

# A Copula Statistic for Measuring Nonlinear Dependence with Application to Feature Selection in Machine Learning

Mohsen Ben Hassine

Department of Computer Science,  
University of El Manar, Tunis,  
Tunisia

Lamine Mili

Bradley Department of Electrical and  
Computer Engineering, Northern  
Virginia Center, Virginia Tech, Falls  
Church, VA 22043, USA

Kiran Karra

Bradley Department of Electrical and  
Computer Engineering, VTRC-A,  
Arlington, VA 22203, USA

**Abstract**—Feature selection in machine learning aims to find out the best subset of variables from the input that reduces the computation requirement and improves the predictor performance. This paper introduces a new index based on empirical copulas, termed as the Copula Statistic (CoS) to assess the strength of statistical dependence and for testing statistical independence. It shows that this test exhibits higher statistical power than other indices. Finally, applying the CoS features selection in machine learning problems, which allow a demonstration of the good performance of the CoS.

**Keywords**—Copula; multivariate dependence; nonlinear systems; feature selection; machine learning

## I. INTRODUCTION

Measures of statistical dependence among random variables and signals are paramount in many scientific applications: engineering, signal processing, finance, biology and machine learning to cite a few. They allow one to find clusters of data points and signals, test for independence to make decisions, and explore causal relationships. The classic measure of dependence is provided by the correlation coefficient, which was introduced in 1895 by Karl Pearson. Since it relies on moments, it assumes statistical linear dependence. However, in biology, ecology and finance, and other fields, applications involving nonlinear multivariate dependence prevail. For such applications, the correlation coefficient is unreliable. Hence, researchers have initiated many proposals in order to address this deficiency [1]–[5]. Reshef *et al.* [6], [7] introduced the Maximal Information Coefficient (*MIC*) and later the Total Information Coefficient (*TIC*), Lopes-Paz *et al.* [8] proposed the Randomized Dependence Coefficient (*RDC*), and Ding *et al.* [9], [10] put forth the Copula Correlation Coefficient (*Ccor*). Additionally, Székely *et al.* [11] proposed the distance correlation (*dCor*). These metrics are able to measure monotonic and non-monotonic dependencies between random variables, but each has strengths and shortcomings [12]–[18]. Feature selection in machine learning is a typical battlefield to appraise the quality and the reliability of a dependence index, it is to find out the best subset of variables (from the input) that reduces the computation requirement and feed up the predictor algorithm for optimal performance [19], [20].

In this paper, a new index based on copulas, termed the Copula Statistic (*CoS*), for measuring the strength of nonlinear statistical dependence and for testing for statistical independence is introduced. The *CoS* ranges from zero to one and attains its lower and upper limit for the independence and the functional dependence case, respectively.

Monte Carlo simulations are carried out to estimate bias and standard deviation curves of the *CoS*, to assess its power when testing for independence. The simulations reveal that for large sample sizes, the *CoS* is approximately normal and approaches Pearson's  $\rho_P$  for the Gaussian copula and Spearman's  $\rho_S$  for many copulas. The *CoS* is shown to exhibit strong statistical power in various functional dependencies as compared to many other indices. Finally, the *CoS* is applied to feature selection problem to unveil bivariate dependence.

The paper is organized as follows. Section II proves two new and essential theorems on copulas used to derive the *CoS* index. Section III introduces a relative distance function and proves several of its properties. Section IV defines the *CoS* and provides an algorithm that implements it. Section V investigates the statistical properties of the *CoS* and treats the case of bivariate dependence. Section VI compares the performance of the *CoS* with the *dCor*, *RDC*, *Ccor*, and the *MICe* in measuring bivariate functional and non-functional dependencies between synthetic datasets. It also shows how the *CoS* can unveil statistical dependence in real datasets of breast tumor and proceed with an in-depth analysis in order to find out the best feature subset for this problem.

## II. BIVARIATE COPULA

In the following, attention is restricted to two-dimensional copulas to develop a statistical index, the *CoS*, in the bivariate dependence case. To define the *CoS* of two continuous random variables  $X$  and  $Y$  with copula  $C(u, v)$ , three definitions of bivariate dependencies are provided, from weaker to stronger versions, as introduced by Lehmann [21]. Then, three theorems are stated which help build the foundation for the *CoS*.

**Definition 1:** Two random variables,  $X$  and  $Y$ , are said to be concordant (or discordant) if they tend to simultaneously take large (or small) values.

A more formal definition is as follows. Let  $X$  and  $Y$  be two random variables taking two pairs of values,  $(x_i, y_i)$  and  $(x_j, y_j)$ .  $X$  and  $Y$  are said to be concordant if  $(x_i - x_j)(y_i - y_j) > 0$ ; they are said to be discordant if the inequality is reversed.

**Definition 2:** Two random variables,  $X$  and  $Y$ , defined on the domain  $\mathfrak{D} = \text{Range}(X) \times \text{Range}(Y)$  are said to be Positively Quadrant Dependent (PQD) if

$$P(X \leq x, Y \leq y) \geq P(X \leq x) P(Y \leq y),$$

that is,  $C(u, v) \geq \Pi(u, v)$  and Negative Quadrant Dependent (NQD) if

$$P(X \leq x, Y \leq y) < P(X \leq x) P(Y \leq y),$$

that is,  $(u, v) \leq (u, v)$  for all  $(x, y) \in \mathfrak{D}$ .

**Definition 3:** Two random variables,  $X$  and  $Y$ , are said to be comonotonic (respectively countermonotonic) if  $Y = f(X)$  almost surely and  $f(\cdot)$  is an increasing (respectively a decreasing) function.

In short, two random variables are monotonic if they are either comonotonic or countermonotonic.

**Theorem 1:** (Fréchet [13]: Let  $X$  and  $Y$  be two continuous random variables. Then,

a)  $X$  and  $Y$  are comonotonic if and only if the associated copula is equal to its Fréchet-Hoeffding upper bound, that is,  $C(u, v) = M(u, v) = \min(u, v)$ ;

b)  $X$  and  $Y$  are countermonotonic if and only if the associated copula is equal to its Fréchet-Hoeffding lower bound, that is,  $C(u, v) = W(u, v) = \max(u + v - 1, 0)$

c)  $X$  and  $Y$  are independent if and only if the associated copula is equal to the product copula, that is,  $C(u, v) = \Pi(u, v) = uv$ .

In the following theorems and corollaries, it is assumed that  $X$  and  $Y$  are continuous random variables and related via a function  $f(\cdot)$ , that is,  $Y = f(X)$ , where  $f(\cdot)$  is continuous and differentiable over the range of  $X$ .

**Theorem 2:** Let  $X$  and  $Y$  be two continuous random variables such that  $Y = f(X)$  almost surely, and let  $C(u, v)$  be the copula value for the pair  $(x, y)$ . The function  $f(\cdot)$  has a global maximum at  $(x_1, y_{\max})$  with a copula value  $C(u_1, v_1)$  or a global minimum at  $(x_2, y_{\min})$  with a copula value  $C(u_2, v_2)$  if and only if

- a)  $C(u_1, v_1) = M(u_1, v_1) = W(u_1, v_1) = \Pi(u_1, v_1) = u_1$ ; (1)  
b)  $C(u_2, v_2) = M(u_2, v_2) = W(u_2, v_2) = \Pi(u_2, v_2) = 0$ . (2)

The proof of Theorem 2 is given in the appendix. For a general definition of the copula, the reader is referred to Nelsen [13].

**Corollary 1:** Let  $X$  and  $Y$  be two continuous random variables such that  $Y = f(X)$ , almost surely. If  $f(\cdot)$  is a periodic function, then (1) and (2) holds true at all the global maxima and global minima, respectively.

The proof of Corollary 1 directly follows from Theorem 2. This corollary is demonstrated in Fig. 1, which displays the graph of the projections on the  $(u, C(u, v))$  plane of the

empirical copula  $C(u, v)$  associated with a pair  $(X, Y)$ , where  $X$  is uniformly distributed over  $[-1, 1]$ , and  $Y = \sin(2\pi X)$ . It is observed that at each one of the four optima of the sine function,  $C(u, v) = M(u, v) = W(u, v) = \Pi(u, v)$ .

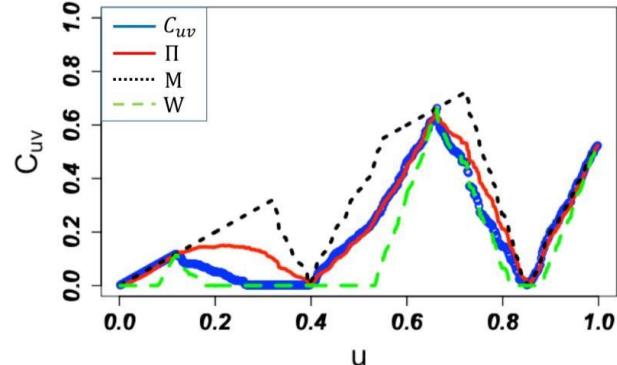


Fig. 1. Graph (in blue dots) of the projections on the  $(u, C(u, v))$  plane of the empirical copula  $C(u, v)$  associated with a pair of random variables  $(X, Y)$ , where  $X \sim U(-1, 1)$  and  $Y = \sin(2\pi X)$ . The  $u$  coordinates of the data points are equally spaced over the unity interval. Similar graphs are shown for the  $M(u, v)$ ,  $W(u, v)$  and  $\pi(u, v)$  copulas.

