Kybernetika

José M. González-Barrios; María M. Hernández-Cedillo

Sample d-copula of order m

Kybernetika, Vol. 49 (2013), No. 5, 663--691

Persistent URL: http://dml.cz/dmlcz/143517

SAMPLE *D*-COPULA OF ORDER *M*

José M. González-Barrios and María M. Hernández-Cedillo

In this paper we analyze the construction of d-copulas including the ideas of Cuculescu and Theodorescu [5], Fredricks et al. [15], Mikusiński and Taylor [25] and Trutschnig and Fernández-Sánchez [33]. Some of these methods use iterative procedures to construct copulas with fractal supports.

The main part of this paper is given in Section 3, where we introduce the sample d-copula of order m with $m \geq 2$, the central idea is to use the above methodologies to construct a new copula based on a sample. The greatest advantage of the sample d-copula is the fact that it is already an approximating d-copula and that it is easily obtained. We will see that these new copulas provide a nice way to study multivariate data with an approximating copula which is simpler than the empirical multivariate copula, and that the empirical copula is the restriction to a grid of a sample d-copula of order n. These sample d-copulas can be used to make statistical inference about the distribution of the data, as shown in Section 3.

Keywords: d-copulas, fractal copulas, sample d-copulas of order m

Classification: 60A10, 60E05, 62E10

1. INTRODUCTION

The construction of multivariate families of copulas for $d > 2$ is of great interest, because they are used in modeling multivariate data in several fields such as Economics, Biology, Hydrology, etc. The problem is that there are only a few known families that are used in practice.

In Cuculescu and Theodorescu [5], they introduce a new family of copulas which they call self-similar copulas, in dimension two, using an iterated procedure. These ideas were substantially improved in Fredricks et al. [15] in dimension $d = 2$, and quite recently in Trutschnig and Fernández-Sánchez [33] these results are generalized to $d \geq 3$.

In this paper we will follow the original ideas given in Cuculescu and Theodorescu [5], giving generalizations to larger dimensions $d \geq 3$. The main advantage of constructing the d-copulas with their ideas is the fact that in every step we already obtain a d-copula, which allows to approximate any given d-copula. These d-copulas correspond to the checkerboard d-copulas defined in Mikusiński and Taylor [25].

Recall that a d-copula is a function $C : [0, 1]^d \to [0, 1]$ for some integer $d \geq 2$ which satisfies:

i) $C(u_1, \ldots, u_d) = 0$ if there exists at least one $i \in \{1, \ldots, d\}$ such that $u_i = 0$.

ii) $C(1, \ldots, 1, u_i, 1, \ldots, 1) = u_i$ for every $i \in \{1, \ldots, d\}$ and for every $u_i \in [0, 1]$.

iii) C is a d-increasing function, that is, for any d-box $R = \Pi_{i=1}^{d}[u_i, v_i]$ such that $R \subset [0,1]^d$ we have that

$$V_C(R) := \sum_{\{\underline{\mathbf{c}}\in[0,1]^d \mid \underline{\mathbf{c}}\in \mathrm{Vert}(R)\}} \mathrm{sgn}(\underline{\mathbf{c}})C(\underline{\mathbf{c}}) \geq 0, \tag{1}$$

where

$$\mathrm{sgn}(\underline{\mathbf{c}}) = \begin{cases} 1, & \text{if } c_i = u_i \text{ for an even number of } i's \\ -1, & \text{if } c_i = u_i \text{ for an odd number of } i's. \end{cases}$$

Therefore, C is a d-copula if and only if C is the restriction to $[0,1]^d$ of a distribution function of a d-dimensional random vector $\underline{\mathbf{U}} = \langle U_1, \dots, U_d\rangle$ with standard uniform $U(0,1)$ marginal distribution functions.

In particular, if $d = 2$ it is common to say that C is a copula instead of a 2-copula. We will state a very well known result when $d = 2$, see for example Nelsen [26]. If we define

$$W(u,v) = \max\{u+v-1, 0\} \quad \text{and} \quad M(u,v) = \min\{u,v\} \quad \text{for every} \quad \langle u,v\rangle \in [0,1]^2.$$

Then W and M are copulas called the Fréchet–Hoeffding lower and upper bounds respectively, that satisfy:

$$W(u,v) \leq C(u,v) \leq M(u,v) \qquad \text{for every} \quad \langle u,v\rangle \in [0,1]^2.$$

For $d > 2$ we also know that if we define on $[0,1]^d$, $W^d(u_1,\dots,u_d) = \max\{u_1 + \cdots + u_d - d + 1, 0\}$ and $M^d(u_1,\dots,u_d) = \min\{u_1,u_2,\dots,u_d\}$. Then for every d-copula C,

$$W^d(u_1,\dots,u_d) \leq C(u_1,\dots,u_d) \leq M^d(u_1,\dots,u_d) \qquad \text{for every} \quad \langle u_1,\dots,u_d\rangle \in [0,1]^d.$$

But in this case, W^d is not a d-copula, even though M^d is always a d-copula. However, the inequality above is sharp for every $\langle u_1,\dots,u_d\rangle \in [0,1]^d = I^d$, in the sense that there exists always a d-copula C such that the left equality holds.

An important result relating a continuous d-distribution function H and its marginals is:

Sklar's Theorem. Let H be a continuous d-distribution function with margins $F_1, F_2, \dots, F_d$. Then there exists a unique d-copula C such that for every $\underline{\mathbf{x}} = \langle x_1, x_2, \dots, x_d\rangle \in \overline{\mathbb{R}}^d$

$$H(x_1,x_2,\dots,x_d) = C(F_1(x_1), F_2(x_2), \dots F_d(x_d)).$$

See for example [26, 32], or [11].

In the second section we start by stating the main results in Fredericks et al. [15] and we relate their concept of transformation matrices to doubly stochastic matrices. Then we generalize the results given in Cuculescu and Theodorescu [5], to any dimension $d > 2$. We also mention that the family of fractal d-copulas is dense in the family of all d-copulas for any $d \geq 2$. Finally we analyze the multivariate extension of Fredricks et al. [15] given in Trutschnig and Fernández-Sánchez [33].

In the third section we will use the results in Fredricks et al. [15] and Trutschnig and Fernández-Sánchez [33] to introduce the sample d-copula of order m for $m \geq 2$ based on a sample of size $n \geq m$, which is very easy to calculate. We find some of its basic properties and possible applications in Statistics. We will see that the sample d-copula of order m is a strongly consistent estimator of the d-copula which generated the data. We will also propose some possible statistical applications of the sample d-copula.

The third section also studies some of the basic probability properties of the sample d-copula of order m including its close relation to the multivariate distribution with parameters n the sample size, and v positive parameters whose sum is one and $m \leq v < m^d$. We will also include new applications of the sample d-copula in Statistics.

In the last section we include some important remarks.

2. D-COPULAS WITH FRACTAL SUPPORTS

In Fredricks et al. [15] using techniques of iterated function systems (IFS) the authors construct for dimension $d = 2$ a large class of copulas. They first consider a **transformation matrix**, that is a real nonnegative matrix $T_{n\times m} = (t_{ij})_{\langle i,j\rangle\in I_n\times I_m}$,where $I_n = \{1, \ldots, n\}$, such that $\max\{n, m\} \geq 2$, $\sum_{i,j} t_{ij} = 1$, $\sum_{i\in I_n} t_{ij} > 0$ for every $j \in I_m$ and $\sum_{j\in I_m} t_{ij} > 0$ for every $i \in I_n$. Define two partitions of $[0, 1]$, $\{p_0, p_1, \ldots, p_n\}$ and $\{q_0, q_1, \ldots, q_m\}$, by letting $p_0 = 0 = q_0$, and for $i \in I_n$ let $p_i = \sum_{i'=1}^{i} \sum_{j\in I_m} t_{i'j}$, and for $j \in I_m$ let $q_j = \sum_{j'=1}^{j} \sum_{i\in I_n} t_{ij'}$. Define

$$R_{ij} = (p_{i-1}, p_i] \times (q_{j-1}, q_j] \quad \text{for every} \quad \langle i, j\rangle \in I_n \times I_m,$$

where if $i = 1$ or $j = 1$ we take closed intervals instead of right open intervals. Of course, $\{R_{ij}\}_{\langle i,j\rangle\in I_n\times I_m}$ is a partition of I^2. Let C be a copula and define a transformation $T(C)$ using the partition of I^2 and the transformation matrix T, where for each $\langle i, j\rangle \in I_n \times I_m$, $T(C)$ spreads mass t_{ij} on R_{ij} rescaling the whole mass of C, that is, if $\langle u, v\rangle \in R_{ij}$ let

$$T(C)(u, v) = \sum_{i'<i,j'<j} t_{i'j'} + \frac{u - p_{i-1}}{p_i - p_{i-1}} \sum_{j'<j} t_{ij'} + \frac{v - q_{j-1}}{q_j - q_{j-1}} \sum_{i'<i} t_{i'j} + t_{ij} C\left(\frac{u - p_{i-1}}{p_i - p_{i-1}}, \frac{v - q_{j-1}}{q_j - q_{j-1}}\right), \quad (2)$$

where empty sums are defined to be zero. Then $T(C)$ is always a copula. If we define iteratively

$$T^2(C) = T(T(C)) \quad \text{and} \quad T^{n+1}(C) = T(T^n(C)) \quad \text{for every} \quad n > 2.$$

In fact, $T^n(C) = (\otimes^n T)(C)$, where $\otimes^n$ is the tensor product of T with itself n times. It is easy to see by induction that if T is a transformation matrix of order $n \times m$ then $\otimes^k T$ is also a transformation matrix of order $n^k \times m^k$ for every $k \geq 2$. Then we have that for any transformation matrix T there exists a unique copula copula C_T, such that $T(C_T) = C_T$. Moreover, $C_T = \lim_{n\to\infty} T^n(C)$ for any copula C. Since C_T does not depend on the copula C, we may restrict to the limit of the sequence $\{T^n(\Pi)\}_{n\geq 1}$. In fact, they call C *invariant* if $C = C_T$ for some transformation matrix T.

They also observe that if $\pi_1 = \{p_0, p_1, \ldots, p_n\}$ and $\pi_2 = \{q_0, q_1, \ldots, q_m\}$ are any partitions of $[0, 1]$, and we define $t_{ij} = (p_i - p_{i-1})(q_j - q_{j-1})$ for every $\langle i, j\rangle \in I_n \times I_m$, then $T = (t_{ij})_{\langle i,j\rangle\in I_n\times I_m}$ is a transformation matrix which generates the partitions π_1 and π_2 and has $C_T = \Pi^2$ the product copula.

Recall that for every $k \geq 2$ a square real matrix $\mathbf{P} = (p_{ij})_{i,j=1}^k$ is a **doubly stochastic matrix** if and only if $p_{ij} \geq 0$ and $\sum_{j=1}^k p_{ij} = \sum_{i=1}^k p_{ij} = 1$ for every $i, j \in \{1, 2, \ldots, k\}$.

Define

$$\mathcal{T} = \{T_{n\times m} \mid T_{n\times m} \text{ is a transformation matrix, with } n, m \geq 2 \text{ and } t_{ij} \in \mathbb{Q} \text{ for every } \langle i, j\rangle \in I_n\times I_m\}.$$

Then we have the following result

Lemma 2.1. Let $T_{n\times m} \in \mathcal{T}$ then there exist $k \geq 2$ and $P_{k\times k} = (p_{ij})_{i,j=1}^{k}$ a double stochastic matrix, such that if we define $S_{k\times k} = (p_{ij}/k)_{i,j=1}^{k}$ then $T(\Pi^2) = S(\Pi^2)$.

The proof follows directly by considering k the least common multiple of the denominators of $(t_{ij})_{\langle i,j\rangle\in I_n\times I_m}$.

The last lemma can be used in every single step of the construction of C_T. However, it is clear that the support of the limit copula C_T may be different from C_S. In fact, in Section 3 we will use only square matrices T with rational entries, and the first step in the construction of C_T. So, we can think of T as a doubly stochastic matrix times a positive integer. For some results on doubly stochastic matrices see for example Sherman [30] or Marcus [23].

Example 2.2. As an easy example of Lemma 2.1 consider the transformation matrix

$$T = \begin{pmatrix} 0 & 1/3 \\ 2/3 & 0 \end{pmatrix}.$$

Then if we define

$$S = \begin{pmatrix} 0 & 0 & 1/3 \\ 1/6 & 1/6 & 0 \\ 1/6 & 1/6 & 0 \end{pmatrix}.$$

We have that $P = 3 \cdot S$ is a doubly stochastic matrix and $T(\Pi^2) = S(\Pi^2)$.

Now we generalize the results given in Cuculescu and Theodorescu [5].

