Copulas in Statistics

Tchilabalo Abozou Kpanzou (tchilabalo@aims.ac.za)
African Institute for Mathematical Sciences (AIMS)

Supervised by Tertius De Wet
University of Stellenbosch

May 24, 2007

Abstract

Copulas are a useful tool for understanding relationships among multivariate variables, and are important tools for describing the dependence structure between random variables, with different copulas representing different dependencies. Since the study of dependence is critically important in Statistics in order to carry out reliable analyses, understanding and applying results from copulas can be very beneficial.

In this project the current status of copulas is investigated, but prior to that an understanding of copulas and their properties is developed. Different applications to Statistics are considered. In particular, the application in Extreme Value Theory is investigated. This area of application has grown tremendously during the last decade and is an area of rapid development of new ideas. In addition, methods of estimating copulas and simulating multivariate outcomes from copulas are described.

Contents

1. Introduction

The study of the relationship between two or more random variables remains an important problem in Statistical Science. For example, when two lives are subject to failure, such as under a joint life insurance, we are concerned with joint distribution of lifetimes. Another example is that when we simulate the distribution of a scenario in a financial security system, we need to understand the distribution of several random variables interacting together, not individually. Unfortunately, the modelling of many of these problems has often been based on a set of simplified assumptions.

Suppose that we have a problem with two important factors, described by two dependent random variables X and Y. Suppose that we know the marginal distributions of X and Y, and also the linear correlation coefficient

$$\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}},$$

where $\mathrm{Cov}(X,Y)$ is the covariance of X and Y, $\mathrm{Var}(X)$ is the variance of X, and $\mathrm{Var}(Y)$ is the variance of Y. (See [17], pages 22 to 27 for more information about variance, covariance and correlation coefficient.) Do we have enough information to describe the joint behaviour of X and Y? The answer is "No", because marginal distributions and the linear correlation coefficient do not necessarily determine the joint distribution. This can be possible for elliptical distributions, whose density is constant on ellipses, but outside the elliptical world, it is not true.

A convenient way to express joint distributions of two or more random variables is provided by copulas. In fact, a copula separates the joint distribution into two contributions: the marginal distributions of the individual variables and the interdependency of the probabilities. The objective of this essay is to explore the meaning of copulas and their role in statistical applications.

In Chapter 2 we define copulas and we give some examples. Some important properties are given in Chapter 3. In that chapter we also provide ways to generate copulas. It is common practice to build a probabilistic model for a real-world scenario in circumstances where there is only limited data on which to develop the model. In cases of dependent multivariate data, copulas provide a useful tool to assist in the model building process. Based on the available data, a copula can be estimated and that is discussed in Chapter 4, where the maximum likelihood and inference functions for margins (IFM) methods are treated. We also present the asymptotic results associated with these methods. The aim of Chapter 5 is to indicate the importance of copulas in applications. These are given in the fields of Survival of Multiple Lives and Extreme Value Theory. In Chapter 6 we discuss methods of simulation of multivariate outcomes from a copula and we give a numerical example. Copula Theory has grown rapidly and is now popular, especially in Risk Management because of its usefulness. However there are statistical problems that can not be solved using copulas. In the conclusion, we also discuss some failures of copulas.

2. Copula Function

2.1 What are Copulas?

The term Copula comes from the Latin noun which means "a link, tie, bond" (see [21]) referring to joining together. With this meaning, a copula is defined as a function that joins multivariate distribution functions to their one-dimensional marginal distribution functions. It is a multivariate distribution function defined on the unit n-cube $[0,1]^n$, with uniformly distributed marginals. Before we give a formal definition of a copula, let us define the H-volume of an n-box.

Definition 2.1. Let S_1, S_2, ..., S_n be nonempty subsets of $\overline{\mathbb{R}}$, where $\overline{\mathbb{R}}$ is the extended real line $[-\infty, \infty]$, and let H be an n-dimensional real function whose domain, $\mathrm{Dom}\, H$, is given by $\mathrm{Dom}\, H = S_1 \times S_2 \times ... \times S_n$. Let $B = [a, b]$ be an n-box all of whose vertices are in $\mathrm{Dom}\, H$. The H-volume of B is given by

$$V_H(B) = \sum_{c \in B} \mathrm{Sign}(c) H(c),$$

where $\mathrm{Sign}(c)$ is given by

$$\mathrm{Sign}(c) = \begin{cases} 1 & \text{if } c_k = a_k, \text{ for an even number of } k\text{'s} \\ -1 & \text{if } c_k = a_k, \text{ for an odd number of } k\text{'s} \end{cases}$$

(Notice that $a = (a_1, a_2, \ldots, a_n),\ b = (b_1, b_2, \ldots,\ b_n),\ c = (c_1, c_2, \ldots,\ c_n)$,
and $B = [a, b] = [a_1, b_1] \times [a_2, b_2] \times \ldots \times [a_n, b_n]$ is well defined if $a_k < b_k$ for all k).

Now let us give a formal definition of copula.

Definition 2.2. An n-dimensional copula is a function $C : [0,1]^n \to [0,1]$, with the following properties:

1. C is grounded, it means that for every $u = (u_1, u_2, \ldots,\ u_n) \in [0,1]^n$, $C(u) = 0$ if at least one coordinate u_i is zero, $i = 1, 2, \ldots,\ n$,
2. C is n-increasing, it means that for every $u \in [0,1]^n$ and $v \in [0,1]^n$ such that $u \leq v$, the C-volume $V_C([u, v])$ of the box $[u, v]$ is non-negative,
3. $C(1, \ldots, 1, u_i, 1, \ldots, 1) = u_i$, for all $u_i \in [0,1]$, $i = 1, 2, \ldots, n$.

For $n = 2$ this definition is reduced to the one following, which is easy to deal with.

Definition 2.3. A two-dimensional (bivariate) copula is a function $C : [0,1]^2 \to [0,1]$, with the following properties:

1. C is grounded: for all u, $v \in [0,1]$, $C(u, 0) = 0$ and $C(0, v) = 0$

2. C is 2-increasing: for all u_1,u_2,v_1,$v_2 \in [0,1]$ such that $u_1 \leq u_2$ and $v_1 \leq v_2$,

$$C(u_2, v_2) - C(u_2, v_1) - C(u_1, v_2) + C(u_1, v_1) \geq 0$$

3. For all u,$v \in [0,1]$, $C(u,1) = u$ and $C(1,v) = v$.

Remark 2.4. The word copula was first employed in a mathematical or statistical sense by Abe Sklar (1959). (See [21].) The term has recently become popular in financial and insurance applications.

2.2 Examples of Copulas

In this section we give some examples of copulas.

Example 1. Marshal-Olkin family (1967)

If $\alpha, \beta \in [0,1]$, then the function $C_{\alpha,\beta} : [0,1]^2 \to [0,1]$, defined by

$$C_{\alpha,\beta}(u,v) = \min(u^{1-\alpha}v, uv^{1-\beta}),$$

is a bivariate copula function. This two-parameter family of copulas is the Marshal-Olkin family (see [18]).

Example 2. Bivariate Pareto copula

This copula is defined by the following formula (see [9])

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

where α is a parameter ($\alpha \in \mathbb{R} \setminus \{0\}$).

Example 3. Farlie-Gumbel-Morgenstern family (1960)

If $\theta \in [-1,1]$, then the function C_θ defined on $[0,1]^2$ by

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

is a one-parameter bivariate copula. This family is known as the Farlie-Gumbel-Morgenstern family (see [23]).

Example 4. Cuadras-Augé family of copulas.

Let $\theta \in [0,1]$. The function C_θ defined by

$$C_\theta(u,v) = [\min(u,v)]^\theta [uv]^{1-\theta} = \begin{cases} uv^{1-\theta}, & \text{if } u \leq v \\ u^{1-\theta}v, & \text{if } u \geq v \end{cases},$$

is a copula function. This family is known as the Cuadras-Augé family of copulas (see [21], page 15).

Example 5. The Gaussian Copula

This copula is simply derived from a multivariate Gaussian distribution function Φ_Σ with mean zero and correlation matrix Σ by transforming the marginals by the inverse of the standard normal distribution function Φ. It is given by (see [20])

$$C(x_1, x_2, \ldots, x_d) = \Phi_\Sigma(\Phi^{-1}(x_1), \Phi^{-1}(x_2), \ldots, \Phi^{-1}(x_d)).$$

Example 6. The t-Copula

The t-copula is derived in the same way as the Gaussian copula. Given a multivariate centered t-distribution function $t_{\Sigma,\nu}$ with correlation matrix Σ, ν degrees of freedom and with marginal distribution function t_ν, this copula is given by (see [20])

$$C(x_1, x_2, \ldots, x_d) = t_{\Sigma,\nu}(t_\nu^{-1}(x_1), t_\nu^{-1}(x_2), \ldots, t_\nu^{-1}(x_d)).$$

2.3 Bivariate Extreme Value Copulas

This family of copulas is obtained by using bivariate extreme-value distributions. A bivariate extreme value copula has the form

$$C_A(u, v) = \exp\left[\log(uv) A\left\{\frac{\log(u)}{\log(v)}\right\}\right],$$

where the dependence function, A, defined on $[0, 1]$ is convex and such that

$$\max(t, 1-t) \leq A(t) \leq 1, \text{ for all } t \in [0, 1].$$

The most common parametric models of bivariate extreme value copulas are given in Table 2.1.

