphi_\theta$, $\tilde{C} = \tilde{C}_\theta$ and $K_C = K_\theta$, where $u = S_1(t_1) = \bar{F}_1(t_1)$ and $v = S_2(t_2) = \bar{F}_2(t_2)$, mentioned that F_1 and F_2 are completely known.

Noted that $M_k(V|H) = M_k(\theta|H)$, then, we can distinguish the following form of the k^{th}-moments:

$$M_k(\theta|h_{(c_1,c_2)}) = \frac{\left(\tilde{C}_\theta\left(c_1, c_2\right)\right)^k}{k+1} - k\left(\tilde{C}_\theta\left(c_1, c_2\right)\right)^{k-1} \varphi_\theta\left(\tilde{C}_\theta\left(c_1, c_2\right)\right) \int_0^1 \frac{v_\theta^{k-1}}{\varphi'_\theta\left(v_\theta\tilde{C}_\theta\left(c_1, c_2\right)\right)} dv_\theta + k\left(\tilde{C}_\theta\left(c_1, c_2\right)\right)^{k-1} \int_0^1 \frac{v_\theta^{k-1}\varphi_\theta\left(v_\theta\tilde{C}_\theta\left(c_1, c_2\right)\right)}{\varphi'_\theta\left(v_\theta\tilde{C}_\theta\left(c_1, c_2\right)\right)} dv_\theta,$$

for unknown $\theta \in \mathbb{R}^d$. Given the empirical version of moment estimator under doubly censored presented by:

$$\hat{M}_k = \hat{M}_k(\hat{V}|h_{(c_1,c_2)}) = \frac{1}{n}\sum_{i=1}^{n}\left\{\tilde{C}_n\left(\hat{S}_i(t_i)|H\right)\right\}^k,$$

for $k \geq 1$ where $\hat{V} = \tilde{C}_n$ is the survival empirical copula given by formula (5.2.3). By analogy, as the natural estimators of moments copula it is necessary to solve the equation system given below:

$$\begin{cases} M_1(\theta|h_{(c_1,c_2)}) = \hat{M}_1 \\ M_2(\theta|h_{(c_1,c_2)}) = \hat{M}_2 \\ \vdots \\ M_d(\theta|h_{(c_1,c_2)}) = \hat{M}_d. \end{cases}$$

To obtain the unique solution $\hat{\theta}^{CCM} = (\hat{\theta}_1, ..., \hat{\theta}_d)$ called the censored copula moment (CCM) estimator of θ.

Example 10 ***Application: illustrative example***

In particular, in the bivariate case, the Gumbel model of one-parameter is given by:

$$C_\alpha\left(u, v\right) = \exp\left(-\left((-\ln u)^\alpha + (-\ln v)^\alpha\right)^{\frac{1}{\alpha}}\right),$$

with the generator: $\varphi_\alpha(t) = (-\ln t)^\alpha$, $\alpha \in [1, +\infty[$. *Consequently, by considering the case of two parameters, the preceding model becomes:*

$$C_{\alpha,\beta}(u,v) = \left(\left(\left(u^{-\alpha}-1\right)^\beta + \left(v^{-\alpha}-1\right)^\beta\right)^{\frac{1}{\beta}} + 1\right)^{-\frac{1}{\alpha}},$$

with the generator: $\varphi_{\alpha,\beta}(t) = (t^{-\alpha}-1)^\beta$, *where* $\alpha > 0$ *and* $\beta \geq 1$ *(see [7]). Obviously, by the use of (5.1.1), we obtain the survival copula of the Gumbel family given by:*

$$\tilde{C}_{\alpha,\beta}(u,v) = u+v-1+\left(\left(\left((1-u)^{-\alpha}-1\right)^\beta + \left((1-v)^{-\alpha}-1\right)^\beta\right)^{1/\beta} + 1\right)^{-1/\alpha}$$

Hence, as an application of our results proved previously we can reach the following bivariate censoring models using equation 1 in Corollary (5.2.1).

For $k \geq 1$, $\alpha > 0$ *and* $1 \leq \beta \leq 2$, *the* k^{th} *moments of the Gumbel's survival copula, is given by:*

$$\begin{aligned} M_k((\alpha,\beta)|H) &= E\left(V^k \middle| h_{(c_1,c_2)}\right) \\ &= \frac{m^k}{k+1} + \frac{k\left(m^{-\alpha}-1\right)^\beta}{\alpha^2\beta m}\beta_{m^\alpha}\left(\beta + \frac{k+1}{\alpha}, 2-\beta\right) \\ &+ \frac{km^{k-1}}{\alpha\beta}\left(\frac{m^{\alpha+1}}{k+\alpha+1} - \frac{m}{k+1}\right), \end{aligned}$$

in which $\beta_{m^\alpha}(x,y)$ *is the Beta function and* $m = \tilde{C}(c_1,c_2)$ *is the ordinary copula. If we simplify more the previous formula we will obtain the following writing:*

$$\begin{aligned} M_k((\alpha,\beta)|h_{(c_1,c_2)}) = &\ \frac{m^k}{k+1} + \frac{k}{\alpha\beta} \times \left(\frac{m^{k+\alpha}}{k+\alpha+1} - \frac{m^k}{k+1}\right. \\ &\left. - \frac{(\beta-1)\left(m^{-\alpha}-1\right)^\beta}{\alpha m^{\alpha+1}} \frac{\Gamma(1-\beta)\,\Gamma\left(\frac{1}{\alpha}(k+\alpha\beta+1)\right)}{\Gamma\left(\frac{1}{\alpha}(k+2\alpha+1)\right)}\right), \end{aligned}$$

where $\Gamma(x)$ *is the Gamma function. In particular, the two first moments are given by:*

$$\begin{cases} M_1((\alpha,\beta)|h_{(c_1,c_2)}) = \frac{1}{2}m + \frac{\left(m^{-\alpha}-1\right)^\beta}{\alpha^2\beta m}\beta_{m^\alpha}\left(\beta+\frac{2}{\alpha}, 2-\beta\right) + \frac{1}{\alpha\beta}\left(\frac{m^{\alpha+1}}{\alpha+2} - \frac{m}{2}\right) \\ M_2((\alpha,\beta)|h_{(c_1,c_2)}) = \frac{1}{3}m^2 + \frac{2\left(m^{-\alpha}-1\right)^\beta}{\alpha^2\beta m}\beta_{m^\alpha}\left(\beta+\frac{3}{\alpha}, 2-\beta\right) + \frac{1}{\alpha\beta}\left(\frac{m^{\alpha+1}}{\alpha+3} - \frac{m}{3}\right) \end{cases}$$

Which can further simplify as well:

$$\begin{cases} M_1((\alpha,\beta)|h_{(c_1,c_2)}) = \frac{1}{2}m + \frac{1}{\alpha\beta}\left\{\frac{m^{\alpha+1}}{\alpha+2} - \frac{1}{2}m - \frac{(\beta-1)\left(m^{-\alpha}-1\right)^\beta}{\alpha m^{\alpha+1}} \frac{\Gamma(1-\beta)\Gamma\left(\frac{1}{\alpha}(\alpha\beta+2)\right)}{\Gamma\left(\frac{2}{\alpha}(\alpha+1)\right)}\right\} \\ M_2((\alpha,\beta)|h_{(c_1,c_2)}) = \frac{1}{3}m^2 + \frac{2}{\alpha\beta}\left\{\frac{m^{\alpha+2}}{\alpha+3} - \frac{1}{3}m^2 - \frac{(\beta-1)\left(m^{-\alpha}-1\right)^\beta}{\alpha m^{\alpha+1}} \frac{\Gamma(1-\beta)\Gamma\left(\frac{1}{\alpha}(\alpha\beta+3)\right)}{\Gamma\left(\frac{1}{\alpha}(2\alpha+3)\right)}\right\} \end{cases}$$

However, the CCM estimator of $\theta = (\alpha,\beta)$ *is the unique solution of the system:*

$$\begin{cases} M_1(\theta|h_{(c_1,c_2)}) = \hat{M}_1 \\ M_2(\theta|h_{(c_1,c_2)}) = \hat{M}_2 \end{cases}$$

5.4 Simulation study

To illustrate the performances of the proposed estimator, a simulation study is carried out based on the Monte Carlo method for right-censored sampling. First, we generate a bivariate survival distribution of the Gumbel copula model where the margins are assumed to be Pareto(λ),

$$F(t) = 1 - t^{-\lambda}, \;\; t \geq 0$$

The distribution of survival times T_1, T_2, and the censoring times C_1, C_2 are all assumed to be Pareto of parameters $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ respectively. If we suppose that the corresponding percentage of observed data is equal to $p_1 = \frac{\lambda_2}{\lambda_1+\lambda_2}$ for the first sample, then we can choose the values 0.3 for λ_1 and $0.95, 0.90, 0.85, 0.80$ for p_1, next we solve the equation $p_1 = \frac{\lambda_2}{\lambda_1+\lambda_2}$ to get the pertaining λ_2-values. In this path, we fix λ_3 and $p_2 = \frac{\lambda_4}{\lambda_3+\lambda_4}$ by the same previous values to find λ_4 by the same way.

Since the quality of the estimate is assessed by evaluating the bias (relative Bias) and the root mean square error (RMSE), then for the two samples both we generate 1000 replicas for each common size n varied for $n = 30, 50, 100, 500, 1000, 2000$, to pick our final performance as empirical evidence of the results gained across all replicates. Besides, for a wide set of parameters of the true survival copula $\tilde{C}_{\alpha,\beta}$ the simulation procedure based on Section (5.3) is repeated for each sample.

The selection of true survival copula parameter values (α, β) must be significant, i.e. each couple of parameters consists a value of one of the dependency measurements. So, if we consider Kendall's τ as an association index then, it can be expressed as a function of the dependency parameter in Archimedean copula models. In this case, we should select the parameter values of $\tilde{C}$ that correspond to specified values of τ by using the transformed of the underlying survival copula. Since the link between Kendall's τ and $\tilde{C}$ is usually formulated by

$$\tau_{\alpha,\beta} = 4E(V_{\alpha,\beta}) - 1,$$

where $V_{\alpha,\beta} = \tilde{C}_{\alpha,\beta}(u, v)$, then to generate data, we select values for survival copula parameters that corresponding to Kendall's tau values 0.05 (low association), 0.5 (mean association) and 0.7 (high positive association).

For the Gumbel survival copula of two parameters, the performance of the estimator proposed is summarized in Tables (5.1-5.3). The results obtained for different values of Kendall's τ are quite good in the three cases of dependence considered (0.05, 0.5, 0.7) and by considering different censoring percentages. In each table, τ_1 and τ_2 are represented respectively the Kendall's tau value before and after censoring.

From the three tables, we deduce that the estimator proposed to have a good performance and works quite well if we compare it with other methods used before on the copulas estimation. By the way, the performance of survival copula estimate based on the moments method is justified,

Table 5.1 – *Moments estimator performance based on Gumbel survival copula generated from 1000 replications with Pareto margins and shape parameter 0.3. Re.Bias and RMSE of the estimators are calculated for different censoring values and weak dependence.*

$\tau = 0.05\,,\ \alpha = 0.1 \rightarrow \beta = 1.00$

1% of censoring												
N	$n = 30$		$n = 50$		$n = 100$		$n = 500$		$n = 1000$		$n = 2000$	
$(\hat{\alpha}, \hat{\beta})$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\alpha}$	$\hat{\beta}$
Re.Bias	-0.0563	0.2852	-0.0537	0.2119	-0.0539	0.2515	-0.0536	0.2403	-0.051	0.2461	-0.0546	0.1972
RMSE	0.0649	0.0116	0.0624	0.0117	0.0629	0.0119	0.0620	0.0117	0.0606	0.0118	0.0631	0.0116
τ_1	0.04336		0.05059		0.0477		0.04733		0.04892		0.04791	
τ_2	0.04384		0.04992		0.0477		0.04711		0.04838		0.03699	
c_1	0.03188		0.01975		0.00974		0.00208		0.00105		0.00043	
c_2	0.03134		0.01956		0.00947		0.00195		0.00104		0.00041	
5% of censoring												
Re.Bias	-0.0539	0.2072	-0.0526	0.2490	-0.0551	0.2450	-0.0544	0.2339	-0.0520	0.2457	-0.0540	0.2475
RMSE	0.0625	0.0115	0.0613	0.0115	0.0638	0.0118	0.0628	0.0115	0.0610	0.0115	0.0629	0.0116
τ_1	0.04685		0.04920		0.04682		0.04851		0.05031		0.04969	
τ_2	0.04240		0.04914		0.04524		0.04656		0.04771		0.04782	
c_1	0.03090		0.01852		0.00932		0.00195		0.001		0.00053	
c_2	0.03135		0.01953		0.00948		0.00192		0.001		0.00051	
10% of censoring												
Re.Bias	-0.0526	0.2387	-0.0538	0.2264	-0.0548	0.2455	-0.0526	0.2294	-0.0547	0.2380	-0.0546	0.2186
RMSE	0.0616	0.0117	0.0627	0.0119	0.0637	0.0117	0.0617	0.0120	0.0634	0.0118	0.0635	0.0116
τ_1	0.0567		0.04865		0.04794		0.05103		0.04924		0.04989	
τ_2	0.04909		0.04385		0.04431		0.04669		0.04515		0.04565	
c_1	0.02813		0.01722		0.0089		0.00179		0.00098		0.00052	
c_2	0.0286		0.01670		0.00867		0.00175		0.00098		0.00048	
20% of censoring												
Re.Bias	-0.0531	0.2136	-0.0532	0.2003	-0.0530	0.1926	-0.0524	0.2049	-0.0532	0.2038	-0.0518	0.1965
RMSE	0.0619	0.0117	0.0617	0.0118	0.0620	0.0118	0.0615	0.0114	0.0623	0.0115	0.0611	0.0117
τ_1	0.0524		0.05195		0.05116		0.04761		0.04974		0.05035	
τ_2	0.03684		0.04077		0.04059		0.03969		0.04155		0.04125	
c_1	0.02514		0.01573		0.00828		0.00171		0.00088		0.00045	
c_2	0.02540		0.01537		0.00785		0.00169		0.00081		0.00045	
25% of censoring												
Re.Bias	-0.0518	0.1861	-0.0531	0.2058	-0.0526	0.2109	-0.0519	0.1782	-0.0534	0.1950	-0.0546	0.1972
RMSE	0.0610	0.0116	0.0625	0.0115	0.0612	0.0114	0.0606	0.0118	0.0622	0.0116	0.0631	0.0116
τ_1	0.04636		0.04423		0.04691		0.04842		0.04888		0.04791	
τ_2	0.0384		0.03729		0.03637		0.03693		0.03795		0.03699	
c_1	0.02441		0.01444		0.00754		0.00148		0.00078		0.00043	
c_2	0.02402		0.01445		0.00698		0.00161		0.00078		0.00041	

Table 5.2 – *Moments estimator performance based on Gumbel survival copula generated from 1000 replications with Pareto margins and shape parameter 0.3. Re.Bias and RMSE of the estimators are calculated for different censoring values and moderate dependence.*

$\tau = 0.5\,,\ \alpha = 0.2 \rightarrow \beta = 1.82$

1% of censoring												
N	$n = 30$		$n = 50$		$n = 100$		$n = 500$		$n = 1000$		$n = 2000$	
$(\hat{\alpha}, \hat{\beta})$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\alpha}$	$\hat{\beta}$
Re.Bias	-0.0263	0.3759	-0.0263	0.3735	-0.0259	0.3480	-0.0256	0.3648	-0.0252	0.3649	-0.0250	0.3695
RMSE	0.0302	0.0063	0.0302	0.0063	0.0299	0.0064	0.0299	0.0064	0.0293	0.0063	0.0292	0.0064
τ_1	0.49995		0.50077		0.49672		0.5008		0.50048		0.50033	
τ_2	0.49148		0.49273		0.48865		0.49225		0.49224		0.49208	
c_1	0.03163		0.01859		0.00977		0.00195		0.00102		5e-04	
c_2	0.03296		0.01876		0.00972		0.00194		0.00103		0.00053	
5% of censoring												
Re.Bias	-0.0267	0.3538	-0.0266	0.3554	-0.0255	0.3759	-0.0257	0.3415	-0.0255	0.3500	-0.0265	0.3566
RMSE	0.0309	0.0064	0.0306	0.0062	0.0296	0.0062	0.0299	0.0063	0.0298	0.0065	0.0306	0.0064
τ_1	0.50274		0.50408		0.49962		0.50013		0.50015		0.50143	
τ_2	0.46311		0.46417		0.45925		0.46042		0.45939		0.46095	
c_1	0.02847		0.01856		0.00946		0.00183		0.00099		0.00051	
c_2	0.02926		0.01803		0.00940		0.00183		0.00094		0.00049	
10% of censoring												
Re.Bias	-0.0259	0.3875	-0.0261	0.3658	-0.0255	0.3319	-0.0261	0.3312	-0.0262	0.3260	-0.0265	0.3301
RMSE	0.0300	0.0065	0.0303	0.0065	0.0298	0.0063	0.0302	0.0065	0.0302	0.0065	0.0304	0.0062
τ_1	0.49346		0.50026		0.50191		0.49985		0.50019		0.4999	
τ_2	0.41385		0.41922		0.42175		0.42134		0.42191		0.42137	
c_1	0.0279		0.01841		0.00902		0.00178		0.00092		0.00049	
c_2	0.02938		0.01794		0.00901		0.00177		0.00089		0.00046	
20% of censoring												
Re.Bias	-0.0261	0.3413	-0.0256	0.2964	-0.0252	0.3214	-0.0259	0.3021	-0.0250	0.3065	-0.0264	0.3017
RMSE	0.0303	0.0063	0.0298	0.0065	0.0295	0.0063	0.0302	0.0063	0.0292	0.0065	0.0304	0.0063
τ_1	0.50244		0.4976		0.49889		0.50007		0.50027		0.49999	
τ_2	0.35571		0.34813		0.35141		0.35171		0.35227		0.35153	
c_1	0.02586		0.01614		0.00838		0.00164		0.00081		0.00041	
c_2	0.02487		0.01648		0.00797		0.0016		0.00081		0.00042	
25% of censoring												
Re.Bias	-0.0251	0.2793	-0.0259	0.3205	-0.0266	0.2833	-0.0254	0.2982	-0.0253	0.2869	-0.0256	0.2838
RMSE	0.0293	0.0063	0.0299	0.0062	0.0305	0.0064	0.0296	0.0064	0.0295	0.0064	0.0298	0.0065
τ_1	0.49657		0.50095		0.50224		0.50036		0.50043		0.50059	
τ_2	0.31482		0.32157		0.31815		0.31958		0.32107		0.32089	
c_1	0.02483		0.01532		0.00737		0.00156		0.00079		4e-04	
c_2	0.02584		0.01444		0.00729		0.00152		0.00074		4e-04	