**Theorem 3:** Let  $X$  and  $Y$  be two continuous random variables such that  $Y = f(X)$ , almost surely where  $f(\cdot)$  has a single optimum and let  $C(u, v)$  be the copula value for the pair  $(x, y)$ . Then,  $C(u, v) = M(u, v)$  if and only if  $df(x)/dx \geq 0$  and  $C(u, v) = W(u, v)$  otherwise.

The proof of Theorem 3 is provided in the appendix. Theorem 3 is illustrated in Fig. 2.

### III. THE RELATIVE DISTANCE FUNCTION

A metric of proximity of the copula to the upper or the lower bounds with respect to the  $\Pi$  copula is defined and its properties are investigated.

**Definition 4:** The relative distance function,  $\lambda(C, (u, v))$ :  $[0, 1] \rightarrow [0, 1]$ , is defined as

- a)  $\lambda(C, (u, v)) = (C(u, v) - uv)/(Min(u, v) - uv)$  if  $C(u, v) \geq uv$ ;  
b)  $\lambda(C, (u, v)) = (C(u, v) - uv)/(Max(u+v-1, 0) - uv)$  if  $C(u, v) < uv$ .

A graphical illustration for the relative distance is shown in Fig. 3.

**Theorem 4:**  $\lambda(C, (u, v))$  satisfies the following properties:

- a)  $0 \leq \lambda(C, (u, v)) \leq 1$  for all  $(u, v) \in \mathbb{I}^2$ ;  
b)  $\lambda(C, (u, v)) = 0$  for all  $(u, v) \in \mathbb{I}^2$  if and only if  $C(u, v) = uv$ ;

c) If  $Y = f(X)$  almost surely, where  $f(\cdot)$  is monotonic, then  $\lambda(C, (u, v)) = 1$  for all  $(u, v) \in \mathbb{I}^2$ ;

d) If  $Y = f(X)$  almost surely, then  $\lambda(C, (u, v)) = 1$  at the global optimal points of  $f(\cdot)$ .

**Proof:** Property a) follows from Definition 5 and (3) while properties b), c) and d) follow from Definition 5 and Theorem 1 and 2. ■

**Corollary 2:** If  $Y = f(X)$  almost surely, where  $f(\cdot)$  has a single optimum, then  $\lambda(C, (u, v)) = 1$  for all  $(u, v) \in \mathbb{I}^2$ .

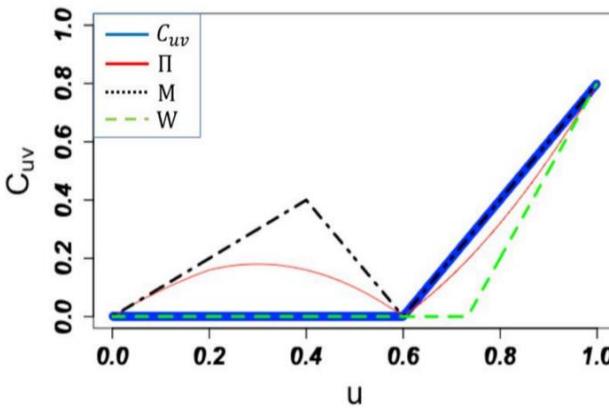


Fig. 2. Graph (blue circles) of the projections on the  $(u, C(u,v))$  plane of  $C(u,v)$  associated with  $X \sim U(-5,5)$  and  $Y = f(X) = (X-1)^2$ . The  $u$  coordinates of the data points are equally spaced. The minimum of the function  $f(\cdot)$  is associated with  $u = 0.6$  and  $C(u,v) = 0$ . Similar graphs are shown for  $M(u,v)$  (dotted black),  $W(u,v)$  (dashed green), and  $\Pi(u,v)$  (solid red). Here,  $C(u,v) = W(u,v)$  for  $0 \leq u \leq 0.6$ , which corresponds to  $f'(x) \leq 0$ , and  $C(u,v) = M(u,v)$  for  $0.6 \leq u \leq 1$ , which corresponds to  $f'(x) \geq 0$ .

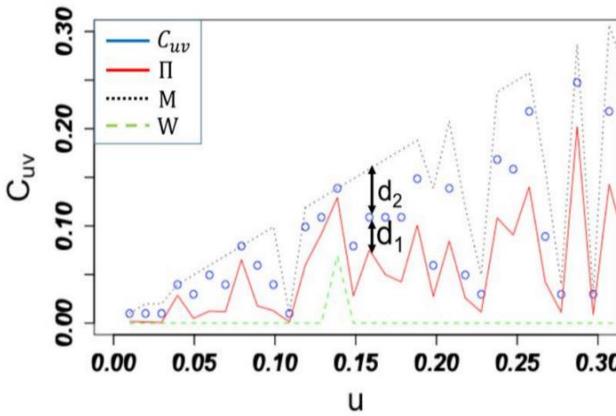


Fig. 3. Graph (blue circles) of the projections on the  $(u, C(u,v))$  plane drawn from the Gaussian copula  $C(u,v)$  with  $\rho_P = 0.5$ . Similar graphs are shown for  $M(u,v)$  (dotted black),  $W(u,v)$  (dashed green), and  $\pi(u,v)$  (solid red). The empirical relative distance function is given by  $\lambda(C(u,v)) = d_1/d_2$ , where  $d_1$  is the distance from  $C(u,v)$  to  $\Pi(u,v)$  and  $d_2$  is the distance from  $M(u,v)$  to  $\Pi(u,v)$ .

Now, the question that arises is the following: Is  $\lambda(C(u,v)) = 1$  for all  $(u,v) \in \mathbb{I}^2$  when there is a functional dependence with multiple optima, be they global or local? The answer is given by the following two theorems:

**Theorem 5:** If  $Y = f(X)$  almost surely where  $f(\cdot)$  has at least two global maxima or two global minima and no local optima on the domain  $\mathcal{D} = \text{Range}(X) \times \text{Range}(Y)$ , then there exists a non-empty interval of  $X$  for which  $\lambda(C(u,v)) < 1$ .

The proof of Theorem 5 is provided in the appendix.

**Theorem 6:** If  $Y = f(X)$  almost surely, where  $(\cdot)$  has a local optimum, then  $\lambda(C(u,v)) \leq 1$  at that point.

The proof of Theorem 6 is provided in the appendix.

#### IV. THE COPULA STATISTIC

The empirical copula is first defined; then, the copula statistic is introduced, and finally, an algorithm that implements it is provided. One possible definition for the  $CoS$  is the mean of  $\lambda(C(u,v))$  over  $\mathbb{I}^2$ , that is,  $CoS(X,Y) = E[\lambda(C(u,v))]$ . However, according to Theorems 5 and 6,  $CoS \leq 1$  for functional dependence with multiple optima, which is not a desirable property. This prompts a better definition of the  $CoS$  based on the empirical copula as explained next.

##### A. The Empirical Copula

Let  $\{(x_i, y_i), i=1, \dots, n, n \geq 2\}$  be a 2-dimensional data set of size  $n$  drawn from a continuous bivariate joint distribution function,  $H(x, y)$ . Let  $R_{xi}$  and  $R_{yi}$  be the rank of  $x_i$  and of  $y_i$ , respectively. Deheuvels [22] defines the associated empirical copula as

$$C_n(u, v) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}(u_i = \frac{R_{xi}}{n} \leq u, v_i = \frac{R_{yi}}{n} \leq v). \quad (3)$$

The empirical relative distance,  $\lambda(C_n(u,v))$ , satisfies Definition 4 by replacing  $C(u,v)$  with the empirical copula given by (3).

##### B. Defining the CoS Statistic for Bivariate Dependence

Let  $X$  and  $Y$  be two continuous random variables with a copula  $C(u,v)$ . Consider the ordered sequence,  $x_{(1)} \leq \dots \leq x_{(n)}$ , of  $n$  realizations of  $X$ . This sequence yields  $u_{(1)} \leq \dots \leq u_{(n)}$  since  $u_i = R_{xi}/n$  as given by (3). Let  $\mathcal{D}$  be the set of  $m$  contiguous domains  $\{\mathcal{D}_i, i = 1, \dots, m\}$ , where each  $\mathcal{D}_i$  is a  $u$ -interval associated with a non-decreasing or non-increasing sequence of  $C_n(u_{(i)}, v_j)$ ,  $i = 1, \dots, n$ . Let  $C_i^{\min}$  and  $C_i^{\max}$  respectively denote the smallest and the largest value of  $C_n(u,v)$  on the domain  $\mathcal{D}_i$ . Let  $\gamma_i$  be defined as:

$$\gamma_i = \begin{cases} 1 & \text{at a local optimum of } Y = f(X) \text{ on } \mathcal{D}_i, \\ \frac{\lambda(C_i^{\min}) + \lambda(C_i^{\max})}{2}, & \text{otherwise.} \end{cases} \quad (4)$$

**Definition 5:** Let  $n_i$  denote the number of data points in the  $i$ -th domain  $\mathcal{D}_i$ ,  $i = 1, \dots, m$ , while letting a boundary point belong to two contiguous domains,  $\mathcal{D}_i$  and  $\mathcal{D}_{i+1}$ . Then, the copula statistic is defined as:

$$CoS(X, Y) = \frac{1}{n+m-1} \sum_{i=1}^m n_i \gamma_i \quad (5)$$

**Corollary 3:** The  $CoS$  of two random variables,  $X$  and  $Y$ , has the following asymptotic properties:

- a)  $0 \leq CoS(X, Y) \leq 1$ ;
- b)  $CoS(X, Y) = 0$  if and only if  $X$  and  $Y$  are independent;
- c) If  $Y = f(X)$  almost surely, then  $CoS(X, Y) = 1$ .