2.4 Archimedean Copulas

Definition 2.5. A copula is an Archimedean copula if it can be expressed in the form

$$C_\phi(u_1, u_2, \ldots, u_n) = \phi^{-1}\left\{\phi(u_1) + \phi(u_2) + \ldots + \phi(u_n)\right\},$$

where $\phi : [0, 1] \to [0, \infty)$ is a bijection such that $\phi(1) = 0$ and

$$(-1)^i \frac{d^i}{dx^i}\phi^{-1}(x) > 0, \; i \in \mathbb{N} \text{ (see [10])}.$$

ϕ is called the generator of the copula C_ϕ.

One key characteristic of Archimedean copulas is the fact that all the information about n-dimensional dependence structure is contained in a univariate generator ϕ. So the Archimedean representation allows to reduce the study of a multivariate copula to a single univariate function.

Some important families of Archimedean copulas are given in Table 2.2.

Table 2.1: Families of bivariate extreme value copulas

Model	$A_\theta(t)$	$C_{A_\theta}(u,v)$
Gumbel[13]	$\theta t^2-\theta t+1,$ $\theta\in(0,1)$	$uv\exp(-\theta\frac{\log(u)\log(v)}{\log(uv)})$
Gumbel-Hougaard	$[t^{\frac{1}{1-\theta}}+(1-t)^{\frac{1}{1-\theta}}]^{1-\theta},$ $\theta\in(0,1)$	$\exp\left\{-[\lvert\log(u)\rvert^{\frac{1}{1-\theta}}+\lvert\log(v)\rvert^{\frac{1}{1-\theta}}]^{1-\theta}\right\}$
Galambos	$1-[t^{-\theta}+(1-t)^{-\theta}]^{-\frac{1}{\theta}},$ $\theta\in(0,\infty)$	$uv\exp\left\{(\lvert\log(u)\rvert^{-\theta}+\lvert\log(v)\rvert^{-\theta})^{-\frac{1}{\theta}}\right\}$
Generalised Marshall-Olkin[19]	$\max\{1-\theta_1 t,1-\theta_2(1-t)\},$ $(\theta_1,\theta_2)\in(0,1)^2$	$u^{1-\theta_1}v^{1-\theta_2}\min(u^{\theta_1},v^{\theta_2})$

Table 2.2: Families of bivariate Archimedean copulas

Family	Generator $\phi(t)$	Bivariate copula $C_\phi(u,v)$
Independence	$-\log(t)$	uv
Clayton[1], Cook-Johnson[2], Oakes[22]	$\frac{t^{-\alpha}-1}{\alpha},$ $\alpha\in(0,\infty)$	$(u^{-\alpha}+v^{-\alpha}-1)^{-\frac{1}{\alpha}}$
Gumbel[12], Hougaard[14]	$(-\log(t))^\alpha,$ $\alpha\in[1,\infty)$	$\exp\left\{-[(-\log(u))^\alpha+(-\log(v))^\alpha]^{\frac{1}{\alpha}}\right\}$
Frank[8]	$\log(\frac{e^{\alpha t}-1}{e^\alpha-1}),$ $\alpha\in R\setminus\{0\}$	$\frac{1}{\alpha}\log\left\{1+\frac{(e^{\alpha u}-1)(e^{\alpha v}-1)}{e^\alpha-1}\right\}$

3. Some Properties of Copulas

3.1 Sklar's Theorem

The importance of copulas in statistics is described in Sklar's theorem. In this sense, this theorem is considered as the central theorem of copula theory.

Theorem 3.1. *(Sklar). See [21], page 17.*

Let H be an n-dimensional distribution function with marginals $F_1, F_2, \ldots, F_n$. Then there exists an n-copula C such that for all $x_1, x_2, \ldots, x_n \in \overline{\mathbb{R}}$,

$$H(x_1, x_2, \ldots, x_n) = C(F_1(x_1), F_2(x_2), \ldots, F_n(x_n)) \tag{3.1}$$

Conversely, if C is an n-copula and $F_1, F_2, \ldots, F_n$ are distribution functions, then the function H defined by Equation (3.1) is an n-dimensional distribution with marginals $F_1, F_2, \ldots, F_n$. Furthermore, if the marginals are all continuous, then C is unique. Otherwise C is uniquely determined on $\operatorname{Ran} F_1 \times \operatorname{Ran} F_2 \times \ldots \times \operatorname{Ran} F_n$, where $\operatorname{Ran} F_i$ is the range of the function F_i.

For $n = 2$, we have the corresponding theorem in two dimensions.

Theorem 3.2. *(Sklar in two dimensions)*

Let H be a joint distribution function with the marginals F and G. There exists a copula C such that for all x and y in $\overline{\mathbb{R}}$,

$$H(x, y) = C(F(x), G(y)). \tag{3.2}$$

If F and G are continuous, then the copula C is unique; otherwise it is uniquely determined on $\operatorname{Ran} F \times \operatorname{Ran} G$. Conversely, if C is a copula, and F, G are distribution functions, then the function H defined by Equation (3.2) is a distribution function with marginals F and G (see [21], page 18).

With this important theorem we see that the copula function is one of the most useful tools for dealing with multivariate distribution functions with given or known univariate marginals.

We now focus on bivariate copulas.

3.2 Continuity, Differentiability and Invariance

Theorem 3.3. *(Continuity)*

Let C be a bivariate copula. Then for all $u_1, u_2, v_1, v_2 \in [0, 1]$ such that $u_1 < u_2$ and $v_1 \leq v_2$,

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

which means that C is uniformly continuous in its domain (see [21]).

Proof. Let $u_1, u_2, v_1, v_2 \in [0,1]$ such that $u_1 < u_2$ and $v_1 \leq v_2$. Let γ_1 be the track passing through the points (u_1, v_1) and (u_2, v_1), and let γ_2 be a track passing through the points (u_2, v_1) and (u_2, v_2). There exist copulas C_{γ_1} and C_{γ_2} such that

$$C(u_1, v_1) = C_{\gamma_1}(u_1, v_1),\ C(u_2, v_2) = C_{\gamma_2}(u_2, v_2),\ C(u_2, v_1) = C_{\gamma_1}(u_2, v_1) = C_{\gamma_2}(u_2, v_1)$$

Therefore,

$$\begin{aligned} |C(u_2, v_2) - C(u_1, v_1)| &\leq |C(u_2, v_2) - C(u_2, v_1)| + |C(u_2, v_1) - C(u_1, v_1)| \\ &= |C_{\gamma_2}(u_2, v_2) - C_{\gamma_2}(u_2, v_1)| + |C_{\gamma_1}(u_2, v_1) - C_{\gamma_1}(u_1, v_1)| \\ &\leq |v_2 - v_1| + |u_2 - u_1|, \end{aligned}$$

the last inequality following from the fact that copulas satisfy Lipschitz's condition (see Lemma 6.1.9 in Schweizer and Sklar [24]). □

Theorem 3.4. *(Differentiability)*

Let C be a bivariate copula. For any $v \in [0,1]$, the partial derivative $\frac{\partial C}{\partial u}(u, v)$ exists for almost all $u \in [0,1]$, and for such v and u,

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

Similarly, for any $u \in [0,1]$, the partial derivative $\frac{\partial C}{\partial v}(u, v)$ exists for almost all $v \in [0,1]$, and for such u and v,

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

Furthermore, the functions $u \mapsto \frac{\partial C}{\partial v}(u, v)$ and $v \mapsto \frac{\partial C}{\partial u}(u, v)$ are well-defined and non-decreasing almost everywhere on $[0,1]$ (see [21]).

Theorem 3.5. *(Invariance)*

Copulas are invariant under strictly monotone transformations of the random variables.

Proof. Let X_1 and X_2 be continuously distributed random variables with copula C, and let T_1, T_2 be strictly increasing transformation functions. Our aim is to prove that $T_1(X_1)$ and $T_2(X_2)$ have the same copula as X_1 and X_2. Let F_1 and F_2 be distribution functions of X_1 and X_2 respectively, and let T_1^{-1} and T_2^{-1} be the inverse functions of T_1 and T_2 respectively. Let G_1 and G_2 be the distribution functions of $T_1(X_1)$ and $T_2(X_2)$ respectively, and let C_T be the copula for $T_1(X_1)$ and $T_2(X_2)$. We have for $i \in \{1, 2\}$,

$$G_i(x_i) = P[T_i(X_i) \leq x_i] = P[X_i \leq T_i^{-1}(x_i)] = F_i(T_i^{-1}(x_i)).$$

Therefore,

$$\begin{aligned} C_T(G_1(x_1), G_2(x_2)) &= P[T_1(X_1) \leq x_1, T_2(X_2) \leq x_2] \\ &= P[X_1 \leq T_1^{-1}(x_1), X_2 \leq T_2^{-1}(x_2)] \\ &= C(F_1(T_1^{-1}(x_1)), F_2(T_2^{-1}(x_2))) \\ &= C(G_1(x_1), G_2(x_2)). \end{aligned}$$

Hence $C_T = C$ in $[0,1]^2$, which means that copulas are invariant under strictly increasing transformations of random variables. Similarly one can verify that copulas are invariant under strictly decreasing transformations of random variables. □

3.3 Frechet-Hoeffding Bounds

Theorem 3.6. *For every copula C and every $(u,v) \in [0,1]^2$,*

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

$W(u,v) = \max(u+v-1,0)$ and $M(u,v) = \min(u,v)$ are themselves copulas (see [21], Theorem 2.2.3, page 11).

