Revista Colombiana de Estadística
January 2019, Volume 42, Issue 1, pp. 61 to 80
DOI: http://dx.doi.org/10.15446/rce.v42n1.71780

A Method to Select Bivariate Copula Functions

Un método para seleccionar funciones cópula bivariadas

José Rafael Tovar Cuevas[1,a], Jennyfer Portilla Yela[1,2,b], Jorge Alberto Achcar[3,c]

[1]Escuela de Estadística, Facultad de Ingeniería, Universidad del Valle, Cali, Colombia

[2]Departamento de ciencias naturales y matemáticas, Pontificia Universidad Javeriana, Cali, Colombia

[3]Faculdade de Medicina, Universidade de São Paulo, Ribeirão Preto SP, Brazil

Abstract

Copula functions have been extensively used in applied statistics, becoming a good alternative for modeling the dependence of multivariate data. Each copula function has a different dependence structure. An important issue in these applications is the choice of an appropriate copula function model for each case where standard classical or Bayesian discrimination methods could be not appropriate to decide by the best copula. Considering only the special case of bivariate data, we propose a procedure obtained from a recently dependence measure introduced in the literature to select an appropriate copula for the statistical data analyses.

***Key words*:** Copula functions; Discrimination of copulas; Dependence measure; Ledwina measure; Selection method.

Resumen

Las funciones de la cópula se han utilizado ampliamente en las estadísticas aplicadas, convirtiéndose en una buena alternativa para modelar la dependencia de los datos multivariados. Cada función de la cópula tiene una estructura de dependencia diferente. Un tema importante en estas aplicaciones es la elección de un modelo de función de cópula apropiado para cada caso en el que los métodos de discriminación clásicos o bayesianos estándar no sean apropiados para decidir por la mejor cópula. Considerando solo el caso especial de datos bivariados, proponemos un procedimiento obtenido a partir de una medida de dependencia recientemente introducida en la literatura para seleccionar una cópula apropiada para los análisis de datos estadísticos.

***Palabras clave*:** Discriminación de cópulas; Funciones de cópula; Medida de dependencia; Medida de Ledwina; Método de selección.

[a]PhD. E-mail: jose.r.tovar@correounivalle.edu.co
[b]PhD. E-mail: jennyfer.portilla@javerianacali.edu.co
[c]PhD. E-mail: achcar@fmrp.usp.br

1. Introduction

In many different areas of knowledge such as medicine, engineering, economy and ecology, it is possible to have a set of observations obtained from variables whose natural behavior has some dependence structure. To model this dependence, there are many statistical techniques, models and indexes introduced in the literature as for instance frailty models, correlation coefficients, concordance coefficients, etc. (See Goethals, Janssen & Duchateau 2008).

Since the introduction of Sklar's theorem (Sklar 1959), many parametric, nonparametric and semiparametric models were proposed derived from different copula functions assuming different probability distributions, including methods for constructing models for copulas from different probability distributions (e.g., see, Durante & Sempi 2015, Nelsen 2006), most of which are parametric models. Copula functions have been used extensively in different applications including sea storm data (Corbella & Stretch 2013); analysis of the dependence structure between the stocks in different foreign exchange markets (Wang, Wu & Lai 2013); operational risk management (Arbenz 2013); risk evaluation of droughts (Zhang, Xiao, Singh & Chen 2013); risk assessment of hydroclimatic variability (Janga Reddy & Ganguli 2012); modeling wind speed dependence (Xie, Li & Li 2012); the dependence between crude oil spot and futures markets (Chang 2012); and stochastic modeling of power demand (Lojowska, Kurowicka, Papaefthymiou, van der Sluis et al. 2012).

Copula function are used to link marginal distributions with a joint distribution. For specified univariate marginal distribution functions $F_1(t_1)$, $F_2(t_2)$, $\ldots, F_m(t_m)$, the function $C(F_1(t_1), F_2(t_2), \ldots, F_m(t_m)) = F(t_1, t_2, \ldots, t_m)$, which is defined using a copula function C, results in a multivariate distribution. On the other hand, any multivariate distribution function F can be written in the form of a copula function; i.e, if $F(t_1, t_2, \ldots, t_m)$ is a joint multivariate distribution function with univariate marginal distribution functions $F_1(t_1)$, $F_2(t_2), \ldots, F_m(t_m)$, there is a copula function $C(u_1, u_2, \ldots, u_m)$, so that:

$$F(t_1, t_2, \ldots, t_m) = C(F_1(t_1), F_2(t_2), \ldots, F_m(t_m)) \tag{1}$$

If every F_i is continuous, then C is unique and $u_i = F_i$. For the special case of bivariate distributions, $m = 2$.

The approach to formulating a multivariate distribution using a copula is based on the concept that a simple transformation can be made of each marginal variable so that each transformed marginal variable has a uniform distribution. Then the dependence structure can be expressed as a multivariate distribution on these obtained uniforms, and a copula is a multivariate distribution with marginally uniform random variables. Consequently there are many families of copulas that differ in the details of their dependence structure. In the bivariate case, let T_1 and T_2 be two random variables with continuous distribution functions F_1 and F_2. The probability integral transformation can be applied separately to both random variables to define $U_1 = F_1(t_1)$ and $U_2 = F_2(t_2)$, where U_1 and U_2 have uniform $(0, 1)$ distributions, but are usually dependent if T_1 and T_2 are dependent (Independent T_1 and T_2 imply that U_1 and U_2 are independent). Specifying dependence

between T_1 and T_2 is the same as specifying dependence between uniform random variables U_1 and U_2. In this case, there is a need to specify a bivariate distribution between two uniform distributions, i.e, a copula.

