UNIVERSIDADE DE SÃO PAULO
ESCOLA DE ENGENHARIA DE SÃO CARLOS

Willian Darwin Junior

Modeling Copulas with Bayesian Networks

São Carlos

2021

Willian Darwin Junior

Modeling Copulas with Bayesian Networks

Tese apresentada à Escola de Engenharia de São Carlos da Universidade de São Paulo, para obtenção do título de Doutor em Ciências - Programa de Pós-Graduação em Engenharia Elétrica.

Área de concentração: Sistemas Dinâmicos

Advisor: Prof. Dr. Carlos Dias Maciel

VERSÃO CORRIGIDA

São Carlos

2021

Ficha catalográfica elaborada pela Biblioteca Prof. Dr. Sérgio Rodrigues Fontes da EESC/USP com os dados inseridos pelo(a) autor(a).

Darwin, Jr, Willian
D228m Modelagem de Cópulas por meio de Redes Bayesianas / Willian Darwin, Jr; orientador Carlos Dias Maciel. São Carlos, 2021.

Tese (Doutorado) - Programa de Pós-Graduação em Engenharia Elétrica e Área de Concentração em Sistemas Dinâmicos -- Escola de Engenharia de São Carlos da Universidade de São Paulo, 2021.

1. cópula. 2. rede bayesiana. 3. modelagem. 4. cópula empírica. 5. normalização não-linear. I. Título.

Eduardo Graziosi Silva - CRB - 8/8907

FOLHA DE JULGAMENTO

Candidato: Engenheiro **WILLIAN DARWIN JÚNIOR.**

Título da tese: “Modelagem de cópulas por meio de Redes Bayesianas”.

Data da defesa: 23/02/2021.

Comissão Julgadora **Resultado**

Prof. Associado **Carlos Dias Maciel** Aprovado
(Orientador)
(Escola de Engenharia de São Carlos – EESC/USP)

Prof. Titular **Jorge Alberto Achcar** Aprovado
(Faculdade de Medicina de Ribeirão Preto/FMRP-USP)

Prof. Dr. **Carlos Henrique Costa Ribeiro** Aprovado
(Instituto Tecnológico de Aeronáutica/ITA)

Dr. **Jorge Eduardo de Schoucair Jambeiro Filho** Aprovado
(Receita Federal do Brasil)

Prof. Titular **Alexandre Cláudio Botazzo Delbem** Aprovado
(Instituto de Ciências Matemáticas e de Computação/ICMC-USP)

Coordenador do Programa de Pós-Graduação em Engenharia Elétrica:
Prof. Associado **João Bosco Augusto London Junior**

Presidente da Comissão de Pós-Graduação:
Prof. Titular **Murilo Araujo Romero**

This Thesis is dedicated to my late parents Willian and Rosa,
to my wife Giselda and to my daughters Ana Carolina and Flávia.

ACKNOWLEDGEMENTS

My most profound love and gratitude to my parents, Willian and Rosa, who were gone during my research, but whose light is still here so that I can keep reflecting them in my attitudes and character.

That same amount of love and gratitude for my girls, Giselda, my wife, and Ana Carolina and Flavia, my daughters, for being so supportive and faithful in my abilities even when I was in doubt and weak.

To my sisters, Ana Cristina and Patricia, for their affection and comprehension for not being as present as I wished along those years of dedication throughout distance.

To Prof. Carlos Dias Maciel, professor and also a friend, a very especial thank for his guidance, patience, wisdom, and technical support along all that extremely gratifying period.

Many thanks to all friends who still are or were in LPS-EESC-USP, in special to Daniel, Fernando, Henrique, Jonas, Jordao, Luis, Mateus, Matheus, Michel, Pedro, Rafael, Rodrigo, Talysson, Victor, Vitor Barth, and Vitor Ribeiro.

Thank very much to all professors and professionals at EESC-USP, especially Marisa and Daniel, who always were there to help us in whatever we needed.

Thanks to all friends and colleagues in RFB, in special to Ana Paula Gervasio, Carlos A. Souza, Rafael Lima, Tatiana Carvalho, Carlos Riboldi, Sergio Mazzetti, Cláudia Andrade, Daniela Caldas, and Ísis Karoline, for the friendship, support and comprehension during my atempt to balance work and research. A very especial thank to Jorge Jambeiro, who also accepted my invitation to join us in this sub-cycle closing event.

Finally, my thanks to everyone who directly or indirectly contributed to this research.

"With caution judge of probability.
Things deemed unlikely, e'en impossible,
experience oft hath proved to be true."
Attributed to William Shakespeare

ABSTRACT

DARWIN JUNIOR, W. **Modeling Copulas with Bayesian Networks**. 2021. 221p. Tese (Doutorado) - Escola de Engenharia de São Carlos, Universidade de São Paulo, São Carlos, 2021.

Bayesian networks are extensively studied in machine learning and there is a significant growing interest on copulas in scientific literature beyond Statistics, but it is still uncommon to join those conceptual artifacts. Our research proposes an initial stage approach for combining those concepts in probabilistic modeling by splitting the model in two coupled elements, individual marginal distributions and a copula, reserving the Bayesian network modeling only to the copula portion and liberating the marginal distributions modeling to be done by any chosen strategy according to the data, without interfering in the dependence modeling. We compared two different marginal modeling techniques for the first stage of the modeling: a standard Bayesian inference using Mont Carlo Markov chain (MCMC) and a sample reducing. The results showed good performance in both cases in the sense of preserving the same structure scoring tendency as the traditional approach for discrete Bayesian networks and pointed to the viability of modeling copulas using Bayesian networks for samples with enough number of instances, which was the premise of this research. For helping in the data analysis stage of the methodology, a general data analysis and visualization software tool, designated LPSCopModel, was developed for providing variables description and concordance indexes, MCMC parametric distribution fitting and an empirical copula profile as a first glance at the dependence structure.

Keywords: copula, Bayesian network, sample reducing, empirical copula, MCMC, Bayesian inference, non-linear normalization.

RESUMO

DARWIN JUNIOR, W. **Modelagem de Cópulas por meio de Redes Bayesianas**. 2021. 221p. Tese (Doutorado) - Escola de Engenharia de São Carlos, Universidade de São Paulo, São Carlos, 2021.

Redes bayesianas vem sendo extensivamente estudadas em Aprendizado de Máquina e há um significativo crescimento no interesse por cópulas na literartura científica além da Estatística, porém ainda é rara a junção desses dois artefatos conceituais. Nossa pesquisa propõe uma abordagem em estágio inicial para combinar esses dois conceitos de modelagem probabilística pela separação do modelo em dois elementos acoplados, as distribuições marginais individuais e uma cópula, reservando a modelagem por redes bayesianas apenas para a parte relativa à cópula e liberando a modelagem das distribuições marginais para ser feita por qualquer estratégia escolhida conforme o‘s dados, sem que isso interfira na modelagem das dependências. Nós comparamos duas técnicas para a modelagem das distribuições marginais para o primeiro estágio da modelagem: inferência bayesiana padrão usando Monte Carlo Markov chain (MCMC) e redução amostral (”sample reducing”). Os resultados mostraram um bom desempenho em ambos os casos no sentido de preservar a mesma tendência para a avaliação de estruturas que apresentada pela abordagem tradicional de redes bayesianas discretas e apontou para a viabilidade de modelar cópulas usando redes bayesianas para amostras com número suficiente de instâncias, que foi uma das premissas dessa pesquisa. Para auxiliar no estágio de análise dos dados, uma aplicação de análise e visualização geral de dados, denominada LPSCopModel, foi desenvolvida para prover uma descrição das variáveis e índices de concordância, um ajuste paramétrico de distribuições usando MCMC e um primeiro vislumbre da estrutura de dependências a partir de uma cópula empírica.

Palavras-chave: cópula, rede bayesiana, redução amostral, cópula empírica, MCMC, inferência bayesiana, normalização não-linear.

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

BDeu	Likelihood-equivalence Bayesian Dirichlet uniform score function
BIC	Bayesian Information Criterion score function
BN	Bayesian Network
BN-C	Bayesian Networks copula modeling
CBN	Copula Bayesian Network
CPT	Conditional Probability Table
CSV	comma-separated values file format
DAG	Direct Acyclic Graph
DBF	dBase database file format
HCBN	Copula Network Classifier
IC	Indictive Causation Algorithm
K2	Bayesian Dirichlet K2 score function
LpsCopModel	Software tool developed for multivariate analysis and copula modeling
MCMC	Monte Carlo Markov Chain
MF	Marginal Distribution Fitting
PCC	Pair Copula Construction
PDF	Portable Document Format
SR	Sample Reducing
USP	Universidade de São Paulo
USPSC	Campus USP de São Carlos

CONTENTS

1 INTRODUCTION

This text is the result of research in the field of **probabilistic models**. Probabilistic models are extensively studied in the scientific literature from many different perspectives. For our purpose of probabilistic modeling multivariate data of many possible types, model generality was a strong constraint, and the copula theory provides that requisite. In parallel, we work in a probabilistic modeling research group with a specialization in Bayesian networks, so it would be natural to follow a path where Bayesian networks would play a key role. Besides, we have a great interest in the associations and dependencies among features because they are essential to the real problems we are motivated by, and both copulas and Bayesian networks pay special attention to them. The fourth important ingredient to this recipe is the multidisciplinary intrinsic character of our research, which is intended to be applied to many areas, like tax administration, energy distribution, biological and medical researches. That mixture of factors led us to concentrate efforts on studying possibilities of a combination between copula theory and Bayesian networks.

1.1 Research Subject and Contribution

Our research proposes a non-linear normalization approach in probabilistic multivariate modeling by Bayesian networks, combining those concepts by splitting the model in two coupled elements, individual marginal distributions and a Bayesian network modeled copula.

One great advantage of this method is the possibility of isolating any erratic random variable behavior from the further random vector dependence analysis, as a kind of "noise filtering" first stage for the modeling. As another benefit, each random variable individual behavior can be best considered and modeled in an independent previous stage which can count in every sample value, so any one-dimensional consolidated modeling technique can be used in its full potential.

We compared two different marginal modeling techniques for the first stage of the modeling: a standard Bayesian inference using Mont Carlo Markov chain (MCMC) and a so-called sample reducing based on fitting an empirical distribution with some calibration for avoiding sample overfitting. The results showed good performance in both cases in the sense of preserving the same structure scoring tendency as the traditional approach for discrete Bayesian networks and pointed out to the viability of modeling copulas using Bayesian networks.

For helping in the data analysis stage of the methodology, a general data analysis and visualization software tool, designated LPSCopModel, was developed for providing variables description and concordance indexes, MCMC parametric distribution fitting, and an empirical copula profile as a first glance at the dependence structure. This software was first used for helping in acquiring the descriptive analysis which was further taken as one of the central elements in our publication on healthcare issues on elderly femur fractures in (PETERLE *et al.*, 2020).

Finally, the non-linear normalization technique here proposed, although specifically used in Bayesian network modeling, may be also useful for many other Machine Learning or Artificial Intelligence approaches given its generality and applicability to multivariate modeling with no prior dimensional constraint. Therefore, we encourage its adoption in any situation our research peers find feasible.

1.2 Research Themes: Bayesian networks and copulas

The main themes within our research are Bayesian networks and copulas.

A Bayesian network (BN) (PEARL, 1988) (NEAPOLITAN, 2003) (KOLLER; FRIEDMAN, 2009) is a graphical representation of a joint probability distribution that encodes dependencies in a graphical structure and a corresponding set of conditional distributions parameters. That graphical element adopts the structure of a directed acyclic graph (DAG) whose nodes and edges stand for the random variables and their corresponding probabilistic dependencies, respectively, while the quantitative element is the set of all conditional probability distributions (CPD) attached to each node conditioned on its parents in the DAG, the product of all those conditional distributions providing the complete joint distribution model.

Just as for joint distributions, modeling BNs with a large number of variables remains challenging (ZHAO *et al.*, 2017), mainly because the number of candidate networks (DAGs) is a super-exponential function of the number of nodes (ROBINSON, 1977), as though probabilistic inference using BNs is also NP-hard (in a general way) (COOPER, 1990). Consequently, identifying the optimal directed topology is NP-hard (CHICKERING; GEIGER; HECKERMAN, 1994), leading researchers to adopt sub-optimal strategies (VILLANUEVA; MACIEL, 2011) (CHEN; DARWICHE; CHOI, 2018) which focus on inference tasks efficiency in prejudice of the real structure of influences among variables. In this context, algorithms used to learn BN structure from data adopt two main approaches

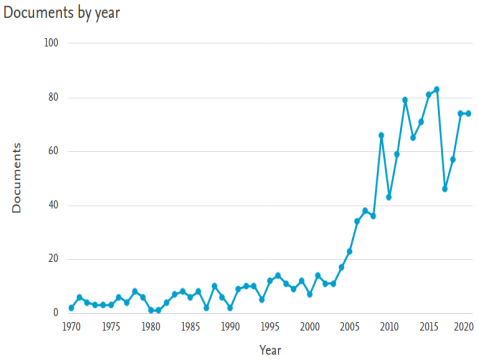
(or a combination of them) (SCUTARI; GRAAFLAND; GUTIÉRREZ, 2019): score-based algorithms that are optimization algorithms which search for a network having a favorable trade-off between data fitness and structural complexity, measured by a score function, and constrained-based algorithms that use conditional independence tests and derive from the inductive causation (IC) algorithm proposed in Verma e Pearl (1991). Hence, our proposed methodology must be based on one, or even both, of those two elements: scores and conditional independence.

By its turn, a copula is a probabilistic model responsible to describe exclusively the associations within a given phenomenon random variables set, without concerning each variable individual behavior. The viability and utility of that kind of modeling structures derive from Sklar (1959)'s theorem which sustains that every joint distribution can be decoupled into the set of the random variables marginal distributions and a function taking them as inputs and describing their dependence relationship. Therefore, the copula is the ultimate conceptual probabilistic structure for completely modeling the dependence structure of any given model, and this is the clinch to associate copulas and Bayesian network structures.

1.3 Literature Review

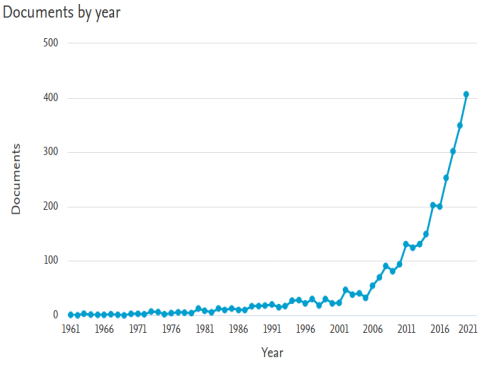
It is important to remark that both themes are very contemporary in many knowledge areas, from theoretical to application fields, like Mathematics, Economics, Finance, Engineering, Computer Science, Environmental Sciences, Social Sciences, Healthcare, etc. Only in the past two years (from 2019 to now), there are a total of $2,730$ documents published in Scopus database on the subject of copulas, some already with over 50 citations. There are examples in asymmetric tail dependence analysis (stock market case) (ECHAUST, 2021), terrestrial vegetation vulnerability (JI *et al.*, 2020), intensity and duration of cold episodes (CHATRABGOUN *et al.*, 2020), hydrological risk assessment (LIU *et al.*, 2020), and software packages (JIANG; CAO; DENG, 2020) (YUAN; HU, 2019). For Bayesian networks, the scenery is even stronger, as in that same database for the same period, we have found $9,374$ published documents mentioning the theme, some with already almost 150 citations. Publications include, for example, flood resilience (SEN; DUTTA; LASKAR, 2021), product development in supply chain (GOSWAMI; DAULTANI; DE, 2021), systems life estimation (subsea pipeline case) (CAI *et al.*, 2020), z-network (JIANG; CAO; DENG, 2020), structure learning algorithms (SCUTARI; GRAAFLAND; GUTIÉRREZ, 2019), and atrial fibrillation strategies (medical application) (LOPES *et al.*, 2019).

Deepening this literature review, we observe that copula theory is an ascending topic in scientific literature today, as so are Bayesian networks, and the same occurs for both themes together in the same research, as we can see in Figure 1.3. It can be noticed that, although the relevance is totally out of questioning, the simultaneous presence of both themes together occurs in a still relatively low number of texts. Therefore, the combination of copulas and Bayesian networks is still a fertile field for research in present days.



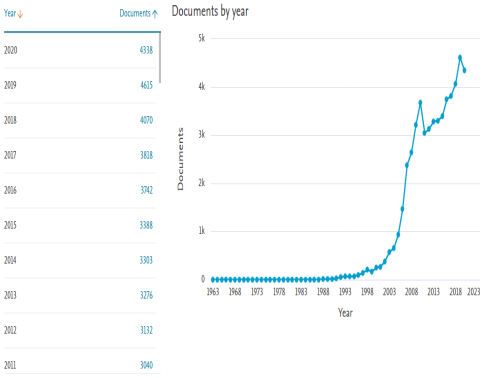
(a) Citations of Sklar works

Fonte: Scopus



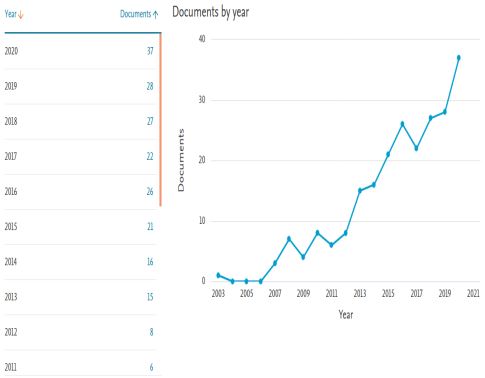
(b) Presence of term "copula"

Fonte: Scopus



(c) Presence of term "Bayesian networks"

Fonte: Scopus



(d) Presence of both terms together

Fonte: Scopus

Figure 1 – The relevance of copula theory (also represented by its main result in Sklar's work), Bayesian networks, and both together in the same paper in scientific literature today. Searching considered the presence of key terms in the title, abstract, or keywords. Both graphics are from Scopus scientific publications database and show how interest in copula theory and Bayesian networks manifest an exponential growth in the last decade, although there are still low numbers for publications dealing with both themes.

If now we focus on what exactly scientists have been working on in the copula field in the last decades and what they are up to now, we will see a great job from pioneers along many years followed by a more collective effort in recent years. After Sklar (1959)'s already mentioned remarkable breakthrough, Deheuvels (1979) came up with the empirical copula definition, bringing a very handy tool for starting analyzing multivariate

dependence structures, which we even used along with our research, and Schweizer e Wolff (1981) extended the field a bit more by bringing some light to invariant properties of joint distributions depending only on the copula and presented some non-parametric measures of dependence.

More than two decades further, Owzar e Sen (2003) published a review on copulas theory, describing copulas main families, and reporting copula applications so far. Researchgate platform registers an unpublished work (BONDÁR *et al.*, 2005) which is one good example of a growing number of studies applying copula in weather analysis, claiming that standard regression analysis would be inappropriate for tail-dependent structures usually found in weather phenomena, while, contemporaneously, Fermanian (2005) published goodness-of-fit tests for copulas and Schmidt (2006) wrote an interesting overview chapter on copulas for a book. On books, a remarkable job is materialized on Nelsen (2006), whose manuscript was by far the most referential base text in our research.

Choi, Noh e Du (2007) compares Rosenblatt and Nataf transformations efficiency in reliability-based design optimization (RBDO) and, as a sequel, Noh, Choi e Du (2007) presents a very interesting study proposing a new transformation for RBDO using non-Gaussian copula associated with Rosenblatt transformation for obtaining the joint cumulative density function.

Bouzebda (2012) brings up new light on empirical copulas analyzing the strong approximation character of bootstrapped empirical copula processes, and further this same author proposes another approximation using Poisson bridges (BOUZEBDA, 2017); another perspective on empirical copulas is given by Segers, Sibuya e Tsukahara (2017) with a beta empirical copula and the study of conditions for a Bernstein polynomial to be a copula. Copula theory is also used for estimators for multivariate density using mixtures of normal distributions to model the joint dependence as in Tran *et al.* (2014), for graphical goodness-of-fit tests of dependence as in Mächler (2011), or for assessing Granger causality by a Bernstein approximation of empirical conditional copula densities as in Hu e Liang (2014).

Diks *et al.* (2014) compares different predictive copulas using Kullback-Leibler information criterion and applies those results in a government bounds case. About the hard-dealing elements of singular copula component and its support, Durante, Fernández-Sánchez e Trutschnig (2015) brings some existence results and constructions that can help in real-life discontinuous problems.

In terms of applications, copula theory is very frequently applied to financial markets (ECHAUST, 2021), but applications also have grown in other areas, as healthcare (JOVANOVIC *et al.*, 2018) (ACHCAR; MARTINEZ EDSON ZANGIACOMI ANDTOVAR CUEVAS, 2016), hydrology (LIU *et al.*, 2020) and weather forecasting (CHATRABGOUN *et al.*, 2020).

One of the most remarkable researches in combining copulas and Bayesian networks may be Elidan (2010a) because it proposes a new model constructed by integrating elements of those two modeling fields by defining a composite structure named Copula Based Network (CBN). A Copula Bayesian Network (CBN) is a triplet $C = (\mathcal{G}, \theta_c, \theta_f)$ that encodes the joint density $f_X(x)$, where θ_c is a set of local copula densities functions $c_i(F(x_i), F(pa_{ik}))$ that are associated with the nodes of $\mathcal{G}$ that have at least one parent, and θ_f is the set of parameters representing the marginal densities $f(x_i)$. More two publications followed that first one extending the original research: one (ELIDAN, 2010b) applying CBN to missing data situations and other (ELIDAN, 2012) that creates the Copula Network Classifier (CNC) as a modeling network with a structure similar to CBN but now with an additional parameter proper for dealing with discrete random variables, considering original CBN was unable to model them due to its copula limitation to continuous variables. Further, Elidan (2013) publishes a review as a book section to advocate the gap between copulas and Bayesian networks, as "the fields of machine learning in general and probabilistic graphical models, in particular, have been ignorant of the framework of copulas". A generalization of the CBN concept was proposed in Karra e Mili (2016) introducing the Hybrid CBN (HCBN) with both continuous and discrete random variables.

Another similar and very interesting research line in associating copulas and BNs was conducted by Bauer e Czado (2016), Bauer, Czado e Klein (2012), Kurowicka (2012), Hanea, Kurowicka e Cooke (2006), and Bedford e Cooke (2002) with their proposal of a new type of multivariate statistical model specified by a directed acyclic graph (DAG) featuring a specific factorization based on vine and pair-copula constructions (PCCs) and hence involving only univariate distributions and bivariate copulas. The vine copula further became a trending topic concerning multivariate copulas, and even a Python package, called "pyvine", was recently made available by Yuan e Hu (2019). Other related contributions were made, for example, by Zhang e Shi (2017) with an application of CBN in Biology and genomic data and by Zilko *et al.* (2015) applying a mixed discrete and continuous CBN in railway disruption lengths forecast.

1.4 Methodology

All that considered, we can now present a general description of our proposed methodology for joining the Bayesian network and copula in a single model. First of all, although we have always been concerned about all the conceptual and theoretical formal Statistics background on probability distributions and copulas, our main perspective is from the machine learning point-of-view, therefore we do not intend to suggest a Bayesian network element for copula theory but instead considering basic copula concepts as tools for Bayesian networks modeling. Therefore, we propose a methodology consisting of two stages for modeling Bayesian networks: first we proceed to the probabilistic decoupling of all random variables individual behaviors from dependencies using copula main result and then we model only the dependence relationship among those variables via Bayesian network.

The first stage, which consists basically of modeling univariate random variables, can be implemented in many ways: traditional Statistics parametric fitting, Bayesian parametric fitting, empirical fitting, sample transformation, and so on. From a Statistics perspective, this can be done by marginal distribution fitting, i.e., searching a population probability distribution that is likely to give such a sample as the one to be modeled, which we are implementing by a Monte Carlo Markov chain (MCMC) Bayesian inference approach. On the other hand, assuming a machine learning perspective, we can implement the first stage by applying an adequate non-linear transformation or normalization to our sample (instead of trying to identify its population probability distribution), sometimes called sample reducing in the literature, but also known as plotting position (CUNNANE, 1979), and consequently filtering individual behaviors as noise for further dependence analysis by Bayesian network modeling. This is done by applying to the sample a transformation equivalent to its (unidentified) population probability distribution.

The second stage is the same for both and consists of modeling the Bayesian network for the non-linear normalized sample, alternatively by marginal fitting or sample reducing methods. As the searching for an optimal network structure for the model is still an open problem, we will be satisfied with analyzing a set of candidates and obtaining coherent results in their relative scoring, that coherence being detailed according to each individual dataset and its multivariate previous analysis.

To help in the pre-processing, data analyzing, and parametric and non-parametric distribution fitting, we have developed a software application with visualization facilities that goes from data filtering and slicing to marginal distribution fitting and empirical copula overview. This tool is more detailed throughout this text.

It must be noticed that, although having many things in common with the before mentioned copula BNs initiatives already registered in the literature (CBN and PCC based), the proposed methodology for modeling copula with Bayesian networks is different from both CBN and PCC approaches in a fundamental point: while in those models conditional copulas are used as building blocks of the network structure, in our proposed BN copula modeling the joint distribution copula is integrally taken as the raw material for the BN modeling.

For contextualization purposes in terms of the theoretical progression we adopted, we must register that this research followed a path strongly induced by the real problem it intended to approach, starting with the search for a sensible data masking technique with special attention to rare events, both related to demands in the tax administration business. From there it became inevitable to cross ways with risk analysis, while a healthcare prediction issue led us back to Bayesian networks. While studying risk analysis and rare events (STRAUB; PAPAIOANNOU; BETZ, 2016), the concept of the Rosenblatt transformation appeared as a technique for transforming an arbitrary form random vector from its original space to the space of independent standard normal random variables, or, in the words of Rosenblatt himself, the Rosenblatt transform original version is "a simple transformation of an absolutely continuous k-variate distribution $F(x_1, ..., x_k)$ into the uniform distribution on the k-dimensional hypercube" (ROSENBLATT, 1952). Rosenblatt even shows the analytical equation for the transform in the case of a bivariate normal joint distribution.

Further studies in risk analysis foundations ended up in our first contact with the even more general concept of copula. Vose (2009) offers a broad view with many quantitative risk analysis methods and techniques and yet most pages in the dependence modeling chapter are devoted to copulas. While the Rosenblatt transformation proposes a total space shift for the joint distribution from an arbitrary distribution class into a uniform distribution, the copula theory states that any joint distribution can be split into the composition of a dependence function, called copula, and its marginal distributions (SKLAR, 1959). This outstanding theoretical result implies that even the most erratic random variables individual behavior can be isolated from their dependence analysis, and also that random vectors dependence patterns can be compared no matter their individual natures.

Another solid step in this upward road, which proved to be mandatory as an important tool for a good comprehension of copulas, was the study of Probability from a more formal perspective than that of our basic Engineering courses on Statistics. We talk

here about probabilistic models defined on sets, algebras, and measure spaces, and designed upon measure theory elements such as Borel sets and Lebesgue measure (SHIRYAEV, 1996). Some Mathematics foundations were also helpful, specially real function analysis and measure theory (LIMA, 1976) (LIMA, 1981). All that ground was essential for a reasonable understanding of concepts like the support of a copula, probability mass distribution, and decomposition of copulas into absolutely continuous and singular components. They are also relevant for theory distinctions in dealing with continuous and discrete random variables.

1.5 Text Structure

We adopted a strategy of keeping in the main text only the essential elements to give it both concision and coherence, prioritizing a friendly and light reading for it to be more easily accessible to people from different areas of knowledge. Therefore, those materials which would be interesting for the more acute reader in one or more aspects of the context were placed in a proper appendix. That appendix also give support to reproductibility and tools usage description purposes.

This text is structured in chapters, each one treating a different context. Introduction stands for presenting motivation, context, relevance, and historical backgrounds of this research and specific knowledge areas it is founded upon. Theory Foundations presents the theoretical concepts and context on which these text proposals were built upon. Those proposals are detailed in the Proposed Methodology chapter, which also describes how those proposals were implemented and which software and hardware were used to build models, make simulations, and extract results from them. Then the main results are presented and discussed in Results and Discussion. Finally, the Conclusion chapter represents a closing to the reading by offering an overview of the research, its main results, and its possible contributions to the specific area of knowledge it is inserted in or for other areas by an interdisciplinary perspective, in parallel with some suggestions for further researches. The last sections of the manuscript are the References with the bibliography cited along with the text, and the already mentioned Appendix.

2 THEORETICAL FOUNDATIONS

2.1 Basic Concepts

To establish the theoretical grounds upon which this study will settle its analysis and results, first we have to start from basic concepts. For matters of formality which will be important further, we adopt here the more formal approach of probability theory based on set and measure theories where a probability model or a probability space is defined upon three elements: a sample space Ω, a σ-algebra $\mathcal{F}$ of subsets of Ω, and a probability P on $\mathcal{F}$ (SHIRYAEV, 1996). In this chapter we will expressly mention only core results for keeping reading easier, but the interested reader can have a deeper view in Appendix "Theoretical Reference" or directly consulting the corresponding bibliography.

Definition 1 ***Sample space*** *(Ω) is the set of all possible elementary outcomes ω that might be observed from an experiment.*

Definition 2 *An* ***event*** *is any subset $A \subset \Omega$.*

Definition 3 *A system $\mathcal{F}$ of subsets of Ω is a* ***σ-algebra*** *if:*

(a) $\Omega \in \mathcal{F}$,

(b) $A_n \in \mathcal{F} \implies \cup A_n \in \mathcal{F}, \cap A_n \in \mathcal{F}$,

(c) $A \in \mathcal{F} \implies \bar{A} \in \mathcal{F}$

Definition 4 *The pair of a space Ω together with a σ-algebra $\mathcal{F}$ of its subsets is a* ***measurable space*** $(\Omega, \mathcal{F})$.

Definition 5 *Let $(\Omega, \mathcal{F})$ be a measurable space. A set function $P = P(A)$, $A \in \mathcal{F}$, taking values in $[0, \infty]$, with $P(\Omega) = 1$, is a* ***probability measure*** *or a* ***probability*** *if, for all pairwise disjoint subsets $A_1, A_2, ...$ of $\mathcal{F}$ with $\sum A_n \in \mathcal{F}$:*

$$P\left(\sum_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} P(A_n) \tag{2.1}$$

Definition 6 *An ordered triple $(\Omega, \mathcal{F}, P)$ is called a* ***probabilistic model*** *or a* ***probabilistic space*** *when:*

(a) Ω *is a set of points* ω,

(b) $\mathcal{F}$ is a σ-algebra of subsets of Ω,

(c) P is a probability on $\mathcal{F}$

Here Ω is the sample space or space of elementary events, the sets A in $\mathcal{F}$ are events, and $P(A)$ is the probability of the event A.

Definition 7 *A **distribution function** $F = F(x)$ on the real line R is a function satisfying:*

1. *$F(x)$ is nondecreasing;*
2. *$F(-\infty) = \lim_{x \to -\infty} F(x) = 0$;*
3. *$F(+\infty) = \lim_{x \to +\infty} F(x) = 1$;*
4. *$F(x)$ is continuous on the right and has a limit on the left at each $x \in R$.*

Definition 8 *The **mean** or **expected value** of a random variable X is:*

1. *if X is discrete: $\mu_X = E[X] = \sum_{x \in R_X} x.p_x(x)$*
2. *if X is continuous: $\mu_X = E[X] = \int_{-\infty}^{+\infty} x.f_X(x)dx$*

Definition 9 *The **variance** of a random variable X is $\sigma_X^2 = E[(X - E[X])^2]$ and the **standard deviation** is $\sigma_X = \sqrt{\sigma_X^2}$.*

All those concepts can be extended to consider a set of variables instead of a single one (LARSON, 1982):

Definition 10 *(LARSON, 1982) Let $\mathbf{X} = (X_1, X_2, ..., X_n)$ be a rule associating an n-tuple with each element ω of a sample space S. Then $\mathbf{X}$ is called an **n-dimensional random vector**. Probability function $p_{X_1,..,X_n}(x_1, ..., x_n)$ and distribution function $F_{X_1,..,X_n}(x_1, ..., x_n)$ of $\mathbf{X}$ are also called **joint probability function** and **joint distribution function** in this multivariate case.*

Definition 11 *(LARSON, 1982) The **joint distribution function ($F_X(x)$)** for a random vector X is a function which gives the value of $P(X_1 \leq x_1, ...X_n \leq x_n)$ for any real vector x.*

Definition 12 *(LARSON, 1982) (adapted) Let* $\mathbf{X} = (X_1, X_2, ..., X_n)$ *be an n-dimensional random vector. Then the* ***marginal probability function*** *for* X_k, $k = 1, ..., n$, *is:*

$$p_{X_k}(x_k) = P(X_k = x_k) = \sum_{x_{i_1}} \sum_{x_{i_2}} ... \sum_{x_{i_{n-1}}} p_{X_1,..,X_n}(x_{i_1}, ..., x_{i_{k-1}}, x_k, x_{i_k}, ..., x_{i_{n-1}}),$$
$$i_j = 1, ..., k-1, k+1, ..., n \tag{2.2}$$

Definition 13 *(LARSON, 1982) The* ***marginal distribution function*** *(*$F_{X_i}(x)$*) for a random variable* X_i *of a random vector* $\mathbf{X}$ *is a function which gives the value of* $P(X_i \leq x_i)$ *for any real value* x_i *where* P *is the probability function for* $\mathbf{X}$*.*

Definition 14 *(LARSON, 1982) Let* $\mathbf{X} = (X_1, X_2, ..., X_n)$ *be an n-dimensional discrete random vector with probability function* $p_{X_1,..,X_n}(x_1, ..., x_n)$ *. Then the* ***conditional probability function*** *for* X_k, $k = 1, ..., n$, *, given* $X_k = x$, *is:*

$$p_{X_{i_1},...,X_{i_{k-1}},X_{i_k},...,X_{i_{n-1}}|X_k}(x_{i_1}, ..., x_{i_{k-1}}, x_{i_k}, ..., x_{i_{n-1}}|x) =$$
$$\frac{p_{X_1,..,X_n}(x_{i_1}, ..., x_{i_{k-1}}, x_k, x_{i_k}, ..., x_{i_{n-1}})}{p_{X_k}(x)},$$
$$i_j = 1, ..., k-1, k+1, ..., n \tag{2.3}$$

Definition 15 *(LARSON, 1982) Let* $\mathbf{X} = (X_1, X_2, ..., X_n)$ *be an n-dimensional continuous random vector with density function* $f_{X_1,..,X_n}(x_1, ..., x_n)$ *. Then the* ***conditional probability function*** *for* X_k, $k = 1, ..., n$, *given* $X_k = x$, *is:*

$$f_{X_{i_1},...,X_{i_{k-1}},X_{i_k},...,X_{i_{n-1}}|X_k}(x_{i_1}, ..., x_{i_{k-1}}, x_{i_k}, ..., x_{i_{n-1}}|x) =$$
$$\frac{f_{X_1,..,X_n}(x_{i_1}, ..., x_{i_{k-1}}, x_k, x_{i_k}, ..., x_{i_{n-1}})}{f_{X_k}(x)},$$
$$i_j = 1, ..., k-1, k+1, ..., n \tag{2.4}$$

Definition 16 *(LARSON, 1982) The random variables in the vector* $\mathbf{X} = (X_1, X_2, ..., X_n)$ *are* ***independent*** *if and only if,* $\forall x_1, x2, ..., x_n$*:*

- *(a) if* X *is discrete,* $p_{X_1,X_2,...,X_n}(x_1, x_2, ..., x_n) = p_{X_1}(x_1).p_{X_2}(x_2)...p_{X_n}(x_n)$

- *(b) if* X *is continuous,* $f_{X_1,X_2,...,X_n}(x_1, x_2, ..., x_n) = f_{X_1}(x_1).f_{X_2}(x_2)...f_{X_n}(x_n)$

2.2 Marginal Distribution Fitting

Identifying the marginal distributions from a joint distribution is crucial when dealing with copulas. Hence, we will first review some methods of doing so.

2.2.1 Empirical Distribution Fitting

The empirical fitting (VAART, 1998) is an unsophisticated but practical method for obtaining a marginal distribution approximation directly from the sample:

$$F_k^n(x) = \frac{1}{n}.\sum_{i=1}^{n} \mathbf{1}(X_k^i \leq x) \tag{2.5}$$

where **1** is the **indicator function** which is 1 if the argument expression is true and 0 otherwise.