##### C. Algorithmic implementation of the Copula Statistic

Given a two-dimensional data sample of size  $n$ ,  $\{(x_j, y_j), j=1, \dots, n, n \geq 2\}$ , the algorithm that calculates the  $CoS$  consists of the following steps:

- 1) Calculate  $u_j$ ,  $v_j$  and  $C_n(u,v)$  as follows:  
a.  $u_j = \frac{1}{n} \sum_{j=1}^n \mathbf{1}\{k \neq j : x_k \leq x_j\}$ ;

- b.  $v_j = \frac{1}{n} \sum_{j=1}^n \mathbf{1}\{k \neq j: y_k \leq y_j\}$ ;  
c.  $C_n(u, v) = \frac{1}{n} \sum_{j=1}^n \mathbf{1}\{u_j \leq u, v_j \leq v\}$ ;
- 2) Order the  $x_j$ 's to get  $x_{(1)} \leq \dots \leq x_{(n)}$ , which results in  $u_{(1)} \leq \dots \leq u_{(n)}$  since  $u_j = R_{x_j}/n$ , where  $R_{x_j}$  is the rank of  $x_j$ ;
- 3) Determine the domains  $\mathfrak{D}_i$ ,  $i = 1, \dots, m$ , where each  $\mathfrak{D}_i$  is a  $u$ -interval associated with a non-decreasing or non-increasing sequence of  $C_n(u_{(j)}, v_p)$ ,  $j = 1, \dots, n$ .
- 4) Determine the smallest and the largest value of  $C_n(u, v)$ , denoted by  $C_i^{\min}$  and  $C_i^{\max}$ , and find the associated  $u_i^{\min}$  and  $u_i^{\max}$  for each domain  $\mathfrak{D}_i$ ,  $i = 1, \dots, m$ .
- 5) Calculate  $\lambda(C_i^{\min})$  and  $\lambda(C_i^{\max})$ ;
- 6) If  $\lambda(C_i^{\min})$  and  $\lambda(C_i^{\max})$  are equal to one, go to Step 8;
- 7) Calculate the absolute difference between the three consecutive values of  $C_n(u_{(i)}, v_p)$  centered at  $u_i^{\min}$  (respectively at  $u_i^{\max}$ ) and decide that the central point is a local optimum if (i) both absolute differences are smaller than or equal to  $1/n$ ; and (ii) there are more than four points within the two adjacent domains,  $\mathfrak{D}_i$  and  $\mathfrak{D}_{i+1}$ ;
- 8) Calculate  $\gamma_i$  given by (16);
- 9) Repeat Steps 2 through 7 for all the  $m$  domains,  $\mathfrak{D}_i$ ,  $i = 1, \dots, m$ ;
- 10) Calculate the  $CoS$  given by (17).

## V. STATISTICAL PROPERTIES OF THE $CoS$

The finite-sample bias of the  $CoS$  is analyzed for the independence case, then a statistical test of bivariate independence is developed.

### 1) Finite-Sample Bias of the $CoS$

Table 1 displays the sample means and the sample standard deviations of the  $CoS$  for independent random samples generated from three monotonic copulas. As observed, the  $CoS$  has a bias for small to medium sample sizes. Fig. 4(a) shows a bias curve given by  $CoS = 8.05 n^{-0.74}$ , fitted to 19 mean bias values for  $Gauss(0)$  using the least-squares method. It is observed that the  $CoS$  bias becomes negligible for a sample size larger than 500. Fig. 4(b) shows values taken by the sample standard deviation  $\sigma_n$  of  $CoS$  for increasing sample size,  $n$ , and for  $Gauss(0)$ . A fitted curve obtained is also displayed; it is expressed as  $\sigma_n = 2.99 n^{-0.81}$ .

### 2) Independence Test

One common practical problem is to test the independence of random variables. To this end, hypothesis testing can be applied to the  $CoS$  based on Corollary 3b). The goal is to test the null hypothesis,  $H_0$ : the random variables are independent, against its alternative,  $H_1$ . The  $CoS$  is standardized under  $H_0$  to get

$$z_n = \frac{CoS - \mu_{n0}}{\sigma_{n0}} \quad (6)$$

Where,  $\mu_{n0}$  and  $\sigma_{n0}$  are the sample mean and the sample standard deviation of the  $CoS$ , respectively. Note that as observed in Fig. 4(a) and (b), for a number of samples  $n$  larger than 500,  $\mu_{n0}$  becomes negligible and  $\sigma_{n0}$  is approximately equal to 0.01.

Hypothesis testing consists of choosing a threshold  $c$  at a significance level  $\alpha$  under  $H_0$  and then applying the following

decision rule: if  $|z_n| \leq c$ , accept  $H_0$ ; otherwise, accept  $H_1$ . Table 2 displays Type-II errors of the statistical test applied to the  $CoS$  for  $Gauss(0)$  for sample sizes ranging from 100 to 3000. It is observed that Type II-errors decrease as increases for a given  $n$  and sharply decrease with increasing  $n$ .

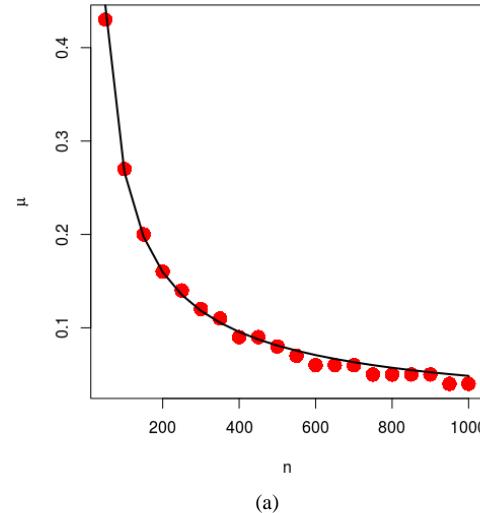
TABLE I. SAMPLE MEANS AND SAMPLE STANDARD DEVIATIONS OF THE  $CoS$  FOR THE GAUSSIAN, GUMBEL, AND CLAYTON COPULA IN THE INDEPENDENCE CASE

$n$	$Gauss(0)$ $\rho_P = 0$		$Gumbel(1)$ $\rho_S = 0$		$Clayton(0)$ $\rho_S = 0$	
	$\mu_n$	$\sigma_n$	$\mu_n$	$\sigma_n$	$\mu_n$	$\sigma_n$
100	0.28	0.08	0.28	0.08	0.28	0.08
500	0.08	0.02	0.08	0.03	0.08	0.02
1000	0.04	0.01	0.04	0.01	0.05	0.01
2000	0.02	0.01	0.02	0.01	0.02	0.01
3000	0.02	0.01	0.02	0.01	0.02	0.01

For monotonic dependence, simulation results show that  $CoS = 1$  for all  $n \geq 2$ . For non-monotonic dependence, there is a bias that becomes negligible when the sample size is sufficiently large. As an illustrative example, Table 3 displays the sample mean,  $\mu_n$ , and the sample standard deviation,  $\sigma_n$ , of the  $CoS$  for increasing sample size,  $n$ , for the sinusoidal dependence,  $Y = \sin(a X)$ . It is observed that as the frequency of the sine function increases, the sample bias,  $1 - \mu_n$ , increases for constant  $n$ .

Table 4 displays  $\mu_n$  and  $\sigma_n$  of the  $CoS$  calculated for increasing  $n$  and for different degrees of dependencies of two dependent random variables following the Gaussian copula. It is interesting to note that for  $n \geq 1000$ , the  $CoS$  is nearly equal to the Pearson's  $\rho_P$  for the Gaussian copula and to the Spearman's  $\rho_S$  for other copulas.

### 3) Bivariate Dependence



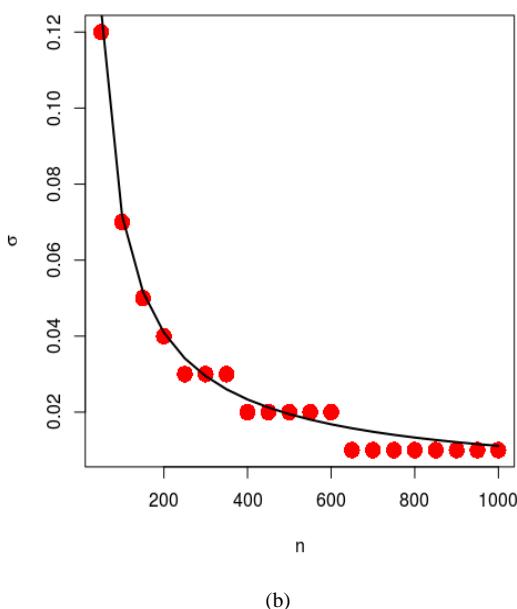


Fig. 4. (a) Bias mean values and (b) standard deviation values (red solid circles) for the  $CoS$  along with fitted curves (solid lines) using the least-squares method for the independence case.