Within each application field, it is needed in general, a suitable measure of the strength of dependence of the two random variables in each application. Dependence scalar measure indexes or global measures of dependence for two random variables have been studied by many authors as Jogdeo (1982), Lancaster (1982), Drouet & Kotz (2001), Balakrishnan & Lai (2009) and in many cases the use of scalar dependence bivariate indexes could be not the best way to represent complex dependence structure, so other dependence structures have been introduced in the literature, see Kowalczyk, Pleszczynska et al. (1977), Bjerve & Doksum (1993), Drouet & Kotz (2001) and Bairamov, Kotz & Kozubowski (2003). Considering the case of two dependent random variables, generally it is very difficult to establish non linear dependence structures using indexes making necessary the use of copula functions but, it is very common to have difficulties to decide on the best copula function to be fitted by the data since the literature has not yet presented many discrimination methods for copula models. In a recent paper, Ledwina (2015) proposed a new function valued measure of dependence for two random variables T_1 and T_2 also presenting its basic properties.This proposed measure has a simple form which explores only cumulative distribution functions taking values in the $[-1, 1]$ interval treating both variables symmetrically. The correlation order or the equivalent concordance order is the quadrant order restricted to the class of distributions where fixed margins are preserved. The Ledwina dependence measure that assumes two random variables T_1 and T_2 is expressed as:

$$q(t_1, t_2) = \frac{F(t_1, t_2) - F_1(t_1)F_2(t_2)}{w(t_1, t_2)} \quad for \ \ (t_1, t_2) \in D \tag{2}$$

where

$$w(t_1, t_2) = \sqrt{F_1(t_1)F_2(t_2)(1 - F_1(t_1))(1 - F_2(t_2))}$$

and $D = \{(t_1, t_2) : 0 < F_i(t_i) < 1\}$ $F_i(t_i) = P(T_i \leq t_i)$ are the marginal distribution functions of T_i, $i = 1, 2$ and $F(t_1, t_2) = P(T_1 \leq t_1, T_2 \leq t_2)$ is the joint distribution function for T_1 and T_2.

A simple empirical estimator for the Ledwina dependence giving in (2) is proposed considering the bivariate data set of n pairs $(t_{1i}, \ t_{2i})$, $i = 1, \ldots, n$ replacing $F(t_1, t_2)$, $F_1(t_1)$ and $F_2(t_2)$ with their respective empirical estimates,

$$\begin{aligned} F_n(t_1, t_2) &= \frac{Number \ of \ obs \ T_1 \leq t_1, \ Number \ of \ obs \ T_2 \leq t_2}{n} \\ F_n(t_1) &= \frac{Number \ of \ obs \ T_1 \leq t_1}{n} \\ F_n(t_2) &= \frac{Number \ of \ obs \ T_2 \leq t_2}{n} \end{aligned} \tag{3}$$

for fixed values t_1 and t_2.

Given that q treats both variables T_1 and T_2 symmetrically, knowledge of q and the marginal distributions allows one to recover the joint distribution function for T_1, T_2. Properties of q and further details can be found en (Ledwina 2015).

Using equation (2) it is possible to obtain a copula-based measure of dependence assuming joint distribution functions $F(t_1, t_2)$ with continuous margins $F_1(t_1)$ and $F_2(t_2)$, as follows:

$$q(u_1, u_2) = \frac{C(u_1, u_2) - u_1 u_2}{w(u_1, u_2)}, \qquad (u_1, u_2) \in [0, 1]^2 \tag{4}$$

where,

$$\sqrt{w(u_1, u_2) = [u_1 u_2 (1 - u_1)(1 - u_2)]}$$

$$u_1 = F_1(t_1) \;\; u_2 = F_2(t_2) \;\; C(u_1, u_2) = F(t_1, t_2)$$

In this paper, it is proposed an index obtained by a modification of the Ledwina measure and with the evaluation of its performance considering five copula functions. This paper is organized as follows: in Section 2, it is introduced some special copula functions; in Section 3, it is proposed a method for discrimination of different copula functions developed with the Ledwina dependence measure; in Section 4, it is presented the results obtained from a simulation study and in Section 5, it is presented some concluding remarks.

2. Some Special Copula Functions

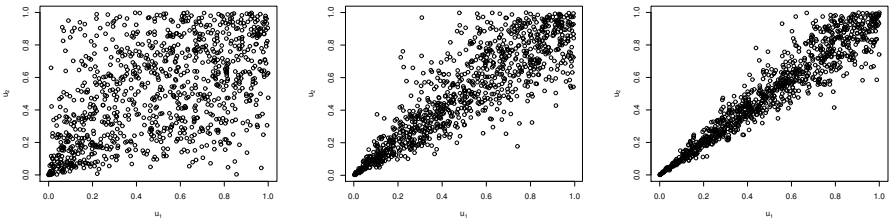
In this section, it is introduced some copula functions that are explored in the present study. In all cases, $u_1 = F_1(t_1) = P(T1 \leq t_1)$, $u_2 = F_2(t_2) = P(T_2 \leq t_2)$; $C(u_1, u_2) = F(t_1, t_2) = P(T_1 \leq t_1, T_2 \leq t_2)$ and θ is the dependence parameter.

2.1. Clayton Copula

The Clayton (1978) copula function models asymmetrical data structures with high dependence in the left tail indicating an expanding cloud. The Clayton copula is know as the Pareto bivariate copula since it is possible to obtain it from the survival function of the the bivariate Pareto distribution (Hutchinson and Lad 1990). Additionally this copula function is considered as a special case of the Lomax copula function. The Clayton copula function has the following analytical structure:

$$C(u_1, u_2) = \left(u_1^{-\theta} + u_2^{-\theta} - 1 \right)^{-\frac{1}{\theta}} \tag{5}$$

for $\theta \in (-1, \infty) \backslash \{0\}$. When $\theta \to \infty$ the dependence is perfect and positive and if $\theta \to 0$ the variables are independent.



(a) $\theta = 1$ (b) $\theta = 5$ (c) $\theta = 10$

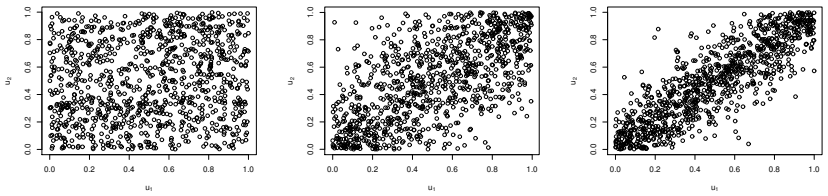
FIGURE 1: Plots of data under different Clayton dependence structures.