Recall that a d-**dimensional square matrix P**, for $d \geq 2$, is an array of real numbers of the form $\mathbf{P} = (p_{i_1 i_2\cdots i_d})_{i_1,\ldots i_d=1}^{k}$ for some $k \geq 2$. We will say that **P** is d-**dimensionally stochastic** if and only if $0 \leq p_{i_1 i_2\cdots i_d} \leq 1$ for every $i_1, i_2, \ldots, i_d \in \{1, \ldots, k\}$, and for every $1 \leq j_1 < j_2 < \cdots < j_{d-1} \leq d$ we have that

$$\sum_{i_{j_1}=1}^{k} \sum_{i_{j_2}=1}^{k} \cdots \sum_{i_{j_{d-1}}=1}^{k} p_{i_1 i_2\cdots i_d} = 1, \tag{3}$$

where the remaining index is fixed and taken in $\{1, \ldots, k\}$. In this case it is clear that

$$\sum_{i_1=1}^{k} \sum_{i_2=1}^{k} \cdots \sum_{i_d=1}^{k} p_{i_1 i_2\cdots i_d} = k.$$

Of course a 2-dimensionally stochastic matrix is a doubly stochastic matrix. Let $C^1_{i_1,i_2,\ldots,i_d} = p_{i_1 i_2\cdots i_d}/k$ for every $i_1, \ldots, i_d \in \{1, \ldots k\}$, and for every $A \in \mathcal{B}([0,1]^d)$ and for every $n \geq 1$ define

$$\mu_n(A) = k^{dn} \sum_{i_1,\ldots,i_d=1}^{k^n} C^n_{i_1,\ldots,i_d}\, \lambda^d \left(\left(\left[\frac{i_1 - 1}{k^n}, \frac{i_1}{k^n} \right] \times \cdots \times \left[\frac{i_d - 1}{k^n}, \frac{i_d}{k^n} \right] \right) \cap A \right), \tag{4}$$

where for every $i_1, \ldots i_d \in \{1, \ldots, k^n\}$, for every $i'_1, \ldots i'_d \in \{1, \ldots, k\}$ and for every $n \geq 1$,

$$C^{n+1}_{k(i_1-1)+i'_1,\ldots,k(i_d-1)+i'_d} = \frac{p_{i'_1\cdots i'_d}}{k^{d-1}} \cdot C^n_{i_1,\ldots,i_d}. \tag{5}$$

Here $\mathcal{B}([0,1]^d)$ is the Borel σ-algebra and λ^d is the Lebesgue measure. Then we have a multivariate extension of Cuculescu and Theodorescu [5], its proof follows as in Trutschnig and Fernández-Sánchez [33].

Proposition 2.3. Let $d \geq 2$, let $\mathbf{P}$ be a d-dimensional square matrix of order $k \geq 2$, which is d-dimensionally stochastic. Let $n \geq 1$ and define μ_n as in equation (4), then $([0,1]^d, \mathcal{B}([0,1]^d), \mu_n)$ is a probability space. Besides, if we define

$$C_n(u_1, \dots, u_d) = \mu_n([0,u_1] \times \cdots \times [0,u_d]) \quad \text{for every} \quad u_1, \dots u_d \in [0,1]. \tag{6}$$

Then C_n is a d-copula for every $n \geq 1$. If we define $\mu_{\mathbf{P}} = \lim_{n\to\infty} \mu_n$, then $\mu_{\mathbf{P}}$ exists with respect to weak convergence and it is a probability measure on $([0,1]^d, \mathcal{B}([0,1]^d))$. Evenmore, $\mu_{\mathbf{P}}$ induces a d-copula $C_{\mathbf{P}}$ by defining

$$C_{\mathbf{P}}(u_1, \dots, u_d) = \mu_{\mathbf{P}}([0,u_1] \times \cdots \times [0,u_d]) \quad \text{for every} \quad u_1, \dots u_d \in [0,1]. \tag{7}$$

If $\mathbf{P}$ includes zeros then $\mu_{\mathbf{P}}$ is a singular measure.

In the case $d = 2$ with $0 < a < 1$

$$\mathbf{P} = \begin{pmatrix} a/2 & (1-a)/2 \\ (1-a)/2 & a/2 \end{pmatrix}.$$

Then $2\mathbf{P}$ is doubly stochastic and $C_{\mathbf{P}}$ is a singular copula if $a \neq 1/2$ as observed in Cuculescu and Theodorescu [5], see also [12].

Now we will see that for $d \geq 2$ the set of d-copulas given in (7) is dense in the family of all copulas with respect to the supremum distance, when we consider the set of d-dimensionally stochastic matrices.

Theorem 2.4. Let C be a d-copula for some $d \geq 2$, then for every $\epsilon > 0$ there exists $\mathbf{P}$ a d-dimensionally stochastic matrix such that if we construct the copula $C_{\mathbf{P}}$ defined in equation (7)

$$d_{\sup}(C, C_{\mathbf{P}}) = \sup_{u_1,\dots,u_d \in [0,1]} |C(u_1, \dots, u_d) - C_{\mathbf{P}}(u_1, \dots, u_d)| < \epsilon. \tag{8}$$

The proof of this theorem follows from Mikusiński and Taylor [25], since in each step of the construction of $C_{\mathbf{P}}$ we obtain a checkerboard approximation. Evenmore, they prove that the convergence of the checkerboard approximations to the d-copula C holds in a stronger mode denoted by ∂-convergence which implies uniform convergence.

In Cuculescu and Theodorescu [5], for dimensions greater than or equal to three they only say “For $q \geq 2$ copulas analogous to $\mu_{\mathbf{P}}$ may also be defined (particularly one concentrated on Menger’s sponge)…”. Here q is the dimension. This statement is not correct as can be seen in Hernández-Cedillo [17]. In [33] the authors provide an example of a 3-copula which has a *Menger’s sponge like set* support.

Finally, we give the generalization of transformation matrices in dimension d found in Trutschnig and Fernández-Sánchez [33]. Let $I_n = \{1, 2, \dots, n\}$ for $n \geq 1$. For $d \geq 2$, let $m_1, \dots, m_d \in \mathbb{N}$ and define $\mathcal{I}^d = \Pi_{i=1}^d I_{m_i}$. Let τ be a probability measure on $(\mathcal{I}^d, 2^{\mathcal{I}^d})$, then we call τ a **generalized transformation matrix** if for every $j \in \{1, \dots, d\}$ and for every $k \in \{1, \dots, m_j\}$

$$\sum_{\underline{\mathbf{i}} \in \mathcal{I}^d, i_j = k} \tau(\underline{\mathbf{i}}) > 0, \tag{9}$$

where $\underline{\mathbf{i}} = \langle i_1, \dots, i_{j-1}, i_j = k, i_{j+1}, \dots, i_d \rangle \in \mathcal{I}^d$.

Observe that equation (9) is a natural extension of the conditions of transformation matrices in Fredricks et al. [15], and that τ can be written as a d-dimensional matrix T, by writing

$$\tau(\underline{\mathbf{i}}) = t_{i_1, i_2, \dots, i_d} \quad \text{if} \quad \underline{\mathbf{i}} = \langle i_1, i_2, \dots, i_d \rangle \in \mathcal{I}^d. \tag{10}$$

In the remaining of this section we will only use the case $m_1 = m_2 = \cdots = m_d = m \geq 2$. In this case, if all $\tau(\underline{\mathbf{i}})$ are rationals, we observe that equation (9) is equivalent to saying that τ induces the existence of an $m_0 \geq m$ and a d-dimensional square matrix T_0 such that $m_0 \cdot T_0$ is a d-dimensionally stochastic matrix, as defined after Example 2.2. This is a consequence of an obvious multivariate extension of Lemma 2.1 for any τ probability measure with rational values on $(\mathcal{I}^d, 2^{\mathcal{I}^d})$ in equation (10).

All the results at the beginning of this section about the construction of copulas in Fredericks et al. [15], can be easily generalized to dimensions $d \geq 3$.

Let $m \geq 2$ and for every $\underline{\mathbf{i}} = \langle i_1, i_2, \dots, i_d \rangle \in \mathcal{I}_m := \Pi_{j=1}^d I_m$ define

$$R_{\underline{\mathbf{i}}} = \left(\frac{i_1 - 1}{m}, \frac{i_1}{m} \right] \times \left(\frac{i_2 - 1}{m}, \frac{i_2}{m} \right] \times \cdots \times \left(\frac{i_d - 1}{m}, \frac{i_d}{m} \right], \tag{11}$$

where if for some $j \in \{1, \dots, d\}$, $i_j = 1$, then we take closed intervals instead of left open intervals. Then $\{R_{\underline{\mathbf{i}}}\}_{\underline{\mathbf{i}} \in \mathcal{I}_m}$ is a partition of $[0,1]^d$ that we will call the **uniform partition of** $[0,1]^d$.

Let C be a d-copula and define for every $\underline{\mathbf{i}} = \langle i_1, i_2, \dots, i_d \rangle \in \mathcal{I}_m$

$$t_{i_1, i_2, \dots, i_d} = V_C(R_{\underline{\mathbf{i}}}) \quad \text{and} \quad T^C = (t_{i_1, i_2, \dots, i_d})_{i_1, \dots, i_d = 1}^m . \tag{12}$$

Then T^C is a square d-dimensional matrix with nonnegative entries, which generates the checkerboard approximation given in [25] and it is a generalized transformation matrix, because by equation (10), if we take any $j \in \{1, \dots, d\}$ and any $k \in \{1, \dots, m\}$ then by the definition of d-copula

$$\begin{aligned}
\sum_{\underline{\mathbf{i}} \in \mathcal{I}_m, i_j = k} \tau(\underline{\mathbf{i}}) &= \sum_{i_1=1}^m \cdots \sum_{i_{j-1}=1}^m \sum_{i_{j+1}=1}^m \cdots \sum_{i_d=1}^m V_C(R_{\langle i_1, \dots, i_{j-1}, k, i_{j+1}, \dots, i_d \rangle}) \\
&= V_C\left([0,1] \times \cdots \times [0,1] \times \left[\frac{k-1}{m}, \frac{k}{m} \right] \times [0,1] \cdots \times [0,1] \right) \\
&= C(1, \dots, 1, k/m, 1, \dots, 1) - C(1, \dots, 1, (k-1)/m, 1, \dots, 1) \\
&= \frac{1}{m} > 0.
\end{aligned} \tag{13}$$

Observe that in equation (13) for every d-copula C, for every $j \in \{1, \dots, d\}$ and for every $k \in \{1, \dots, m\}$, $\sum_{\underline{\mathbf{i}} \in \mathcal{I}_m, i_j = k} \tau(\underline{\mathbf{i}}) = 1/m$ only depends on m.

Also observe that $m \cdot T^C$ is a d-dimensionally stochastic square matrix.

Now, if we have $T = (t_{i_1, \dots, i_d})_{i_1, \dots, i_d = 1}^m$ a generalized transformation square d-dimensional matrix, define $p_{1,0} = p_{2,0} = \cdots = p_{d,0} = 0$, and for every $j \in \{1, \dots, d\}$ and for every $k \in \{1, \dots, m\}$ define

$$p_{j,k} = \sum_{i_j=1}^k \sum_{i_1=1}^m \cdots \sum_{i_{j-1}=1}^m \sum_{i_{j+1}=1}^m \cdots \sum_{i_d=1}^m t_{i_1, \dots, i_d}. \tag{14}$$

Then $0 = p_{j,0} < p_{j,1} < \cdots < p_{j,m-1} < p_{j,m} = 1$ is the partition with $m + 1$ points of $[0, 1]$ induced by T, which corresponds to the j^{th} coordinate.

If we take a d-copula C and we define $T(C)(u_1, \ldots, u_d)$ for $\langle u_1, \ldots, u_d\rangle \in [0, 1]^d$ using a similar formula as the one used in dimension 2 in equation (2), then $T(C)$ is always a d-copula, see also Trutschnig and Fernández-Sánchez [33]. In particular if $C = \Pi^d$ the product d-copula $T(\Pi)(u_1, \ldots, u_d)$ has a simpler expression. For example if $d = 3$ and $\langle u_1, u_2, u_3\rangle \in R_{\langle i_1,i_2,i_3\rangle} = R_{\underline{\mathbf{i}}}$ $= (p_{1,i_1-1}, p_{1,i_1}] \times (p_{2,i_2-1}, p_{2,i_2}] \times (p_{3,i_3-1}, p_{3,i_3}]$ for some $\underline{\mathbf{i}} \in \mathcal{I}_m$ then

$$\begin{aligned}
T(\Pi)(u_1, u_2, u_3) \quad = \quad & \sum_{i<i_1,j<i_2,k<i_3} t_{i,j,k} + \left(\frac{u_1 - p_{1,i_1-1}}{p_{1,i_1} - p_{1,i_1-1}}\right) \sum_{j<i_2,k<i_3} t_{i_1,j,k} + \left(\frac{u_2 - p_{2,i_2-1}}{p_{2,i_2} - p_{2,i_2-1}}\right) \sum_{i<i_1,k<i_3} t_{i,i_2,k} \\
& + \left(\frac{u_3 - p_{3,i_3-1}}{p_{3,i_3} - p_{3,i_3-1}}\right) \sum_{i<i_1,j<i_2} t_{i,j,i_3} + \left(\frac{u_1 - p_{1,i_1-1}}{p_{1,i_1} - p_{1,i_1-1}}\right)\left(\frac{u_2 - p_{2,i_2-1}}{p_{2,i_2} - p_{2,i_2-1}}\right) \sum_{k<i_3} t_{i_1,i_2,k} \\
& + \left(\frac{u_1 - p_{1,i_1-1}}{p_{1,i_1} - p_{1,i_1-1}}\right)\left(\frac{u_3 - p_{3,i_3-1}}{p_{3,i_3} - p_{3,i_3-1}}\right) \sum_{j<i_2} t_{i_1,j,i_3} \\
& + \left(\frac{u_2 - p_{2,i_2-1}}{p_{2,i_2} - p_{2,i_2-1}}\right)\left(\frac{u_3 - p_{3,i_3-1}}{p_{3,i_3} - p_{3,i_3-1}}\right) \sum_{i<i_1} t_{i,i_2,i_3} \\
& + t_{i_1,i_2,i_3} \left(\frac{u_1 - p_{1,i_1-1}}{p_{1,i_1} - p_{1,i_1-1}}\right)\left(\frac{u_2 - p_{2,i_2-1}}{p_{2,i_2} - p_{2,i_2-1}}\right)\left(\frac{u_3 - p_{3,i_3-1}}{p_{3,i_3} - p_{3,i_3-1}}\right).
\end{aligned} \tag{15}$$

Observe that from equation (15) it is clear that $T(\Pi^3)$ is a 3-copula which assigns uniform mass t_{i_1,i_2,i_3} to each box $R_{\langle i_1,i_2,i_3\rangle}$ for every $\langle i_1, i_2, i_3\rangle \in \{1, \ldots, m\}$. Therefore, the generalized transformation matrix T can be thought as the weighted density of the 3-copula $T(\Pi^3)$, given by $t_{\underline{\mathbf{i}}}/\lambda^d(R_{\underline{\mathbf{i}}})$ for every $\underline{\mathbf{i}} \in \mathcal{I}_m$, induced by the partitions and the 3-boxes that they generate. Of course the d-dimensional case includes 2^d terms that can be easily generalized.

