Article **Journal Computational Simulation**
December 2023, Vol.7 No.18 17-29

Fast procedure to compute empirical and Bernstein Copulas

Procedimiento rápido para calcular cópulas empíricas y de Bernstein

HERNÁNDEZ-MALDONADO, Victor Miguel*[1], ERDELY, Arturo[2], DIAZ-VIERA, Martin Alberto[3] and RIOS-SOLIS, Leonardo[4]

[1]*Dirección Adjunta de Innovación y Conocimiento (DAIC).*
[2]*Universidad Nacional Autónoma de México, FES Acatlán.*
[3]*Instituto Mexicano del Petróleo. Universidad Nacional Autónoma de México. Instituto de Geofísica*
[4]*Metabolic and Bioprocessing. Engineering School of Natural and Environmental Sciences, Newcastle University, Newcastle Upon Tyne, UK. Honorary Associate Professor, Institute for Bioengineering, Centre for Systems and Synthetic Biology, The University of Edinburgh, Scotland, UK.*

ID 1st Author: *Victor Miguel, Hernández-Maldonado* / **ORC ID**: 0000-0002-9306-8535, **CVU CONAHCYT ID**: 174514

ID 1st Co-author: *Arturo, Erdely* / **ORC ID**: 0000-0003-1653-8342

ID 2nd Co-author: *Martin Alberto, Diaz-Viera* / **ORC ID**: 0000-0001-5811-6186

ID 3rd Co-author: *Leonardo, Rios-Solis* / **ORC ID**: 0000-0002-4387-984X

DOI: 10.35429/JCS.2023.18.7.17.29 Received: November 20, 2023; Accepted December 30, 2023

Abstract

In this work, a novel technique for efficient computation of bivariate empirical copulas and, by extension, non-parametric copulas. The algorithm addresses discrete and finite equations, integrating mathematical-statistical components. It introduces two novel concepts: Propagation and Overlapping, to enhance computations and their comprehension during empirical copula construction. The algorithm is presented in pseudo-code for its implementation in any programming language. Comparative performance assessments demonstrate computing speeds ranging from 60 to 250 times faster than the standard algorithm across multiple case studies. Recent research highlights the utility of copulas in Artificial Intelligence (AI) techniques for enhanced predictions [I]. Existing studies center on parametric copulas, underscoring the significance of introducing a methodology for non-parametric copula implementation because this approach facilitates precise modeling of non-linear relationships among random variables, offering substantial improvements over conventional techniques, and boosting its integration, within the realm of artificial intelligence.

Empirical Copula, Bernstein Copula, Fast Algorithm

Resumen

En este trabajo se presenta una técnica novedosa para el cálculo eficiente de cópulas empíricas bivariadas y, por extensión, cópulas no paramétricas. El algoritmo aborda ecuaciones discretas y finitas, integrando componentes matemático-estadísticos. Introduce dos conceptos novedosos: Propagación y Superposición, para mejorar los cálculos y su comprensión durante la construcción de cópulas empíricas. El algoritmo se presenta en pseudocódigo para su implementación en cualquier lenguaje de programación. Las evaluaciones comparativas de rendimiento demuestran velocidades de computo que van de 60 a 250 veces más rápidas que el algoritmo estándar en múltiples estudios de casos. Investigaciones recientes destacan la utilidad de las cópulas en técnicas de Inteligencia Artificial (IA) para predicciones mejoradas [I]. Los estudios existentes se centran en cópulas paramétricas, lo que subraya la importancia de introducir una metodología para la implementación de cópulas no paramétricas porque este enfoque facilita el modelado preciso de relaciones no lineales entre variables aleatorias, ofreciendo mejoras sustanciales sobre las técnicas convencionales e impulsando su integración dentro del ámbito. de inteligencia artificial.

Cópula empírica, Cópula de Bernstein, Algoritmo rápido

Citation: HERNÁNDEZ-MALDONADO, Victor Miguel, ERDELY, Arturo, DIAZ-VIERA, Martin Alberto and RIOS-SOLIS, Leonardo. Fast procedure to compute empirical and Bernstein Copulas. Journal Computational Simulation. 2023. 7-18: 17-29

* Correspondence to the author (e-mail: vmhernann@gmail.com)
† Researcher contributing as first author.

1. Introduction

An algorithm characterized by precise computational outcomes is deemed commendable, but one that achieves both precision and efficiency attains a superior status. Preeminent among algorithms is the one that necessitates minimal computational time and memory resources, while concurrently delivering exacting results, as elucidated in previous studies [II, III].

Traditionally, an algorithm's performance is evaluated through the meticulous scrutiny of these two metrics, as expounded upon in the work by [IV]

1. **Memory efficiency**. Quantified as the requisite memory allocation, colloquially termed space complexity, pertains to the memory utilization encompassing the following components:

a) Instruction space is influenced by the compiler, its configuration parameters, and the architectural attributes of the target computer's central processing unit (CPU).

b) Data space is susceptible to variability contingent upon the program's allocation of dynamic memory, the presence of static variables, and the scale of the data.

c) Data space can be modulated through the program's dynamic memory allocation practices, data volume, and static variable characteristics.

2. **Time efficiency**. The temporal resources required for the execution of a program or function, often referred to as time complexity, are of paramount concern. Expedient task completion is a desirable attribute for a program or function; nevertheless, the actual temporal duration of execution is contingent upon a multitude of influential factors:

a) The computational speed of the computer, encompassing attributes beyond mere clock speed, including CPU architecture, I/O subsystem efficiency, and other pertinent factors.

b) The intrinsic characteristics of the compiler, in conjunction with its configurable settings and associated options.

c) The scale of the data set, exemplified by tasks involving extensive or compact data structures.

d) The inherent attributes of the data under consideration, such as the positional index of a name within a sequential search operation (e.g., first or last occurrence).

The primary aim of this research endeavor is to formulate an algorithm capable of computationally generating a bivariate empirical copula with optimal spatiotemporal efficiency, minimizing both memory utilization and computational time. Subsequently, we employ the outcomes of this methodology to ascertain the Bernstein Copula, leveraging parallel computing methodologies for enhanced computational throughput.