2.2. Frank Copula

The Frank copula function (Frank 1979) is appopriate to model weak dependence structures with positive linear trend. This copula function has the following analytical structure:

$$C(u_1, u_2) = -\frac{1}{\theta} \log \left(1 + \frac{\left(e^{-\theta u_1} - 1\right)\left(e^{-\theta u_2} - 1\right)}{e^{-\theta} - 1} \right) \tag{6}$$

for $\theta \in (-\infty, \infty) \setminus \{0\}$. The maximum dependence value is reached when $\theta \to \infty$ and the minimum when $\theta \to -\infty$. When $\theta \to 0$ it is possible to assume independence between the variables.



(a) $\theta = 1$ (b) $\theta = 5$ (c) $\theta = 10$

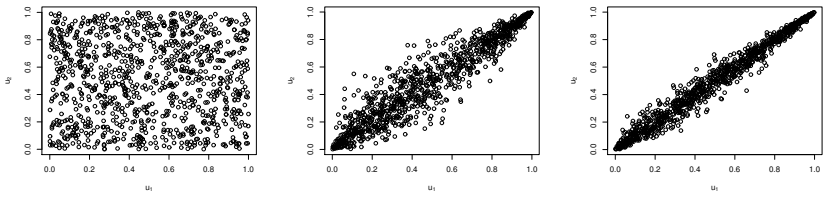
FIGURE 2: Plots of data under different Frank dependence structures.

2.3. Gumbel-Hougaard copula

The Gumbel-Hougaard copula function introduced by for details see Gumbel (1960a), Gumbel (1961) and (Hougaard 1986) is useful to model data structures with strong dependence in upper tail and weak dependence in lower tail where it is expected that the upper data show a strong correlation and the lower data shows weakly correlation. The analytical form of the Gumbel-Hougaard copula function is defined as:

$$C(u_1, u_2) = e^{-\left[(-\log(u_1))^{\theta}+(-\log(u_2))^{\theta}\right]^{\frac{1}{\theta}}} \tag{7}$$

with $\theta \geq 1$. The perfect dependence is obtained when $\theta \to 0$ and if $\theta = 1$ the is possible to assume independence between the variables.



(a) $\theta = 1$ (b) $\theta = 5$ (c) $\theta = 10$

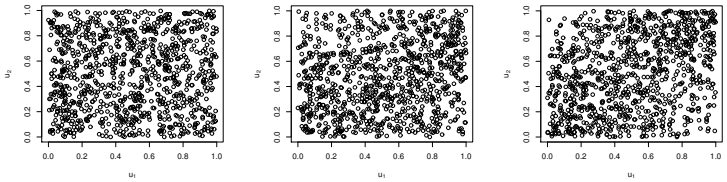
FIGURE 3: Plots of data under different Gumbel-Hougaard dependence structures.

2.4. Farlie-Gumbel-Morgenstern Copula (FGM Copula)

The first reference on the FGM copula functions family is the Eyraud in 1938 (For details see Nelsen 2006). This copula can be considered when the data set to be analyzed shows a weak and non linear dependence structure (Meintanis 2007). In this case, the scater plots obtained from the data are very similar with the plots obtained from data sets of independent variables. The FGM copula (Nelsen 2006) is defined by,

$$C(u_1, u_2) = u_1 u_2[1 + \theta(1 - u_1)(1 - u_2)] \tag{8}$$

for $-1 \leq \theta \leq 1$. Therefore, it is possible to assume independence between the variables when $\theta = 0$. This copula function models weak linear dependence structures



(a) $\theta = 0.2$ (b) $\theta = 0.5$ (c) $\theta = 0.9$

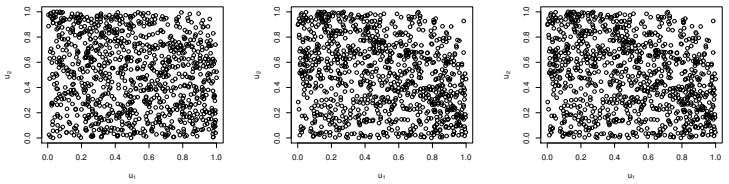
FIGURE 4: Plots of data under different FGM dependence structures.

2.5. Gumbel-Barnett Copula (GB Copula)

The GB copula introduced by Gumbel (1960a) and Barnett (1980), has an analytical structure defined by,

$$C(u_1, u_2) = u_1 + u_2 - 1 + (1 - u_1)(1 - u_2)exp[-\theta ln(1 - u_1)ln(1 - u_2)] \quad (9)$$

for $0 \leq \theta \leq 1$. Independence corresponds to $\theta = 0$. Two random variables whose dependence structure can be fitted with a bivariate Gumbel distribution do not present linear correlations. This copula function can be obtained from tjoint bivariate Gumbel distribution with standard exponential marginal distributions.



(a) $\theta = 0.2$ (b) $\theta = 0.5$ (c) $\theta = 0.9$

FIGURE 5: Plots of data under different GB dependence structures.

3. Algorithms to Simulate Data with Dependence Structure Type Copula

The algorithms used to simulate data from the copula function structures presented in section 2 were obtained from different sources. The FGM data were simulated using an algorithm published in the Johnson & Kotz (1972) book, algorithms to simulate data with Clayton, Frank and Gumbel-Hougaard were obtained from studies published by Hofert (2008). To simulate data from the Gumbel-Barnett copula function, an algorithm was developed by the authors based on the results presented by Gumbel (1960b).