Even if equation (15) seems quite complicated, it is easy to program in a computer when we have the 3-dimensional generalized transformation square matrix T of order m. We have written a short program in language **R** which computes $T(\Pi)(u_1, u_2, u_3)$ for any given vector $\langle u_1, u_2, u_3\rangle \in [0, 1]^3$.

3. SAMPLE *D*-COPULA OF ORDER *M*

Now we use the ideas of Section 2 to define the **sample *d*-copula of order** *m* in two settings.

3.1. Sample *d*-Copula of Order *m* for a *d*-Copula *C*

Let $m \geq 2$ and assume that we take an independent sample of size n, where $n \geq m$, from a d-copula C, let us denote the sample by

$$U_n = \{\underline{\mathbf{x_1}}, \cdots, \underline{\mathbf{x_n}}\}, \tag{16}$$

where $\underline{\mathbf{x_k}} = \langle x_{k,1} \ldots x_{k,d}\rangle \in [0, 1]^d$ for every $k \in \{1, \ldots, n\}$.

Define for every $\underline{\mathbf{i}} = \langle i_1, \ldots, i_d\rangle \in \mathcal{I}_m$ using equation (11)

$$s^n_{i_1,\ldots,i_d} = \frac{|R_{\underline{\mathbf{i}}} \cap U_n|}{n}, \tag{17}$$

where $|\cdot|$ denotes the cardinality of a set. Define

$$S^n_m = \left(s^n_{i_1,\dots,i_d}\right)^m_{i_1,\dots,i_d=1}. \tag{18}$$

Then it is clear that S^n_m is a square d-dimensional matrix such that

$$\sum_{i_1,\dots,i_d=1}^{m} s^n_{i_1,\dots,i_d} = 1. \tag{19}$$

Define

$$\mathcal{S}^+ = \{S^n_m \,|\, S^n_m \text{ is a generalized transformation matrix}\}. \tag{20}$$

If we assume that $S^n_m \in \mathcal{S}^+$ then define for every $j \in \{1,\dots,d\}$ the partitions of $[0,1]$, $\pi^n_j := \{p^n_{j,0},\dots,p^n_{j,m}\}$ given in equation (14), and define the **sample d-copula of order** m, denoted by C^n_m by

$$C^n_m(u_1,\dots,u_d) = \begin{cases} S^n_m(\Pi)(u_1,\dots,u_d) & \text{if} \quad S^n_m \in \mathcal{S}^+, \\ \Pi^d(u_1,\dots,u_d) & \text{if} \quad S^n_m \notin \mathcal{S}^+, \end{cases} \tag{21}$$

for every $\langle u_1,\dots,u_d\rangle \in [0,1]^d$.

If we are given a sample from a d-copula C of size $n \geq m$, but we do not have any information about C except the sample, then the terms $s^n_{i_1,\dots,i_d}$ from the d-dimensional matrix S^n_m, give us the relative frequencies of the sample vectors that belong to $R_{\underline{\mathbf{i}}}$ for every $\underline{\mathbf{i}} \in \mathcal{I}_m$, see equation (11), which gives us a partition of $[0,1]^d$. So, it seems natural to spread these frequencies uniformly on the transformed version of $R_{\underline{\mathbf{i}}}$ under the partitions π_j, that is why we select Π^d the product d-copula to define the sample d-copula in equation (21). This idea is very common in Statistics, for example, the empirical distribution function assigns uniform mass $1/n$ to each observed point or vector. On the other hand if S^n_m is not a generalized transformation matrix, as defined above, we define the sample d-copula as Π^d, the reason for this selection is the fact that for dimension $d = 2$ and $m = 2$, if S^n_2 is not a transformation matrix, then there exists at least one column or one row such that the sum of its entries is zero, and as observed in Fredricks et al. [15], if T is a column or row vector then $T(\Pi^2) = \Pi^2$, in the remaining case, that is when only one entry in T is non zero, then we could define $T = (1)$, and in this case $T(\Pi^2) = \Pi^2$, even when T is not a transformation matrix. For larger dimension $d \geq 3$, we refer the reader to the gluing method in Siburg and Stoimenov [31].

If S^n_m is not a generalized transformation matrix then we recommend first to try with smaller values of m, and in the case that the value of m that makes S^n_m a generalized transformation matrix (if it exists) is too small for the required statistical methodology, then see Remark 3.10.

Example 3.1. We generated four samples from the copula Π^2 given by $\langle 0.13587, 0.78362\rangle$, $\langle 0.29310, 0.21312\rangle$, $\langle 0.66104, 0.73981\rangle$ and $\langle 0.88332, 0.43167\rangle$, and we took $m = 2$, then we obtained that $s^4_{1,1} = s^4_{1,2} = s^4_{2,1} = s^4_{2,2} = 1/4$. So, S^4_2 is clearly a transformation matrix which generates the uniform partition given in equation (11), and C^4_2 is simply Π^2, that is, we recover the original 2-copula. We will return to this example in Subsection 3.2.

Now we analyze some of the main properties of S^n_m.

Proposition 3.2. Let $m \geq 2$, let C be a d-copula and let $U_n = \{\underline{\mathbf{x_1}}, \cdots, \underline{\mathbf{x_n}}\}$ be an independent sample of size $n \geq m$ from C, define $q_{\underline{\mathbf{i}}} = V_C(R_{\underline{\mathbf{i}}})$ for every $\underline{\mathbf{i}} \in \mathcal{I}_m$. Then the square d-dimensional matrix $n \cdot S_m^n$ has associated a multinomial distribution with parameters n and $\{q_{\underline{\mathbf{i}}}\}_{\underline{\mathbf{i}} \in \mathcal{I}_m^C}$, where $\mathcal{I}_m^C = \{\underline{\mathbf{i}} \in \mathcal{I} \,|\, q_{\underline{\mathbf{i}}} > 0\}$. Besides, we have that $0 \leq q_{\underline{\mathbf{i}}} \leq 1/m$ and $\sum_{\underline{\mathbf{i}} \in \mathcal{I}_m, i_j = k} q_{\underline{\mathbf{i}}} = 1/m$ for every $j \in \{1, \ldots, d\}$ and for every $k \in \{1, \ldots, m\}$.

Proof. Since $\{R_{\underline{\mathbf{i}}}\}_{\underline{\mathbf{i}} \in \mathcal{I}_m}$ in equation (11) is a partition of $[0,1]^d$, then for every $k \in \{1, 2, \ldots, n\}$ there exists a unique $\underline{\mathbf{i}} \in \mathcal{I}_m$ such that $\underline{\mathbf{x_k}} \in R_{\underline{\mathbf{i}}}$. Observe that from equation (19) $\sum_{\underline{\mathbf{i}} \in \mathcal{I}_m} n \cdot s_{\underline{\mathbf{i}}}^n = n$. Now, define $q_{\underline{\mathbf{i}}} = t_{i_1, \ldots, i_d}$, as in equation (12), for every $\underline{\mathbf{i}} = \langle i_1, \ldots, i_d \rangle \in \mathcal{I}_m$. If $q_{\underline{\mathbf{i}}} = 0$ then $P(\underline{\mathbf{x_k}} \in R_{\underline{\mathbf{i}}}) = 0$ for every $k \in \{1, \ldots, n\}$. So, if we let $\mathcal{I}_m^C = \{\underline{\mathbf{i}} \in \mathcal{I}_m \,|\, q_{\underline{\mathbf{i}}} > 0\}$, then using the independence of the sample

$$P\left(S_m^n = \left(s_{i_1, \ldots, i_d}^n\right)_{i_1, \ldots, i_d = 1}^m\right) = \left(\frac{n!}{\Pi_{\underline{\mathbf{i}} \in \mathcal{I}_m^C} (n \cdot s_{\underline{\mathbf{i}}}^n)!}\right) \Pi_{\underline{\mathbf{i}} \in \mathcal{I}_m^C} \, q_{\underline{\mathbf{i}}}^{n \cdot s_{\underline{\mathbf{i}}}^n}. \tag{22}$$

Therefore, S_m^n has the desired distribution.

The restrictions on the values of the $p_{\underline{\mathbf{i}}}$ follow from equation (13). □

Depending on C we can have some simplifications in Proposition 3.2, for example:

Corollary 3.3. If $C = \Pi^d$ in Proposition 3.2, then

$$P\left(S_m^n = \left(s_{i_1, \ldots, i_d}^n\right)_{i_1, \ldots, i_d = 1}^m\right) = \left(\frac{n!}{\Pi_{\underline{\mathbf{i}} \in \mathcal{I}_m} (n \cdot s_{\underline{\mathbf{i}}}^n)!}\right) \left(\frac{1}{m^d}\right)^n, \tag{23}$$

and if $C = M^d$, where $M^d(u_1, \ldots, u_d) = \min\{u_1, \ldots, u_d\}$, then

$$P\left(S_m^n = \left(s_{i_1, \ldots, i_d}^n\right)_{i_1, \ldots, i_d = 1}^m\right) = \left(\frac{n!}{\Pi_{\underline{\mathbf{i}} \in \mathcal{I}_m^{M^d}} (n \cdot s_{\underline{\mathbf{i}}}^n)!}\right) \left(\frac{1}{m}\right)^n. \tag{24}$$

Proof. If $C = \Pi^d$ just observe that $q_{\underline{\mathbf{i}}} = 1/m^d > 0$ for every $\underline{\mathbf{i}} \in \mathcal{I}_m$, and if $C = M^d$ then $q_{\underline{\mathbf{i}}} = 1/m$ if and only if $R_{\underline{\mathbf{i}}} = ((k-1)/m, k/m]^d$ for some $k \in \{1, \ldots, m\}$. So, in this case, $I_m^M = \{\underline{\mathbf{i}} \in \mathcal{I}_m \,|\, \underline{\mathbf{i}} = \langle k, \ldots, k \rangle$ for some $k \in \{1, \ldots, m\}\}$. □

Now, we state a result about the values of n and m.

Lemma 3.4. Let $m \geq 2$ and let $C = \Pi^d$ the product d-copula, and assume that the sample size of the sample U_n satisfies that $n = m$. Then

$$P(S_m^m \in \mathcal{S}^+) = \frac{(m!)^d}{(m^d)^m}. \tag{25}$$

Proof. First, let $d = 2$ and $m = n$, in this case, $I_m = \{1, \ldots, m\}^2$ and if we define $S_m^m = (s_{i_1, i_2}^m)_{i_1, i_2 = 1}^m$ as in equation (18), then $s_{i_1, i_2}^m = |R_{\langle i_1, i_2 \rangle} \cap U_m|/m$. So, there are at most m vectors, say $\underline{\mathbf{i}}^1, \ldots, \underline{\mathbf{i}}^m \in \mathcal{I}_m$ such that $|R_{\underline{\mathbf{i}}^l} \cap U_m|/m = 1/m > 0$ for every $l \in \{1, \ldots, m\}$. From Fredricks et al.

[15], we know that $S_m^m \in \mathcal{S}^+$ if each column and each row of S_m^m have a positive element. But, since there are at most m entries in S_m^m which are different from zero, then there must be exactly one entry different from zero in each row and in each column. Since S_m^m is a square matrix of order $m \times m$, we can do this in $m!$ forms. So using Corollary 3.3 equation (23)

$$P(S_m^m \in \mathcal{S}^+) = m! \cdot \left(\frac{m!}{1! \cdots 1!} \right) \left(\frac{1}{m^2} \right)^m = \frac{(m!)^2}{m^{2m}}.$$

For $d > 2$ we proceed in a similar way. We know that S_m^m is a d-dimensional square matrix of order m, so, $S_m^m \in \mathcal{S}^+$ if and only if there is exactly one entry of S_m^m different from zero in each coordinate. In this case, proceeding as in the case $d = 2$, in the first coordinate we can select the non zero entry in m^{d-1} forms, for the second coordinate we have $(m-1)^{d-1}$ forms, etc. Then using Corollary 3.3, equation (23) again

$$P(S_m^m \in \mathcal{S}^+) = \Pi_{l=1}^m (l)^{d-1} \cdot \left(\frac{m!}{1! \cdots 1!} \right) \left(\frac{1}{m^d} \right)^m = \frac{(m!)^d}{(m^d)^m},$$

which finishes the proof. □

Remark 3.5. From the proof of Lemma 3.4, it is clear that if the sample size n is less than m, that is, $n < m$, then $\mathcal{S}^+ = \emptyset$, that is why we asked for the condition $n \geq m$ in the definition of a sample d-copula of order m.

Now we give some asymptotic results about C_m^n.