1.1. Why non-parametrical copulas?

This discourse marks the inception of a critical evaluation concerning the advantages and disadvantages associated with the utilization of multivariate distributions featuring asymmetrical marginal probability distributions. A pivotal concept in this context is the copula, serving as an intermediary link between a multivariate probability distribution and its univariate marginal probability distributions, as meticulously elucidated by [V]. In general, the adoption of copula-based methodologies provides a direct and highly efficient framework for elucidating the inter-dependencies among stochastic variables, as comprehensively discussed by [VI].

Copulas represent a valuable tool for comprehending the dependencies among various outcomes. By establishing a connection between univariate marginal distributions and their comprehensive multivariate counterparts, these analytical instruments facilitate a profound understanding of these associations. Originating in *1959* within the realm of probabilistic metric spaces, copulas have since evolved into a pivotal analytical tool, as underscored by [II].

ISSN 2523-6865

Notably, there has been a conspicuous surge in the literature dedicated to unraveling their statistical properties and practical applications in recent times. This versatile apparatus, characterized by its intrinsic capability to model and estimate joint multivariate distributions, proves indispensable across a wide spectrum of disciplines. Due to their adaptability, copulas provide a potent means of characterizing the underlying dependency structure among random variables, as expounded upon by [VII]. Nonparametric copulas offer a distinct advantage over alternative statistical methodologies by enabling the replication of nonlinear joint distributions that may defy linear or Gaussian assumptions. These copulas, notable for their non-restrictive treatment of marginal constraints, possess the capacity to capture nonlinear relationships effectively.

The ongoing academic discourse revolves around the persistent deliberation between non-parametric copulas and their parametric counterparts, as evidenced by the works of [VII], [VIII], [IX], [XI], and [XII]. Generally, [XIII] advocates for a critical perspective on the computational efficiency of the pseudo-likelihood inference method, which necessitates the use of parametric copulas alongside empirical distributions for the marginal variables, as elaborated upon by [XIV]. Within this context, Feng Lin briefly hints at the possibility of discovering more efficient estimators for elliptical copulas, as exemplified in [XV]. Nevertheless, it is worth noting recent advancements in non-parametric probability estimation, as underscored by [XVI]. Emphasizing the challenges involved, [XVII] argue that obtaining exact estimations of risk measures remains elusive due to the intricate interplay between the available data and the selected model. While acknowledging the inherent limitations of each model, it is imperative to recognize that certain models may provide a more precise representation than others.

Copulas serve as a widely adopted framework for modeling dependence structures. However, in certain scenarios, the appropriate selection of a specific parametric copula remains uncertain based solely on the available data, as elucidated by [VIII] Various methodologies have been proposed for estimating copula functions.

One such method employs a parametric approach, wherein a particular copula family with estimable parameters is chosen through maximum likelihood estimation. This parametric technique finds extensive practical utility due to its convenience and simplicity, as underscored by [XVIII]. An alternative approach is the semi-parametric method, which combines a parametric copula model with a non-parametric model for the marginal distributions, as introduced by [IX]. A third approach to copula estimation embraces a wholly non-parametric methodology, which, unlike parametric methods, obviates the need for specifying a copula model by relying solely on observed data. The chief advantage of this approach, as emphasized by [XIX], lies in its adaptability to accommodate diverse forms of dependence structures.

The central focus of our research methodology in this investigation centers on the adoption of non-parametric copulas. We have developed an innovative methodology that leverages the Bernstein copula representation to generate multivariate probability distributions. Our proposed approach involves the use of efficient and expeditious procedures for constructing empirical copulas and Bernstein copulas within predefined multivariate distributions with fixed marginal characteristics. Specifically, we employ a mixture distribution framework to represent the Bernstein copula. Furthermore, we introduce a Calculations Number Reduction (CNR) procedure to estimate the empirical copula and employ a parallel computing strategy to expedite the computation of the Bernstein copula, as detailed in [XX]. It is essential to emphasize that non-parametric copulas entail substantial computational demands.

1.2. The calculation of the empirical copula

The essence of statistical science lies in the conceptualization and quantification of relationships among random variables. In *1959*, Sklar introduced and coined the term ‘copula functions’, unveiling a foundational connection between the individual distribution functions of a random vector and its joint distribution, specifically through the incorporation of a copula, as extensively documented by [V].

ISSN 2523-6865

The copula function effectively encapsulates the dependence structure among variables within the joint distribution, while the marginal distributions elucidate the behavior of each individual variable in isolation. Sklar's theorem asserts that any joint probability distribution can be expressed as a copula function evaluated within the univariate marginal probability distributions.

Consider a random vector (X, Y) with a joint probability distribution function denoted as $F_{X,Y}$. This function is mathematically expressed as $F_{X,Y}(x, y) = P(X \leq x, Y \leq y)$. Additionally, we define the marginal continuous distribution functions as $F_X(x) = F_{X,Y}(x, +\infty)$ and $F_y(y) = F_{X,Y}(+\infty, y)$. This specific function is recognized as the product copula. According to Sklar's Theorem, as presented in [V], there exists a unique copula function $C\_\{X, Y\}$ such that:

$$F(x, y) = C_{X,Y}(F_X(x), F_Y(y)) \quad (1)$$

This result holds substantial importance in the field of bivariate statistics and modeling, enabling the separate modeling of marginal probability distributions and the characterization of their dependence structure. Copulas have demonstrated their versatility in a wide array of domains, including finance [XXI], as well as engineering and environmental science [XXII].

It is crucial to emphasize that the marginal distributions lack information concerning the interactions among random variables, thus consolidating the entirety of dependence information within the underlying copula function. As a result, it becomes feasible to quantify measures of dependence exclusively through copulas, independent of the information provided by the marginal distributions.

Furthermore, it is noteworthy that random variables X and Y are independent continuous variables if and only if their joint distribution function can be expressed as the product of their respective marginal distributions, denoted as $F_{X,Y} = F_X(x) \cdot F_Y(y)$. Consequently, it can be deduced that the unique underlying copula characterizing independence is:

$$\prod(u, v) = uv. \quad (2)$$

This function is also referred to as the product copula, as mentioned in [VIII]. Consequently, the copula serves as the repository for all information regarding the dependency structure of continuous stochastic variables, and the evaluation of independence among such variables can be determined through their copula.