3.1. Data with Copula Clayton Dependence

The data with dependence structure type Clayton copula were simulated using the algorithm as follows:

- Set a value for the parameter of dependence θ
- Generate $u_1 \sim U(0, 1)$ and $w \sim U(0, 1)$
- Replace in $u_2 = \left[\left(w^{-\frac{\theta}{\theta+1}} - 1\right) u_1^{-\theta} - 1\right]^{-\frac{1}{\theta}}$ the u_1 and w values
- The u_1 and u_2 values have dependence with type Clayton copula function

3.2. Data with Dependence Type Copula Frank

The data with dependence structure type Frank copula were simulated using the algorithm as follows:

- Set a value for the parameter of dependence θ
- Generate $u_1 \sim U(0,1)$ and $w \sim U(0,1)$
- Replace in $u_2^* = -\frac{1}{\theta}\log\left(-\frac{w(e^{-\theta}-1)}{e^{-\theta u_1}(w-1)-w}+1\right)$ the u_1 and w values
- The u_1 and u_2 values have dependence with type Frank copula function

3.3. Data with Dependence Type Copula Gumbel-Hougaard

The data with dependence structure type Gumbel-Hougaard copula were simulated using the algorithm as follows:

- Set a value for the parameter of dependence θ
- Generate an observation x from a positive stable distribution

$$X \sim st\left(\frac{1}{\theta}, 1, \left(\cos\left(\frac{\pi}{2\theta}\right)\right)^{\theta}, \mathbf{1}_{\{\theta=\mathbf{1}\}}, 1\right)$$

- Generate $v_1 \sim U(0,1)$ and $v_2 \sim U(0,1)$
- Let $u_i = \exp\left(-\left(-\frac{\log\ v_i}{x}\right)^{\frac{1}{\theta}}\right); \quad i \in \{1,2\}$ where the pair (v_1, v_2) have a dependence structure Gumbel-Hougaard
- Repeat m times the previous steps to obtain a vector of pairs of data with Gumbel-Hougaard dependence structure

3.4. Data with Dependence Type Copula Fgm

The data with dependence structure type FGM copula were simulated using the algorithm as follows:

- Set a value for the parameter of dependence θ
- Generate $v_1 \sim U(0,1)$ and $v_2 \sim U(0,1)$
- Let $u_1 = v_1$
- Compute

$$A = \theta(2u_1 - 1) - 1$$

and

$$B = \left(1 - 2\theta\,(2u_1 - 1) + \theta^2(2u_1 - 1)^2 + 4\theta v_2\,(2u_1 - 1)\right)^{\frac{1}{2}}$$

- Let $u_2 = \dfrac{2v_2}{B - A}$
- Repeat m times the previous steps to obtain a vector of pairs of data with FGM dependence structure

3.5. Data with Dependence Type Copula Gumbel-Barnett

In this case, the algorithm used was:

- Set a value for the parameter of dependence θ
- Generate $u_2 \sim U(0,1)$ and $w \sim U(0,1)$
- Replace u_2 in $y = -\log(1 - u_2)$
- Obtain a value of x as solution of the non-linear equation $1-(1+\theta x)e^{-(1+\theta y)x} - w = 0$
- Let $u_1 = 1 - exp(-x)$ and $u_2 = 1 - exp(-y)$ one pair of values with Gumbel-Barnett dependence structure
- Repeat m times the previous steps to obtain a vector of pairs of data with Gumbel-Barnett dependence structure

4. Relationship Between the Kendall's Tau and the Copula Parameter of Dependence

In general, for some copula function families it is possible to have $\tilde{\theta}_n = g(\tau)$ being g a differentiable function. The relationship between the concordance measure Kendall's tau τ and the copula function parameter can be very important to estimate the dependence parameter using the moments method. According to Nelsen (2006) it is possible to obtain an expression for the kendall's tau from the copula function as follows;

$$\tau = 4 \int_0^1 \int_0^1 C_{\mathbf{Y}}(u_1, u_2) c_{\mathbf{Y}}(u_1, u_2)\, du_1 du_2 - 1 \tag{10}$$

Solving (10) for each copula function considered, it is obtained in Table 1 the results of interest.

Where:

$$D_1(\theta) = \frac{1}{\theta} \int_0^{\theta} \frac{t}{e^t - 1} dt$$

TABLE 1: Kendall's tau and copula function parameters

Copula function	τ	$\hat{\tau}$	$\hat{\theta}_\tau$
Clayton	$\frac{\theta}{\theta+2}$	$[0,1]\backslash\{0\}$	$\frac{2\hat{\tau}}{1-\hat{\tau}}$
Frank	$1+\frac{4}{\theta}\left[D_1(\theta)-1\right]$	$[-1,1]\backslash\{0\}$	It has not close form
Gumbel-Hougaard	$1-\frac{1}{\theta}$	$[0,1]$	$\frac{1}{1-\hat{\tau}}$

5. Estimation of the Copula Dependence Parameter

Given that the copula function is a expression of a multivariate probability distribution , under a statistical point of view, it is possible to associate a likelihood function given the data and to use standard maximum likelihood or Bayesian methods to get estimates for the dependence parameter.

5.1. Maximum Likelihood Estimation

To obtain the maximum likelihood estimates (MLE), it is needed to have density function,

$$c_{\mathbf{Y}}(u_1,u_2)=\frac{\partial C_{\mathbf{Y}}(u_1,u_2)}{\partial u_1 \partial u_2} \tag{11}$$

In this way, the MLE are obtained maximizing the log likelihood function

$$\ell_n(\theta)=\sum_{i=1}^{n}\log\left(c_{\mathbf{Y}}(\hat{F}_{Y_1}(y_{1i}),\hat{F}_{Y_2}(y_{2i}))\right) \tag{12}$$

For each considered copula function introduced in section 2, the log likelihood function was derived as follows:

- **Clayton copula function**

$$\ell_n(\theta)=n\log\left(\theta+1\right)-(2+\theta^{-1})\sum_{i=1}^{n}\left(u_{1i}^{-\theta}+u_{2i}^{-\theta}-1\right)-(\theta+1)\sum_{i=1}^{n}\log\left(u_{1i}u_{2i}\right)$$

- **Gumbel-Hougaard copula function**

$$\begin{aligned}\ell_n(\theta)=&(\theta-1)\left[\sum_{i=1}^{n}\log\left(-\log\left(u_{1i}\right)\right)+\sum_{i=1}^{n}\log\left(-\log\left(u_{2i}\right)\right)\right]\\&-\sum_{i=1}^{n}\left(\log\left(u_{1i}\right)\log\left(u_{2i}\right)-\left(\left(-\log\left(u_{1i}\right)\right)^{\theta}+\left(-\log\left(u_{2i}\right)\right)^{\theta}\right)^{\theta-1}\right)\\&+\sum_{i=1}^{n}\log\left((\theta-1)+\left(\left(-\log\left(u_{1i}\right)\right)^{\theta}+\left(-\log\left(u_{2i}\right)\right)^{\theta}\right)^{\theta^{-1}}\right)\\&+\left(\theta^{-1}-2\right)\sum_{i=1}^{n}\log\left(\left(-\log\left(u_{1i}\right)\right)^{\theta}+\left(-\log\left(u_{2i}\right)\right)^{\theta}\right)\end{aligned}$$