Theorem 3.6. *Let $m \geq 2$, $n \geq m$ and let U_n be an independent sample of size n from a d-copula C for some fixed $d \geq 2$. Define C_m^n as in equation (21). Let S_m^n the d-dimensional square matrix induced by the sample U_n given in equations (17) and (18). Then for every $\underline{\mathbf{i}} = \langle i_1, \ldots, i_d \rangle \in \mathcal{I}_m$ with m fixed,*

$$\lim_{n \to \infty} s_{i_1, \ldots, i_d}^n = V_C(R_{\underline{\mathbf{i}}}) \quad \text{almost surely.} \tag{26}$$

The elements in the partitions $\{p_{j,0}^n, p_{j,1}^n, \ldots, p_{j,m}^n\}$ given in equation (14) satisfy that for every $j \in \{1, \ldots, d\}$ and for every $k \in \{0, 1, \ldots, m\}$,

$$\lim_{n \to \infty} p_{j,k}^n = \frac{k}{m} \quad \text{almost surely.} \tag{27}$$

Therefore, if we define the grid $K_m = \{0, 1/m, 2/m, \ldots, (m-1)/m, 1\}^d$, the sample d-copula C_m^n is such that

$$\lim_{n \to \infty} C_m^n(u_1, \ldots, u_d) = C(u_1, \ldots, u_d) \quad \text{for every} \quad \langle u_1, \ldots, u_d \rangle \in K_m \quad \text{almost surely.} \tag{28}$$

Finally, if we also let $m \to \infty$ with values of $m \approx n^{1/2d}$ we have that

$$C_m^n \quad \text{converges uniformly and almost surely to} \quad C. \tag{29}$$

Proof. Let $m \geq 2$ and $d \geq 2$ be fixed integers, let C be a d-copula and let U_n be a random sample from C of size $n \geq m$. Let $s^n_{i_1,\ldots,i_d}$ be defined as in equation (17), and observe that $s^n_{i_1,\ldots,i_d}$ can be written as

$$s^n_{i_1,\ldots,i_d} = \sum_{j=1}^{n} \frac{1_{R_{\langle i_1,\ldots,i_d \rangle}}(\underline{\mathbf{x}}_{\mathbf{j}})}{n} \quad \text{for every} \quad \underline{\mathbf{i}} = \langle i_1, \ldots, i_d \rangle \in \mathcal{I}_m, \tag{30}$$

where 1_A is the indicator function of A. Using the strong law of large numbers (SLLN), we have that for every $\underline{\mathbf{i}} = \langle i_1, \ldots, i_d \rangle \in \mathcal{I}_m$

$$\lim_{n\to\infty} s^n_{i_1,\ldots,i_d} = P(\underline{\mathbf{x}} \in R_{\langle i_1,\ldots,i_d \rangle}) = V_C(R_{\underline{\mathbf{i}}}) \quad \text{almost surely,} \tag{31}$$

which proves (26). Now using equations (13), (14) and (26), we have that for every $j \in \{1, \ldots, d\}$ and for every $k \in \{1, \ldots, m\}$,

$$\begin{aligned} \lim_{n\to\infty} p^n_{j,k} &= \sum_{i_j=1}^{k} \sum_{i_1=1}^{m} \cdots \sum_{1_{j-1}=1}^{m} \sum_{i_{j+1}=1}^{m} \cdots \sum_{i_d=1}^{m} \lim_{n\to\infty} s^n_{i_1,\ldots,i_d} \\ &= \sum_{i_j=1}^{k} \sum_{i_1=1}^{m} \cdots \sum_{1_{j-1}=1}^{m} \sum_{i_{j+1}=1}^{m} \cdots \sum_{i_d=1}^{m} V_C(R_{\langle i_1,\ldots,i_d \rangle}) \\ &= \sum_{i_j=1}^{k} \frac{1}{m} = \frac{k}{m} \quad \text{almost surely.} \end{aligned} \tag{32}$$

So, (27) holds. Now, using equations (21), (2) , (15) and their generalizations, it is clear that (28) holds.

Finally, equation (29) follows from the upper bounds given by the Polya urn scheme, see Table 5 below, and the multivariate normal approximation of the multinomial distribution. □

Observe that from equation (28), if we let $C_m = \lim_{n\to\infty} C^n_m$, then C_m coincides with μ_1 in equation (4), for $k = m$ and $\mathbf{P} = (p_{i_1,\ldots,i_d})^m_{i_1,\ldots,i_d=1}$, where $p_{i_1,\ldots,i_d} = m \cdot V_C(R_{\langle i_1,\ldots,i_d \rangle})$ for every $\langle i_1, \ldots, i_d \rangle \in \mathcal{I}_m$.

In the definition of C^n_m the sample d-copula of order m given in equation (21), it is very important to check when the d-dimensional square matrix S^n_m belongs to $\mathcal{S}^+$, in terms of the sample size n and the generating copula C. In order to evaluate $P(S^n_m \in \mathcal{S}^+)$, we will use $C = \Pi^d$ and a simulation procedure to approximate its value. We already have an exact value of $P(S^m_m \in \mathcal{S}^+)$, given in Lemma 3.4, when $n = m$, which is the limit case. We give a preliminary study of $P(S^n_m \in \mathcal{S}^+)$ for $d = 2$ and with $m = 2, 3, 4$ and for $d = 3, 4$ with $m = 2, 3$ for $n \geq m$, in the case $C = \Pi$, using 100,000 simulations. Observe that the case $C = \Pi^d$ is the uniform case, hence the most "spread" case among the d-copulas. See Tables 1 and 2, where the values of $P(S^n_m \in \mathcal{S}^+)$ for several values of $n \geq m$ are approximated via simulations. Observe that even for small values of n the probability of obtaining a generalized transformation matrix is close to one. Also observe that probabilities, in the limit case $n = m$, given in Lemma 3.4 are approximated very accurately.

In order to compare behaviors, we obtained 100,000 simulations in dimensions $d = 2$ and $d = 3$ and different sample sizes n from several families, such as M^2, M^3, W^2, Frank, Clayton,

value of n	d=2 and m=2	d=2 and m=3	d=2 and m=4
2	0.24980	-	-
3	0.56013	0.04890	-
4	0.76562	0.19765	0.00867
5	0.88051	0.37881	0.05392
10	0.99588	0.89792	0.61116
15	0.99984	0.98655	0.89724
20	1	0.99849	0.97408
25	1	0.99979	0.99378
30	1	0.99998	0.99870
35	1	1	0.99950
40	1	1	0.99990
45	1	1	0.99995
50	1	1	1

Tab. 1. Approximations of $P(S_m^n \in \mathcal{S}^+)$ for $d = 2$ and $m = 2, 3, 4$.

value of n	d=3 and m=2	d=3 and m=3	d=4 and m=2	d=4 and m=3
2	0.12505	-	0.06192	-
3	0.42239	0.01112	0.31648	0.00249
5	0.82229	0.23426	0.76920	0.14455
10	0.99381	0.85042	0.99238	0.80913
15	0.99980	0.97938	0.99980	0.97177
20	0.99998	0.99759	0.99998	0.99625
25	1	0.99971	1	0.99959
30	1	0.99995	1	0.99997
35	1	1	1	0.99999
40	1	1	1	1

Tab. 2. Approximations of $P(S_m^n \in \mathcal{S}^+)$ for $d = 3, 4$ and $m = 2, 3$.

Normal with different parameters to compare the behavior of $P(S_m^n \in \mathcal{S}^+)$ to the samples coming from the product copula of dimension $d = 2$ and $d = 3$. Some of these results can be seen in Table 3 and Table 4. From Tables 1 and 3 we observe that for very small values of n the product copula gives smaller probabilities of $P(S_m^n \in \mathcal{S}^+)$ than the other distributions, we also have that M^2 produces the largest probabilities compare to the other distributions. However, for values of n between 30 and 50 the probabilities are quite similar for all the distributions. Similar observations can be obtained from Tables 2 and 4.

We simulated several extra examples with copulas with Spearman's rho varying from -1 to 0, and we obtained very similar results. For example for $d = 2$, W^2 gives very similar results as M^2, as expected.

For a further exploration of these results we recommend to see the algorithms to generate samples of d-copulas in Mai and Scherer [22].

Another way of finding an upper bound for $P(S_m^n \in \mathcal{S}^+)$ is to use the Polya approach. Con-

value of n	Clayton $\theta = 2$	Frank $\theta = 5$	Normal $\rho = 0.5$	M^2
4	0.01393	0.01272	0.01037	0.09364
5	0.06941	0.06674	0.05859	0.23416
10	0.62622	0.62193	0.61671	0.78161
15	0.90017	0.89780	0.89658	0.94685
20	0.97561	0.97532	0.97600	0.98712
25	0.99433	0.99395	0.99428	0.99694
30	0.99866	0.99860	0.99863	0.99934
35	0.99948	0.99967	0.99967	0.99988
40	0.99991	0.99992	0.99993	0.99997
45	0.99999	0.99999	0.99995	0.99998
50	0.99999	0.99999	0.99999	0.99998

Tab. 3. Approximations of $P(S_m^n \in \mathcal{S}^+)$ for $d = 2$ and $m = 4$ under different distributions.

value of n	Clayton $\theta = 2$	Frank $\theta = 5$	Normal $\rho = 0.5$	M^3
3	0.02657	0.02285	0.01597	0.21962
5	0.29490	0.28744	0.25843	0.61540
10	0.86428	0.86021	0.85690	0.94941
15	0.98036	0.98036	0.97966	0.99249
20	0.99725	0.99750	0.99737	0.99903
25	0.99971	0.99966	0.99962	0.99989
30	0.99999	0.99995	0.99999	0.99999
35	1	0.99999	1	1
40	0.99999	1	1	1

Tab. 4. Approximations of $P(S_m^n \in \mathcal{S}^+)$ for $d = 3$ and $m = 3$ under different distributions.

sider k boxes and $n \geq k$ balls, for each ball we select uniformly one of the boxes and the ball is placed inside that box, we repeat independently the procedure for the n balls. We want to find the probability that at the end of this procedure there are no empty boxes, let us call this event E_k^n. This probability is known as the Maxwell-Boltzmann occupancy problem formula given by

$$P\left(E_k^n\right) = \sum_{j=0}^{k} (-1)^j \binom{k}{j} \left(1 - \frac{j}{k}\right)^n, \tag{33}$$

see for example Mahmoud [21] page 37. Observe that if we have a sample of size n from the copula product Π^d and we take $2 \leq m \leq n$, then the m^d boxes used in the construction of the empirical copula of size m have the same probability $1/m^d$. If we assume that $n \geq m^d$ and $k = m^d$, in the occupancy problem above it is clear that the matrix $S_m^n \in \mathcal{S}^+$ if every box has at least one ball (observation). Then $P(E_k^n) \leq P(S_m^n \in \mathcal{S}^+)$. So, if we find a value of n, depending on k, such that $P(E_k^n) \approx 1$, then we have that S_m^n is generalized transformation matrix with very high probability. We proposed to use $n(k)$ the minimum value of n such that $P(E_k^{n(k)}) \geq 0.99999995$, if we use the language **R** and we obtain a probability satisfying this condition it is reported as 1.

We obtained the values of $n(k)$ for $1 \leq k \leq 150$, see some values on Table 5, and we fit linear and non linear models to check its behavior. We found that a linear model is a good approximation and that for large values of k the estimated line remains above the real values of $n(k)$. Observe that from Tables 1 and 2 the value of n such that $P(S_m^n \in \mathcal{S}^+)$ is close to one is actually smaller than the values of $n(k)$ where $k = m^d$ in the Polya urn scheme even for small values of k.

value of k	4	8	9	16	25	27	32	49	64	81	100
value of $n(k)$	64	142	162	304	491	533	639	1005	1332	1708	2131

Tab. 5. Values of $n(k)$ such that of $P(E_k^{n(k)}) \approx 1$ for k of the form m^d.

In the remaining part of this section we will study some important statistical applications.

Assume that $d = 2$ and recall that the main concordance measures are Kendall's tau and Spearman's rho. If C is a copula we know that

$$\tau_C = 4\int_0^1\int_0^1 C(u,v)\,\mathrm{d}C(u,v) - 1 \quad \text{and} \quad \rho_C = 12\int_0^1\int_0^1 uv\,\mathrm{d}C(u,v) - 3, \tag{34}$$

see for example [26], equations 5.1.7 and 5.1.15b. Let $2 \leq m \leq n$ and let $U_n = \{\underline{\mathbf{x_1}}, \ldots, \underline{\mathbf{x_n}}\}$ be a sample of size n of a copula C, or a modified sample of a joint continuous distribution $H(x,y)$. Define $s^n_{i_1,\ldots,i_d}, S^n_m, \mathcal{S}^+$ and C^n_m as in equations (17), (18), (20) and (21), and assume that $S^n_m = (s^n_{ij})^m_{i,j=1}$ is a transformation square matrix of order m. Using the same notation as in Fredericks et al. [15] it is easy to see that for every $i, j \in \{0, 1, \ldots, m\}$

$$\begin{aligned}\int\int_{R_{ij}} C^n_m(u,v)\,\mathrm{d}C^n_m(u,v) &= \int_{q_{j-1}}^{q_j}\int_{p_{i-1}}^{p_i} C^n_m(u,v)\frac{s_{ij}}{(p_1-p_{i-1})(q_j-q_{j-1})}\,\mathrm{d}u\mathrm{d}v \\ &= \sum_{i'<i}\sum_{j'<j} s_{i'j'}s_{ij} + \sum_{j'<j}\frac{s_{ij'}s_{ij}}{2} + \sum_{i'<i}\frac{s_{ij'}s_{ij}}{2} + \frac{s_{ij}^2}{4},\end{aligned} \tag{35}$$

and

$$\begin{aligned}\int\int_{R_{ij}} uv\,\mathrm{d}C^n_m(u,v) &= \int_{q_{j-1}}^{q_j}\int_{p_{i-1}}^{p_i} uv\frac{s_{ij}}{(p_1-p_{i-1})(q_j-q_{j-1})}\,\mathrm{d}u\mathrm{d}v \\ &= \frac{s_{ij}}{4}(p_{i-1}+p_i)(q_{j-1}+q_j).\end{aligned} \tag{36}$$

Using (35) and (36) we can prove the following:

Lemma 3.7. Let $d = 2$ and let $U_n = \{\underline{\mathbf{x_1}}, \ldots, \underline{\mathbf{x_n}}\}$ be a sample of size n of a copula C, or a modified sample of a joint continuous distribution $H(x,y)$. Define $s^n_{i_1,\ldots,i_d}, S^n_m, \mathcal{S}^+$ and C^n_m as in equations (17), (18), (20) and (21), and assume that $S^n_m = (s^n_{ij})^m_{i,j=1}$ is a transformation square matrix of order m. Then

$$\tau_{C^n_m} = \sum_{i=1}^{m-1}\sum_{j=1}^{m-1}\sum_{i'=i+1}^{m}\sum_{j'=j+1}^{m} s^n_{ij}s^n_{i'j'} - \sum_{i=1}^{m}\sum_{j=2}^{m}\sum_{i'=i+1}^{m}\sum_{j'=1}^{j-1} s^n_{ij}s^n_{i'j'}, \tag{37}$$

and

$$\rho_{C^n_m} = 3\left(\sum_{i=1}^{m}\sum_{j=1}^{m} s^n_{ij}(p_{i-1}+p_i)(q_{j-1}+q_j) - 1\right). \tag{38}$$

Besides,

$$\tau_{C^n_m} \in \left[-\left(1-\frac{1}{m}\right), \left(1-\frac{1}{m}\right)\right] \quad \text{and} \quad \rho_{C^n_m} \in \left[-\left(1-\frac{1}{m^2}\right), \left(1-\frac{1}{m^2}\right)\right]. \tag{39}$$

Observe that if n is a multiple of m and $s^n_{ii} = 1/m$ for every $i \in \{1,\ldots,m\}$, with $s^n_{ij} = 0$ if $i \neq j$, then the upper bounds in (39) are attained. For example if $m = 2$, and we consider a copula C such that $V_C(R_{\langle 1,1\rangle}) = 1/2 = V_C(R_{\langle 2,2\rangle})$. Here we consider two extreme cases let $C_1(u,v) = M^2(u,v) = \min\{u,v\}$ for every $\langle u,v\rangle \in I^2$ and $C_2(u,v) = \max\{0, u+v-1/2\}$ if $\langle u,v\rangle \in R_{\langle 1,1\rangle}$, $C_2(u,v) = 1/2 + \max\{0, u+v-3/2\}$ if $\langle u,v\rangle \in R_{\langle 2,2\rangle}$ and $C_2(u,v) = 0$ otherwise, that is, C_2 is a shuffle of M^2. In this case using (34) it is easy to see that $\tau_{C_2} = 0$ and $\rho_{C_2} = 1/2$, and obviously $\tau_{M^2} = \rho_{M^2} = 1$. In general, for any $m > 2$ if we let $C_1 = M^2$ and if we define C_2 to be a shuffle of M that behaves like W^2 on each $R_{\langle i,i\rangle}$ for every $i \in \{1,\ldots,m\}$ then we have that $\tau_{C_2} = 1 - 2/m$ and $\rho_{C_2} = 1 - 2/m^2$, but $\tau_{M^2} = \rho_{M^2} = 1$. Therefore, the upper bounds in (39) are the average of the minimum and maximum values of τ_C and ρ_C when we only know that $V_C(R_{\langle i,i\rangle}) = 1/m$ for every $i \in \{1,\ldots,m\}$. Of course for the lower bounds we have a similar result. In order to see how the above methodology of estimation of measures of concordance works, we simulated 10000 samples from the normal copula in dimension $d = 2$ of sizes $n = 100$ and $n = 200$ for different values of ρ between -1 and 1. In Table 6 we report the results of the estimations of ρ, when $n = 200$ with $\rho = 0$ and $\rho = 0.5$, and for $m = 5, 7, 12, 15$. Of course when $\rho = 0$ we are sampling from the product copula Π^2.

In Table 6 we can observe that when $\rho = 0$ the expected value of ρ is close to 0 even for small m, and for the case $\rho = 0.5$ the expected values approach 0.5 from the left when m increases and the variances are stable in both cases. We also observed that the variances decrease when the simple size increases form $n = 100$ to $n = 200$. For positive values of ρ the behavior of the expected values and variances is similar to the case $\rho = 0.5$.

	$\rho = 0$		$\rho = 0.5$	
m	$E(\hat\rho)$	Var$(\hat\rho)$	$E(\hat\rho)$	Var$(\hat\rho)$
5	0.000203	0.004563	0.43255	0.003174
7	-0.000567	0.004817	0.45464	0.003079
12	-0.001174	0.004945	0.47136	0.003124
15	0.000470	0.004974	0.47417	0.003079

Tab. 6: Estimations of ρ for $n = 200$ and real values $\rho = 0$ and $\rho = 0.5$.

m	$E(\hat\rho)$	Var$(\hat\rho)$	min$(\hat\rho)$	max$(\hat\rho)$	upper bound of $\hat\rho$
2	0.74626	0.00002713	0.69532	0.75	0.75
5	0.95760	0.00000292	0.94329	0.95997	0.96
9	0.98614	0.00000058	0.98008	0.98760	0.98765
12	0.99189	0.00000026	0.98574	0.99192	0.99305
15	0.99460	0.00000013	0.99212	0.99542	0.99555

Tab. 7: Estimations of ρ for $n = 200$ when $\rho = 1$.

In Table 7 we first observe that if $\rho = 1$ we are sampling from the copula M^2 then the expected values approach 1 quickly when m increases. Evenmore, if we observe the values of the minima and maxima they approach the upper bound of $\rho_{C_m^n}$ in equation (39), which reflects in smaller variances. Besides, the worst case in the 10000 simulations when $n = 200$ and $m = 15$ is 0.99212 which is very close to one. For negative values of ρ we obtained similar results.

As a second statistical application we proposed a method for the estimation of a parameter when we are sampling from a parametric family $\{C_\theta | \theta \in \Theta\}$ with $\Theta \subset \mathbb{R}$.

For some parametric families $\{C_\theta | \theta \in \Theta\}$ in dimension d, it is possible to make good estimation of the parameter θ using the sample d copula of order m, even in the case when $m = 2$. For example, in the case of some multivariate parametric Archimedean copulas, we observe that $V_{C_\theta}([0, 1/2]^d)$ is a continuous strictly monotone function f of the parameter $\theta \in \Theta \subset \mathbb{R}$ for any $d \geq 2$. This is the case for example in the families Clayton with $\theta \in (0, \infty)$, Frank with $(0, \infty)$, Ali–Mikhail–Haq with $\theta \in [0, 1)$, Gumbel–Hougaard with $\theta \in [1, \infty)$, etc, see [26] Table 4.1. Then by estimating $V_{C_\theta}([0, 1/2]^d)$ using $s^n_{1,1,\ldots,1}$ in the generalized transformation matrix S^n_m as defined in equation (18), we can find a unique value of $\hat{\theta}$ such that $f(\hat{\theta}) = s^n_{1,1,\ldots,1}$. In general, we need to find f^{-1} in order to give $\hat{\theta}$, but in many cases f^{-1} may not have an analytic expression. However, it can be approximated very accurately with a numerical procedure.

Observe that when $m = 2$ from Proposition 3.2, we are trying to estimate the value of $\theta \in \Theta$ such that $f(\theta) = V_{C_\theta}(R_{\underline{i_0}}) = p_{\underline{i_0}} = C_\theta(1/2, 1/2, \ldots, 1/2)$, where $\underline{i_0} = \langle 1, 1, \ldots, 1 \rangle$, based on a sample from C_θ of size n. We know from the basic properties of the multinomial distribution that the number of observations that fall in the d-box $R_{\underline{i_0}}$, let us say $X_{\underline{i_0}}$, is distributed as a binomial with parameters n and $p_{\underline{i_0}}$. Therefore, $s^n_{1,1,\ldots,1} = X_{\underline{i_0}}/n$ is distributed as a rescaled binomial with values in $\{0, 1/n, 2/n, \ldots, 1\}$, and for n large enough $s^n_{1,1,\ldots,1}$ is a good estimator of $f(\theta)$, hence $\hat{\theta} = f^{-1}(s^n_{1,1,\ldots,1})$ is a good estimator of θ. The procedure of estimation follows the next steps:

1. Find the direct image $f[\Theta] = \{f(\theta) = C_\theta(1/2, \ldots, 1/2) | \theta \in \Theta\} \subset I$ for the family $\{C_\theta | \theta \in \Theta\}$.

2. Given a sample $U_n = \{\underline{x_1}, \ldots, \underline{x_n}\}$ find the value of $s^n_{1,1,\ldots,1}$ in the construction of the sample d-copula of order $m = 2$. If $s^n_{1,1,\ldots,1} \in \operatorname{int} f[\Theta]$ proceed with the next steps.

3. If f^{-1} has an analytic expression define $\hat{\theta} = f^{-1}(s^n_{1,1,\ldots,1})$, and we are done. In other case, if Θ is bounded, give a fine grid of Θ, to approximate $f[\Theta]$, otherwise give a fine grid of a bounded subset Θ_0 of Θ such that $s^n_{1,1,\ldots,1} \in f[\Theta_0]$ and it is close to $f[\Theta]$, and use a linear interpolation to estimate $f^{-1}(s^n_{1,1,\ldots,1}) = \hat{\theta}$.

As an application of this methodology we use the Frank family of copulas for $d = 2$ and $d = 3$. In the case $d = 2$ it is easy to see that $f[\Theta] = (0, 1/4) \cup (1/4, 1/2)$ and $f(\theta) = C_\theta(1/2, 1/2)$ is a strictly increasing function which is symmetric with respect to the point $\langle 0, 1/4 \rangle$. If $d \geq 3$ then $f[\Theta] = [1/2^d, 1/2)$ since $\theta \geq 0$. In these cases f^{-1} has no analytic expression, so, we use the grid construction defined above to estimate θ.

We generated 5000 samples of different sizes $n = 500, 1000, 10000, 50000, 100000$ from the Frank copula with parameters $\theta = 2$ and $\theta = 5$. In Tables 8, 9, 10 and 11 we can see the basic statistics of the estimations for the 5000 samples.

n	$E(\hat{\theta})$	$\mathrm{Var}(\hat{\theta})$	$\min(\hat{\theta})$	$\max(\hat{\theta})$
500	2.03938	0.582948	-0.51333	5.54950
1000	2.01002	0.286775	0.35233	3.97850
10000	2.00103	0.027756	1.42433	2.65066
50000	2.00187	0.005590	1.74500	2.27800
100000	1.99985	0.002830	1.81266	2.19766

Tab. 8: Estimations of θ for the Frank copula with $d = 2$ and $\theta = 2$.

n	$E(\hat{\theta})$	$\mathrm{Var}(\hat{\theta})$	$\min(\hat{\theta})$	$\max(\hat{\theta})$
500	2.00465	0.215213	0.37775	3.63266
1000	2.00282	0.103440	0.94150	3.33600
10000	2.00082	0.010230	1.61925	2.42125
50000	2.00155	0.002118	1.85150	2.20225
100000	2.00017	0.001008	1.88675	2.10825

Tab. 9: Estimations of θ for the Frank copula with $d = 3$ and $\theta = 2$.

n	$E(\hat{\theta})$	$\mathrm{Var}(\hat{\theta})$	$\min(\hat{\theta})$	$\max(\hat{\theta})$
500	5.17081	1.950315	1.57509	13.84323
1000	5.07970	0.891967	2.29133	10.10000
10000	5.01109	0.083791	3.90100	6.12500
50000	5.00001	0.016473	4.58750	5.53900
100000	5.00227	0.008161	4.69700	5.32500

Tab. 10: Estimations of θ for the Frank copula with $d = 2$ and $\theta = 5$.

n	$E(\hat{\theta})$	$\mathrm{Var}(\hat{\theta})$	$\min(\hat{\theta})$	$\max(\hat{\theta})$
500	5.07597	0.678271	2.66900	10.07200
1000	5.04366	0.319974	3.16400	7.27900
10000	5.00105	0.031167	4.40900	5.67650
50000	5.00303	0.006224	4.73633	5.28200
100000	5.00040	0.002999	4.82266	5.22500

Tab. 11: Estimations of θ for the Frank copula with $d = 3$ and $\theta = 5$.

As can be observed in Tables 8, 9, 10 and 11, the average estimation of θ is good in all cases, and as expected the variances decrease as n increases. The minima and maxima of the estimations are relatively far from each other when the sample size is $n = 500$. So, we do not recommend to use this methodology for small n. It is also very important to observe that, as expected from the binomial distribution and the central limit theorem, the estimation of θ is quite good for $n = 100000$, but if we try to use the empirical distribution function when $d = 3$, we would need an array of 10^{15} terms, which is needed to perform calculations in order to estimate

θ, which no computer can handle. However, in our tables the elapsed time for each simulation was 15.98 seconds for $\theta = 2$ and $d = 3$, and 16.25 seconds for $\theta = 5$ and $d = 3$.