Table 5.3 – *Moments estimator performance based on Gumbel survival copula generated from 1000 replications with Pareto margins and shape parameter 0.3. Re.Bias and RMSE of the estimators are calculated for different censoring values and strong dependence.*

$\tau = 0.7\,,\, \alpha = 0.4 \rightarrow \beta = 2.78$

1% of censoring												
N	$n = 30$		$n = 50$		$n = 100$		$n = 500$		$n = 1000$		$n = 2000$	
$(\hat{\alpha}, \hat{\beta})$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\alpha}$	$\hat{\beta}$
Re.Bias	-0.0131	0.4582	-0.0129	0.4197	-0.0128	0.4053	-0.0127	0.4127	-0.0128	0.4195	-0.0125	0.4171
RMSE	0.0150	0.0041	0.0149	0.0042	0.0147	0.0040	0.0147	0.0040	0.0148	0.0430	0.0145	0.0042
τ_1	0.70344		0.69913		0.70007		0.70007		0.70013		0.69997	
τ_2	0.69002		0.68812		0.68855		0.68788		0.68815		0.68794	
c_1	0.03001		0.0189		0.00982		0.0019		0.00096		5e-04	
c_2	0.03035		0.01927		0.01		0.00191		0.0095		0.00048	
5% of censoring												
Re.Bias	-0.0126	0.4252	-0.0127	0.4063	-0.0127	0.3972	-0.0126	0.3973	-0.0126	0.4056	-0.0124	0.4001
RMSE	0.0146	0.0041	0.0147	0.0042	0.0147	0.0042	0.0146	0.0042	0.0146	0.0042	0.0144	0.0041
τ_1	0.69937		0.69845		0.69828		0.70019		0.7001		0.70065	
τ_2	0.64116		0.63732		0.6397		0.64107		0.64168		0.64204	
c_1	0.03068		0.02042		0.00955		0.00194		0.00098		0.00049	
c_2	0.0305		0.01966		0.00974		0.00186		0.00095		0.00049	
10% of censoring												
Re.Bias	-0.0123	0.3847	-0.0125	0.3756	-0.0127	0.3768	-0.0127	0.3927	-0.0129	0.3889	-0.0121	0.3860
RMSE	0.0144	0.0041	0.0145	0.0042	0.0146	0.0041	0.0147	0.0042	0.0149	0.0041	0.0142	0.0043
τ_1	0.69714		0.70026		0.69879		0.69974		0.70095		0.70013	
τ_2	0.58936		0.58693		0.58613		0.58613		0.58814		0.58743	
c_1	0.03007		0.01752		0.00928		0.00183		0.00088		0.00045	
c_2	0.02926		0.01711		0.00886		0.00186		0.00092		0.00047	
20% of censoring												
Re.Bias	-0.0128	0.3923	-0.0125	0.3671	-0.0125	0.3364	-0.0132	0.3458	-0.0130	0.3441	-0.0127	0.3445
RMSE	0.0148	0.0042	0.0146	0.0041	0.0148	0.0041	0.0151	0.0041	0.0149	0.0041	0.0147	0.0041
τ_1	0.70236		0.70103		0.69985		0.70113		0.70053		0.70069	
τ_2	0.4952		0.49110		0.48840		0.49066		0.48991		0.4895	
c_1	0.02444		0.01543		0.00820		0.0016		0.00082		4e-04	
c_2	0.02485		0.01492		0.00829		0.00155		0.00077		0.00039	
25% of censoring												
Re.Bias	-0.0126	0.2926	-0.0128	0.3569	-0.0126	0.3280	-0.0126	0.3417	-0.0122	0.334	-0.0126	0.3247
RMSE	0.0147	0.0043	0.0149	0.0041	0.0146	0.0042	0.0146	0.0040	0.0142	0.0041	0.0146	0.0041
τ_1	0.69894		0.69874		0.70112		0.70002		0.70018		0.70029	
τ_2	0.44299		0.43503		0.44424		0.44622		0.44462		0.4453	
c_1	0.02306		0.01406		0.00691		0.00147		0.00073		0.00038	
c_2	0.02331		0.01521		0.00749		0.00147		0.00072		0.00036	

through the adoption of relative bias (Re.Bais) and RMSE discourse, when we can see all their values are sufficiently decreased for each case of small and even large samples (are almost close to zero). Even so, the value of Kendall's tau after censoring (τ_2) remains close to its original theoretical value given by τ_1 , which means that the variables remain dependent despite the censorship.

5.5 Discussion

In this paper, we elaborate a semi-parametric estimation method of a survival copula based on Archimedean models, but in specific conditions on the data. Indeed, under different censoring (singly or doubly), the results of our estimator were presented with an analytical form which overcame the problem that occurs usually by other methods. As an application of the considered method, we have chosen the Gumbel model, given T_1 and T_2 as doubly right-censored variables. In the simulation part, three cases of dependence are considered, where the results can validate the use of the method proposed. Consequently, this method is preferable if we compare it with the maximum likelihood method, because of its easy mathematical form.

Our main result for these studies is based on the copula approaches and the survival analysis, in which the correlation between two survival time variables was detected. Therefore, our research results open a vast area of application, notably in real life, when there are two related events defined under specific situations. This will be discussed in an interesting new paper that we are currently working on. Based on the outcomes of Gripkova and Lopez (2015) [39], Lopez and Saint-Pierre (2012) [72], research our results can be applied for left and right censoring. This is one of our current research topics and the idea has been developed in another paper that is also under preparation.

Acknowledgement. The authors would like to extend their gratitude to the editor of the journal and to the reviewers for their valuable advice.

5.6 Appendix

Proof of corollary (5.2.1). In order to prove the result of Corollary (5.2.1) we need to use the results given in Theorem (5.2.1) and we start by equation 1. For $k > 1$ the k^{th} moments is given by:

$$\mathbb{E}(V^k \Big| T_1 > c_1, T_2 > c_2) = \int_0^{\tilde{C}(c_1,c_2)} v^k dF_1(v, c_1, c_2)$$

Using the conditional distribution of $(V| T_1 > c_1, T_2 > c_2)$, given in Theorem (5.2.1) we get:

$$\begin{aligned}
\mathbb{E}(V^k \Big| T_1 &> c_1, T_2 > c_2) = \\
&= \frac{1}{\tilde{C}(c_1,c_2)} \int_0^{\tilde{C}(c_1,c_2)} v^k \left\{ 1 - \frac{(\varphi'(v))^2 - \varphi''(v)\left(\varphi(v) - \varphi\left(\tilde{C}(c_1,c_2)\right)\right)}{(\varphi'(v))^2} \right\} dv \\
&= \frac{1}{\tilde{C}(c_1,c_2)} \int_0^{\tilde{C}(c_1,c_2)} v^k dv \\
&- \frac{1}{\tilde{C}(c_1,c_2)} \int_0^{\tilde{C}(c_1,c_2)} v^k \frac{(\varphi'(v))^2 - \varphi''(v)\left(\varphi(v) - \varphi\left(\tilde{C}(c_1,c_2)\right)\right)}{(\varphi'(v))^2} dv \\
&= I_1 - I_2,
\end{aligned}$$

by the way, I_1 have to simplify as follows:

$$I_1 = \frac{1}{\tilde{C}(c_1,c_2)} \int_0^{\tilde{C}(c_1,c_2)} v^k dv = \frac{\tilde{C}(c_1,c_2)^k}{k+1}.$$

On otherhand, to simplify I_2 we pass directly to integration by parts, and we have:

$$\begin{aligned}
I_2 &= \frac{1}{\tilde{C}(c_1,c_2)} \int_0^{\tilde{C}(c_1,c_2)} v^k \frac{(\varphi'(v))^2 - \varphi''(v)\left(\varphi(v) - \varphi\left(\tilde{C}(c_1,c_2)\right)\right)}{(\varphi'(v))^2} dv \\
&= \frac{1}{\tilde{C}(c_1,c_2)} \left(\left[v^k \frac{\varphi(v) - \varphi\left(\tilde{C}(c_1,c_2)\right)}{\varphi'(v)} \right]_0^{\tilde{C}(c_1,c_2)} \right. \\
&\left. -k \int_0^{\tilde{C}(c_1,c_2)} v^{k-1} \frac{\varphi(v) - \varphi\left(\tilde{C}(c_1,c_2)\right)}{\varphi'(v)} dv \right) \\
&= -\frac{k}{\tilde{C}(c_1,c_2)} \int_0^{\tilde{C}(c_1,c_2)} v^{k-1} \frac{\varphi(v) - \varphi\left(\tilde{C}(c_1,c_2)\right)}{\varphi'(v)} dv.
\end{aligned}$$

it follows after changing variables that:

$$\begin{aligned}
I_2 &= -k\left(\tilde{C}(c_1,c_2)\right)^{k-1} \int_0^1 v^{k-1} \frac{\varphi(v\tilde{C}(c_1,c_2)) - \varphi\left(\tilde{C}(c_1,c_2)\right)}{\varphi'(v\tilde{C}(c_1,c_2))} dv \\
&= -k\left(\tilde{C}(c_1,c_2)\right)^{k-1} \int_0^1 v^{k-1} \frac{\varphi(v\tilde{C}(c_1,c_2)}{\varphi'(v\tilde{C}(c_1,c_2))} dv \\
&+ k\left(\tilde{C}(c_1,c_2)\right)^{k-1} \varphi\left(\tilde{C}(c_1,c_2)\right) \int_0^1 \frac{v^{k-1}}{\varphi'(v\tilde{C}(c_1,c_2))} dv.
\end{aligned}$$

For equations 2 and 3 (in the case of singly-censored data), we see its proofs and approaches details in a future article which is under preparation. □

Proof of theorem (5.2.2). If we consider the survival copula $\tilde{C}$ and its empirical version $\tilde{C}_n$, we have

$$\begin{aligned}\tilde{C}_n(u,v) - \tilde{C}(u,v) &= u + v - 1 + C_n(1-u, 1-v) - \tilde{C}(u,v) \\ &= C_n(1-u, 1-v) - C(1-u, 1-v),\end{aligned}$$

hence, by a change of variables $w_1 = 1 - u$ and $w_2 = 1 - v$, we get

$$\tilde{C}_n(u,v) - \tilde{C}(u,v) = C_n(w_1, w_2) - C(w_1, w_2),$$

where (w_1, w_2) remain belongs to the interval I^2. So, we can concluded that $n^{\frac{1}{2}}\left(\tilde{C}_n - \tilde{C}\right)$ also converges weakly in $l^\infty(I^2)$ to the limiting approach L. Assuming the application ζ represented on $R_I(I^2)$ the set of all functions defined on I^2 whose total variation is bounded by 1 and which are continuous from above and with discontinuities of the first kind, given by

$$\zeta(h) = \int_{I^2} g(w_1, w_2) dh(w_1, w_2),$$

which is Hadamard differentiable on $R_I(I^2)$, (see [89]). Because $n^{\frac{1}{2}}(C_n - C)$ converges weakly to the limiting approach L, then, by using delta method we get

$$\begin{aligned}n^{\frac{1}{2}}\left\{\int_{I^2} g dC_n(w_1, w_2) - \int_{I^2} g dC(w_1, w_2)\right\} &= n^{\frac{1}{2}}\left\{\int_{I^2} g d\tilde{C}_n(u,v) - \int_{I^2} g d\tilde{C}(u,v)\right\} \\ &= n^{\frac{1}{2}}\left\{\zeta(\tilde{C}_n(u,v)) - \zeta(\tilde{C}(u,v))\right\} \\ &= \bar{\zeta} \\ &\Leftrightarrow \bar{\zeta} \quad \underrightarrow{D} \quad \zeta'_c(L)\end{aligned}$$

where $\zeta'_c(L) = \int_{I^2} g(u,v) d\left(L(u,v)\right)$ is the derivative of ζ in the point c (see [62]). □

Part II

Survival Copula parameters estimation for Archimedean family under singly censoring

Survival Copula parameters estimation for Archimedean family under singly censoring

Idiou Nesrine[1] Benatia Fatah[2]

Abstract

Given $(Z_i, \delta_i) = \left\{\min(T_i, C_i), I_{(T_i<C_i)_{i=1,2}}\right\}$, as dependent or independent right-censored variables. As a logical continuation of results established by N.IDIOU et al (2021) [68], a particular case of right-censoring has well detailed, as well as the empirical survival copula has also evaluated in this case of singly-censoed data. As an application, two Archimedean copula models have been chosen to illustrate our theoretical results. A simulation study follows, which sheds light on the behavior of the process estimation method shown that the proposed estimator performs well in terms of relative bias and RMSE. The methodology of the proposed estimator is also illustrated by using real lifetime data from the Diabetic Retinopathy Study, where its efficiency and robustness are observed.

Keywords: Archimedean copulas, Individually censoring data, Moment estimator, Survival copula, Semi-parametric estimation.

Résumé

Soit $(Z_i, \delta_i) = \left\{\min(T_i, C_i), I_{(T_i<C_i)_{i=1,2}}\right\}$, en tant que dépendantes ou indépendantes variables censurées à droite. Dans la suite logique des résultats établis par N.IDIOU et al (2021) [68], un cas particulier de la censure à droite est bien détaillée, ainsi que la copule de survie empirique a également évaluée dans ce cas de données censurées individuellement. Comme application, deux modèles de copules d'Archimède ont été choisis pour illustrer notre résultats théoriques. Une étude de simulation suit, qui met en lumière le comportement de la méthode d'estimation de processus, a montré que l'estimateur proposé fonctionne bien en termes de biais relatif et de RMSE. La méthodologie de l'estimateur proposé est également illustrée en utilisant des données de durée de vie réelles du Diabetic et de rétinopathie, où son efficacité et sa robustesse sont observées.

Mots clés : copules d'Archimède, Données censurées simplement, estimateur des moments, copule de survie, Estimation semi-paramétrique.

AMS Subject Classification: 62G05, 62G20.

5.7 Introduction

In the medical domain, researchers were mostly confronted with competing risk issues, that is, event times may be dependent and they are censoring each other [94], [4], [16], [46], [47]. Likewise, in survival analyses

it is popular to observe two or more lifetimes for the same customer, patient, or equipment. For example, the lifetimes of a pair of organs can be observed in a pair of kidneys, an ear, or an eye in patients, or the lifetimes of engines in a two-engine vehicle. In most cases, these variables are related and this pattern of bivariate data is well-suited to the copula model, particularly the Archimedean one.

As an outcome, we suggest that the two failure times T_1 and T_2 can be modelled by an Archimedean copula model and it is subject to dependence or independence right-censoring with the censoring vector (C_1, C_2), we also propose that the vector (C_1, C_2) follows an arbitrary bivariate continuous distribution. Hence, we can only observe $Z_i = \min(T_i, C_i)$, $\delta_i = I_{\{T_i \leq C_i\}_{i=1,2}}$ where $I_{(.)}$ represents the indicator function.