Consider a set of n observations denoted as $S = \{\hat{u}_1, \hat{v}_1), \cdots, (\hat{u}_n, \hat{v}_n)\}$, which are drawn from a random vector (X, Y). We may obtain empirical estimates for the marginal distributions X, Y can be derived using the following procedure:

$$\hat{F}_n(x) = \frac{1}{n}\sum_{k=1}^{n} \mathbb{I}\{x_k \leq x\}, \hat{G}_n(y) = \frac{1}{n}\sum_{k=1}^{n} \mathbb{I}\{x_k \leq x\} \quad (3)$$

Where $\mathbb{I}$ represents the indicator function, which yields a value of 1 when its input is true and 0 otherwise. According to [XXIII], it is commonly accepted that the empirical distribution $\hat{F}_j$ serves as a consistent estimator of $\hat{F}_j$, signifying that $\hat{F}_j$ converges to $\hat{F}_j$ almost surely as the sample size n approaches infinity, denoted as $n \to \infty$, for all values of t.

Similarly, the bivariate empirical copula definition aligns with prior studies by [VII] and [XXIV].

$$C_n\left(\frac{i}{n}, \frac{j}{n}\right) = \frac{1}{n}\sum_{k=1}^{n} \mathbb{I}_A\{x_k \leq x_i, y_k \leq y_i\} \quad (4)$$

In this context, x_i denotes the order statistic corresponding rank $i \in \{1, \ldots, n\}, C_n\left(\frac{i}{n}, 0\right) = 0 = C_n(0, \frac{j}{n})$and $\mathbb{I}_A$ is the indicator function of condition A.

Sklar's theorem provides a universally applicable framework for constructing a joint distribution function by leveraging the copula function. This copula function serves as a pivotal tool for disentangling marginal probability distributions from correlations, effectively capturing and representing the dependency structure, thereby establishing itself as an indispensable feature of copulas.

Sklar's Theorem extends beyond the constraint of continuous marginals. In the context of simulating continuous random variables, the utilization of the marginal empirical distribution function Eq. (3) For every (u,v) is unsuitable, as it assumes a stepwise nature, resulting in discontinuities.

ISSN 2523-6865

This necessitates the adoption of a smoothing method. Given that the primary objective of employing copulas is to simulate a target variable based on one or more explanatory variables, obtaining a smooth estimate of the marginal quantile function becomes crucial. The quantile function, defined as $Q(u) = F^{-1}(u) = infx: F(x) \geq u\}$, where $0 \leq u \leq 1$, can be realized through the application of Bernstein polynomials, as elucidated by [XXV].

$$\tilde{Q}_n(u) = \sum_{k=1}^{n} \frac{1}{2}\left(x_{(k)} + x_{k+1}\right)\binom{n}{k}u^k(1-u)^{n-k} \quad (5)$$

To obtain a smooth estimation of the underlying copula, we utilize the Bernstein copula, described in [XI].

$$\tilde{C}(u,v) = \sum_{i=0}^{n}\sum_{j=0}^{n} C_n\left(\frac{i}{n},\frac{j}{n}\right)\binom{n}{i}u^i(1-u)^{n-i}\binom{n}{j}v^j(1-v)^{n-j} \quad (6)$$

For every (u, v) in the unit hypercube $[0,1]^m$ and $C_n\left(\frac{i}{n},\frac{j}{n}\right)$ is the empirical copula, defined in Eq. (4) Sancetta, A., and Satchell, S. (2004) [XI].

Usually, when employing equation (4) to construct a two-dimensional empirical copula, a precise yet relatively straightforward computational method requires the examination of each cell within the matrix at least once, involving multiple operations at each instance. This approach is notably inefficient for addressing the problem. The Currently Visited Cell *(CVC)* refers to the particular matrix location under scrutiny at the current moment. The following section elucidates the procedural intricacies and outlines the minimal computational requirements associated with this approach.

1.3. The standard method goes as follows:

In summary, the procedure entails the following steps:

– Initially, organize the empirical data set based on the secondary variable and record this ordering.

– Subsequently, sort the empirical data based on the primary variable and record this ordering as well.

– Initiate a nested loop to iterate through the entire matrix of the empirical copula, calculating the new value for each cell.

– Count the number of observations in the original data set before reaching the coordinates of the Currently Visited Cell *(CVC)*, followed by performing the operation $(observations \,/\, n)$.

– These steps represent the standard application of equation (4). However, it is crucial to note that this method exhibits relatively slow computational efficiency.

In the following section, we will introduce our accelerated approach, primarily based on minimizing computational operations during the activation of the *(CVC)*.

2. Empirical Copula fast calculation (Algorithm)

1. Initially, it is imperative to have an empirical data structure, herein referred to as a data set, which is characterized by two columns, denoted as $(m = 2)$, and comprises a total of ***n*** *rows*. This data set functions as a representation of bivariate data, wherein one of the columns serves as the secondary variable, while the other column represents the primary variable.

2. Sort the data set in ascending order based on the secondary variable while preserving the underlying dependency structure related to the primary variable. Document this arrangement by assigning integer values in ascending order as follows: Introduce a new column labeled as *"sortedSVID"* and assign the sequential values from 1 to n to each row within the data set.

3. We systematically apply the same procedure to the primary variable by sorting the data set in ascending order and then introducing a novel column labeled *"sortedPVID"*, in which values are assigned incrementally from 1 to n. These sequential steps meticulously preserve the underlying dependency structure, as the *"sortedSVID"* and *"sortedPVID"* columns will serve as the coordinates of the observations during the construction of the empirical copula.

ISSN 2523-6865

4. We establish a square matrix with dimensions of $[(n+1)\ \mathrm{x}(n+1)]$ dedicated to the empirical copula, initializing its elements to zero.

5. Leveraging the ordered data set, meticulously arranged to preserve the dependency structure through the organization based on primary and secondary variables, we employ unique identification *'ID'* values as exact coordinates within the empirical copula matrix to represent individual observations.

6. Following its creation, the matrix is systematically traversed, with the *'ID'* column values serving as precise coordinates. This process initiates a *propagation and overlapping process* when an observation is encountered.

2.1. The propagation and overlapping concepts

Initially, we provide a data set consisting of 20 elements, as detailed in Table (1), with the intention of constructing its empirical copula. The scatter plot illustrating this data set is visually presented in Figure 1.