As a third application we propose a simple goodness-of-fit test. Let us assume that we take a sample of size n coming from a d-copula C and we take $2 \le m \le n$ a fixed integer. Let $R_{\underline{\mathbf{i}}}$ for $\underline{\mathbf{i}} \in \mathcal{I}_m$ be the partition of I^d in the construction of the sample d-copula, and assume that $S^n_m = \left(s^n_{i_1,\ldots,i_d}\right)^m_{i_1,\ldots,i_d=1}$ is a generalized transformation matrix. From Proposition 3.2 we know that the square d-dimensional matrix S^n_m has a multinomial distribution with parameters n and $\{q_{\underline{\mathbf{i}}}\}_{\underline{\mathbf{i}}\in\mathcal{I}^C_m}$, where $\mathcal{I}^C_m = \{\underline{\mathbf{i}} \in \mathcal{I}_m \mid q_{\underline{\mathbf{i}}} > 0\}$. Therefore, we want to test the simple hypothesis

$$H_0 : n \cdot S^n_m \rightsquigarrow \text{Mult}\left(n, \{q_{\underline{\mathbf{i}}}\}_{\underline{\mathbf{i}}\in\mathcal{I}^C_m}\right), \tag{40}$$

against the alternative composite hypothesis $H_1 : n \cdot S^n_m \not\rightsquigarrow \text{Mult}\left(n, \{q_{\underline{\mathbf{i}}}\}_{\underline{\mathbf{i}}\in\mathcal{I}^C_m}\right)$. In the literature there are several proposals for a goodness-of-fit test for the multinomial distribution, see for example [4] or [27]. In order to prove H_0 vs H_1 we used the most common statistics, that is, Pearson's X^2, which has asymptotically a chi-squared distribution with $k-1$ degrees of freedom, where k denotes de the cardinality of $\mathcal{I}^C_m$. Here we present a couple of examples:

In the first one we choose the case $d = 2$, $m = 2$, and the Frank copula with $\theta = 10$, and different values of n, in this case $q_{\langle 1,1\rangle} = q_{\langle 2,2\rangle} = 0.43136$ and $q_{\langle 1,2\rangle} = q_{\langle 2,1\rangle} = 0.06844$. In Table 12 we give the basic statistical results of 10000 simulations with values of $n = 5, 25, 100, 250, 500$ and $n = 1000$. As we can observe, even for n as small as 25, the expected values of $s^n_{i,j}$ for $i, j \in \{1, 2\}$ are close to the real ones, with smaller variances as n increases as expected. In Table 13 we present the results of the number of rejections of H_0 at level $\alpha = 0.05$ for $n = 100, 250, 500, 1000$ and the mean of the p-values of the 10000 tests. From these results we can see that the test performs as expected when $\alpha = 0.05$. In order to check the power of the test, depending on θ the parameter of the Frank copula, we performed 10000 tests of H_0 as above, for $n = 1000$ with different values of θ varying from $\theta = 6$ up to $\theta = 15$ taking integer values, the number of rejections of H_0 in order were 9997 for $\theta = 6$, 9815 for $\theta = 7$, 6692 for $\theta = 8$, 1860 for $\theta = 9$, 1147 for $\theta = 11$, 3577 for $\theta = 12$, 6646 for $\theta = 13$, 8816 for $\theta = 14$ and 9709 for $\theta = 15$.

Observe that the null hypothesis (40) does not characterize a unique copula, but only gives the volumes of the d-boxes needed in the construction of a d-sample copula of order m. However, when m is large enough (40) approximates closely the underlying copula C, by Theorem 3.6. We also performed simulations for $d = 3$ for different families of 3-copulas obtaining similar results. Of course, we can also use different statistics to test (40), for example the ones proposed in [4] or [27].

n	$E(s^n_{1,1})$	Var$(s^n_{1,1})$	$E(s^n_{1,2})$	Var$(s^n_{1,2})$	$E(s^n_{2,1})$	Var$(s^n_{2,1})$	$E(s^n_{2,2})$	Var$(s^n_{2,2})$
5	0.4359	0.03509	0.0668	0.01200	0.0658	0.01196	0.4315	0.03426
25	0.4312	0.00973	0.0681	0.00263	0.0690	0.00254	0.4316	0.00986
100	0.4431	0.00244	0.0688	0.00063	0.0685	0.00064	0.4319	0.00244
250	0.4312	0.00099	0.0683	0.00025	0.0685	0.00025	0.4320	0.00101
500	0.4312	0.00049	0.0688	0.00013	0.0686	0.00013	0.4314	0.00048
1000	0.4314	0.00025	0.0686	0.00006	0.0687	0.00006	0.4313	0.00024

Tab. 12: Estimations of $s^n_{ij}, i, j \in 1, 2$ for the Frank copula with $d = 2$, $m = 2$ and $\theta = 10$.

n	Number of rejections	Mean of p-value
100	485	0.50105
250	487	0.49428
500	508	0.49923
1000	492	0.50140

Tab. 13: Rejections of H_0 for the Frank copula with $d = 2$, $m = 2$ and $\theta = 10$.

As a second example we generated 10000 simulations of the copula M^2 for $m = 3$ and different values of n between 5 and 1000. In this case $q_{\langle 1,1\rangle} = q_{\langle 2,2\rangle} = q_{\langle 3,3\rangle} = 1/3$ and zero in any other case. On Table 14 we report the basic statistics of the simulations. On Table 15 we report the number of rejections of H_0 for $n \geq 25$, and observe that even for $n = 25$ we obtain nice results.

n	$E(s^n_{1,1})$	$\mathrm{Var}(s^n_{1,1})$	$E(s^n_{2,2})$	$\mathrm{Var}(s^n_{2,2})$	$E(s^n_{3,3})$	$\mathrm{Var}(s^n_{3,3})$
5	0.33288	0.019545	0.33364	0.019482	0.33347	0.019482
25	0.33371	0.008919	0.33326	0.008833	0.33302	0.008880
100	0.33331	0.002206	0.33327	0.002223	0.33340	0.002229
250	0.33340	0.000889	0.33319	0.000889	0.33340	0.000889
500	0.33329	0.000444	0.33331	0.000445	0.33339	0.000445
1000	0.33324	0.000222	0.33336	0.000222	0.33339	0.000223

Tab. 14: Estimations of $s^n_{ij}, i, j \in 1, 2$ for the copula M^2 with $d = 2$ and $m = 3$.

n	Number of rejections	Mean of p-value
25	477	0.49930
100	543	0.49839
250	496	0.49869
500	498	0.50014
1000	507	0.49868

Tab. 15: Rejections of H_0 for the M^2 copula with $d = 2$ and $m = 3$.

As a fourth application we propose a methodology to test two simple hypotheses, that is,

$$H_0 : \underline{X} \rightsquigarrow C_0 \quad \text{VS} \quad H_1 : \underline{X} \rightsquigarrow C_1 \tag{41}$$

where C_0 and C_1 are two completely determined d-copulas which are obviously different. Let $m \geq 2$ fixed and denote by $\mathrm{Vol}_0(R_{\underline{\mathbf{i}}})$ and $\mathrm{Vol}_1(R_{\underline{\mathbf{i}}})$ to the volumes of the uniform partition given in (11) of $[0, 1]^d$ under H_0 and H_1 respectively. Observe that since $C_0 \neq C_1$ then it is clear that there exist $m \geq 2$ and $\underline{\mathbf{i}}_1, \underline{\mathbf{i}}_2 \in \mathcal{I}_m$ such that $\mathrm{Vol}_0(R_{\underline{\mathbf{i}}_1}) \neq \mathrm{Vol}_1(R_{\underline{\mathbf{i}}_1})$ and $\mathrm{Vol}_0(R_{\underline{\mathbf{i}}_2}) \neq \mathrm{Vol}_1(R_{\underline{\mathbf{i}}_2})$. In several typical cases $m = 2$ satisfies the above condition.

Let $U_n = \{\underline{\mathbf{x_1}}, \ldots, \underline{\mathbf{x_n}}\}$ be a random sample from the d-copula C_0 and obtain the matrix S^n_m as defined in equations (17) and (18). If S^n_m is a generalized transformation matrix, using Proposition 3.2 and the classical Neyman–Pearson's theorem, see for example [19], we can find the likelihood ratio $L(\theta_0; n \cdot S^n_m)/L(\theta_1; n \cdot S^n_m)$, where $\theta_0 = \{\mathrm{Vol}_0(R_{\underline{\mathbf{i}}})\}_{\underline{\mathbf{i}}\in\mathcal{I}_m}$ and $\theta_1 = \{\mathrm{Vol}_1(R_{\underline{\mathbf{i}}})\}_{\underline{\mathbf{i}}\in\mathcal{I}_m}$, which is given by

$$T^d_{m,n} = \Pi_{\underline{\mathbf{i}}\in\mathcal{I}_m}\left(\frac{\mathrm{Vol}_0(R_{\underline{\mathbf{i}}})}{\mathrm{Vol}_1(R_{\underline{\mathbf{i}}})}\right)^{n\cdot s^n_{\underline{\mathbf{i}}}}. \tag{42}$$

Then the best critical region to test H_0 VS H_1 is given by

$$D = \{U_n \mid U_n \text{ is a sample of size } n \text{ from } C_0 \text{ such that } T^d_{m,n} \leq K_\alpha\},$$

where $P(T^d_{m,n} \leq K_\alpha | H_0 \text{ is true}) = \alpha$ and $0 < \alpha < 1$ is the probability of Type 1 error.

Since the distribution of $T^d_{m,n}$ is not known, we can estimate K_α for $\alpha = 0.01, 0.05$ and $\alpha = 0.1$, by simulating samples from C_0 a large number of times L, we estimate the required quantiles of the distribution of $T^d_{m,n}$. We recommend to use $L \geq 50000$ to estimate the values of K_α.

We calculated $T^d_{m,n}$ 10,000 times to test all possible couples of the following 2-copulas: Clayton(6), Frank(14.1385) and Gumbel(4), which correspond to three copulas with the same Kendall's tau where $\tau = 0.75$. We used $m = 6$ and $m = 8$, $n = 150$ and $L = 1,000,000$. We compare our results to the best percentage of rejections, out of seven different statistics given in Genest et al. [16], Table 3, when $m = 6$ and $m = 8$ in Table 16.

Copula under H_0	True copula	$m = 6$	$m = 8$	Best percentage in [16]
Clayton	Clayton	5.15	5.20	4.9-5.4
Clayton	Frank	99.63	100	99.9
Clayton	Gumbel	100	100	99.9
Frank	Clayton	99.73	99.99	96.6
Frank	Frank	5.01	5.08	4.5-5.2
Frank	Gumbel	80.69	93.91	81.9
Gumbel	Clayton	100	100	99.9
Gumbel	Frank	80.24	93.51	83.8
Gumbel	Gumbel	5.19	4.88	4.4-5.2

Tab. 16: Percentage of Rejections of H_0 for $n = 150$ and $\alpha = .05$.

From Table 16 we can see that when $m = 6$ we have similar results as in Genest et al. [16]. However, when $m = 8$ we improve in all cases the percentages of rejections in [16]. When the true copula coincides with the copula under H_0 we report the percentage average of the test when H_0 is fixed, and in column 5 we report the lower and upper percentages in [16].

We are in the process of making all the comparisons with [16] in the preprint *Testing simple hypotheses using the sample d-copula of order m*. For example, in the comparison for Plackett(68.46996) VS Frank(14.1385) both with $\tau = 0.75$, we obtained 90.3 % of rejections while the best percentage reported out of the seven statistics used in Genest et al. [16] is 18.5 %, that is, we improved their power by more than 70 %.

It is very important to observe that it may be possible to make Bayesian inference. By Proposition 3.2, we know that the square d-dimensional matrix S^n_m needed in the construction of the d sample copula of order m follows a multinomial distribution, with restrictions on the values of the $p_{\underline{\mathbf{i}}}$ for $q_{\underline{\mathbf{i}}} \in \mathcal{I}_m^C$. So, we could try to extend the classical approach of considering a Dirichlet prior for the parameters in order to obtain the posterior distribution based on a sample, as in [2]. But, this is material for future research.

Now we study the general setting of the sample d-copulas.

3.2. Sample d-Copula of Order m for a Continuous d-Distribution Function

Let $m, d \geq 2$ be fixed integers and let H be a continuous d-distribution function in $\mathbb{R}^d$. Let $V_n = \{\underline{\mathbf{z_1}}, \ldots, \underline{\mathbf{z_n}}\}$ be a random sample from H of size $n \geq m$. Let $U_n = \{\underline{\mathbf{x_1}}, \ldots, \underline{\mathbf{x_n}}\}$ be the usual **modified sample** or pseudo sample, that is, if $j \in \{1, \ldots, n\}$ and $\underline{\mathbf{z_j}} = \langle z_{j,1}, \ldots, z_{j,d}\rangle \in \mathbb{R}^d$, define for $k \in \{1, \ldots, d\}$

$$R_{j,k} = \sum_{l=1}^{n} 1_{\{z_{l,k} \leq z_{j,k}\}}, \tag{43}$$

where $R_{j,k}$ is the rank of the observation $z_{j,k}$ for l varying between 1 and n. Now define for every $j \in \{1, \ldots, n\}$, $\underline{\mathbf{x_j}} = \langle x_{j,1}, \ldots, x_{j,d}\rangle$ where

$$x_{j,k} = \frac{R_{j,k}}{n} \quad \text{and for every} \quad k \in \{1, \ldots, d\}. \tag{44}$$

Then

$$U_n = \{\underline{\mathbf{x_1}}, \ldots, \underline{\mathbf{x_n}}\} \subset \{1/n, \ldots, (n-1)/n, 1\}^d \subset [0,1]^d. \tag{45}$$

Of course, from the continuity assumption on H, the ranks in the definition of $\underline{\mathbf{x_j}}$ are all different for every $j \in \{1, \ldots, n\}$ almost surely.

Recall that the **empirical d-copula** for the modified sample U_n is defined for every $\langle u_1, \ldots, u_d\rangle \in [0,1]^d$ by

$$C^n(u_1, \ldots, u_d) = \frac{1}{n} \sum_{j=1}^{n} 1_{\{x_{j,1} \leq u_1, \ldots, x_{j,d} \leq u_d\}}, \tag{46}$$

see for example Nelsen [26]. Observe that the empirical d-copula is not a d-copula. For example, if $\underline{\mathbf{u}} = \langle u_1, \ldots, u_d\rangle$ and $0 < u_1 < 1/n$ then $C^n(u_1, \ldots, u_d) = 0$. In fact, since $C^n(u_1, \ldots, u_d) = 0$ if for some $j \in \{1, \ldots, d\}$, $u_j = 0$. Then it is well known that the restriction of C^n to the grid $\{0, 1/n, \ldots, (n-1)/n, 1\}$ is a d-subcopula.