Sometimes, the problem in right-censoring is how to model the dependence concept among bivariate censoring vectors (T_1, T_2) and (C_1, C_2), when both variables are censored at the same time (see [62]). The issue now is how to construct the dependency structure between this vector when only one variable is right-censored. Let's look at the bivariate pattern (T_1, T_2), with the joint distribution function (df) $F(t_1, t_2) = P(T_1 \leq t_1, T_2 \leq t_2)$, which can be presented by the following form $F(t_1, t_2) = C(F_1(t_1), F_2(t_2))$, where F_1, F_2 are two continuous margins and C is the related copula function ordinarily known for all (u, v) in $\mathbb{I}^2$ by:

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

when $F^{-1}(u) = \inf\{x \in \mathbb{R} : F(x) \geq u\}$ is the generalized inverse of a non-decreasing function F. A joint survival function S of (T_1, T_2) is said to have an Archimedean association dependence structure if for all $t_1, t_2 \geq 0$, it can be interpreted as follows $S(t_1, t_2) = \varphi^{-1}(\varphi(S_1(t_1)) + \varphi(S_2(t_2)))$, where φ is a continuous and convex function defined on $\mathbb{I} \to [0, \infty]$ with $\varphi(1) = 0$, and S_1, S_2 are the marginal survival functions of T_1 and T_2 respectively (see [71], [32], and [33]).

In the context of multivariate survival analysis, many models have been proposed to model multivariate survival data among them, Archimedean copulas models (see [4], [92] and [93]). Specifically, for the couple $(u, v) \in \mathbb{I}^2$, an Archimedean copula is noted as $C(u, v) = \varphi^{-1}(\varphi(u) + \varphi(v))$, where φ^{-1} is the inverse function of φ and φ usually called the Archimedean generator of C. Hence, the function $\tilde{C}$ define from $\mathbb{I}^2 \to \mathbb{I}$, which couples S_1 and S_2, known by the survival copula of (T_1, T_2) via

$$\tilde{C}(u, v) = u + v - 1 + C(1 - u, 1 - v), \tag{5.7.1}$$

(Nelsen, (2006) [67]). Supposing that (T_1, T_2) follows an Archimedean copula, where S_1 and S_2 are the marginal survival functions, Genest and Rivest (1993) [33], have proved that $U = \frac{\varphi(S_1(T_1))}{\varphi(S_1(T_1)) + \varphi(S_2(T_2))}$ and $V = \tilde{C}(S_1(T_1), S_2(T_2)) = \varphi^{-1}(\varphi(S_1(T_1)) + \varphi(S_2(T_2)))$, are independently distributed random variables when U follows a uniform distribution on $\mathbb{I}$ and V follows a so-called Kendall distribution with the density function: $k_C(t) = \frac{\varphi(t)\varphi''(t)}{(\varphi'(t))^2}$ defined on $(0, 1]$, as a function of t depends on the unidentified parameter θ.

The main aim of this paper is to present a new semi-parametric estimation procedure and its application to health-related survival data, given (T_1, T_2) as individually censored. General formulas for all possible parameters estimate of a survival copula $\tilde{C}$ are also presented under the assumption that the copula is Archimedean.

Important results are reviewed in section 2, where general formulas are proposed for the marginal survival functions of T_1 and T_2. As an application of our results, a simple way of the estimation of the unknown parameters is declared in section 3, where an estimator of V based on the classical moments method is proposed, followed by two examples of Clayton and Gumbel copula models. Under the Archimedean dependence structure assumption for censored data, a simulation study evaluates the performance of our estimator presented in Section 4, relatively on bias and RMSE, where the robustness and efficiency of the estimator are proven. In section 5, we illustrate the methodology presented in section 3 on real data from the Diabetic Retinopathy Study, which is available in the "survival" package (see [73] and [86]), of the R software. Our paper ends with some discussions in Section 6.

5.8 Important results

Assume that (T_1, T_2) are two positive random variables whose distributions can be modelled by an Archimedean copula either dependently or independently right-censored by a censoring vector (C_1, C_2) that follows an arbitrary bivariate continuous distribution. Take the available observation in the case of absence data $(Z_{1i}, Z_{2i}, \delta_{1i}, \delta_{2i})_{1\leq i\leq n}$: the independent copies of a non-negative random variable of the vector $(Z_1, Z_2, \delta_1, \delta_2)$. As a result, the variable $Z_i = \min(T_i, C_i)$ is only observed when $T_i \leq C_i$ for $i = 1, 2$, then $\delta_i = I_{\{T_i\leq C_i\}_{i=1,2}}$ equal to one which represents the indicator function of censored data.

Considering only T_1 is right-censored, in other words, $C_2 = \infty$ almost surely, which is a particular situation from another case of doubly right-censored (see [62]). In this case, the empirical distribution function is

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

this model was studied by Stute (1993) [85], who suggested the empirical distribution function $\tilde{F}_n$ given by:

$$\tilde{F}_n(t_1, t_2) = \frac{1}{n}\sum_{i=1}^{n} \frac{\delta_{1i}}{\hat{S}_1\left(Z_{1i}^{-}\right)} 1_{\{Z_{1i}\leq t_1, Z_{2i}\leq t_2\}}, \tag{5.8.1}$$

which is a consistent estimator of F, known as a particular model situation from the case given in section (5.2), where we can take $\tilde{C}\left(\hat{S}_1\left(Z_{1i}\right), \hat{S}_2\left(Z_{2i}\right)\right) = \hat{S}_1\left(Z_{1i}\right)\hat{S}_2\left(Z_{2i}\right)$ and for any right continuous function $M\left(t^{-}\right)$, when M defined from $\mathbb{R}$ in $\mathbb{R}$ we set $M\left(t^{-}\right) = \lim_{n\to\infty} M(t - \frac{1}{n})$ the left-hand limit of M at t when it exists (see [68]).

Recognizing it was provided that

$$\hat{S}_1(t) = \prod_{k/Z'_{1k}<t} (1 - \frac{\sum_{i=1}^{n} 1_{\{Z_{1i}=Z'_{1k},\delta_{1i}=0\}}}{\sum_{i=1}^{n} 1_{\{Z_{1i}\geq Z'_{1k}\}}}),$$

where $\hat{S}_1$ as the Kaplan-Meier estimate of S_1 and $((Z'_{1k})_{1\leq k\leq m}, m \leq n)$ is the distinct values of $(Z_{1i})_{1\leq i\leq n}$.

Suppose that the copula C is twice continuously differentiable and the variable T_1's support is lower than the variable T_2's support. Following Gribkova and Lopez (2015) [39] and noted that

$$F_{1n}(t_1) = \lim_{t_2\to\infty} F_n(t_1, t_2), \quad F_{2n}(t_2) = \lim_{t_1\to\infty} F_n(t_1, t_2),$$

the empirical copula function C_n have estimated by:

$$C_n(u, v) = \frac{1}{n} \sum_{i=1}^{n} \frac{\delta_{1i}}{\hat{S}_1(Z_{1i}^-)} 1_{\{F_{1n}(Z_{1i})\leq u, F_{2n}(Z_{2i})\leq v\}}, \quad (u, v) \in [0, 1]^2$$

The weak convergence of C_n has proved under some assumptions (see [68]). Hence, the empirical survival copula of such form:

$$\tilde{C}_n(u, v) = u + v - 1 + \frac{1}{n} \sum_{i=1}^{n} \frac{\delta_{1i}}{\hat{S}_1(Z_{1i}^-)} 1_{\{\bar{F}_{1n}(Z_{1i})\geq u, \bar{F}_{2n}(Z_{2i})\geq v\}} \tag{5.8.2}$$

where $(u, v) \in [0, 1]^2$. The reader is invited to take a look on the references mentioned below ([68] and [42]).

Following [68], the asympthotic normality of the empirical survival copula $\tilde{C}_n$, can be proven for a singly censored under some assumptions and by the same manner as seen in Theorem (5.2.2) . Because the dependence between T_i and C_i, $i = 1, 2$ can be modeled by an Archimedean copula, Wang and Oakes (2008), proved that the distribution function of V formulated by

$$F(v, c_1, c_2) = \frac{1}{\tilde{C}(c_1, c_2)} \left\{ v - \frac{\varphi(v) - \varphi(\tilde{C}(c_1, c_2))}{\varphi'(v)} \right\}, \ 0 \leq v \leq \tilde{C}(c_1, c_2), \tag{5.8.3}$$

where T_1 and T_2 are both right-censored [95]. By analogy, when only one variable is censored the distribution function of V become as follows:

- $F_1(v, c_1, t_2) = \frac{\varphi'(\tilde{C}(c_1,t_2))}{\varphi'(v)}, 0 \leq v \leq \tilde{C}(c_1, t_2)$, when only T_1 is right-censored
- $F_2(v, t_1, c_2) = \frac{\varphi'(\tilde{C}(t_1,c_2))}{\varphi'(v)}, 0 \leq v \leq \tilde{C}(t_1, c_2)$, when only T_2 is right-censored.

Proof. see [95]. □

By the way used (5.8.3), the k^{th} moments of V in the case of doubly right censoring have established by:

$$\begin{aligned}\mathbb{E}(V^k \Big| T_1 > c_1, T_2 > c_2) = & \frac{\left(\tilde{C}(c_1, c_2)\right)^k}{k+1} \\ & - k\left(\tilde{C}(c_1, c_2)\right)^{k-1} \varphi\left(\tilde{C}(c_1, c_2)\right) \int_0^1 \frac{v^{k-1}}{\varphi'\left(v\tilde{C}(c_1, c_2)\right)} dv \\ & + k\left(\tilde{C}(c_1, c_2)\right)^{k-1} \int_0^1 \frac{v^{k-1}\varphi\left(v\tilde{C}(c_1, c_2)\right)}{\varphi'\left(v\tilde{C}(c_1, c_2)\right)} dv,\ k \geq 1\end{aligned}$$

Proof. see [68]. □

We can recall this corollary based on N.IDIOU et al's results (see [68]).

Corollary 5.8.1 (IDIOU, N. et al 2021) *Let (T_1, T_2) be a random pair whose distribution can be modelled by an Archimedean copula. Assuming that (T_1, T_2) is subject to dependent or independent right censoring by a censoring vector (C_1, C_2) that follows an arbitrary bivariate continuous distribution, then we have:*

1. *For $k \geq 1$, the k^{th} moments of V when only T_1 is right-censored is*

$$\begin{aligned}\mathbb{E}(V^k \Big| T_1 > c_1, T_2 = t_2) = & \left(\tilde{C}(c_1, t_2)\right)^k \\ & - k\left(\tilde{C}(c_1, t_2)\right)^k \varphi'\left(\tilde{C}(c_1, t_2)\right) \int_0^1 \frac{v^{k-1}}{\varphi'\left(v\tilde{C}(c_1, t_2)\right)} dv.\end{aligned}$$

2. *For $k \geq 1$, the k^{th} moments of V when only T_2 is right-censored is*

$$\begin{aligned}\mathbb{E}(V^k \Big| T_1 = t_1, T_2 > c_2) = & \left(\tilde{C}(t_1, c_2)\right)^k \\ & - k\left(\tilde{C}(t_1, c_2)\right)^k \varphi'\left(\tilde{C}(t_1, c_2)\right) \int_0^1 \frac{v^{k-1}}{\varphi'\left(v\tilde{C}(t_1, c_2)\right)} dv.\end{aligned}$$

5.9 Parameters estimation under singly right censored variable

We propose a simple way of estimating the unknown parameters for Archimedean copula models. We set up the procedure based on the classical moments method, that we have seen before in section(5.3) [68]. Assume that $Z_{1:n} < ... < Z_{n:n}$, the order statistics, pertaining to the sample $\{Z_i, \delta_i; 1 \leq i \leq n\}$ with their associated concomitants $\delta_{[i:n]}, ..., \delta_{[n:n]}$.

Then, $\delta_{[j:n]} = \delta_i$ if $Z_{j:n} = Z_i$ for $1 \leq j \leq n$. Since we are focusing on the datasets that contain extreme values which include distributions such as Burr, Fréchet, generalized Pareto...etc. However, the selected Pareto model is well known as a heavy-tailed censored data model and it is obvious that the heavy-tailed distribution class plays a significant role in the theory of extreme value. Then it would be natural to assume that both survival functions $S_1 = 1 - F_1$ and $S_2 = 1 - F_2$ are regularly varying at infinity with tail indices $\gamma_1 > 0$ and $\gamma_2 > 0$ respectively. In another word, if we

assume that both F_1 and F_2 are heavy-tailed (mentioned that F_1 and F_2 are completely known), so there exist two constants $\gamma_1 > 0$ and $\gamma_2 > 0$ such that:

$$\lim_{t\to\infty} \frac{S_1(tx)}{S_1(t)} = x^{\frac{-1}{\gamma_1}} \quad \text{and} \quad \lim_{t\to\infty} \frac{S_2(tx)}{S_2(t)} = x^{\frac{-1}{\gamma_2}}, \quad \text{for } x > 0$$

By a logical sequence and since F_1 and F_2 are heavy-tailed, then the censoring distribution is assumed to be heavy tailed too (i.e: the CDF of the observed Z's noted by H and given by $\bar{H} = S_1 S_2$ is heavy-tailed too), hence:

$$\lim_{t\to\infty} \frac{\bar{H}(tx)}{\bar{H}(t)} = x^{\frac{-1}{\gamma}}, \quad \text{for } x > 0$$

Therefore, the extreme value index of the distribution function (d.f) of (Z, δ) denoted by γ and given by $\gamma = \frac{\gamma_1\gamma_2}{\gamma_1+\gamma_2}$.

Let (T_1, T_2) two random variables whose distribution can be modelled by an Archimedean copula and is subject to dependent or independent singly right-censoring, $V = \tilde{C}\left(S_1(t_1), S_2(t_2)\right)$ is a conditionally distributed variable follows a so-called Kendall distribution K_C with the density function: $k_C(t) = \frac{\varphi(t)\varphi''(t)}{(\varphi'(t))^2}$, defined on $(0, 1]$.

We take only T_1 right-censored and we noted $M_k(V|c_1, t_2)$ the k^{th}-moments of V, then:

$$M_k(V|c_1, t_2) = E(V^k|T_1 > c_1, T_2 = t_2), \quad \text{for } k \geq 1$$

Relying on the results in Corollary (5.8.1) we have:

$$\begin{aligned} M_k(V|c_1, t_2) = E(V^k|T_1 > c_1, T_2 = t_2) &= \left(\tilde{C}\left(c_1, t_2\right)\right)^k \\ -k\left(\tilde{C}\left(c_1, t_2\right)\right)^k \varphi'\left(\tilde{C}\left(c_1, t_2\right)\right) &\int_0^1 \frac{v^{k-1}}{\varphi'\left(v\tilde{C}\left(c_1, t_2\right)\right)} dv. \end{aligned} \tag{5.9.1}$$

Assuming that V belongs to a parametric family $V_\theta = \tilde{C}_\theta(u, v)_{\theta\in\mathbb{R}^d}$, then it follows that $\varphi = \varphi_\theta$ and $K_C = K_\theta$, for the unknown parameter $\theta \in \mathbb{R}^d$.

If we suppose that $M_k(V|c_1, t_2) = M_k(\theta|c_1, t_2)$, the equation (5.9.1) can be written as:

$$M_k(\theta|c_1, t_2) = \left(\tilde{C}_\theta\left(c_1, t_2\right)\right)^k - k\left(\tilde{C}_\theta\left(c_1, t_2\right)\right)^k \varphi'_\theta\left(\tilde{C}_\theta\left(c_1, t_2\right)\right) \int_0^1 \frac{v_\theta^{k-1}}{\varphi'_\theta\left(v_\theta\tilde{C}_\theta\left(c_1, t_2\right)\right)} dv_\theta.$$

Because of the copula symmetry, the equation (2) in Corollary (5.8.1) via :

$$M_k(\theta|t_1, c_2) = \left(\tilde{C}_\theta\left(t_1, c_2\right)\right)^k - k\left(\tilde{C}_\theta\left(t_1, c_2\right)\right)^k \varphi'\left(\tilde{C}_\theta\left(t_1, c_2\right)\right) \int_0^1 \frac{v_\theta^{k-1}}{\varphi'\left(v_\theta\tilde{C}_\theta\left(t_1, c_2\right)\right)} dv_\theta.$$

which shows the k^{th}moments of V, when only T_2 is right-censored. Given, the empirical version of the moment estimator presented by $\hat{M}_k(\hat{V}|H_j)$:

$$\hat{M}_k(\hat{V}|H_j) = \frac{1}{N}\sum_{i=1}^{n}(\tilde{C}_n(\hat{S}_i\left(t_i\right))|H_j)^k, \text{ for } k \geq 1, \ j = 1, 2.$$

Where $\hat{V}$ is the survival empirical copula $\tilde{C}_n$ and H_j represent each case of censoring. Then, as the natural estimators of moments copula, it is necessary to solve the equation system given by:

$$M_k(\theta|H_j) = \hat{M}_k(\hat{V}|H_j), \text{ for } \theta = (\theta_1, ..., \theta_d) \text{ and } j = 1, 2.$$

To obtain the unique solution of $\hat{\theta}^{SCCM} = (\hat{\theta}_1, ..., \hat{\theta}_d)$ called the singly censored copula moment (SCCM) estimator of θ.

5.10 Application: illustrative examples

From now, only T_1 is considered as a censored variable. Therefore, two models evaluated, the first is for the Clayton model of one-parameter and the second is for the Gumbel model of two parameters.