ID	PV	SV
1	0.029702970	0.04
2	0.079207921	0.19
3	0.168316832	0.10
4	0.188118812	0.64
5	0.267326733	0.49
6	0.366336634	0.78
7	0.405940594	0.30
8	0.485148515	0.37
9	0.534653465	0.58
10	0.594059406	0.92
11	0.653465347	0.17
12	0.702970297	0.35
13	0.732673267	0.70
14	0.762376238	0.14
15	0.792079208	0.87
16	0.831683168	0.27
17	0.851485149	0.63
18	0.881188119	0.46
19	0.940594059	0.98
20	0.950495050	0.07

Table 1 A data set of 20 values

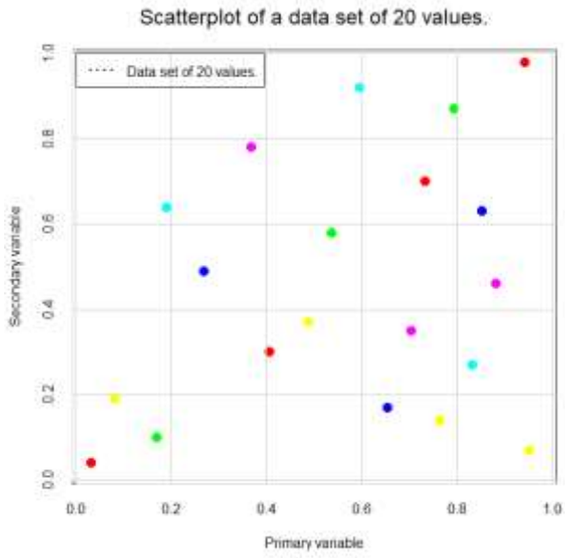


Figure 1 Scatter plot of a data set of 20 bivariate observations. This graph was created in R-Project Software.

– The concept of propagation involves the iterative replication of an observation's value, represented as $(1/n)$, within consecutive cells. To implement this concept, it is essential to determine the precise origin coordinates of the observation within the empirical copula matrix and then incrementally add the value $(1/n)$ to that specific cell. Figure 2 visually illustrates the determination of the observation's origin, highlighted within the black background cell. The propagation concept entails the repetitive application of this value $(1/n)$ in an upward direction along both dimensions $(x\ and\ y)$ until reaching the terminus of the empirical copula matrix, as demonstrated by the light yellow background cells in Figure 2.

ISSN 2523-6865

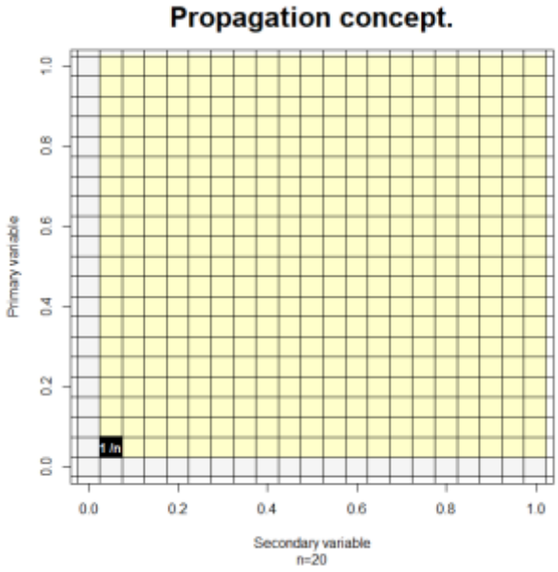


Figure 2 A bivariate empirical copula was created using Table 1 information, incorporating the propagation concept with *light yellow background cells*. The *black background cell* marks its starting point, and $(1/n)$ is systematically propagated upward in the copula matrix. This graph was created in R-Project Software.

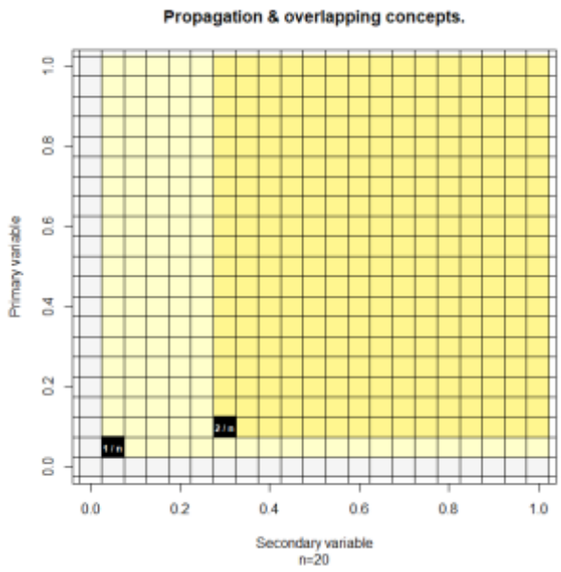


Figure 3 This figure illustrates propagation *light yellow background* starting from a black cell with $(1/n)$. Overlapping is shown with *dark yellow cells*, signifying a second black cell with $(2/n)$. This graph was created in R-Project Software.

– The concept of overlapping pertains to situations where two or more observations coincide at the same location. It is highly probable that during the process of extracting coordinates from the empirical copula's data set and its subsequent propagation, instances of overlapping will occur within the empirical copula matrix. For example, as illustrated in Figure 3, such an occurrence is observed in the second cell marked with a black background. When such overlapping transpires, the value ***(1/n)*** associated with the cell where the new observation is positioned is combined with the previously propagated value *(1/n)*, resulting in *(2/n)*. This gives rise to two consequences: firstly, the initiation of a new propagation process starting from the newly inserted observation, and secondly, the application of the same overlapping protocol, signifying the accumulation of the preexisting value with the newly introduced value, denoted as $CVC = [((xn)/n) + (1/n)]$. This dynamic is exemplified by the dark yellow background in Figure (3).

Our prior analysis has led us to an optimal algorithm that aims to minimize the number of operations performed on each accessed cell. This implies that each cell within the empirical copula matrix is accessed only once. Accordingly, we have devised an algorithm capable of concurrently achieving these two objectives. In essence, it systematically traverses every cell within the empirical copula matrix, executing a single operation that encompasses both propagation and overlapping simultaneously.