TABLE II. TYPE-II ERRORS OF THE STATISTICAL TEST OF BIVARIATE INDEPENDENCE BASED ON  $CoS$  FOR  $GAUSS(0)$

$N$	$\mu_{n0}$	$\sigma_{n0}$	Type-II error for $\rho_n = 0.1$	Type-II error for $\rho_n = 0.3$
100	0.28	0.08	97%	46%
500	0.08	0.02	27%	0%
1000	0.04	0.01	0%	0%
2000	0.02	0.01	0%	0%
3000	0.02	0.01	0%	0%

TABLE III. SAMPLE MEANS AND SAMPLE STANDARD DEVIATIONS OF THE  $COS$  FOR THREE SINUSOIDAL FUNCTIONS OF INCREASING FREQUENCY

$N$	$Sin(x)$		$Sin(5x)$		$Sin(14x)$	
	$\mu_n$	$\sigma_n$	$\mu_n$	$\sigma_n$	$\mu_n$	$\sigma_n$
100	1.00	0.00	0.91	0.10	0.67	0.10
500	1.00	0.00	0.99	0.03	0.88	0.07
1000	1.00	0.00	1.00	0.01	0.96	0.04
3000	1.00	0.00	1.00	0.00	1.00	0.01
5000	1.00	0.00	1.00	0.00	1.00	0.00

## VI. COMPARATIVE STUDY

In this section, bivariate synthetic datasets and multivariate datasets of breast tumor cells are analyzed.

### A. Synthetic Datasets

In this section, the performances of the  $CoS$ ,  $dCor$ ,  $RDC$ ,  $Ccor$ , and of the  $MICe$  for various types of statistical dependencies is compared. Székely *et al.* [11] define the distance correlation,  $dCor$ , between two random vectors,  $X$  and  $Y$ , with finite first moments as

$$dCor(X, Y) = \begin{cases} \frac{v^2(X, Y)}{\sqrt{v^2(X)v^2(Y)}} & \text{for } v^2(X)v^2(Y) > 0, \\ 0 & \text{for } v^2(X)v^2(Y) = 0, \end{cases} \quad (7)$$

where  $v^2(X, Y)$  is the distance covariance. Lopes-Paz *et al.* [9] define the  $RDC$  as the largest canonical correlation between  $k$  randomly chosen nonlinear projections of the copula transformed data. Ding *et al.* [9], [10] define the copula correlation ( $Ccor$ ) as half of the  $L_1$  distance between the copula density and the independence copula density. As for the  $MIC$ , it is defined by Reshef *et al.* [6] as the maximum taken over all  $x$ -by- $y$  grids  $G$  up to a given grid resolution, typically  $x, y < n^{0.6}$ , of the empirical standardized mutual information,  $I_G(x, y)/\log(\min\{x, y\})$ , based on the empirical probability distribution over the boxes of a grid  $G$ . Formally,

$$MIC(X, Y) = \max \left\{ \frac{I_G(x, y)}{\log(\min\{x, y\})} \right\}. \quad (8)$$

### 1) Bias Analysis for Non-Functional Dependence

A bias analysis is performed for the  $MICe$ , the  $Ccor$ , the  $CoS$ , the  $RDC$ , and the  $dCor$  and using three data samples drawn from a bivariate Gaussian copula with  $\rho_p(X, Y) = 0.2$ , 0.5 and 0.8, which models a weak, medium and strong dependence, respectively.

TABLE IV. SAMPLE MEANS AND SAMPLE STANDARD DEVIATIONS OF THE  $COS$  FOR THE NORMAL COPULA

$N$	$Gauss(0.1)$ $\rho_p = 0.1$		$Gauss(0.3)$ $\rho_p = 0.3$	
	$\mu_n$	$\sigma_n$	$\mu_n$	$\sigma_n$
100	0.33	0.09	0.49	0.09
500	0.14	0.05	0.36	0.05
1000	0.11	0.03	0.33	0.04
2000	0.09	0.02	0.32	0.03

The sample sizes range from 50 to 2000, in steps of 50. From Fig. 5, it is observed that unlike the *MICe* and *Ccor*, the *CoS*, *RDC*, and the *dCor* are almost equal to for large sample size.

### 2) Functional Dependence

Another series of simulations are conducted to compare the performance of the *MICe*, the *Ccor*, the *CoS*, the *RDC*, and the *dCor* when they are applied to four data sets drawn from an affine, polynomial, periodic, and circular bivariate relationship with an increasing level of white Gaussian noise. Described in [23], the procedure is executed with  $N = n = 1000$ , where  $n$  is the number of realizations of a uniform random variable  $X$  and  $N$  is the number of times the procedure is executed.

It is inferred from Table 5 that while the *CoS*, *dCor*, *Ccor* steadily decrease as the noise level  $p$  increases, the *MICe* sharply decreases as  $p$  grows from 0.5 to 2 and then reaches a plateau for  $p > 2$ . The *RDC* also decreases steadily with an increase in noise level for the functional dependencies considered, except for the quadratic dependence where it maintains a high power even under heavy noise level.

### 3) Ripley's Forms and Copula's Induced Dependence

Table 6 reports values of the *MICe*, the *Ccor*, the *CoS*, the *RDC*, and the *dCor* for Ripley's forms, and copula-induced dependencies for a sample size  $n = 1000$  averaged over 1000 Monte-Carlo simulations. The values of the Spearman's  $\rho_s$  for Gumbel(5), Clayton(-0.88), Galambos(2), and BB6(2, 2) copulas are calculated using the copula and the CDVine toolboxes of the software package R. As for the four Ripley's forms displayed in Fig. 6, a linear congruential generator using the Box-Muller transformation is used to generate several bivariate sequences with nonlinear dependencies.

Table 6 shows that the *CoS*, *MICe*, *RDC*, and *Ccor* correctly reveal some degree of nonlinear dependence for Ripley's form 2, with the *Ccor* detecting the highest level of dependence and the *dCor* the lowest level. It is observed that the *Ccor* is the only metric to correctly reveal some degree of nonlinear dependence for Ripley's form 3.

Furthermore, unlike the *MICe* values, the *dCor* and the *CoS* values are very close to the Pearson's  $\rho_P$  value for the Gaussian copula and to the Spearman's  $\rho_s$  values for the Gumbel, Clayton, Galambos and BB6 copulas.

### B. Statistical Power Analysis

Finally, following Simon and Tibshirani [23], the power of the statistical tests based on the *CoS*, *dCor*, *RDC*, *TICe*, and the *Ccor* for bivariate independence subject to increasing additive Gaussian noise levels is tested. Six noisy functional dependencies at a noise level  $p$  ranging from 10% to 300% are considered. They include a linear, a quadratic, a cubic, a fourth-root, a sinusoidal, and a circular dependence. The results are shown in Fig. 7.

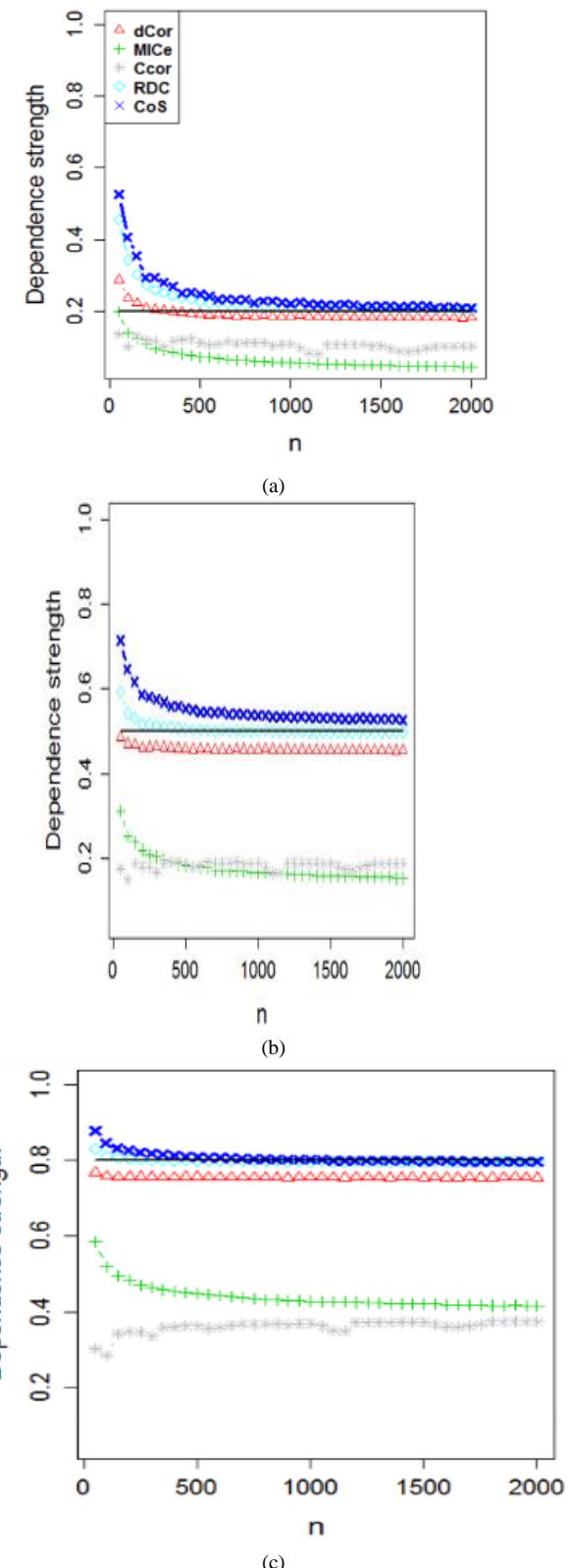


Fig. 5. Bias curves of the *CoS*, *MICe*, *dCor*, *RDC*, and *Ccor* for the bivariate Gaussian copula with  $\rho_P(X,Y) = 0.2, 0.5$  and  $0.8$ , which are displayed in a), b), and c), respectively, and for sample sizes that vary from 50 and 2000 with steps of 50.