- **Clayton model**

  For the Clayton model of one-parameter, the survival copula is known by

  $$\tilde{C}_\alpha(u, v) = u + v - 1 + ((1-u)^{-\alpha} + (1-v)^{-\alpha} - 1)^{\frac{-1}{\alpha}},$$

  with generator $\varphi_\alpha(t) = t^{-\alpha} - 1$, $\alpha > 0$. Applying Corollary (5.8.1), we can simplify the estimating equations as follow:

  $$\mathbb{E}(V^k \Big| T_1 > c_1, T_2 = t_2) = (m)^k - km^k \varphi'(m) \int_0^1 \frac{v^{k-1}}{\varphi'(vm)} dv, \quad (5.10.1)$$

  for $k > 0$ and when $m = \tilde{C}(c_1, t_2)$, represent the ordinary copula. If we simplify more the formula (6) we can obtain:

  $$\mathbb{E}(V^k \Big| T_1 > c_1, T_2 = t_2) = (m)^k - km^{k-\alpha-1} \int_0^1 \frac{v^{k-1}}{(vm)^{-\alpha-1}} dv$$

  By an elementary calculation, we get the kth moments

  $$M_k(\alpha) = m^k - \frac{km^{k-1}}{k+\alpha+1},$$

  where $m = \tilde{C}(c_1, t_2)$. Hence, for $k = 1$ the first moments is normally given by:

  $$M_1(\alpha) = m - \frac{1}{\alpha+2}$$

  Then, as the natural estimators of moments copula, it is necessary to solve the equation system given by:

  $$M_1(\alpha) = \hat{M}_1$$

  which allows us easily find the unique estimator of α given by:

  $$\hat{\alpha} = 2 - \frac{1}{m - \hat{M}_1}$$

- **Gumbel model**

  For the Gumbel model of two parameters, where the data are singly right-censored (suppose that is only T_1 in this case). As a result, by using the procedure given in section (5.9), the two first moments M_1, M_2 are given by:

$$\begin{cases} M_1(\theta|\, c_1, t_2) = m - \frac{(\beta-1)(m^{-\alpha}-1)^{\beta}}{\alpha m^{2\alpha+1}} \frac{\Gamma(1-\beta)\Gamma\left(\frac{1}{\alpha}(\alpha\beta+2)\right)}{\Gamma\left(\frac{2}{\alpha}(\alpha+1)\right)} = M_1(\alpha,\beta) \\ M_2(\theta|\, c_1, t_2) = m^2 - \frac{2(\beta-1)(m^{-\alpha}-1)^{\beta}}{\alpha m^{2\alpha+1}} \frac{\Gamma(1-\beta)\Gamma\left(\frac{1}{\alpha}(\alpha\beta+3)\right)}{\Gamma\left(\frac{1}{\alpha}(2\alpha+3)\right)} = M_2(\alpha,\beta) \end{cases}$$

  and the estimator $\hat{\theta}$ of θ is the unique solution of the system:

$$\begin{cases} M_1(\theta) = \hat{M}_1 \\ M_2(\theta) = \hat{M}_2 \end{cases}$$

We will talk about these two models (Clayton and Gumbel) in more detail in the upcoming simulation section when we will see how they contrast.

5.11 Simulation studies

Taking into consideration that only T_1 to be right-censored. A simulation study was carried out to evaluate the performance of the proposed estimators, based on the Monte Carlo procedure under the Clayton and Gumbel Archimedean dependence assumption. The results are shown in Tables (5.4-5.8), we first generate bivariate data from the Clayton and Gumbel models of T_1 and T_2 with Pareto margins of parameters γ_1 and γ_2 respectively. We also generate the censoring variable C_1 whose marginal distribution is a Pareto with γ_c parameter.

We suppose that $\gamma_1 = \gamma_2 = 0.3$ and that the corresponding percentage of observed data is given by $p_1 = \frac{\gamma_c}{\gamma_1+\gamma_c}$, we choose parameter values corresponding to p_1 values $0.95, 0.90, 0.85, 0.80$, and we solve the equation $p_1 = \frac{\gamma_c}{\gamma_1+\gamma_c}$ to get the pertaining γ_c-values. Based on the parameters estimate procedure in Section (5.9), 1000 replicas to be generated for each common size n varied for $n = 30, 50, 100, 500, 1000$, to pick our final performance as empirical evidence of the results gained across all replicates.

Table (5.4), describes the results obtained for the Clayton model of one-parameter (5.10.1), with unit Pareto margins of shape parameter (0.3), whose estimator looked with:

$$\hat{\alpha} = 2 - \frac{1}{m - \hat{M}_1},$$

where, we can see the R.Bias and the RMSE are very close to zero. Once the rate of dependence τ is increased, we see an improvement in the results of the estimated parameters $\hat{\alpha}$ due to a large decrease in R.Bias and RMSE, which are inversely proportional readings.

Table 5.4 – *Moments estimator performance based on Clayton survival copula of one parameter under singly right-censored variable.*

$\tau = 0.05$, $\alpha = 0.1$								
N	$n = 30$		$n = 100$		$n = 500$		$n = 1000$	
censure	R.Bias	RMSE	R.Bias	RMSE	R.Bias	RMSE	R.Bias	RMSE
5%	-0.0547	0.0640	-0.0549	0.0631	-0.0535	0.0626	-0.0542	0.0625
10%	-0.0532	0.0621	-0.055	0.0632	-0.0529	0.0620	-0.0548	0.0642
15%	-0.0554	0.0642	-0.0539	0.0632	-0.0532	0.0623	-0.0541	0.0630
20%	-0.0536	0.0624	-0.0525	0.0614	-0.0530	0.0619	-0.0553	0.0638
$\tau = 0.5$, $\alpha = 0.2$								
5%	-0.0257	0.0298	-0.0253	0.0291	-0.0255	0.0299	-0.027	0.0311
10%	-0.0262	0.0304	-0.0260	0.0299	-0.0264	0.0305	-0.0267	0.0308
15%	-0.0262	0.0302	-0.0254	0.0297	-0.0258	0.0299	-0.0254	0.0295
20%	-0.0257	0.0299	-0.0257	0.0300	-0.0257	0.0299	-0.0261	0.0302
$\tau = 0.7$, $\alpha = 0.4$								
5%	-0.0130	0.0149	-0.0125	0.0145	-0.0131	0.0150	-0.0127	0.0148
10%	-0.0126	0.0145	-0.0126	0.0146	-0.0123	0.0144	-0.0129	0.0149
15%	-0.0132	0.0151	-0.0124	0.0144	-0.0125	0.0145	-0.0126	0.0146
20%	-0.0127	0.0147	-0.0130	0.0149	-0.0127	0.0148	-0.0126	0.0146

Now, by considering the second model of the Gumbel survival copula of two parameters, where the two first moments are formulated as (5.10). Given Kendall's tau:

$$\tau_{\alpha,\beta} = 4E(V_{\alpha,\beta}) - 1,$$

as an association index (a function of the dependency parameter in Archimedean copula models), we select the survival copula parameter values (α, β) that correspond to specified values of τ by using the select values $0.05, 0.5$ and 0.7 of Kendall's tau dependence assumption values and the transformed of the underlying survival Gumbel copula

$$V_{\alpha,\beta} = u + v - 1 + \left(\left(\left((1-u)^{-\alpha} - 1 \right)^{\beta} + \left((1-v)^{-\alpha} - 1 \right)^{\beta} \right)^{1/\beta} + 1 \right)^{-1/\alpha}$$

as shown in Table (5.5).

Tables (5.6-5.8) shows the results obtained of SCCM estimator $(\hat{\alpha}, \hat{\beta})$ of (α, β) based on survival copula under the censored variable T_1, generated from the Gumbel copula model of two parameters given in section (5.10) with unit Pareto margins of shape parameter (0.3).

By looking at three different values of dependency weak (0.05) moderate (0.5) and strong (0.7), the R.Bias and the RMSE of the two parameters estimate $\hat{\alpha}$ and $\hat{\beta}$ were calculated and are usually given lower values especially when the dependency increases.

Table 5.5 – *The true parameters of the survival Gumbel copula transformed using Kendall's tau.*

τ	α	β
0.05	0.1	1.00
0.5	0.2	1.82
0.7	0.4	2.78

Table 5.6 – *Moments estimator performance based on Gumbel survival copula of two parameters under singly right censored variable (T_1) generated from 1000 replications with unit Pareto margins and shape parameter (0.3). Relative bias and RMSE of the estimators a are calculated for different censoring values and for weak dependence.*

$$\tau = 0.05,\ \alpha = 0.1 \Rightarrow \beta = 1.00$$

Sample Size	c_1	$\hat{\alpha}$		$\hat{\beta}$	
		R.Bias	RMSE	R.Bias	RMSE
% of censoring 20					
30	0.11681	-0.05381	0.06285	0.13100	0.01145
50	0.07223	-0.05428	0.06293	0.09702	0.01185
100	0.03696	-0.05455	0.06302	0.12730	0.01179
500	0.00796	-0.05241	0.06136	0.11346	0.01171
1000	0.00411	-0.05310	0.06181	0.11126	0.01178
% of censoring 15					
30	0.15716	-0.05364	0.06258	0.14563	0.01153
50	0.09968	-0.05506	0.06414	0.17947	0.01192
100	0.05382	-0.05405	0.06289	0.14270	0.01125
500	0.01168	-0.05338	0.06219	0.12725	0.01156
1000	0.00591	- 0.05223	0.06111	0.13337	0.01216
% of censoring 10					
30	0.22207	-0.05170	0.06076	0.16483	0.01172
50	0.14458	-0.05233	0.06110	0.16214	0.01164
100	0.08509	-0.05438	0.06300	0.14554	0.01151
500	0.01801	-0.05562	0.06412	0.16037	0.01148
1000	0.00892	-0.05353	0.06223	0.14950	0.0117
% of censoring 5					
30	0.31617	-0.05347	0.06244	0.17334	0.01160
50	0.25135	-0.05312	0.06190	0.17205	0.01181
100	0.15138	-0.05363	0.06259	0.17993	0.01146
500	0.03666	-0.05295	0.06245	0.1544	0.01162
1000	0.01841	-0.05247	0.06124	0.15832	0.01170

Table 5.7 – *Moments estimator performance based on Gumbel survival copula of two parameters under singly right censored variable* (T_1) *generated from 1000 replications with unit Pareto margins and shape parameter 0.3. Relative bias and RMSE of the estimators a are calculated for different censoring values and for moderate dependence.*

$\tau = 0.5,\ \alpha = 0.2 \Rightarrow \beta = 1.82$

Sample Size	c_1	$\hat{\alpha}$		$\hat{\beta}$	
		R.Bias	RMSE	R.Bias	RMSE
% of censoring 20					
30	0.12019	-0.02632	0.03015	0.06926	0.00642
50	0.07406	-0.02598	0.03007	0.08484	0.00645
100	0.03829	-0.02558	0.02983	0.08158	0.00641
500	0.00804	-0.02591	0.03011	0.10276	0.00637
1000	0.00394	-0.02579	0.02979	0.08549	0.00643
% of censoring 15					
30	0.15611	-0.02605	0.03027	0.0989	0.00657
50	0.10008	-0.02498	0.02906	0.09554	0.00633
100	0.05478	-0.02615	0.03037	0.12623	0.00641
500	0.01088	-0.02536	0.02957	0.09632	0.00631
1000	0.00546	-0.02605	0.03012	0.08785	0.00632
% of censoring 10					
30	0.21223	-0.2556	0.02972	0.17075	0.00637
50	0.14818	-0.02544	0.02948	0.09582	0.00644
100	0.08047	-0.02526	0.02945	0.13741	0.00644
500	0.01811	-0.02714	0.03121	0.09872	0.00636
1000	0.00915	-0.02675	0.03072	0.09565	0.00637
% of censoring 5					
30	0.30813	-0.02537	0.02939	0.16263	0.00644
50	0.23892	-0.02488	0.02905	0.19983	0.00646
100	0.15392	-0.02639	0.03044	0.09317	0.00640
500	0.03543	-0.02568	0.02995	0.10464	0.00630
1000	0.01844	-0.02517	0.02931	0.10486	0.00637

Table 5.8 – *Moments estimator performance based on Gumbel survival copula of two parameters under singly right censored variable* (T_1) *generated from 1000 replications with unit Pareto margins and shape parameter 0.3. Relative bias and RMSE of the estimators a are calculated for different censoring values and for strong dependence.*

$\tau = 0.7,\ \alpha = 0.4 \Rightarrow \beta = 2.78$

Sample Size	c_1	$\hat{\alpha}$		$\hat{\beta}$	
		R.Bias	RMSE	R.Bias	RMSE
% of censoring 20					
30	0.11466	-0.01275	0.01476	0.14167	0.00377
50	0.07528	-0.0126	0.01465	0.10573	0.00393
100	0.03747	-0.01287	0.01483	0.08515	0.00389
500	0.00787	-0.01251	0.01446	0.11032	0.00391
1000	0.00421	-0.01285	0.01478	0.08713	0.00385
% of censoring 15					
30	0.15831	-0.01260	0.01459	0.14577	0.00394
50	0.10332	-0.01282	0.01485	0.10507	0.00387
100	0.05540	- 0.01303	0.01502	0.13646	0.00385
500	0.01108	-0.01221	0.01427	0.10486	0.00391
1000	0.00590	-0.01276	0.01473	0.10388	0.00389
% of censoring 10					
30	0.21203	-0.01244	0.01443	0.12436	0.00384
50	0.14910	-0.01261	0.01460	0.16765	0.00384
100	0.08393	-0.01285	0.01475	0.15731	0.00389
500	0.01785	-0.01252	0.01455	0.1095	0.00395
1000	0.00872	-0.01297	0.01509	0.11251	0.00392
% of censoring 5					
30	0.31116	-0.0126	0.01457	0.1568	0.00391
50	0.24405	-0.01236	0.01448	0.17432	0.00389
100	0.15254	-0.01259	0.01466	0.10309	0.00383
500	0.03668	-0.01306	0.01504	0.11537	0.00387
1000	0.01838	-0.01271	0.01469	0.10606	0.00386

5.12 Application to a Real Data Set

In this part of the paper, we examine the performance of our estimation procedure given in section (5.3), for a real data set of diabetic retinopathy, which is available in the R software via the survival package (see Ophthalmol, AM.J (1976) and Therneau, T.M (2015)). Diabetic retinopathy is a disease that affects people with diabetes and can outcome in vision loss and blindness. In the study, a significant number of diabetic patients (times of follow-up for 197 diabetic patients under 60 years old, who are at high-risk of vision loss) was followed for an extended period. The primary aim of the study was to assess the efficacy of photocoagulation as a treatment for proliferative retinopathy. For each patient, one eye was treated with laser photocoagulation, and the other eye was taken as a control.

a. Scatter plots b. Cens and obs data for T_1 c. Cens and obs data for T_2

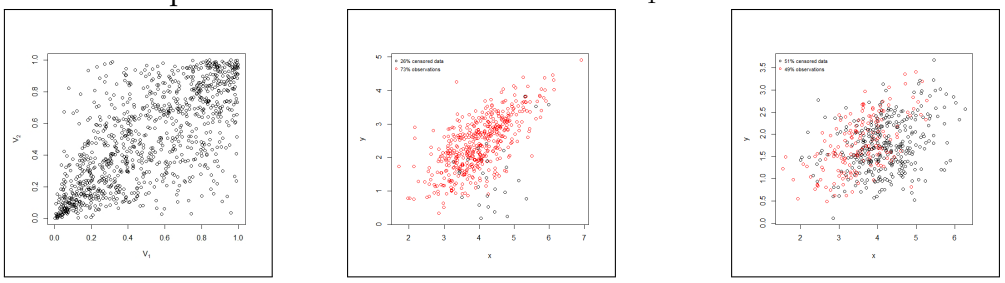

Figure 5.2 – *Censored and observed points for each T_1 and T_2 separately of bivariate survival Gumbel copula.*

To model this data, the evaluation that piques our interest is concerned by the case when the two variables are both censored. To fit the failure times (T_1, T_2), we use a bivariate Gumbel family of two parameters with extreme value margins (Pareto $(\gamma = 0.3)$) for both T_1 and T_2. Taking T_1 as the time to a visual loss for the treatment eye and T_2 the time to visual loss for the control eye.

In the R software (the survival package), the percentage of uncenored times for T_1 is 73% (143 observations) and 49% (96 observations) for T_2. To model this data in our case, we performed a censoring adjustment to the percentage of censoring time, assumed for $5, 10, 15, 20\%$ for T_1 and T_2 who have the same tau of censorship the purpose of which is to prove the efficiency of the estimator in several cases of censoring.