Proof. Let C be a bivariate copula. Let X and Y be random variables with copula C. Let F and G be distribution functions of X and Y respectively, and let H be the joint distribution function. We have

$$P[X \leq x, Y \leq y] \leq P[X \leq x], \text{ and } P[X \leq x, Y \leq y] \leq P[Y \leq y],$$

so $P[X \leq x, Y \leq y] \leq \min(P[X \leq x], P[Y \leq y])$

Moreover, $P[X \leq x, Y \leq y] = P[X \leq x] + P[Y \leq y] + P[X > x, Y > y] - 1.$

Since $P[X > x, Y > y] \geq 0$, we have

$$P[X \leq x] + P[Y \leq y] - 1 \leq P[X \leq x] + P[Y \leq y] + P[X > x, Y > y] - 1,$$

which means that, $P[X \leq x] + P[Y \leq y] - 1 \leq P[X \leq x, Y \leq y]$.

Therefore, $\max(P[X \leq x] + P[Y \leq y] - 1, 0) \leq P[X \leq x, Y \leq y]$.

It follows that, $\max(F(x) + G(y) - 1, 0) \leq H(x,y) \leq \min(F(x), G(y))$, for all x and y,

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

The proof that $W(u,v) = \max(u+v-1,0)$ and $M(u,v) = \min(u,v)$ are copulas can be found in [25]. □

3.4 Copulas and Association

This section contains different ways in which copulas can be used in the study of dependence between random variables.

3.4.1 Kendall's Tau

Kendall's tau measure of a pair (X, Y), distributed according to H, is defined as the difference between the probabilities of concordance and discordance for two independent pairs (X_1, Y_1) and (X_2, Y_2) each with distribution H; that is

$$\tau = P[(X_1 - X_2)(Y_1 - Y_2) > 0] - P[(X_1 - X_2)(Y_1 - Y_2) < 0]. \tag{3.3}$$

3.4.2 Spearman's Rho

Let (X_1, Y_1), (X_2, Y_2) and (X_3, Y_3) be three independent random vectors, copies of a random vector (X, Y), with a common joint distribution function H. The Spearman's rho associated with (X, Y), distributed according to H, is defined by

$$\rho = 3P[(X_1 - X_2)(Y_1 - Y_3) > 0] - P[(X_1 - X_2)(Y_1 - Y_3) < 0]. \tag{3.4}$$

Remark 3.7. If C is the copula associated with (X, Y), distributed according to H, then Kendall's tau and Spearman's rho can be written in the forms (see [23]):

$$\tau = 4 \int_0^1 \int_0^1 C(u, v) dC(u, v) - 1, \tag{3.5}$$

$$\rho = 12 \int_0^1 \int_0^1 (C(u, v) - uv) du dv. \tag{3.6}$$

3.4.3 Schweizer and Wolff's Sigma

If we replace the function, $(u, v) \mapsto C(u, v) - uv$, in Equation (3.6) by its absolute value, then we obtain Schweizer and Wolff's Sigma given by (see [23])

$$\sigma = 12 \int_0^1 \int_0^1 |C(u, v) - uv| du dv. \tag{3.7}$$

3.5 Tail Dependence

Definition 3.8. Let X and Y be random variables with distribution functions F and G respectively. Let $U = F(X)$ and $V = G(Y)$.

The coefficient of upper tail dependence is defined as

$$\lambda_U = \lim_{u \to 1^-} P[V > u | U > u], \tag{3.8}$$

provided this limit exists ($\lambda_U \in [0,1]$).

The coefficient of lower tail dependence is defined as

$$\lambda_L = \lim_{u \to 0^+} P[V \leq u | U \leq u], \tag{3.9}$$

provided this limit exists ($\lambda_L \in [0,1]$).

Interpretation 3.9. The coefficients λ_U and λ_L are interpreted as follow:

1. If $\lambda_U = 0$, then X and Y are independent in the upper tail.
2. If $\lambda_U \in (0,1]$, then X and Y are dependent in the upper tail.
3. If $\lambda_L = 0$, then X and Y are independent in the lower tail.
4. If $\lambda_L \in (0,1]$, then X and Y are dependent in the lower tail.

Proposition 3.10. *Let C be a copula associated with (X,Y).*

If $\lim_{u \to 1^-} (\frac{1-2u+C(u,u)}{1-u})$ *and* $\lim_{u \to 0^+} (\frac{C(u,u)}{u})$ *exist, then λ_U and λ_L are given by*

$$\lambda_U = \lim_{u \to 1^-} (\frac{1-2u+C(u,u)}{1-u}) \text{ and } \lambda_L = \lim_{u \to 0^+} (\frac{C(u,u)}{u}).$$

Remark 3.11. We now find λ_U and λ_L for the Archimedean copulas.

Let C be an Archimedean copula generated by ϕ, i.e, $C(u,v) = \phi^{-1}(\phi(u) + \phi(v))$.

Using [1]l'Hopital's rule and the fact that $(\phi^{-1})'(y) = \frac{1}{\phi'(\phi^{-1}(y))}$, λ_U and λ_L are given by

$$\lambda_U = 2 - 2 \lim_{u \to 1^-} \frac{\phi'(u)}{\phi'(\phi^{-1}(2\phi(u)))},$$

$$\lambda_L = 2 \lim_{u \to 0^+} \frac{\phi'(u)}{\phi'(\phi^{-1}(2\phi(u)))}.$$

[1]L'Hopital's rule: Let c be either a finite number or ∞.

$$\text{If } \lim_{x \to c} f(x) = 0 \text{ and } \lim_{x \to c} g(x) = 0, \text{ then } \lim_{x \to c} \frac{f(x)}{g(x)} = \lim_{x \to c} \frac{f'(x)}{g'(x)}.$$

3.6 Methods of Generating Copulas

In this section we present some methods of constructing bivariate copulas. We particularly focus on two illustrations: the Marshall-Olkin Bivariate Exponential family and the Bivariate Pareto Model. To start, let us define the survival function and the survival copula.

Definition 3.12. For a pair (X, Y) of random variables with joint distribution function H, the joint survival function is defined by

$$\overline{H}(x, y) = P[X > x, Y > y]. \tag{3.10}$$

The marginals of $\overline{H}$ are the functions $\overline{H}(\infty, y)$ and $\overline{H}(x, \infty)$ which are univariate survival functions $\overline{F}$ and $\overline{G}$, where F and G are the distribution functions of X and Y respectively.

Definition 3.13. If C is a copula for X and Y, then the survival copula of X and Y is the function $\widehat{C} : [0, 1]^2 \to [0, 1]$, given by (see [21], page 32)

$$\widehat{C}(u, v) = u + v - 1 + C(1 - u, 1 - v). \tag{3.11}$$

Also if $\overline{C}$ is the joint survival function for two uniform $(0, 1)$ random variables U and V whose joint distribution function is the copula C, then we have (see [21], page 33)

$$\overline{C}(u, v) = 1 - u - v + C(u, v) = \widehat{C}(1 - u, 1 - v). \tag{3.12}$$

3.6.1 The Inversion Method

Let H be a bivariate distribution function with continuous marginals F and G. A copula C can be constructed by using Sklar's Theorem through the relation

$$C(u, v) = H(F^{-1}(u), G^{-1}(v)). \tag{3.13}$$

Using the survival function $\overline{H}$, we can also construct a survival copula by the relation

$$\widehat{C}(u, v) = \overline{H}(\overline{F}^{-1}(u), \overline{G}^{-1}(v)), \tag{3.14}$$

where $\overline{F}$ and $\overline{G}$ are taken as in Definition 3.12.

Let us now use this method to construct the Marshal-Olkin Bivariate Exponential family and the Bivariate Pareto Model.

Example 7. We consider a two-component system such as a two engine aircraft. The components are subject to "Shocks", which are always "fatal" to one or both of the components. For example one of the two aircraft engines may fail, or both of them could be destroyed simultaneously. Let X and Y denote the lifetimes of the components 1 and 2, respectively. The survival function $\overline{H}$ is given by $\overline{H}(x, y) = P[X > x, Y > y]$, the probability that the component 1 survives beyond time x and that the component 2 survives beyond time y. The "Shocks" to the two components are assumed to form three independent Poisson processes with (positive) parameters λ_1, λ_2 and

λ_{12}, depending on whether the shock kills only component 1, only component 2, or both the two components simultaneously. The times Z_1, Z_2 and Z_{12} of occurrence of these three shocks are independent exponential random variables with parameters λ_1, λ_2 and λ_{12}, respectively. So we have $X = \min(Z_1, Z_{12})$, $Y = \min(Z_2, Z_{12})$, and then for all nonnegative numbers x and y,

$$\overline{H}(x,y) = P[Z_1 > x]P[Z_2 > y]P[Z_{12} > \max(x,y)] \tag{3.15}$$

$$= \exp\{-\lambda_1 x - \lambda_2 y - \lambda_{12}\max(x,y)\}. \tag{3.16}$$

The marginal survival functions are $\overline{F}(x) = \exp\{-(\lambda_1+\lambda_{12})x\}$ and $\overline{G}(y) = \exp\{-(\lambda_2+\lambda_{12})y\}$; and then X and Y are exponential random variables with parameters $\lambda_1 + \lambda_{12}$ and $\lambda_2 + \lambda_{12}$, respectively. To construct the survival copula $\widehat{C}$, let us first express $\overline{H}(x,y)$ in terms of $\overline{F}(x)$ and $\overline{G}(y)$. Using the relation $\max(x,y) = x + y - \min(x,y)$, we get

$$\begin{aligned}\overline{H}(x,y) &= \exp\{-(\lambda_1+\lambda_{12})x - (\lambda_2+\lambda_{12})y + \lambda_{12}\min(x,y)\} \\ &= \overline{F}(x)\overline{G}(y)\min\{\exp(\lambda_{12}x), \exp(\lambda_{12}y)\}.\end{aligned}$$