2.2. Algorithm 1. Pseudo-code for fast computation

The proposed universal pseudo-code offers a readily accessible and efficient solution for expediting computations of the empirical copula. It possesses adaptability across various programming languages and enhances the overall efficiency of copula-based modeling. The comprehensive explanation and practical demonstration using a data set consisting of 20 values empower users with the confidence to incorporate this pseudo-code into their projects, thereby driving progress in research domains where copulas and Artificial Intelligence techniques play a central role in enabling accurate predictions and improving the modeling of complex relationships.

ISSN 2523-6865

The Pseudo-code:

The presented pseudo-code algorithm is tailored for rapid empirical copula computation, specifically targeting bivariate data sets. Furthermore, it incorporates an illustrative example featuring a data set comprising 20 values (Table 1).

INPUT: Empirical data set model, be sure to place the secondary variable in the first column and the primary in the second one.

OUTPUT: Matrix of the Empirical Copula.

computingEmpCop() **{**

1. # Populate a new array with the Empirical data set.
   datasetModelMatirx ← Empirical data set

2. # Retrieve and update essential data set parameters.
   #Assign the quantity of rows.
n ← Get *(size (datasetModelMatirx[Rows,]))*

Assign the number of Columns.
m ← Get *(size (datasetModelMatirx[, Cols]))*

Set the secondary variable column number.
svcn←Set (1)
Set the primary variable column number.}
svcn←Set (2)

3. # Create a full-of-zeros 2D matrix named *datasetModelIDMatrix* that will hold the original data set and two empty columns.
Dimension *datasetModelIDMatrix* [n, (m+2))] ← 0

Populate the *datasetModelIDMatrix* with the empirical model's data set, leaving the final two columns blank.
Set datasetModelIDMatrix [n, m] ←
datasetModelMatirx [n, m]

We possess a matrix comprising four columns, which we shall designate as follows: Column 1 for the Secondary Variable (*SV*), Column 2 for the Primary Variable (*PV*), Column 3 for the Secondary Variable ID (SVID), and Column 4 for the Primary Variable ID (PVID). Our next action involves sorting the *dataset Model IDMatrix* based on Columns 1 and 2. We will then set this order in Cols 3 and 4, respectively, while steadfastly preserving the underlying dependence structure.

4. # Sort datasetModelIDMatrix in terms of column *(SV)*, without losing the dependence structure.
SORT *(datasetModelIDMatrix [n,1])*

Now you document this sequence by populating column 3, SVID in ascending order from 1 *to n*.
Set *datasetModelIDMatrix [n, 3]*←(1,…, n)

5. # Sort *datasetModelIDMatrix* in terms of column *(PV)*, without losing the dependence structure.
SORT *(datasetModelIDMatrix [n,2])*

Now you can store this sequence by populating column *4*, *PVID* in ascending order from *1 to n*.

Set *datasetModelIDMatrix [n, 4]*←(1,…, n)
Upon the careful population of the *SVID* and *PVID* columns, while simultaneously preserving their underlying dependency structure, these columns represent coordinates of observations within the matrix of the Empirical Copula.

6. # The Empirical Copula's matrix is created with an initial value of zero for all its elements.
Dimension *empCop[(n+1), (n+1))]* ← *0*

7. # To begin with let's focus on the Fréchet-Hoeffding limits, populating both upper limits *(x, y)* of the Empirical Copula's matrix.
Fill Both *French-Hoffding* limits.
FOR *x* = 1 to *(n+1)*
empCop[j, (n+1)] = j/n Secondary variable lim
empCop[(n+1), j] = j/n 'Primary variable lim
END

8. # To populate the interior of the Empirical Copula's matrix, we suggest the creation of an integer variable (pvcoo) that begins with an initial value of 0. This variable will systematically store the coordinate value of the Primary Variable (PV) when the Seconday Variable (SV) has already been determined.
SET **pvcoo** as Integer
SET **pvcoo** ← 0

9. # Set up a nested loop structure to systematically traverse all elements within the matrix of the Empirical Copula. The outer loop will move horizontally (in the x-direction, by columns), while the inner loop will do vertically

ISSN 2523-6865

(in y-direction, by rows), simultaneously addressing both propagation and overlapping.

```
FOR SVID = 1 to (n)
  # Read the PVID coordinate value (from
  datasetModelIDMatrix) and put it into 'pvcoo'
  variable.
  SET pvcoo = datasetModelIDMatrix [SVID, 4]
  # The "Propagation and Overlapping concepts"
  are key components of this proposal, they
  occur within the first nested loop. They involve
  the repetition or "propagation" of the (1/n)
  value across all rows of the empirical copula
  matrix, Figure (4). It is crucial to note that
  Overlapping takes place at this stage, indicated
  by: "empCop[xi, SVID-1]" next:
  FOR j = pvcoo to (n)
   empCop[j , SVID] = (1.0/(Rows)) + empCop[j , SVID - 1]
  END
  # During this second nested loop, only the
  values in the row where the first loop did not
  proceed will be updated.
  FOR j = 1 to (pvcoo-1)
   empCop [j] [SVID] = empCop [j] [SVID] + empCop[j , SVID -1]
  END
END
# The computation of the Empirical Copula's
matrix has concluded. It is now time to return
this matrix to use it in the Bernstein copula
calculation.
return (empCop) }
```

2.3. A simple example

We utilize the same data-set comprising 20 elements to construct its empirical copula following the previously delineated pseudocode. For a visual representation, please refer to Figure 4.