We ran the algorithm presented in section (5.3), by considering Kendall's tau as the association index (a function of the dependency parameter in this application is considered to be the correlation between the two visual loss times (T_1, T_2) for the treatment eye and the control eye). To assess the performance of the considered estimator, we have used the RMSE and the relative bias (R.Bais) define by:

$$\text{R.Bais} = \frac{1}{N}\frac{\left[\sum_{i=1}^{N}\hat{\theta}_i - \theta\right]}{\theta}, \quad \text{RMSE}=\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\hat{\theta}_i - \theta\right)^2}$$

Table 5.9 – *Relative bias and RMSE of Moments estimator based on a Gumbel survival copula model from the Diabetic Retinopathy study data.*

$$\tau = 0.7\,,\, \alpha = 0.1 \Rightarrow \beta = 2.78$$

sample Size	% of cens	$\hat{\alpha}$ R.Bias	$\hat{\alpha}$ RMSE	$\hat{\beta}$ R.Bias	$\hat{\beta}$ RMSE	Assoc τ before cens	Assoc τ after cens
$n = 50$	5%	-0.0126	0.0146	0.3941	0.0042	0.6967	0.6412
	10%	-0.0126	0.0146	0.4147	0.0041	0.7056	0.5925
	15%	-0.0128	0.0148	0.3547	0.0041	0.7002	0.5373
	20%	-0.0125	0.0145	0.3440	0.0042	0.7001	0.4930
$n = 100$	5%	-0.0125	0.0146	0.4243	0.0043	0.7004	0.6415
	10%	-0.0122	0.0143	0.3818	0.0041	0.6994	0.5849
	15%	-0.0133	0.0152	0.3613	0.0042	0.7016	0.5376
	20%	-0.0130	0.0149	0.3585	0.0042	0.6982	0.4859
$n = 500$	5%	-0.0125	0.0146	0.4102	0.0042	0.7007	0.6426
	10%	-0.0131	0.0150	0.3784	0.0043	0.7000	0.5872
	15%	-0.0126	0.0146	0.3648	0.0041	0.7000	0.5369
	20%	-0.0130	0.0150	0.3415	0.0042	0.6998	0.4885
$n = 1000$	5%	-0.0127	0.0147	0.4077	0.0042	0.7003	0.6409
	10%	-0.0128	0.0148	0.3841	0.0042	0.7001	0.5881
	15%	-0.0125	0.0145	0.3573	0.0042	0.6999	0.5361
	20%	-0.0125	0.0145	0.3388	0.0042	0.7003	0.4891

where $\hat{\theta}_i$ is the CCM estimator (from the considered model) of θ. In Figure (5.2), (a): shows the scatter plots of the survival Gumbel copula with two parameters, (b) and (c): shows the censored and observed data for each variable T_1 and T_2 respectively.

Table (5.9) shows the relative bias (R.Bias) and the RMSE of the parameters estimates under different doubly right-censoring values. The correlation between these two times was supposed to be strong $\tau = 0.7$ (it can be assumed to be lower in the same way and it also gives good results for the estimator). This association dependency value presented before and after censoring (Assoc τ before cens, Assoc τ after cens). For this data set, the estimator gave the smaller relative bais and RMSE values, which proves its effectiveness.

5.13 Conclusion and perspective

In this paper, we have presented a semi-parametric estimation method of a survival copula $\tilde{C}$ based on the classical method of moments under individually censored of (T_1, T_2). As a logical continuation of results established by Idiou et al (2020) [68], general formulas are given for marginal survival copula $\tilde{C}$ of such data by the assumption that their underlying copula is Archimedean. Two models are proposed for this study, the Clayton model of one-parameter and the Gumbel model of two-parameters proved our theoretical results obtained. Under the Archimedean dependence structure assumption for censored data, a simulation study evaluates the performance of our estimator, relative bias, and RMSE formulas for estimator are evaluated. This study shows that the new estimator

works well, where the values obtained are tending towards zero for each case of small and even large samples. The methodology presented in section 5.3, was applied to real data from the Diabetic Retinopathy Study, which is available in the "survival" package [73],[86], of the R software. For this data set, the estimator gave the smaller relative bias and RMSE values, which proves its effectiveness and robustness.

Consequently, this method is preferable if we compare it with the maximum likelihood method and other methods ([5], [71]), because of its easy analytical mathematical form.

Our main result for this study is based on the copula approaches and the survival analysis, under the Archimedean dependence structure assumption for censored data. Based on these results, we can establish a new methods checking process of Archimedean copula models for singly right-censored data. This is one of our recent research areas and the idea was already established in another paper that is under preparation.

5.14 Appendix

Proof of Corollary (5.8.1)

For $k > 1$ the k^{th} moments is defined by:

$$\mathbb{E}(V^k \Big| T_1 > c_1, T_2 = t_2) = \int_0^{\tilde{C}(c_1,t_2)} v^k dF_1(v, c_1, t_2),$$

based on theorem given by Wang, we use the conditional distribution of V when only T_1 is censored $(V | T_1 > c_1, T_2 = t_2)$, we have

$$\begin{aligned}\mathbb{E}(V^k \Big| T_1 > c_1, T_2 = t_2) &= \int_0^{\tilde{C}(c_1,t_2)} v^k dF_1(v, c_1, t_2) \\ &= \int_0^{\tilde{C}(c_1,t_2)} v^k \left\{ \frac{-\varphi''(v)\,\varphi'\left(\tilde{C}(c_1,t_2)\right)}{(\varphi'(v))^2} \right\} dv \\ &= I\end{aligned}$$

To simplify I we pass directly to integration by parts, and we have:

$$I = \left(\left[v^k \frac{\varphi'\left(\tilde{C}(c_1,t_2)\right)}{\varphi'(v)} \right]_0^{\tilde{C}(c_1,t_2)} - k \int_0^{\tilde{C}(c_1,t_2)} v^{k-1} \frac{\varphi'\left(\tilde{C}(c_1,t_2)\right)}{\varphi'(v)} dv \right),$$

it follows by changing variables:

$$I = (\tilde{C}(c_1,t_2))^k - k(\tilde{C}(c_1,t_2))^k \varphi'\left(\tilde{C}(c_1,t_2)\right) \int_0^1 \frac{v^{k-1}}{\varphi'(v\tilde{C}(c_1,t_2))} dv$$

Which is the k^{th} moments of the variable V, where only T_1 is censored. Because of the copula's symmetry, the same proof may be used to obtain the k^{th} moments of the variable V using equation (2) of Corollary (5.8.1), where only T_2 is censored.

Copulas and frailty models in multivariate survival data 6

Sommaire

In this chapter, we are interested by copula modeling and its applications in the analysis of multivariate survival data. We have used the frailty model for bivariate survival data by considering Archimedean copulas. Our main idea in this chapter focused to introducing the dependence between the survival times $T_1, ..., T_d$, using an unobserved random variable W, called frailty model with variable latent. We then focused on the particular cases of Clayton-Oakes copulas and the model with frailty gamma-type.

Copulas and frailty models in multivariate survival data

Idiou Nesrine[1], Benatia Fatah[2], Mesbah Mounir[3]

Abstract

In mathematics and statistics areas, the modeling of Copulas and their estimation is one of the most important research fields. Copulas have recently become a very important average in the structure dependence modeling between marginal distributions and joint distribution of a couple of random variables. In this paper, we are interested by Copula modeling and its applications in the analysis of multivariate survival data. In particular, the Archimedean copula familly. As well, our main idea is also focused to introducing the dependence between the survival times $T_1, ..., T_d$, using an unobserved random variable W, called frailty model with variable latent. This paper ended with an application presented on bivariate survival data in biostatistics fields, analyzed by the Copula procedure of the SAS software.

Index Term Copula, Survival analysis, Fraitly model, Archimedean Copulas.

Résumé

Dans le domaine des mathématiques et des statistiques, la modélisation des copules et leur estimation est l'un des domaines de recherche les plus importants. Les copules sont récemment devenues une moyenne très importante dans la modélisation de la dépendance de structure entre les distributions marginales et la distribution conjointe d'un copule de variables aléatoires. Dans cet article, nous présentons une synthèse des travaux récents portant sur cette théorie et ses applications à l'analyse des données de survie multivariée. De plus, notre idée principale est également axée sur l'introduction de la dépendance entre les temps de survie $T_1, ..., T_d$, en utilisant une variable aléatoire non observée W, appelée modèle de fragilité avec variable latente. Enfin, à titre d'illustration, une application sur des données de survie bivariée issue de la littérature dans des domaines de biostatistique est présentée, et analysée par la procédure Proc Copula du logiciel SAS.

Terme d'indice Copule, Analyse de survie, Modèle de fragilité, Copules archimédiennes.

6.1 Introduction

Recently, considerable attention has been paid to the problem of inference about copulas, the term copula comes from the Latin word " copũlae ",

which means a bond, bond, or union. Among the most important statistical works in copula theory are those of Hoeffding, (1940) [43], (1941) [44], who used copulas to study measures of non-parametric associations. He thus obtained optimal inequalities, providing upper and lower bounds for particular versions of copulas, cited in the Theorem (bounds of Fréchet Hoeffding (1957)) [29]. The monographs Deheuvels (1979) [19], Cook and Johnson (1981) [17], Cherubini et al. (2004) [14], Nelsen (2006) [67], Joe (1997) [49] and Genest (1993) [33] summarize some extent activities in this area.

Basically, a copula function is a function which joins or couples multivariate distribution functions to their univariate marginal distribution functions, which is confirmed by Sklar (1959) [84], in a theorem bearing his name, shows that, under certain conditions, there is a unique copula function C such as :

$$F(x_1, ..., x_d) = C(F(x_1), ..., F(x_d)) \tag{6.1.1}$$

In various fields of statistics, this function is critical for modeling dependency as (finance, actuarial science, and more recently in biology and health, etc.). In this context, particularly in the area of health, the study of the links between the dates of occurrence of a disease, its possible date of recovery, relapse, and death, several multivariate survival models taking into account the dependence between random variables are based on the notion of copulas, often without making explicit reference to them. The nature of these problems in survival analysis leads to constructing a family's model of multivariate survival functions from univariate marginal survival functions.

Our aim for this article is to introduce the copula approach to multivariate survival modeling, this approach appears implicitly in Clayton (1978) [16], who was one of the first to suggest a survival analysis of the bivariate association model, and in Marshall and Olkin (1988) [61].

In the epidemiological context, frailty models are models involving an individual parameter. These epidemiological models assume that a subject may be more brittle than another and therefore have a greater risk of death or another pathological event. These fragility models were introduced by Lancaster (1979) [54], who used a proportional risk model. Vaupel et al (1979) [91], suggested the application of the gamma model with another frailty model. Gamma distributions were used due to mathematical attractiveness. They are well known and have simple densities. The application of these models in survival is contemporary (Clayton, (1978)) [16]. The main idea is to introduce a dependence between the survival d-times $T_1, ..., T_d$, using an unobserved random variable W.

The article is organized as follows. The second section, is devoted to a brief presentation of univariate and multivariate survival models, and copulas specific to them. Section three, is concacred for the copula model using the Clayton model. In section four, we consider the approach of multivariate survival models with frailty variable and present the copulas associated with this type of model. In the fifth section, we use the Proc

Copula procedure of the SAS software to analyze real data of recurrent durations in hemodialysis.

6.2 Survival models

Let T be a survival time (also called variable of interest), from now we noted F, f, S, the distribution function, the density function, and the survival function for a fixed time t respectively. Usually in survival modeling, one of the main concepts is the instantaneous hazard rate or the risk function $\lambda(t)$ for a fixed instant t given by the formula (4.1.1) (Lancaster, 1990) [52]. It can be interpreted as the instantaneous rate of death, another expression of λ is known by:

$$\lambda(t) = \frac{f(t)}{S(t)},$$

the link between the cumulative hazard function Λ and S is given by the following relation:

$$S(t) = \exp(-\Lambda(t)),$$

which allows us to write:

$$f(t) = \lambda(t)\exp(-\Lambda(t)) = \lambda(t)S(t).$$

Another important concept is the baseline hazard function $\lambda_0(t)$ (Frees and Valdez (1998) [28]). It intervenes in particular in the widely used model known by Cox's model (Cox (1972) [18]), it entails modeling the risk function $\lambda(t)$, thus:

$$\lambda(t) = \exp(X\beta^T)\lambda_0(t),$$

where X is a vector of covariates, and β is a vector of parameters associated with these covariates. This regression model (a proportional risks model) makes it possible to analyze the distribution of durations as a function of covariates. He is part of a larger family $\lambda(t) = g\left(X\beta^T\right)\lambda_0(t)$, where $g(.)$ any function, presented in Cox's model as the exponential function.

One of the interests of this model is the interpretation of the parameters. Take the example of a covariate X_j which can take two values: 0 if the individual is taking treatment A and 1 if he is taking treatment B. The coefficient β_j, or rather $exp(\beta_j)$ is the instantaneous risk of death relative to treatment B compared to treatment A. In the multivariate case, the survival function $S(t)$ is defined by:

$$S(t_1, ..., t_d) = P[T_1 > t_1, ..., T_d > t_d],$$

where $T_1, ...T_d$ are a d-survival times. The univariate marginal survival functions $S_j(t_j)$, are noted by:

$$\begin{aligned} S_j(t_j) &= P\{T_j > t_j\} \\ &= S(0, ..., 0, t_j, 0, ..., 0) \end{aligned}$$

In this article we assume that the survival times are continuous and take their values in $\mathbb{R}_+$. Noting that the relation between the multivariate

survival function S and the multivariate distribution function F, is not so trivial that in the univariate case:

$$S(t_1, ..., t_d) \neq 1 - F(t_1, ..., t_d).$$

On the other hand, the hazard rate and the multivariate hazard function are given respectively by:

$$\begin{aligned}\lambda(t_1, ..., t_d) &= \lim_{\max \Delta_{j \to 0}} \frac{P[t_1 \leq T_1 \leq t_1 + \Delta_1, ...|T_1 \geq t_1, ...]}{\Delta_1 ....\Delta_d} \\ &= \frac{f(t_1, ..., t_d)}{S(t_1, ..., t_d)}\end{aligned}$$

$$\Lambda(t_1, ..., t_d) = \int_0^{t_1} ... \int_0^{t_d} \lambda(s_1, ..., s_d) ds_1 ... ds_d$$

So, we can conclude that the relation between S and Λ cannot be formulated simply as in the univariate case. For example, we obtain in the bivariate case:

$$S(t_1, t_2) = S_1(t_1) S_2(t_2) e^{-\Lambda(t_1, t_2)}$$

The construction of a multivariate survival function is not convenient when using functions risk directly, as it is generally based on complex conditional risk rates (Shaked and Shanthikumar (1987) [82]).

6.3 Copula models

a. 1 Survie bivariée

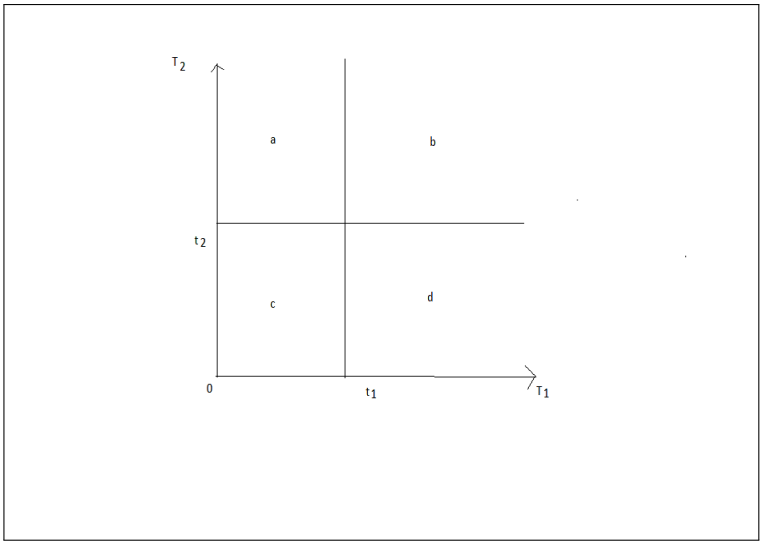

Figure 6.1 – $F_1(t_1) = a + c; F_2(t_2) = c + d; S(t_1, t_2) = b; F(t_1, t_2) = c$

A multivariate distribution can be constructed through the use of copulas in a survival setting. In most cases, we are looking for the lifetime of statistical members in a certain population, which gives particular importance to this copula. Here we define a particular copula that is associated

with this notion of survival and we are mainly interested by the bivariate case. Let's $F(x_1, x_2)$ the joint distribution function of the random pair (X, Y), where $C_{X,Y}$ is the copula of (X, Y), then the survival function of the couple (X, Y) is:

$$\begin{aligned} S(x,y) &= P(X > x, Y > y) \\ &= 1 - P(X \leq x \text{ or } Y \leq y) \\ &= 1 - F(x) - G(y) + F(x,y) \\ &= S_1(x) + S_2(y) - 1 + C_{X,Y}\left(1 - \overline{F}(x), 1 - \overline{G}(y)\right), \end{aligned}$$

where $S_1(x) = 1 - F(x)$ and $S_2(y) = 1 - G(y)$ the marginal survival functions of X and Y respectivly. So if we define a function $\widetilde{C}$ in $I^2 \to I$:

$$\tilde{C}(u,v) = u + v - 1 + C(1-u, 1-v) \tag{6.3.1}$$

We obtain

$$S(x,y) = \widetilde{C}\left(S_1(x), S_2(y)\right).$$

Noted that: $S(x,y) \neq 1 - F(x,y)$. The application of Sklar's theorem to survival functions is immediate: any multivariate survival function is written $S(t_1, .., t_d) = \tilde{C}(S_1(t_1), ..., S_d(t_d))$, where $\tilde{C}$ is a copula, and S_1,...,S_d are the marginal survival functions.