Now we set $\overline{F}(x) = u$, $\overline{G}(y) = v$, $\alpha = \frac{\lambda_{12}}{\lambda_1+\lambda_{12}}$ and $\beta = \frac{\lambda_{12}}{\lambda_2+\lambda_{12}}$. Then the previous relation gives us

$$\widehat{C}(u,v) = uv\min(u^{-\alpha}, v^{-\beta}) = \min(u^{1-\alpha}v, uv^{1-\beta}). \tag{3.17}$$

This leads to a two-parameter family of copulas given by

$$C_{\alpha,\beta}(u,v) = \min(u^{1-\alpha}v, uv^{1-\beta}) \tag{3.18}$$

$$= \begin{cases} u^{1-\alpha}v, & if\ u^\alpha \geq v^\beta \\ uv^{1-\beta}, & if\ u^\alpha \leq v^\beta \end{cases} \tag{3.19}$$

This family is the Marshall-Olkin family of copulas. It is also known as the Generalised Cuadras-Augé family of copulas.

Example 8. Bivariate Pareto Model.

Here we consider a random variable X that, given a risk classification parameter γ, can be modeled as an exponential distribution; that is (see [9])

$$P[X \leq x|\gamma] = 1 - e^{-\gamma x}.$$

If γ has a gamma distribution, then the marginal distribution of X is Pareto. That is, if γ is gamma (α,λ) then

$$F(x) = 1 - (1 + \frac{x}{\lambda})^{-\alpha}.$$

Now suppose, conditional on the risk class γ, that X_1 and X_2 are independent and identically distributed. Assuming that they come from the same risk class γ induces a dependency. The joint distribution is

$$F(x_1,x_2) = P[X_1 \leq x_1, X_2 \leq x_2] \tag{3.20}$$

$$= 1 - (1+\frac{x_1}{\lambda})^{-\alpha} - (1+\frac{x_2}{\lambda})^{-\alpha} + (1+\frac{x_1+x_2}{\lambda})^{-\alpha} \tag{3.21}$$

$$= F_1(x_1) + F_2(x_2) - 1 + [(1-F_1(x_1))^{-\frac{1}{\alpha}} + (1-F_2(x_2))^{-\frac{1}{\alpha}}]^{-\alpha}. \tag{3.22}$$

This yields the copulas function

$$C(u,v) = u + v - 1 + [(1-u)^{-\frac{1}{\alpha}} + (1-v)^{-\frac{1}{\alpha}}]^{-\alpha}. \tag{3.23}$$

3.6.2 A Way to Generate Archimedean Copulas

An Archimedean copula is known once one knows its generator. Therefore, to generate it, we just need to construct its generator. Genest and Riviest (1993) provided a procedure for identifying an Archimedean copula (see [9]). To start, let us assume that we have available a random sample of bivariate observations, $(X_{11}, X_{21}), (X_{12}, X_{22}), \ldots, (X_{1n}, X_{2n})$. Assume that the distribution function has an Archimedean copula C_ϕ. Our aim is to identify the form of ϕ. We consider an intermediate pseudo-observation Z_i (defined in 2.a below), that has distribution function $K(z) = P[Z_i \leq z]$. Genest and Riviest (1993) (see [9]) showed that K is related to an Archimedean copula through the relation $K(z) = z - \frac{\phi(z)}{\phi'(z)}$. To identify ϕ, we use the following algorithm:

Algorithm 3.14. *Generating Archimedean copula.*

1. *Estimate Kendall's correlation coefficient using the usual estimate*

$$\tau_n = \binom{n}{2}^{-1} \sum_{i<j} \text{Sign}[(X_{1i} - X_{1j})(X_{2i} - X_{2j})]. \tag{3.24}$$

2. *Construct a nonparametric estimate of K as follows:*

   a. *define the pseudo-observations*

$$Z_i = \frac{\{\textit{number of } (X_{1j}, X_{2j}) \textit{ such that } X_{1j} < X_{1i} \textit{ and } X_{2j} < X_{2i}\}}{n-1}, \tag{3.25}$$

   b. *construct the estimate K_n of K as $K_n(z) =$ proportion of $Z_i's \leq z$.*

3. *Since K has to satisfy the relation*

$$K(z) = z - \frac{\phi(z)}{\phi'(z)}, \tag{3.26}$$

   we obtain an estimate ϕ_n of ϕ, by solving the equation

$$z - \frac{\phi_n(z)}{\phi_n'(z)} = K_n(z). \tag{3.27}$$

Remark 3.15. Some other methods of constructing copulas are illustrated in [21]. There are geometric methods (see examples in [21], pages 59 to 86) and algebraic methods (see examples in [21], pages 89 to 99).

4. Estimation of Copulas

An estimation approach is proposed for models for a multivariate response with covariates when each of the parameters (either univariate or a dependence parameter) of the model can be associated with a marginal distribution. In this chapter we give three ways to estimate a copula. We also discuss confidence bands and asymptotic theory.

4.1 Methods of Estimating Copulas

To start, let us make the following assumptions and notations. We assume that the copula we have to estimate belongs to a family $\{C_\theta,\ \theta \in \Theta\}$, where Θ is the space of parameters. Consider a copula-based parametric model for the random vector $Y = (Y_1, Y_2, \ldots, Y_d)$, with cumulative distribution function,

$$F(y; \alpha_1, \alpha_2, \ldots, \alpha_d; \theta) = C(F_1(y_1; \alpha_1), F_2(y_2; \alpha_2), \ldots, F_d(y_d; \alpha_d); \theta),$$

where $F_1, F_2, \ldots, F_d$ are univariate cumulative distribution functions with respective parameters $\alpha_1, \alpha_2, \ldots, \alpha_d$. We assume that C has density c (mixed derivatives of order d), and by f_j we denote the marginal probability density of Y_j, for $j \in \{1, 2, \ldots, d\}$. Then Y has the density (see [15]):

$$f(y; \alpha_1, \ldots, \alpha_d; \theta) = c(F_1(y_1; \alpha_1), F_2(y_2; \alpha_2), \ldots, F_d(y_d; \alpha_d); \theta) \prod_{j=1}^{d} f_j(y_j; \alpha_j) \tag{4.1}$$

For a sample of size n with observed random vectors $Y_1, Y_2, \ldots, Y_n$, we consider the d log-likelihood functions for the univariate marginals,

$$L_j(\alpha_j) = \sum_{i=1}^{n} \log f_j(y_{ij}; \alpha_j), j = 1, 2, \ldots, d, \tag{4.2}$$

and the log-likelihood function for the joint distribution,

$$L(\alpha_1, \alpha_2, \ldots, \alpha_d; \theta) = \sum_{i=1}^{n} \log f(y_i; \alpha_1, \ldots, \alpha_d; \theta). \tag{4.3}$$

Once one estimates the parameter θ, one has an estimate of the copula.

4.1.1 The Inference Method for Marginals

The inference function for marginals (IFM) method consists of doing d separate optimisations of the univariate likelihoods, followed by an optimisation of the multivariate likelihood as a function of the dependence parameter vector. It consists of the following two steps:

1. the log-likelihoods $L_1(\alpha_1), L_2(\alpha_2), \ldots, L_d(\alpha_d)$, of the d univariate marginals are separately maximised to get estimates $\widehat{\alpha}_1, \widehat{\alpha}_2, \ldots, \widehat{\alpha}_d$ of $\alpha_1, \alpha_2, \ldots, \alpha_d$, respectively,
2. the function $L(\widehat{\alpha}_1, \widehat{\alpha}_2, \ldots, \widehat{\alpha}_d; \theta)$ is maximised over θ to get an estimate $\widehat{\theta}$ of θ.

That is, under regularity conditions, $(\widehat{\alpha}_1, \widehat{\alpha}_2, \ldots, \widehat{\alpha}_d, \widehat{\theta})$ is the solution of

$$\left(\frac{\partial L_1}{\partial \alpha_1}, \frac{\partial L_2}{\partial \alpha_2}, \ldots, \frac{\partial L_d}{\partial \alpha_d}, \frac{\partial L}{\partial \theta}\right) = \underline{0} \tag{4.4}$$

The IFM method is useful for models with the closure property of parameters associated with or being expressed in lower-dimensional marginals (see [15]).

4.1.2 The Maximum Likelihood Method

This method obtains estimates $\widehat{\alpha}_1, \widehat{\alpha}_2, \ldots, \widehat{\alpha}_d, \widehat{\theta}$, by solving the equation

$$\left(\frac{\partial L}{\partial \alpha_1}, \frac{\partial L}{\partial \alpha_2}, \ldots, \frac{\partial L}{\partial \alpha_d}, \frac{\partial L}{\partial \theta}\right) = \underline{0} \tag{4.5}$$

simultaneously. Contrast this with Equation (4.4). An example of the bivariate case can be found in [9], page 14.