- **Frank copula function**

$$\ell_n(\theta) = n\left(\log(\theta) + \log\left(1 - e^{-\theta}\right)\right) - \theta \sum_{i=1}^{n}(u_{1i} + u_{2i}) \\ -2\sum_{i=1}^{n}\log\left(\left(e^{-\theta} - 1\right) + \left(e^{-\theta u_{1i}} - 1\right)\left(e^{-\theta u_{2i}} - 1\right)\right)$$

- **FGM copula function**

$$\ell_n(\theta) = \sum_{i=1}^{n}\log\left(1 + \theta\left(1 - 2u_{1i}\right)\left(1 - 2u_{2i}\right)\right)$$

- **Gumbel-Barnett copula function**

$$\ell_n(\theta) = -\theta\sum_{i=1}^{n}\left(\log\left(1 - u_{1i}\right)\log\left(1 - u_{2i}\right)\right) \\ + \sum_{i=1}^{n}\log\left(\left(\theta\log\left(1 - u_{1i}\right) - 1\right)\left(\theta\log\left(1 - u_{2i}\right) - 1\right) - \theta\right)$$

5.2. Bayesian Estimation

To estimate the dependence parameters using Bayesian methods, it was assumed non informative prior distributions considering the range of values in the parametric space. In general, the posterior distribution for the dependence parameter has the form: $\pi(\theta \mid \mathbf{u}_1, \mathbf{u}_2) \propto \pi(\theta)L(\theta \mid \mathbf{u}_1, \mathbf{u}_2)$ where $\pi(\theta)$ is the prior distribution and $L(\theta \mid \mathbf{u}_1, \mathbf{u}_2)$ is the likelihood function. For all cases, a $Uniform(a, b)$ distribution was assumed as a non informative distribution for the dependence parameter. To obtain values for the hyperparameters (a,b), it was assumed the existence of an expert in the subject of study, whose knowledge can be expressed through Kendall's tau. When the expert opinion for the dependence level was weak it was assumed $\tau \in (0, 0.33)$ if in according with the expert, the expected dependence was moderate then $\tau \in (0.33, 0.66)$ and for a strong dependence it was assumed $\tau \in (0.66, 1)$. Using the relationship between the Kendall's tau and dependence parameter presented in (10), it was obtained the intervals showed in Table 2.

TABLE 2: Prior intervals for dependence parameter and non informative prior distributions.

Copula function	Weak	Moderate	Strong	Non informative prior
Clayton	[0,0.98)	[0.98,3.88)	[3.88,10)	$Uniform(0, 10)$
Frank	[0,3.26)	[3.26,9.78)	[9.78,24)	$Uniform(0, 24)$
Gumbel Hougaard	[1,1.49)	[1.49,2.94)	[2.94,10)	$Uniform(1, 10)$

To estimate the dependence parameter of the FGM and Gumbel Barnett copula functions it was assumed a Uniform(0,1) prior distribution. For the FGM copula function, it was considered a positive range of values in the parameter space a

commonly approach assumed for this copula function. The obtained posterior distributions are:

- **Clayton copula**

$$\pi\left(\theta|\underset{\sim}{u_1},\underset{\sim}{u_2}\right) \propto \frac{(\theta+1)^n}{10}\prod_{i=1}^{n}\left(u_{1i}^{-\theta}+u_{2i}^{-\theta}-1\right)u_{1i}^{-(\theta+1)}u_{2i}^{-(\theta+1)}$$

- **Frank copula**

$$\pi\left(\theta|\underset{\sim}{u_1},\underset{\sim}{u_2}\right) \propto \frac{\theta^n\left(1-e^{-\theta}\right)^n e^{-\theta\left(\sum_{i=1}^{n}(u_{1i}+u_{2i})\right)}}{10\prod_{i=1}^{n}\left(\left(e^{-\theta}-1\right)+\left(e^{-\theta u_{1i}}-1\right)\left(e^{-\theta u_{2i}}-1\right)\right)^2}$$

- **Gumbel-Hougaard copula**

$$\pi\left(\theta|\underset{\sim}{u_1},\underset{\sim}{u_2}\right) \propto \frac{k_1}{10}\prod_{i=1}^{n}\frac{\left(-\log\left(u_{1i}\right)\right)^{\theta}\left(-\log\left(u_{2i}\right)\right)^{\theta}}{u_{1i}u_{2i}}$$

where:

$$k_1 = \left((\theta-1)+\left(\left(-\log\left(u_{1i}\right)\right)^{\theta}+\left(-\log\left(u_{2i}\right)\right)^{\theta}\right)^{\theta^{-1}}\right)\left(\left(-\log\left(u_{1i}\right)\right)^{\theta}+\left(-\log\left(u_{2i}\right)\right)^{\theta}\right)^{\left(\theta^{-1}-2\right)}$$

- **FGM copula**

$$\pi\left(\theta|\underset{\sim}{u_1},\underset{\sim}{u_2}\right) \propto \frac{1}{10}\prod_{i=1}^{n}\left(1+\theta\left(1-2u_{1i}\right)\left(1-2u_{2i}\right)\right)$$

- **Gumbel-Barnett copula**

$$\pi\left(\theta|\underset{\sim}{u_1},\underset{\sim}{u_2}\right) \propto k_2\prod_{i=1}^{n}\left(\left(\theta\log\left(1-u_{1i}\right)-1\right)\left(\theta\log\left(1-u_{2i}\right)-1\right)-\theta\right)$$

where:

$$k_2 = \frac{1}{10}e^{-\theta\sum_{i=1}^{n}\left(\log(1-u_{1i})\log(1-u_{2i})\right)}$$

6. The Goodness-Of-Fit (GOF) Method

This method to evaluate the goodness of fit of data to copula functions, was developed by Deheuvels (1981) and Genest & Rémillard (2004). The idea is to compare the empirical copula C_n defined as:

$$C_n(u_1, u_2) = \frac{1}{n} \sum_{i=1}^{n} I\left(U_{1i} \leqslant u_1, U_{2i} \leqslant u_2\right) \tag{13}$$

with the parametric copula function $C_(u, v)$ using a statistical hypothesis test. In this way, the authors define the Cramer-Von Mises Statistic as:

$$S_n = \sum_{i=1}^{n} \left(C_n(\hat{u}_{1i}, \hat{u}_{2i}) - C(\hat{u}_{1i}, \hat{u}_{2i})\right)^2 \tag{14}$$

and the asymptotic distribution of the test statistics S_n was derived from the process $C_n(u_1, u_2)$ depending on the unknown distribution $C(\hat{u_1}, \hat{u_2})$.