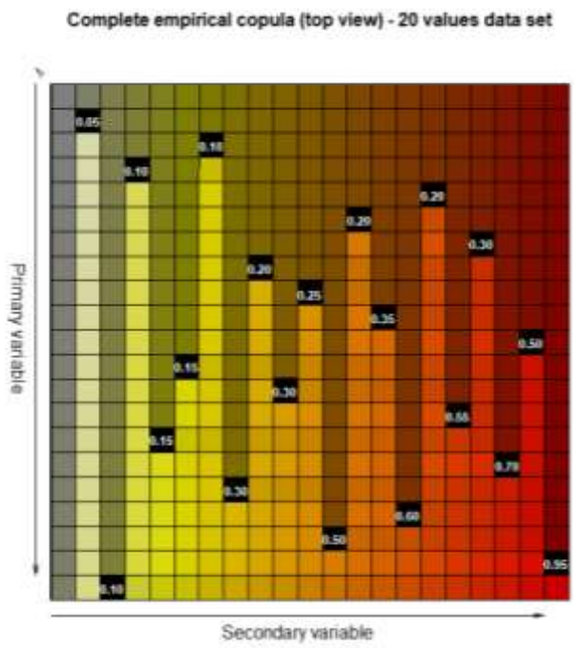


Figure 4 This graphical representation encapsulates the complete procedure for constructing the two-dimensional Empirical Copula's matrix, as introduced in this illustration. Propagation points are indicated by a white font color against a black background, while different colors correspond to increasing values resulting from overlapping processes. This graph was created in R-Project Software.

A bivariate empirical copula is represented by a two-dimensional unit square matrix, where each node contains a real numerical value, specifically denoted by the presence of an observation as *(1/n)*. These values are computed using the propagation and overlapping concepts, as previously explained. It is important to note that both of these concepts extend across all dimensions of the hypercube, as illustrated in Figure 2. In cases where two or more propagations overlap, their values are additive, as shown in Figure 3. Following the prescribed pseudo-code outlined in this proposal, this process continues until the occurrence of the final propagation, as depicted in Figure 4.

Within the unit square, each observation contributes *(1/n)*, which accumulates with others until the completion of the Empirical Copula's matrix. A visual representation of equation (1) is give out in Figure 4. Figure 5 provides a three-dimensional depiction, where discrete and incrementally ascending steps within the copula are visible for the same dataset, ultimately reaching the final value of *1.00*.

ISSN 2523-6865

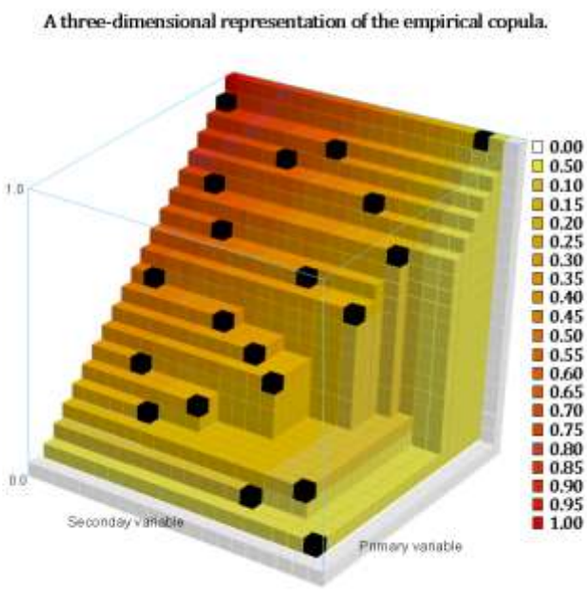


Figure 5 Here is the Empirical Copula, visualized in a three-dimensional format as depicted in Figure 4. It is noteworthy this representation does not involve continuous data; instead, it consists of a collection of discrete values meticulously arranged in a stepped pattern, distinguished by the presence of black points. It is essential to recognize this unique characteristic of the matrix when performing an analysis of it. This graph was created in 3ds max Software.

3. Practical observations and some results

Equation (4) outlines a rigorous procedural framework that imposes specific constraints requiring adherence. As illustrated in Figure (4), the majority of cells within the empirical copula matrix display an incremental, progression in their values, transitioning from *0* to *(observations/n)* in a systematic sequential manner, advancing from top to bottom and from left to right. It is crucial to emphasize that each cell within the empirical copula matrix is influenced by at least one observation that extends to the matrix's extremity, as visually depicted in Figure 1.

Consequently, we can deduce that the influence of an observation extends uniformly across the copula, propagating from its initial occurrence to the matrix's boundary. Moreover, in cases where multiple observations overlap, their values are summed, and these cumulative sums subsequently propagate throughout the copula matrix, as illustrated in Figures 3 and 4.

When computing equation (4) for a *2D* or *m-dimensional* empirical copula, it is essential to systematically traverse each cell within its matrix.

This process can be inherently time-intensive due to the presence of nested loops and the multitude of computations required within each cell. To provide a quantitative assessment of the computational complexity inherent in strict or standard implementations of equation (4) in the 2D scenario, we present a mathematical model in equation (7). Additionally, we introduce an alternative model in equation (8) to quantify the computational requirements associated with this proposed approach. These models serve as valuable tools for estimating execution time and the computational resources necessary to perform such calculations in practical applications.

$$\left[\left((2*n)+2\right)+\left((3*n)+(4*n)\right)\right]*(n+1)^2 \qquad (7)$$

$$n^2-(2*n)+1 \qquad (8)$$

Table (2) presents a comparative analysis of computations employing both of these methodologies, with a sample size of *n=20*.

Method	n	Calculations number
Standard	20	61,782
This proposal	20	361

Table 2 Comparison between the calculations performed using these two methods when the sample size is 20.

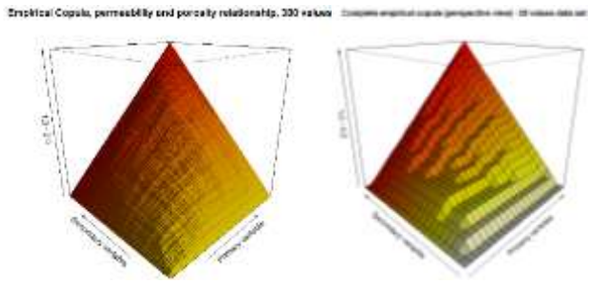
Table 3 illustrates case studies involving empirical copula computations using the Bernstein copula, encompassing both the standard method and the proposed approach. These cases begin with a simple example comprising a dataset of *20* values *(n=20)* and subsequently expand to include larger sample sizes, such as scenarios involving *380* and *3696* values for modeling the relationship between porosity and permeability, as well as *1081* values for investigating the sentiment-intensity relationship in the field of natural language processing.