4.1.3 The Empirical Copula Function

Here we give a non parametric method for getting a bivariate copula. Consider a sample $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$, iid copies of a random vector (X, Y). The bivariate empirical distribution function (see [7], page 182) associated with (X, Y) is

$$H_n(x, y) = \frac{1}{n} \sum_{i=1}^{n} I_{\{X_i \leq x, Y_i \leq y\}},$$

with marginals

$$F_n(x) = H_n(x, -\infty) = \frac{1}{n} \sum_{i=1}^{n} I_{\{X_i \leq x\}} \text{ and } G_n(y) = H_n(-\infty, y) = \frac{1}{n} \sum_{i=1}^{n} I_{\{Y_i \leq y\}},$$

where I_A is the indicator function of the set A.

Then (see [25]) the empirical copula function is given by

$$C_n(u, v) = H_n(F_n^{-1}(u), G_n^{-1}(v)) \tag{4.6}$$

$$= \frac{1}{n} \sum_{k=1}^{n} I_{\{X_k \leq F_n^{-1}(u), Y_k \leq G_n^{-1}(v)\}}. \tag{4.7}$$

Nelsen (see [21], page 219) defined this copula as

$$C_n(\frac{i}{n}, \frac{j}{n}) = \frac{\text{number of pairs } (x, y) \text{ in the sample with } x \leq x_{(i)}, y \leq y_{(j)}}{n}, \tag{4.8}$$

where $x_{(i)}$ and $y_{(j)}$, $1 \leq i, j \leq n$, denote the order statistics of the sample.

Note that the empirical copula function based on $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$, is the same as that based on uniform $[0, 1]$ random variables $(U_1, V_1), (U_2, V_2), \ldots, (U_n, V_n)$, where $U_i = F(X_i)$ and $V_i = G(Y_i)$, $i \in \{1, 2, \ldots, n\}$ (see [25]).

4.1.4 Estimating Archimedean Copulas

The following method was proposed by Genest and Rivest [11].

Consider a sample $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$, which are iid copies of (X, Y), and assume that the copula C associated with (X, Y) is Archimedean with parameter α. To construct an estimate of α, Genest and Rivest [11] used the observed value of Kendall's tau. In fact, for Archimedean copulas, Kendall's tau can be conveniently computed via the identity

$$\tau = 1 + 4 \int_0^1 \frac{\phi(t)}{\phi'(t)} dt. \tag{4.9}$$

Let us consider the usual estimate of Kendall's tau given by (see [9])

$$\widehat{\tau} = \binom{n}{2}^{-1} \sum_{i<j} \text{Sign}[(X_i - X_j)(Y_i - Y_j)]. \tag{4.10}$$

Since τ is expressed in terms of ϕ (Equation (4.9)), and ϕ is a function of α, an estimate $\widehat{\alpha}$ of α is obtained by solving the equation

$$\widehat{\tau} = 1 + 4 \int_0^1 \frac{\phi(t)}{\phi'(t)} dt, \tag{4.11}$$

for α.

4.2 Confidence Bands

In the previous section we saw how to estimate copula parameters. However, that gives only a point estimate of the parameter. We also need a so-called confidence interval for the parameter in order to better understand the confidence we can have in our estimate.

Assume that it is possible to define two statistics t_1 and t_2 (functions of the sample only) such that, θ being a parameter of interest, we have

$$P[t_1 \leq \theta \leq t_2] = 1 - \alpha,$$

where α is some fixed probability called the confidence coefficient. Then the interval $[t_1,t_2]$ is called a $1-\alpha$ confidence interval for θ. t_1 and t_2 are respectively called the upper and the lower limits of the confidence interval. The parameter α is also called the confidence level.

Confidence bands can be created from the fitted copulas by simulating samples the same size as the original, and creating envelopes of the simulated descriptive functions. This gives an idea of the reasonableness of the fit. It makes more sense to do this for the best fitting copula once that has been determined.

4.3 Asymptotic Theory

In this section we present asymptotic results associated with the methods of estimating copula parameters. We present the iid case and an approach of dealing with covariates.

4.3.1 Independent and Identically Distributed Case

Here we assume that the regularity conditions for asymptotic maximum likelihood theory hold for the multivariate model as well as for all its marginals.

Let $\eta = (\alpha_1, \alpha_2, \ldots, \alpha_d; \theta)$ be the row vector of parameters and let Ψ be the row vector of inference functions of the same dimension as η. Let $Y, Y_1, Y_2, \ldots, Y_n$, be iid with density $f(.;\eta)$. Suppose that the estimator $\widehat{\eta} = (\widehat{\alpha}_1, \widehat{\alpha}_2, \ldots, \widehat{\alpha}_d; \widehat{\theta})$ is given by

$$\sum_{i=1}^{n} \Psi(Y_i, \widehat{\eta}) = 0,$$

and let $\frac{\partial \Psi^T}{\partial \eta}$ be the matrix with (j,k) components $\frac{\partial \Psi_j(y,\eta)}{\partial \eta_k}$. Joe and Xu [15] showed that the asymptotic covariance matrix of $n^{\frac{1}{2}}(\widehat{\eta} - \eta)^T$, called the Godambe information matrix, is

$$V = D_{\Psi}^{-1} M_{\Psi} (D_{\Psi}^{-1})^T, \tag{4.12}$$

where $D_{\Psi} = E\left[\frac{\partial \Psi^T(Y,\eta)}{\partial \eta}\right]$, and $M_{\Psi} = E\left[\Psi^T(Y,\eta)\Psi(Y,\eta)\right]$.

4.3.2 Inclusion of Covariates

Here we assume that we have independent, non-identically distributed random vectors Y_i, $i = 1, 2, \ldots, n$, with densities $f_i(.;\alpha)$, where $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_d, \theta)$. In order to include covariates we assume that $\alpha_j = a_j(x, \gamma_j)$, $j = 1, 2, \ldots, d$, and $\theta = t(x, \gamma_{d+1})$, where $a_1, a_2, \ldots, a_d, t$ are link functions. Instead of $f(y; \alpha_1, \alpha_2, \ldots, \alpha_d, \theta)$ in the case without covariates, we now

consider the density

$$\begin{aligned} f_{Y|x}(y|x;\gamma) &= f(y; a_1(x,\gamma_1), a_2(x,\gamma_2), \ldots, a_d(x,\gamma_d), t(x,\gamma_{d+1})) \\ &= c(F_1(y_1;\alpha), F_2(y_2;\alpha), \ldots, F_n(y_n;\alpha)) \prod_{i=1}^{n} f_i(y_i;\alpha), \end{aligned} \tag{4.13}$$

where F_i is the marginal distribution function of Y_i, $i = 1, 2, \ldots, n$, and $\alpha = (a_1(x,\gamma_1), a_2(x,\gamma_2), \ldots, a_d(x,\gamma_d), t(x,\gamma_{d+1}))$.

The estimate $\widehat{\gamma} = (\widehat{\gamma}_1, \widehat{\gamma}_2, \ldots, \widehat{\gamma}_d, \widehat{\gamma}_{d+1})$ of $\gamma = (\gamma_1, \gamma_2, \ldots, \gamma_d, \gamma_{d+1})$, is obtained by the maximum likelihood method under the following conditions (see [15]):

1. mixed derivatives of Ψ of first and second order are dominated by integrable functions,
2. products of these derivatives are uniformly integrable,
3. the link functions are twice continuously differentiable with first and second order derivatives bounded away from zero,
4. covariates are uniformly bounded, the sample covariance matrix of the covariates is strictly positive definite,
5. a Lindeberg-Feller type condition holds.

If all these conditions hold then the asymptotic normality result has the form (see [15])

$$n^{-\frac{1}{2}} V_n^{-\frac{1}{2}} (\widehat{\gamma} - \gamma)^T \xrightarrow{d} N(0, I),$$

where $V_n = D_n^{-1} M_n (D_n^{-1})^T$, with

$$D_n = n^{-1} \sum_{i=1}^{n} E\left[\frac{\partial \Psi^T(Y_i,\gamma)}{\partial \gamma}\right] \text{ and } M_n = n^{-1} \sum_{i=1}^{n} E\left[\Psi^T(Y_i,\gamma)\Psi(Y_i,\gamma)\right].$$

Note that this approach allows to extend asymptotic theory to the case of random vectors with covariates.

Remark 4.1. This result can also be extended to random covariates. See [15].

5. Statistical Applications of Copulas

Copulas have applications in many fields, particularly in statistical science. In this chapter we give some examples of these applications.

5.1 Survival of Multiple Lives

In epidemiology and actuarial studies, people often examine the joint mortality of pattern of groups of more than a single individual, for example husband and wife, or parents and children.

Before we give an example let us define the hazard function for a random survival time by (see [9])

$$h(t) = -\frac{\partial \log S(t)}{\partial t} = \frac{f(t)}{S(t)},$$

where F is the distribution function of the lifetime T, and S, defined by

$$S(t) = P[T > t] = 1 - F(t),$$

is the survival function of T.

Using Cox's (1972) proportional hazards model, the hazard function is represented as

$$h(t, z) = e^{\beta z} b(t),$$

where $b(t)$ is the so-called "baseline" hazard function and β is a vector of regression parameters.