Although empirical fitting has its limitations in terms of performance, that technique was chosen for its simplicity and methodological coherence with the empirical copula to be further described in this text, considering the immediate objective here of building the complete modeling methodology from real data to simulated sample in a simple version. For future development in this research it is within consideration to apply more consistent fitting techniques, like parametric distribution fitting, Bayesian inference, sum of distributions, kernel density estimation and others present in literature.

2.2.2 Bayesian Inference based on MCMC

Another more sophisticated method for fitting a univariate distribution to a sample is by using Bayesian inference and sampling that distribution by an asymptotic method. For that we are going to need to dive into three theoretical concepts: the Monte Carlo method, the Markov Chain process and sampling methods such as Gibbs' or No-U-Turn-Sampler (NUTS). Nevertheless, those three theoretic tools description are reserved to the corresponding Appendix, for keeping this text concise.

Basically, Monte Carlo Markov Chain (MCMC) algorithms are based in assuming a group of conditional distributions whose composition results in the joint distribution of all the random variables involved, there included both independent variables, covariates and distribution parameters, and then using a sampling methodology for walking in steps on a Markov chain until enough convergence is achieved for some parameters, each step consisting of sampling parameters from its marginal distributions through conditional distributions and obtaining a new posterior joint distribution from priors and likelihood at the given sampled parameters (GELMAN; RUBIN, 1992) (HAND, 2007) (ANDRIEU *et al.*, 2003) (ZHAO; SHANG; LIN, 2016). Many different sampling methodologies are available, such as Gibbs sampling, Metropolis-Hastings sampling, a family based in Hamiltonian algorithms and so on (ANDRIEU *et al.*, 2003).

Algorithm convergence derives from Monte Carlo and Markov chain convergence results and the sampling methodology responds essentially for the convergence speed and computational costs (ANDRIEU *et al.*, 2003).

One of the most efficient and used in recent literature is the No-U-Turn Sampler (NUTS) (HOFFMAN; GELMAN, 2011), implemented in the Python package "pymc3" (SALVATIER; WIECKI; FONNESBECK, 2016) and used in our research. Algorithm 1 shows a naïve version of that algorithm as illustration.

Algoritmo 1: Naive No-U-Turn Sampler - main function

Result: Posterior n-sample with burn-in and stable parts

Given $\theta^0, \epsilon, \mathcal{L}, M$;

for $m = 1$ **to** M **do**

  Resample $r^0 \sim \mathcal{N}(0, I)$;

  Resample $u \sim Uniform([0, exp\{\mathcal{L}(\theta^{m-1}) - \frac{1}{2}.r^0.r^0\}])$;

  Initialize

  $\theta^- = \theta^{m-1}, \theta^+ = \theta^{m-1}, r^- = r^0, r^+ = r^0, j = 0, \mathcal{C} = \{(\theta^{m-1}, r^0)\}, s = 1$;

  while $s = 1$ **do**

    Choose a direction $v_j \sim Uniform(\{-1, 1\})$;

    if $v_j = -1$ **then**

      $\theta^-, r^-, \_, \_, C', s' \leftarrow BuildTree(\theta^-, r^-, u, v_j, j, \epsilon)$;

    else

      $\_, \_, \theta^+, r^+, C', s' \leftarrow BuildTree(\theta^+, r^+, u, v_j, j, \epsilon)$;

    end

    if $s' = 1$ **then**

      $C \leftarrow C \cup C'$;

    end

    $s \leftarrow s'.\mathcal{I}[(\theta^+ - \theta^-).r^- \geq 0].\mathcal{I}[(\theta^+ - \theta^-).r^+ \geq 0]$;

    $j \leftarrow j + 1$;

  end

  Sample θ^m, r uniformly at random from C;

end

In this algorithm and in its associated function presented in Algorithm 2, M is the number of samples, ϵ is the step size parameter of a "leapfrog" integrator, $\mathcal{L}$ is the logarithm of the joint density of the variables of interest θ (up to a normalizing constant), r^t and θ^t denote the values of the momentum and position variables r and θ at time t, $\mathcal{I}$ denotes the identity matrix, $\mathcal{N}(\mu, \Sigma)$ denotes a multivariate normal distribution with mean μ and covariance matrix Σ, and C is a finite set of candidate position-momentum states. In Hoffman e Gelman (2011) words, this sampling model can interpreted in physical terms

as a fictitious Hamiltonian system where θ denotes a particle's position in D-dimensional space, r^d denotes the momentum of that particle in the dth dimension, $\mathcal{L}$ is a position-dependent negative potential energy function, $\frac{1}{2}r.r$ is the kinetic energy of the particle, and $logp(\theta, r)$ is the negative energy of the particle, and simulate the evolution over time of the Hamiltonian dynamics of this system via the "leapfrog" integrator.

Algoritmo 2: Naive No-U-Turn Sampler - BuildTree function

def *BuildTree*$(\theta, r, u, v, j, \epsilon)$**:**

  if $j = 0$ **then**

    Base case - take one leapfrog step in the direction v;

    $\theta', r' \leftarrow Leapfrog(\theta, r, v\epsilon)$;

    $C' \leftarrow \begin{cases} \{(\theta', r')\}, \ u \leq exp\{\mathcal{L}(\theta') - \frac{1}{2}.r'.r'\} \\ \emptyset, \ otherwise \end{cases}$

    $s' \leftarrow \mathcal{I}[u < exp\{\Delta_{max} + \mathcal{L}(\theta') - \frac{1}{2}.r'.r'\}]$;

    return $\theta', r', \theta', r', C', s'$;

  else

    Recursion - build the left and right subtrees;

    $\theta^-, r^-, \theta^+, r^+, C', s' \leftarrow BuildTree(\theta, r, u, v, j-1, \epsilon)$;

    if $v = -1$ **then**

      $\theta^-, r^-, \_, \_, C'', s'' \leftarrow BuildTree(\theta^-, r^-, u, v_j, j-1, \epsilon)$;

    else

      $\_, \_, \theta^+, r^+, C'', s'' \leftarrow BuildTree(\theta^+, r^+, u, v_j, j-1, \epsilon)$;

    end

    $s' \leftarrow s'.s''.\mathcal{I}[(\theta^+ - \theta^-)\mathring{u}r^- \geq 0].\mathcal{I}[(\theta^+ - \theta^-)\mathring{u}r^+ \geq 0]$;

    $C' \leftarrow C' \cup C''$;

    return $\theta^-, r^-, \theta^+, r^+, C', s'$;

  end

2.3 Copulas

2.3.1 Copula Definition

The theory of copulas is based on decoupling random variables individual behavior and their dependencies, and that is where the "copula" name comes from. An n-dimensional copula is an n-increasing function defined in the $[0, 1]^n$ hypercube, meaning it is electable to represent a joint cumulative probability function for an n-dimensional random vector where all univariate components have a uniform marginal distribution. The existence of such decomposition for all n-dimensional random vector is granted by the main result in copula theory, the Sklar's theorem.

Definition 17 *(NELSEN, 2006) an* ***n-dimensional copula*** *or* ***n-copula*** *is a function* C *from* $I^n = [0,1]^n$ *to* $I = [0,1]$*, for which:*

1. *if* $\boldsymbol{u}$ *in* I^n *has at least one coordinate equal to zero, then* $C(\boldsymbol{u}) = 0$ *(grounded);*
2. *if* $\boldsymbol{u}$ *in* I^n *has all but* u_k *equal to one, then* $C(\boldsymbol{u}) = u_k$ *(uniform marginals);*
3. *for every* $\boldsymbol{a}, \boldsymbol{b}$ *in* I^n *such that* $a_k \leq b_k$ *for all* k*, then* $V_c([\boldsymbol{a}, \boldsymbol{b}]) \geq 0$ *(n-increasing);*

where $B = [\boldsymbol{a}, \boldsymbol{b}]$ *is the n-box* $[a_1, b_1] \times ... \times [a_n, b_n]$*,* $V_c(B)$ *is the* $C-$*volume given by* $V_{\boldsymbol{c}}(B) = \sum sgn(\boldsymbol{c}).C(\boldsymbol{c})$ *over all vertices* $\boldsymbol{c}$ *of* B*, where* $sgn(\boldsymbol{c})$ *is* 1 *for* $\boldsymbol{c}$ *having an even number of coordinates taken from* $\boldsymbol{a}$ *or* -1 *otherwise.*

As an example, Figure 2 shows a 2-dimensional copula surface (which is also a particular case of joint distribution, as any copula).

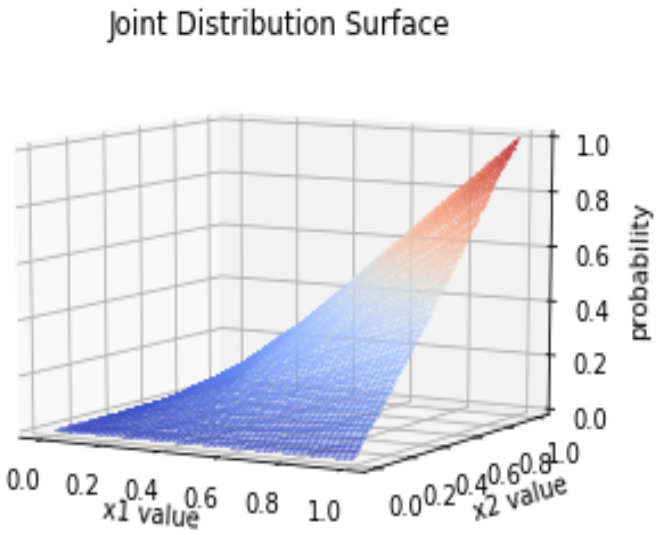


Figure 2 – 2-dimensional copula example surface.

Theorem 1 (Sklar's theorem in n-dimensions) *(NELSEN, 2006) for every n-dimension distribution function* H *with marginal distributions* $F_1, ..., F_n$ *there exists a n-copula* C *such that for all* x *in* R^n*:*

$$H(x_1, ..., x_n) = C(F_1(x_1), ..., F_n(x_n)) \tag{2.6}$$

To visually explain the concept of copula and the interpretation of Sklar's theorem, we present Figure 3 where the usually black-box joint distribution is opened up to show its two-stage modeling as proposed by Sklar's theorem and copula theory. One important point is that copula theory is all based on distribution functions instead of the usual probability densities perspective.

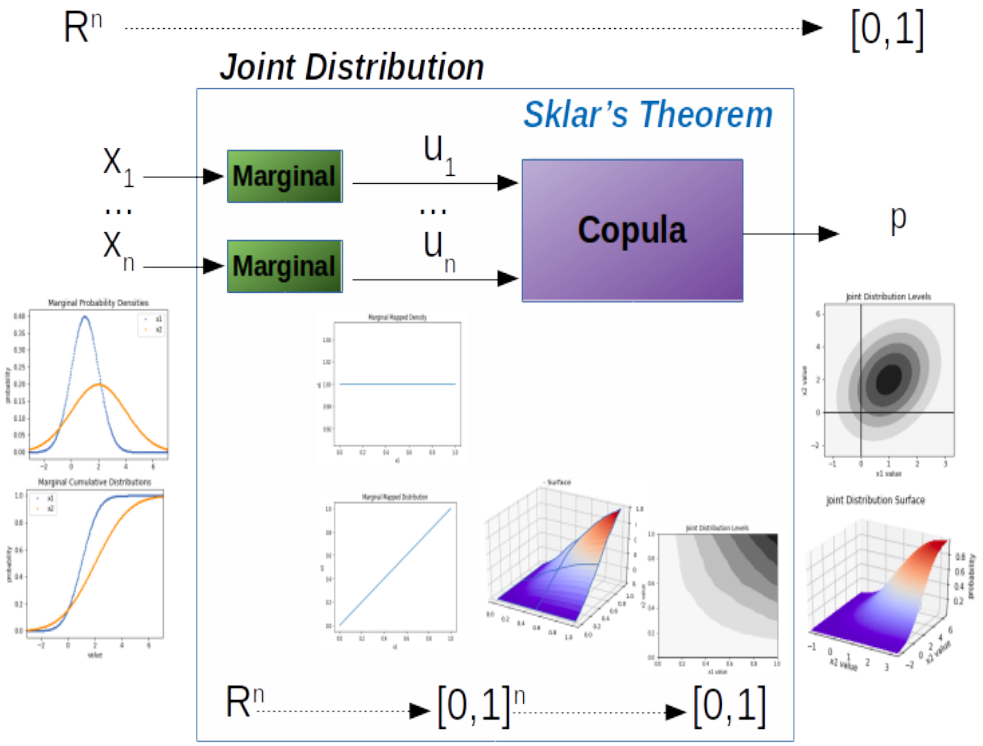


Figure 3 – Copula definition and probabilistic decoupling visualization. Sklar's theorem assures that any joint distribution can be modeled in a two-stages approach: marginal distributions mapping and copula modeling.

Definition 18 *(NELSEN, 2006) Let* $\{x_1, x_2, ..., x_n\}$ *be a sample from a random variable* X*. Then* $\{x_{(1)}, x_{(2)}, ..., x_{(n)}\}$*, where* $i < j \implies x_i \leq x_j$*, denotes an order statistic from the sample.*

Definition 19 *(NELSEN, 2006) An* ***bivariate empirical copula*** *for a sample of size* n *is the function:*

$$C_n(\frac{i}{n}, \frac{j}{n}) = \frac{\#\{(x, y) \in sample | x \leq x_{(i)}, y \leq y_{(j)}\}}{n}, \tag{2.7}$$

where $x_{(i)}, y_{(j)}$ *denote order statistics from the sample.*

An example of empirical copula is given in Figure 4.

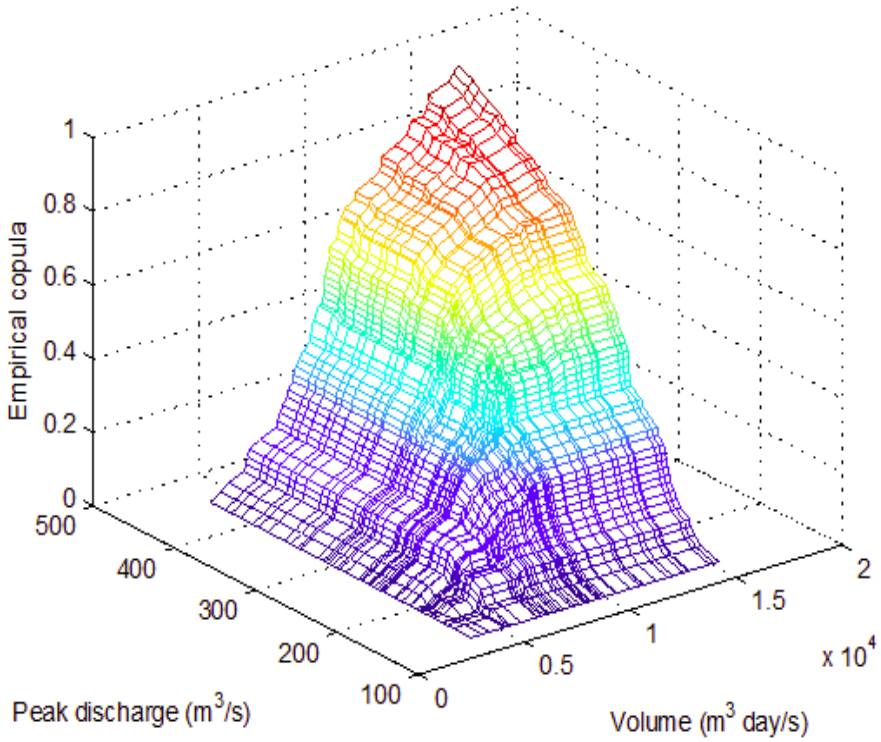


Figure 4 – Empirical copula example.

This can be generalized to the multivariate case (STRELEN, 2009) as:

Definition 20 *An* ***d-dimensional empirical copula*** *for a sample of size n is the function:*

$$C^n(u_1, .., u_d) = \frac{1}{n}.\sum_{i=1}^{n} \mathbf{1}(\tilde{U}_1^i \leq u_1, .., \tilde{U}_d^i \leq u_d), \tag{2.8}$$

where each $\tilde{U}_j^i$ *are the pseudo copula observations defined by:*

$$(\tilde{U}_1^i, ..., \tilde{U}_d^i) = (F_1(X_1^i), ..., F_n(X_n^i)) \tag{2.9}$$

It is essential to notice for the purpose of this research that a copula is not necessarily well-defined in correspondence to a specific random vector. As we are meant to deal with phenomena which comprises simultaneously discrete and continuous modeling features (federation unity and income in tax administration applications, gender and age in healthcare applications, and so on), thus discrete, or at least singular, random variables are to be inescapably considered as among our models and they do not define a unique copula but a whole family of copulas which can compose with their marginal distributions to give the phenomena unique joint distribution. For the copula to be unique it is mandatory for all marginal distributions to be continuous, otherwise the copula is only uniquely determined on $RanF_1 \times ... \times RanF_n$, where $RanF_i$ stands for the range of F_i in its image set.

That same core concept on copulas formal definition implies that there is no discrete part in a copula; copulas can be only absolutely continuous, singular, or a composition of both. That occurs because all discrete nature is bared by the marginal distributions in their domains and their composition with the copula activates only a discrete part of the copula domain, hence every copula with that same subset mapping in common is eligible for representing the joint distribution.

Definition 21 *(ELIDAN, 2013) Let C be an n-copula with marginal distributions $F_1, ..., F_n$ on a random vector X and corresponding marginal densities $f_1, ..., f_n$, which means its joint distribution is defined by $F_X(x_1, ..., x_n) = C(F_1(x_1), ..., F_n(x_n))$. If C has n'th order partial derivatives, then the* ***copula density*** *is defined by*

$$c(F_1(x_1), ..., F_n(x_n)) = \frac{\partial^n C(F_1(x_1), ..., F_n(x_n))}{\partial F_1(x_1), ..., \partial F_n(x_n)} \tag{2.10}$$

and the joint density can be derived from the copula density using the derivative chain rule

$$f_X(x) = c. \prod_i f_i(x_i) \tag{2.11}$$

An example of a commonly used n-copula is the Gaussian n-copula (ELIDAN, 2013) defined by

$$C_\Sigma(\{F_i(x_i)\}) = \Phi_\Sigma(\Phi^{-1}(F_1(x_1)), ..., \Phi^{-1}(F_1(x_n))) \tag{2.12}$$

where Φ is the standard normal distribution and Φ_Σ is a zero mean normal distribution with correlation matrix Σ.

2.3.2 Referential Copulas and Correlated Results

We adopt in this text the term "referential copulas" for those copulas which are conceptually constructed as reference for canonical dependence relations: complete dependence or complete independence. Those reference copulas derive from the following expressed results.

Theorem 2 (copula boundaries) *Let C be a copula. Then:*

$$\forall (u, v) \in DomC,\ max(u + v - 1,\ 0) \leq C(u, v) \leq min(u, v) \tag{2.13}$$

Indeed, both the inferior and superior bounds in that theorem are themselves copulas and, along with a third important copula, the product one, complete the referential copulas defined as follows.

Definition 22 *The* ***Fréchet-Hoeffding lower bound copula*** W *is the copula defined as* $W(u, v) = max(u + v - 1,\ 0), \forall (u, v) \in I^2$.

Definition 23 *The* ***Fréchet-Hoeffding upper bound copula*** M *is the copula defined as* $M(u, v) = min(u, v), \forall (u, v) \in I^2$.

Definition 24 *The* ***product copula*** Π *is defined as* $\Pi(u, v) = u.v, \forall (u, v) \in I^2$.

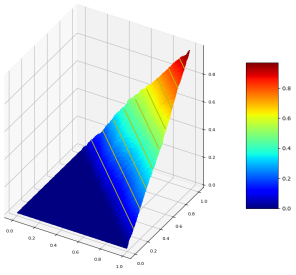
For the complete dependence case, in two dimensions, the Fréchet-Hoeffding copulas derived from the corresponding inequalities are those references, while the product copula is the reference for the complete independence case.

In a intuitive sense a complete independent copula means that all random variables are independent and given any set of fixed values for some, the probabilities for the others remain homogeneously distributed among all possible values. In contrast, complete positive dependent or *comonotonic* copula represents variables that grows altogether; and, in the inverse perspective, complete negative dependent or *countermonotonic* copula represents variables that decrease altogether. When limited to two variables, it is also possible to define the complete negative dependent copula for that case in which every growth in one variable corresponds to a decrease in the other and vice-versa, but this concept is not trivially extendable for more than two variables, although there also is a corresponding lower bound for $n \geq 3$ but which is not a copula nor can be associate to that simple negative dependence intuition.

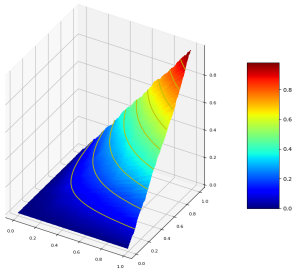
For the bivariate case, (NELSEN, 2006) remarks that the Fréchet-Hoeffding bounds suggests a partial order on the set of copulas, which can be extended under certain adaptation to the multivariate case, and he presents a definition for this order:

Definition 25 *If* C_1 *and* C_2 *are copulas, it is said that* C_1 *is* ***smaller than*** C_2 *(or* C_2 *is* ***larger than*** C_1*),* $C_1 \prec C_2$ *(*$C_1 \succ C_2$*), if* $\forall u, v \in I,\ C_1(u, v) \leq C_2(u, v)$.

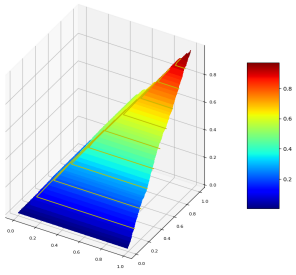
The three reference copulas are graphically represented in Figure 5.



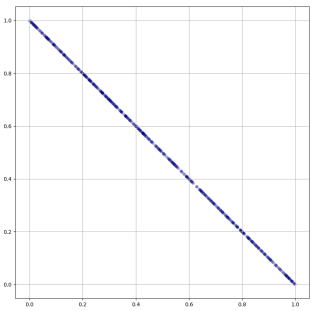
(a1) W copula 3D graph



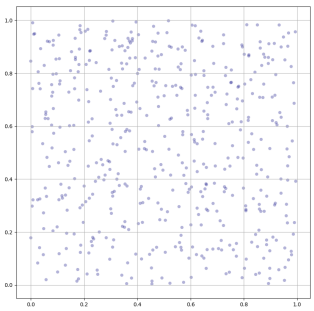
(a2) Π copula 3D graph



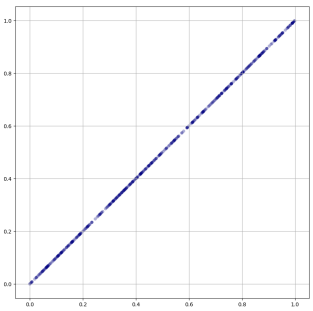
(a3) M copula 3D graph



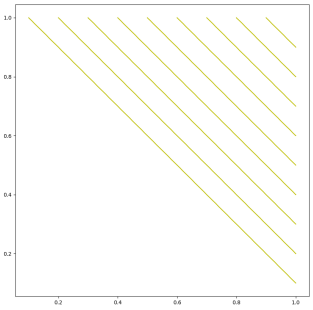
(b1) W copula scatterplot



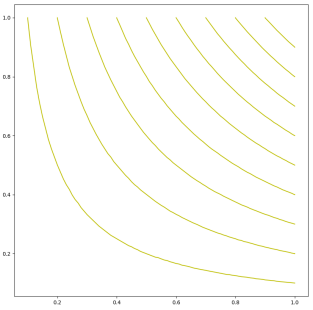
(b2) Π copula scatterplot



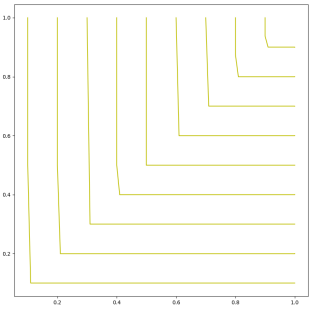
(b3) M copula scatterplot



(b1) W copula level curves



(b2) Π copula level curves



(b3) M copula level curves

Figure 5 – Comparison of the reference copulas: W Copula, the product Π Copula and the M Copula. (a) shows the 3D graphs with the 0.1 to 0.9 probability level curves in grey. (b) shows the scatterplot of each curve. (c) presents the level curves itselves projected in the base plan.

Referential copulas can be used to visually analyze the general pair dependence tendencies of a given copula by the corresponding 2-dimensional projection observing how much projected sample points in a neighborhood of a given probability level describe or not a pattern near one of the reference copulas projection, as in Figure 6. Each probability level boundaries is projected as a right triangle, where the cathetus are the upper boundary (complete positive dependence), the hypotenuse is the lower boundary (complete negative

dependence), and the iso-product curve is the complete independence reference.

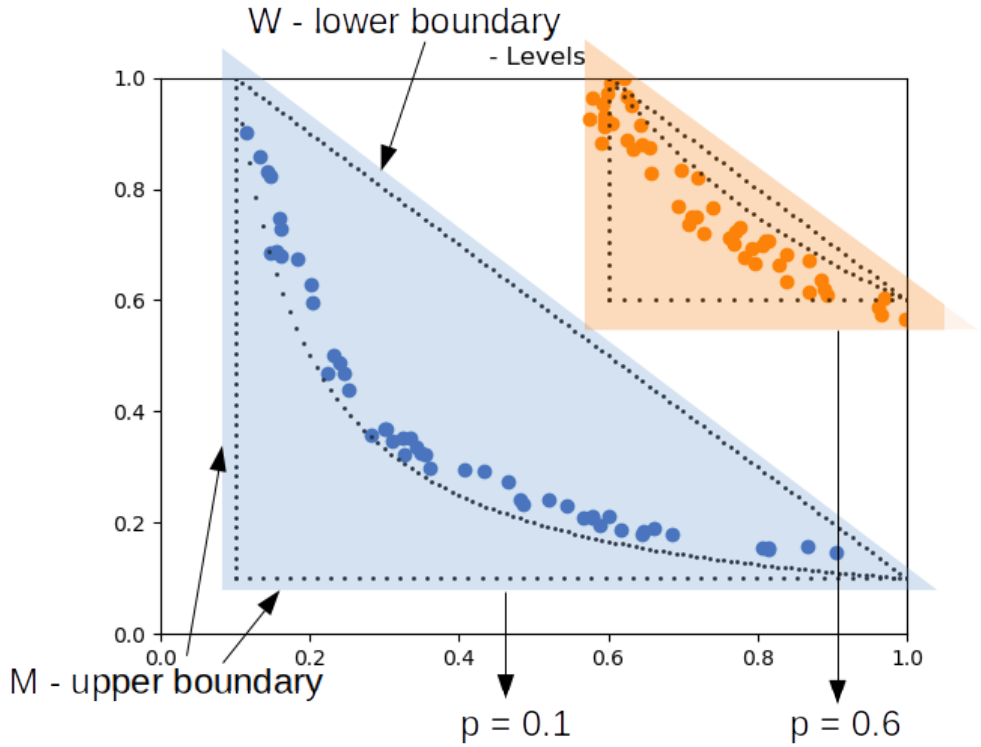


Figure 6 – Copula level boundaries for each probability. Figure shows p=0.1 and p=0.6 cases, so chosen to avoid triangles superposition.

2.4 Bayesian Networks

Another paramount theoretical element used in our research is the Bayesian network, which is a graphical approach to statistical modeling.

A Bayesian network is a model composed by a directed acyclic graph (DAG) $\mathcal{G}$ and a joint probability distribution P where $(\mathcal{G}, P)$ satisfies the Markov condition and so the joint distribution P can be decomposed in a product of conditional distributions defined by $\mathcal{G}$ as explained in the following paragraphs.

Definition 26 *(ELIDAN, 2013) A* ***Markov Network (MN)*** *is an undirected graphical model which uses an undirected graph $\mathcal{H}$ that encodes the independencies $I(\mathcal{H}) = \{(X_i \perp X \backslash \{X_i\} \cup Ne_i | Ne_i)\}$, where Ne_i are the neighbors of X_i in $\mathcal{H}$, which means that each node is independent of all others given its neighbors in $\mathcal{H}$, also known as the* ***Markov condition****.*

Theorem 3 (Hammersley-Clifford Theorem) *(ELIDAN, 2013) Let $\mathcal{C}$ be the set of cliques in $\mathcal{H}$, where a clique is a set of nodes such that each node is connected to all others in the set. For positive densities, if the independence statements encoded by $\mathcal{H}$ hold in*

$f_X(x)$, *then* ***the joint density decomposes according to the graph structure***

$$f_X(x) = \frac{1}{Z} \prod_{c \in \mathcal{C}} \phi_c(x_c) \tag{2.14}$$

where X_c is the set of nodes in the clique c, and $\phi_c : \Re^{|c|} \to \Re^+$ is any positive function over the values of these nodes. Z is a normalizing constant called the partition function. The converse composition theorem also holds.

Theorem 4 (Product induced by independence structure) *(ELIDAN, 2013) Let T be an undirected tree structured graph (i.e., a graph with no cycles) and let E denote the set of edges in T that connect two vertices. If the independencies $I(T)$ defined by T hold in $f_X(x)$, then*

$$f_X(x) = [\prod_i f_i(x_i)]. \prod_{(i,j) \in \mathcal{E}} \frac{f_{ij}(x_i, x_j)}{f_i(x_i).f_i(x_i)} \tag{2.15}$$

Theorem 5 (Copula Bivariate Decomposition) *(ELIDAN, 2013) Let T be an undirected tree structured graph and let E denote the set of edges in T that connect two vertices. If the independencies $I(T)$ defined by T hold in $f_X(x)$, then*

$$c_T(.) = \frac{f_X(x)}{\prod_i f_i(x_i)} = \prod_{(i,j) \in \mathcal{E}} \frac{f_{ij}(x_i, xj)}{f_i(x_i).f_i(x_i)} = \prod_{(i,j) \in \mathcal{E}} c_{ij}(F_i(X_i), F_j(X_j)) \tag{2.16}$$

where $c_T(.)$ is used to denote a copula density associated to the structure T and c_{ij} is used to denote the bivariate copula corresponding to the edge (i, j). The converse composition holds.

Definition 27 *(NEAPOLITAN, 2003) A* ***directed graph*** *is a pair (V, E), where V is a finite, nonempty set whose elements are called* ***nodes*** *(or vertices), and E is a set of ordered pairs of distinct elements of V whose elements are called* ***edges*** *(or arcs).*

Definition 28 *(NEAPOLITAN, 2003) A directed graph G is called a* ***directed acyclic graph (DAG)*** *if it contains no path from a node to itself (directed cycles).*

Definition 29 *(NEAPOLITAN, 2003) Given a DAG $\mathcal{G} = (V, E)$ and nodes X and Y in V, Y is called a* ***parent*** *of X if there is an edge from Y to X, Y is called a* ***descendent*** *of X and X is called an ancestor of Y if there is a path from X to Y.*

Definition 30 *(NEAPOLITAN, 2003) Suppose we have a joint probability distribution P of the random variables in some set V and a DAG $\mathcal{G} = (V, E)$. We say that $(\mathcal{G}, P)$ satisfies the* ***Markov condition*** *if for each variable $X \in V$, X is conditionally independent I_P of the set of all its nondescendents ND_X given the set of all its parents Pa_X,for which we adopt the notation $I_P(X, ND_X|Pa_X)$.*

Theorem 6 (Product of Conditional Distributions) *(NEAPOLITAN, 2003) If* $(\mathcal{G}, P)$ *satisfies the Markov condition, then* P *is equal to the* ***product of its conditional distributions*** *of all nodes given values of their parents, whenever these conditional distributions exist, that is*

$$P(x_n, x_{n-1}, ..., x_1) = P(x_n|pa_n).P(x_{n-1}|pa_{n-1})...P(x_1|pa_1),\ P(Pa_i) \neq 0, 1 \leq i \leq n \tag{2.17}$$

Theorem 7 (DAG Markov Condition) *(NEAPOLITAN, 2003) Let a DAG* $\mathcal{G}$ *be given in which each node is a random variable, and let a discrete conditional probability distribution of each node given values of its parents in* $\mathcal{G}$ *be specified. Then the product of these conditional distributions yields a joint probability distribution* P *of the variables, and* $(\mathcal{G}, P)$ *satisfies the Markov condition.*

Definition 31 *(NEAPOLITAN, 2003) Let* P *be a joint probability distribution of the random variables in some set* V*, and* $\mathcal{G} = (V, E)$ *be a DAG. We call* $(\mathcal{G}, P)$ *a* ***Bayesian network*** *if* $(\mathcal{G}, P)$ *satisfies the Markov condition. From Theorem 6,* P *is the product of its conditional distributions in* $\mathcal{G}$*, and this is the way* P *is always represented in a Bayesian network. Furthermore, from Theorem 7, if we specify a DAG* $\mathcal{G}$ *and any discrete conditional distributions (and many continuous ones), we obtain a Bayesian network. This is the way Bayesian networks are constructed in practice.*

Figure 7 shows an example of Bayesian network.

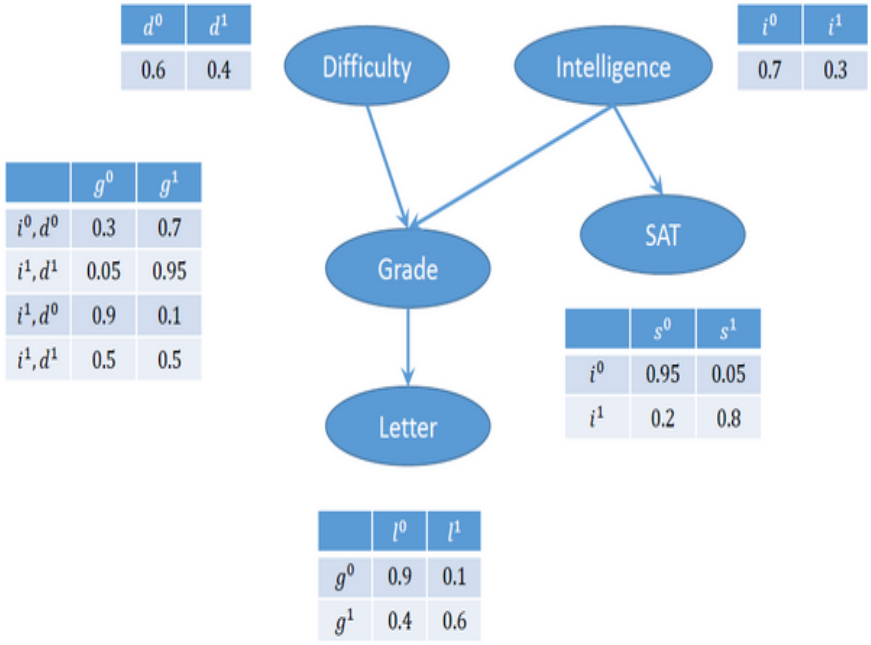


Figure 7 – Bayesian network example (GORDON *et al.*, 2014). It represents the probabilistic relations among variables in a student grading problem with the following features: student intelligence, discipline difficulty, discipline grade for that student, student SAT and a positive letter of recommendation for that student by the discipline teacher.

The contribution of this graphical decomposition is that estimation and learning are simplified by the compact representation, but at a trade-off of strong independence assumptions. Some further approaches to overcome those strong premises are (KIRSHNER, 2009) mixture of all copula trees model proposal, at the cost of some loss of flexibility by parameter sharing constraints, and (SILVA; GRAMACY, 2009) Bayesian approach of a mixture of some trees with flexible priors on all components of the model.

Theorem 8 (Product of conditional densities) *(ELIDAN, 2013) If the independences encoded by $\mathcal{G}$ hold in f_X, then*

$$f_X(x) = \prod_{i=1}^{n} f_{X_i|Pa_i}(x_i|pa_i) \tag{2.18}$$

and the converse composition theorem is valid, i.e., a product of any local conditional densities defines a valid joint density with the independences encoded by the DAG $\mathcal{G}$ associated to that product.

Although that decomposition and graphical representation simplifies the joint distribution modeling, it is still very far from a simple problem. Finding a Bayesian

network best structure is a NP-hard (non-deterministic polynomial-time hardness) problem, which can be taken as being super-exponentially time expensive for an exhaustive search algorithm. Figure 8 shows graphically the number of possible structures for Bayesian networks in terms of the number of variables. Observe that for 20 variables, this number is above 10^{70}.