7. An Approach to Decide by an Appropriate Copula Function Using the Ledwina Dependence Measure

To discriminate the best copula function among k different functions that could be candidates to be fitted by the sample data (n pairs of observations), it is proposed from a modification of (2) the following discrimination index given by,

$$I(model[j]) = \sqrt{\sum_{i=1}^{n} (q(u_{1i}, u_{2i}) - q^*(u_{1i}, u_{2i}))^2} \qquad j = 1, 2, ...k \tag{15}$$

where;

$$q(u_{1i}, u_{2i}) = \{[C(u_{1i}, u_{2i}) - (u_{1i}u_{2i})]w(u_{1i}, u_{2i})\} \tag{16}$$

$$q^*(u_{1i}, u_{2i}) = \{[C_n(u_{1i}, u_{2i}) - (u_{1i}u_{2i})]w(u_{1i}, u_{2i})\} \tag{17}$$

$$w(u_{1i}, u_{2i}) = \frac{1}{\sqrt{(u_{1i}u_{2i}(1 - u_{1i})(1 - u_{2i})}} \qquad i = 1, \ldots, n$$

Given that to compute q it is necessary to estimate $C(u_1, u_2)$, the dependence parameter θ must be estimated using some statistical method reported in the literature (for instance, maximum likelihood). To obtain q^* it is needed to compute $F_n(t_1, t_2)$, $F_n(t_1)$ and $F_n(t_2)$ as in equation (3). When the k indexes are computed, the model with minimum value for (15) is choosed as the best model to be fitted by the data.

8. A simulation Study

Using the algorithms in section 3, we conducted a simulation study to evaluate the performance of our proposed index. We carried out our procedure 1000 times considering three different sample sizes ($n = 50$, 100, 500, 1000) to simulate vectors of pairs of observations $(t_{1i}, t_{2i} \quad i = 1, 2, \ldots, n)$ for each of the five copula functions introduced in section 2. With each data set, it was estimated the dependence parameter using maximum likelihood and Bayesian methods. From these estimates, it is estimated the cumulative probability $C(u_1, u_2)$ associated to each pair of observations. This procedure was used considering the simulated data from each assumed copula function. With the values of $C(u_{1i}, u_{2i})$ it is computed $q(u_{1i}, u_{2i})$ applying equation (3). Each data set was used to compute the empirical copula and to obtain $q^*(u_{1i}, u_{2i})$. Finally, it was computed the proposed index given by equation (13) considering each copula function.

The obtained results were compared with those obtained using the Goodness Of Fit (GOF) method to select copula functions, where the null hypothesis is the model fitted by the data set (See details in Kojadinovic, Yan & Holmes 2011). The GOF procedure could be problematic when there is rejection of the null hypothesis which indicates a particular copula function but there is no indication of the best copula to be fitted by the data set. Observe that this hypothesis test must be applied for each proposed copula function. For each simulated data set it was fitted all copula functions introduced in section 2 and the procedure was carried out 1000 times. Following, it was counted the number of times that each procedure identified the true copula model.

8.1. Selection of Values for the Copula Dependence Parameter

(Weiss 2011) studied different dependence levels measured with Kendall's tau. In accordance with those authors, we decided to use three dependence levels within the range of values that the dependence parameter can take for each copula function: Weak dependence ($\tau = 0.2$), moderate dependence ($\tau = 0.5$) and strong dependence ($\tau = 0.8$). For each established tau value, it was obtained the associated value of the copula parameter in each of five considered models. For the dependence parameters of the Gumbel-Barnett and FGM copula functions it was used the same values assumed by Tovar & Achcar (2012), Tovar & Achcar (2013). See Table 3.

Table 3: Values assumed for the copula dependence parameters in the simulation study.

	Weak	**Moderate**	**Strong**
Clayton	0.50	2.00	8.00
Frank	1.86	5.73	18.19
Gumbel Hougaard	1.25	2.00	5 .00
FGM	0.20	0.50	0.90
Gumbel Barnett	0.20	0.50	0.90

9. Results of the Simulation Study

Table 5 shows the results obtained when using the proposed method to select the best copula function among the five assumed copulas. The percentage of times that the method selected the copula, changes in sample size, the level of dependence and the method that was used to estimate the dependence parameter are reported. Table 3 also presents the results obtained using the GOF method for the classifications under the same study scenarios. When the structure of the data dependence was modeled using in a Clayton copula function, the proposed method had no problems in selecting the correct copula a good percentage of the time (regardless of the level of dependence or sample size) although with small amounts of data ($n = 50$) a minimum amount of misclassified samples were observed. When comparing the classification rates with those obtained using the GOF method, it can be said that, contrary to what was observed for the proposed method, the GOF procedure requires large amounts of data (500 or more) to obtain high percentages of correct classifications for all levels of dependence.

If the data have a Frank copula dependence structure, our procedure identifies the copula less than 50% of the time when the dependence is weak and the sample size is below 100. With larger sample sizes, the percentage of correct classification increases, coming close to the unit when the sample size is 1000. For moderate and strong dependences, the percentages of a good classification improve noticeably; and when the sample size is 1000, no classification errors were observed. For this copula function, the inferential method used to estimate the dependence parameter of the copula might have a negligible effect on the percentage of correct responses. For this copula, the GOF method can identify the real copula function only when there are large amounts of data in the sample, regardless of the strength of the dependence between the variables. For the data with a weak Gumbel-Hougaard type of dependence, the proposed method obtained a good percentage of classification when the samples were over 500 and the parameter was been estimated using the maximum likelihood method. For those cases where we fitted the copula function using a Bayesian estimate, the results were quite poor.