6.3.1 Example: Clayton model

Clayton [1978] considers a bivariate association model for an ordered pair of individuals. For (T_1 and T_2) the ages of the first and second member of the pair, Clayton introduces a function $\theta(t_1, t_2)$ defined as:

$$\theta(t_1, t_2) = \frac{\lambda(t_1 | T_2 = t_2)}{\lambda(t_1 | T_2 \geq t_2)}$$

This function is interpreted as the ratio of the risk rate of the conditional distribution of T_1 given $T_2 = t_2$, to that of T_1, given $T_2 \geq t_2$.

$$\begin{aligned} \lambda(t_1 | T_2 = t_2) &= -\frac{\partial_1 S_1(t_1 | T_2 = t_2)}{S_1(t_1 | T_2 = t_2)} \\ &= \frac{\partial_{1,2} S(t_1, t_2)}{\partial_1 S(t_1, t_2)} \end{aligned}$$

and

$$\lambda(t_1 | T_2 \geq t_2) = -\frac{\partial_1 S(t_1, t_2)}{S(t_1, t_2)},$$

then

$$\theta(t_1, t_2) = -\frac{\partial_{1,2} S(t_1, t_2) \times S(t_1, t_2)}{\partial_1 S(t_1, t_2) \times \partial_2 S(t_1, t_2)} = -\frac{f(t_1, t_2) \times S(t_1, t_2)}{\partial_1 S(t_1, t_2) \times \partial_2 S(t_1, t_2)}$$

Clayton assumes that $\theta(t_1, t_2)$ is constant and equal to a parameter θ, $\theta > 0$. Then, we have:

$$\frac{\partial_{1,2} S(t_1, t_2)}{S(t_1, t_2)} - \theta \frac{\partial_1 S(t_1, t_2)}{S(t_1, t_2)} \times \frac{\partial_2 S(t_1, t_2)}{S(t_1, t_2)} = 0$$

The survival function $S(t_1, t_2)$ is therefore the solution of the non-linear partial differential equation of second order:

$$\partial_{1,2}\pi(t_1, t_2) + (\theta - 1)\,\partial_1\pi(t_1, t_2) \times \partial_2\pi(t_1, t_2) = 0$$

where $\pi(t_1, t_2) = -\ln\left(S(t_1, t_2)\right)$. Clayton showed that the solution is of the form:

$$S(t_1, t_2) = \left[1 + (\theta - 1)\left(a_1\left(t_1\right) + a_2\left(t_2\right)\right)\right]^{-\frac{1}{\theta-1}}$$

where a_1 and a_2 are two non-decreasing functions satisfying $a_1(0) = a_2(0) = 0$, we will give the canonical representation of this survival function. The univariate marginals of $S(t_1, t_2)$ are respectively:

$$S_1\left(t_1\right) = S_1\left(t_1, 0\right) = \left[1 + (\theta - 1)\left(a_1\left(t_1\right)\right)\right]^{-\frac{1}{\theta-1}}$$

and

$$S_2\left(t_2\right) = S_1\left(0, t_2\right) = \left[1 + (\theta - 1)\left(a_2\left(t_2\right)\right)\right]^{-\frac{1}{\theta-1}}$$

noted $S_j\left(t_j\right) = u_j$, $j = 1; 2$, it is easy to show that:

$$t_j = a_j^{-1}\left\{\frac{u_1^{1-\theta} - 1}{\theta - 1}\right\}$$

then

$$S_j^{-1}\left(u_j\right) = a_j^{-1}\left\{\frac{u_1^{1-\theta} - 1}{\theta - 1}\right\},$$

where $a_j^{-1}(.)$ and $S_j^{-1}(.)$ the reciprocal functions (and not inverse power) of a_j and S_j respectively. The survival copula associated with the Clayton model is therefore:

$$\begin{aligned}
\tilde{C}\left(u_1, u_2\right) &= S(S_1^{-1}\left(u_1\right), S_2^{-1}\left(u_2\right)) \\
&= \left[1 + (\theta - 1)\left\{\frac{u_1^{1-\theta} - 1}{\theta - 1} + \frac{u_2^{1-\theta} - 1}{\theta - 1}\right\}\right]^{-\frac{1}{\theta-1}} \\
&= \left\{u_1^{1-\theta} + u_2^{1-\theta} - 1\right\}^{-\frac{1}{\theta-1}}
\end{aligned}$$

By noting $\alpha = 1 - \theta$, we find the Archimedean copula of Clayton.

6.4 Frailty model

The frailty model, often known as the frailty model or model with frailty, is a conditional risk model containing a multiplicative factor. In the context of health, this term implies that one patient may be more brittle than another, exposing them at a higher risk of death (or worsening of his disease) than one other.

In this model, a random parameter (called the frailty parameter) is introduced that could be shared by a group of patients (group effect).

Integrating this frailty parameter (having a function of appropriate density and its corresponding Laplace transform) in the bivariate survival distribution conditional yields the joint survival function from the conditional risk model, where the joined survival functions take the form of an Archimedean copula. On the foundation it is sometimes claimed that the frailty model corresponds to a specific model of Archimedean copulas based on this observation (Manatunga and Oakes (1999) [1], Viswanathan and Manatunga (2001) [90], Andersen (2005) [2]). This assertion, though, is confusing because the two modeling approaches are so different in nature.

The main idea is to introduce the dependence between the survival times $T_1, ..., T_d$, using an unobserved random variable W, called frailty. This corresponds to the modeling of a variable latent (or hidden). Conditional on the frailty W of distribution G, the survival times are assumed to be independent. Then, the conditional survival function is given by:

$$\begin{aligned} S(t_1, ..., t_d|w) &= P[T_1 > t_1, ..., T_d > t_d|W = w] \\ &= \prod_{j=1}^{d} P[T_j > t_j|W = w] \\ &= \prod_{j=1}^{d} S_j[t_j|W = w]. \end{aligned}$$

Thus, the unconditional survival function can be determined as:

$$\begin{aligned} S(t_1, .., t_n) &= E(E(S(t_1, .., t_n|W)) \\ &= \int S(t_1, .., t_n|w)dG(w). \end{aligned}$$

We require Marshall and Olkin's Theorem [61], in order to have a more interesting representation of frailty models.

Theorem 6.4.1 (Marshall et Olkin (1988) [61]) *Let $F_1, ...F_d$ the univariate distribution functions, and G a d-variable distribution function such that $G(0, ..., 0) = 1$, with the univariate marginal function G_j, $j = 1, ..., d$. Noted the Laplace transform of G, φ and also of G_j is φ_j. Let C a d-varied distribution function with all univariate marginals uniform over $\mathbb{I}$, if $H_j(x) = \exp(-\varphi_j^{-1}(F_j(x)))$, then:*

$$F(x_1, ..., x_d) = \int C[H_1(x_1)^{w_1}, ..., H_d(x_d)^{w_d}]dG(w_1, ..., w_d) \tag{6.4.1}$$

is a d-variable distribution function with marginals $F_1, ...F_d$.

Marshall and Olkin (1988), afterward have studied a particularly interesting and simple case of (6.4.1), (see [35], page 15). Then, the expression (6.4.1)) becomes:

$$F(x_1, ..., x_d) = \varphi_1(\varphi_1^{-1}(F_1(x_1)) + ... + \varphi_1^{-1}(F_d(x_d))) \tag{6.4.2}$$

It is a particular case of an Archimedean copula where the generator φ is the inverse of the Laplace transform. We can now state the definition of the survival functions of frailty.

Definition 6.4.1 *A survival function is said to be frailty if it is written as:*

$$S(t_1,..,t_d) = \tilde{C}(S_1(t_1),...,S_d(t_d))$$

where $\tilde{C}$ is an Archimedean copula with a generator that corresponds to the inverse of the Laplace transform of the frailty variable's distribution W. The generator is the inverse of a Laplace transform in more generic terms.

6.4.1 Bivariate survival copula and frailty model

The modeling of bivariate survival data is the emphasis of this section. Consider two "survival times" (T_1, T_2), which correspond to two durations for establishing a diagnosis carried out on the same individual by two different techniques, for example, T_1 for radiography (radiographic (RX)) and T_2 for ultrasound (l'échographie (US)). Let $S_1(t)$ and $S_2(t)$ be the marginal survival functions for each method (Goethals et al, [37]). A frailty model is given by:

$$\lambda_{ij}(t) = w_i \lambda_{j,w_i}(t),$$

- $\lambda_{ij}(t)$ the instantaneous risk function at time t for an individual $i = 1,...,n$, with diagnostic technique $j = 1, 2$.
- $\lambda_{j,w}(t)$ the risk function at time t for an individual whose frailty is equal to w and the technique diagnostic j.
- w_i the term of the frailty of individual i.

To define the copula models and frailty models, we need a particular Archimedean copula family where the generator φ is the inverse of the Laplace transform:

$$C(u,v) = \varphi\left\{\varphi^{-1}(u) + \varphi^{-1}(v)\right\},$$

where $\varphi(0) = 1$, and $\varphi^{-1}(.)$ is the inverse of this generator, so we only need a function family $\varphi(.)$. Let $g_W(.)$ be the density of the frailty r.v., defined on the support $[0, \infty[)$ and $\varphi_W(s)$, its Laplace transform given by:

$$\begin{aligned}\varphi_W(s) &= E\left\{\exp(-sw)\right\} \\ &= \int_0^\infty \exp(-sw) g_W(w) dw.\end{aligned}$$

Thus, the conditional survival function of W, is written as:

$$S_W(t_1,t_2) = \tilde{C}(S_{1,W}(t_1), S_{2,W}(t_2)),$$

which can also presented by:

$$S_W(t_1,t_2) = \varphi_W\left\{\varphi_W^{-1}(S_{1,W}(t_1)) + \varphi_W^{-1}(S_{2,W}(t_2))\right\} \tag{6.4.3}$$

For the frailty model, the conditional survival function is given by:

$$S_{W_i}(t_1,t_2) = \exp[-w_i\left\{\Lambda_{1,w_i}(t_1) + \Lambda_{2,w_i}(t_2)\right\}], \quad i = 1,...,n$$

where $\Lambda_{j,w_i}(t) = \int_0^t \lambda_{j,w_i}(s)ds,\ j = 1, 2$, the cumulative hazard function.

As a result, by integrating the frailty in relation to the frailty density a frailty models joint survival function can be expressed as:

$$\begin{aligned} S(t_1, t_2) &= \int_0^\infty S_W(t_1, t_2) g_W(w) dw \\ &= E[\exp\{-W(\Lambda_{1,w}(t_1) + \Lambda_{2,w}(t_2))\}] \end{aligned}$$

Then, $S(t_1, t_2)$ becomes:

$$S(t_1, t_2) = \varphi\{\Lambda_{1,w}(t_1) + \Lambda_{2,w}(t_2)\} \tag{6.4.4}$$

Since the marginal survival function can be written as:

$$S_j(t) = \varphi\{\Lambda_{j,W}(t)\} \Rightarrow \Lambda_{j,w}(t) = \varphi^{-1}(S_j(t)) \tag{6.4.5}$$

hence, if we replace (6.4.5) in (6.4.4), the joint survival function of the frailty model becomes:

$$S(t_1, t_2) = \varphi\left\{\varphi^{-1}(S_1(t)) + \varphi^{-1}(S_2(t))\right\} \tag{6.4.6}$$

From (6.4.3) and (6.4.6), we notice that the two models are different in nature because the copula used in the joint survival functions in (6.4.3) and (6.4.6) is the same but the marginal survival functions are not even the same.

6.4.2 Clayton-Oakes copula and gamma frailty model

The copula function for the Clayton-Oakes model, is the joint survival function of a frailty model whose Laplace transform is that of an r.v. with gamma distribution, i.e. $\varphi_\theta(s) = (1 + \theta s)^{\frac{-1}{\theta}}$.

Consider again the example of the durations corresponding to two diagnostic techniques $j = 1; 2$ (see [34]). Laplace transform of a gamma density that have a single parameter θ and its inverse $\varphi^{-1}(s)$ are given by :

$$\varphi_\theta^{-1}(s) = \frac{(s^{-\theta} - 1)}{\theta},\ \ \theta \geqq 0$$

The joint survival copula function of the Clayton Oakes model can be calculated directly using (6.4.3):

$$\begin{aligned} S_c(t_1, t_2) &= C_\theta(S_{1,c}(t_1), S_{2,c}(t_2)) \\ &= [\{S_{1,c}(t_1)\}^{-\theta} + \{S_{2,c}(t_2)\}^{-\theta} - 1]^{\frac{-1}{\theta}} \end{aligned}$$

The joint survival function for this frailty model becomes:

$$S_m(t_1, t_2) = [1 + \theta\{\Lambda_{1,u}(t_1) + \Lambda_{2,u}(t_2)\}]^{-\frac{1}{\theta}}$$

which allows us to write:

$$S_m(t_1,t_2) = \left\{1 + [(S_{1,m}(t_1))^{-\theta} - 1] + [(S_{2,m}(t_2))^{-\theta} - 1]\right\}^{-\frac{1}{\theta}}$$
$$= \left\{(S_{1,m}(t_1))^{-\theta} + (S_{2,m}(t_2))^{-\theta} - 1\right\}^{-\frac{1}{\theta}}$$

This expression resembles the copula form shown previously, but

$$S_{j,m}(t) \neq S_{j,c}(t), \quad \forall j = 1,2.$$

See Genest and Werker (2002), for more details.

6.5 Application to hemodialysis data

In this section, we use hemodialysis data published by McGilchrist and Aisbett (1991), and analyzed by frailty models. The occurrence of infections in patients with renal failure who were on hemodialysis motivated this study, this is said by purifying toxins from the blood produced by the body and eliminating them through an artificial filter.

The catheter, a hollow plastic tube that the doctor inserts into a vein during hemodialysis, can become infected. The catheter is removed and the infection is cured after an infection is discovered. A catheter is returned for the next hemodialysis, etc. For each patient, the duration between each date insertion of the catheter and subsequent infection are observed. Only two observations per patient are considered. There may also be censoring if during the study period one or both infections do not happen.

We will not deal with the censored case here and will consider all times as observed. McGilcrist and Aisbett treat the censored case with a frailty model. For this, we use the new SAS procedure (ProcCopula), to estimate the association parameter and then select the copula that best fits the data. Future work will examine the copula model's processing of these bivariate data in the presence of censoring.

The histograms in Figure (6.2), are those of the marginal durations of the two recurrence times. The continuous curves represent the theoretical fitted densities assumed to be Weibull's. We acknowledge a pretty good fit.

a. Recurrence 1 b. Recurrence 2

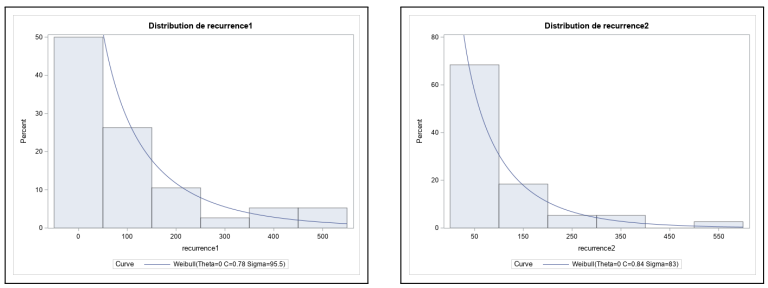

Figure 6.2 – *Empirical histogram (Weibull) and fitted densities of the two recurrence times.*

The corresponding empirical marginal distribution functions at recurrences 1 and 2 are shown in Figure (6.3). The updated Weibull distribution functions from the estimated parameters are represented by the continuous lines.

a. Recurrence 1 b. Recurrence 2

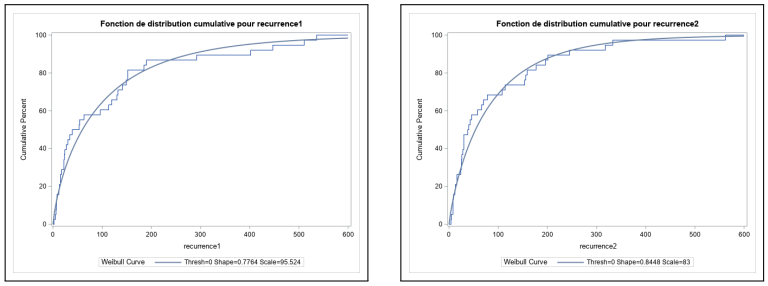

Figure 6.3 – *Empirical distribution functions and their estimates (Weibull) of the two recurrence times.*

The Pearson, Spearman and Kendall (tau) correlation coefficient estimates are 0.07522 (with a degree of significance $p = 0.6535$), $0.01040(p = 0.9506)$ and $0.01004(p = 0.9298)$ respectively. The Clayton copula parameter θ is estimated to be $0,00000010536712$.

a. Scatter plots b. Clayton Copula: outlines c. Clayton Copula: area

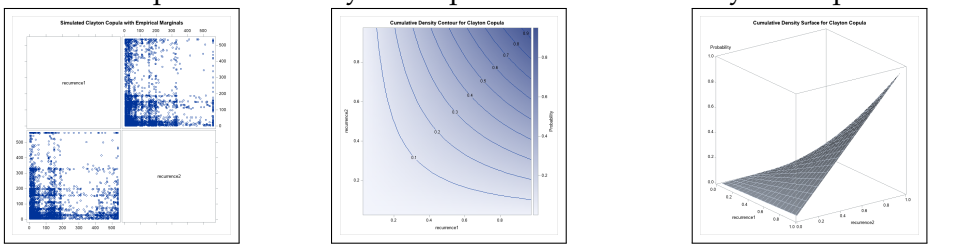

Figure 6.4 – *Bivariate empirical distribution and associated graphics of Clayton Copula.*

The point clouds ((t_1, t_2) and (t_2, t_1)) (a)) are visualized in Figure (6.4), as well as the contour lines of the bivariate distribution (b) and the surface corresponding to it (c) estimated by a Clayton copula. Frank's and Gumbel's copulas can be obtained in the same way.