TABLE V. SAMPLE MEANS OF THE *COS*, *DCOR* AND THE *MICE* FOR SEVERAL DEPENDENCE TYPES AND ADDITIVE NOISE LEVELS

Noise level p Type of dependence	0.5	1	2	3	
Affine: $Y = 2X+1$	<i>CoS</i>	0.86	0.72	0.41	0.29
	<i>dCor</i>	0.91	0.71	0.46	0.35
	<i>MICe</i>	0.88	0.46	0.26	0.22
	<i>RDC</i>	0.95	0.74	0.60	0.59
	<i>Ccor</i>	0.63	0.47	0.34	0.30
$4^{\text{th}}$ -order Polynomial: $Y=(X^2-0.25)(X^2-1)$	<i>CoS</i>	0.64	0.41	0.29	0.26
	<i>dCor</i>	0.41	0.35	0.31	0.30
	<i>MICe</i>	0.79	0.54	0.49	0.48
	<i>RDC</i>	0.95	0.93	0.92	0.91
	<i>Ccor</i>	0.72	0.63	0.60	0.59
Periodic: $Y = \cos(X)$	<i>CoS</i>	0.53	0.46	0.28	0.23
	<i>dCor</i>	0.35	0.27	0.17	0.13
	<i>MICe</i>	0.78	0.40	0.22	0.19
	<i>RDC</i>	0.85	0.67	0.43	0.36
	<i>Ccor</i>	0.57	0.41	0.29	0.26

### C. Feature Selection Applied to Breast Cancer Data

In order to reduce computation time, improve prediction performance and reducing irrelevant data in machine learning applications, the feature selection presents the all-important step required to choose the optimal subset of data.

The dependent variables provide useless information about the classes and thus serve as noise for the predictor. The rule of thumb here is that best feature selection must include independent features that have a strong dependence with the class or the label considered. The dimensionality reduction is a part of most known methods in machine learning such as filter, wrapper end embedded methods. Pearson correlation coefficient and mutual information are largely used in feature selection; nevertheless, the results are still unsuitable. A serious alternative here is using the CoS index to work out the feature selection problem.

A useful data for this purpose is the Wisconsin Diagnostic Breast Cancer (WDBC) data, available on UCI machine learning repository. The extraction of breast tumor tissue is performed using a fine needle aspiration (FNA). The procedure begins by obtaining a small drop of the fluid in hand by examining the characteristics of individual cells and important contextual features such as the radius of the nucleus, the compactness, the smoothness, among others.

A dataset of 569 cells (malignant and benign) and 30 input features is obtained [24]. Among the 30 features, 20 considered are computed from the others; hence, only 10 features are considered as initial subset. Table 6 reports the CoS measures for all pairwise feature dependence.

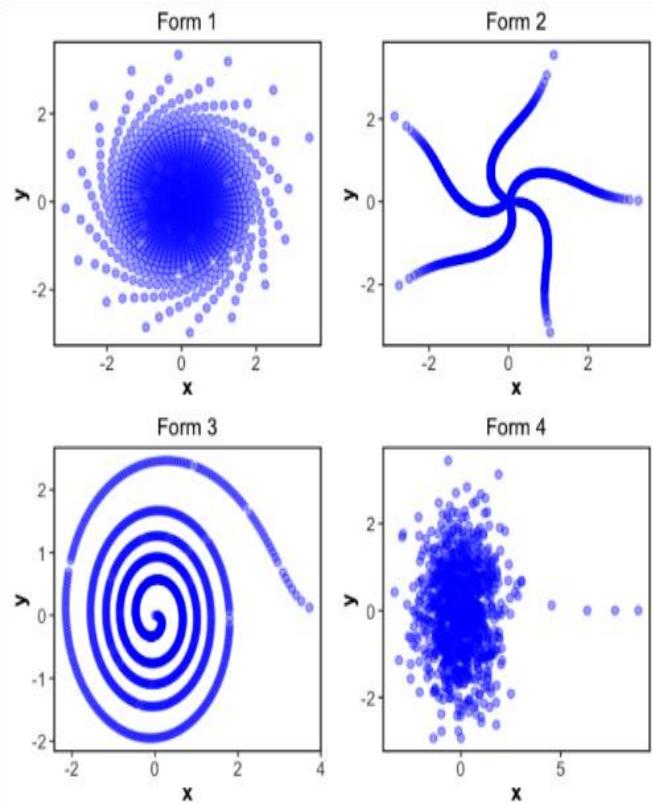


Fig. 6. Plots of four Ripley's forms generated using a linear congruential generator followed by the Box-Muller transformation. The parameters of the congruential generator,  $x_{i+1} = (a x_i + c) \bmod M$ , are as follows: Form 1:  $a = 65, c = 1, M = 2048$ ; Form 2:  $a = 1229, c = 1, M = 2048$ ; Form 3:  $a = 5, c = 1, M = 2048$ ; Form 4:  $a = 129, c = 1, M = 2^{64}$ .

TABLE VI. DEPENDENCE INDICES FOR COPULA DEPENDENCIES AND RIPLEY'S FORMS

Type of Dependence	<i>CoS</i>	<i>dCor</i>	<i>MICe</i>	<i>RDC</i>	<i>Ccor</i>
Ripley's form 1	0.01	0.02	0.02	0.02	0.01
Ripley's form 2	0.52	0.19	0.42	0.42	0.84
Ripley's form 3	0.14	0.08	0.12	0.13	0.26
Ripley's form 4	0.03	0.04	0.03	0.08	0.09
Gaussian(0.1)	0.11	0.10	0.04	0.13	0.10
Gumbel(5)	0.92	0.93	0.72	0.96	0.62
Clayton(-0.88)	0.90	0.87	0.68	0.88	0.75
Galambos(2)	0.82	0.79	0.48	0.86	0.42
BB6(2,2)	0.84	0.83	0.57	0.92	0.48

Using 0.90 as a threshold to decide a total dependence, the subset is reduced to only 7 features. If the choice is spanned to a threshold of 0.85, the subset length is further reduced to five features [25]. Fig. 8 displays the scatters of the final subset empirical copulas while Fig. 9 displays the heat maps.