For a moderate GH dependence, the rating capacity of the proposed method is good with sample sizes over 100, regardless of the procedure used in the estimation of the dependence parameter. If the GH dependence is strong but there are only 100 or fewer data, it is necessary to get maximum likelihood estimates in order to obtain good ranking results. For dependence structures with this copula function, the GOF method had difficulties in identifying the correct copula. Its capacity for identification is good only when there are amounts of data higher than 500 and the GH dependence is moderate or strong. For weak dependence structures using a Gumbel-Barnett copula function,the proposed method of classification has better performance with maximum likelihood estimators and sample sizes of at least 100 observations although in this case it is possible to get rates up to 20% of poor classifications. For moderate-level dependence with a sample size of at least 500 observations, it is possible to obtain a high probability of identifying the correct copula function; and if the dependence structures are strong, the rating capacity is quite good when the parameter is estimated using maximum likelihood, regardless

of the sample size. The GOF method cannot identify this dependence structure so in most cases, it classified data with a GB dependence as if they had an FGM or Frank dependence.

When it was evaluated situations with a dependence structure modeled by a FGM copula function, it was observed that in all cases the ratios of correct classification were under 40% when based on a maximum likelihood estimator for the parameter. If the estimate was obtained using Bayesian methods, the results could be worse. For this copula function, the rates of identification using the GOF method were close to zero in all cases. When the data have a FGM dependence, this method tends to classify them as a Frank or Gumbel-Barnett type of dependence.

TABLE 4: Proportions of accuracy in the identification of copula functions using the proposed method and the GOF method.

		Weak				**Moderate**				**Strong**			
	n	**50**	**100**	**500**	**1000**	**50**	**100**	**500**	**1000**	**50**	**100**	**500**	**1000**
Clayton	**ML**	0.97	1.00	1.00	1.00	0.99	0.99	1.00	1.00	0.99	1.00	1.00	1.00
	Bayes	0.96	1.00	1.00	1.00	0.99	0.99	1.00	1.00	0.99	1.00	1.00	1.00
	GOF	0.22	0.65	0.93	0.95	0.18	0.76	0.92	0.96	0.12	0.63	0.95	0.99
Frank	**ML**	0.36	0.40	0.77	0.95	0.65	0.89	1.00	1.00	0.66	0.83	1.00	1.00
	Bayes	0.40	0.72	0.90	0.97	0.59	0.90	1.00	1.00	0.87	1.00	1.00	0.49
	GOF	0.01	0.03	0.11	0.21	0.17	0.56	0.96	0.96	0.49	0.85	0.95	0.99
Gumbel-Hougaard	**ML**	0.45	0.59	0.93	0.98	0.69	0.84	0.99	1.00	0.93	1.00	1.00	1.00
	Bayes	0.24	0.17	0.02	0.00	0.56	0.69	0.97	0.99	0.20	0.24	1.00	0.96
	GOF	0.04	0.04	0.57	0.86	0.25	0.56	0.99	0.93	0.22	0.59	0.96	0.96
FGM	**ML**	0.21	0.23	0.27	0.27	0.17	0.30	0.41	0.41	0.22	0.24	0.42	0.42
	Bayes	0.49	0.12	0.00	0.00	0.44	0.12	0.01	0.01	0.29	0.29	0.30	0.30
	GOF	0.00	0.02	0.01	0.01	0.00	0.01	0.01	0.01	0.01	0.01	0.01	0.01
Gumbel-Barnett	**ML**	0.41	0.80	0.80	0.85	0.78	0.78	0.96	0.97	0.91	0.92	1.00	1.00
	Bayes	0.40	0.60	0.65	0.65	0.70	0.75	0.95	0.96	0.30	0.30	0.40	0.30
	GOF	0.00	0.01	0.01	0.01	0.00	0.00	0.01	0.01	0.01	0.02	0.01	0.01

10. Conclusions

In this paper it was proposed a new method to select the copula function that best fits a bivariate data set. This methodology was based on a modification of a previous dependence measure introduced by Ledwina (2015), which uses a weighted difference between cumulative probabilities based on the assumption of dependence and independence, respectively. In this study it was conducted a simulation study to evaluate the performance of our empirical procedure in order to obtain an index; in this way, it was compared the results with those obtained using the inferential procedure proposed by Deheuvels (1981) and Genest & Rémillard (2004) (i.e., the GOF method). In our study it was evaluated the true classification capacity using data sets of pairs of observations simulated from the five copula functions commonly used in the literature and presented in section 2. The performance of the proposed procedure was better when compared to the GOF

method for all data sets. The method easily identifies data from copula functions with an analytical structure such as the Clayton copula regardless some additional topics as the sample size or the dependence level. For structures type Frank copula function, the performance of the method depends of the sample size and it works better for moderate and strong dependence.If the dependence structure is like Gumbel-Hougaard copula function, to obtain a correct classification it is necessary to have sample sizes greater than 100. With dependence structures like Gumbel-Barnett copula function, the observed percentages of correct classification are better when the maximum likelihood estimation method is used to obtain the proposed index. If a Uniform(0,1) prior distribution is assumed for the dependence parameter and the dependence level is strong, the method does not classify correctly, and the performance of the measure could be improved if informative prior distributions are considered but this topic is beyond the scope of this paper and will be topic of a new study. In general, when the dependence structure shares similarities with other copula functions, the correct classification depends on different aspects such as sample sizes and the strength of dependence. The proposed procedure fails to identify data with dependence that can be modeled using the FGM copula function. It can be deduced that very weak linear dependence structures (close to independence) would be difficult to identify using the proposed method, given that, the procedure measures the difference between the observed probabilities assuming a proposed copula and the independence copula weighted by a factor that in the FGM copula case is a part of its structure. In this way, it is possible to deduce that for this copula functions family, it is necessary to conduct a detailed study on the performance of discrimination measures as it was proposed in this study. In our study, we used the GOF method to compare the classification capacity of the proposed method since we considered that method is the most general among those we found in the literature. It is important to point out that other methods have been proposed in the literature but for some specific copula function families (For instance see, Topçu 2016).