With this relation, if we set $\gamma = e^{\beta z}$, then we have (see [9])

$$h(t, z) = \gamma b(t) = -\frac{\partial \log S(t|\gamma)}{\partial t}.$$

Integrating this relation, we obtain

$$\log S(t|\gamma) = -\gamma \int_0^t b(s)ds,$$

and then

$$S(t|\gamma) = \exp(-\gamma \int_0^t b(s)ds).$$

Therefore $S(t|\gamma)$ is given by

$$S(t|\gamma) = B(t)^\gamma,$$

where $B(t) = \exp(-\int_0^t b(s)ds)$, is the survival function corresponding to the baseline hazard. γ is called frailty because the larger it is, the smaller $S(t|\gamma)$ is.

Now let us assume that p lifetimes $T_1, T_2, \ldots, T_p$, are independent, given the frailty γ. The joint multivariate survival function is defined as

$$P[T_1 > t_1, T_2 > t_2, \ldots, T_p > t_p] = E_\gamma \left(P[T_1 > t_1, T_2 > t_2, \ldots, T_p > t_p|\gamma] \right).$$

But

$$\begin{aligned} P[T_1 > t_1, T_2 > t_2, \ldots, T_p > t_p|\gamma] &= P[T_1 > t_1|\gamma] \cdot P[T_2 > t_2|\gamma] \cdot \ldots \cdot P[T_p > t_p|\gamma] \\ &= B_1(t_1)^\gamma \cdot B_2(t_2)^\gamma \cdot \ldots \cdot B_p(t_p)^\gamma \\ &= [B_1(t_1) \cdot B_2(t_2) \cdot \ldots \cdot B_p(t_p)]^\gamma. \end{aligned}$$

So we have

$$\begin{aligned} P[T_1 > t_1, T_2 > t_2, \ldots, T_p > t_p] &= E_\gamma([B_1(t_1) \cdot B_2(t_2) \cdot \ldots \cdot B_p(t_p)]^\gamma) \\ &= E_\gamma\{\exp[\gamma \log(B_1(t_1) \cdot B_2(t_2) \cdot \ldots \cdot B_p(t_p))]\} \end{aligned}$$

For the Laplace transform of γ, defined by $E_\gamma(e^{-S\gamma})$, we find that $E_\gamma(e^{-s\gamma}) = \exp(-s^\alpha)$ if γ has a positive stable distribution of order α. This gives

$$S_i(t_i) = \exp\{-[-\log B_i(t_i)]^\alpha)\},$$

and thus (see [9], page 6)

$$P[T_1 > t_1, T_2 > t_2, \ldots, T_p > t_p] = \exp\{-\gamma[-\log B_1(t_1) - \log B_2(t_2) - \ldots - \log B_p(t_p)]\}$$

$$= \exp\{[(-\log S_1(t_1))^{\frac{1}{\alpha}} + (-\log S_2(t_2))^{\frac{1}{\alpha}} + \ldots + (-\log S_p(t_p))^{\frac{1}{\alpha}}]^\alpha\}.$$

This relation is a copula expression. Once one knows the properties of this copula, the relationship between the lifetimes $T_1, T_2, \ldots, T_p$, can be well understood.

5.2 Copulas in Extreme Value Theory

The aim of Extreme Value Theory is to model extreme risks, that is events which occur with only a small probability, but may have catastrophic consequences. This theory is used in Finance, Insurance, Engineering and several other areas. The recent method of understanding Multivariate Extreme Value Theory is that of copulas. In fact, the classical results in this theory may be written in terms of copulas (see [16]).

Let $(X_{j,1}, X_{j,2})$, $j = 1, 2, \ldots$, be a sequence of random variables in $\mathbb{R} \times \mathbb{R}$, with joint distribution function H and marginals F_{X_i}, $i = 1, 2$. Let (M_1, M_2) be the componentwise maxima, i.e, $M_i = \max\{X_{1,i}, X_{2,i}, \ldots, X_{N,i}\}$, $i = 1, 2$. The bivariate version of Fisher-Tippet theorem (see [4], Theorem 1.5) states that, if there exist sequences of normalising constants $a_{iN} > 0$, and $b_{iN} \in \mathbb{R}$, $i = 1, 2$, such that the joint distribution,

$$P\left[\frac{M_1 - b_{1N}}{a_{1N}} \leq x_1, \frac{M_2 - b_{2N}}{a_{2N}} \leq x_2\right] = H^N(a_{1N}x_1 + b_{1N}, a_{2N}x_2 + b_{2N}),$$

converges in distribution as $N \to \infty$ to a proper distribution $W(x_1, x_2)$ with nondegenerate marginals, then $W(x_1, x_2)$ is a bivariate extreme value distribution (see [21], page 28, for the bivariate extreme value distribution). This means that H belongs to the maximum domain of

attraction of W (MDA(W)) and so $F_{X_i} \in$ MDA(W_i), where W_i, $i = 1, 2$, are extreme value distributions. (See [7], page 128 for information about MDA.) Moreover (see [16]) $W(x_1, x_2)$ must satisfy the max-stability relation: for all $N \geq 1$, there exist $a_{iN} > 0$, and $b_{iN} \in \mathbb{R}$, $i = 1, 2$, such that

$$W^N(x_1, x_2) = W(a_{1N}x_1 + b_{1N}, a_{2N}x_2 + b_{2N}).$$

Note that the normalised sequences do not affect the marginal limiting distributions, which are unique up to affine transformations. The following theorem shows that these sequences have influence on the affine transformations of the marginals, but do not affect the copula.

Theorem 5.1. *Let $(X_{11}, X_{12}), (X_{21}, X_{22}), \ldots, (X_{N1}, X_{N2}), \ldots$, be a sequence of independent and identically distributed version of (X_1, X_2), with joint distribution H. Assume that there are normalising sequences $a_{iN} > 0, a'_{iN} > 0$, $b_{iN}, b'_{iN} \in \mathbb{R}$, $i = 1, 2$, such that*

$$H^N(a_{1N}x_1 + b_{1N}, a_{2N}x_2 + b_{2N}) \to W(x_1, x_2), \text{ as } N \to \infty,$$

and

$$H^N(a'_{1N}x_1 + b'_{1N}, a'_{2N}x_2 + b'_{2N}) \to W'(x_1, x_2), \text{ as } N \to \infty,$$

for two nondegenerate distributions $W(x_1, x_2)$ and $W'(x_1, x_2)$. Then the marginal distributions of $W(x_1, x_2)$ and $W'(x_1, x_2)$ are unique up to an affine transformation, i.e, there are $\alpha_{X_1}, \alpha_{X_2}, \beta_{X_1}, \beta_{X_2}$, such that $W_1(x_1) = W'_1(\alpha_{X_1}x_1 + \beta_{X_1})$ and $W_2(x_2) = W'_2(\alpha_{X_2}x_2 + \beta_{X_2})$.

Further, the dependence structures of $W(x_1, x_2)$ and $W'(x_1, x_2)$ are equal, i.e, the copulas are equal ($C_W = C_{W'}$).

With this result, copulas can be well used to understand extreme events. Let C be the copula associated with H. From the continuity of extreme value distributions it follows that there exists a unique copula C_* such that

$$W(x_1, x_2) = C_*(W_1(x_1), W_2(x_2)).$$

Deheuvels [6] showed that (see [4])

$$C_*(u_1, u_2) = \lim_{N \to \infty} C^N(u_1^{\frac{1}{N}}, u_2^{\frac{1}{N}}), \tag{5.1}$$

and then C_* satisfies the relation

$$C_*(u_1^t, u_2^t) = C_*^t(u_1, u_2), \text{ for all } t > 0. \tag{5.2}$$

Equation (5.2) defines an extreme value copula. Copula modelling of extremes may also be addressed by modelling joint excesses over high thresholds. Important applications are found in the fields of Insurance, Environment, Finance. For example, in Finance, risk management theory includes extreme value techniques. Particularly one non-linear measure of dependence at extreme levels, the tail dependence coefficient (see Section 3.5) has become very popular. Many examples of modelling classical multivariate extreme value theory in terms of copulas can be found in [4].

6. Simulation and Data Example

Simulation is a widely used tool, using the power of the computer, for experimenting with complex stochastic models. It can be used for example to generate a sequence of data from a multivariate distribution, in order to study the properties of functions of such data. In this chapter we describe how to use the copula construction to simulate outcomes from a multivariate distribution, and we give a numerical example for the bivariate case.

6.1 Simulation of Multivariate Outcomes

In our simulation, we consider the situation of given marginals and a given copula. We assume the marginal distributions $F_1, F_2, \ldots, F_n$ are continuous so that the copula C specifies a unique multivariate distribution $F(x_1, x_2, \ldots, x_n) = C(F_1(x_1), F_2(x_2), \ldots, F_n(x_n))$ (see Theorem 3.1). Our aim is to generate a random vector $(U_1, U_2, \ldots, U_n)$, from C, since $(F_1^{-1}(U_1), \ldots, F_n^{-1}(U_n))$ has the distribution F. We present two algorithms for this simulation: the method of recursive simulation using univariate conditional distributions, and the algorithm to generate random samples from distributions with Archimedean copulas.