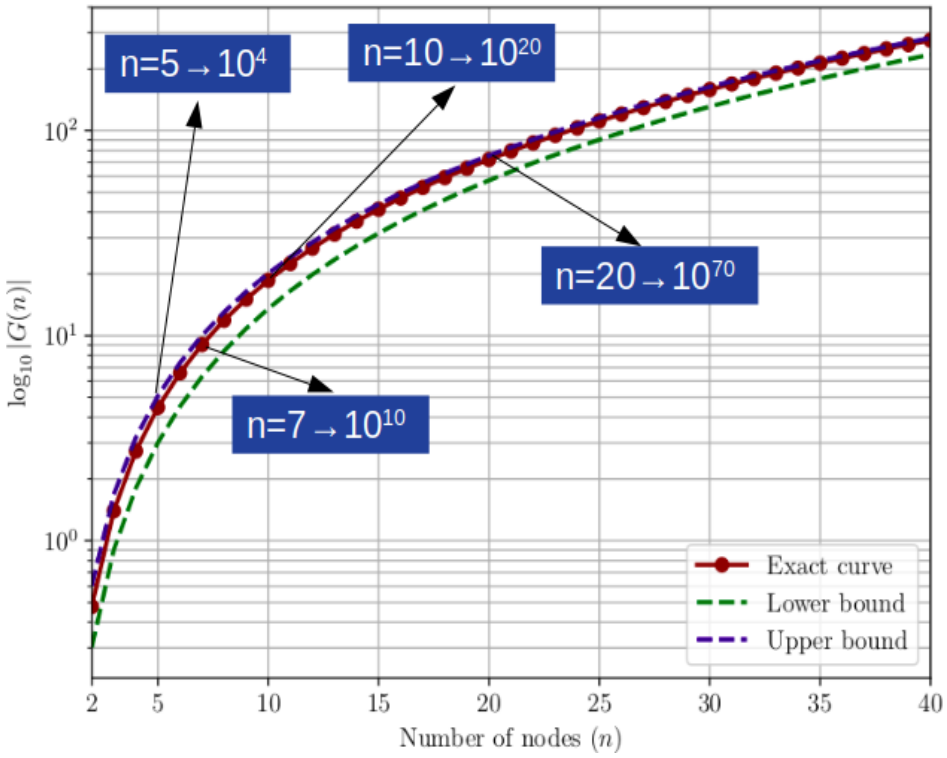


Figure 8 – Number of DAGs as the number of nodes increases according to Robinson's recurrence (ROBINSON, 1977). Graphics from Gross *et al.* (2019).

Therefore, one approach for Bayesian networks modeling is to treat it as a search for the optimal network structure and parameters among all possible structures for a given dataset. That problem is usually decomposed in a structure search followed by a parametric computation for the chosen structure, formulated as follows.

Definition 32 ***BN learning*** *is the optimization problem that, given a dataset D, find the BN $B = (\mathcal{G}, \Theta)$ that maximizes $P(B|D) = P(D|B).P(B) = P(D|\mathcal{G}, \Theta).P(\Theta|\mathcal{G}).P(\mathcal{G})$.*

Definition 33 ***Structure learning*** *is the part of BN learning focused on finding the network structure $\mathcal{G}$ that maximizes $P(\mathcal{G}|D)$*

- $P(\mathcal{G}|D) \propto P(D|\mathcal{G}).P(\mathcal{G})$
- $P(D|\mathcal{G}) = \int_\Theta P(D|\mathcal{G}, \Theta).P(\Theta|\mathcal{G}).d\Theta$

The optimization problem needs to be instrumented by a score function which associates to each possible structure a corresponding score measuring how good is that structure to represent the given dataset. In this research we choose a score, very used in literature, called Bayesian Dirichlet equivalence with uniform prior metric (BIELZA; NAGA, 2014) - BDeu as the referential score in our structure learning stages.

Definition 34 ***BDeu scoring*** *is a scoring measure for Bayesian network structures which assumes $P(\mathcal{G})$ to be a uniform distribution and $P(\Theta|\mathcal{G})$ to be a Dirichlet distribution resulting in the following equation (HEKERMAN; GEIGER; CHICKERING, 1995) for computing a network structure score for a given dataset:*

$$P(D|\mathcal{G}) = \prod_{i=1}^{n} \prod_{j=1}^{q_i} \frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij} + N_{ij})} . \prod_{k=1}^{r_i} \frac{\Gamma(\alpha_{ijk} + N_{ijk})}{\Gamma(\alpha_{ijk})} \tag{2.19}$$

where i stands for each structure node, j for each state of each node, and k for each node parents instance.

The CBN mentioned in Chapter 1 has the formal definition presented in Definifition 35, based on 1.

Lemma 1 (copula conditional density) *(ELIDAN, 2013) Let $f(x|\boldsymbol{y})$, with $\boldsymbol{y} = \{y_1, y_2, ..., y_n\}$, be a conditional density function. There exists a copula density function $c(F(x), F_1(y_1), ..., F_k(y_k))$ such that*

$$f(x|\boldsymbol{y}) = R_c(F(x), F_1(y_1), ..., F_k(y_k)).f_X(x), \tag{2.20}$$

where R_c is the copula ratio

$$R_c(F(x), F_1(y_1), ..., F_k(y_k)) = \frac{c(F(x), F_1(y_1), ..., F_k(y_k))}{\frac{\partial^k C(1, F_1(y_1), ..., F_k(y_k))}{\partial F_1(y_1) ... \partial F_k(y_k)}} \tag{2.21}$$

and R_c is defined to be 1 when $\boldsymbol{Y} = \emptyset$. The converse is also true: for any copula, $R_c(F(x), F_1(y_1), ..., F_k(y_k)).f_X(x)$ defines a valid conditional density.

Definition 35 *(ELIDAN, 2013) A Copula Bayesian Network (CBN) is a triplet $\mathcal{C} = (\mathcal{I}, \Theta_C, \Theta_f)$ that defines $f_X(x)$. $\mathcal{I}$ encodes the independencies $\{(X_i \perp \boldsymbol{ND}_i|\boldsymbol{Pa}_i)\}$, assumed to hold in $f_X(x)$. Θ_C is a set of local copula functions $C_i(F(x_i).F(\boldsymbol{pa}_{i1})...F(\boldsymbol{pa}_{ik_i})$ that are associated with nodes of $\mathcal{I}$ that have at least one parent. In addition, Θ_f is the set of parameters representing the marginal densities $f_i(x_i)$ (and distributions $F_i(x_i)$). The joint density $f_X(x)$ then takes the form*

$$f_X(x) = \prod_{i=1}^{n} R_{c_i}(F(x_i).F(\boldsymbol{pa}_{i1})...F(\boldsymbol{pa}_{ik_i})).f_i(x_i). \tag{2.22}$$

Table 1 (ELIDAN, 2013) presents a summary of the different copula-based multivariate models and its application with the author's observations at the time (2013).

Table 1 – Summary of the different copula-based multivariate models extracted from (ELIDAN, 2013) with that author's observations.

Model	**Variables**	**Structure**	**Copula**	**Comments**
Vines	< 10 in practice	Conditional dependence	Any bivariate	well understood general purpose framework
Nonparametric BBN	100s 100s	vines vines	Gaussian in practice	mature application
Tree-averaged	10s	Mixture of trees	Any bivariate	requires only bivariate estimation
Nonparanormal	100-1000s	MN	Gaussian	high-dimensional estimation with theoretical guarantees
Copula networks	100s	BN	Any	flexible at the cost of partial control over marginals
Copula processes	∞ (replications)	-	Multivariate	Nonparametric generalization of Gaussian processes

3 METHODOLOGY

3.1 General Modeling Methodology

According to (HAIR *et al.*, 1998) (pp. 25-27), a structured approach to multivariate model building may follow some steps:

1. define the research problem, objectives, and multivariate technique to be used;
2. develop the analysis plan;
3. evaluate the assumptions underlying the multivariate technique;
4. estimate the multivariated model and assess overall model fit;
5. interpret the variates; and
6. validate the multivariate model.

Still based on (HAIR *et al.*, 1998), the first step establishes the analysis starting point as the conceptual model development, consisting in defining the research problem and analysis objectives in theoretical terms before specifying any variables or measures.

From the application point of view, one of our original research problems was, for example, to generate likely simulated samples from a model based on an initial real sample acquired from a given phenomenon. The goal was to supply other researchers, data scientists and analysts with the simulated samples for them to make more specific analysis and detect behaviors or relations that are representative and useful for dealing with that phenomenon. In parallel with that goal, we also want to use the modeling tool for analyzing phenomena of interest, like healthcare system and tax issues.

As to the multivariate technique to be used, the copula approach is the chosen one, as already mentioned at Chapter 1, because the idea of treating separately each random variable isolated behavior from their relationships seemed very promising. The challenge here will be to adventure, even not deeply, in multivariate copula, a field not yet totally consolidated. At this point, we decided to get advantage of our relative previous knowledge on BN for modeling joint distributions and try to apply this modeling technique to copula modeling, considering copulas are also joint distributions.

The methodology to be adopted is represented by the block diagram showed at Figure 9 concerning the sample data simulation case. In the context of this research, the final stages of getting samples from the models and their comparison was not conducted because it was not mandatory for our purposes.

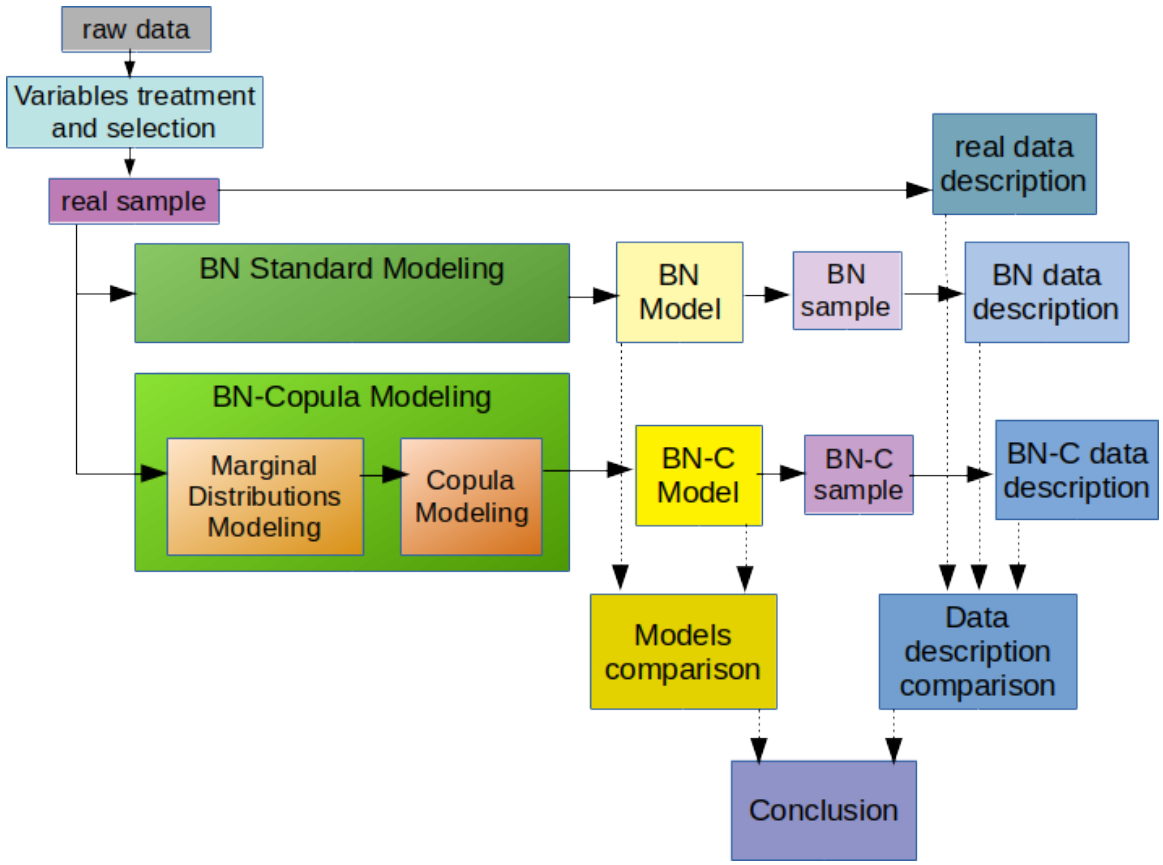


Figure 9 – Methodology Block Diagram. Basically, the methodology have three phases: data pre-processing for generating the original sample, model extraction (occasionally generating simulated samples) and a final comparison for conclusion about the model efficiency. Regarding data description, this stage is done with the help of the software tool LpdCopModel and consists in both Descriptive Statistics measures and graphics and an empirical copula modeling for dependence visualization.

3.1.1 Data Pre-processing

After drafting the conceptual model, next step is to acquire the data upon which the analysis is going to take place.

At this point, specialists knowledge must be considered to determine scope and means to integrate the model, making it a supervised and context-based stage of the modeling able to reduce an initial set of many variables simultaneously acquired from a sophisticated database to a smaller one based on the specialist's experience on their

relevance to the model or its objectives.

In some of the applications focused by this research, such as both public healthcare system and tax administration analysis, the data is available in a complex and heterogeneous repository among similar or even unrelated information, being more efficient to acquire a considerable volume and diversity of data in a massive extraction and then selecting the representative share for the ongoing analysis and modeling. In such cases, the process must include a selection stage somewhere between the data acquisition and the detailed analysis of the relevant data.

3.1.2 Data Acquisition

The data acquisition stage is a non-structured one, in the sense that brute data can be in any specific form, would it be readily usable tables, files, text or even images.

In this research, data sources varied from html tables available in web pages and downloadable text files to data base coded DBF files. Thence, data acquisition consisted in downloading data specific formatted files, text or character oriented information for further parsing into usable databases.

As mentioned before, all the data were collected from Brazilian Government open databases which are expanding since a huge transparency initiative was conducted in the last two decades in the country.

The first step was to identify the best data available to suit that subject, which leaded to very promising data sources in both matters. For the application facing the behavior of tax income, a data source managed by the Brazilian Federal tax service, the "Receita Federal do Brasil - RFB", which is a web page accessible by an open data menu option ("Dados Abertos") in that institution front page, heading to the "Dados Econômico-Tributários e Aduaneiros da Receita Federal" page [1], where one can obtain access to a large amount of consolidated economic, tax and customs data. The best suitable data found in that site were those from the individual tax returns and the enterprise application forms, aggregated by county and so guaranteeing privacy protection naturally associated to that kind of data. In the matter of public health system, data was acquired from DATASUS systems [2], a big public database managed by the Brazilian public health government branch. DATASUS is a Brazilian health care public dataset which broadcasts

[1] https://www.gov.br/receitafederal/pt-br/acesso-a-informacao/dados-abertos
[2] http://www2.datasus.gov.br/DATASUS/index.php?area=02

many kinds of reports available to the general public and it gives access to its granular datasets.

3.1.3 Data Formatting

Occasionally, raw data need first to be formatted before being manipulated by any processing. This is the case with converting numbers from different language standards, for example.

Furthermore, it is frequent that data collected from many different sources must be merged into a single database, hence it has to be normalized regarding one unified key.

In the case of the two real datasets here considered, data was initially processed for cleaning and aggregating values which were originally segregated by classes (for example, tax returns separated into complete and simplified types of form) and then sequentially imported into a code manageable data frame.

3.1.4 Data Selection

As data sources are in general plural and heterogeneous, acquired data may contain information in excess or irrelevant to the subject of study, therefore aggregating undesirable and unnecessary complexity to the analysis.

It is advisable to count on an expert help to reduce the original dataset to a more concise one, regarding restraining analysis complexity to its minimum. Therefore, in order to drive efforts more to the methodology than to the application itself in this first moment, the originally collected data was subjected to a feature reduction by a tax specialist view, who excluded the less relevant variables based on his/her experience.

3.1.5 Modeling Extraction

The strategy used in this research was driven by previous initial studies in risk analysis and rare events which further leaded to the copula modeling strategy. Although some general analysis using aggregated data are applicable for studying a phenomenon, the strength of rare events and risk analysis techniques would be more fruitful the more granular were the data, which reinforces the motivation for obtaining a granular dataset. Afterwards, the methodology resulted in a model based on a copula function to be applied to the ensemble of marginal distributions of granular data samples.

3.1.5.1 Random Variables Attribution

To each selected feature must correspond an appropriate random variable. First we must discriminate between numeric and categorical features and then, among the numeric ones, between discrete and continuous.

A numeric continuous feature is naturally associated to a random variable taking the same values presented by the feature; nevertheless, it can be also associated to a discrete variable by splitting the feature domain into a number of ranges, as if it was a categorical feature. A numeric discrete feature with a great number of possible values can also be treated by a smaller number of ranges or taken as continuous by approximation, if the cardinality is sufficient large. Finally, both low cardinality numeric discrete and categorical features must be treated as discrete random variables, and there are a variety of techniques for that, such as binarization, dummy variables and one-hot encoding.

3.1.5.2 Marginal Distributions Modeling

After associating an adequate random variable to each feature, one ends up with a corresponding set of random variables to which corresponds a set of yet unknown marginal distributions as prior defined, and the task now is to fit a distribution model for each variable based on the sample values.

(VOSE, 2009) (pp. 263-300) dedicates a full chapter on the subject of "Fitting distributions to data" and its importance for the risk analyst. He explains that this task can be done from two sources, available data and experts opinion, but, considering that our subject is to propose a methodology applicable in various fields and not to treat any specific dataset, this research is going to restrain itself to fit distributions to available data. The referenced author also mentions that data can be fit to empirical (non-parametric) or parametric distributions and the fitting can consider many approaches concerning its complexity, like a first-order distribution based only on variability or a second-order distribution taking into account both variability and uncertainty.

We are going to consider some different techniques to fit distributions to individual random variables for matters of comparison: empirical non-parametric fitting, parametric Bayesian MCMC sampling and a sample reducing mapping, as will be further presented.

3.1.5.3 Copula Modeling

The model construction consists in starting at a given dataset of samples from the modeling subject to first determine the marginal distributions for each variable, and then identifying the corresponding copula that compound with those functions gives a

reasonable approximation of the former dataset joint distribution function.

As copula traditional parametric approaches have restrictions, such as forcing similar correlation orders among all variables as a reflex of the used function family characteristics, and difficulties that would not allow the generality intended for the model or impose severe costs to the modeling, we decided to initially adopt the more general concept of empirical copula (NELSEN, 2006), as treated in Chapter 2.

Again, just as in the case of marginal distribution fitting, the empirical approach has its limitations in terms of performance, but no determinant compromise is expected that would overcome the advantage of keeping the modeling simple, allowing to focus on the complete modeling instead of in its precision. Further development in the research line implies trying any other copula modeling also for this early stage which can improve the modeling, including graphical approaches, machine learning structures or even parametric copula families fitting if viable.

In a latter stage of this research, tests were conducted for modeling copula using Bayesian networks and the results are presented in the corresponding section.

3.1.6 Model Interpretation and Validation

Model validation was conducted by running the modeling process on test datasets and comparing the previously known data generation model with the model acquired by the methodology.

As already mentioned, the modeling methodology was motivated from real-life problems concerning analyzing tax administration and public healthcare system issues, so the methodology was applied to real datasets from both areas and the resulting models can be used to extract relations and profiles from those datasets. The resulting analysis can help checking if the model was adherent to reality and would survive a real-world examination, and simultaneously to see if it could already provide new insights on the underlying phenomena.

3.2 Proposed Modeling Methodology

3.2.1 Premises

Before starting presenting the methodology itself, it is important to settle grounds for it by establishing the premises upon which it was developed.

First of all, **effectiveness** must be a north, in the sense that we are to focus on short-term applications and any method here developed must have a practical application without further great improvement needs. As already mentioned, this research aims to provide more a **multidisciplinary** approach, although innovative in some extent, for machine learning than focus on specific theoretical development in copula or Bayesian networks. In consequence, we privileged **simplicity** in dealing with each stage of the methodology, mainly in the sense that we always adopted well-established techniques in every step not specifically under testing, preferably those already implemented by reliable software packages, as though in making choices we also go for the simpler way. In developing our models, we also prefer approaches that provide more **generality** to them to more specific techniques, regardless of casual advantages. Another essential point is that we are always assuming **large enough samples** so that sample size will not compromise intermediate results throughout any alternative implementations of the methodology, as we are not going to treat small sample cases or missing data, although small samples can be also submitted to a majority of the presented methods still with effective results. Both **numeric and categorical features** must be comprised by the methodology in order to be useful to the most common real-world datasets. We adopt here a **Machine Learning perspective** where we aim approximated models and not exact theoretical models, meaning a trade-off between precision and model complexity will always be considered by the standard of giving practical results. Finally, this research aims to provide the **viability and potential** of the proposed methodology, without the presumption of exhausting its adherence grades or modeling performance in particular cases.

3.2.2 Essential Elements

The essence of the proposed methodology is to apply Sklar's theorem probabilistic decomposition to Machine Learning, specifically to Bayesian networks, and to verify empirically the impacts of this approach in relation to a standard procedure. This immediately implies in modeling copulas with Bayesian networks, as the title of this text suggests.

Sklar's theorem establishes that any n-dimensional joint distribution can be decomposed into n marginal distributions and an n-dimensional copula (or simply an n-copula), as represented in Figure 10.

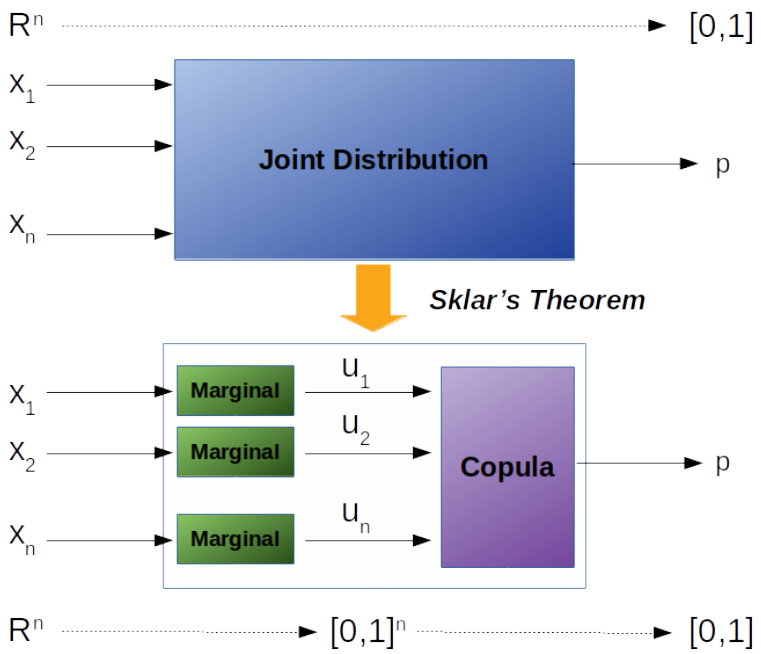


Figure 10 – Diagram representing the application of Sklar's theorem to transform an n-dimensional joint distribution in a composition of all n marginal distributions and a n-copula. X_i stands for the original random variables, while u_i is its corresponding cumulative probability, and p the final cumulative probability for the vector $X = (X_1, ..., Xn)$.

Figure 11 shows the first stage for the copula modeling, which is the mapping from each individual sample value to the corresponding accumulated probability.

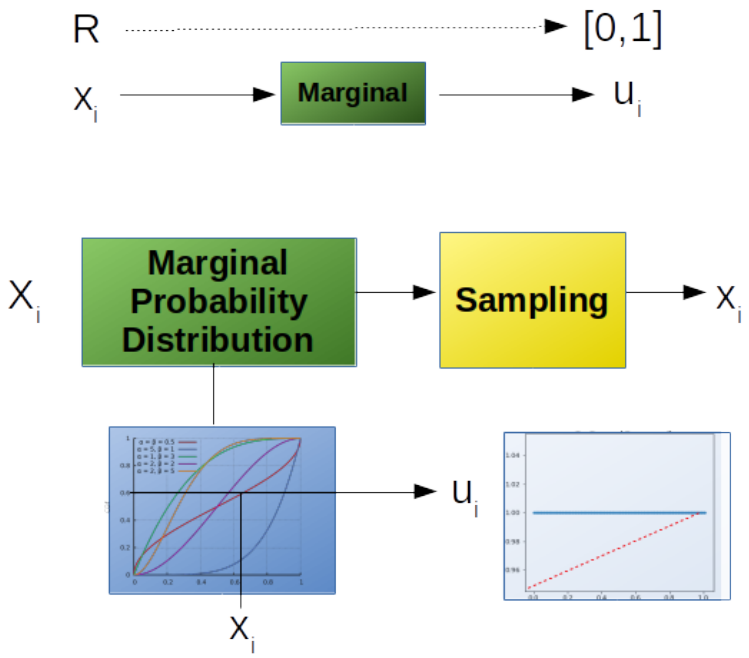


Figure 11 – Diagram representing a random variable marginal probability distribution and the mapping between its sample values and the corresponding probability values.

One of the classical methods for mapping a sample to the corresponding marginal distribution - besides more basic techniques based on sample statistics instead of considering all sample values individually, like the method of moments - is to use the complete sample to fit it to an adequate probability distribution, or, for continuous random variables, to the equivalent probability density distribution, given a parametric distribution model (see Figure 12).

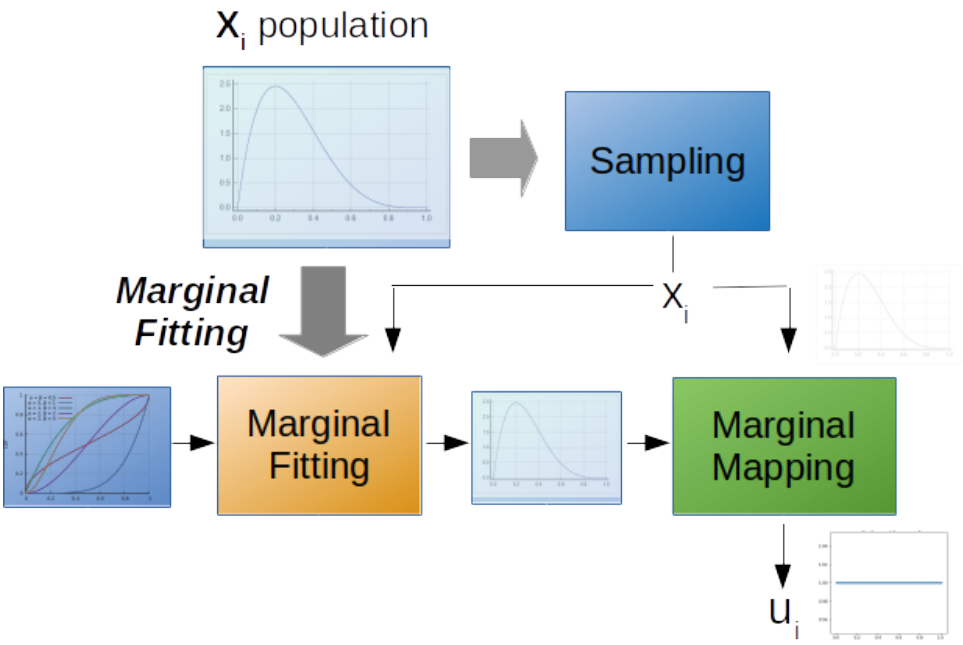


Figure 12 – Diagram representing a random variable marginal probability distribution fitting and the mapping between sample values and the fitted distribution.

For distribution fitting there are again many techniques, and one of the most popular in scientific applications is the Bayesian inference using MCMC sampling, which is the one we choose for our research. Details on this technique and the specific implementations were also detailed in Chapter 2.

In parallel to the abstract population approach derived from a Statistics-driven perspective, we can also observe the problem by an empirical sample-centered perspective, which is usual in Machine Learning (and also in Statistics, although less popular than the parametric approach). In that sense, instead of trying to discover what would be the population original probability distribution, we can keep it unknown and focus on what would be the sample resulting probability mapping without assuming any premises on the population distribution. That is exactly what the empirical distribution method is up to, by equally distributing the probability mass among the sample instances.

The basic theoretical solution for this problem is the well-known empirical distribution, mentioned by (VAART, 1998), which associates to each sample instance value an equal amount of accumulated probability mass, also called in literature the plotting position, by the equation:

$$F_k^n(x) = \frac{1}{n}.\sum_{i=1}^{n} \mathbf{1}(X_k^i \leq x) \tag{3.1}$$

where **1** is the **indicator function** which is 1 if the argument expression is true and 0 otherwise.

Cunnane (1979) has studied the plotting position characteristics for many possible calibrations in that formula to correct some distortions like bias and variance among estimates. His procedure from the hypothetical unknown distribution to the empirical distribution and sample mapping (probability percent) are represented in Figure 13, and his conclusions are summarized in Table 2, where he references the following general formula for the plotting position:

$$F[X \leq x_{(i)}] = \frac{(i - \alpha)}{(N + 1 - 2.\alpha)} \tag{3.2}$$

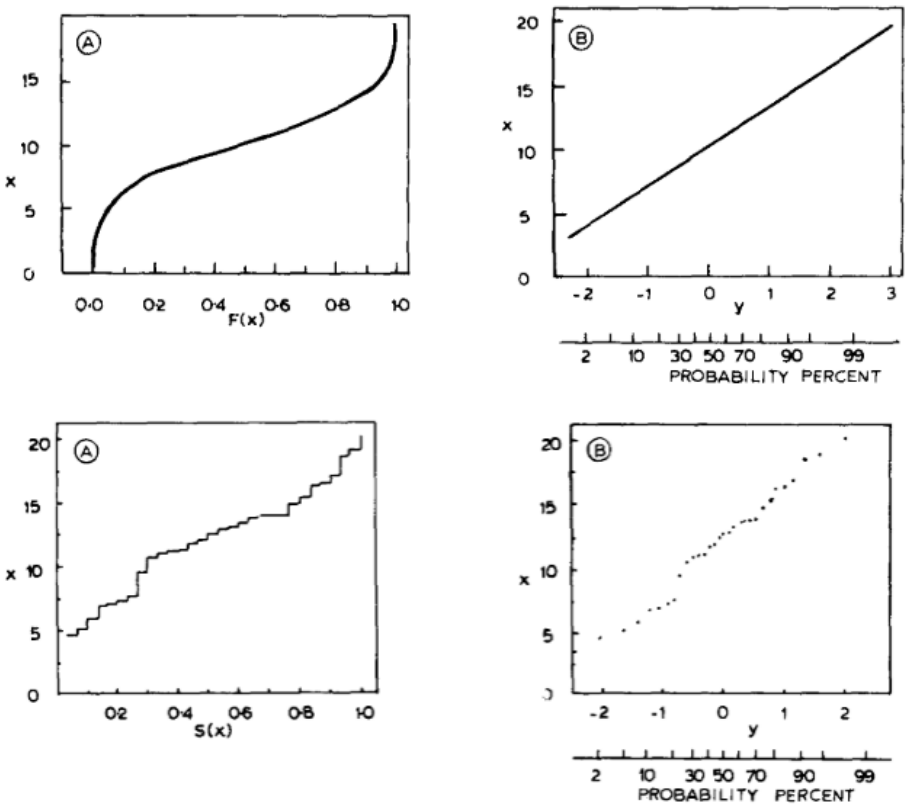


Figure 13 – Sample probability-variate relationship extracted from Cunnane (1979).

Table 2 – Sample empirical fitting comparison among formulae proposed in literature.

Name	**Formulae**	α	**Application**
Hazen	$(i-0.5)/N$	0.5	good for both normal and Gumbel distributions
Weibull	$i/(N+1)$	0.0	uniform distribution;
Blom	$(i-0.375)/(N-0.25)$	0.375	unbiased in normal distribution
Gringorten	$(i-0.44)/(N+0.12)$	0.44	unbiased in Gumbel distributions
Cunnane	$(i-0.4)/(N+0.2)$	0.4	compromise for all distributions

Therefore, we adopted here the sampling empirical fitting using the general case formula, with $\alpha = 0.4$:

$$F[X \leq x_{(i)}] = \frac{(i-0.4)}{(N+0.2)} \tag{3.3}$$

Next step, after each sample vector component transformation from its original one-dimensional space to the $[0, 1]$ probability space, is the Bayesian network modeling of the resulting reduced d-dimension sample vector.

Again using our simplicity premise, we adopt discrete Bayesian network modeling, because it avoids the complexity of mixing network discrete and continuous variables by previously discretizing all of them for treating everyone as discrete, and also because dealing with conditional probability tables instead of conditional probability distributions is much more consolidated and easy, with this option causing no harm to the conclusions we are interested in.

3.2.3 Random Variables Attribution

The first step in statistical modeling is attributing random variable to each feature of interest. For features whose values already come from an at least approximately numeric continuous set, such as money, distances, weights, the association is immediate from the given feature to a continuous random variable represented by its values in the Real line. For categorical or numeric discrete features where the association to a continuous random variable is not possible, there are other strategies available.

As we will be going to deal with copulas in further modeling stages, and as in copula theory mainstream it is usual to deal with random variables of different natures reflected on singular and absolutely continuous copula components, and concerned to the curse of dimensionality impact on copulas, we have opted not to split categorical or

discrete features into binary dummy random variables. Therefore, categorical and discrete features were mapped into Real random variables as well, but, unlike the natural mapping provided for continuous features, this mapping demanded a more careful interpretation.

Another consequence of the copula approach is that we are going to work from the (cumulative) distribution function perspective, instead of the more common probability density distribution one. This pushed us to do the feature-random variable association considering the cumulative probability at each point. Again, the continuous case has a natural correspondence, because $P[X \leq x]$ remains naturally corresponding to feature values that are inferior or equal to x, while the interpretation for the discrete/categorical case is tougher.

To accomplish a certain homogeneity to all random variable mapping, we have established a similar mapping in all cases, but with different interpretations, - which will be relevant only in the marginal distribution modeling stage - as described, where Ω is the sample space, ω each individual sample element in that space, X the random variable associated to that space, and $P[x]$ the probability of event x:

1. For continuous features:

   $X : \Omega \to R,\ \omega \to x = \omega$ (feature value in the real world);

   $X : \Omega \to R,\ P[X \leq x] = P[X(\omega) \leq x]$;

2. For discrete/categorical features with a natural order relation in the real world:

   $X : \Omega \to R,\ \omega \to x$ picked from $\{1, 2, .., n\}$ obeying the pre-existing natural order where n is the total number of categories;

   $X : \Omega \to R,\ P[X \leq x] = P[\omega : X(\omega) \in \{1, 2, ..., x\}]$;

3. For discrete/categorical features with no natural order relation in the real world:

   $X : \Omega \to R,\ \omega \to x$ randomly picked from $\{1, 2, .., n\}$ where n is the total number of categories;

   $X : \Omega \to R,\ P[X \leq x] = P[\omega : X(\omega) \in \{1, 2, ..., x\}]$;

It must be registered that, on which concerns discrete/categorical features with no natural order among their categories, this association would introduce in the modeling an artificial order relation with no real meaning, but it has no negative or noise effect in the copula modeling, because the order considered in this methodology past the marginal fitting stage regards only to the cumulative probability assumptions, which stands accordingly both for the real world feature and the random variable in this mapping. In

contrast, if we were to model using Machine Learning methodologies based on densities and elementary mass probabilities this could cause severe damages, because the artificial order introduced by the mapping would be interpreted by some models as a natural order between elements of Ω in the real world.

As it is a somewhat tricky aspect, let us go a little bit further on this matter. For example, if we are modeling country populations for n different countries and we have only those two features, the categorical feature country and the numeric feature population, we could associate random numeric variables by assigning an integer from 1 to n to each on of those n countries. Of course if we take the average of the country random variable, it would mean nothing, because there is no intrinsic natural order among countries, but the cumulative probability $P[X \leq 5] = 0.6$ would have the well-definite meaning that the probability that a sample comes from any of the countries represented by the numbers 1 to 5 is 60%. Instead, if we were modeling using other Machine Learning techniques, such as neural networks or clustering, directly from that mapping, the model could incorporate that artificial order inserted by the mapping in noisy meanings into the model.

3.2.4 Marginal Distribution Modeling

Marginal Distribution Empirical Modeling

For starting this line of research, the empirical fitting (VAART, 1998) was taken as a first approach, as follows:

$$F_k^n(x) = \frac{1}{n}.\sum_{i=1}^{n} \mathbf{1}(X_k^i \leq x) \tag{3.4}$$

where **1** is the **indicator function** which is 1 if the argument expression is true and 0 otherwise.