Run	Method	n	Calc	t[s]	Regs 0.1 0.5 0.9 t[s]
1	Standard	20	61,782	0.05	0.89
1	This proposal	20	361	0.001	0.03
2	Standard	380	386,129,022	0.06	22.46
2	This proposal	380	143,641	0.0012	0.16
3	Standard	3696	353,613,561,842	0.67	11,140
3	This proposal	3696	13,653,025	0.0134	179.15
4	Standard	1081	8,858,870,672	0.19	394.46
4	This proposal	1081	1,166,400	0.0038	1.55

Table 3 A performance evaluation is conducted across four distinct computation scales denoted by the variable (*n*). Column *5* provides the time consumed expended in the computation of the empirical copula, while column *6* denotes the temporal expenditure involved in executing three quantile regressions utilizing Bernstein copulas

ISSN 2523-6865

In Figure 6, they are compared two empirical copulas of very different sizes, the left side (6a) depicts a 3D perspective view of an empirical copula, which may appear continuous but is actually not. It displays the complex relationship between permeability and porosity in heterogeneous porous media, as observed in a data set comprising *n=380* values; [XXVI]. To model this complex relationship, a non-parametric Bernstein copula was employed to fit the bivariate empirical distribution and simulate joint behavior through efficient computation of both the empirical and Bernstein copulas. Meanwhile, on the right side (6b), another empirical copula is shown from the same perspective, but it consists of only 20 values, resulting in a considerable reduction in size. The copula in (6b) exhibits a noticeable staggered pattern that is less prominent in (6b).



a. b.

Figure 6 a. Permeability and Porosity relationship, *380* values data-set; b. Random *20* values data-set. This graph was created in R-Project Software

A practical application of this proposal can be centered on modeling the relationship between sentiment and intensity in the field of natural language processing. Sentiment/analysis, widely employed in this domain, heavily relies on the sentiment-intensity dictionary, which is a valuable resource. Sentiment/analysis presents a significant challenge and is gaining momentum in text processing, particularly in the Spanish language, where resources for polarity classification tasks are limited.

The copula approach is utilized to model sentiment/intensity relationships, thereby enhancing the classification process. The relationship is visualized in Figure 7, presenting both a scatter plot of the dataset and its smoothed counterpart at different quantile values (utilizing the Bernstein copula). A comprehensive explanation of Bernstein's copula usage and implementation can be found in [XXVI].

Figure 7 effectively demonstrates how the copula captures complex dependencies in natural language processing, representing a spectrum of values ranging from the ***0.1*** to the ***0.9*** quantile, as well as a median regression (quantile ***0.5***). This illustrates the copula's ability to encompass all data points from the empirical model, spanning from the lowest to the highest values. A critical consideration in this analysis was the time and computational workload involved. For the empirical copula, this approach required approximately *50* to *60* times fewer calculations, while for the Bernstein copula, it resulted in a reduction of computational load by a factor of *254*, as shown in Table *2*.

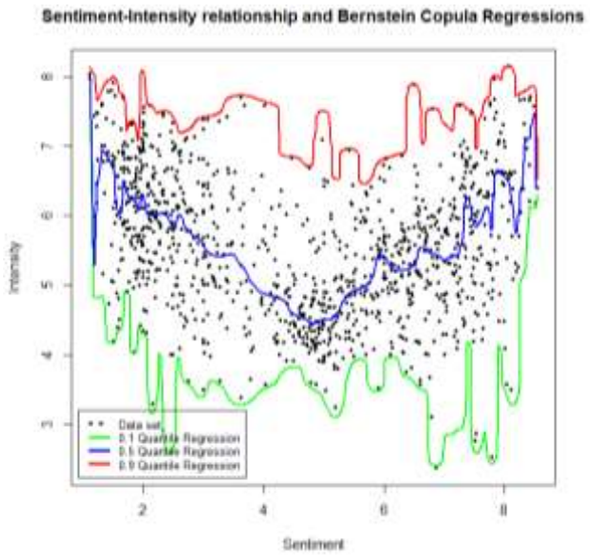


Figure 7 The copula demonstrates its ability to accurately model intricate relationships by depicting an inter-quantile range between quantile 0.1 and 0.9, along with a median regression ***quantile=0.5***. It is highlighted that this effectively captures all empirical information from the lowest to the highest value in the given empirical model.

4. Conclusions

The empirical copula is a discrete and finite mathematical-statistical function. However, its computational complexity tends to increase as the size of the data set grows, denoted by ***'n'***. This expansion results in a larger matrix size and necessitates a greater number of calculations per cell.

To tackle this challenge, we have introduced two novel concepts: propagation and overlapping. These concepts have demonstrated remarkable efficiency in expediting the computation of the empirical copula, leading to substantial time savings.

ISSN 2523-6865

Our approach can achieve computation speeds that are up to *60* times faster compared to standard methods. In the case of the Bernstein copula, which we have specifically investigated in this study, the speed enhancement can be even more impressive, exceeding *254* times.

While parallel computing has exhibited certain advantages in the context of the Bernstein copula, relying exclusively on brute force methods is not the most optimal approach. Consequently, our primary aim has been to devise a more efficient algorithm capable of swiftly executing the calculations associated with the empirical copula.

Through the implementation of this novel approach, we managed to decrease the number of computations necessary for the empirical copula by approximately *50* to *60* times. Regarding the Bernstein copula, the reduction in computational burden ranged from *30* to *254* times, as illustrated in Table 3 (last case).

This proposal introduces the potential integration of non-parametric copulas with Artificial Intelligence (AI) methodologies to enhance predictive capabilities. Existing research has predominantly concentrated on parametric copulas Ren *et al.*, 2022 [I]. However, non-parametric copulas excel in modeling nonlinear relationships, providing a more precise depiction of the dependencies among random variables. This surpasses the capabilities of conventional approaches and lays the foundation for improved predictive modeling.

5. Acknowledgement

Full financing of this research is acknowledged to INFOTEC.