6.1.1 Method Using Univariate Conditional Distributions

We first introduce the notation

$$C_i(u_1, u_2, \ldots, u_i) = C(u_1, u_2, \ldots, u_i, 1, \ldots, 1),\ i = 2, 3, \ldots, n-1,$$

to present i-dimensional marginal distributions of $C(u_1, u_2, \ldots, u_n)$. The conditional distribution of U_i, given the values of $U_1, U_2, \ldots, U_{i-1}$, can be written in terms of derivatives and densities of the i-dimensional marginals, that is

$$\begin{aligned} C_i(u_i|u_1, u_2, \ldots, u_{i-1}) &= P\left[U_i \leq u_i | U_1 = u_1, U_2 = u_2, \ldots, U_{i-1} = u_{i-1}\right] \\ &= \frac{\frac{\partial^{i-1} C_i(u_1, u_2, \ldots, u_i)}{\partial u_1 \partial u_2 \ldots \partial u_{i-1}}}{\frac{\partial^{i-1} C_{i-1}(u_1, u_2, \ldots, u_{i-1})}{\partial u_1 \partial u_2 \ldots \partial u_{i-1}}}, \end{aligned} \tag{6.1}$$

provided the numerator and the denominator of (6.1) exist. For the case in which we can calculate these conditional distributions, this suggests that we should use the following algorithm (see [5]).

Algorithm 6.1. *Consider that the simulation of a value from $C_i(u_i|u_1, u_2, \ldots, u_{i-1})$ could be done by simulating u from the uniform $U(0,1)$ and then calculate $C_i^{-1}(u_i|u_1, u_2, \ldots, u_{i-1})$, if necessary by numerical root finding.*

1. *Simulate a value u_1 from $U(0,1)$.*
2. *Simulate a value u_2 from $C_2(u_2|u_1)$.*

3. *Simulate a value* u_3 *from* $C_3(u_3|u_1, u_2)$.

$\vdots$

n. *Simulate a value* u_n *from* $C_n(u_n|u_1, u_2, \ldots, u_{n-1})$.

6.1.2 Archimedean Construction

In this section we examine an algorithm to generate random samples from distributions with Archimedean copulas. To start, let us review some important theorems.

Theorem 6.2. *See [5]*
Let C *be an Archimedean copula generated by* ϕ. *Let* $K_C(t)$ *denote the* C*-measure of the set*

$$\left\{(u, v) \in [0, 1]^2 : C(u, v) \leq t\right\},$$

or equivalently, of the set

$$\left\{(u, v) \in [0, 1]^2 : \phi(u) + \phi(v) \geq \phi(t)\right\}.$$

Then for any $t \in [0, 1]$,

$$K_C(t) = t - \frac{\phi(t)}{\phi'(t)}.$$

Proposition 6.3. *Let* U *and* V *be uniform random variables whose joint distribution function is the Archimedean copula* C *generated by* ϕ. *Then the distribution function* K_C, *given by Theorem 6.2, is the distribution function of the random variable* $C(U, V)$.

The following theorem, from Genest and Rivest [11], is an extension of Proposition 6.3.

Theorem 6.4. *Under the hypotheses of Proposition 6.3, the joint distribution function* $H(s, t)$ *of the random variables* $S = \frac{\phi(U)}{\phi(U)+\phi(V)}$ *and* $T = C(U, V)$ *is given by*

$$H(s, t) = sK_C(t), \text{ for all } (s, t) \in [0, 1]^2. \tag{6.2}$$

A result of Theorem 6.4 is the following algorithm (see [5]).

Algorithm 6.5. *This algorithm generates random samples* (u, v) *whose joint distribution function is an Archimedean copula* C *generated by* ϕ.

1. *Generate two independent standard uniform random numbers* q *and* s.
2. *Set* $t = K_C^{-1}(q)$, *where* K_C^{-1} *denotes the quasi-inverse of the distribution function* K_C, *given in Theorem 6.2.*
3. *Set* $u = \phi^{-1}(s\phi(t))$ *and* $v = \phi^{-1}((1 - s)\phi(t))$.
4. *The desired pair is* (u, v).

6.2 Data Example

The data we consider in this section can be found in Cook and Weisberg [3]. We analyse two variables from this data: the flight range factor (RGF) and the sustained load factor (SLF), for 22 aircrafts. The two variables, $X = RGF$ and $Y = SLF$, constitute bivariate observations $(X_1, Y_1), (X_2, Y_2), \ldots, (X_{22}, Y_{22})$, given in Table 6.1.

Table 6.1: Bivariate data constituted by $X = RGF$ and $Y = SLF$

X	3.30	3.64	4.87	4.72	4.11	3.75	3.97	4.65	3.84	4.92
Y	0.10	0.10	2.90	1.10	1.00	0.90	2.40	1.80	2.30	3.20

3.82	4.32	4.53	4.48	5.39	4.99	4.50	5.20	5.65	5.40	4.20	6.45
3.50	2.80	2.50	3.00	3.00	2.64	2.70	2.90	2.90	3.20	2.90	2.00

We first calculate the linear correlation (or Pearson's correlation) coefficient, Kendall's tau and Spearman's rho.

6.2.1 Correlations Between the Variables

Table 6.2 contains three correlation coefficients between the variables X and Y. The linear correlation coefficient shows that there is weak positive association between X and Y. With Kendall's tau we can say that there is a weak degree of concordance.

Table 6.2: Correlation relations between the variables $X = RGF$ and $Y = SLF$

Pearson correlation	Kendall's tau	Spearman's rho
0.4655	0.3532	0.4618

The dependence structure between X and Y is shown in Figure 6.1, where we present the scatterplot of Y versus X and that of X versus Y.

6.2.2 Estimation of the Copula Associated with the Data

We assume that the copula C_α associated with the data is Archimedean, generated by ϕ, and depending on a parameter α. Our aim is to estimate α. We use the method described in Section 4.1.4. We consider Kendall's tau obtained in Section 6.2.1. Taking C_α in the Clayton family, the generator is given by $\phi(t) = \frac{t^{-\alpha}-1}{\alpha}$, and so $\phi'(t) = -t^{-\alpha-1}$. Therefore, $\frac{\phi(t)}{\phi'(t)} = -\frac{1}{\alpha}(t - t^{\alpha+1})$, and then using this last relation we find

$$1 + 4\int_0^1 \frac{\phi(t)}{\phi'(t)}dt = \frac{\alpha}{\alpha+2}.$$

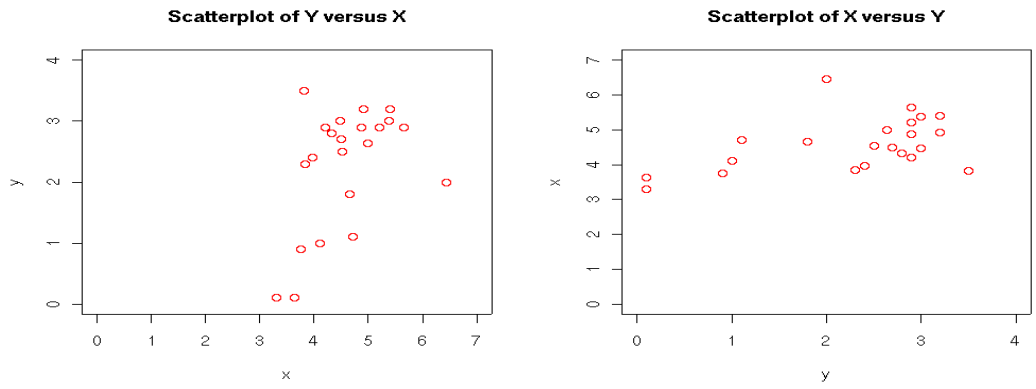

Figure 6.1: Dependence structure of (X, Y).

With the Gumbel-Hougaard family we have $\phi(t) = (-\log(t))^{\alpha}$ and $\phi'(t) = -\frac{\alpha(-\log(t))^{\alpha-1}}{t}$.

Therefore,

$$1 + 4\int_0^1 \frac{\phi(t)}{\phi'(t)}dt = \frac{\alpha - 1}{\alpha}.$$

Solving equation (4.11) for α, with the above relations, we obtain estimates of the parameter α given in Table 6.3.

Table 6.3: Estimates of the parameter α for the Clayton and Gumbel-Hougaard families

Family	Estimate $\widehat{\alpha}$ of α
Clayton	1.0924
Gumbel-Hougaard	1.5462

Therefore, an estimate of the generator ϕ is given by

$$\phi(t) = \frac{t^{-1.0924} - 1}{1.0924}, \tag{6.3}$$

for the Clayton family, and

$$\phi(t) = (-\log(t))^{1.5462}, \tag{6.4}$$

for the Gumbel-Hougaard family.

Finally an estimate of the copula associated with the data is given as follows:

$$\widehat{C}(u, v) = (u^{-1.0924} + v^{-1.0924} - 1)^{-\frac{1}{1.0924}}, \tag{6.5}$$

if C_α belongs to the Clayton family and

$$\widehat{C}_\alpha(u, v) = \exp\left\{-[(-\log(u))^{1.5462} + (-\log(v))^{1.5462}]^{\frac{1}{1.5462}}\right\}, \tag{6.6}$$

if C_α belongs to the Gumbel-Hougaard family.

The plot of the Clayton copula obtained is shown in Figure 6.2.