The fitting of the obtained empirical distribution to the fitted sample can be checked by comparing a detailed sampling of that distribution and the sample itself. For better checking it is wise to use a remarkably superior sampling rate than the one associated to the orignal sample, signifying the sampling to have a much superior number of instances. This leads to a much smaller granularity in the distribution curve graph than in the sample and will reflect in the corresponding graphs as a horizontal line pattern for the more rarefied part of the probability distribution (the tail concerning higher values) while the distribution curve remains decreasing. For example, if the sample has $5,000$ instances and the marginal distribution is discretized by a $50,000$ sampling, the distribution histogram will have unit steps about one-tenth the one for the sample.

Although empirical fitting has severe limitations in terms of performance, that technique was chosen for its simplicity and methodological coherence with the empirical copula, considering the immediate objective here of building the complete modeling methodology from real data to simulated sample in a simple version. Latter in the research, this fitting technique was improved to the sample reducing further explained.

Bayesian Inference using MCMC

Growing in sophistication, another method for modeling marginal distributions is the Bayesian inference one based on Markov-Chain Monte Carlo (MCMC) which consists of an iterative process of sampling from an evolving distribution within some rules of improvement until achieving stability enough to consider that the sample approximately refers to the posterior itself.

This method depends on some premises: the assumption of a given parametric distribution with unknown parameters for the target random variable and initial prior distributions for its parameters.

For matters of simplicity, as stated by one of this research premises, we are assuming the same parametric distribution family for all random variable cases, a **beta distribution**. For this to be possible, first all variables are normalized from its original domain into the $(0, 1)$ interval by applying a transformation from the sample range to the interval $(0.05, 0.95)$. Now taking advantage of the large enough samples premise, which provides us with a strong likelihood to estimate the posterior distribution out of no prior information, we adopted uniform distributions on $(0, 1)$ for our prior distributions for the parameters a and b of the beta distribution to be estimated. And finally, as we are estimating isolated random variables, there is no point in considering covariates.

It must be highlighted that this technique refers to continuous random variables only. For discrete random variables, the marginal distribution fitting is much simpler and it is done by the usual fitting of a multinomial distribution based on sample frequencies.

Non-linear Normalization by Sample Reducing

Whenever analyzing random vectors, it is usual to normalize each feature so that its range fits a limited interval and to allow statistical profile comparison among different

features. That kind of normalization has an intrinsic linear nature and is based on position and scale fixed factors for each feature, usually the sample range or average and its variation or standard deviation. Such a normalization do not change the sample histogram profile (beyond position and scale) and the inherent random variable probability distribution nature.

Inspired by the Sklar's theorem, it is possible to mentally visualize what would be the transformed sample after submitted as an input to the actual random marginal distribution. This operation could be interpreted as a non-linear normalization and, applied to all feature random variables, the problem of modeling the joint distribution would have turned into the problem of modeling the joint distribution of uniform random variables or, equivalently, of modeling a copula function from a sample representing the cumulative probabilities of the original random variables.

Therefore, this non-linear normalization is the transformation to be applied to the sample to "stretch" it over the domain of the original random variable according to its probability mass as if it were to come from a uniform distribution. If the sampling mechanism was a perfect one, in the sense that the sampling experiment was perfectly symmetrical in probability, that would mean that this perfect sample would split the domain in equal probability mass intervals. If we also apply a position/scale to the variable domain to transform it to the interval $[0, 1]$, then we would have for a sample of size n the transformed sample $S_t = \{0, 1/(n-1), 2/(n-1), ..., 1\}$. For our real sample, the equivalent is to take the normalized rank-order, except for extreme values in both left and right sides.

It is important to observe that although the resultant image sample after the transformation is previously determined by the set S_t, what will define this non-linear normalization are the domain-side points that are transformed into them.

Obviously, fitting an empirical or a parametric distribution to the original sample are methods to estimate the probability mass distribution of a random variable, but they focus on the underlying distribution. Here the proposal is to focus on the sample itself, and, for that to gain a level of generality, we will have to assume some kind of regularization premise.

While for the distribution-oriented perspective the regularization premises where of global nature, for the sample-oriented perspective we shall adopt a local-wise premise, such as the one of sample reducing (CUNNANE, 1979), already treated previously.

3.2.5 Copula Modeling

In order to allow comparison, along with the proposed Bayesian network copula modeling, we prospected also two other traditional copula modeling: empirical copula and parameter copula. Next sections treat each one of those modeling techniques.

Empirical Copula Modeling

Before any other attempt of modeling a copula, it is very helpful to visualize what is the dependence profile among the random variables association for the subject phenomenon. This can be made by computing some of the concordance and dependence indices treated in Chapter 2 and Appendix A.

A first direct approach is to compute those indices globally throughout the entire joint distribution domain. This will lead to a notion of the overall strength degree of concordances and dependencies among variables. Nevertheless, it has to be registered that most of them are bivariate defined and some of them may not have natural extensions to the multivariate case. Therefore, it is easier to study groups of pairwise dependence than the complete distribution dependence profile.

A more indirect approach can be adopted by computing those indices in certain regions to identify different degrees of association in different regions of the domain.

A very useful tool for measure those indices from the marginals-transformed sample perspective is the empirical copula, as defined in Chapter 2.

All those indices and association analysis can point towards a specific profile that is most suitable to a given family of parametric copula allowing a further fitting based on parameter estimation, which will be discussed in the next subsection.

Parametric Copula Modeling

An option for the copula modeling stage could be using a parametric modeling for the transformed data after mapping it to the marginal distributions by using any of the methods previously discussed.

This parametric modeling consists basically of analyzing the dependence profile

of the joint distribution from its marginal-transformed sample and identifying if there is a copula parametric family with a matching dependence profile to be fitted to the transformed data.

Then, that family parameter can be estimated from the corresponding indices computed from the transformed sample.

Although there are many bivariate parametric copulas, the multivariate case is still a field in development and the copula families covering that field are fewer, the most used being the Gaussian n-copula and the series of Vine n-copulas.

For fitting data into a given parametric copula family, usually there is an association between a sample concordance or dependence measure(s) and an element of that family with its parameters estimated directly from the measure(s).

Although it was initially considered as an option, the copula parametric modeling was not used in this research simulations because of its complexity, what would compromise our time and the simplicity premise without a proportional benefit within our scope. Many parameter copula families are described in the Appendix A.

Bayesian Network Copula Modeling

In this research we propose modeling copulas using Bayesian networks. As Bayesian networks are models to represent joint distributions, and copulas are also joint distributions itself for uniform random variables which happen to be the marginal distributions of the original random variables, then Bayesian networks can also be used to model copulas.

Our approach is to use consolidated methods for modeling Bayesian networks, avoiding unnecessary complexity. As Bayesian network modeling is still an open field itself, we will restrict ourselves to prospect within a set, mixing educated-guessed structures and random ones, which are the best fitted network structures to our data by applying well-established structure scores.

Therefore, we have chosen three very used Bayesian network scores which are also implemented in a widespread software package, BIC, K2 and BDeu, with BDeu taken as the final reference score. As to the structures picking, we are going to use two internally developed routines built to implement a random structure generator and a reference

structures generator based on typical configurations (sequential, naïve Bayes, binary tree, etc) and pairwise concordance strength.

With the constructed set of possible structures for a given dataset, we proceed to those structures scoring in relation to that dataset sample and pick up the best scored structures. Applying that procedure to each normalization method, we then compare the resulting structures.

3.3 Developed Tools

Visualization of each step in the process of modeling presented itself as an essential tool for granting good progress. For that matter, as many calibrations and trials were needed for feature fittings and copula modeling, it was soon realized that a graphical interactive interface which showed the modeling procedure step-by-step would be very helpful, leading to the idea of developing the LpsCopModel software, a graphical interface for the entire modeling process, starting from data¨set choosing and data acquisition, going through features marginal distribution fitting and finishing with the copula modeling itself for completion (only empirical copula in the current version).

Although the LpsCopModel software was used in our research for more specific purposes, it is expected to be considerably helpful in many other scientific researches which need to acquire simplified models from datasets before further analysis, especially whenever there is a main concern about dependence and concordance between features and variables. Material evidence of this software relevance are many other similar initiatives in copula modeling like Paprotny *et al.* (2020) and the Data to AI Lab at MIT projects "SDV" and "copulas" Patki, Wedge e Veeramachaneni (2016), recently released in its 0.3.3 version in Sep 18, 2020, each covering different aspects, while the ones emphasized here are visualization and panoramic analysis. Nevertheless, as the software has been built under an open-platform approach, it can be easily incremented for allowing originally not included parametric distributions or copula families.

The LpsCopModel software covers the entire workflow from choosing among previously downloaded datasets to that data complete copula modeling (empirical copula, in this version). The process is split into five sequential stages: dataset selection and acquisition, data filtering and slicing, data description, marginal distributions fitting and copula modeling.

As a complete example of using the software, let us consider the use of LPSCopModel for one of the experiments with real data in this research, a DATASUS (Brazilian public healthcare system data) subset taken considering a random proportional 1% sample of all hospital admissions in Brazil through the years 2008 to 2010.

From the descriptive analysis (Figures 14 and 15), it can be seen that the subset contains 206,110 patients from all federations unities in Brazil. The most number of admissions corresponds to female patients (SEXO=3) and, from the map (in yellow), the maximum values for mean days in the hospital (DIAS_PERM) occur in Rio de Janeiro (RJ, Southeast) and Rio Grande do Norte (RN, Northeast). Also, the concordance table points

towards strong concordance between days in hospital (DIAS_PERM) and cost (US_TOT).

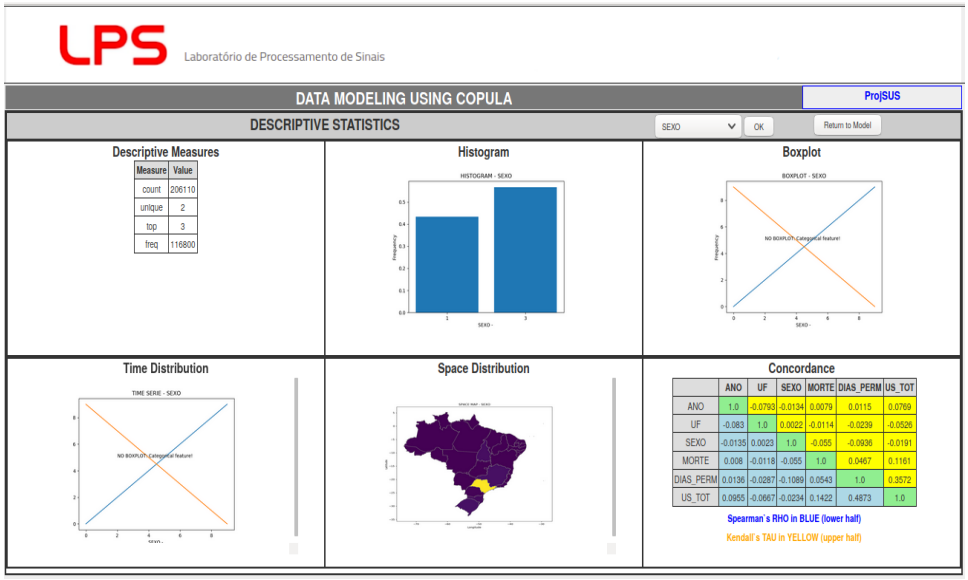


Figure 14 – Categorical feature descriptive example: hospital admissions by gender (1-male, 3-female). Measures (number of samples, gender with greater occurrences and its frequency), histogram, geographical distribution and concordance with other features. Box-plot based figures are not displayed for categorical features.

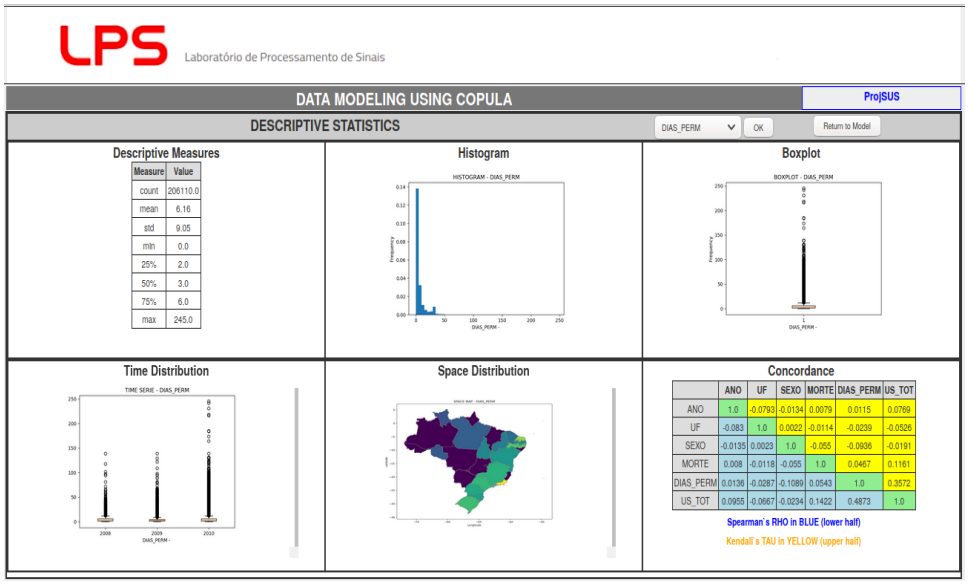


Figure 15 – Numeric feature descriptive example: days in hospital between admittance and discharge. Measures (number of samples, mean, standard deviation, minimum, maximum, quantiles), histogram, box-plots, box-plot time series, geographical distribution and concordance with other features.

Marginal distribution modeling of death (MORTE) and cost in US dollars (US_TOT) features resulted, respectively, in a multinomial with a 0.04271 probability of death and a

Beta distribution with 0.82089 for alpha and 98.5276 for beta, with standard deviations of 0.00222 and 0.35579, as Figures 16 and 17 shows. It can also be noticed that those results for cost were obtained from an MCMC ran on 5,000 samples after a 1,000 tunning stage and a significant convergence was achieved.

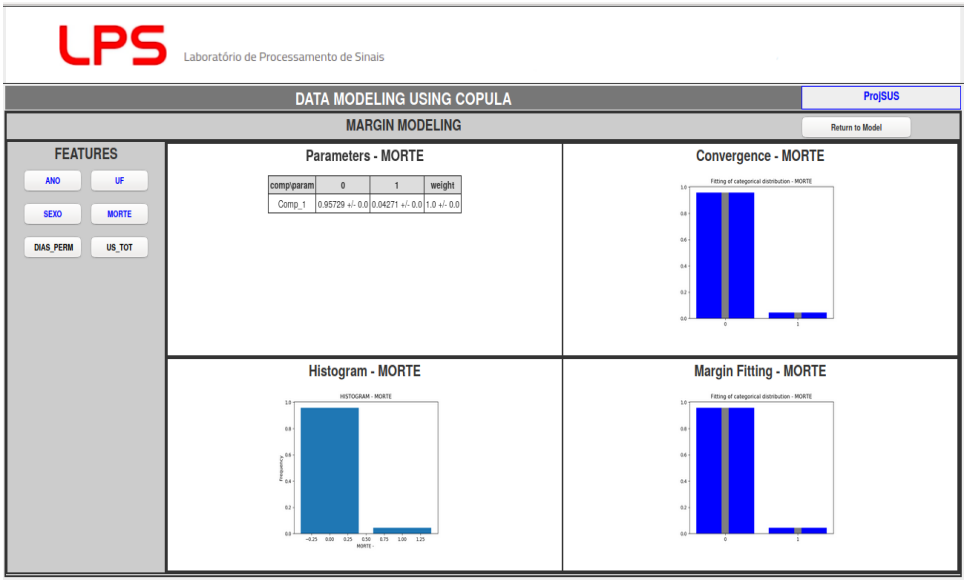


Figure 16 – Categorical feature fitting example: death in hospital. Parameters show probability by frequency estimation for each category (death in hospital or discharged alive).

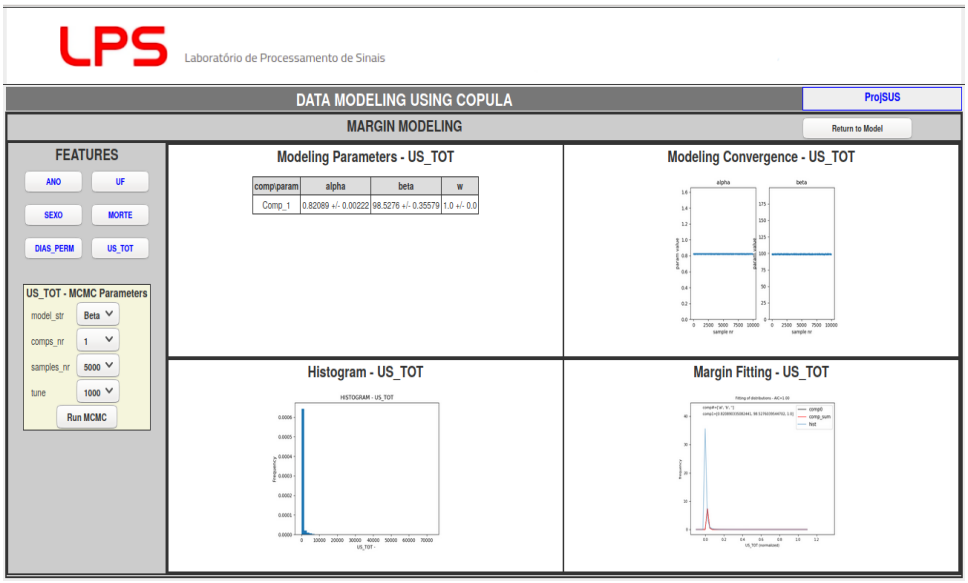


Figure 17 – Numeric feature fitting example: hospital treatment total costs in US dollars. Costs are very concentrated in low-cost area. In this case, beta fitting using MCMC (pymc3 package) resulted in a smoothed fitting for that number of samples and a spiky profile.

Finally, Figure 18 presents three two-dimension projections of feature pairs and allows to identify a growing degree of positive dependence between cost and days in hospital as the surface initially follows a W-copula similar pattern but goes toward a more M-copula pattern at the top.

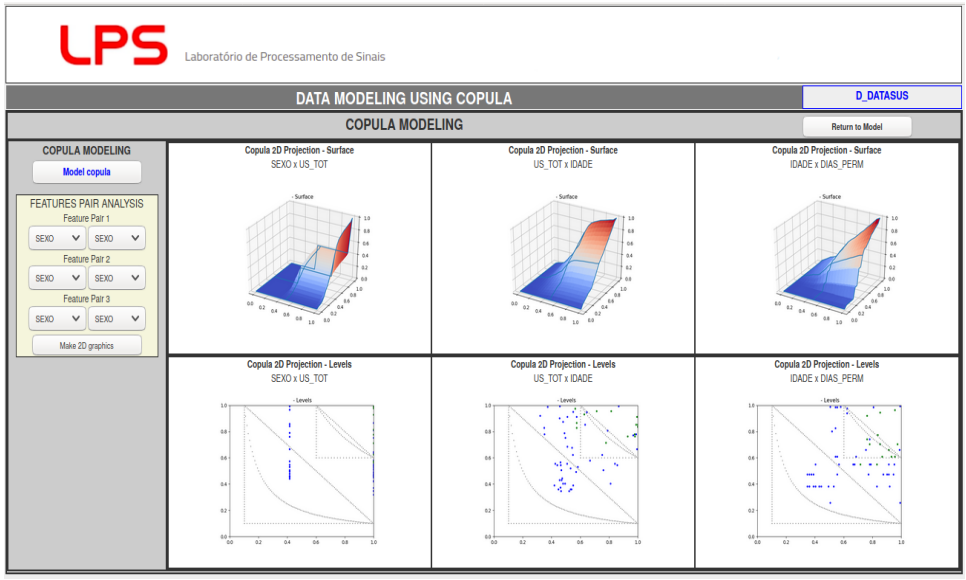


Figure 18 – Empirical copula modeling page showing features three pairs copula non smoothed flat projection surfaces and discrete footprints. Users can choose any three pairs for the corresponding copula projection to be displayed.

The complete version of the software is very newly available even in the research group where it was developed but it has also been used in previous versions before the copula module was completed for descriptive health data analysis in a publication in early 2020 (PETERLE *et al.*, 2020) and also for helping in analyzing a multimodal distributed health feature regarding a disease study conducted by a multidisciplinary team. In both cases, it helped a lot by saving time and producing global and systematic phenomenon visualization for each research team.

As a general tool for modeling, this software is intended to be used in a widespread range of areas, wherever modeling is involved, especially when focusing concordance and dependence issues, with no previous restrictions.

LpsCopModel has a systemic approach and is intended to contribute in saving time and giving a panoramic visualization of a phenomenon by producing a first model based on MCMC parametric marginal distributions fitting and empirical copula modeling, where users can try some modeling options and grade complexity according to their needs.

It was developed for general use and therefore any dataset which can be registered in CSV format or converted to a pandas data frame can be input to be treated. In the same way, every figure and results can be exported, from parameters values and intervals to MCMC traces.

Simultaneously, by being open-source and modular, users can also upgrade by easily implementing by themselves new model options they happen to need or evolving the copula modeling to parametric models. In parallel, it is intended by the developer group to continue improving the software aggregating new modeling facilities, and enhancing the user interface.

3.4 Software and Equipment

The work has been done in the Signal Processing Laboratory in the Dept. of Electrical and Computer Engineering from the Engineering School at São Carlos, University of São Paulo, Brazil.

The equipment there available for this research is:

1. one portable computer HP 64-bit Intel® Core™ i3-7100U CPU @ 2.40GHz × 4, 7.7GiB, Intel® HD Graphics 620 (Kaby Lake GT2), with operational system Ubuntu 16.04 LTS, used for developiong algorithms, doing tests and registrying results and documents;

2. a cluster of computers for parallel processing on Ubuntu Server 14.04.1 LTS using Python 2.7.6 and IPython 1.2.1 (not used yet, but available for using in next phases for intensive processing).

In parallel with the operational systems and platforms already described, the following specific software packages were used:

1. Python 3.5.2

2. IPython 2.4.1

3. Scipy and Numpy libraries from SciPy.org

4. Matplotlib 1.4.0

5. Pebl - Python Environment for Bayesian Learning 1.0.2, from University of Michigan's Systems Biology Lab, avialable via MIT license;

6. pandas - for dataframes
7. chaospy - for copulas
8. tensorflow - for neural networks
9. keras - for enveloping tensorflow userfriendly
10. statsmodels - distributions models
11. vincent - for geographic maps online
12. altair - for geographic maps on local computer
13. geopandas - for geographic maps association to pandas ´dataframes
14. pymc3 - for MCMC processing
15. pypmg - for BN modeling and scoring

4 RESULTS AND DISCUSSION

For testing the methodology we have chosen five groups of datasets, aiming to comprise both simulated datasets for initial controlled testing specific modeling features, and real datasets. The first three simulated groups consist of four datasets on the same general profile but differing on the dependence type among its random variables: independent, positive dependent, negative dependent (in a generalized sense, when mode than two variables), and intermediate dependent. The last two groups consist of one dataset each and reflect real data from healthcare and tax administration phenomena, respectively.

- **Experiment 01** - Bivariate (2D) Unimodal - 4 datasets:
  - independent random variables
  - positive dependent random variables
  - negative dependent random variables
  - intermediate dependent random variables
- **Experiment 02** - Bivariate (2D) Trimodal - 4 datasets:
  - independent random variables
  - positive dependent random variables
  - negative dependent random variables
  - intermediate dependent random variables
- **Experiment 03** - Multivariate (6D) Unimodal - 4 datasets:
  - independent random variables
  - positive dependent random variables
  - negative dependent random variables
  - intermediate dependent random variables
- **Real Case 01** - Brazilian Public Healthcare System (7D) - 1 dataset
- **Real Case 02** - Brazilian Counties Tax Revenue (11D) - 1 dataset

4.1 Experiment 01 - Bivariate Unimodal Phenomenon

We have started by the most simple test possible with enough substance, and it is the case of a bivariate continuous unimodal distribution with independent components.

Let a given phenomenon family be represented by two continuous random variables $x1$ and $x2$, whose samples are generated by a bivariate unimodal normal model with means 1.0 and 2.0 and standard deviations equally 1.0 and 2.0. Its marginal probability densities are showed graphically in Figure 19.

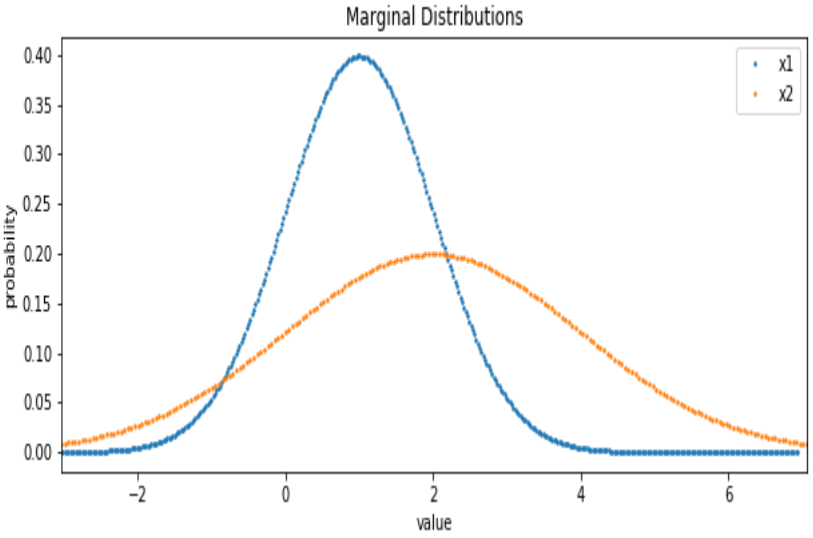
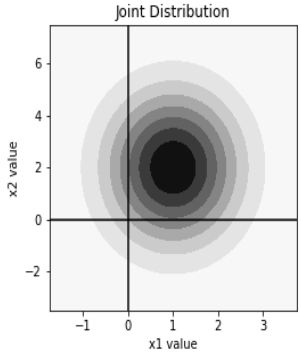


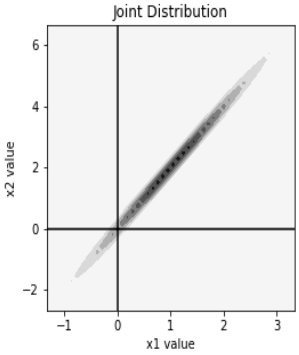
Figure 19 – Bivariate independent normal unimodal experimental dataset - marginal probability densities.

It can be seen that $x1$ and $x2$ random variables range throughout the entire Real line, while many samples are concentrated on 1.0 and 2.0 neighborhoods, respectively.

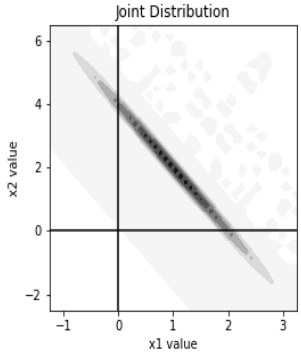
Now we consider four possible members of that family regarding their variables dependence relation: one where the variables are totally independent (correlation 0.0), one where they are almost totally positive dependent (correlation +0.999), one where they are almost totally negative dependent (correlation −0.999) and one where they are at a random intermediate grade of dependence (correlation randomly chosen at 0.3 to be between 0.1 and 0.9, positive or negative). The corresponding joint probability densities are graphically represented by its level curves in Figure 20.



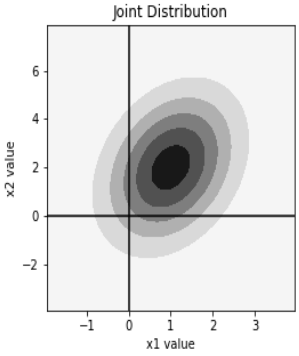
(a) Independent variables.



(b) Positive dependent variables.



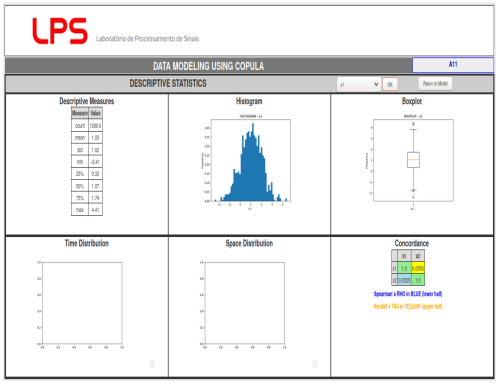
(c) Negative dependent variables.



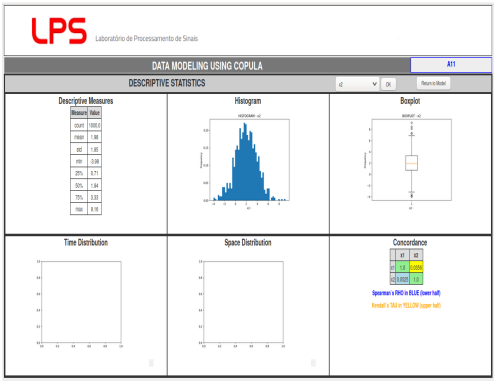
(d) Intermediate degree dependent variables.

Figure 20 – Bivariate normal experimental datasets - joint probability densities level curves.

With our test dataset already made, we can now run the entirely methodology on it. We start by using the analysis software tool for giving a general overview of data and modeling. As a first step, Figure 21 show a statistical description of both $x1$ and $x2$ features as presented by LpsCopModel tool in the case of the independent dataset and Figure 22 shows each one of those graphics for $x1$, while Table 3 presents a consolidation of the variables statistics measures, for better readability. The other three datasets have similar results. In parallel, Table 4 shows the concordance measures presented in the corresponding tool screen tables for all four datasets.

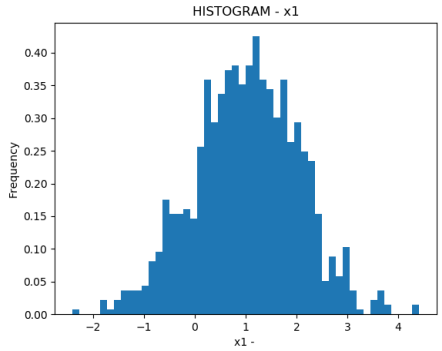


(a) Feature $x1$ descritpive statistics.

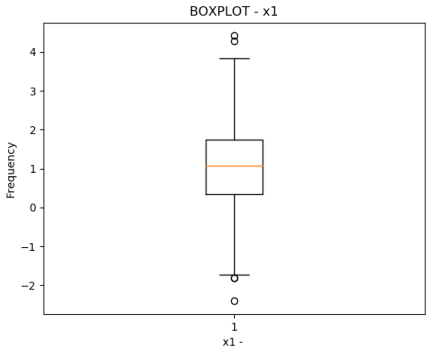


(b) Feature $x2$ descritpive statistics.

Figure 21 – Bivariate normal unimodal independent experimental dataset - descriptive statistics screen in LpsCopModel tool.



(a) Feature $x1$ histogram.



(b) Feature $x1$ boxplot.

Figure 22 – Bivariate normal unimodal experimental dataset - descriptive statistics individual graphics.

Table 3 – Descriptive measure numbers for the bivariate normal unimodal experimental dataset random variables.

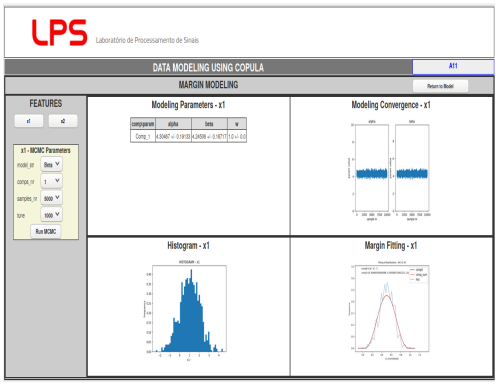
Measure	$x1$ **value**	$x2$ **value**
count	1,000	1,000
mean	1.03	1.98
std dev	1.02	1.95
min	−2.41	−3.98
25%	0.33	0.71
50%	1.07	1.94
75%	1.74	3.33
max	4.41	9.16

Table 4 – Corcondance pairwise values for the bivariate normal unimodal experimental datasets random variables. Rho stands for Spearman's rho and Tau for Kendall's tau.

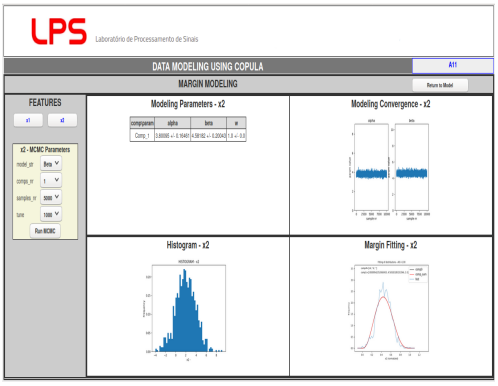
Dataset	Variables Pair	Rho	Tau
Independent	$x1$-$x2$	0.0525	0.0356
Positive dependent	$x1$-$x2$	0.9989	0.9724
Negative dependent	$x1$-$x2$	-0.9988	-0.9717
Intermediate dependent	$x1$-$x2$	0.3035	0.2050

4.1.1 MCMC Marginal Distribution Fitting

Still using LpsCopModel tool, a step further is to model each feature marginal distribution with a Bayesian parametric MCMC technique. This can be done by choosing an adequate parametric distribution, a prior distribution (usually a non-informative one, in the absence of specialists previous knowledge) and running an MCMC modeling. For this test, a non-informative uniform prior were taken and the parametric distribution was a (two-parameter) Beta distribution. Before modeling, samples were normalized from its original Gaussian range to fit the Beta range, as explained in Chapter 3. The results are presented in Figures 23 and 24.

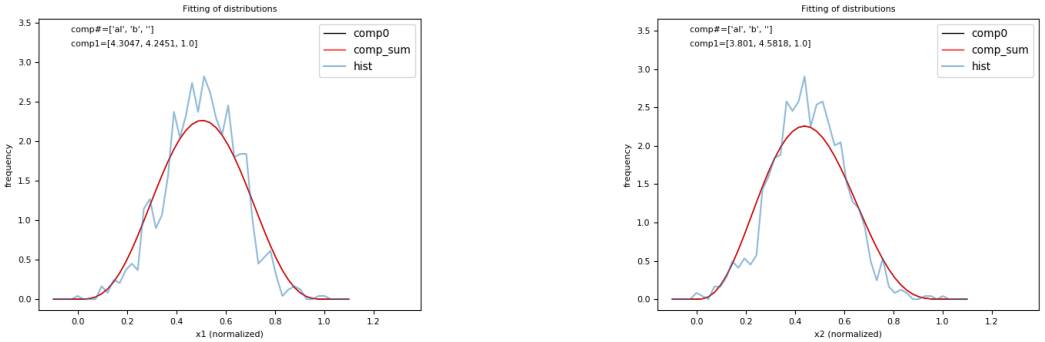


(a) Margin fitting for variable $x1$.



(b) Margin fitting for variable $x2$.

Figure 23 – Bivariate normal unimodal independent experimental dataset - marginal fitting with a parametric Beta and non-informative uniform prior by Bayesian MCMC modeling. Tool screen blocks show, respectively, distribution estimated parameters with standard deviation, convergence graphics for two chains, original histogram and fitted distribution plot. Marginal fitting for all three other datasets are similar.



(a) Margin fitting for variable $x1$.

(b) Margin fitting for variable $x2$.

Figure 24 – Bivariate normal unimodal independent experimental dataset - marginal fitting graphics with a parametric Beta and non-informative uniform prior by Bayesian MCMC modeling. Marginal fitting for all three other datasets are similar.

All the marginal distribution fitting parameters for each bivariate unimodal dataset are presented at Table 5.

Table 5 – Margin fitting normalized parameters for the bivariate unimodal datasets.

Dataset	Variable	Distribution	Parameter a	Parameter b
Independent	$x1$	Beta	4.9081	4.3839
	$x2$	Beta	3.4180	2.9951
Positive dependent	$x1$	Beta	4.1335	4.2661
	$x2$	Beta	3.4625	3.2525
Negative dependent	$x1$	Beta	3.9057	4.5626
	$x2$	Beta	3.3627	3.8832
Random dependent	$x1$	Beta	3.6006	3.4188
	$x2$	Beta	3.9705	4.1767

4.1.2 Empirical Copula Modeling

In the possession of marginal distribution parametric models, an empirical copula modeling is a good instrument for a first general view of dependence profiles. The LpsCopModel tool easily provides this analysis in a pairwise perspective. Figure 25 shows such an overview for the independent dataset on the software screen, while Figure 26 presents the graphics for all four datasets (independent, positive dependent, negative dependent and intermediate dependent).