For $m \geq 2$, $n \geq m$ and U_n a modified random sample from a continuous d-distribution function H. Define $s^n_{i_1,\ldots,i_d}, S^n_m, \mathcal{S}^+$ and C^n_m as in equations (17), (18), (20) and (21).

In this case the structure of the modified sample U_n simplifies significantly the structure of the sample d-copula of order m, as can be seen in the following:

Theorem 3.8. Let U_n be a modified random sample obtained from an original random sample V_n of H a continuous d-distribution function in $\mathbb{R}^d$. Define $s^n_{i_1,\ldots,i_d}, S^n_m, \mathcal{S}^+$ and C^n_m as in equations (17), (18), (20) and (21). Then

$$S^n_m \in \mathcal{S}^+ \quad \text{for every} \quad 2 \leq m \leq n, \tag{47}$$

that is, S^n_m is always a generalized transformation square matrix. Besides, if $2 \leq m \leq n$ and we define for every $j \in \{1, \dots, d\}$ the partitions $\pi^n_j = \{0 = p_{j,0}, p_{j,1}, \dots, p_{j,m-1}, p_{j,m} = 1\}$ given in equation (14), then

$$p_{j,k} = \frac{\lfloor \frac{k \cdot n}{m} \rfloor}{n} \quad \text{for every } j \in \{1, \dots, d\} \text{ and for every } k \in \{1, \dots, m\}, \tag{48}$$

where $\lfloor a \rfloor$ denotes the greatest integer less than or equal to a. In particular, when $n = m$, $\pi^n_j = \{0, 1/n, \dots, (n-1)/n, 1\}$ for every $j \in \{1, \dots, d\}$.

Even more, when $n = m \geq 2$ the sample d-copula C^n_n is such that

$$C^n_n(u_1, \dots, u_d) = C^n(u_1, \dots, u_d) \text{ for every } \langle u_1, \dots, u_d \rangle \in \{0, 1/n, \dots, (n-1)/n, 1\}^d, \tag{49}$$

that is, we recover the empirical d-copula defined in equation (46) on the grid $\{0, 1/n, \dots, 1\}^d$.

Proof. Let U_n be a modified random sample obtained from an original random sample V_n of H a continuous d-distribution function in $\mathbb{R}^d$ and define $s^n_{i_1,\dots,i_d}$, S^n_m, $\mathcal{S}^+$ and C^n_m as in equations (17), (18), (20) and (21).

Of course, it is enough to see that equation (47) holds for the limit case, that is, when $n = m \geq 2$. So Assume that $n = m \geq 2$, in this case, from equations (43) and (44), we know that $x_{j,k} = R_{j,k}/n$ for every $j, k \in \{1, \dots, n\}$. But, since all the ranks are different with probability one, we have that the matrix $S^n_n = (s^n_{i_1,\dots,i_d})^n_{i_1,\dots,i_d=1}$, given in equation (18), satisfies that for every $j \in \{1, \dots, d\}$ and for every $k \in \{1, \dots, n\}$

$$\sum_{i_1=1}^{n} \cdots \sum_{i_{j-1}=1}^{n} \sum_{i_{j+1}=1}^{n} \cdots \sum_{i_d=1}^{n} s^n_{i_1,\dots,i_{j-1},i_j=k,i_{j+1},\dots,i_d} = \frac{1}{n}. \tag{50}$$

Therefore, S^n_n is a d-dimensional square matrix which is a generalized transformation matrix, that is, $S^n_n \in \mathcal{S}^+$. So, (47) holds.

Now, assume that m is such that $2 \leq m \leq n$ and define for every $j \in \{1, \dots, d\}$ the partitions $\pi^n_j = \{0 = p_{j,0}, p_{j,1}, \dots, p_{j,m-1}, p_{j,m} = 1\}$ given in equation (14). Then we know that

$$p_{j,k} = \sum_{i_j=1}^{k} \sum_{i_1=1}^{m} \cdots \sum_{i_{j-1}=1}^{m} \sum_{i_{j+1}=1}^{m} \cdots \sum_{i_d=1}^{m} s^n_{i_1,\dots,i_{j-1},i_j,i_{j+1},\dots,i_d}.$$

Now using the sample size n, the partition of $[0,1]^d$ given by $\{R_{\underline{\mathbf{i}}}\}_{\underline{\mathbf{i}} \in \mathcal{I}_m}$, see equation (11), and by equation (30), we have that there are $\lfloor (k \cdot n)/m \rfloor$ points in the regions defined by $p_{j,k}$, where $\lfloor a \rfloor$ is the greatest integer less than or equal to a. Therefore, (48) holds.

Finally, if we assume that $n = m \geq 2$, using the definition of the d-sample copula of order n, C^n_n given in equation (21), the definition of the empirical copula in equation (46), the partition given in equation (11), together with (15) and its generalizations. It is easy to see that equation (49) also holds. □

Observe that in the last Theorem, if n is a multiple of m, then by equation (48), $p_{j,k} = k/m$ for every $j \in \{1, \dots, d\}$ and for every $k \in \{1, \dots, m\}$, that is, we recover the original partition of $[0,1]^d$. In the case that n is not a multiple of m the partition given in equation (48) is still a good approximation of the original partition given by $\{R_{\underline{\mathbf{i}}}\}_{\underline{\mathbf{i}} \in \mathcal{I}_m}$ in equation (11).

Remark 3.9. We know that the empirical copula C^n is a d-subcopula, then if we use the multivariate extension of Lemma 2.3.5 in [26], used in the proof of Sklar's theorem, we can extend the empirical copula C^n to a d-copula C^n_* using standard multilinear interpolation. Observe that in this case $C^n_* = C^n_n$ the sample d-copula of order n.

Returning to Example 3.1 and by Remark 3.9 the density of the 2-copula $C^4_* = C^4_4$ is given by

$$c^4_*(u,v) = \begin{cases} 4 & \text{if} \quad \langle u,v\rangle \in R_{1,4} \cup R_{2,1} \cup R_{3,3} \cup R_{4,2} \\ 0 & \quad\;\;\; \text{otherwise,} \end{cases}$$

where $R_{1,4} = [0,1/4] \times (3/4,1], R_{2,1} = (1/4,2/4] \times [0,1/4], R_{3,3} = (2/4,3/4] \times (2/4,3/4]$ and $R_{4,2} = (3/4,1] \times (1/4,2/4]$. If we take the copula C^4_* which has the above density then the sup distance between C^4_* and the real one Π^2 is 3/16. If we use any other distance it is obvious that the distance between the sample 2-copula of order 2 C^4_2 and the real one is always zero.

Remark 3.10. If we are sampling from a d-copula C and S^n_m is not a generalized transformation matrix for the value of m required in a statistical procedure, then we recommend to obtain the modified sample and, using Theorem 3.8, the modified S^n_m is always a generalized transformation matrix, even in the case $m = n$. However, we also recommend to add a warning saying that the original sample did not allow us to obtain the sample d-copula of order m, this only happens for relatively small sample sizes n or large values of m.

The statistical procedures presented in Section 3.1 can be used for modified samples. For example, in the case of the concordance measures Kendall's tau and Spearman's rho, we can observe that if we have two continuous random variables X and Y, such that $Y = f(X)$, where f is a strictly increasing function almost surely, then it is well known that the copula $C_{X,Y}$ is the M^2 copula. But, in this case, it is obvious to see that if we have an independent random sample of size n of $\langle X,Y\rangle$, and we take $m = n$, then $\tau_{C^n_n} = 1 - 1/n$ and $\rho_{C^n_n} = 1 - 1/n^2$ with probability one, which correspond to the upper bounds in (39). So, even for small values of n both measures are close to one.

In order to see how the estimation procedure in Section 3.1 works for modified samples, we generated 10000 samples of different sizes n of a joint distribution with exponential margins and corresponding copula Frank with parameter $\theta = 5$ and $d = 2$. We use $n = 200, 500, 1000, 5000$ and $n = 10000$. In general the results had the same behavior as the one in Table 10, providing good estimators of θ.

As an application of the hypothesis testing of (40) with modified samples we generated 10000 samples of $Z = \langle X_1, X_2, X_3\rangle$ of three independent normal variables with corresponding variances 1, 4 and 9, with different sample sizes $n = 500, 1000, 10000$ and $n = 100000$. Then we obtained the modified samples for each simulation, and we calculate the corresponding 3-dimensional transformation matrices S^n_3 for $m = 3$. Finally we tested the hypothesis $H_0 : q_{\langle i,j,k\rangle} = 1/27 = 0.0370370$ for $i,j,k \in \{1,2,3\}$, corresponding to independence.

Instead of giving large tables we report only the extreme cases for each sample size for the twenty seven 3-boxes included in each test.

For $n = 500$ the minimal expected value observed was 0.03683, the maximal expected value was 0.03727, the minimal value observed was 0.006 and the maximal was 0.078, and the maximal variance was 0.0000554. For $n = 1000$ the minimal expected value observed was 0.03690, the maximal expected value was 0.03726, the minimal value observed was 0.013 and the maximal was 0.065, and the maximal variance was 0.0000277. For $n = 10000$ the minimal expected value observed was 0.03700, the maximal expected value was 0.03706, the minimal value observed was 0.0297 and the maximal was 0.044, and the maximal variance was 0.0000028. For $n = 100000$ the minimal expected value observed was 0.037020, the maximal expected value was 0.037056, the minimal value observed was 0.03499 and the maximal was 0.03935, and the maximal variance was 0.00000028. From these results it is clear that the estimations of the $q_{\underline{\mathbf{i}}}$ largely improve as n increases.

When we implement the hypothesis testing we observed that when $\alpha = 0.05$, the number of rejections of (40) was quite small for $n \geq 500$.

We also applied the test with $m = 2$ and $d = 3$ with similar results.

As a last application we give a new of test of independence for multivariate normal data. We generated 10000 samples of sizes $n = 1000, 10000$ and $n = 100000$ from a tetravariate normal with mean $\mu = \underline{0}$ and variances $\sigma_1 = \sigma_2 = \sigma_3 = \sigma_4 = 1$ and covariance matrix

$$\begin{pmatrix} 1 & \rho_{1,2} & \rho_{1,3} & \rho_{1,4} \\ \rho_{1,2} & 1 & \rho_{2,3} & \rho_{2,4} \\ \rho_{1,3} & \rho_{2,3} & 1 & \rho_{3,4} \\ \rho_{1,4} & \rho_{2,4} & \rho_{3,4} & 1 \end{pmatrix}.$$

We want to test the hypothesis

$$H_0 : \rho_{i,j} = 0 \text{ for every } i, j \quad \text{VS} \quad H_1 : \text{there exist} \, i, j \text{ such that} \, \rho_{i,j} \neq 0.$$

We first obtained the modified samples of the tetravariate normal, then we took $m = 2$ and we obtained the sample 4-copulas of order 2. We observe that H_0 holds if and only if the coordinates are independent. So, the copula associated to the observations is the product copula Π^4, in this case we have that we can test alternatively

$$H_0 : V(R_{\underline{\mathbf{i}}}) = \frac{1}{2^4} \text{ for every } \underline{\mathbf{i}} \in \mathcal{I}_2 \quad \text{VS} \quad H_1 : \text{ there exists } \underline{\mathbf{i}} \in \mathcal{I}_2 \text{ such that } V(R_{\underline{\mathbf{i}}}) \neq \frac{1}{2^4}.$$

To find the power of this test we took $\rho_{1,2} \neq 0$ and $\rho_{i,j} = 0$ for any other i, j. The results of these tests for $n = 1000, 10000$ and $n = 100000$ are given in Tables 17, 18 and 19. We used Pearson's χ^2 to test the hypotheses.

In Tables 17, 18 and 19 $p - v$ stands for "p-value". In Table 17 we observe that when the sample size is $n = 1000$ the number of rejections of H_0 increases rapidly when $\rho_{1,2}$ varies from 0.1 up to 0.4, as expected the mean of the p-value decreases as well as its variances. It is important to notice that we obtained practically the same results when we took $\rho_{i_0,j_0} \neq 0$ for any $i_0 \neq j_0$ with $i_0, j_0 \in \{1, 2, 3, 4\}$ and the remaining $\rho' s = 0$. We also observed that if we let another $\rho_{i,j} \neq 0$ with $\{i, j\} \neq \{1, 2\}$ the number of rejections increases even faster. This last observation holds also for Tables 18 and 19. In Table 18 we observe that when the sample size is $n = 10000$ the number of rejections of H_0 increases when $\rho_{1,2}$ varies from 0.03 up to 0.14, observe also that we would reject independence when $\rho_{1,2} = 0.14$ even for $\alpha = 0.01$. In Table 19 we observe that when the sample size is $n = 100000$ the number of rejections of H_0 increases rapidly when $\rho_{1,2}$ varies from 0.01 up to 0.0425, the final observation in Table 17 holds for Table 18.

$\rho_{1,2}$	Rejections	$E(p - v)$	$\mathrm{Var}(p - v)$	$\min(p - v)$	$\max(p - v)$
0.1	676	0.5113	0.0947	$6.5\ e^{-6}$	0.9999
0.2	5690	0.1147	0.0327	$8.2\ e^{-11}$	0.9983
0.3	9811	0.0040	0.0040	$1.1\ e^{-15}$	0.6424
0.4	10000	$1.2\ e^{-5}$	$1.4\ e^{-7}$	$8.4\ e^{-26}$	0.0295

Tab. 17: Rejections of H_0 for $n = 1000$ with $\rho_{1,2} \neq 0$.