Table (6.1) summarizes the results concerning the association parameter θ estimated under different choices of copula models (Clayton, Frank, and Gumbel). The Akaike Criterion (AIC) allows you to select the optimal model for your data (smallest AIC value). The choice of Gumbel's copula looks the best.

Table 6.1 – *Associated estimation parameter for three models of Copula under Akaike penalization criterion*

Copule	θ	Std. Error	AIC
Clayton	$1,054 \times 10^{-7}$	0	2,0
Frank	0,150408	1,031073	1,97873
Gumbel	1,039736	0,127503	1,89611

The two recurrence times are weakly correlated. These bivariate data could be analyzed as univariate data. We have presented here a succinct analysis of the association between the two recurrence times. We did not analyze the effects of the covariates. It would be interesting to compare the analysis of the effect of these covariates by considering the data as univariate to that taking into account the bivariate aspect modeled by a Gumbel copula, retained here by the AIC. It would also be interesting to compare the covariate effects to those obtained using a model Cox regression with frailty.

The SAS programs used are given in the appendix.

6.6 Conclusion and perspectives

This article's purpose was to present copulas and their essential properties, as well as their application to survival data. Copulas can thus be used to model the relationships between the components of multivariate survival in a simple and natural way. An additional tool often used for modeling multivariate survival data is the introduction of parameters individual random factors often interpreted as parameters of frailty. In this work, we have used this model for bivariate survival data considering Archimedean copulas. We then focused on the particular cases of Clayton-Oakes copulas and the model with gamma-type frailty. For each of these two models, the copulas used for the functions of bivariate survival are the same. However, the marginal survival functions are modeled in ways different.

Then we moved on to health-related survival data applications. Survival data can also come from reliability studies in the industry. The presence of censorship, particularly in the univariate case, is a significant challenge in the study of this sort of data. In the bivariate case, this question remains largely open. Data analysis of bivariate survival and censored by copulas is the subject of ongoing work.

In the field of insurance (actuarial) and more generally in finance, the theory of copulas has been very successful, and in connection with the theory of extreme values, many copulas have been built and used.

This work also overlooks statistical inference issues, which are often complex, especially in presence of censorship. Recent research in the field of copulas is more often statistical than theoretical nature, F. Lounas (2011) [59], Deheuvels (1979) [19], Genest (1987) [34], Genest and Rivest (1993) [33], are excellent references on this subject.

In the realm of medicine, survival data, or continuous quantitative data, is widely employed for diagnostic or prognostic purposes. Quite often we also find data of a nature qualitative and discreet (Fontaine, 2017) [26]. This article prompts several perspectives for future work and the development of copula-based models for this type of data.

In the field of copulas, research has also been applied to the development of computer aspects, through the publication of programs in the R language and recently, SAS with "Proc Copula". We used "Proc Copula" to process the hemodialysis data. Unfortunately, this procedure cannot process censored data. Future and useful work will be writing a macro SAS going in this direction.

The frailty variables, considered here, are latent, not observed, but one-dimensional. In the example presented, this variable characterized the effect of the individual on the recovery time currency. These individuals could come from several hospitals. The differential effect, not observed, of these centers would then be a latent variable. A future research perspective would be the modeling and analysis by copulas of multivariate lifetimes with latent variables themselves multivariate. Processing such data is extremely complicated, regardless of computer type.

6.7 Appendix

Program SAS: Proc Copula applied on recurrency hemodialysis data

```
proc univariate data=recurrences; var recurrence1 recurrence2;
histogram recurrence1 recurrence2/weibull;run;
proc univariate data=recurrences;var recurrence1 recurrence2;
cdfplot recurrence1 recurrence2/weibull;run;
proc corr data=recurrences kendall pearson spearman;
var recurrence1 recurrence2;run;
proc copula data=recurrences;var recurrence1 recurrence2;
fit clayton/marginals=empirical;
simulate /ndraws = 5000 seed = 12345678 marginals=empirical
plots = (distribution=cdf) out = fic1;run;
proc copula data=recurrences; var recurrence1 recurrence2;
fit frank/marginals=empirical;
simulate /ndraws = 5000 seed = 12345678 marginals=empirical
plots = (distribution=cdf) out = fic1; run;
proc copula data=recurrences; var recurrence1 recurrence2;
fit gumbel/marginals=empirical;
simulate /ndraws = 5000 seed = 12345678 marginals=empirical
plots = (distribution=cdf) out = fic1; run;
```

Outlook

We close this thesis with some research perspectives that could be the subject of future research.

- Based on the outcomes of Gripkova and Lopez's (2015)[39], Lopez and Saint-Pierre's (2012)[72], research, our results presented in chapter four can be applied for mixed censoring. This is one of our current research topics and the idea is to develop another document that is under preparation.
- Based on the results given in chapter four, we can establish a new semi-parametric method checking process of Archimedean copula models for various censoring patterns (singly or doubly censored). This is one of our recent areas of research and the idea to establish in another document that is also under preparation.
- The frailty variables, considered in chapter five, are latent, unobserved, but one-dimensional. A future research perspective would be the modeling and analysis by copulas of multivariate survival times with latent variables themselves multivariate.
- Establish the asymptotic normality of the estimator obtained in chapter four, by considering the two cases of censoring proposed.
- Study the same method of moments with other types of censored data (mixed censoring, censored by intervals,...).

Bibliography

[1] A.K. Manatunga, and Oakes, D. (1999). Parametric Analysis for Matched Pair Survival Data, Springer Link.

[2] Andersen, E.W. (2005). Two-stage estimation in copula models used in family studies, Lifetime Data Analysis, 11(3), pp. 333-350.

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

[4] Bandeen-Roche, K.J., Liang, K.-Y. (1996). Modelling failure-time associations in data with multiple levels of clustering. Biometrika 83, 29-39.

[5] Bhattacharyya, G.K., JOHNSON, R. A. (1973).: Maximum Likelihood Estimation and Hypothesis Testing in the Bivariate Exponential Model of Marshall and Olki. Journal of the American Statistical Association, 68(343), pp. 704-706.

[6] Benatia.F et al. (2011). A semiparametric estimation procedure for multi-parameter Archimedean copulas based on the L-moments method. Africa Statistika, Numéro: 335-345.

[7] Brahimi, B., Necir, A. (2012). A semiparametric estimation of copula models based on the method of moments. Journal of Statistical Methodology V9, Issue 4, July 2012, Pages 467-477.

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

[9] Cai, Z. (1998). Asymptotic properties of kaplan-meier estimator for censored dependent data. Statistics & Probability Letters, 37: 381-389.

[10] Cai, Z. (2001). Estimating a distribution function for censored time series data. J. Multivariate Anal, 78: 299-318.

[11] Cai, Z. and Roussas, G. (1992). Uniform strong estimation under -mixing with rates. Statistics & Probability Letters, 15-47.

[12] Capéraà, P., Fougères, A.L. and Genest, C. (2000). Bivariate Distributions with Given Extreme Value Attractor. Journal of Multivariate Analysis 72, 30-49.

[13] Chaieb L., Lajmi M. (2006). Estimation de la dépendance et de lois marginales dans des modèles pour l'analyse des durées de vies multidimensionnelles. Thèse Ph. D., Université de LAVAL, QUEBEC.

[14] Cherbini, U., Luciano, E., Vecchiato, W. (2004). Copula Methods in Finance, Wiley.

[15] Chine, A., Benatia.F, (2017).Bivariate copulas estimation using the trimmes L-moments methode. Africa Statistika, Numéro: 1185-1197.

[16] Clayton, D.G. (1978). A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence, Biometrika, 65(1), 141-151.

[17] Cook, R.D. and Johnson, M.E. (1981). A family of distributions for modelling nonelliptically symmetric multivariate data, Royal Statistical Society. 43, 210-218.

[18] Cox, D.R. (1972). Regression models and life-tables (with discussion), Journal of the Royal Statistical Society -Series B, 34, 187-220

[19] Deheuvels, P. (1979). La fonction de dépendance empirique et ses propriétés. Acad. Roy. Belg. Bull. Cl. Sci. 65, 274-292.

[20] Diehl, S. et Stute, W. (1988). Kernel density and hazard function estimation in the presence of censoring. Journal of Multivariate Analysis, 25: 299-310.

[21] D. Lynden-Bell (1971). A method of allowing for known observational selection in small samples applied to 3CR quasars. Monthly Not. R. Astronomical Soc. 155, (1), 95-118.

[22] Fermanian, J.D., D. Radulovic, and M. Wegkamp. (2004). Weak convergence of empirical copula processes. Bernoulli 10 (5):847-60.

[23] Foldes, A., and Rejto, L. (1981a). A LIL type result for the product limit estimator. Probability Theory and Related Fields, 56(1): 75-86.

[24] Foldes, A. et Rejto, L. (1981b). Strong uniform consistency for nonparametric survival curve estimators from randomly censored data. The Annals of Statistics, 9:122, 129.

[25] Foldes , A., Rejto, L. and Winter, B. (1981). Strong consistency properties of nonparametric estimators for randomly censored data, ii : estimation of density and failure rate. Period. Math. Hungar.

[26] Fontaine, C. (2017). Utilisation de copules paramétriques en présence de données observationnelles : cadre théorique et modélisations. Médecine humaine et pathologie. Université de Montpellier.

[27] F. Proschan, P. Sullo (1976). Estimating the Parameters of a Multivariate Exponential Distribution, Journal of the American Statistical Association, 7(354), pp. 465-472.

[28] Frees, E.W., and Valdez, E.A. (1998). Understanding relationships using copulas, North American Actuarial Journal, 2, 1-25.

[29] Fréchet, M. (1957). Les tableaux de corrélation dont les marges et des bornes sont données. Annales de l'Université de Lyon. Sciences Mathématiques et Astronomie, 20, 13-31.

[30] GenestC. (1987). Frank's family of bivariate distributions. Biometrika, 74, 549-555125.

[31] Genest, C., Ghoudi, K., Rivest, L.P. (1995). A semiparametric estimation procedure of dependence parameters in multivariate families of distributions. Biometrika 82, 543-552.

[32] Genest, C., MacKay, R.J. (1986). The joy of copulas: Bivariate distributions with uniform marginals. Amer. Statist. 40, 280-283.

[33] Genest, C., Rivest, L.P. (1993). Statistical inference procedures for bivariate Archimedean copulas. J. Amer. Statist. Assoc. 88, 1034-1043.

[34] Genest, C., Werker, B.J.M. (2002). Conditions for the asymptotic semiparametric efficiency of an omnibus estimator of dependence parameters in copula models. In: Distributions with Given Marginals and Statistical Modelling. Kluwer, Dordrecht, The Netherlands, 103-112.

[35] Georges, P., Lamy, A.G., Nicolas, E., Quibel, G., and Roncalli, T. (2001). Multivariate survival modelling : a uni...ed approach with copulas. Working Paper, Groupe de Recherche Opérationnelle, Crédit Lyonnais, France.

[36] Ghoudi, K. and Abdous, B. (2005). Non-parametric estimators of multivariate extreme dependence functions. Journal of Nonparametric Statistics, Vol. 17, No. 8, 915–935.

[37] Goethals, K., Janssen, P., and Duchateau, L. (2008). Frailty models and copulas : Similarities and dierences. Journal of Applied Statistics. Vol. 35, No 9, 1071-1079.

[38] Gomes, M. I., and Neves, M. (2011). Estimation of the extreme value index for randomly censored data. Biometrical Lett., 48(1), 122.

[39] Gribkova, S., and O. Lopez. (2015). Non-parametric Copula estimation under bivariate censoring.Scandinavian Journal of Statistics 42 (4): 925- 46.

[40] Gumbel, E.J. (1960). Distributions des valeurs extrême en plusieurs dimensions, Institut de Statistique de l'Université de Paris. 9, 171-173.

[41] Gu, M.G. et Lai, T.L. (1990). Functional laws of the iterated logarithm for the product limit-estimator of a distribution function under random censorship or truncation. The Annals of Probability, 18(1):160-189.

[42] H. Joe (2005). Asymptotic efficiency of th two-stage estimation method for copula-based models, Journal of Multivariate Analysis, 94, pp. 401-419.

[43] Hoeffding, W. (1940). Masstabinvariante Korrelations theorie. Schriften des Matematischen Instituts und des Instituts für Angewandte Mathematik der Universität Berlin5 Heft 3 :179-233

[44] Hoeffding, W. (1941). Masstabinvariante Korrelations masse fur Diskontinuierliche Verteilungen. Archiv fur Mathematische Wirtscafts. 7, 49-70, reprinted as Scale- Invariant correlations for discontinuous distributions in The collected works of Wassily Hoeffding, Springer Verlag, New york (109-132).

[45] Hosking, J.R.M., (1990). L-moments : analysis and estimation of distributions using linear combinations of order statistics. J. Royal Statist. Soc. Sev. B 52, 105-124.

[46] Hougaard, P. (1986a). Survival models for heterogeneous populations derived from stable distributions, Biometrika, 73(2), 387-396.

[47] Hougaard, P. (1986b). A class of multivariate failure time distributions, Biometrika, 73(3), 671-678.

[48] Hougaard, P. (1999). Fundamentals of survival data, Biometrics, 55, 13-22.

[49] Joe, H. (1997). Multivariate Models and Dependence concepts, Chapman & Hall, London.

[50] Kagba, N. (2004). On kernel density estimation for censored data. (Ph.D. thesis), University of California, San Diego.

[51] Kaplan, E. L. et Meier, P. (1958). Nonparametric estimation from incomplete observations. Journal of the American Statistical Association, 53:457{481. 1, 16127

[52] Kendall, M. (1938). A New Measure of Rank Correlation, Biometrika, 30, 81-89.

[53] Kimeldorf, G. and Sampson, A. (1975) One-parameter families of bivariate distributions with fixed marginals. Comm Statist A Theory Methods 4, 293-301.

[54] Lancaster, T. (1979). Econometric methods for the duration of unemployment, Econometrica vol. 47 pp. 939-56, 1979.

[55] Lancaster, T. (1990). The Econometric Analysis of Transition Data, Econometric Society Monographs, 17, Cambridge University Press, Cambridge.

[56] Lehmann, E. L., & Casella, G. (1998). Theory of point estimation. New York: Springer.

[57] Liebscher, E. (2002). Kernel density and hazard rate estimation for censored data under -mixing condition. Ann. Inst. Statist. Math., 34:19-28.

[58] Ling, C.H. (1965). Representation of associative functions. Publ Math. Debrecen. 12, 189-212.

[59] Louna, F. (2011). Modélisation de la dépendance par les copules et applications. Thèse de magister. Université de Tizi Ouzou.

[60] Malevergne Y., Sornette D. (2006). Extreme Financial Risks, From Dependence to Risk Management, Springer-Verlag Berlin Heidelberg.

[61] Marshall, A.W. and I. Olkin (1988). Families of multivariate distributions, Journal of the American Statistical Association, 83, 834-84.

[62] M. Boukeloua (2020). Study of semiparametric copula models via divergences with bivariate censored data. Communications in Statistics Theory and Methods.

[63] McGilchrist, C.A., and Aisbett, C.W. (1991). Regression with frailty in survival analysis. Biometrics 47, 461-466.

[64] Mielniczuk, J. (1986). Some asymptotic properties of kernel estimators of a density function in case of censored data. Int. Stat. Rev., 14(2):766-773.

[65] Morales, D.L., and Quesada, V. (1991). Bayesian survival estimate for incomplete data when the life distribution is proportionally related to the censoring time distribution. Comm. Statist. Theory Methods, 20:831-850.

[66] M. Woodroofe (1985). Estimating a distribution function with truncated data. Ann. Statist. 163–177.

[67] Nelsen, R.B. (2006). An Introduction to Copulas, second ed. Springer, New York.