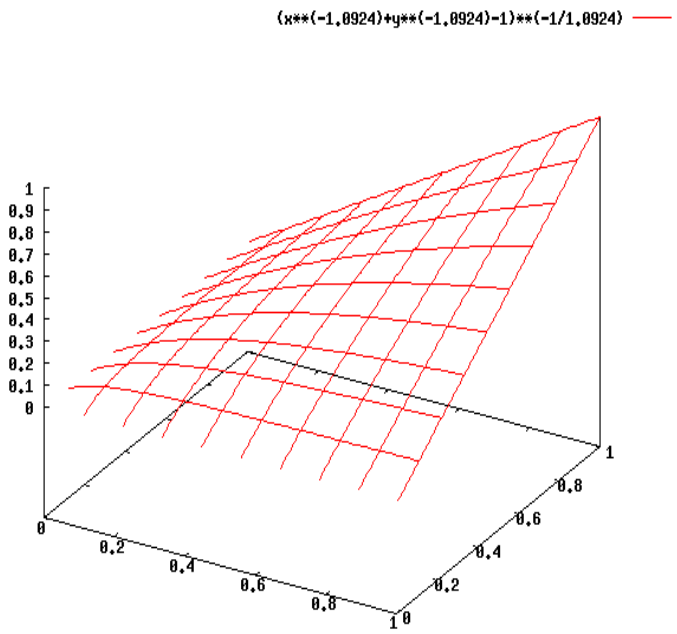

Figure 6.2: Plot of the estimated Clayton copula associated with the data

6.2.3 Simulation

In this section we check whether the method we used to estimate α in section 6.2.2 works with artificial data. To do this, we generate a sequence of data from a bivariate distribution, using Algorithm 6.5. The copula we consider is from the Clayton family, and generated by

$$\phi(t) = \frac{t^{-\alpha} - 1}{\alpha}.$$

We first find the distribution function K_C from Theorem 6.2. Using $\phi'(t) = -t^{-\alpha-1}$, we obtain

$$K_C(t) = (1 + \frac{1}{\alpha})t - \frac{1}{\alpha}t^{\alpha+1}. \tag{6.7}$$

Using $\alpha = 1.0924$ in Algorithm 6.5, we generate the data given in Table 6.4.

Table 6.4: Simulated data from a Clayton copula with $\alpha = 1.0924$

U	0.85	0.04	0.23	0.78	0.93	0.13	0.38	0.30	0.37	0.97
V	0.33	0.05	0.19	0.81	0.84	0.28	0.20	0.70	0.41	0.89

0.40	0.44	0.16	0.52	0.82	0.14	0.27	0.41	0.93	0.23	0.23	0.39
0.29	0.28	0.84	0.14	0.17	0.07	0.41	0.63	0.81	0.10	0.46	0.62

We apply the same method as in Section 6.2.2 to obtain an estimate, $\overline{\alpha}$, of the copula parameter associated with this data. We find $\overline{\alpha} = 1.1095$, and then we estimate the copula:

$$\overline{C}(u, v) = (u^{-1.1095} + v^{-1.1095} - 1)^{-\frac{1}{1.1095}}. \tag{6.8}$$

We simulate 10000 samples the same size as the original one, and we find an estimate $\overline{\alpha}$ for each sample. We then calculate the average $\overline{\overline{\alpha}}$ and the standard error $\overline{\sigma}$. We find $\overline{\overline{\alpha}} = 1.2351$ and $\overline{\sigma} = 0.0108$.

Note that the value of $\overline{\overline{\alpha}}$ is not too close to α, for example an approximate 95% confidence interval for α based on the simulation is $[1.2243, 1.2459]$. This does not include the true α value of 1.0924. Further work on the estimation method needs to be done.

7. Conclusion

One of the most widely used tools to study multivariate outcomes is the copula function. In the case of dependent multivariate data, multivariate copulas provide a usefull tool to assist in the process of model building. In this essay we have shown how to analyse some real-world scenarios using copulas.

We started with a discussion of the significance of the copula function in the second chapter. A copula is a function that relates a multivariate distribution function to its one-dimensional marginal distribution functions. We reviewed its properties, such as the invariance under strictly monotone transformations in Chapter 3. In that chapter, we also looked at some methods of constructing bivariate copulas, for purposes of simplicity.

Since copulas are parametric families, standard techniques such as the maximum likelihood and inference functions for margins (IFM) methods, are useful for estimating their parameters. These methods are described in Chapter 4.

In Chapter 5 we discussed the statistical applications of copulas by giving two illustrations: the Survival of Multiple Lives and the Extreme Value Theory. We considered the case of Survival of Multiple Lives because in epidemiology and actuarial studies, one often examines the joint mortality of linked individuals (for example a family). In the fields of Insurance, Environment and Finance, it is common to model extreme events, and so the Extreme Value Theory is applied.

We showed in Chapter 6 how copulas could be used to simulate multivariate outcomes, an important tool for applied work, where many variables need to be considered. We also provided an example where we showed how one can estimate a copula where bivariate data is at hand.

In the last two decades the theory and applications of copulas, as a field of study, has been developing rapidly. Copulas offer a flexible structure that can be applied in many situations. However, there are statistical problems in handling copulas. One of these problems is the curse of dimensionality. In fact, if one has deeper insight into underlying dependence structure of the data one can try to reduce the dimension of a parametric model by assuming functional dependencies between the parameters, or one might want to use dimension reduction techniques. As suggested by Mikosch [20], the problem of reducing dimensions for distributions related to multivariate extreme value theory is an important problem since it would be inappropriate to deal with the extremes in all components of a multivariate sample if some of the components dominated others. Another problem faced in dealing with copulas is the description of complex space-time dependence structure. In this case copulas are not useful for modelling dependence through time.

Current work in copula theory should help to solve such problems, and develop the theory in order to be applied in each area of study.

In further research, we will deepen the Extreme Value Theory, and we will further investigate estimation and Goodness-of-Fit tests for copulas.

Acknowledgements

I am very grateful to Prof Tertius De Wet for giving me the opportunity to work with him on writing an essay in such a fascinating area.

I wish to express my sincere appreciation to all the lecturers and tutors for their assistance, and to all my friends at AIMS for the time we have spent together.

Finally I would like to thank Prof Fritz Hahne and Prof Neil Turok for giving me the opportunity to study at the African Institute for Mathematical Sciences.

Bibliography

[1] Clayton D.G. A Model for Association in Bivariate Life Tables and its Application in Epidemiological Studies of Familial Tendency in Chronic Disease Incidence. *Biometrika*, 65:141–151, 1978.

[2] Cook R.D and Johnson M.E. A Family of Distributions for Modeling Non-Elliptically Symmetric Multivariate Data. *Journal of the Royal Statistical Society*, B43:210–218, 1981.

[3] Cook R.D. and Weisberg S. *Residuals and Influence in Regression*. Chapman & Hall, New York, 1982.

[4] Damarta S. Extreme Value Theory and Copulas. Master's thesis, Department of Mathematics, ETH Zurich, Switzerland, 2002.

[5] de Matteis R. Fitting Copulas to Data. Master's thesis, Institute of Mathematics, The University of Zurich, 2001.

[6] Deheuvels P. Probabilistic Aspects of Multivariate Extremes. In *Statistical Extremes and Applications*, pages 117–130. Reidel Publishing Compan, 1984.

[7] Embrechts P, Klüppelberg C, and Mikosch T. *Modeling Extremal Events*. Springer, 1997.

[8] Frank M.J. On the Simultaneous Associativity of F(x,y) and x+y-F(x,y). *Acquationes Mathematicae*, 19:194–226, 1979.

[9] Frees E. W. and Valdez E. A. Understanding Relationships Using Copulas. *North American Actuarial Journal*, 1998.

[10] Genest C., Quessy J.F., and Rémillard B. Goodness-of-fit Procedures for Copula Models Based on the Probability Integral Transformation. *Scandinavian Journal of Statistics*, 2005.

[11] Genest C. and Rivest L. Statistical Inference Procedures for Archimedean Copulas. *Journal of the American Statistical Association*, 88:1034–1043, 1993.

[12] Gumbel E.J. Bivariate Exponential Distributions. *Journal of American Statistical Association*, 55:698–707, 1960.

[13] Gumbel E.J. Distribution des Valeurs Extremes en Plusieurs Dimensions. *Publ. Inst. Statist. Univ. Paris 9*, pages 171–173, 1960.

[14] Hougaard P. A class of Multivariate Failure Time Distributions. *Biometrika*, 73:671–678, 1986.

[15] Joe H. and Xu J.J. The Estimation Method of Inference Functions for Margins for Multivariate Models. Technical report, Department of Statistics, University of British Columbia, 1996.

[16] Kolev N., dos Anjos U., and Beatriz Vaz de M. Mendes. Copulas: A Review and Recent Developments. *Stochastic Models, Taylor & Francis Group*, 2006.

[17] Lee P.M. *Bayesian Statistics: An Introduction*. Arnold Publication, 1997.

[18] Marshall A.W and Olkin. A Generalized Bivariate Exponential Distribution. *Applied Probability*, 1967.

[19] Marshall A.W. and Olkin. A Multivariate Exponential Distribution. *Journal of the American Statistical Association*, 62:30–44, 1988.

[20] Mikosch T. Copulas: Tales and facts. *Springer Science + Business Media, LLC*, 2006.

[21] Nelsen R.B. *An Introduction to Copulas*. Springer, second edition, 2006.

[22] Oakes D. A Model for Association in Bivariate Survival Data. *Journal of the Royal Statistical Society*, B44:414–422, 1982.

[23] Quasada-Molina J.J. What are Copulas? *Monografias Del Semi. Matem. Garcia de Galdeono. 27*, 2003.

[24] Schweizer B. and Sklar A. *Probabilistic Metric Spaces*. North-Holland, Amsterdam, 1983.

[25] Veraverbeke N. *Multivariate Data and Copulas*. Hasselt University, Belgium, 31 October 2005. Lectures notes.