$\rho_{1,2}$	Rejections	$E(p - v)$	$\mathrm{Var}(p - v)$	$\min(p - v)$	$\max(p - v)$
0.03	625	0.5278	0.0941	$3.7\ e^{-7}$	0.9999
0.07	6968	0.0729	0.0196	$2.7\ e^{-12}$	0.9576
0.11	9982	0.0005	$2.4\ e^{-5}$	$8.2\ e^{-23}$	0.1906
0.14	10000	$1.5\ e^{-6}$	$1.9\ e^{-9}$	$2.8\ e^{-31}$	0.0035

Tab. 18: Rejections of H_0 for $n = 10000$ with $\rho_{1,2} \neq 0$.

$\rho_{1,2}$	Rejections	$E(p - v)$	$\mathrm{Var}(p - v)$	$\min(p - v)$	$\max(p - v)$
0.01	700	0.5076	0.0959	$1.6\ e^{-6}$	0.9999
0.02	5593	0.1217	0.0357	$1.1\ e^{-10}$	0.9970
0.03	9722	0.0053	0.0006	$1.8\ e^{-17}$	0.5603
0.04	9997	$4.8\ e^{-5}$	$1.7\ e^{-6}$	$5.7\ e^{-28}$	0.0818
0.0425	10000	$3.2\ e^{-6}$	$4.7\ e^{-6}$	$5.4\ e^{-28}$	0.0052

Tab. 19: Rejections of H_0 for $n = 100000$ with $\rho_{1,2} \neq 0$.

It is quite important to provide elapsed times for each individual run, from different sample sizes. In Table 20 we give the average elapsed times of each test using modified samples, and the last column gives us the average times to find the empirical copulas. We observed that for $n \geq 600$ the language **R** can not allocate arrays of the required size (NA). We ran our simulations using the language **R** in a Dell Precision 490 Workstation, we used extensively the copula package in our programs [18]. It is clear that the elapsed times increase linearly with the sample size n, instead of polynomially as it is the case for the empirical copula.

n	seconds for $m = 2$	seconds for $m = 3$	seconds for $m = n$
50	0.007	0.03	113.28
100	0.017	0.06	1799.36
500	0.08	0.25	722061
1000	0.15	0.50	NA
10000	1.57	4.87	NA
100000	15.94	48.28	NA

Tab. 20: Elapsed times for different sample sizes n in dimension $d = 3$.

The empirical d-copula has big restrictions in terms of evaluations in computers, for example if we consider a sample size $n = 1000$ in dimension $d = 4$, then we need an array of 10^{12} entries, and in many situations we have to perform calculations with this array, which generally can not be supported in a computer. The sample d copula of order m only needs an array of m^d entries which is more manageable specially for small m. Since we can use in its definition $2 \leq m \leq n$, we recommend to use $m = 2$ as the first approximation, in many instances the sample d-copula of order 2 gives us some preliminary information about the data, as observed in Section 3.

The sample d-copula of order m can be used in several statistical procedures, such as goodness of fit tests, tests of symmetry, estimation of one or more parameters in parametric models, etc.

Of course by equation (49) in Theorem 3.8, we can also use all the asymptotic results known for the empirical d-copula in the case that $n = m$. In the case that $2 \leq m < n$ we think that the convergence of the sample d-copula of order m has also nice asymptotic properties, but this is a topic for future research.

4. FINAL REMARKS

In the last years many researchers have been proposing methods of constructing multivariate copulas, see for example [9, 28] and [14]. The idea is to provide new families that allow to model multivariate data, since the known models are not numerous enough to do so.

To find multivariate extensions of known results for 2-copulas is of great importance, and lately several papers have been written to achieve this goal. In the case of ordinal sums, see [1] or [26], we have a multivariate extension given in Mesiar and Sempi [24]. For the shuffles, see [26], we have the extension of Durante and Fernández-Sánchez [10]. For extensions in construction of multivariate copulas with a given diagonal, we may cite [20, 29] and [5]. Another interesting references are [6, 7] and [34], see also a recent note on singular copulas in [12].

The importance of the construction of what the authors called self-similar 2-copulas in Cuculescu and Theodorescu [5], was extended in [15] to construct interesting examples of 2-copulas with given fractal supports. For the multivariate extension of the construction of fractal copulas quite recently in Trutschnig and Fernández-Sánchez [33], using the results in [15], give a method using transformation matrices to construct new interesting d-copulas.

In this paper we provide in Proposition 2.3 the multivariate generalization of the construction in Cuculescu and Theodorescu [5].

In Section 3 we introduced the sample d-copula of order m , based on the ideas of the transformation matrices given in [15], and its generalization in [33], in two settings: First when the sample is obtained from a d-copula C, and second when the sample comes from a continuous d-distribution function on $\mathbb{R}^d$.

In the first case, we observed that the sample d-copula has very nice properties and we provided some important asymptotic results. We also observe that even for small values of n the d-dimensional square matrix S^n_m, used in the definition of the sample d-copula of order m, C^n_m in equation (21), is with high probability a generalized transformation matrix. We also provide interesting statistical applications such as a new methodology for estimation of parameters, a goodness-of-fit test results about the concordance measures and a comparison with the empirical copula results when testing two simple hypotheses, improving their powers.

In the second case, for $2 \leq m \leq n$, we proved that d-dimensional square matrix S^n_m, used in the definition of the sample d-copula of order m, C^n_m in equation (21), is always a generalized transformation matrix, which allows us to have a non trivial sample d-copula. We also saw that we can recover the empirical d-copula from C_n^n. We also observed that the statistical applications in the first case can be carried out easily to this case by using the modified samples. We introduce a new independence test for multivariate normal data with nice power.

In both cases, the empirical d-copula of order m can be used to study the statistical properties of the sample, and to try to model the d-copula that "better fits" the observations.
How to choose $m = m(n)$? The selection of m depends on the type of statistical inference required and on the proposed methodology. For example, in our applications we observed that $m = 2$ works fine when we are dealing with parameter estimation for an Archimedean family, or the new independence test for multivariate normal data, proposed in Section 3.2. In the case of the estimation of concordance measures Kendall's tau and Spearman's rho we need relatively large values of m as observed in Section 3.1. Finally, in the new proposal of testing $H_0 : \underline{X} \rightsquigarrow C_0$ VS $H_1 : \underline{X} \rightsquigarrow C_1$, which uses the classical Neyman–Pearson's Theorem, which gives the best critical region, see for example Hogg and Craig (1978) [19]. We compare our results to the ones obtained using the empirical copula in Genest et al. We observe that we need m to be larger than or equal to 6, at least in the case that the sample size is $n = 150$. For $m = 8$ we improved all the percentages of rejections given in Genest et al., in some cases the improvement is quite significant.

The empirical copula first proposed by Deheuvels in [8], which he called "*fonction de dépendance empirique*" has very nice theoretical properties. However, for large samples even in small dimensions it has big problems in applications, because of the limitations of a standard computer. If the sample size n is small as well as the dimension d we can still use all the strong statistical techniques developed for empirical copulas, see for example [3, 13] or [16]. However, if the sample size n is large, let us say $n \geq 100000$, even in small dimensions $d = 2$ or $d = 3$, the statisticians require new tools and methods which can be easily implemented in standard computers, without the need of taking a much smaller subsample. We think the sample d copula of order m may be this new tool.

We believe the sample d-copula of order m may be quite useful in applications, because it is easy to obtain and any computer can handle the arrays needed for its construction, considering medium values of m and values of d not so small, even if the sample size n is quite large. This last fact is a great advantage for any statistician.

You can find all the programs in https://sites.google.com/site/probstatsr, the programs were written using language **R**.

ACKNOWLEDGEMENT

We want to thank two anonymous referees for the their valuable comments and questions which greatly improved this paper.

(Received November 23, 2012)

REFERENCES

[1] C. Alsina, M. J. Frank, and B. Schweizer: Associative Functions: Triangular Norms And Copulas. World Scientific Publishing Co., Singapore 2006.

[2] J. O. Berger and J. M. Bernardo: Ordered group reference priors with application to the multinomial problem. Biometrika *79* (1992), 1, 25–37.

[3] E. Bouyé, V. Durrleman, A. Nikeghbali, G. Riboulet, and T. Roncalli: Copulas for Finance. A Reading Quide and Some Applications. Groupe de recherche opérationnelle, Crédit Lyonnais, Paris 2000.

[4] N. Cressie and T. R. C. Read: Multinomial goodness-of-fit tests. J. Roy. Statist. Soc., Ser. B *46* (1984), 3, 440–464.

[5] I. Cuculescu and R. Theodorescu: Copulas: Diagonals, tracks. Rev. Roumaine Math. Pures Appl. *46* (2001), 6, 731–742.

[6] E. de Amo, M. Díaz Carrillo, and J. Fernández-Sánchez: Measure-preserving functions and the independence copula. Mediterr. J. Math. *8* (2011), 431–450.

[7] E. de Amo, M. Díaz Carrillo, and J. Fernández-Sánchez: Copulas and associated fractal sets. J. Math. Anal. Appl. *386* (2012), 528-541.

[8] P. Deheuvels: La fonction de dépendance empirique et ses propriétés. Un test non paramétrique d'indépendance. Acad. Roy. Belg. Bull. Cl. Sci. *65* (1979), 5, 274–292.

[9] F. Durante, J. J. Quesada-Molina and M. Úbeda-Flores: On a family of multivariate copulas for aggregation processes. Inform. Sci. *177* (2007), 5715–5724.

[10] F. Durante and J. Fernández-Sánchez: Multivariate shuffles and approximation copulas. Statist. Probab. Lett. *80* (2010), 1827–1834.

[11] F. Durante, J. Fernández-Sánchez, and C. Sempi: Sklar's theorem obtained via regularization techniques Nonlinear Anal. *75* (2012), 2, 769–774.

[12] F. Durante, J. Fernández-Sánchez, and C. Sempi: A note on the notion of singular copula. Fuzzy Sets and Systems *211* (2013), 120–122.

[13] J. D. Fermanian, D. Radulović, and M. Wegcamp: Weak convergence of empirical copula. Bernoulli *10* (2004), 5, 847–860.

[14] J. Fernández-Sánchez, R. B. Nelsen, and M. Úbeda-Flores: Multivariate copulas, quasi-copulas and lattices. Statist. Probab. Lett. *81* (2011), 1365–1369.

[15] G. A. Fredricks, R. B. Nelsen, and J. A. Rodríguez-Lallena: Copulas with fractal supports. Insurance Math. Econom. *37* (2005), 42–48.

[16] C. Genest, B. Rémillard, and D. Beaudoin: Goodness-of-fit tests for copulas: A review and a power study. Insurance Math. Econom. *44* (2009), 199–213.

[17] M. M. Hernández-Cedillo: Topics on Multivariate Copulas and Applications. Ph.D. Thesis, Universidad Nacional Autónoma de México 2013, preprint.

[18] M. Hofert, I. Kojadinovic, M. Maechler, and J. Yan: copula: Multivariate Dependence with Copulas. R package version 0.999-5. http://CRAN.R-project.org/package=copula, 2012.

[19] R. H. Hogg and A. T. Craig: Introduction to Mathematical Statistics. Fourth edition. Collier Macmillan International Eds., New York-London 1978.

[20] P. Jaworski: On copulas and their diagonals. Inform. Sci. *179* (2009), 2863–2871.

[21] H. M. Mahmoud: Polya urn models. Texts Statist. Sci. Ser., Chapman and Hall/CRC, New York 2008.

[22] J. F. Mai and M. Scherer: Simulating Copulas: Stochastic Models, Sampling Algorithms and Applications. Series in Quantitative Finance *4*, Imperial College Press, London 2012.

[23] M. Marcus: Some properties and applications of double stochastic matrices. Amer. Math. Monthly *67* (1960), 215–221.

[24] R. Mesiar and C. Sempi: Ordinal sums and idempotent of copulas. Aequat. Math. *79*, (2010), 1–2, 39–52.

[25] P. Mikusiński and M. D. Taylor: Some approximations of n-copulas. Metrika *72* (2010), 385–414.

[26] R. B. Nelsen: An Introduction to Copulas. Lecture Notes in Statist. *139*, second edition, Springer, New York 2006.

[27] T. R. C. Read and N. Cressie: Goodness-of-fit Statistics for Discrete Multivariate Data. Springer Series in Statist., Springer, New York 1988.

[28] J. A. Rodríguez-Lallena and M. Úbeda-Flores: Distribution functions of multivariate copulas. Statist. Probab. Lett. *64* (2003), 41–50.

[29] T. Rychlik: Distributions and expectations of order statistics for possible dependent random variables. J. Multivariate Anal. *48* (1994), 31–42.

[30] S. Sherman: Doubly stochastic matrices and complex vector spaces. Amer. J. Math. *77* (1955), 245–246.

[31] K. F. Siburg and P. A. Stoimenov: Gluing copulas. Comm. Statist. Theory and Methods. *37* (2008), 3124–3134.

[32] A. Sklar: Fonctions de répartition à n dimensions et leurs marges. Publ. Inst. Statist. Univ. Paris *8* (1959), 229–231.

[33] W. Trutschnig and J. Fernández-Sánchez: Idempotent and multivariate copulas with fractal support. J. Statist. Plan. Infer. *142* (2012), 3086–3096.

[34] W. Trutschnig: Idempotent copulas with fractal support. Adv. Comp. Intel. *298*, (2012), 3, 161–170.

José M. González-Barrios, Universidad Nacional Autónoma de México, Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas, Departamento de Probabilidad y Estadística, Circuito Escolar s/n, Ciudad Universitaria, C. P. 04510, México D.F.. México.
e-mail: gonzaba@sigma.iimas.unam.mx

María M. Hernández-Cedillo, Universidad Nacional Autónoma de México, Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas, Departamento de Probabilidad y Estadística, Circuito Escolar s/n, Ciudad Universitaria, C. P. 04510, México D.F.. México.
e-mail: magdaz@gmail.com