Those figures show two different graphics: one for the empirical copula two-variable 2d-projection surface (which in this bivariate dataset case is the own full copula) and another for a sample of that projection footprint in the two variables plane. The latter figure presents two levels, copula image probabilities $p = 0.1$ and $p = 0.6$, and the copula dots in 0.15 large neighborhoods of those levels, where the triangles represent the contour limits determined by the W and M Frechet copulas for copula dots exactly at those p values.

Further those analysis will be useful in helping searching Bayesian network structures for more complex datasets copula modeling.

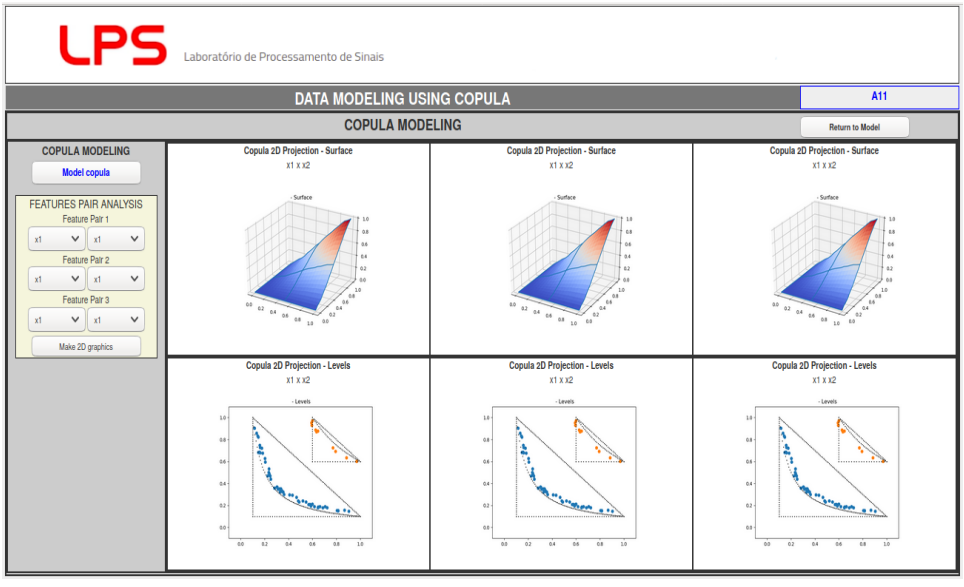
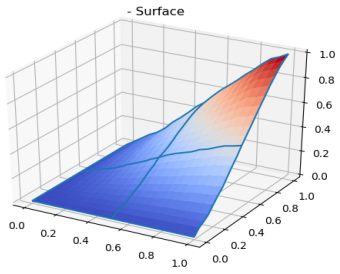
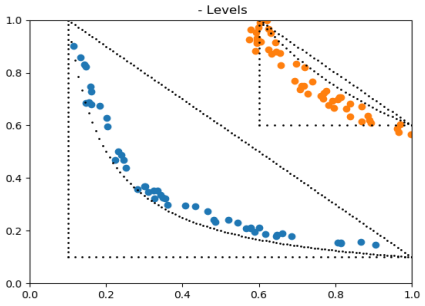


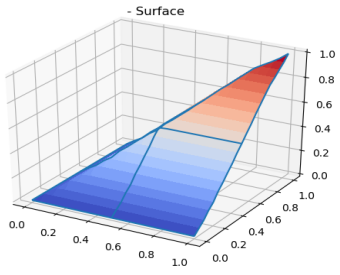
Figure 25 – Bivariate independent experimental dataset - empirical copula figures in LpsCopModel software tool. As there are only two variables in this dataset, all three copula projection figures show the same $x1$-$x2$ projection.



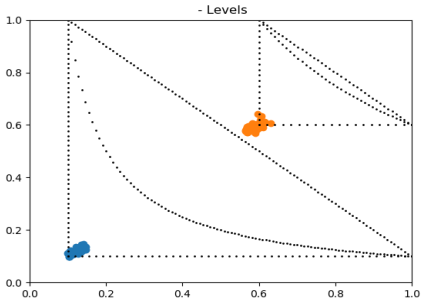
(a1) Independence - copula surface.



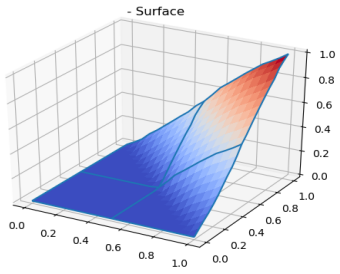
(a2) Independence - level curves.



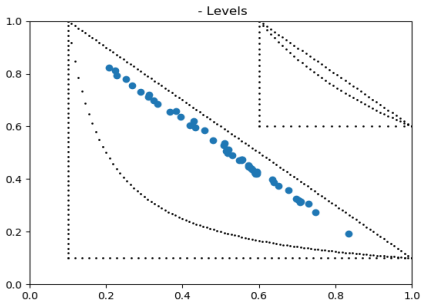
(b1) Positive dependence - copula surface.



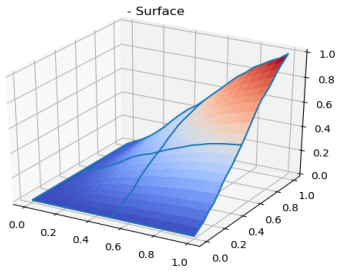
(b2) Positive dependence - level curves.



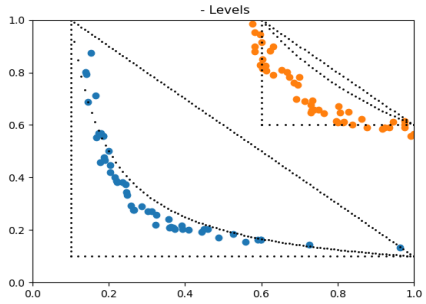
(c1) Negative dependence - copula surface.



(c2) Negative dependence - level curves.



(d1) Random dependence - copula surface.



(d2) Random dependence - level curves.

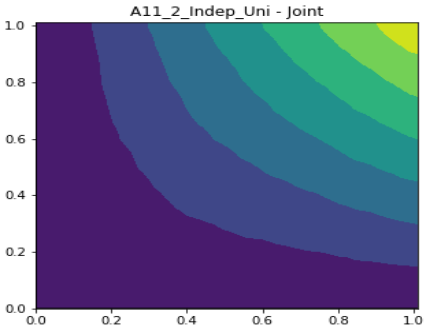
Figure 26 – Bivariate unimodal experimental dataset - empirical copula 2D projection surfaces and level curves. All dots projections in the p-level plane remain in a tight neighborhood of the corresponding limit curve.

Figure 26 shows that all dots projections in the p-level plane remain in a tight neighborhood of the corresponding limit curve: $u.v = p$ for the independent case, main diagonal for the positive dependent case, secondary diagonal for the negative dependent case, and an intermediary curve for the 0.3 correlation case. Particularly, as the limiting triangle goes up for higher levels, it becomes smaller and the product copula projection tends to straighten and to approximate the linear upper limit; thus the difference between negative and independent associations became less visible for higher levels and the intermediate case shows higher level dots further from the corresponding curved limit than in the bigger triangle (lower level). Regarding the positive dependence, dots tend to collapse to a small circle around the right angle vertex as a correspondence to the scatter-plot main diagonal footprint. And, finally, for the negative case, the only dots projection is the one over the secondary diagonal, hence there will be no footprint far from that diagonal, and that is why no dot appears in the smaller triangle.

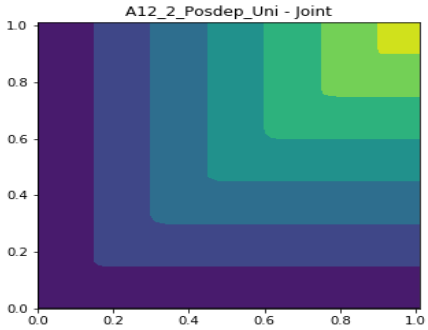
4.1.3 Non-Linear Normalization by Sample Reducing

Our non-linear normalization by sample reducing methodology proposes to apply a probabilistic transformation to constraint the dataset random variables to the $[0.0, 1.0]$ interval and only then stepping up to the Bayesian network modeling stage by the blessings of Sklar's theorem.

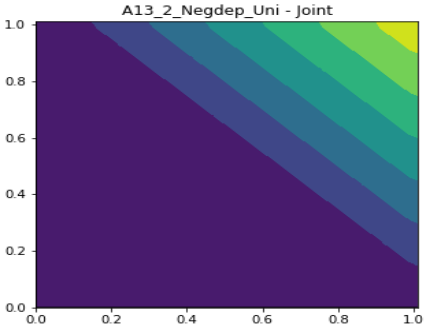
As expected, the marginal distributions became uniform distributions when normalized by sample reducing, the individualization of each variable being materialized in the corresponding denormalizing transformation. By its turn, the reduced random variables joint distribution, which is also the copula of both these normalized variables and the original ones, shows the typical dependence pattern concerning each one of the four cases (Figure 27).



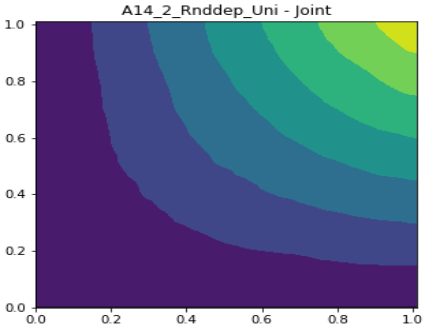
(a) Independence.



(b) Positive dependence.



(c) Negative dependence.



(d) Intermediate degree dependence.

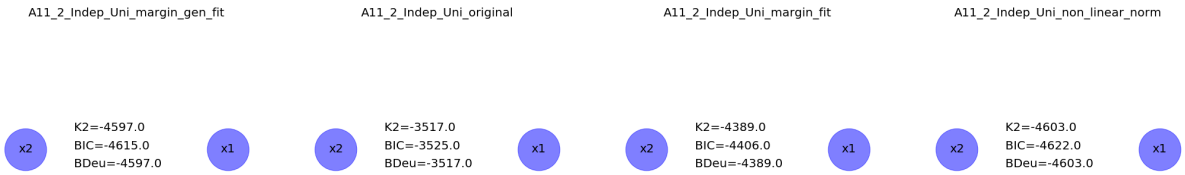
Figure 27 – Bivariate normal experimental dataset - joint distributions level curves.

4.1.4 Bayesian Network Copula Modeling

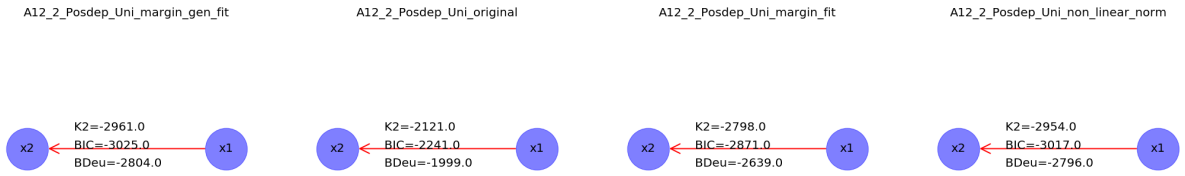
For these bivariate datasets, the Bayesian network structure is no big deal, because there can be only three possible structures: totally disconnected, connected from $x1$ to $x2$ and otherwise. Therefore, no previous dependence analysis need to be conducted for determining structure search strategies and we can focus on the scores for those three possible structures.

The Bayesian network modeling in this research is based on a traditional discrete approach for avoiding introducing another degree of complexity to the analysis. That said, a simple standard linear discretization is placed taking the minimum between 10 and the integer part of the square root of the number of samples as the number of bins.

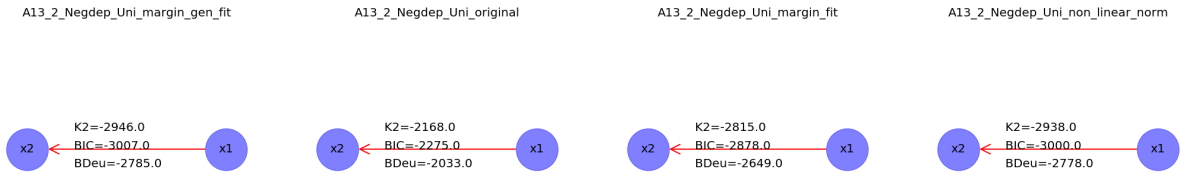
Figure 28 shows the Bayesian network structure best scored and its scores for generator distribution normalization (using the known original marginal distribution for this test dataset), no normalization, MCMC marginal distribution fitting and non-linear normalization. Although all three scores (K2, BIC, and BDeu) are showed, BDeu was taken as reference here and in all subsequent datasets.



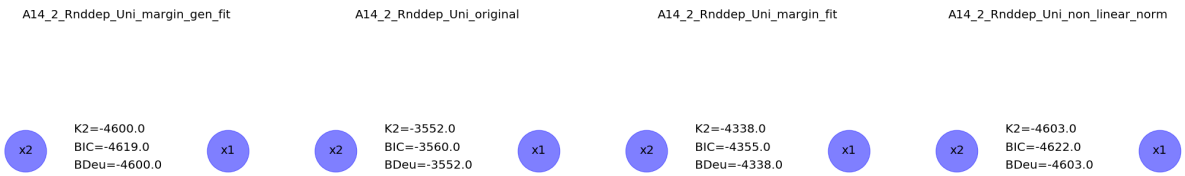
(a) Independent variables.



(b) Positive dependent variables.



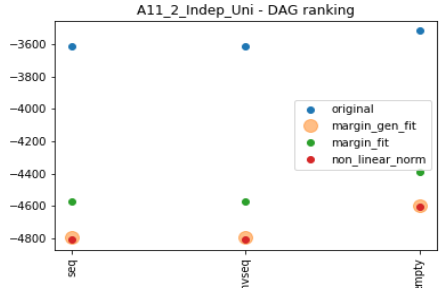
(c) Negative dependent variables.



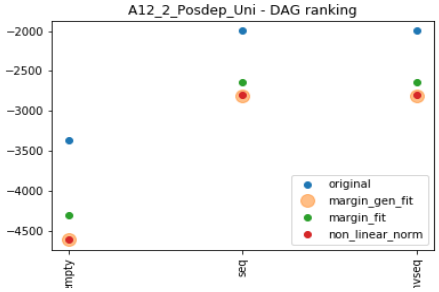
(d) Random dependent variables.

Figure 28 – Comparison of Bayesian networks for all the four normalizations applied to the bivariate normal distribution cases. Each row corresponds to a specific dependence type dataset and the columns to a different normalization, from left to right: real marginal distributions, none, fitted marginal distributions, and non-linear normalization. The figures show also the score measures for comparison. Isolated (non connected) dots stands for a totally unconnected network structure, the one where all random variables are modeled as independent.

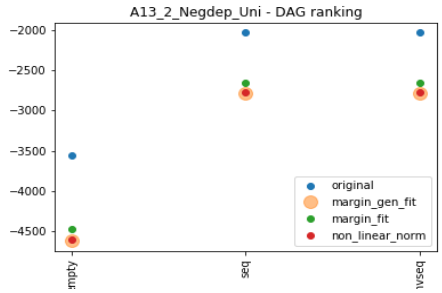
For better comprehension and interpretation of the results, Figure 29 graphically shows where each possible network structure stays in terms of score ranking. It can be clearly noticed that none of the normalization methods has interfered with the score ranking, all methods preserved the ranking order in relation to the no normalization method ranking. A remarkable result was the match between non-linear normalization and generative distribution sample transformations, indicating that, in this experiment, the non-linear was the most representative of the original population behavior.



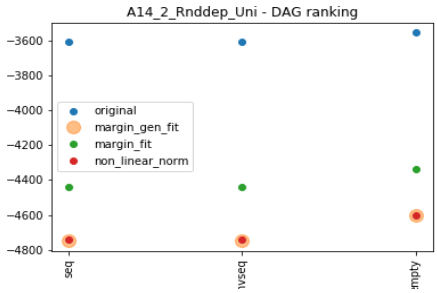
(a) Independent variables.



(b) Positive dependent variables.



(c) Negative dependent variables.



(d) Intermediate dependent variables.

Figure 29 – Comparison of original, marginal distribution fitting and normalized Bayesian networks score ranking for the bivariate normal distribution cases. In the figures, each Bayesian network structure is a colored dot in the graphs where the horizontal axis represents different network structures and the vertical stands for scores. Different dot colors represent different normalization techniques: blue is for no normalization (original dataset), green for marginal distribution fitting normalization, red for the proposed non-linear normalization, and light orange for the generative distribution normalization.

4.2 Experiment 02 - Bivariate Multimodal Phenomenon

Going a degree of complexity further in the bivariate phenomenon, we will analyze the case of a multimodal bivariate distribution, here obtained as a mixture of three bivariate normal distributions.

We will again suppose a family of phenomena with two continuous random variables $x1$ and $x2$, whose samples are generated by a mixture model of three bivariate normal distributions with means $(1.0, 2.0)$, $(4.0, 5.0)$ and $(7.0, 8.0)$ and individual standard deviations $(1.0, 2.0)$ equally for all components, while the covariation values will be established according to each dependence assumption just as in the previous case. The weight vector for the mixture is $(0.35, 0.45, 0.2)$. Its marginal probability densities are showed graphically in Figure 30.

While $x1$ and $x2$ range throughout the entire Real line, samples are concentrated on the three neighborhoods defined by the components means $(1.0, 2.0)$, $(4.0, 5.0)$, $(7.0, 8.0)$. It is noticeable that the $x2$ random variable marginal distribution happens to be approximately unimodal, despite the joint distribution trimodal nature, because its standard deviation is twice that of $x1$, what leads to modes partial superposition in the $x2$ projection.

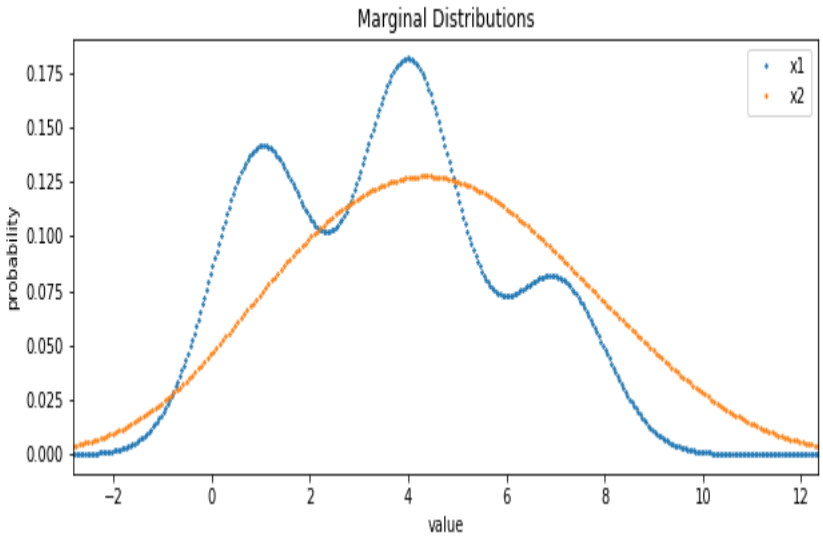
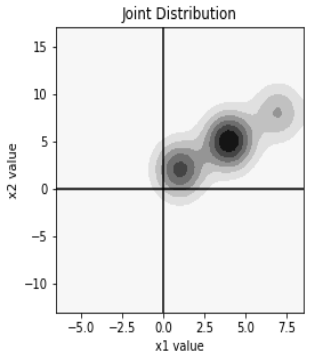


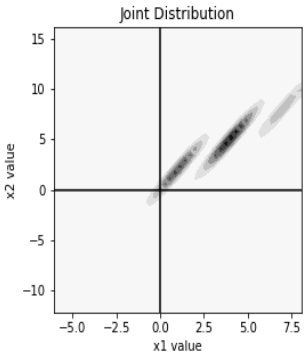
Figure 30 – Bivariate trimodal mixture of normal distributions experimental datasets - marginal probability densities.

Just as in the previous case, we will consider four possible variables dependence relations: total independence, total positive dependence, total negative dependence and an

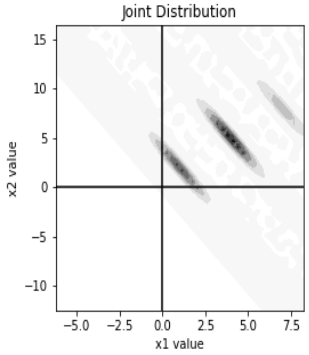
intermediate grade of dependence arbitrarily fixed at 0.3 correlation. The corresponding joint probability densities are graphically represented by its level curves in Figure 31.



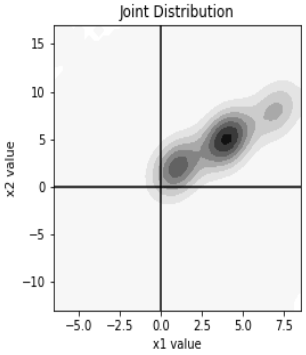
(a) Independent variables.



(b) Positive dependent variables.



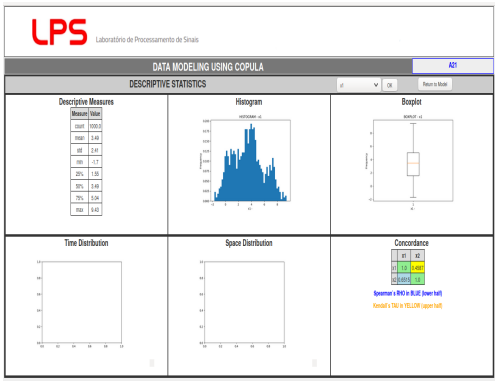
(c) Negative dependent variables.



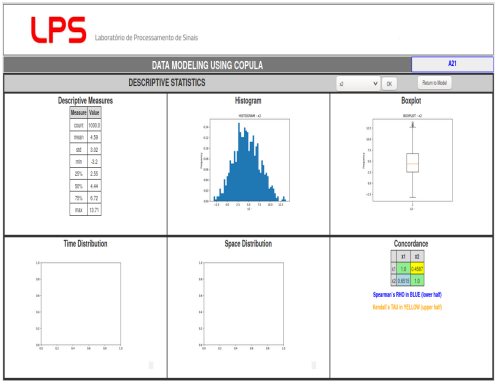
(d) Intermediate degree dependent variables.

Figure 31 – Bivariate trimodal mixture experimental datasets - joint probability densities level curves.

Applying the analysis software tool, a statistical description of all variables is available, and so $x1$ and $x2$ features characteristics can be seen in Figure 32 for the independent case, with the other three cases with similar individual behavior.



(a) Feature $x1$ descritpive statistics.



(b) Feature $x2$ descritpive statistics.

Figure 32 – Bivariate normal trimodal independence experimental datasets - descriptive statistics.

Table 6 presents a consolidation of the variables statistics measures again for the independent case as an example, while Table 7 shows the concordance measures for all four datasets. It is important to remark that the mixture of three joint distributions, although each one with very similar and well-defined dependence relations between their variables, did not assured the same behavior for the resultant mixture joint distribution in all cases. Hence, the mixture of three independent variables (near zero correlations) generated a medium positive dependent distribution (concordance indexes of 0.6515 and 0.4587), while the mixture of negative dependent variables (near -1.0 correlations) ended up in a positive dependent resultant distribution (concordance indexes of $+0.4388$ and $+0.1368$); consequently, we expect more undefined dependence patterns for this datasets than for the previous one.

Table 6 – Descriptive measure numbers for the bivariate normal unimodal experimental dataset random variables.

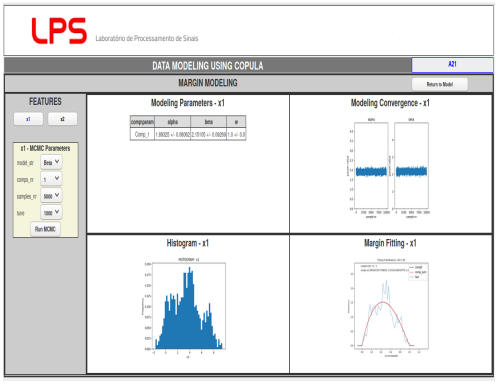
Measure	$x1$ **value**	$x2$ **value**
count	$1,000$	$1,000$
mean	3.49	4.59
std dev	2.41	3.02
min	-1.70	-3.20
25%	1.55	2.55
50%	3.49	4.44
75%	5.04	6.72
max	9.43	13.71

Table 7 – Corcondance pairwise values for the bivariate normal unimodal experimental dataset random variables. Rho stands for Spearman's rho and Tau for Kendall's tau.

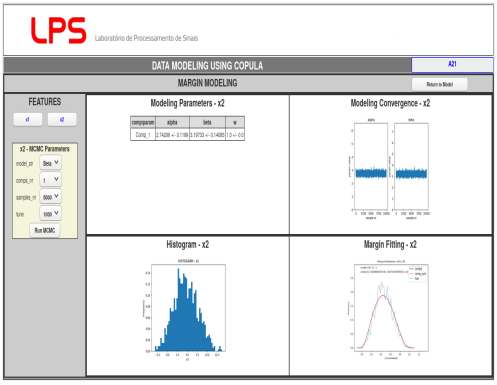
Dataset	**Variables Pair**	**Rho**	**Tau**
Independent	$x1$-$x2$	0.6515	0.4587
Positive dependent	$x1$-$x2$	0.9452	0.8428
Negative dependent	$x1$-$x2$	0.4388	0.1368
Intermediate dependent	$x1$-$x2$	0.7709	0.5709

4.2.1 MCMC Marginal Distribution Fitting

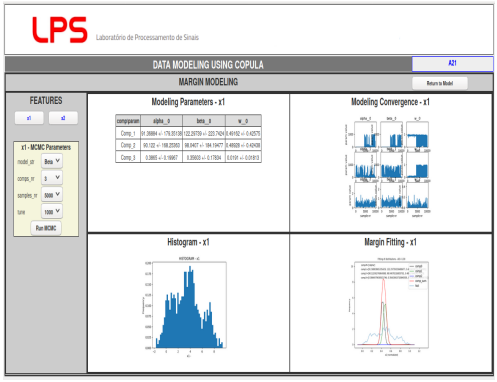
Each feature marginal distribution was then modeled by the MCMC module in the LpsCopModel software modeling with a non-informative uniform prior and a (two-parameter) Beta distribution. As this time we have a trimodal distribution strongly reflected on the variable $x1$, as seen in that variable histogram, we have tried not only a single beta component MCMC modeling, but also some mixture with 3 and 5 components. The results for the mixture simulations did not converge for the same MCMC parameters ($5,000$ samples after a $1,000$ tunning sample), thus the single Beta distribution was chosen, although convergence necessarily does not mean good fitting. All MCMC modeling results are presented in Figures 33 and 34.



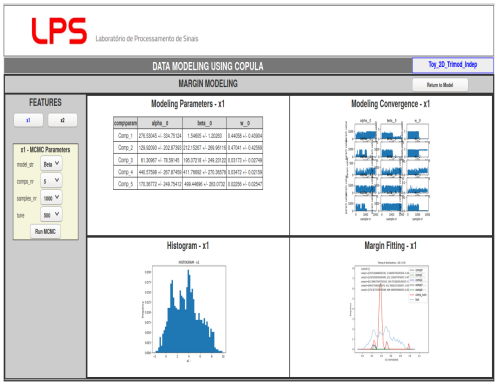
(a) $x1$ single component marginal fitting.



(b) $x2$ single component marginal fitting.

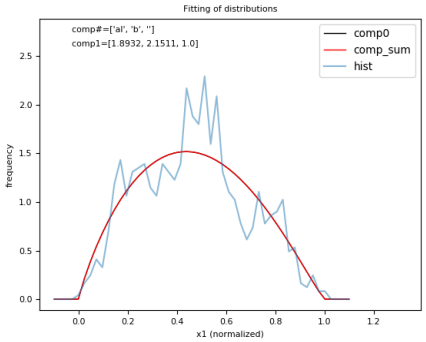


(c) $x1$ three component marginal fitting.

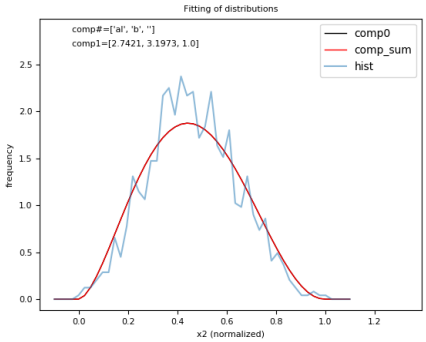


(d) $x1$ five component marginal fitting.

Figure 33 – Bivariate trimodal normal experimental dataset - marginal distribution fitting with a parametric Beta and non-informative uniform prior by Bayesian MCMC modeling. Figures (c) and (d) shows a three and five component MCMC modeling for apparently trimodal variable $x1$, but the fitting using the same MCMC tunning parameters proved to be poor and did not converge for the standard parameters used.



(a) $x1$ single component marginal fitting.



(b) $x2$ single component marginal fitting.

Figure 34 – Bivariate normal experimental dataset - marginal distribution fitting graphics with a parametric Beta and non-informtive uniform prior by Bayesian MCMC modeling.

Table 8 presents all the marginal distribution fitting parameters for each bivariate trimodal dataset.

Table 8 – Margin fitting normalized parameters for the bivariate unimodal datasets.

Dataset	Variable	Distribution	Parameter a	Parameter b
Independent	$x1$	Beta	1.8933	2.1511
	$x2$	Beta	2.7421	3.1973
Positive dependent	$x1$	Beta	2.2364	2.8408
	$x2$	Beta	3.2048	4.0908
Negative dependent	$x1$	Beta	1.9402	2.5321
	$x2$	Beta	3.7469	4.5363
Random dependent	$x1$	Beta	1.5801	1.9041
	$x2$	Beta	3.0843	3.2484

4.2.2 Empirical Copula Modeling

Using LpsCopModel to acquire an empirical copula modeling just as in the previous case, Figure 35 shows such an overview for the independent dataset on the software screen, while Figure 36 presents the graphics for all four datasets (independent, positive dependent, negative dependent and intermediate dependent) of this family.

As expected, the copula profile does reflect the less defined dependence relations caused by a mixture of well-defined individual distributions, but there are still some detectable regularity. The independent mixture shows a well-distributed plot in a pattern which goes along a parallel curve in relation to the independence reference curve, but now somewhere between that reference and the positive reference (the triangles cathetus). A similar behavior is presented by the intermediate dependence mixture. By its turn, the positive dependent mixture still concentrates its dots basically over the triangle positive dependence reference sides, although not on the vertex any more. The negative dependence mixture presents very distinguishable lines of concentration with angles in the same quadrant as the secondary diagonal (negative dependence reference) and in the same number as the components of the mixture (three distinct lines).

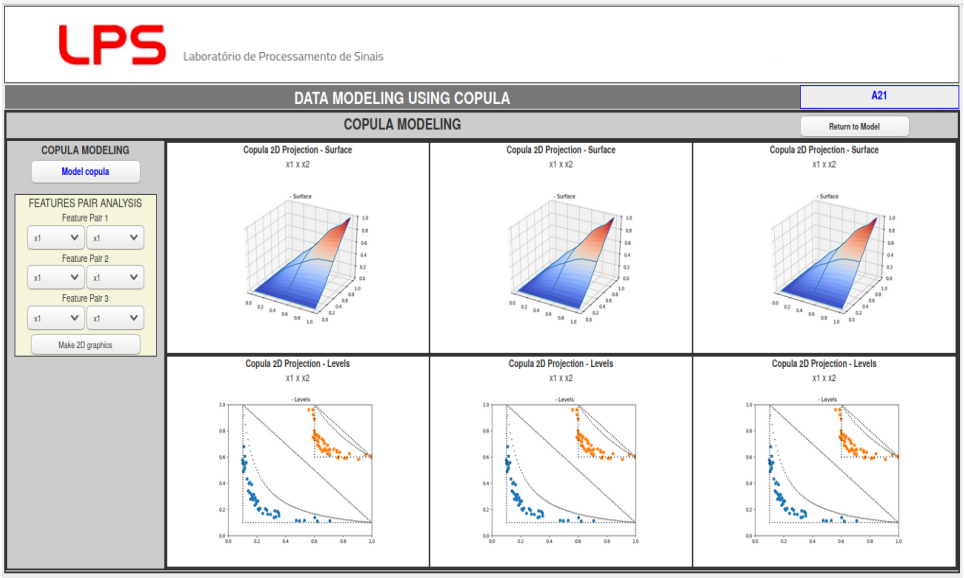
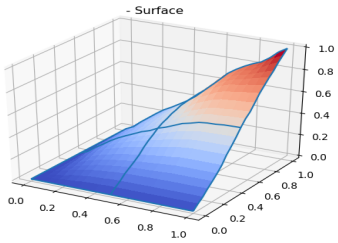
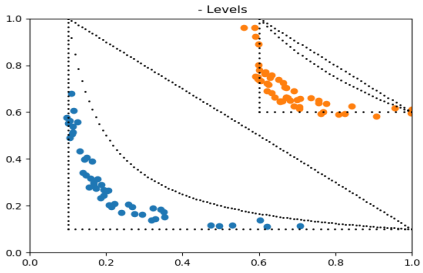


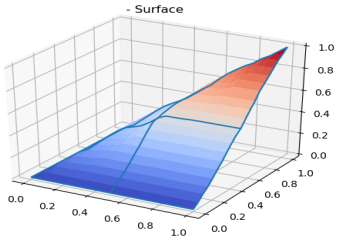
Figure 35 – Bivariate independent experimental dataset - empirical copula figures in LpsCopModel software tool.



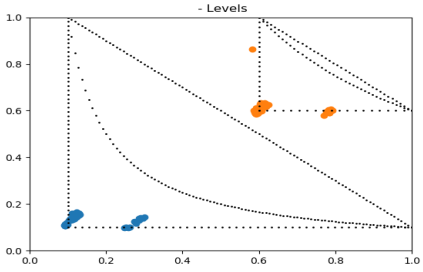
(a1) Independence - surface.



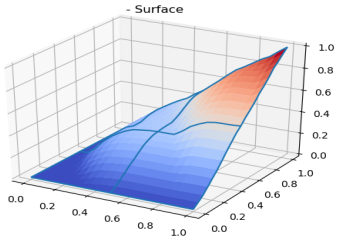
(a2) Independence - level curves.



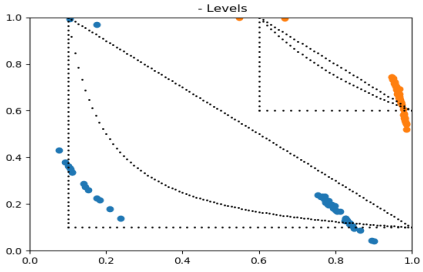
(b1) Positive dependence - surface.



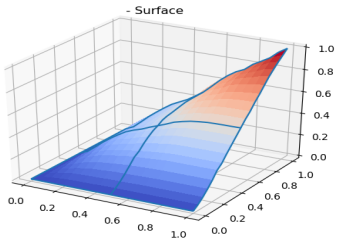
(b2) Positive dependence - level curves.



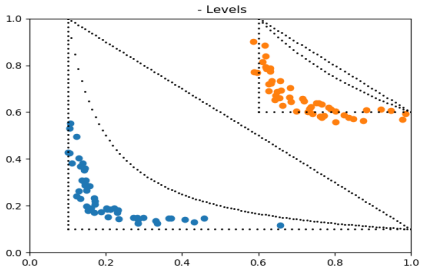
(c1) Negative dependence - surface.



(c2) Negative dependence - level curves.



(d1) Intermediate dependence - surface.



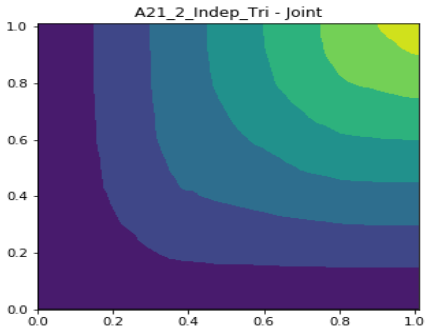
(d2) Intermediate dependence - level curves.

Figure 36 – Bivariate trimodal experimental dataset - empirical copula 2D projection surfaces and level curves.