[68] N.Idiou et al (2021). A semi-parametric estimation of copula models based on moments method under right censoring. Journal of TWMS J. App and Eng. Math. (Accepted and to be published).

[69] N. Idiou, et al (2021). Copulas and frailty models in multivariate survival data. Journal of Biostatistics and Health Sciences. ISTE OpenScience, BHS. Vol. 2, No 1, 13-39.

[70] Oakes D. (1982). A model for association in bivariate survival data. J. Roy. Statist. Soc. Ser. B 44, no. 3, 414-422.

[71] Oakes, D. (1989). Oakes. Bivariate survival models induced by frailties. Journal of the American Statistical Association, 84(406).

[72] O. Lopez and P. Saint-Pierre (2012). Bivariate censored regression relying on a new estimator of the joint distribution function. Journal of129 Statistical Planning and Inference 142 (8):2440-53.

[73] Ophthalmol, Am. J. : The Diabetic Retinopathy Study Research Group. Preliminary report on the ecurrency ect of photocoagulation therapy. 81, 383- 396, 1976.

[74] Parzen, E. (1962). On estimation of a probability density function and mode. Ann. Math. Statist., 33: 1065-1076.

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

[76] Rosenblatt, M. (1956). A central limit theorem and a strong mixing condition. Proc. Nat. Acad. Sci. U. S. A., 42:43-47.

[77] Resnick, S. (2006). Heavy-Tail Phenomena: Probabilistic and Statistical Modeling, Springer.

[78] S.A. Osmetti, P.M. Chiodini (2011). A method of moments to estimate bivariate survival functions: the copula approach. Statictica, anno LXXI, n. 4.

[79] S.A. Osmetti, P.M. Chiodini (2008). Some Problems of the Estimation of MarshallOlkin Copula Parameters, Atti XLIV Riunione Scientifica della Società Italiana di Statistica, Arcavacata diRende.

[80] Schweizer, B. and Sklar, A. (1961). Associative functions and statistical triangle inequalities, Publ. Math. Debrecen., 8, 169-186.

[81] Schweizer, B. and Wolf, E. (1981). On non-parametric measures of dependence for random variables. Ann. Statist., 9, 879-885130.

[82] Shaked, M. and Shanthikumar J.G. (1987). The multivariate hazard construction, Stochastic Processes and their Applications, 24, 241-258.

[83] Shih, J.H. and T.A. Louis (1995). Inferences on the asociation parameter in copula models for bivariate survival data, Biometrics, 51, 1384-1399.

[84] Sklar, A. (1959). Fonctions de répartition à n dimensions et leurs marges, Publ. Inst. Statist. Univ. Paris 8, 229-231.

[85] Stute, W. (1993). Consistent estimation under random censorship when covariables are present. Journal of Multivariate Analysis 45 (1):89-103.

[86] Therneau, T.M. (2015). A Package for Survival Analysis. Version 2.38.

[87] Tsukahara, H. (2005). Semiparametric estimation in copula models. Canad. J.Statist. 33, 357-375.

[88] Turnbull, B. W. (1974). Nonparametric estimation of a survivorship function with doubly censored data. J. Amer. Statist. Assoc., 69(345), 169-173.

[89] van der Vaart, A. W., and J. A. Wellner. (1996). Weak convergence and empirical processes with applications to statistics. New York: Springer.

[90] Viswanathan, B. and Manatunga, A.K. (2001). Diagnostic plots for assessing the frailty distribution in multivariate survival data. Lifetime Data Analysis 7, 143-155.

[91] Vaupel. J.W., Manton, K. G., and Stallard, E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. Demography vol. 16 pp. 439-54.

[92] Wang, W., Wells, M.T. (2000a). Model selection and semiparametric inference for bivariate failure-time data. J. Amer. Statist. Assoc. 449, 62-72.

[93] Wang, W., Wells, M.T. (2000b). Estimation of Kendall's tau under censoring. Statist. Sinica 10, 1199-1218131

[94] Wang, A. (2012). On the non-identifiability property of Archimedean copula models under dependent censoring. Statistics and Probability Letters 621-625.

[95] Wang. A and David.O. (2008). Some properties of the Kendall distribution in bivariate Archimedean copula models under censoring. Statistics & Probability Letters.Volume 78, Issue 16, Pages 2578- 2583.

Software Overview

R software

R is a system for statistical computation and graphics. It provides, among other things, a programming language, high-level graphics, interfaces to other languages, and debugging facilities. This manual details and defines the R language. It is a programming language and environment commonly used in statistical computing, data analytics and scientific research. It is one of the most popular languages used by statisticians, data analysts, researchers and marketers to retrieve, clean, analyze, visualize and present data. Due to its expressive syntax and easy-to-use interface, it has grown in popularity in recent years.

R is open-source and free

R is free to download as it is licensed under the terms of the GNU General Public License. You can look at the source to see what's happening under the hood. There's more, most R packages are available under the same license so you can use them, even in commercial applications without having to call your lawyer.

R is popular – and increasing in popularity

IEEE publishes a list of the most popular programming languages each year. R was ranked 5th in 2016, up from 6th in 2015. It is a big deal for a domain-specific language like R to be more popular than a general-purpose language like C. This not only shows the increasing interest in R as a programming language but also in the fields like Data Science and Machine Learning where R is commonly used.

R runs on all platforms

You can find distributions of R for all popular platforms Windows, Linux and Mac.R code that you write on one platform can easily be ported to another without any issues. Cross-platform interoperability is an important feature to have in today's computing world even Microsoft is making its coveted. NET platform is available on all platforms after realizing the benefits of technology that runs on all systems.

Is R programming an easy language to learn !

This is a difficult question to answer. Many researchers are learning R as their first language to solve their data analysis needs. That's the power of the R programming, it is simple enough to learn as you go. All you need is data and a clear intent to draw a conclusion based on analysis on that data.

In fact, R is built on top of the language S programming that was originally intended as a programming language that would help the student learn to program while playing around with data. It is a dialect of S which was designed in the 1980s and has been in widespread use in the statistical community since. Its principal designer, John M. Chambers, was awarded the 1998 ACM Software Systems Award for S.

The language syntax has a superficial similarity with C, but the semantics are of the FPL (functional programming language) variety with stronger affinities with Lisp and APL. In particular, it allows "computing on the language", which in turn makes it possible to write functions that take expressions as input, something that is often useful for statistical modeling and graphics. R is an implementation of the S programming language combined with lexical scoping semantics, inspired by the Scheme. S was created by John Chambers in 1976 while at Bell Labs. A commercial version of S was offered as S-PLUS starting in 1988. Much of the code written for S-PLUS runs unaltered in R.

In 1991 Ross Ihaka and Robert Gentleman at the University of Auckland, New Zealand, began an alternative implementation of the basic S language, completely independent of S-PLUS, which they began publicizing in 1993. It was named partly after the first names of the first two R authors and partly as a play on the name of S. In 1995, Martin Maechler convinced Ihaka and Gentleman to make R free and open-source software under Version 2 of the GNU General Public License.

However, programmers that come from a Python, PHP, or Java background might find R quirky and confusing at first. The syntax that R uses is a bit different from other common programming languages. While R does have all the capabilities of a programming language, you will not find yourself writing a lot of if conditions or loops while writing code in the R language. There are other programming constructs like vectors, lists, frames, data tables, matrices, etc. that allow you to perform transformations on data in bulk.

Applications of R Programming in Real World

Data Science

Harvard Business Review named data scientists the "sexiest job of the 21st century". Glassdoor named it the "best job of the year" for 2016. With the advent of IoT devices creating terabytes and terabytes of data that can be used to make better decisions, data science is a field that has no other way to go but up. Simply explained, a data scientist is a statistician with an extra asset: computer programming skills. Programming languages like R give data scientists superpowers that allow them to collect data in real time, perform statistical and predictive analysis, create visualizations, and

communicate actionable results to stakeholders. Most courses on data science include R in their curriculum because it is the data scientist's favorite tool.

Statistical computing

R is the most popular programming language among statisticians. It was initially built by statisticians for statisticians. It has a rich package repository with more than 9100 packages with every statistical function you can imagine. R's expressive syntax allows researchers – even those from non-computer science backgrounds to quickly import, clean, and analyze data from various data sources. R also has charting capabilities, which means you can plot your data and create interesting visualizations from any dataset.

Machine Learning

R has found a lot of use in predictive analytics and machine learning. It has various packages for common ML tasks like linear and non-linear regression, decision trees, linear and non-linear classification, and many more. Everyone from machine learning enthusiasts to researchers use R to implement machine learning algorithms in fields like finance, genetics research, retail, marketing, and health care.

SAS software

SAS software, standing for Statistical Analysis System, is a proprietary fourth-generation programming language (L4G) published by SAS Institute since (1976).

Since (2004), SAS has been at version 9, which corresponds to a major evolution in the software because it incorporates a new conceptual brick intended to establish itself in the world of business intelligence software. It is, therefore, necessary to separate SAS Foundation, which represents (L4G) alone, and SAS BI, which integrates specific applications.

The development of SAS began in (1966), with a grant from the NIH to eight (US) universities, to analyze agricultural data. North Carolina State University led this consortium.

In (1972), the (NIH) withdrew from the project, and SAS Institute was founded in (1976) to continue the project. The SAS source code then included 300,000 lines of code on punch cards. The system was completely rewritten in C in the mid (1980s), for version (6) of (SAS1).

Traditional SAS software consists of a set of modules to meet the following needs through programming:

- creation and management of databases.
- analytical processing of databases.
- creation and distribution of summary and listing reports.

Abstract

This thesis combines two interesting branches of statistics: survival analysis and copula theory. The primary objective is to extend the copula theory results via semi-parametric estimation, under censored data. More precisely, we are interested by a copulas semi-parametric estimation, based on the classical moments estimation method, adapted for bivariate censored data. There are various kinds of censoring, we are only look at doubly and singly right-censored data. As theoretical results, general formulas were proved with analytical forms of the obtained estimators. According to early research, many asymptotic results obtained in the framework of non-parametric statistics for right-censored observations are based on the Kaplan Meier estimator, which estimates the survival function. Taking into account the results of Lopez and Saint-Pierre (2012) [72], Gribkova and Lopez (2015) [39], the asymptotic normality of the empirical survival copula was established for the two cases of censoring. The dependence structure between the bivariate survival times was modeled under the assumption that the underlying copula is Archimedean. Accounting for various censoring patterns (singly or doubly censored), a simulation study was performed efficiency and robustness of the new estimator proposed.

Individual random parameters, which are commonly understood as frailty parameters, are another tool frequently employed for modeling multivariate survival data. We implemented this model for two-variable survival data using Archimedean copulas in the final part of the thesis. The frailty variables considered here are latent variables that are not observed, are nevertheless one-dimensional. In the example presented, this variable characterized the effect of the individual on the recurrence time. Then we looked at Clayton-Oakes copulas in particular, and even the model with gamma-type frailty. For each of these two models, the copulas used for the bivariate survival functions are the same. Even so, the marginal survival functions are modeled in different ways. The applications for health-related survival data were next examined.

Keywords: Copula, Archimedean copulas models, Semi-parametric estimation, Moments method, Survival copula, Right censored data, Frailty model.

Résumé

Cette thèse forme une sorte de mariage entre deux branches intéressantes de la statistique: l'analyse de survie et la théorie des Copules. L'objectif principal est d'étendre les résultats de la théorie des Copules sur la base de l'estimation semi-paramétriques dans le cas ou les données sont censurées. Plus précisément, nous nous intéressons à l'estimation semi-paramétrique des copules, en utilisant la méthode classique d'estimation des moments, adaptée pour des données censurées. Il existe plusieurs types de censure, nous nous concentrons uniquement sur les données censurées à droite doublement et simplement. Comme résultats théoriques, nous avons présenté les formules générales de ce nouvel estimateur obtenu avec des formes analytiques. Des travaux existants montre que beaucoup de résultats asymptotiques obtenus dans le cadre de la statistique non paramétrique pour des observations censurées à droite, se basent sur l'estimateur de Kaplan Meier qui estime la fonction de survie. Prise en compte les résultats de Lopez et de Saint-Pierre(2012) [72], Gribkova et Lopez (2015) [39], la normalité asymptotique de la copule de survie empirique a établie pour les deux cas de censure présenté. La structure de dépendance entre les temps de survie bivariées a modélisé en supposant que la copule sous-jacente appartient à une famille des copules Archimidéennes. Prise en compte de divers modèles de censure (simple ou double), une étude de simulation a été réalisée pour chaque cas de censure, éclairer le comportement de la méthode d'estimation, a montré l'efficacité et la robustesse du nouvel estimateur proposé.

Un outil additionnel souvent utilisé pour la modélisation des données de survie multivariée est l'introduction de paramètres aléatoires individuels interprétés souvent comme des paramètres de fragilité. Dans la dernière partie de la thèse, nous avons utilisé ce modèle pour les données de survie à deux variables en considérant des copules Archimédiennes. Les variables de fragilité, considérées ici, sont des variables latentes, non observées, mais unidimensionnelles. Dans l'exemple présenté, cette variable caractérisait l'effet de l'individu sur le temps de récurrence. Nous nous sommes concentrés ensuite aux cas particuliers des copules de Clayton-Oakes et du modèle avec fragilité de type gamma. Pour chacun de ces deux modèles, les copules utilisées pour les fonctions de survie bivariée sont les mêmes. Toutefois les fonctions de survie marginales sont modélisées de façons différentes. Nous nous sommes intéressés ensuite à l'applications pour des données de survie relatives à la santé.

Mots clés : Copule, Modèles de copules archimédiennes, Estimation semi-paramétrique, La méthod des moments, Copule de survie, Données censurées à droite, Modèle de fragilité.

تلخيص

تجمع هذه الأطروحة بين فرعين مهمين من الإحصاء تحليل البقاء ونظرية الكوبولا. الهدف الأساسي هو تمديد نتائج نظريات الكوبولا اعتمادا على طرق التقدير شبه المعلمي، في حالة البيانات الخاضعة للرقابة. بتعبير أدق، نحن مهتمون بتقدير شبه المعلمي، استنادًا إلى طريقة تقدير اللحظات الكلاسيكية حيث البيانات ثنائية المتغير و خاضعة للرقابة. هناك أنواع مختلفة من الرقابة، نحن ننظر فقط إلى البيانات الخاضعة للرقابة على اليمين المزدوجة والفردية. كنتائج نظرية، قمنا بإثبات الصيغ العامة للمقدرات التي تم الحصول عليها وذلك بصيغ تحليلية. وفقًا للابحاث الحديثة ، فإن العديد من النتائج المقاربة التي تم الحصول عليها في إطار الإحصاءات اللامعلمية للملاحظات الخاضعة للرقابة من اليمين تستند إلى مقدر كابلن و ماير، الذي يقدر دالة البقاء على قيد الحياة.

ناخد بعين االاعتبار نتائج لوبيز وسان بيير (2012) ، وغريبكوفا ولوبيز (2015) تم تحديد الحالة الطبيعية المقاربة لكوبولا البقاء على قيد الحياة في حالتي الرقابة المقدمة سابقا. هيكل الارتباط بين فترات البقاء على قيد الحياة ثنائية المتغير تم تصميمها على أساس افتراض أن الكوبولا الأساسية تنتمي الى عائلةأرخميدس. ناخد بعين الاعتبار أنماط رقابة مختلفة (الخاضعة للرقابة الفردية أو المزدوجة)، تم إجراء دراسة محاكاة في كلتا الحالتين وذلك من اجل توضيح سلوك طريقة التقدير، والتي أظهرت كفاءة ومتانة المقدر الجديد المقترح. هناك أداة إضافية تُستخدم غالبًا لنمذجة بيانات البقاء على قيد الحياة متعددة المتغيرات وهي إدخال معلمات عشوائية فردية غالبًا ما يتم تفسيرها على أنها معلمات هشاشة.

في الجزء الأخير من الأطروحة، استخدمنا هذا النموذج لبيانات البقاء على قيد الحياة ذات المتغيرين مع الأخذ في الاعتبار عائلة أرخميدس. متغيرات الهشاشة ، التي نعتبرها هنا، هي متغيرات كامنة، لم يتم ملاحظتها، ولكنها ذات بعد واحد. في المثال المقدم، يميز هذا المتغير تأثير الفرد على وقت التكرار. ثم ركزنا بعد ذلك على الحالة الخاصة من الكوبولا كلايتون اوكس و نموذج الهشاشة من نوع جاما. لكل من هذين النموذجين، فإن الكوبولات المستخدمة لوظائف البقاء على قيد الحياة ذات المتغيرين هي نفسها. ومع ذلك، يتم نمذجة دوال البقاء على قيد الحياة الهامشية بطرق مختلفة. ثم قدمنا تطبيقات بيانات البقاء على قيد الحياة لهذا النمودج المتعلقة بالصحة.

الكلمات المفتاحية : الكوبولا، نماذج كوبولا أرخميدس، التقدير شبه المعلمي، طريقةاللحظات، كوبولا الحياة، البيانات الخاضعة للرقابة على اليمين، نموذج الهشاشة.