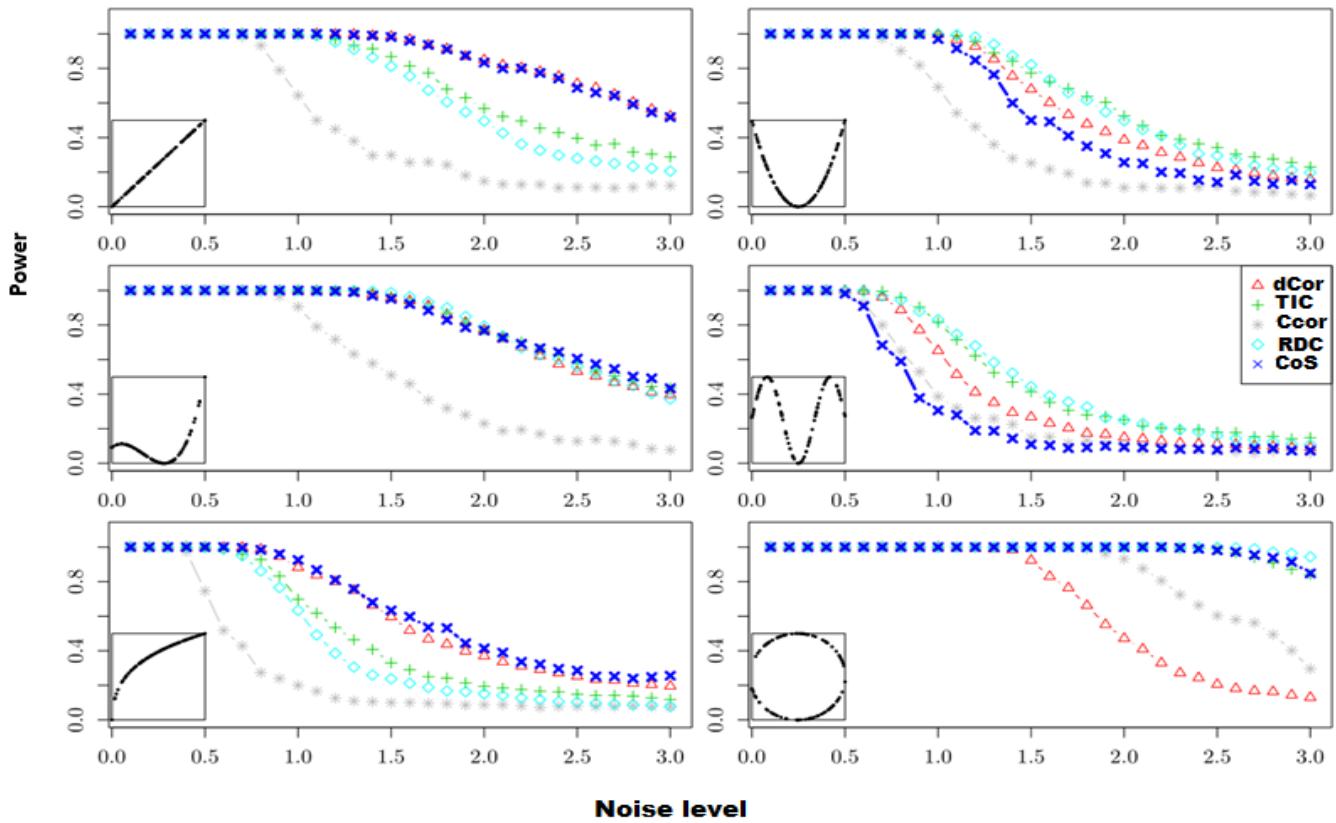


Fig. 7. Displays the power of the tests calculated using a collection of  $N = 500$  data samples, each of size  $n = 500$ , for a significance level  $\alpha = 5\%$  under the null hypothesis. As observed from that figure, the CoS is a powerful measure of dependence for the linear, cubic, circular and rational dependence.

TABLE VII. *CoS* VALUES FOR BIVARIATE DEPENDENCE BETWEEN FEATURES

	Radius	Texture	Perimeter	Area	Smooth	Compact	Concav	Nbrconcav	Sym	Fractal
Radius	1	0.23	<b>0.99</b>	<b>0.99</b>	0.37	0.33	0.60	0.72	0.18	0.41
Texture	-	1	0.24	0.25	0.18	0.27	0.26	0.21	0.17	0.18
Perimeter	-	-	1	<b>0.99</b>	0.40	0.40	0.60	0.78	0.17	0.33
Area	-	-	-	1	0.37	0.30	0.61	0.75	0.15	0.43
Smooth	-	-	-	-	1	0.80	0.75	0.74	0.72	<b>0.85</b>
Compact	-	-	-	-	-	1	<b>0.88</b>	0.84	0.76	0.84
Concav	-	-	-	-	-	-	1	<b>0.93</b>	0.67	0.64
Nbrconcav	-	-	-	-	-	-	-	1	0.56	0.48
Sym	-	-	-	-	-	-	-	-	1	0.71
Fractal	-	-	-	-	-	-	-	-	-	1

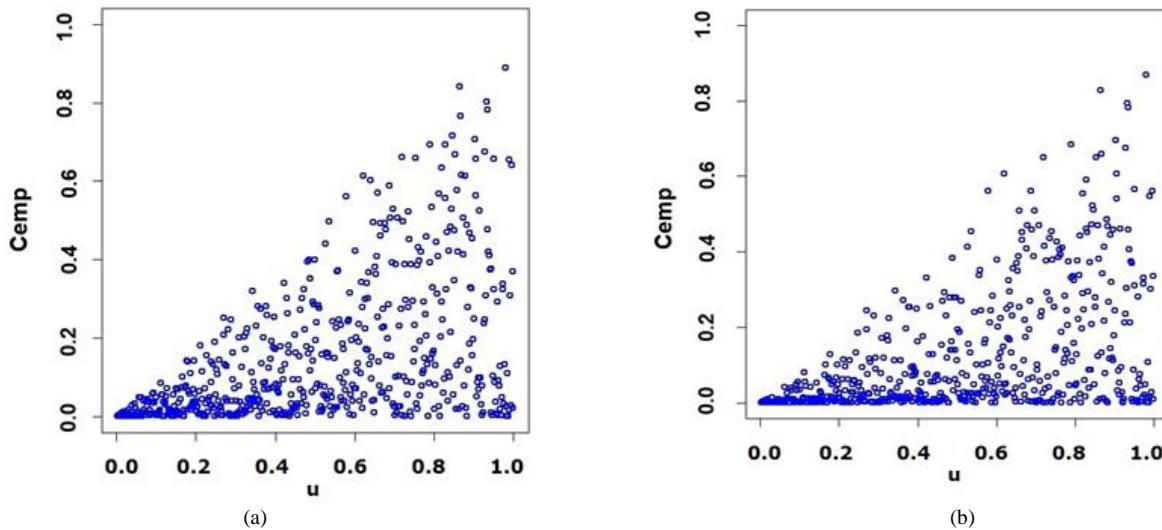


Fig. 8. Scatters of empirical copulas for a) the final feature subset except the perimeter and for b) the final subset, where the CoS values are respectively 0.27 and 0.26.

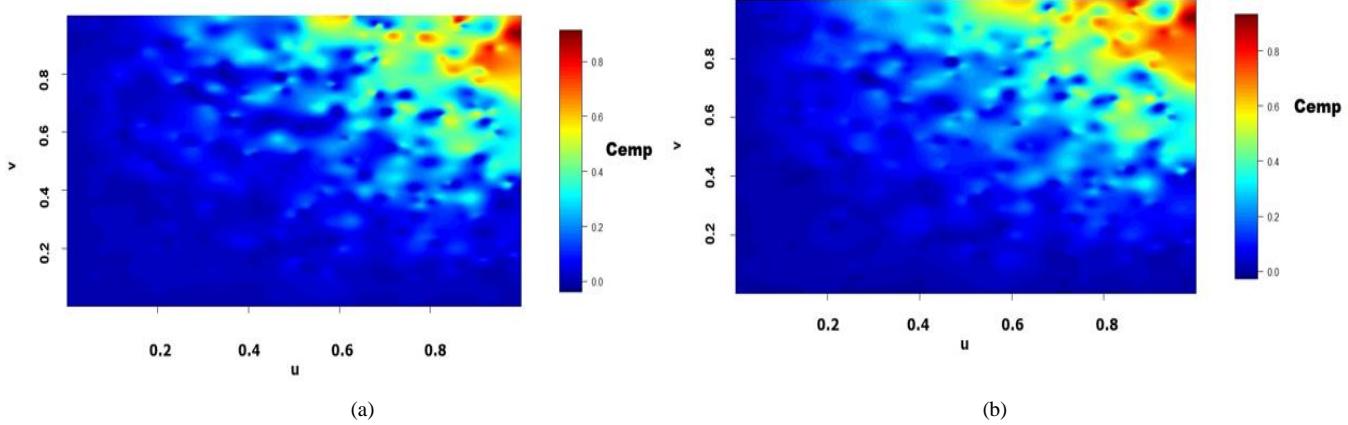


Fig. 9. Heat maps of empirical copulas for a) the final feature subset except the perimeter and for b) the final subset, CoS values are respectively 0.27 and 0.26.

## VII. CONCLUSIONS AND FUTURE WORK

A new reliable statistic for multivariate nonlinear dependence has been proposed and its statistical properties unveiled. In particular, it asymptotically approaches zero for statistical independence and one for functional dependence. Finite-sample bias and standard deviation curves of the *CoS* have been estimated and hypothesis testing rules have been developed to test bivariate independence. The power of the *CoS*-based test has been evaluated for noisy functional dependencies. Monte Carlo simulations show that the *CoS* performs reasonably well for both functional and non-functional dependence and exhibits a good power for testing independence against all alternatives. Good performance of the *CoS* was proved also with other application. Note that the code that implements the *CoS* is available on the GitHub repository.<sup>1</sup> As a future research work, the self-equitability of the *CoS* and other metrics will be assessed under various noise probability distributions and the robustness of the *CoS* to

outliers will be investigated. Furthermore, the *CoS* will be applied to common signal processing and more machine learning problems, including data mining, cluster analysis, and testing of independence. Another interesting property of the *CoS* that is not shared by the *MICe*, *RDC*, *Ccor*, and the *dCor* is its ability to measure multivariate dependence. This property will be investigated as a future work.

## REFERENCES

- [1] C. Spearman, The proof and measurement of association between two things, *The American Journal of Psychology* 15 (1) (1904) 72–101.
- [2] M.G. Kendall, A new measure of rank correlation, *Biometrika* 30 (1/2) (1938) 81–89.
- [3] T. Kowalczyk, Link between grade measures of dependence and of separability in pairs of conditional distributions, *Statistics and Probability Letters* 46 (2000) 371–379.
- [4] F. Vandenbende, P. Lambert, Improved rank-based dependence measures for categorical data, *Statistics and Probability Letters* 63 (2003) 157–163.
- [5] R.B. Nelsen, M. Úbeda-Flores, How close are pairwise and mutual independence? *Statistics and Probability Letters* 82 (2012) 1823–1828.