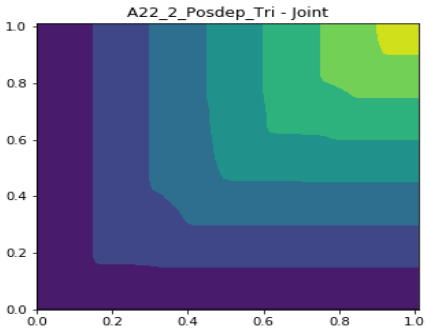
4.2.3 Non-Linear Normalization by Sample Reducing

Again, the margins become uniform distributions when normalized by sample reducing, and, just as before, the normalized joint distribution (also a copula) shows the expected dependence pattern for the independent, positive dependent, and intermediate

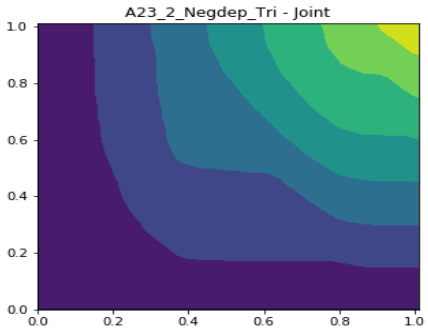
dependent case (Figure 37) although with a more noticeable distortion than for the unimodal case, totally justified as the empirical copula footprint analysis already treated in previous subsection, while for the negative case there was no identifiable pattern, which is also explained by that previous analysis, considering the dots pattern splitting in three different lines which causes a considerable fluctuation in the resulting random variable addition along a secondary axis.



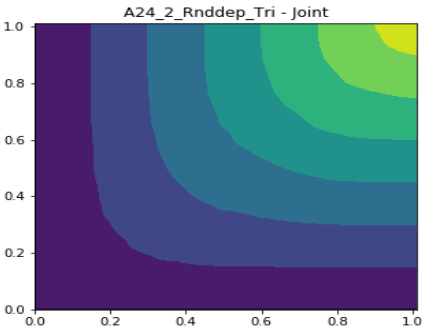
(a) Independent variables.



(b) Positive dependent variables.



(c) Negative dependent variables.



(d) Intermediate degree dependent variables.

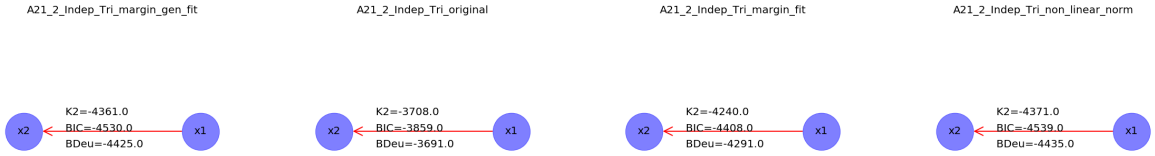
Figure 37 – Bivariate trimodal mixture experimental dataset - normalized joint distributions.

4.2.4 Bayesian Network Copula Modeling

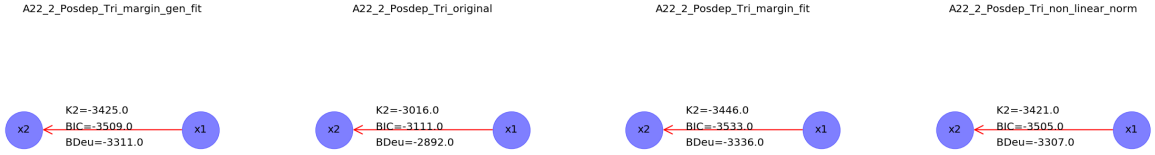
Again, as the number of variables is also 2, there can be only three possible structures and no dependence analysis is needed. Therefore, the Bayesian network structure and its scores for generator, original, MCMC marginal distribution fitting and normalized datasets are those on Figure 38.

Considering that the mixture caused in all four cases a disturbance in the strong original dependence pattern of each component, there is no independence network preva-

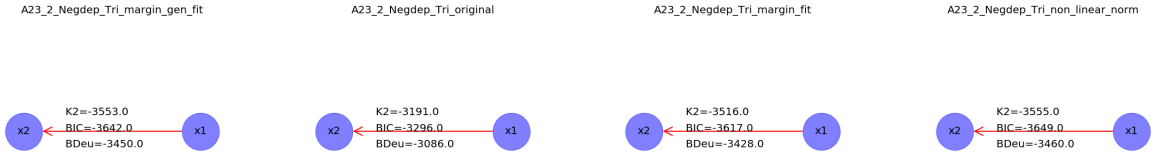
lence in any one of the datasets. Nevertheless, for the interest of this research, the relevant point is that all normalization methods pointed to the same best network structure.



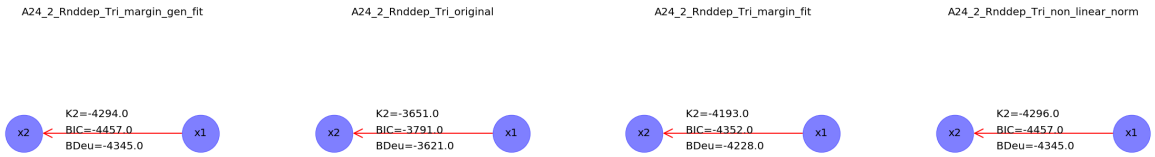
(a) Independent variables.



(b) Positive dependent variables.



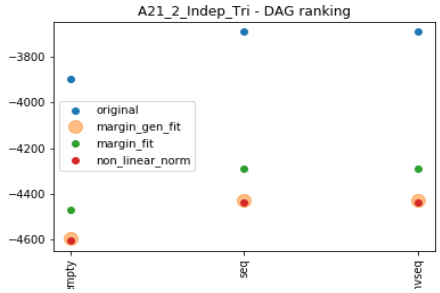
(c) Negative dependent variables.



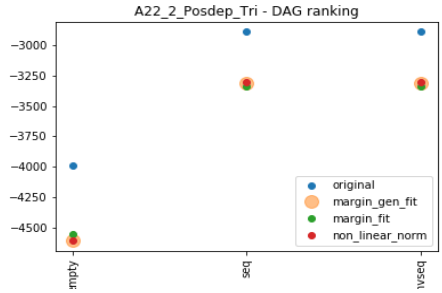
(d) Random dependent variables.

Figure 38 – Comparison of Bayesian networks for all the four normalizations applied to the bivariate normal trimodal distribution cases. Each row corresponds to a specific dependence type dataset and the columns to a different normalization, from left to right: real marginals, none, fitted marginals, and non-linear normalization. The figures show also the score measures for comparison. Isolated (non connected) dots stands for a totally unconnected network structure, the one where all random variables are modeled as independent.

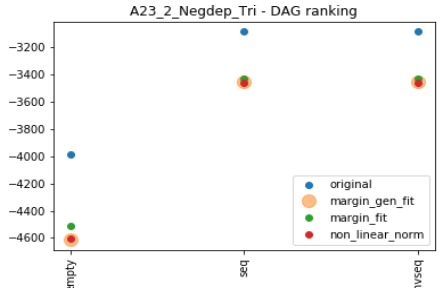
Network structure score ranking is presented in Figure 39. Once more, none of the normalization methods has interfered with the score ranking and all methods kept invariant the ranking order in relation to the original dataset ranking, and again a perceivable match between non-linear normalization and generative distribution sample transformations, this experiment also pointing to the non-linear normalization as the most representative of the original population behavior.



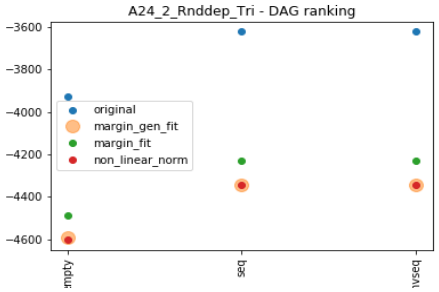
(a) Independent variables.



(b) Positive dependent variables.



(c) Negative dependent variables.



(d) Intermediate dependent variables.

Figure 39 – Comparison of original, marginal distribution fitting and normalized Bayesian networks score ranking for the bivariate trimodal distribution cases. In the figures, each Bayesian network structure is a colored dot in the graphs where the horizontal axis represents different network structures and the vertical stands for scores. Different dot colors represent different normalization techniques: blue is for no normalization (original dataset), green for marginal distribution fitting normalization, red for the proposed non-linear normalization, and light orange for the generative distribution normalization.

4.3 Experiment 03 - Multivariate (Unimodal) Phenomenon

In another augment in complexity, but now in terms of the number of variables, we choose to analyze a phenomenon with features represented by a multivariate distribution with 6 different random variables.

We will now suppose a family of phenomena with six continuous random variables $x1, x2, ..., x6$, whose samples are again generated by normal distributions with means $1.0, 2.0, ..., 6.0$ respectively and individual standard deviations equally $1.0, 2.0, ..., 6.0$, while the covariation values will be established according to each of four dependence cases. Its margins probability densities are showed graphically in Figure 40.

While $x1, x2, ..., x6$ range throughout the entire Real line, many samples are concentrated on $1.0, 2.0, ..., 6.0$ neighborhoods, respectively.

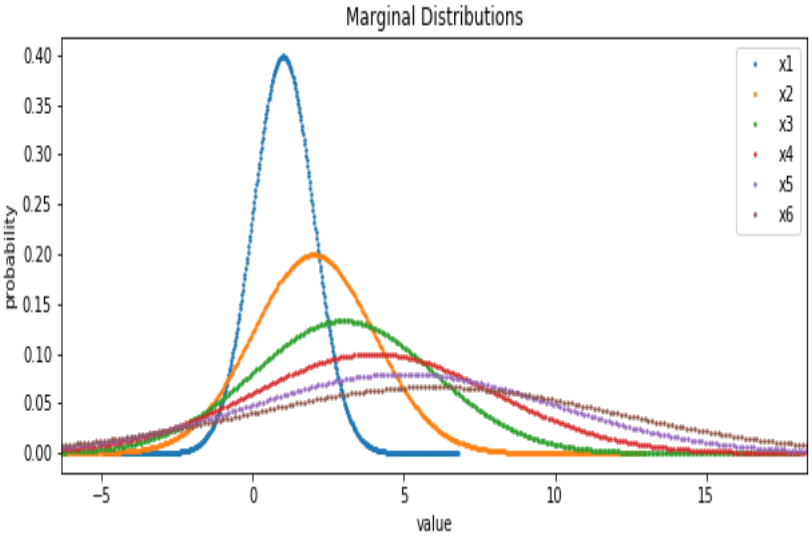
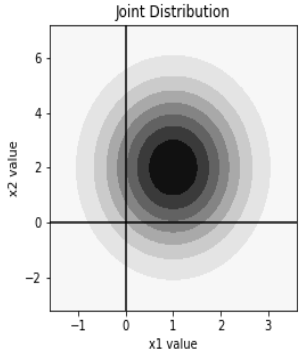
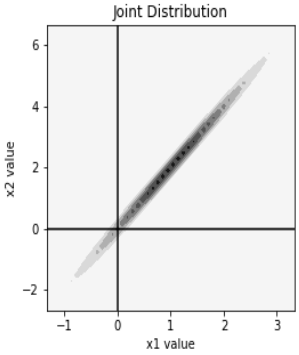


Figure 40 – Multivariate independent normal experimental datasets with 6 random variables - marginal probability densities.

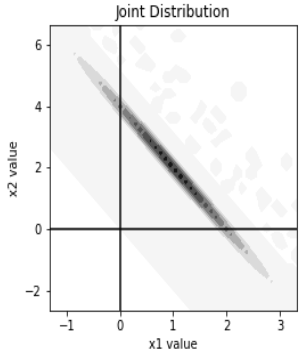
Following our established procedure, we consider four possible members of that family regarding their variables dependence relation: one where the variables are almost totally independent, one where they are almost completely positive dependent, one where half of them are almost completely negative dependent with the other half almost completely positive dependent and one where they are at a intermediate grade of dependence (correlation +0.3, as before). The corresponding joint probability densities are graphically represented by its level curves for $x1$-$x2$ variables pair as example, in Figure 41.



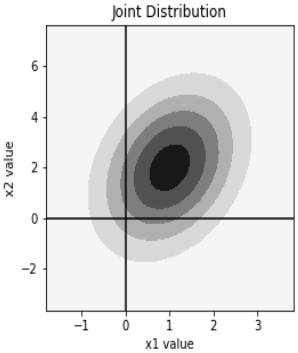
(a) Independent variables.



(b) Positive dependent variables.



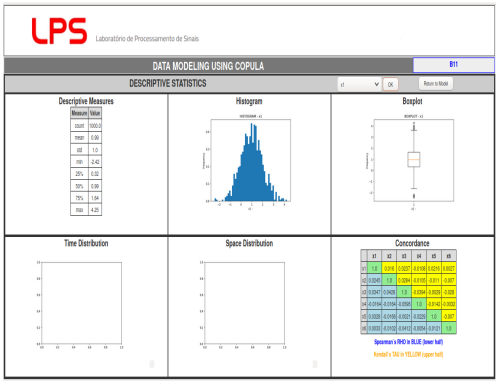
(c) Negative dependent variables.



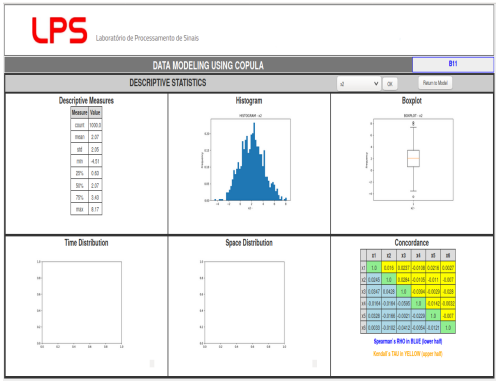
(d) Intermediate degree dependent variables.

Figure 41 – Multivariate normal experimental datasets with 6 random variables - joint probability densities level curves.

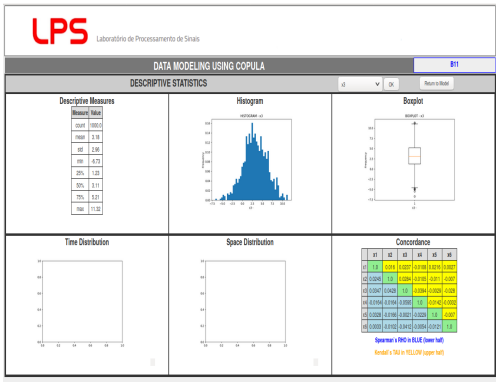
The analysis software tool shows a statistical description of $x1$, $x2$, $x3$, and $x6$ features as examples (Figure 42).



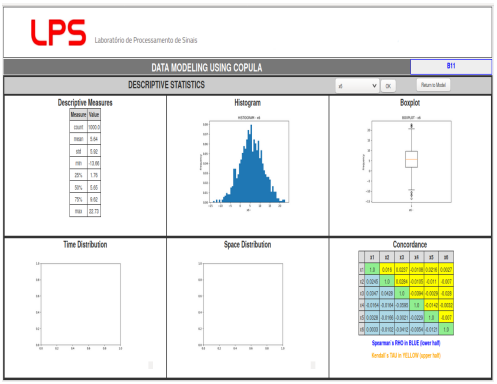
(a) Feature $x1$ descriptive statistics.



(b) Feature $x2$ descriptive statistics.



(a) Feature $x3$ descriptive statistics.



(b) Feature $x6$ descriptive statistics.

Figure 42 – Multivariate normal unimodal experimental datasets - $x1$, $x2$, $x3$ and $x6$ variables descriptive statistics.

Table 9 presents a consolidation of the variables statistics measures, and all other three datasets have similar results. In parallel, Table 10 shows the concordance measures presented in the corresponding tool screen tables for all four datasets.

Table 9 – Descriptive measure numbers for the multivariate normal unimodal independent experimental dataset random variables.

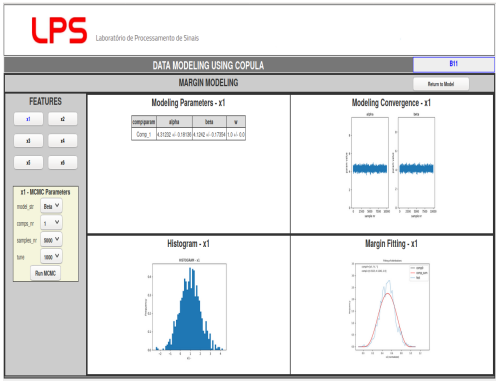
Measure	$x1$ **value**	$x2$ **value**	$x3$ **value**	$x4$ **value**	$x5$ **value**	$x6$ **value**
count	1,000	1,000	1,000	1,000	1,000	1,000
mean	0.99	2.07	3.18	4.01	5.08	5.64
std dev	1.0	2.05	2.96	3.94	4.96	5.92
min	−2.42	−4.51	−6.73	−9.12	−9.98	−13.66
25%	0.32	0.63	1.23	1.31	1.75	1.76
50%	0.99	2.07	3.11	3.89	5.22	5.65
75%	1.64	3.43	5.21	6.71	8.38	9.62
max	4.25	8.17	11.32	16.59	22.57	22.73

Table 10 – Concordance pairwise values for the multivariate normal unimodal experimental dataset random variables. Rho stands for Spearman's rho and Tau for Kendall's tau.

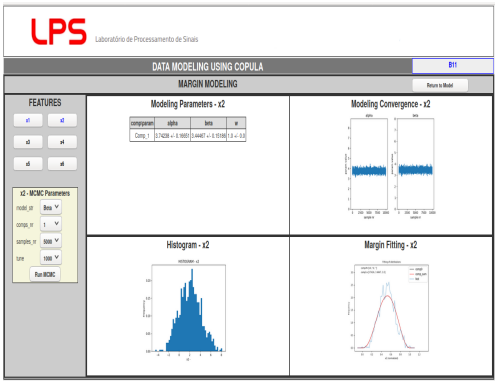
Dataset	**Independent**		**Positive**		**Negative**		**Intermediate**	
Variables Pair	**Rho**	**Tau**	**Rho**	**Tau**	**Rho**	**Tau**	**Rho**	**Tau**
$x1$-$x2$	+0.0245	+0.0016	+0.9987	+0.9711	−0.9987	−0.9702	+0.2037	+0.1373
$x1$-$x3$	+0.0347	+0.0237	+0.9978	+0.9610	+0.9976	+0.9589	+0.0986	+0.0657
$x1$-$x4$	−0.0164	−0.0108	+0.9965	+0.9507	−0.9960	−0.9472	+0.0343	+0.0240
$x1$-$x5$	+0.0328	+0.0216	+0.9955	+0.9433	+0.9948	+0.9403	−0.0176	−0.0120
$x1$-$x6$	+0.0033	+0.0027	+0.9941	+0.9357	−0.9935	−0.9325	+0.0272	+0.0184
$x2$-$x3$	+0.0428	+0.0284	+0.9977	+0.9597	−0.9973	−0.9567	+0.1137	+0.0765
$x2$-$x4$	−0.0164	−0.0105	+0.9961	+0.9480	+0.9960	+0.9472	+0.0482	+0.0325
$x2$-$x5$	−0.0166	−0.0110	+0.9954	+0.9422	−0.9948	−0.9399	−0.0794	−0.0531
$x2$-$x6$	−0.0102	−0.0070	+0.9940	+0.9353	+0.9932	+0.9315	+0.0678	+0.0452
$x3$-$x4$	−0.0595	−0.0394	+0.9961	+0.9480	−0.9961	−0.9478	+0.0669	+0.0447
$x3$-$x5$	−0.0021	−0.0029	+0.9954	+0.9429	+0.9944	+0.9384	−0.0205	−0.0131
$x3$-$x6$	−0.0412	−0.0280	+0.9940	+0.9349	−0.9931	−0.9307	+0.0430	+0.0282
$x4$-$x5$	−0.0229	−0.0142	+0.9953	+0.9418	−0.9950	−0.9408	+0.0269	+0.0178
$x4$-$x6$	−0.0054	−0.0032	+0.9942	+0.9353	+0.9933	+0.9312	+0.0238	+0.0154
$x5$-$x6$	−0.0121	−0.0070	+0.9940	+0.9348	−0.9933	−0.9313	+0.0252	+0.0171

4.3.1 MCMC Marginal Distribution Fitting

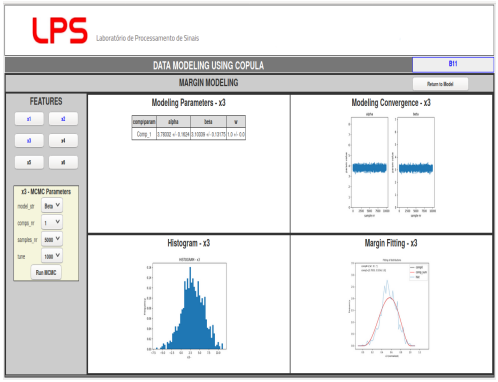
As usual, after running MCMC modeling using LpsCopModel with a non-informative uniform prior and a (two-parameter) Beta distribution for each one of the six random variables, the results are presented in Figures 43 and 44.



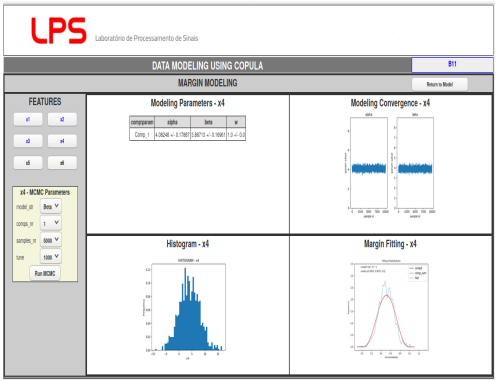
(a) Margin fitting for variable $x1$.



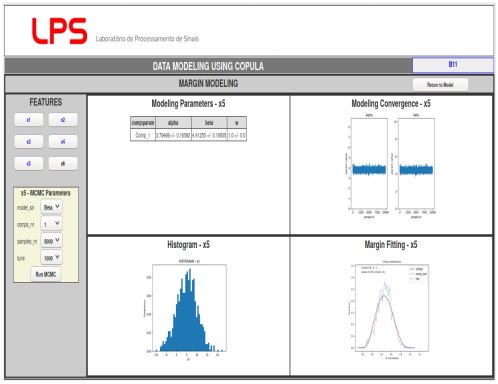
(b) Margin fitting for variable $x2$.



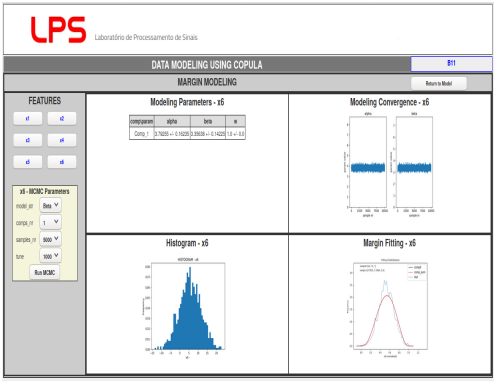
(a) Margin fitting for variable $x3$.



(b) Margin fitting for variable $x4$.

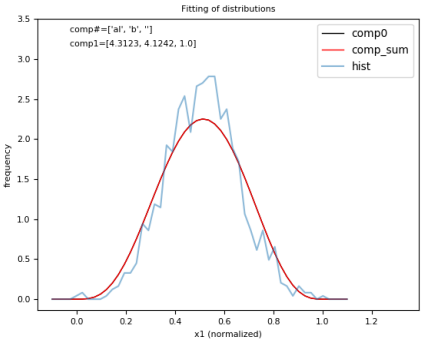


(a) Margin fitting for variable $x5$.

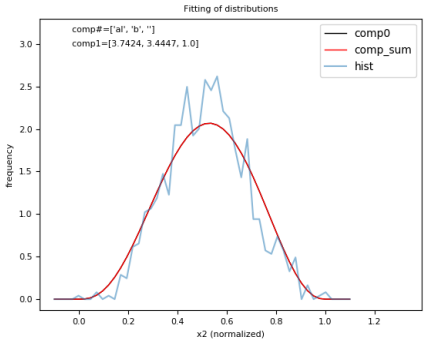


(b) Margin fitting for variable $x6$.

Figure 43 – Multivariate normal experimental dataset - marginal distribution fitting for all variables with a parametric Beta and non-informative uniform prior by Bayesian MCMC modeling.



(a) Margin fitting for variable $x1$.



(b) Margin fitting for variable $x2$.

Figure 44 – Multivariate normal experimental dataset - marginal distribution fitting for two of the variables with a parametric Beta and non-informative uniform prior by Bayesian MCMC modeling.

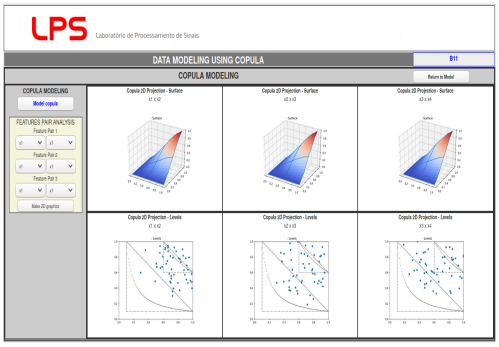
Table 11 presents all the marginal distribution fitting parameters for each multivariate unimodal dataset.

Table 11 – Margin fitting normalized parameters for the multivariate unimodal datasets.

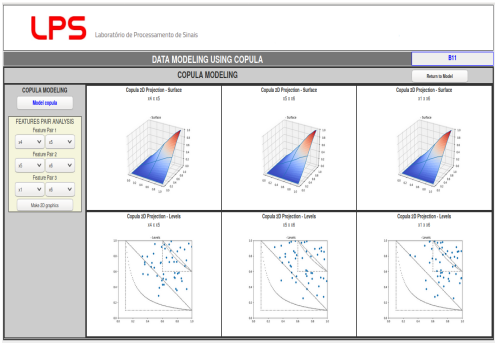
Dataset	Variable	Distribution	Parameter a	Parameter b
Independent	$x1$	Beta	4.3123	4.1242
	$x2$	Beta	3.7424	3.4447
	$x3$	Beta	3.7833	3.1034
	$x4$	Beta	4.0825	3.8871
	$x5$	Beta	3.7950	4.4126
	$x6$	Beta	3.7925	3.3564
Positive dependent	$x1$	Beta	2.9047	3.1630
	$x2$	Beta	3.0807	3.4209
	$x3$	Beta	3.2066	3.3823
	$x4$	Beta	3.0998	3.3502
	$x5$	Beta	3.1180	3.4361
	$x6$	Beta	2.9086	3.3356
Negative dependent	$x1$	Beta	3.2065	3.8754
	$x2$	Beta	4.0219	3.2800
	$x3$	Beta	3.4056	4.0794
	$x4$	Beta	4.1202	3.3780
	$x5$	Beta	3.0309	3.8121
	$x6$	Beta	3.9655	3.1804
Random dependent	$x1$	Beta	4.0195	4.8570
	$x2$	Beta	4.5475	3.6203
	$x3$	Beta	3.4016	3.9101
	$x4$	Beta	4.3740	4.2274
	$x5$	Beta	3.5331	3.4137
	$x6$	Beta	4.4851	4.1576

4.3.2 Empirical Copula Modeling

Figure 45 shows empirical copula modeling with LpsCopModel for the independent dataset on the software screen, while Figure 46 presents the graphics for all four datasets (independent, positive dependent, negative dependent and intermediate dependent) $x1-x2$ copula projection. For six variables, the dimensionallity curse begins to show its claws and the datasets without very strong dependencies sparsity enlarge level boundaries enough to samples not characterizing independence patterns.

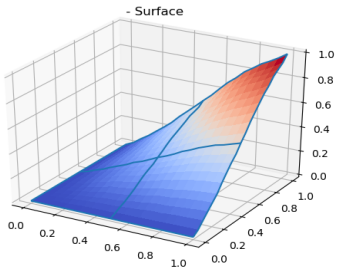


(a) $x1-x2$, $x2-x3$ and $x3-x4$ copula 2D projection surfaces.

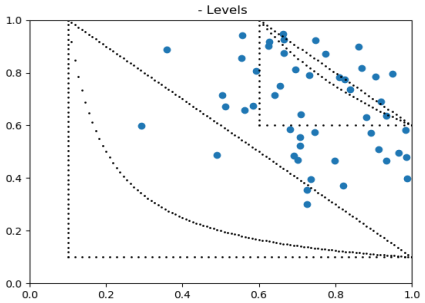


(b) $x4-x5$, $x5-x6$ and $x1-x6$ copula 2D projection surfaces.

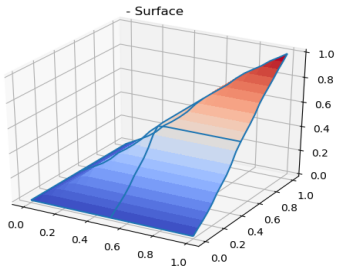
Figure 45 – Multivariate independent experimental dataset - empirical copula figures in LpsCopModel software tool.



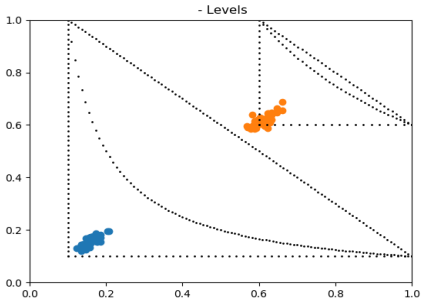
(a1) Independence - surface.



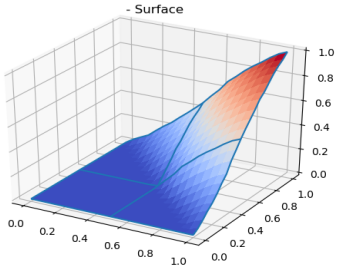
(a2) Independence - level curves.



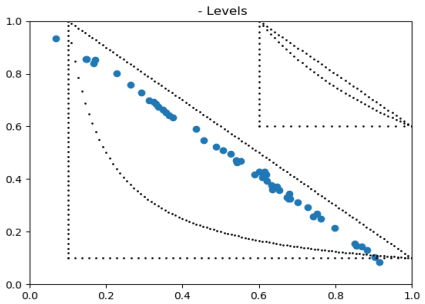
(b1) Positive dependence - surface.



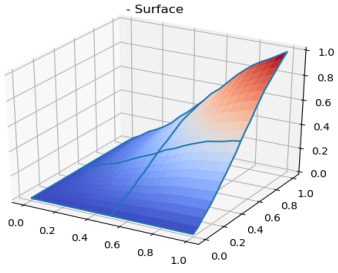
(b2) Positive dependence - level curves.



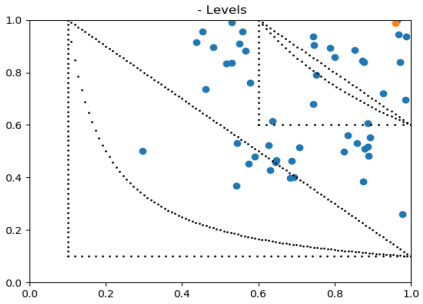
(c1) Negative dependence - surface.



(c2) Negative dependence - level curves.



(d1) Intermediate dependence - surface.

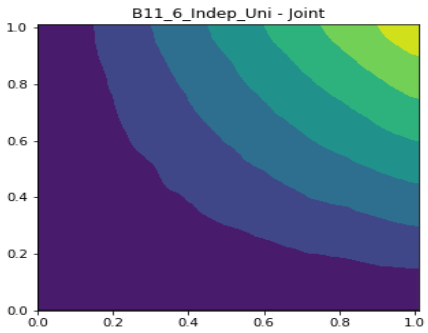


(d2) Intermediate dependence - level curves.

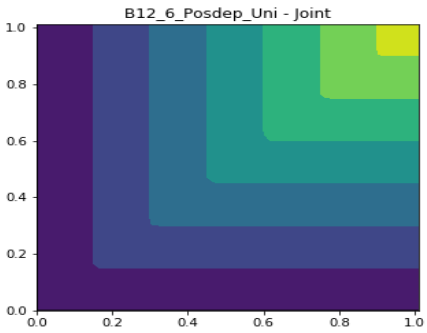
Figure 46 – Multivariate unimodal experimental dataset - empirical copula 2D $x1$-$x2$ projection surfaces and level curves.

4.3.3 Non-Linear Normalization by Sample Reducing

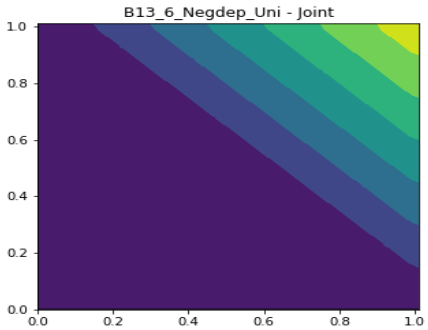
The proposed non-linear normalization turns original marginal distributions into uniform distributions and the joint distribution (copula) shows the typical dependence pattern concerning each one of the four cases (Figure 47). In the case of negative dependence, as it is a multivariate situation with more than two variables, the dataset was originally made by alternating the negative and positive dependence, negative and positive dependence patterns also alternate among variables.



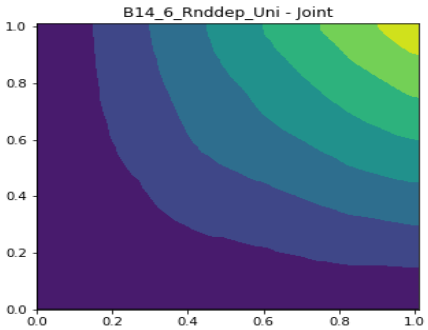
(a) Independent variables.



(b) Positive dependent variables.



(c) Negative dependent variables.



(d) Intermediate degree dependent variables.

Figure 47 – Multivariate unimodal normal experimental dataset - $x1$-$x2$ joint distribution projections.

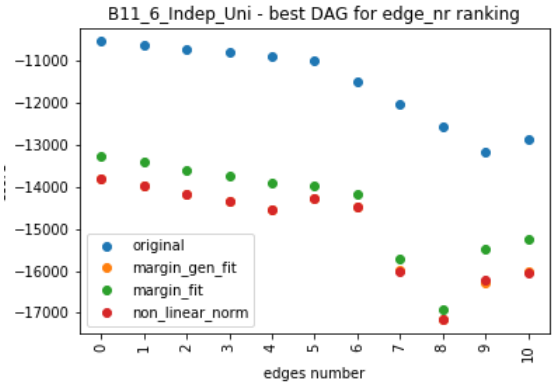
4.3.4 Bayesian Network Copula Modeling

For a multivariate dataset with six variables, the Bayesian network structure is a choice among an outstanding number of possibilities in result of its super-exponential nature. Therefore, the structure searching problem is not a trivial one as before and there are still many researches on this matter, hence reaching the best structure to a dataset is still an open question. At this point, a strategy has to be adopted in order to collect a suitable set of structures for comparing the different BN copula modeling methodologies.

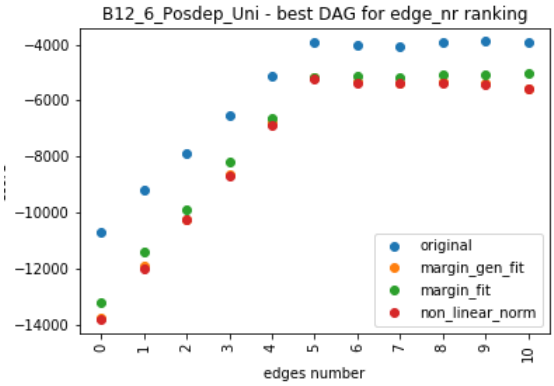
The main goal to obtain a good model for the dataset copula is equivalent to reaching the best Bayesian network structure possible. Therefore, if we could verify that the proposed methodology of normalizing each random variable by its own marginal distribution does not interfere with the structure searching, that goal would have been clearly achieved. That would occur if the scoring relative order of Bayesian networks regarding a dataset is essentially preserved by applying the normalization methodology. Hence, we must test if our methodology preserves BNs scoring relative order in a grade that does not endanger BN structure search, i.e., if the relevant part of the relative scoring order among BNs is the same before and after normalization.

For selecting a set of network structures to be considered we adopted the strategy of taking many structures variations using some referential patterns (empty BN, sequencial BN, naive BN, binary tree BN, etc.) variations and increment that set with a number of randomly chosen BN structures with many different number of edges, thus getting hundreds of possible structures to score.