Acknowledgement

The participation of the second author was financed by a scholarship from the Virginia Gutierrez of Pineda Program for Young Researches and Innovators of the Administrative Department of Science, Technology and Innovation (COLCIENCIAS) in Colombia.

[Received: April 2018 — Accepted: November 2018]

References

Arbenz, P. (2013), 'Bayesian copulae distributions, with application to operational risk management-some comments', *Methodology and computing in applied probability* **15**(1), 105–108.

Bairamov, I., Kotz, S. & Kozubowski, T. (2003), 'A new measure of linear local dependence', *Statistics: A Journal of Theoretical and Applied Statistics* **37**(3), 243–258.

Balakrishnan, N. & Lai, C.-D. (2009), *Continuous Bivariate Distributions*, Springer, Dordrecht.

Barnett, V. (1980), 'Some bivariate uniform distributions', *Communications in statistics-theory and methods* **9**(4), 453–461.

Bjerve, S. & Doksum, K. (1993), 'Correlation curves: measures of association as functions of covariate values', *The Annals of Statistics* pp. 890–902.

Chang, K.-L. (2012), 'The time-varying and asymmetric dependence between crude oil spot and futures markets: Evidence from the mixture copula-based arji-garch model', *Economic Modelling* **29**(6), 2298–2309.

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.

Corbella, S. & Stretch, D. D. (2013), 'Simulating a multivariate sea storm using archimedean copulas', *Coastal Engineering* **76**, 68–78.

Deheuvels, P. (1981), 'An asymptotic decomposition for multivariate distribution-free tests of independence', *Journal of Multivariate Analysis* **11**(1), 102–113.

Drouet, M. & Kotz, S. (2001), *Correlation and dependence*, Imperial College Press, London.

Durante, F. & Sempi, C. (2015), *Principles of copula theory*, Chapman and Hall/CRC.

Frank, M. J. (1979), 'On the simultaneous associativity of $f(x, y)$ and $x + y - f(x, y)$', *Aequationes Mathematicae* **19**(1), 194–226.

Genest, C. & Rémillard, B. (2004), 'Test of independence and randomness based on the empirical copula process', *Test* **13**(2), 335–369.

Goethals, K., Janssen, P. & Duchateau, L. (2008), 'Frailty models and copulas: similarities and differences', *Journal of Applied Statistics* **35**(9), 1071–1079.

Gumbel, E. J. (1960a), 'Bivariate exponential distributions', *Journal of the American Statistical Association* **55**(292), 698–707.

Gumbel, E. J. (1960b), 'Distributions des valeurs extremes en plusiers dimensions', *Institut de statistique de l'Universite? de Paris* **9**, 171–173.

Gumbel, E. J. (1961), 'Bivariate logistic distributions', *Journal of the American Statistical Association* **56**(294), 335–349.

Hofert, M. (2008), 'Sampling archimedean copulas', *Computational Statistics & Data Analysis* **52**(12), 5163–5174.

Hougaard, P. (1986), 'A class of multivanate failure time distributions', *Biometrika* **73**(3), 671–678.

Janga Reddy, M. & Ganguli, P. (2012), 'Risk assessment of hydroclimatic variability on groundwater levels in the manjara basin aquifer in india using archimedean copulas', *Journal of Hydrologic Engineering* **17**(12), 1345–1357.

Jogdeo, K. (1982), 'Concepts of dependence', *Encyclopedia of statistical sciences* **2**, 324–334.

Johnson, N. & Kotz, S. (1972), *Distributions in Statistics: Continuous Multivariate Distributions*, Wileg.

Kojadinovic, I., Yan, J. & Holmes, M. (2011), 'Fast large-sample goodness-of-fit tests for copulas', *Statistica Sinica* pp. 841–871.

Kowalczyk, T., Pleszczynska, E. et al. (1977), 'Monotonic dependence functions of bivariate distributions', *The Annals of Statistics* **5**(6), 1221–1227.

Lancaster, H. O. (1982), 'Dependence, measures and indices', *Encyclopedia of statistical sciences* **2**, 334–339.

Ledwina, T. (2015), Visualizing association structure in bivariate copulas using new dependence function, *in* 'Stochastic Models, Statistics and Their Applications', Springer, pp. 19–27.

Lojowska, A., Kurowicka, D., Papaefthymiou, G., van der Sluis, L. et al. (2012), 'Stochastic modeling of power demand due to evs using copula', *IEEE Transactions on Power Systems* **27**(4), 1960.

Meintanis, S. G. (2007), 'Test of fit for marshall-olkin distributions with applications', *Journal of Statistical Planning and inference* **137**(12), 3954–3963.

Nelsen, R. B. (2006), *An introduction to copulas*, Springer.

Sklar, M. (1959), 'Fonctions de repartition an dimensions et leurs marges', *Institut de statistique de l'Universite? de Paris* **8**, 229–231.

Topçu, Ç. (2016), 'Comparison of some selection criteria for selecting bivariate archimedean copulas', *Afyon Kocatepe University Journal of Sciences and Engineering* **16**, 250–255.

Tovar, J. R. & Achcar, J. A. (2012), 'Two dependent diagnostic tests: Use of copula functions in the estimation of the prevalence and performance test parameters', *Revista Colombiana de Estadística* **35**(3), 331–347.

Tovar, J. R. & Achcar, J. A. (2013), 'Dependence between two diagnostic tests with copula function approach: a simulation study', *Communications in Statistics-Simulation and Computation* **42**(2), 454–475.

Wang, Y.-C., Wu, J.-L. & Lai, Y.-H. (2013), 'A revisit to the dependence structure between the stock and foreign exchange markets: A dependence-switching copula approach', *Journal of Banking & Finance* **37**(5), 1706–1719.

Weiss, G. (2011), 'Copula parameter estimation by maximum-likelihood and minimum-distance estimators: a simulation study', *Computational Statistics* **26**(1), 31–54.

Xie, K., Li, Y. & Li, W. (2012), 'Modelling wind speed dependence in system reliability assessment using copulas', *IET Renewable Power Generation* **6**(6), 392–399.

Zhang, Q., Xiao, M., Singh, V. P. & Chen, X. (2013), 'Copula-based risk evaluation of droughts across the pearl river basin, china', *Theoretical and applied climatology* **111**(1-2), 119–131.