<sup>1</sup> <https://github.com/stochasticresearch/copulastatistic>

- [6] D. N. Reshef, Y.A. Reshef, H. K. Finucane, S. R. Grossman, G. McVean, P. J. Turnbaugh, E. S. Lander, M. Mitzenmacher, P. C. Sabeti, Detecting novel associations in large data sets, *Science* 334 (2011) 1518-1524.
- [7] Y. A. Reshef, D. N. Reshef, P. C. Sabeti, M. M. Mitzenmacher, Equitability, interval estimation, and statistical power, *ArXiv pre-print, ArXiv: 1505.02212* (2015) 1-25.
- [8] D. Lopez-Paz, P. Henning, B. Schlkopf, The randomized dependence coefficient, *Advances in neural information processing systems* 26, Curran Associates, Inc. (2013).
- [9] A. Ding, Y. Li, Copula correlation: An equitable dependence measure and extension of Pearson's correlation, *arXiv:1312.7214* (2015).
- [10] Y. Chang, Y. Li, A. Ding, J.A. Dy, Robust-equitable copula dependence measure for feature selection, *Proceedings of the 19<sup>th</sup> International Conference on Artificial Intelligence and Statistics*, 41 (2016) 84-92.
- [11] G.J. Székely, M.L. Rizzo, N.K. Bakirov, Measuring and testing dependence by correlation of distances, *The Annals of Statistics* 35 (6) (2007) 2769-2794.
- [12] G. Marti, S. Andler, F. Nielsen, P. Donnat, Optimal transport vs. Fisher-Rao distance between copulas for clustering multivariate time series, *arXiv:1604.08634v1* (2016).
- [13] R. B. Nelsen, *An introduction to copula*, Springer Verlag, 2<sup>nd</sup> ed., New York (2006).
- [14] A. Sundaresan, P. K. Varshney, Estimation of a random signal source based on correlated sensor observations, *IEEE Transactions on Signal Processing* (2011) 787-799.
- [15] X. Zeng, J. Ren, Z. Wang, S. Marshall, T. Durrani, Copulas for statistical signal processing (Part I): Extensions and generalization, *Signal Processing* 94 (2014) 691-702.
- [16] X. Zeng, J. Ren, Z. Wang, S. Marshall, T. Durrani, Copulas for statistical signal processing (Part I): Simulation, optimal selection and practical applications, *Signal Processing* 94 (2014) 681-690.
- [17] S. G. Iyengar, P. K. Varshney, T. Damarla, A parametric copula-based framework for hypothesis testing using heterogeneous data, *IEEE Transactions on Signal Processing* 59 (5) (2011) 2308-2319.
- [18] X. Zeng, T. S. Durrani, Estimation of mutual information using copula density function, *Electronics Letters* 47 (8) (2011) 493-494.
- [19] A. L. Blum, P. Langley, Selection of relevant features and examples in machine learning, *Artif Intell* 1997; 97:245-70.
- [20] C. Lazar, J. Taminau, S. Meganck, D. Steenhoff, A. Coletta, C. Molter *et al.*, A survey on filter techniques for feature selection in gene expression microarray analysis, *IEEE/ACM Trans Comput Biol Bioinform* 9 (2012).
- [21] E. L. Lehmann, Some concepts of dependence, *The Annals of Mathematical Statistics* 37 (1966) 1137-1153.
- [22] P. Deheuvels, La fonction de dépendance empirique et ses propriétés: Un test non paramétrique d'indépendance, *Bulletin de la Classe des Sciences Académie Royale de Belgique* 65 (1979) 274-292.
- [23] N. Simon, R. Tibshirani, Comment on "Detecting novel associations in large data sets" in [7], *Science* 334 (2011) 1518-1524.
- [24] W. N. Street, W.H. Wolberg, and O.L. Mangasarian, Nuclear feature extraction for breast tumor diagnosis, *Proceedings of the IS&T/SPIE International Symposium on Electronic Imaging: Science and Technology* 1905 (1993) 861-870.
- [25] G. Chandrashekhar, S. Ferat, A survey on feature selection methods, *Computers and Electrical Engineering*, 40 (1) (2014) 16-28.

## APPENDIX

In this appendix, Lemma 1 is stated and proved and then the proofs of Theorems 2, 3, 5, and 6 are provided.

**Lemma 1:** Let  $X$  and  $Y$  be two continuous random variables with copula  $C(F_1(x), F_2(y)) = H(x, y) = P(X \leq x, Y \leq y)$ . Then it follows:

$$a) P(X \leq x, Y > y) = F_1(x) - C(F_1(x), F_2(y)); \quad (9)$$

$$b) P(X > x, Y \leq y) = F_2(y) - C(F_1(x), F_2(y)); \quad (10)$$

$$c) P(X > x, Y > y) = 1 - F_1(x) - F_2(y) + C(F_1(x), F_2(y)). \quad (11)$$

**Proof of Lemma 1:** Partition the domain  $\mathcal{D} = \text{Range}(X) \times \text{Range}(Y)$  of the joint probability distribution function,  $H(x, y)$ , into four subsets, namely

$\mathcal{D}_1 = \{X \leq x, Y \leq y\}$ ,  $\mathcal{D}_2 = \{X \leq x, Y > y\}$ ,  $\mathcal{D}_3 = \{X > x, Y \leq y\}$  and  $\mathcal{D}_4 = \{X > x, Y > y\}$ . Then it follows :  
 $P(X \leq x, Y \leq y) + P(X \leq x, Y > y) + P(X > x, Y \leq y) + P(X > x, Y > y) = 1$ . (12)

It also follows

$$P(X \leq x, Y \leq y) + P(X \leq x, Y > y) = P(X \leq x), \quad (13)$$

which yields (9), and

$$P(X \leq x, Y \leq y) + P(X > x, Y \leq y) = P(Y \leq y), \quad (14)$$

which yields (10). Substituting the expressions of  $P(X \leq x, Y > y)$  given by (9) and of  $P(X > x, Y \leq y)$  given by (10) into (12) produces (11). ■

**Proof of Theorem 2:** a) Under the assumption that  $Y = f(X)$ , suppose that  $(x_i, y_{max})$  is a global maximum of  $f(.)$ . Then, by definition  $C(F_1(x_i), F_2(y_{max})) = P(X \leq x_i, Y \leq y_{max}) = P(X \leq x_i)$ , implying that  $C(u_i, I) = u_i$ . Additionally,  $\text{Min}(u_i, v_i) = \text{Min}(u_i, I) = u_i$  and  $\text{Max}(u_i, v_i) = \text{Max}(u_i + 1 - 1, 0) = u_i$ , from which (1) follows. To prove the converse under the assumption that  $Y = f(X)$ , suppose that there exists a pair  $(u_i, v_i)$  such that  $C(u_i, v_i) = M(u_i, v_i) = W(u_i, v_i) = \square \Pi(u_i, v_i) = u_i$ . It follows that  $v_i = 1$ , which implies that  $C(u_i, I) = u_i$  and  $C(F_1(x_i), F_2(y_{max})) = P(X \leq x_i, Y \leq y_{max})$ , that is,  $(x_i, y_{max})$  is a global maximum of  $f(.)$ .

b) Suppose that  $Y = f(X)$  and  $(x_2, y_{min})$  is a global minimum. Then, by definition  $C(F_1(x_2), F_2(y_{min})) = P(X \leq x_2, Y \leq y_{min}) = 0$ , implying that  $C(u_2, 0) = 0$ . Additionally,  $W(u_2, v_2) = \min(u_2, 0) = 0$ , and  $M(u_2, v_2) = \max(u_2 + 0 - 1, 0) = 0$ , from which (2) follows. To prove the converse under the assumption that  $Y = f(X)$ , suppose that there exists a pair  $(u_2, v_2)$  such that  $C(u_2, v_2) = M(u_2, v_2) = W(u_2, v_2) = \Pi(u_2, v_2) = u_2 v_2 = 0$ . It follows that either  $u_2 = 0$ , or  $v_2 = 0$ , or  $u_2 = v_2 = 0$ . Consider the first case where  $u_2 = 0$ . It follows that  $C(0, v_2) = 0$ , implying that  $C(F_1(x_{2min}), F_2(y_2)) = P(X \leq x_{2min}, Y \leq y_2) = 0$ . This means that  $(x_{2min}, y_2)$  is a global minimum of  $f(.)$ . Consider the second case where  $v_2 = 0$ . It follows that  $C(u_2, 0) = 0$ , implying that  $C(F_1(x_{2min}), F_2(y_2)) = P(X \leq x_2, Y \leq y_{2min}) = 0$ . This means that  $(x_2, y_{2min})$  is a global minimum of  $f(.)$ . Consider the third case where  $u_2 = v_2 = 0$ . It follows that  $C(0, 0) = 0$ , implying that  $C(F_1(x_{2min}), F_2(y_{2min})) = P(X \leq x_{2min}, Y \leq y_{2min}) = 0$ . This means that  $(x_{2min}, y_{2min})$  is a global minimum of  $f(.)$ . ■

**Proof of Theorem 3:** Suppose that  $Y = f(X)$  almost surely, where  $f(.)$  has a single optimum, which is necessarily a global one. Denote by  $S_1$  and  $S_2$  the non-increasing and the non-decreasing line segments of  $f(.)$ , respectively. Note that  $f(.)$  may have inflection points but may not have a line segment of constant value because otherwise  $Y$  will be a mixed random variable, violating the continuity assumption. Let  $A$  denote a point with coordinate  $(x, y)$  of the function  $f(.)$ . Consider the four subsets  $\mathcal{D}_1 = \{X \leq x, Y \leq y\}$ ,  $\mathcal{D}_2 = \{X \leq x, Y > y\}$ ,  $\mathcal{D}_3 = \{X > x, Y \leq y\}$  and  $\mathcal{D}_4 = \{X > x, Y > y\}$ . Suppose that  $A$  is a point of  $S_1$ . As shown in Fig. 10(a), either  $\mathcal{D}_1 \cap S_1 = \{A\}$  or  $\mathcal{D}_4 \cap S_1 = \emptyset$  depending upon whether  $f(.)$  has a global minimum or a global maximum point, respectively. In the former case,  $P(X \leq x, Y \leq y) = 0$ , implying from (9) that  $C(u, v) = u + v - 1 \geq 0$ . Combining both cases, it follows that for all  $(x, y) \in S_1$ ,  $C(u, v) = \text{Max}(u + v - 1, 0)$ .