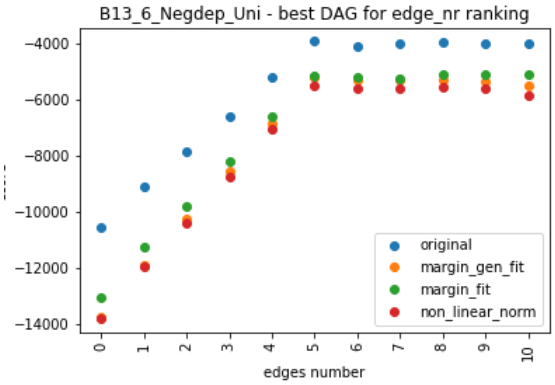
Grouping the network structures by number of edges as a parameter of model complexity, we analyzed the variation of the best score in each group by the number of edges for all four normalization methods, which is presented in Figure 48. All four methods presented very similar tendencies and the best score number of edges matched. Figures 49, 50 , and 51 shows the best Bayesian network structures and its scores for original, MCMC marginal distribution fitting and normalized datasets, and all the methods selected the same structure or a group of four coherent structures in the best four scores.



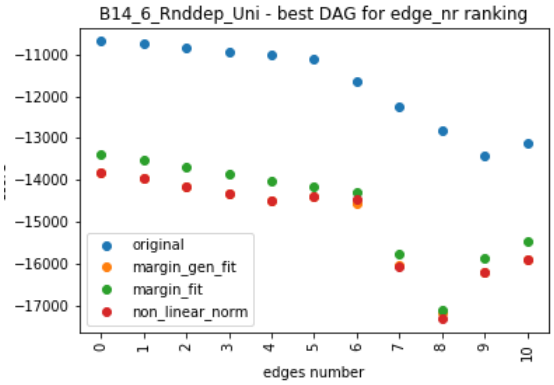
(a) Independent variables.



(b) Positive dependent variables.

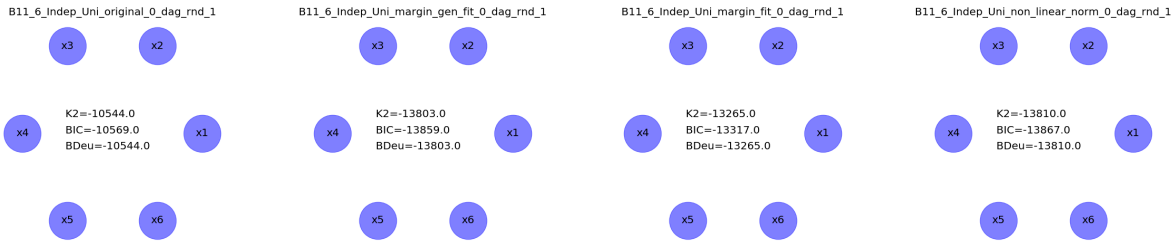


(c) Negative dependent variables.

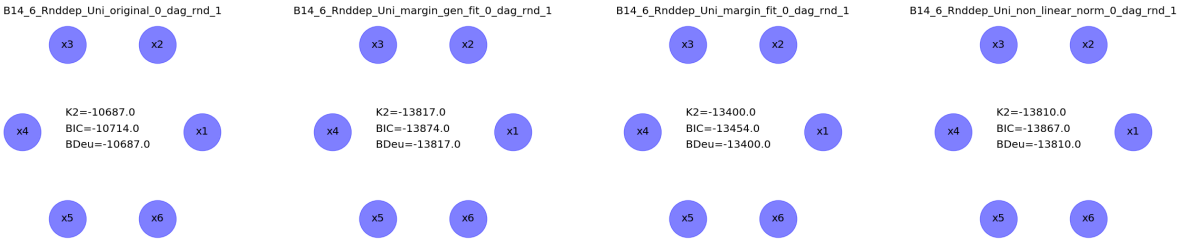


(d) Intermediate dependent variables.

Figure 48 – Comparison of score variation among best Bayesian network score within each number of edges for all four normalization methods for the multivariate normal distribution cases.

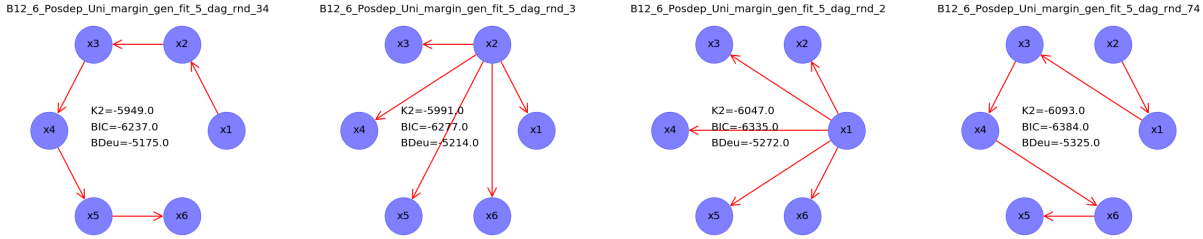


(a) Independent variables.

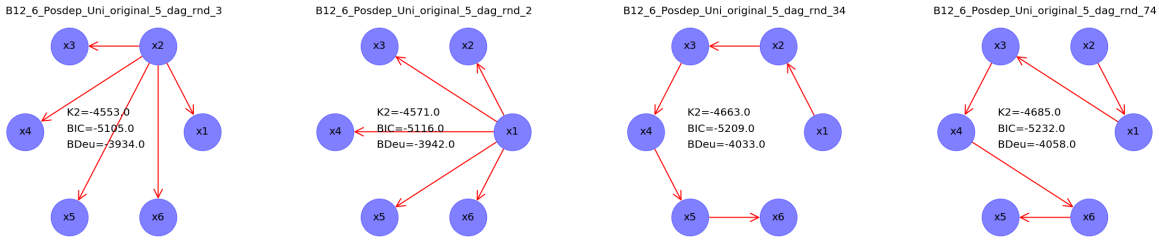


(b) Intermediate dependent variables.

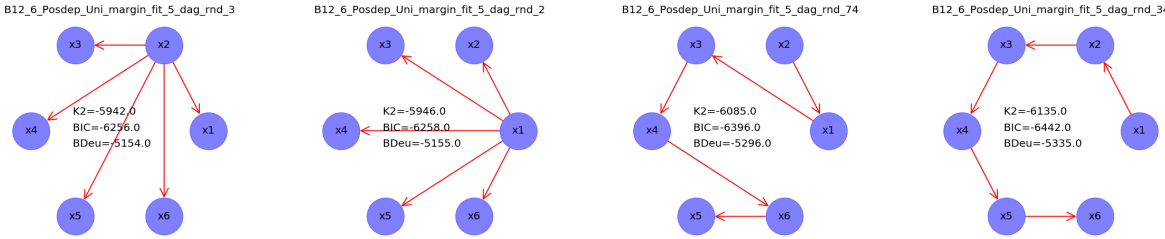
Figure 49 – Comparison of original and normalized Bayesian networks for the multivariate normal distribution cases for independent and intermediate dependent datasets. For those datasets, the best structure was exactly the same for all normalization methods. The figures show also the score measures for comparison. Isolated (non connected) dots stands for a totally unconnected network structure, the one where all random variables are modeled as independent.



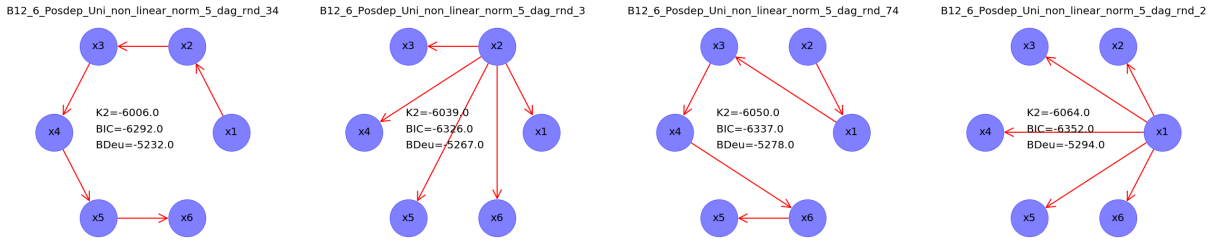
(a) Generator marginal fitting normalization.



(b) No normalization - original data.

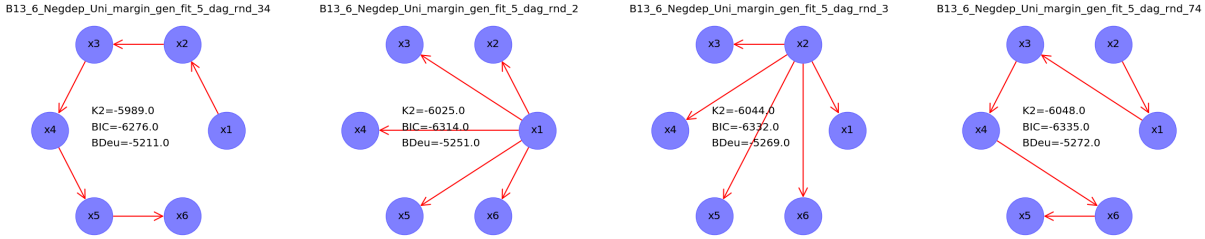


(c) Marginal fitting normalization.

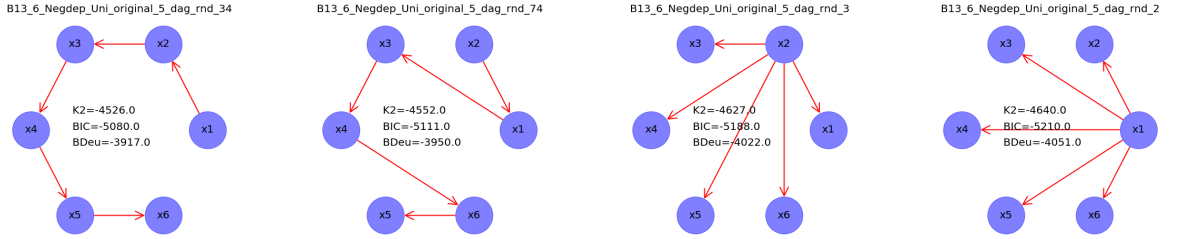


(d) Sample reducing normalization.

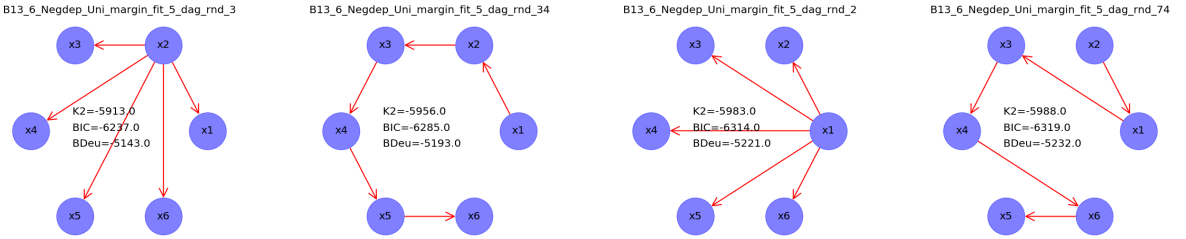
Figure 50 – Comparison of original and normalized Bayesian networks for the multivariate normal distribution cases for the positive dependent dataset. Structures in descending score order from left to right. For this case, there were no exact coincidence, but the four best ranking structures were the same and all reflect a strong positive dependence in distinct structures. The figures show also the score measures for comparison. Isolated (non connected) dots stands for a totally unconnected network structure, the one where all random variables are modeled as independent.



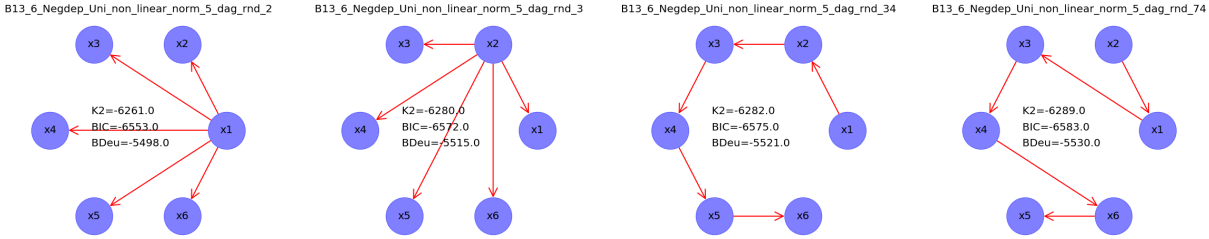
(a) Generator marginal fitting normalization.



(b) No normalization - original data.



(c) Marginal fitting normalization.

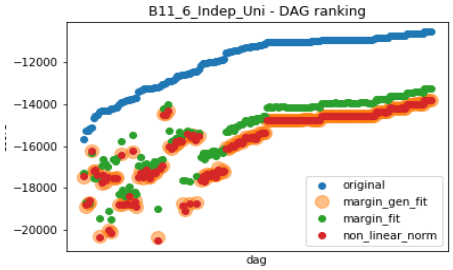


(d) Sample reducing normalization.

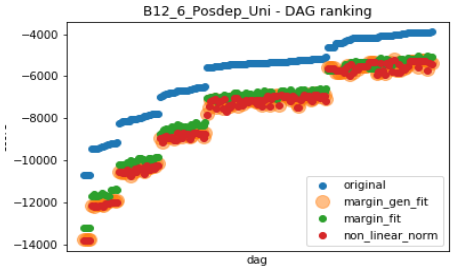
Figure 51 – Comparison of original and normalized Bayesian networks for the multivariate normal distribution cases for the negative dependent dataset. Structures in descending score order from left to right. For this case, there were no exact coincidence, but the four best ranking structures were the same and all reflect a strong positive dependence in distinct structures. The figures show also the score measures for comparison. Isolated (non connected) dots stands for a totally unconnected network structure, the one where all random variables are modeled as independent.

Network structure score ranking for all possible structures in the solution search

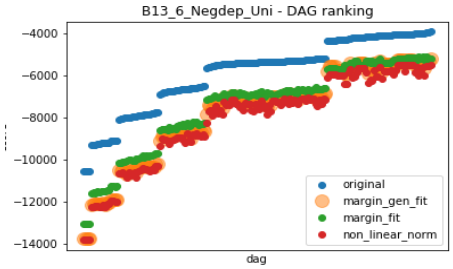
set is presented in Figure 52. It can be clearly noticed that none of the normalization methods has caused relevant quantitative interference in the score ranking and all methods kept the score ranking order tendency in relation to the original dataset ranking. Besides, just as before, the best matching method to the generative distribution normalization (the one made from the distribution used to originally generate the samples) proved to be the non-linear normalization by sample reducing.



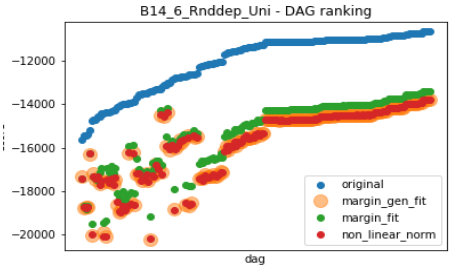
(a) Independent variables.



(b) Positive dependent variables.



(c) Negative dependent variables.



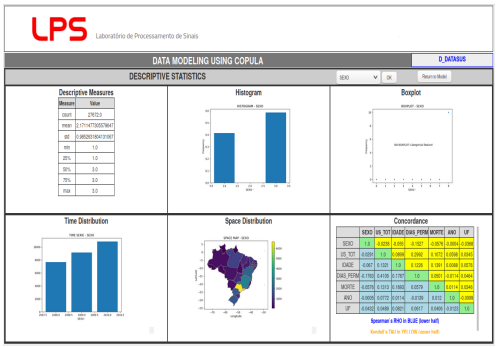
(d) Random dependent variables.

Figure 52 – Comparison of original, marginal distribution fitting and normalized Bayesian networks score ranking for the bivariate normal distribution cases. In the figures, each Bayesian network structure is a colored dot in the graphs where the horizontal axis represents different network structures and the vertical stands for scores. Different dot colors represent different normalization techniques: blue is for no normalization (original dataset), green for marginal distribution fitting normalization, red for the proposed non-linear normalization, and light orange for the generative distribution normalization.

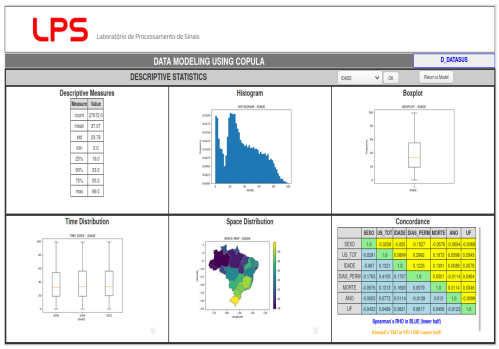
4.4 Real Case 1 - DATASUS General Profile 2008-2010

The first real case we are going to analyze is a healthcare dataset. We started by collecting data from a public available Brazilian nationwide hospital admissions dataset and selecting some features to be observed. We collected individual data from years 2008 to 2010 with 7 chosen features, starting with all admissions along those years and then randomly sampling a 1% subset to avoid size complexity and high processing costs. The features considered were "SEXO" (gender), "IDADE" (age), "DIAS_PERM" (days in hospital), "US_TOT" (costs in US dollars), "MORTE" (death), "ANO" (year), and "UF" (federation state).

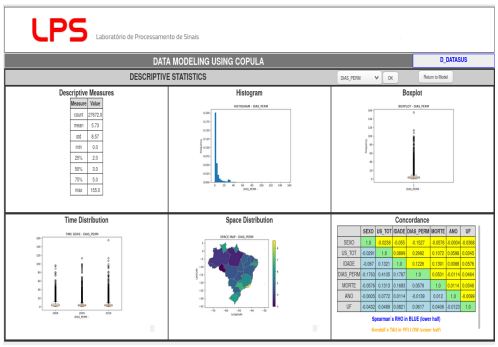
In this case, dataset is real, so the generation data stages do not apply and we go straight to the analysis software tool for giving a general overview of the collected data profile, which is presented in Figure 53.



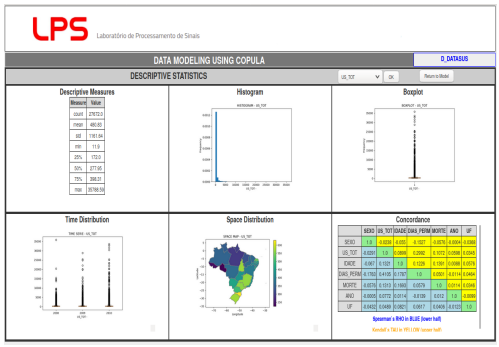
(a) Feature "SEXO".



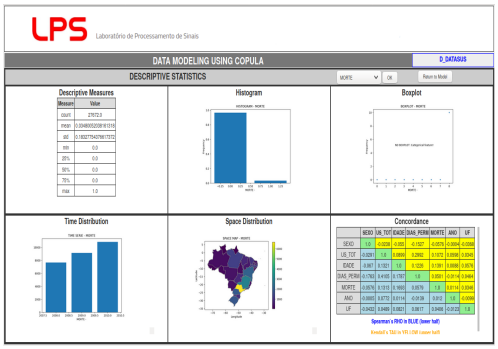
(b) Feature "IDADE".



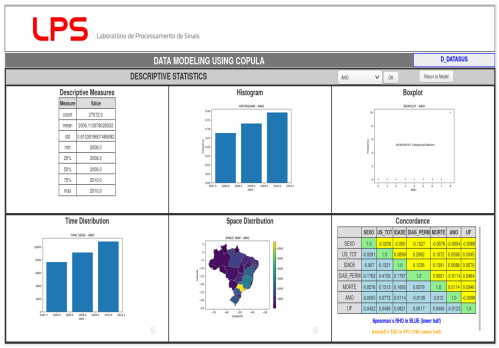
(a) Feature "DIAS_PERM".



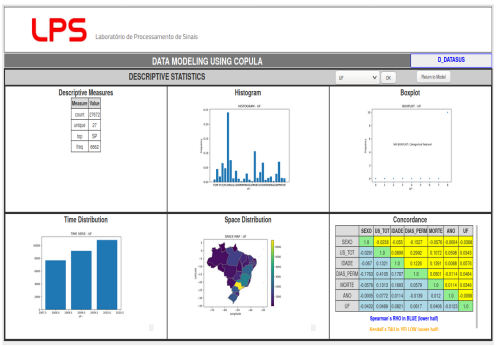
(b) Feature "US_TOT".



(a) Feature "MORTE".



(b) Feature "ANO".

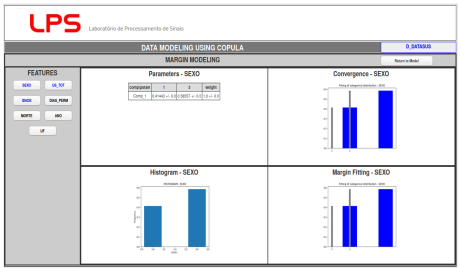


(a) Feature "UF".

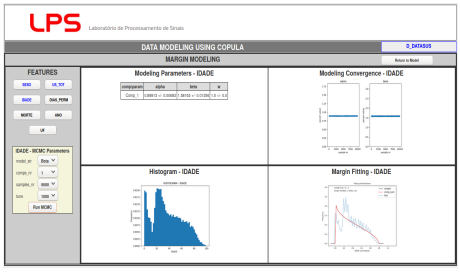
Figure 53 – DATASUS healthcare sample dataset - descriptive statistics LPSCopModel screens.

4.4.1 MCMC Marginal Distribution Fitting

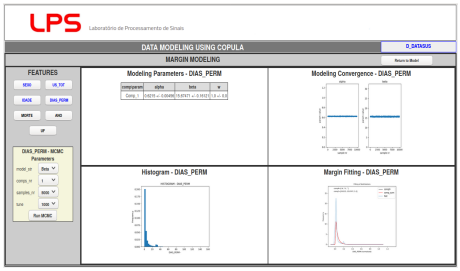
Modeling each feature marginal distribution using a multinomial frequency-oriented for categorical variables and, for numeric variables, an MCMC distribution fitting with the same non-informative uniform prior and Beta distribution premises, the results are presented in Figure 54.



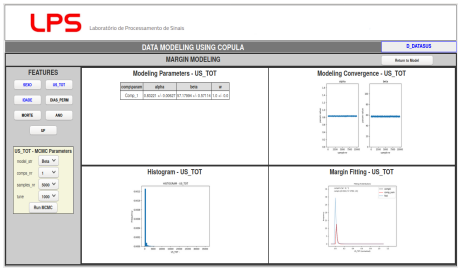
(a) Marginal fitting for variable "SEXO".



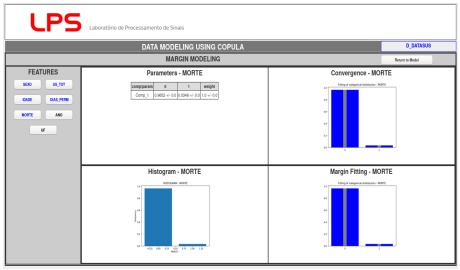
(b) Marginal fitting for variable "IDADE".



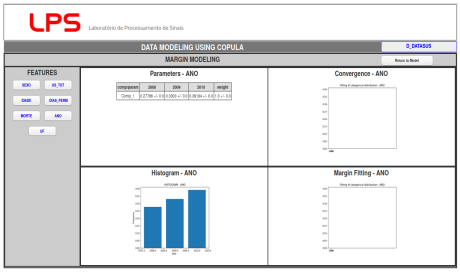
(c) Marginal fitting for variable "DIAS_PERM"



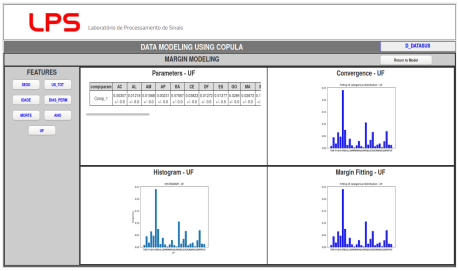
(d) Marginal fitting for variable "US_TOT".



(e) Marginal fitting for variable "MORTE".



(f) Marginal fitting for variable "ANO".



(g) Marginal fitting for variable "UF".

Figure 54 – DATASUS healthcare sample dataset - marginal distribution fitting with multinomial frequency-oriented or parametric Beta and non-informative uniform prior by Bayesian MCMC modeling, for categorical and numeric variables, respectively.

Tables 12, 13, and 14 presents all the marginal distribution fitting parameters for the DATASUS healthcare sample dataset.

Table 12 – Margin fitting normalized parameters for the DATASUS healthcare sample dataset - numeric features.

Variable	Distribution	Param. a	Param. b
IDADE	Beta	0.8891	1.5816
US_TOT	Beta	0.8322	57.1799
DIAS_PERM	Beta	0.6215	15.6747

Table 13 – Margin fitting normalized parameters for the DATASUS healthcare sample dataset - categorical features "SEXO", "MORTE", and "ANO".

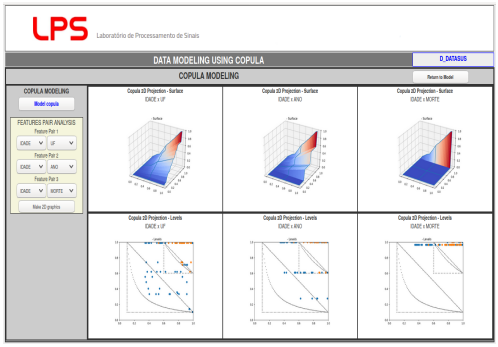
Variable	Feature	Category	Probability
SEXO	Male	1	0.4144
	Female	2	0.5856
MORTE	Alive	1	0.9652
	Dead	2	0.0348
ANO	2008	1	0.2779
	2009	2	0.3303
	2010	3	0.3918

Table 14 – Margin fitting normalized parameters for the DATASUS healthcare sample dataset - categorical feature "UF".

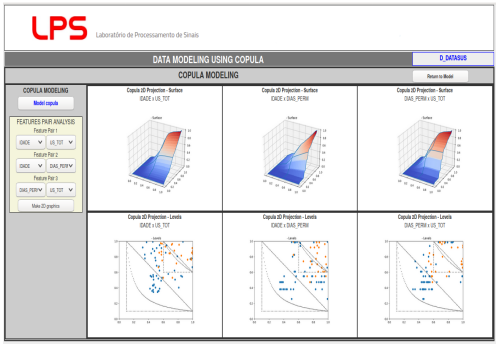
Category	**AC**	**AL**	**AM**	**AP**	**BA**	**CE**	**DF**	**ES**	**GO**
Variable	1	2	3	4	5	6	7	8	9
Probability	0.0031	0.0122	0.0109	0.0023	0.0757	0.0382	0.0127	0.0138	0.0284
Category	**MA**	**MG**	**MS**	**MT**	**PA**	**PB**	**PE**	**PI**	**PR**
Variable	10	11	12	13	14	15	16	17	18
Probability	0.0287	0.1062	0.0114	0.0136	0.0472	0.0187	0.0447	0.0186	0.0697
Category	**RJ**	**RN**	**RO**	**RR**	**RS**	**SC**	**SE**	**SP**	**TO**
Variable	19	20	21	22	23	24	25	26	27
Probability	0.0649	0.0136	0.0067	0.0019	0.0663	0.0336	0.0078	0.2408	0.0086

4.4.2 Empirical Copula Modeling

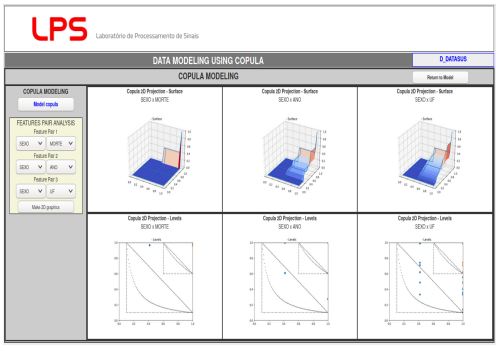
Going on, the empirical copula model is produced as shown in Figures 55 and 56.



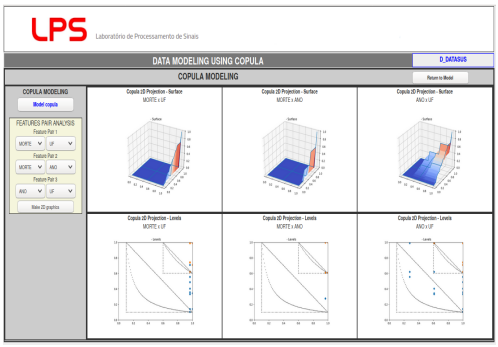
(a) "IDADE" and categorical variables.



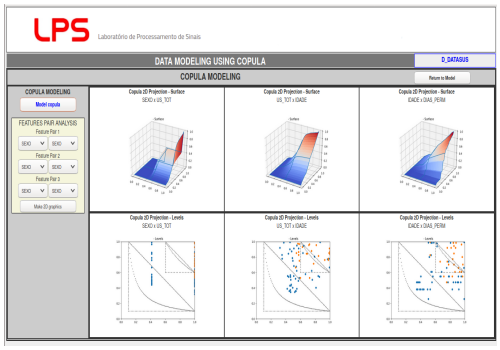
(b) Continuous variables.



(b) Categorical variables.

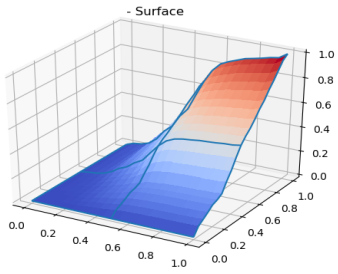


(d) "MORTE" and time and space variables.

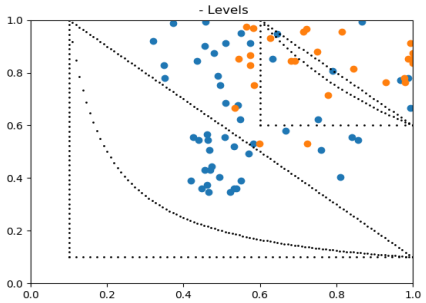


(e) "SEXO" and continuous variables.

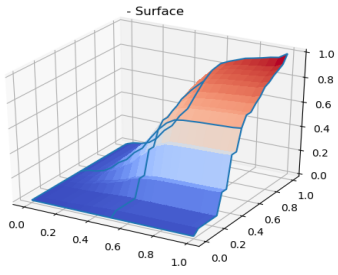
Figure 55 – DATASUS healthcare sample dataset - empirical copula figures in LpsCopModel software tool. The most relevant projections were chosen to be presented.



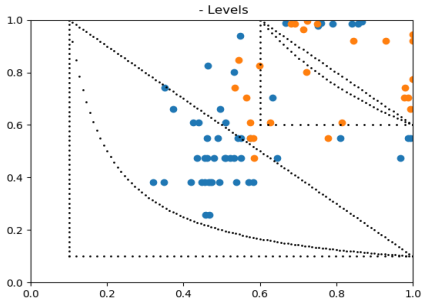
(a1) "IDADE"-"US_TOT" - surface.



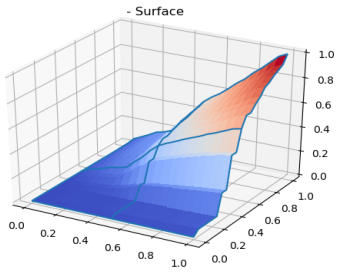
(a2) "IDADE"-"US_TOT" - level curves.



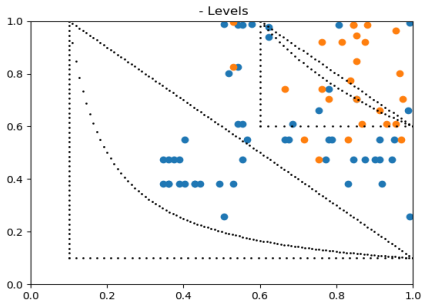
(b1) "IDADE"-"DIAS_PERM" - surface.



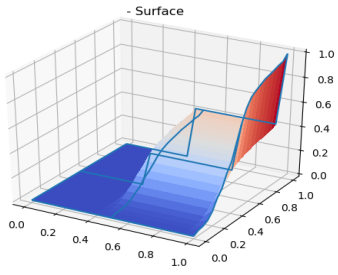
(b2) "IDADE"-"DIAS_PERM" - level curves.



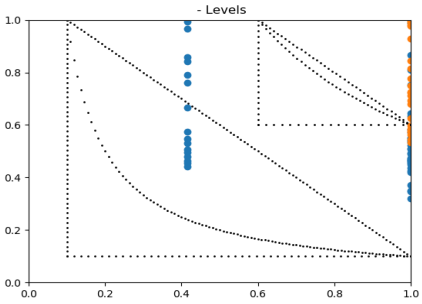
(c1) "US_TOT"-"DIAS_PERM" - surface.



(c2) "US_TOT"-"DIAS_PERM" - level curves.



(d1) "SEXO"-"IDADE" - surface.

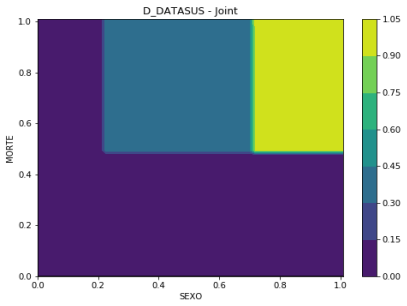


(d2) "SEXO"-"IDADE" - level curves.

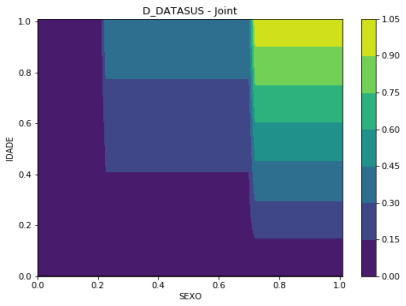
Figure 56 – DATASUS healthcare sample dataset - empirical copula 2D projection surfaces and level curves for relevant variables pair examples.

4.4.3 Non-Linear Normalization by Sample Reducing

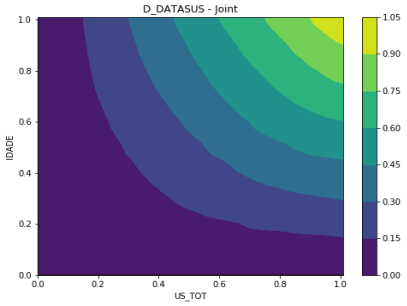
As expected, the marginals became uniform distributions when normalized by sample reducing. As a real dataset, the joint distribution of the non-linear normalized variables shows a different pattern according to the projection variables real behavior which is totally autonomous from the individual variables profile (Figure 57). When categorical variables are involved, the level curves adopt rectangular patterns.



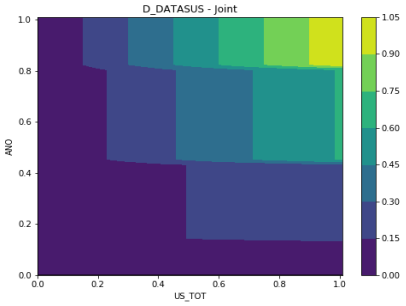
(a) "SEXO"-"MORTE".



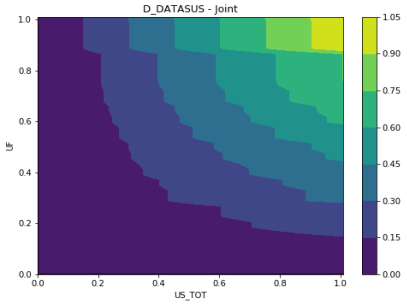
(b) "SEXO"-"IDADE".



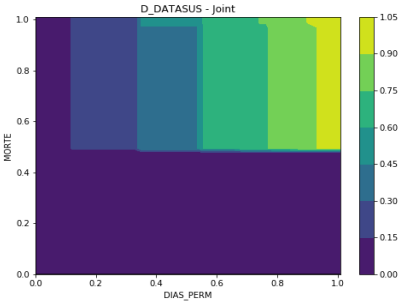
(c) "US_TOT"-"IDADE".



(d) ""US_TOT"-"ANO".



(e) ""US_TOT"-"UF".



(f) "DIAS_PERM"-"MORTE".

Figure 57 – DATASUS healthcare sample dataset - normalized joint distribution projection.

4.4.4 Bayesian Network Copula Modeling

As in the last experiment, here again the number of possible Bayesian network structures is prohibitive and we produced a set with hundreds of possible structures

from referential structures and random generation to be scored. Best network structure score in each edge number group is presented in Figure 58. All normalization methods presented very similar tendencies and the best score number of edges matched in 6. Figure 59 shows the evidenced Bayesian network structures by considering all methods together and its scores for original, MCMC marginal distribution fitting and normalized datasets, considering all the methods top ranked the same set of structures. that one is a structure coherent to the phenomenon associated to the dataset, since it states that age and gender determine both days in hospital and mortality rate, and those latter determine costs.