References

[I] Ren, H., Li, Q., Wu, Q., Zhang, C., Dou, Z., and Chen, J. (2022). Joint forecasting of multi-energy loads for a university based on copula theory and improved LSTM network. Energy Reports, 8, 605-612. https://doi.org/10.1016/j.egyr.2022.05.208

[II] Liang, E., Zhu, H., Jin, X., and Stoica, I. (2019). Neural packet classification. In Proceedings of the ACM Special Interest Group on Data Communication (pp. 256-269). https://doi.org/10.1145/3341302.3342221

[III] Beyer, D., Löwe, S., and Wendler, P. (2019). Reliable benchmarking: requirements and solutions. International Journal on Software Tools for Technology Transfer, 21, 1-29. https://doi.org/10.1007/s10009-017-0469-y

[IV] Kent State University. (2023). https://shorturl.at/jlxDN

[V] Sklar, A. (2010). Fonctions de répartition à n dimensions et leurs marges [republication of mr0125600]. Ann. ISUP, 54(1-2), 3-6. https://hal.science/hal-04094463/document

[VI] Bouezmarni, T., El Ghouch, A., and Taamouti, A. (2011). Bernstein estimator for unbounded density copula. URI: http://hdl.handle.net/10016/14147

[VII] Mikosch, T. (2006). Copulas: Tales and facts--rejoinder. Extremes, 9(1), 55-62. https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=0ebf0c8cb7f79c11a26970b5156cdc505f1ace65

[VIII] Belalia, M., Bouezmarni, T., Lemyre, F. C., and Taamouti, A. (2017). Testing independence based on Bernstein empirical copula and copula density. Journal of Nonparametric Statistics, 29(2), 346-380. https://doi.org/10.1080/10485252.2017.1303063

[IX] Liebscher, E. (2009). Semiparametric estimation of the parameters of multivariate copulas. Kybernetika, 45(6), 972-991. https://dml.cz/bitstream/handle/10338.dmlcz/140022/Kybernetika_45-2009-6_7.pdf

[XI] Sancetta, A., and Satchell, S. (2004). The Bernstein copula and its applications to modeling and approximations of multivariate distributions. Econometric theory, 20(3), 535-562. https://doi.org/10.1017/S026646660420305X

[XII] Díaz-Viera, M. A., Vázquez-Ramírez, D., del Valle-García, R., Erdely, A., and Grana, D. (2020). Bernstein copula-based spatial cosimulation for petrophysical property prediction conditioned to elastic attributes. Journal of Petroleum Science and Engineering, 193, 107382. https://doi.org/10.1016/j.petrol.2020.107382

[XIII] Gudendorf, G., and Segers, J. (2010, May). Extreme-value copulas. In Copula Theory and Its Applications: Proceedings of the Workshop Held in Warsaw, 25-26 September 2009 (pp. 127-145). Berlin, Heidelberg: Springer Berlin Heidelberg. https://link.springer.com/chapter/10.1007/978-3-642-12465-5_6

[XIV] Beirlant, J., Goegebeur, Y., Segers, J., and Teugels, J. L. (2006). Statistics of extremes: theory and applications. John Wiley and Sons. https://shorturl.at/nqBV9

[XV] Lin, F., Peng, L., Xie, J., and Yang, J. (2018). Stochastic distortion and its transformed copula. Insurance: Mathematics and Economics, 79, 148-166. https://doi.org/10.1016/j.insmatheco.2018.01.003

[XVI] Einmahl, J. H., Ferreira, A., de Haan, L., Neves, C., and Zhou, C. (2022). Spatial dependence and space–time trend in extreme events. The Annals of Statistics, 50(1), 30-52. DOI: 10.1214/21-AOS2067 https://projecteuclid.org/journals/annals-of-statistics/volume-50/issue-1/Spatial-dependence-and-spacetime-trend-in-extreme-events/10.1214/21-AOS2067.short

[XVII] Embrechts, P., Wang, B., and Wang, R. (2015). Aggregation-robustness and model uncertainty of regulatory risk measures. Finance and Stochastics, 19, 763-790. https://doi.org/10.1007/s00780-015-0273-z

[XVIII] Nelsen, R. B., Quesada-Molina, J. J., Rodríguez-Lallena, J. A., and Ubeda-Flores, M. (2009). Kendall distribution functions and associative copulas. Fuzzy Sets and Systems, 160(1), 52-57. https://doi.org/10.1016/j.fss.2008.05.001

[XIX] Rémillard, B., Nasri, B., and Bouezmarni, T. (2017). On copula-based conditional quantile estimators. Statistics and Probability Letters, 128, 14-20. https://doi.org/10.1016/j.spl.2017.04.014

[XX] Chen, Q., Yu, C., and Li, Y. (2022). General strategies for modeling joint probability density function of wind speed, wind direction and wind attack angle. Journal of Wind Engineering and Industrial Aerodynamics, 225, 104985. https://doi.org/10.1016/j.jweia.2022.104985

[XXI] Dias, A., Salmon, M., & Adcock, C. (Eds.). (2013). Copulae and Multivariate Probability Distributions in Finance. Routledge. https://doi.org/10.4324/9781315871820

[XXII] Bhatti, M. I., and Do, H. Q. (2019). Recent development in copula and its applications to the energy, forestry and environmental sciences. International Journal of Hydrogen Energy, 44(36), 19453-19473. https://doi.org/10.1016/j.ijhydene.2019.06.015

[XXIII] Chevallier J. Uniform decomposition of probability measures: quantization, clustering and rate of convergence. Journal of Applied Probability. 2018; 55(4):1037-1045. https://doi.org/10.1017/jpr.2018.69

[XXIV] Deheuvels, P. (1979). La fonction de dépendance empirique et ses propriétés. Un test non paramétrique d'indépendance. Bulletins de l'Académie Royale de Belgique, 65(1), 274-292. https://www.persee.fr/doc/barb_0001-4141_1979_num_65_1_58521

[XXV] Perez, J. M., and Palacín, A. F. (1987). Estimating the quantile function by Bernstein polynomials. Computational Statistics and Data Analysis, 5(4), 391-397. https://doi.org/10.1016/0167-9473(87)90061-2

[XXVI] Erdely, A., and Diaz-Viera, M. (2010, May). Nonparametric and semiparametric bivariate modeling of petrophysical porosity-permeability dependence from well log data. In Copula Theory and Its Applications: Proceedings of the Workshop Held in Warsaw, 25-26 September 2009 (pp. 267-278). Berlin, Heidelberg: Springer Berlin Heidelberg. https://link.springer.com/chapter/10.1007/978-3-642-12465-5_13

ISSN 2523-6865