Now, suppose that  $A$  is a point of  $S_2$ . As shown in Fig. 10(b), either  $\mathcal{D}_2 \cap S_2 = \{A\}$  or  $\mathcal{D}_3 \cap S_2 = \emptyset$  depending upon whether  $f(.)$  has a global maximum or a global minimum point, respectively. In the former case,  $P(X \leq x, Y > y) = 0$ , implying from (7) that  $C(u, v) = u$  while in the latter case,  $P(X > x, Y \leq y) = 0$ , implying from (8) that  $C(u, v) = v$ . Combining both cases, it follows from (3) that for all  $(x, y) \in S_2$ ,  $C(u, v) = \min(u, v)$ . ■

**Proof of Theorem 5:** Suppose that  $Y = f(X)$  almost surely, where  $f(.)$  has at least two global maxima and no local optima. As depicted in Fig. 11(a), let  $B$  and  $C$  be two global maximum points of  $f(.)$  with coordinates  $(x_B, y_{max})$  and  $(x_C, y_{max})$ , respectively. This means that there exists  $\Delta x > 0$  such that  $f(x_B + \Delta x) < y_{max}$  and  $f(x_C + \Delta x) < y_{max}$ . Consider a point  $A$  with coordinate  $(x_A, y_A)$  such that  $x_B < x_A < x_B + \Delta x$ ,  $f(x_B - \Delta x) < y_A < y_{max}$  and  $f(x_C - \Delta x) < y_A < y_{max}$ . Denote by  $S_B$  and  $S_C$  the line segments of  $f(.)$  defined over the intervals  $[f(x_B - \Delta x), y_{max}]$  and  $[f(x_C - \Delta x), y_{max}]$ , respectively, which are shown as solid lines in Fig. 11(a). Partition the domain  $\mathcal{D}$  into four subsets,  $\mathcal{D}_1 = \{X \leq x_A, Y \leq y_A\}$ ,  $\mathcal{D}_2 = \{X \leq x_A, Y > y_A\}$ ,  $\mathcal{D}_3 = \{X > x_A, Y \leq y_A\}$  and  $\mathcal{D}_4 = \{X > x_A, Y > y_A\}$ . As observed in Fig. 12(a),  $\mathcal{D}_1 \cap S_B \setminus \{A\} \neq \emptyset$ ,  $\mathcal{D}_2 \cap S_B \neq \emptyset$ ,  $\mathcal{D}_3 \cap S_C \neq \emptyset$ , and  $\mathcal{D}_4 \cap S_C \neq \emptyset$ , yielding  $\square \square C(u, v) < 1$ . A similar proof can be developed for the case where  $f(.)$  has at least two global minima and no local optima. ■

**Proof of Theorem 6:** Suppose that  $Y = f(X)$  almost surely, where  $f(\cdot)$  has a local minimum point, say point A of coordinates  $(x_A, y_A)$  as shown in Fig. 11(b). This means that there exists  $\Delta x > 0$  such that  $f(x_A \pm \Delta x) > y_A$ . As depicted in Fig. 11(b), let  $S_{A1}$  and  $S_{A2}$  denote the line segments of  $f(\cdot)$  defined over  $x_A - \Delta x$  and  $x_A + \Delta x$ , respectively. Consider the four domains,  $\mathfrak{D}_1 = \{X \leq x_A, Y \leq y_A\}$ ,  $\mathfrak{D}_2 = \{X \leq x_A, Y > y_A\}$ ,  $\mathfrak{D}_3 = \{X > x_A, Y \leq y_A\}$  and  $\mathfrak{D}_4 = \{X > x_A, Y > y_A\}$ . As observed in Fig. 11(b),  $\mathfrak{D}_2 \cap S_{A1} \neq \emptyset$  and  $\mathfrak{D}_4 \cap S_{A2} \neq \emptyset$ . Now, because A is by hypothesis a local minimum point, there exist line segments of  $f(\cdot)$  denoted by  $S$  such that  $f(y) < y_A$ . Consequently, one of the following three cases arises: either  $\mathfrak{D}_1 \cap S \setminus \{A\} \neq \emptyset$  and  $\mathfrak{D}_3 \cap S \neq \emptyset$  as depicted in Fig. 11(b), or  $\mathfrak{D}_1 \cap S \setminus \{A\} \neq \emptyset$  and  $\mathfrak{D}_3 \cap S = \emptyset$ , or  $\mathfrak{D}_1 \cap S \setminus \{A\} = \emptyset$  and  $\mathfrak{D}_3 \cap S \neq \emptyset$ . In the first case,  $\square C(u,v) < 1$  while in the last two cases,  $\square C(u,v) = 1$ . A similar proof can be developed for  $f(\cdot)$  with a local maximum point. ■

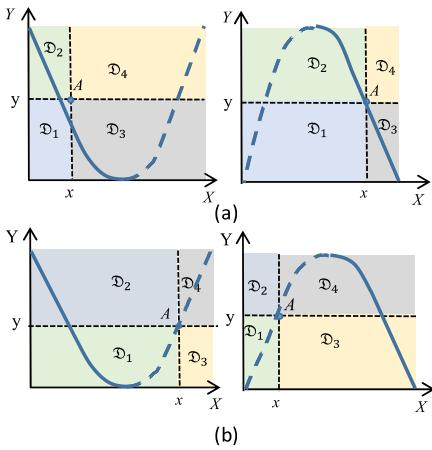


Fig. 10. Graphs of a function  $Y = f(X)$  having a single optimum. A point A with coordinate  $(x,y)$  is located either on the non-increasing part,  $S_1$ , shown as a solid line in (a) or on the non-decreasing part,  $S_2$ , shown as a dashed line in (b) of the function  $f(\cdot)$ . Four domains,  $\mathfrak{D}_1, \dots, \mathfrak{D}_4$ , are delineated by the vertical and horizontal lines at position  $X = x$  and  $Y = y$ , respectively.

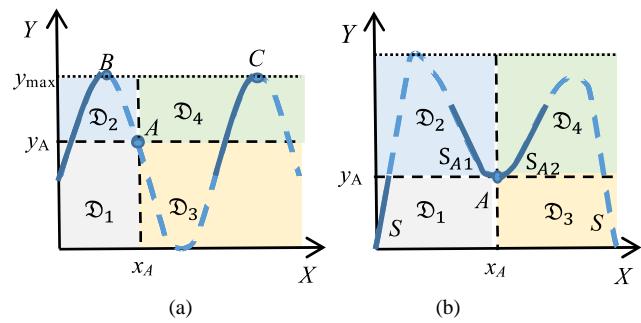


Fig. 11. (a) The graph of a function  $Y = f(X)$  having two global maximum points denoted by B and C, and one global minimum point, with two solid line segments denoted by  $S_B$  and  $S_C$ . (b) The graph of a function  $Y = f(X)$  having one local minimum point denoted by A, with line segments denoted by  $S_{A1}$ ,  $S_{A2}$ , and  $S$ . Four domains,  $\mathfrak{D}_1, \dots, \mathfrak{D}_4$ , are delineated by the vertical and horizontal lines at position  $X = x_A$  and  $Y = y_A$ , respectively.

#### AUTHOR PROFILE

**Mohsen Ben Hassine** received an engineering diploma and the M.S. degree in computer sciences from the École Nationale des Sciences de l'Informatique, Tunis, Tunisia, in 1993 and 1996, respectively. He is currently a graduate teaching assistant at the University of El Manar, Tunis, Tunisia. His research interests include statistical signal processing, mathematical simulations, and statistical bioinformatics.

**Lamine Mili** received an electrical engineering diploma from the Swiss Federal Institute of Technology, Lausanne, in 1976, and a Ph.D. degree from the University of Liege, Belgium, in 1987. He is presently a Professor of Electrical and Computer Engineering at Virginia Tech. His research interests include robust statistics, robust statistical signal processing, radar systems, and power system analysis and control. Dr. Mili is a Fellow of IEEE for contribution to robust state estimation for power systems.

**Kiran Karra** received a B.S. in Electrical and Computer Engineering and an M.S. degree in Electrical Engineering from North Carolina State University and Virginia Polytechnic Institute and State University, in 2007 and 2012, respectively. He is currently a research associate at the Virginia Tech and is studying statistical signal processing and machine learning for his PhD research.