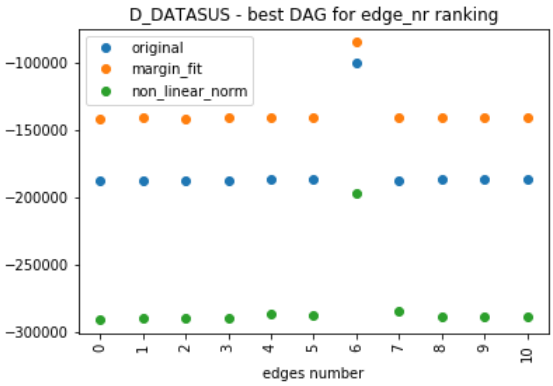


Figure 58 – Comparison of score variation among best Bayesian network score within each number of edges for all four normalization methods for the DATASUS healthcare sample dataset case.

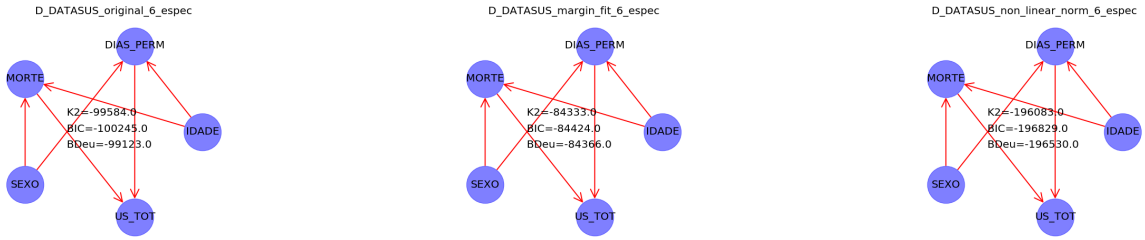


Figure 59 – Comparison of highlighted Bayesian network when all the three normalizations applied to the DATASUS real case are considered together. Each column corresponds to a different normalization, from left to right: none, fitted marginals, and non-linear normalization. The figures show also the score measures for comparison.

Once more, Figure 60 graphically shows where each possible network structure stays in terms of score ranking. It can be clearly noticed that none of the normalization methods has interfered with the score ranking, all methods preserved the ranking order tendency in relation to the no normalization method ranking.

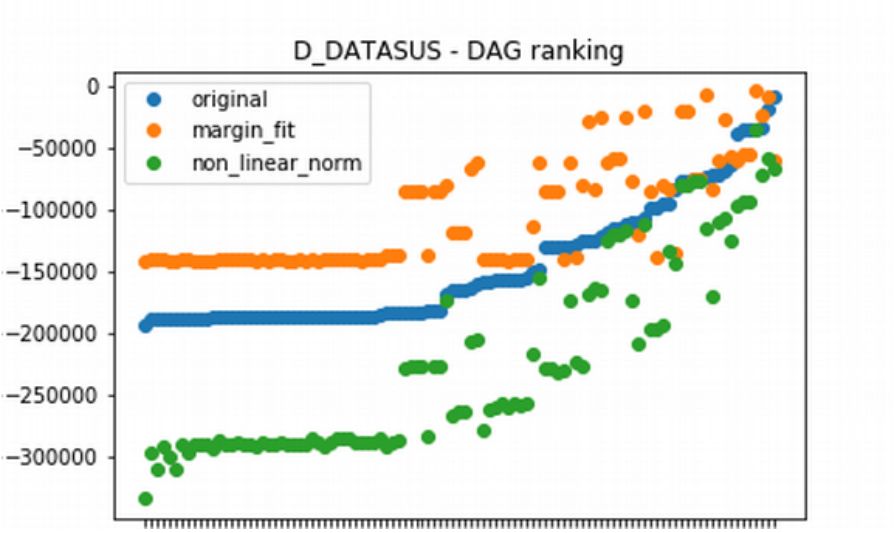


Figure 60 – Comparison of original, marginal distribution fitting and normalized Bayesian networks score ranking for the DATASUS healthcare sample dataset case. In the figures, each Bayesian network structure is a colored dot in the graphs where the horizontal axis represents different network structures and the vertical stands for scores. Different dot colors represent different normalization techniques: blue is for no normalization (original dataset), orange for marginal distribution fitting normalization and green for the proposed non-linear normalization.

4.5 Real Case 2 - Tax Counties Revenue 2011-2015

This subject real dataset was one made up by aggregating some Brazilian Government data organized by county available for the public in general at the web sites of IPEA - Instituto de Pesquisa Economica Aplicada IPEA (2017) and RFB - Receita Federal do Brasil RFB (2017).

IPEA is a Brazilian public institution for economic research which collects data from many sources for its studies and let them available for the general public (Figure 61).

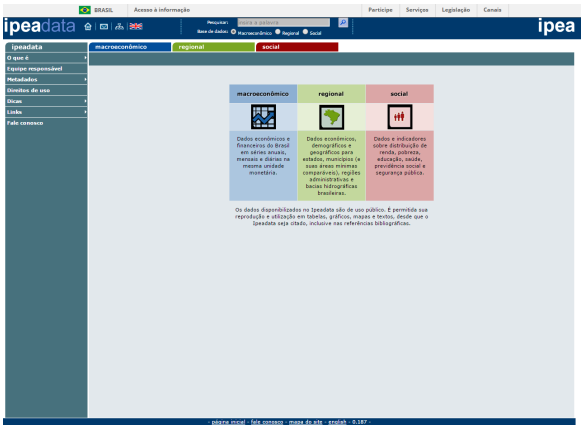
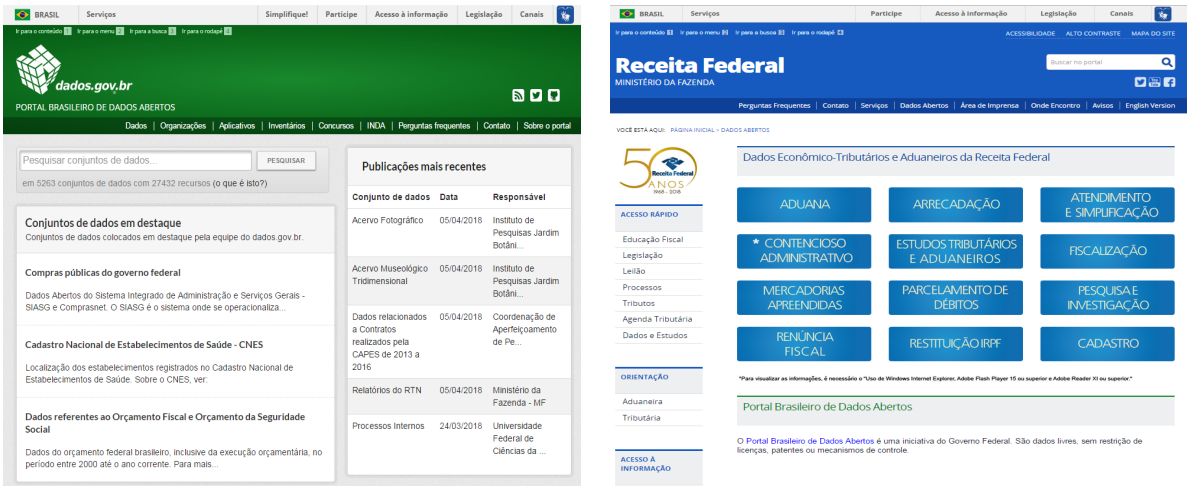


Figure 61 – IPEADATA main webpage. Accessible by http:||ipeadata.gov.br.

The "Dados Abertos" (open data) portal (Figure 62) is a tool provided by the Brazilian government for everybody to access public data and information RFB (2017). All data in that repository has no restriction at all, being available for any use by anyone.



(a) Brazilian Federal Government (b) Brazilian tax administration dept. (RFB)

Figure 62 – Entry webpages for Brazilian public data portal.

Specifically on Tax Administration data, one of the webpages connected to the "Dados Abertos" main webpage is the one from RFB ("Receita Federal do Brasil"), the Brazilian department responsible for tax administration at federal level. At its entrance page, one can access many categories of aggregated tax related data that are not subject to security restrictions or protection, as shown on the left on Figure 62. The data which specifically interested to this research were the ones of municipal revenue, accessible by the webpage in Figure 63.

Figure 63 – Categories of data available at the Brazilian Tax Administration department. Accessible by www.rfb.gov.br

About the collected data itself, a numeric code is associated to each county and every other feature is related to that code in every data source. There are some features which are essentially timeless, such as county area, latitude and longitude, while others are time dependent, like tax income and revenue. For this study, time dimension was excluded by considering all time-dependent features taken in the more recent year in which the data was available, varying from 2011 to 2015 according to the variable and its source.

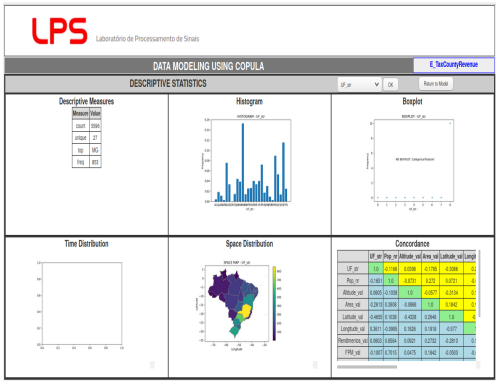
The acquired dataframe consists of $5,596$ counties and eleven features, most of them numeric, as sampled in the following list (Table 4.5), and it refers to various categories and items.

Table 15 – Brazilian counties random sample of 10 instances for illustration.

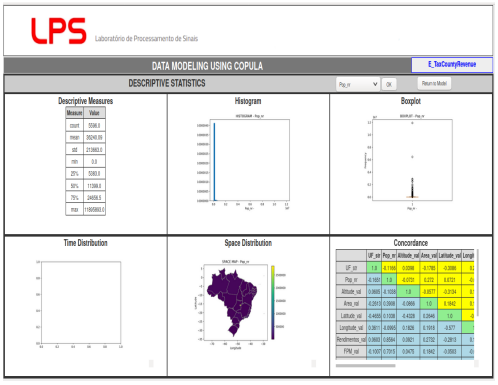
Pop (population)	**Rendimentos** (people revenue)	**FPM** (federal transfers)	**QtdeDecl** (returns number)	**RendTrib** (taxed revenue)	**ReceitaTrib** (taxed income)
0	87260	13714.5	0	4283	161.88
$9.8222e+06$	20812	4104.97	$9.8222e+06$	1728	64.6
0	10588	577.12	0	294	11.44
$5.29319e+06$	3354	0	$5.29319e+06$	266	8.47
$7.35824e+07$	219749	23036.5	$7.35824e+07$	13240	488.51
$5.57875e+06$	5956	1756.48	$5.57875e+06$	938	36.6
$5.57875e+06$	10152	2339.39	$5.57875e+06$	1159	36.48
$5.74776e+06$	2984	271.25	$5.74776e+06$	102	3.39
$4.36817e+06$	6424	1297.88	$4.36817e+06$	466	15.72
$4.45512e+06$	5579	683.27	$4.45512e+06$	324	11.07

The eleven dataset features observed are: "UF" (federation unit, 27 categories), "Pop" (population), "Altitude" (altitude), "Area" (area), "Latitude" (latitude), "Longitude" (longitude), "Rendimentos" (profit or personal revenue), "FPM" (tax income redistribution county share), "ReceitaTrib" (taxable income), "QtdeDecl" (tax return applications number), and "RendTrib" (taxable personal revenue).

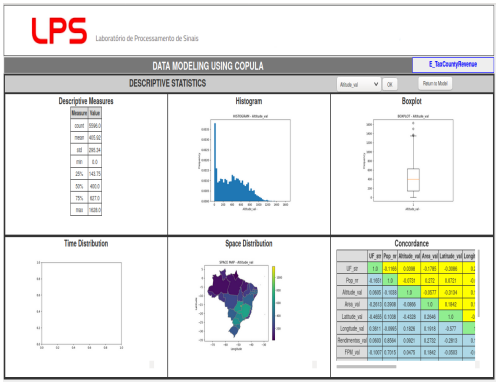
Again, a real dataset dismisses early stages and allows us to go straight to data analysis. The results are presented in Figures 64 and 65.



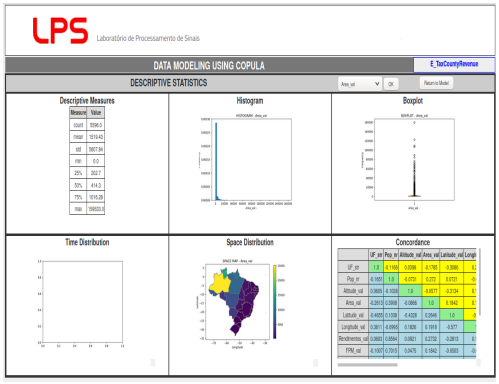
(a) Feature "UF".



(b) Feature "Pop".

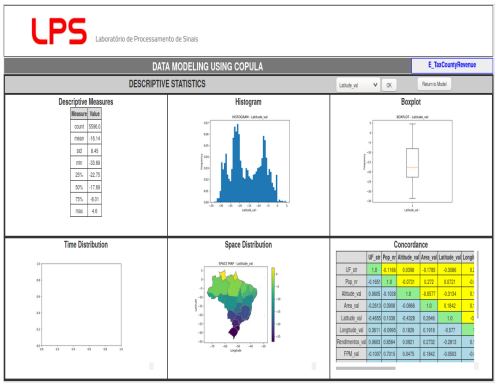


(c) Feature "Altitude".

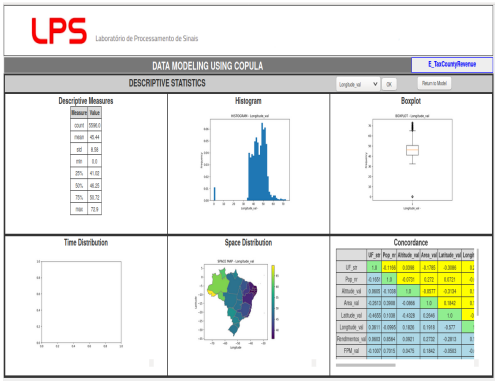


(d) Feature "Area".

"UF"

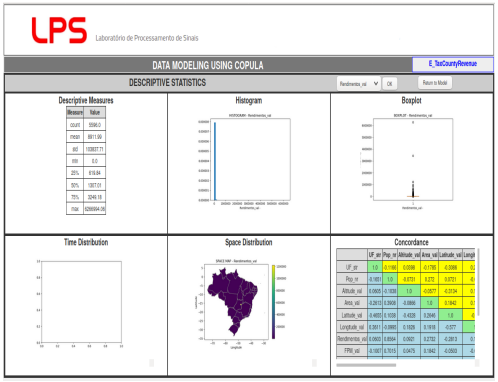


(e) Feature "Latitude".

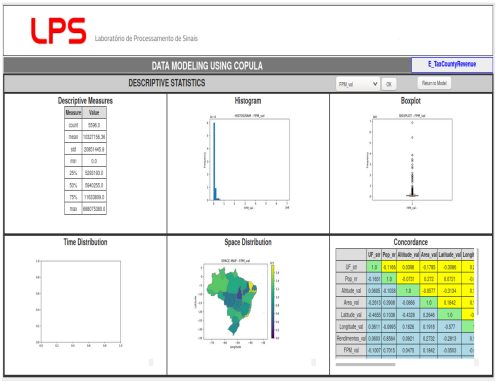


(f) Feature "Longitude".

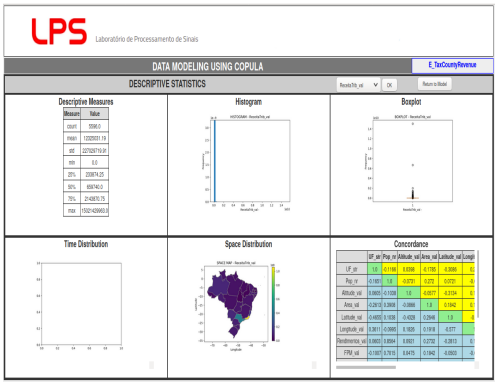
Figure 64 – Brazilian counties tax revenue dataset - descriptive statistics - Part 1.



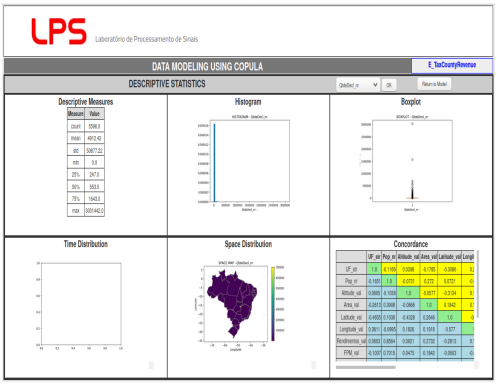
(a) Feature "Rendimentos".



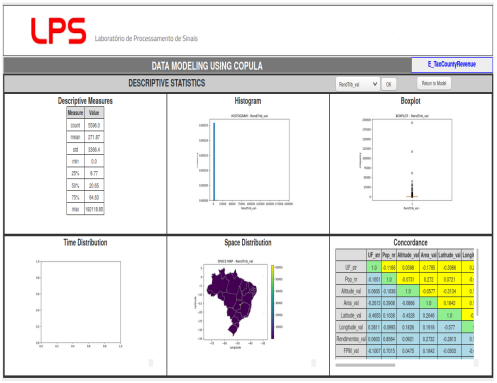
(b) Feature "FPM".



(c) Feature "ReceitaTrib".



(d) Feature "QtdeDecl".

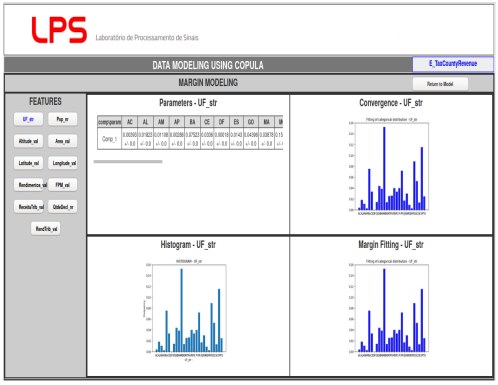


(e) Feature "RendTrib".

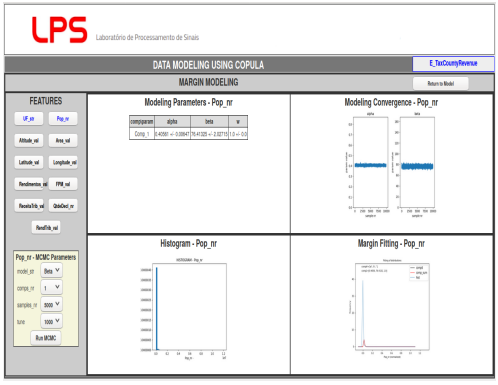
Figure 65 – Brazilian counties tax revenue dataset - descriptive statistics - Part 2.

4.5.1 MCMC Marginal Distribution Fitting

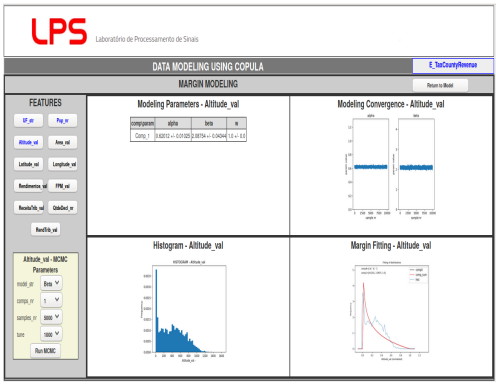
Modeling each categorical (only "UF" in this case) or numeric feature marginal distribution using, respectively, frequency-oriented multinomial and MCMC, with the same non-informative uniform prior and Beta distribution premises, the results are presented in Figures 66 and 67.



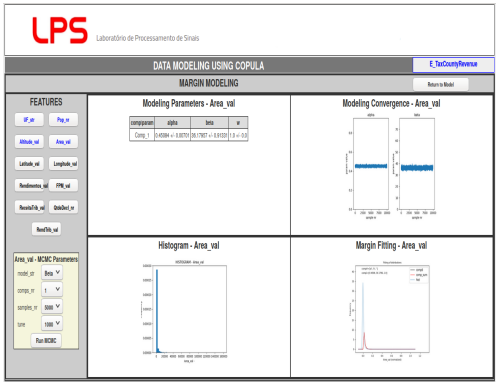
(a) Margin fitting for variable "UF".



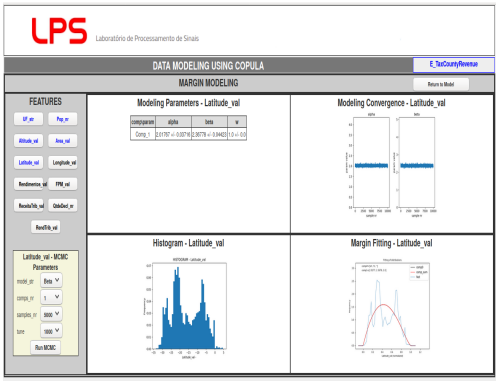
(b) Margin fitting for variable "Pop".



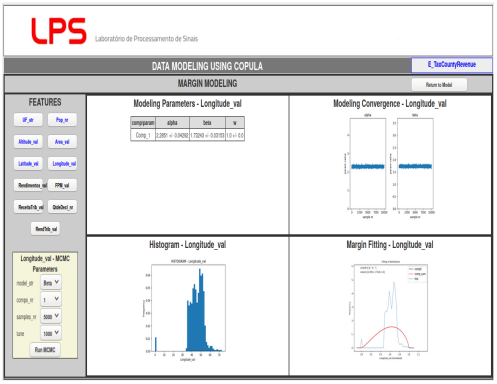
(a) Margin fitting for variable "Altitude".



(b) Margin fitting for variable "Area".

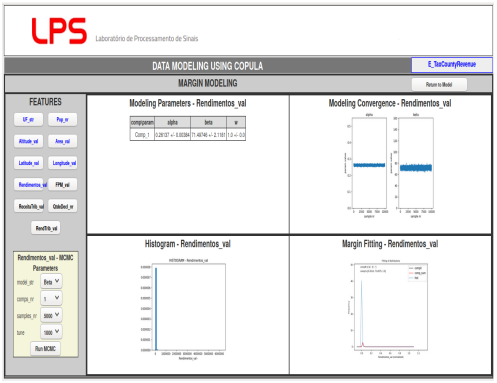


(a) Margin fitting for variable "Latitude".

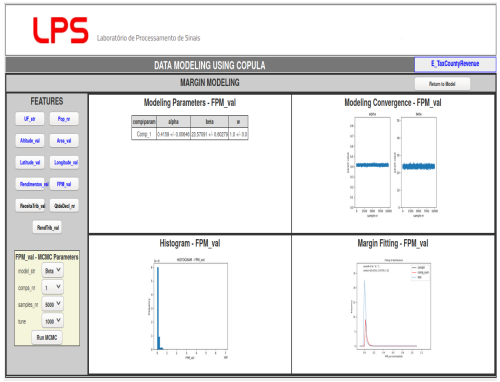


(b) Margin fitting for variable "Longitude".

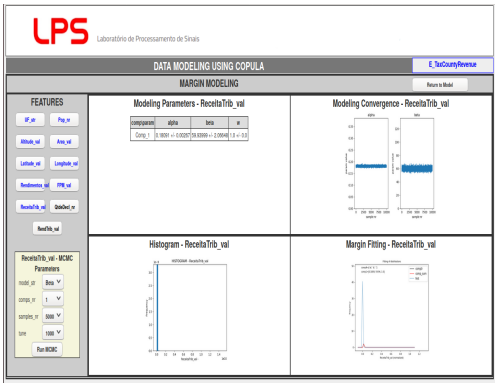
Figure 66 – Brazilian counties tax revenue dataset - marginal distribution fitting with multinomial or parametric Beta and non-informative uniform prior by Bayesian MCMC modeling - Part 1.



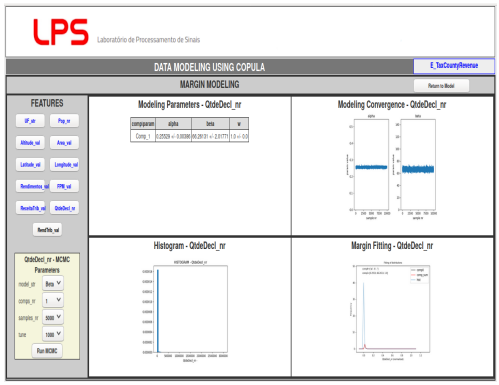
(a) Margin fitting for variable "Rendimentos".



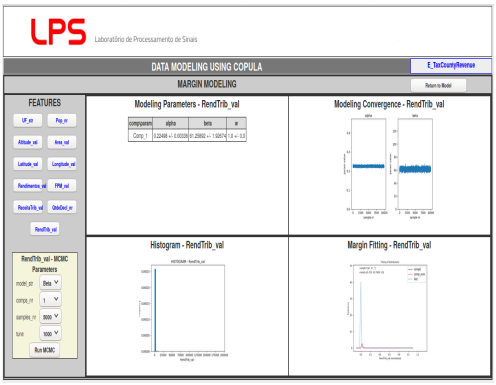
(b) Margin fitting for variable "FPM".



(a) Margin fitting for variable "ReceitaTrib".



(b) Margin fitting for variable "QtdeDecl".

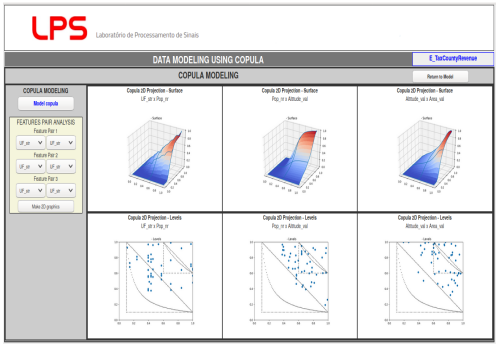


(a) Margin fitting for variable "RendTrib".

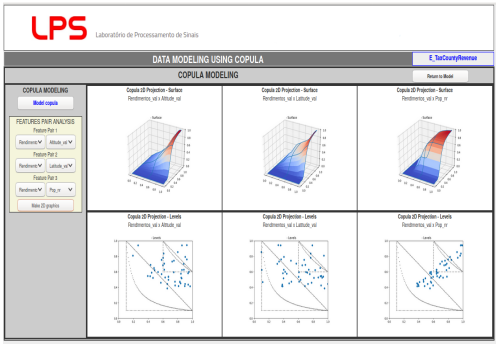
Figure 67 – Brazilian counties tax revenue dataset - marginal distribution fitting with multinomial or parametric Beta and non-informative uniform prior by Bayesian MCMC modeling - Part 2.

4.5.2 Empirical Copula Modeling

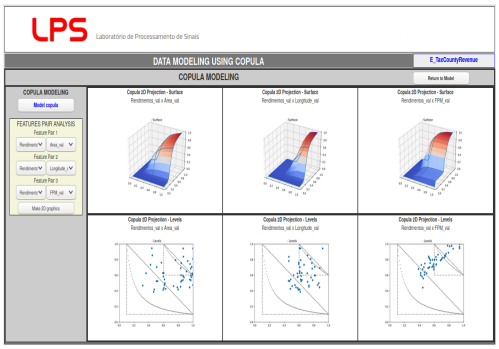
Going on, the empirical copula model is produced as shown in Figure 68 and Figure 69. Again, the considerable number of variables for copula matters leads to great sparsity and spoils empirical copula sample projection patterns.



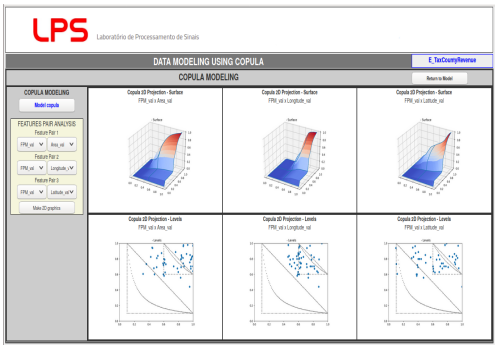
(a) "UF", "Pop", "Altitude", "Area".



(b) "UF", "Pop", "Altitude", "Area".

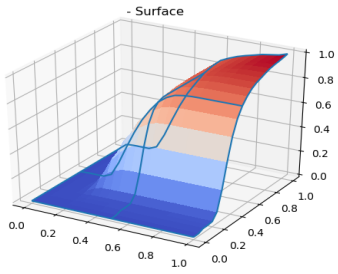


(b) "Rendimento", "Area", "Longitude", "FPM".

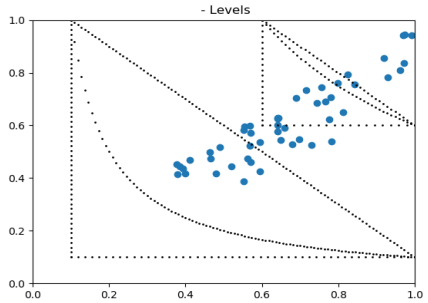


(d) "FPM", "Longitude", "Latitude", "Area".

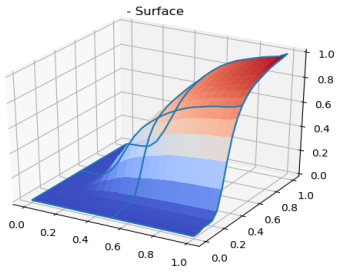
Figure 68 – Brazilian counties tax revenue dataset - empirical copula figures in LpsCopModel software tool. The most relevant projections were chosen to be presented.



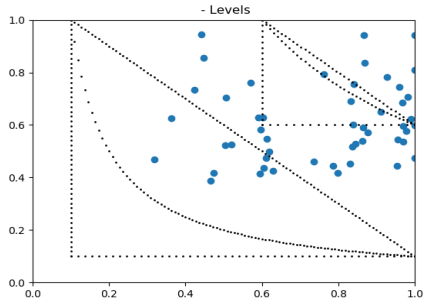
(a1) "Rendimentos"-"Pop" - surface.



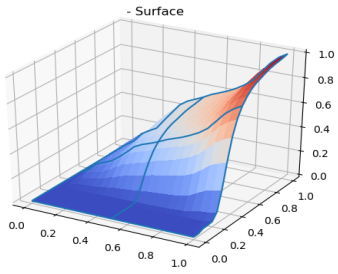
(a2) "Rendimentos"-"Pop" - level curves.



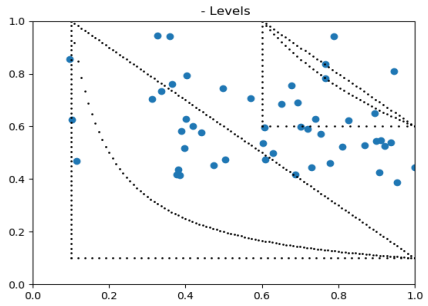
(b1) "Rendimentos"-"Area" - surface.



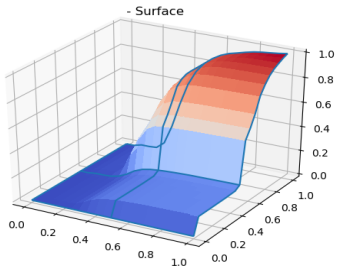
(b2) "Rendimentos"-"Area" - level curves.



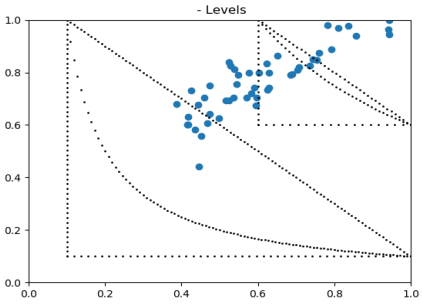
(c1) "Rendimentos"-"Latitude" - surface.



(c2) "Rendimentos"-"Latitude" - level curves.



(d1) "Rendimentos"-"FPM" - surface.

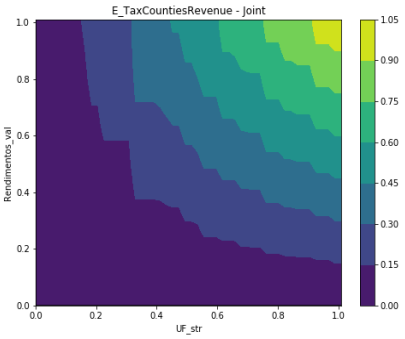


(d2) "Rendimentos"-"FPM" - level curves.

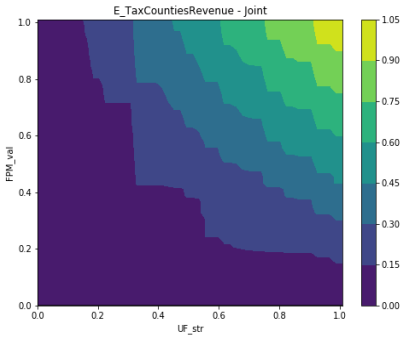
Figure 69 – DATASUS healthcare sample dataset - empirical copula 2D projection surfaces and level curves.

4.5.3 Non-Linear Normalization by Sample Reducing

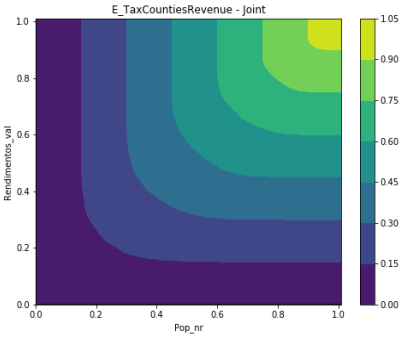
While marginal distributions became uniform distributions when normalized by sample reduction, the joint distribution for this real dataset of the non-linear normalized variables shows a different pattern according to the projection variables real behavior, some regularity patterns show up when strong dependence is present between the variables pair (Figure 70).



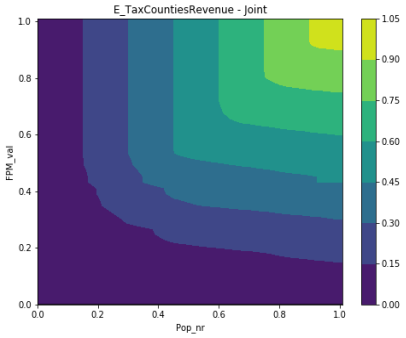
(a) "UF"-"Rendimentos".



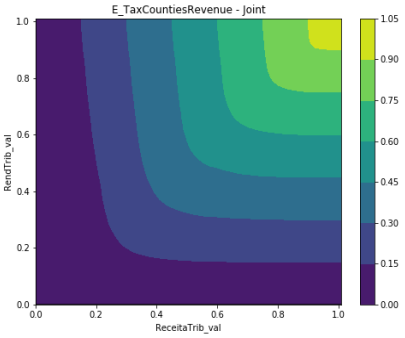
(b) "UF"-"FPM".



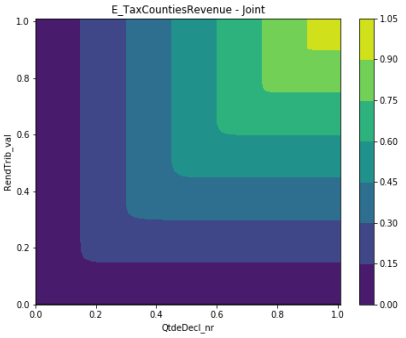
(c) "Pop"-"Rendimentos".



(d) "Pop"-"FPM".



(e) "ReceitaTrib"-"RendTrib".



(f) "QtdeDecl"-"RendTrib".

Figure 70 – Brazilian Counties Tax Revenue dataset - normalized joint distribution projection.

4.5.4 Bayesian Network Copula Modeling

The variation of the best score in each edges number group for all three normalization methods is presented in Figure 71 and all methods presented very similar tendencies as the best score number of edges was the same for all of them and equal to 10. Figure 72 shows the best Bayesian network structures and its scores for original, MCMC marginal distribution fitting and normalized datasets for that number of edges, and, just as before, one structure was highlighted when all the methods were jointly considered.

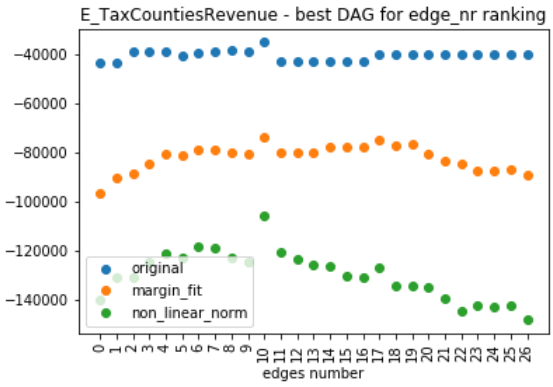


Figure 71 – Comparison of score variation among best Bayesian network score within each number of edges for all four normalization methods for the multivariate normal distribution cases.

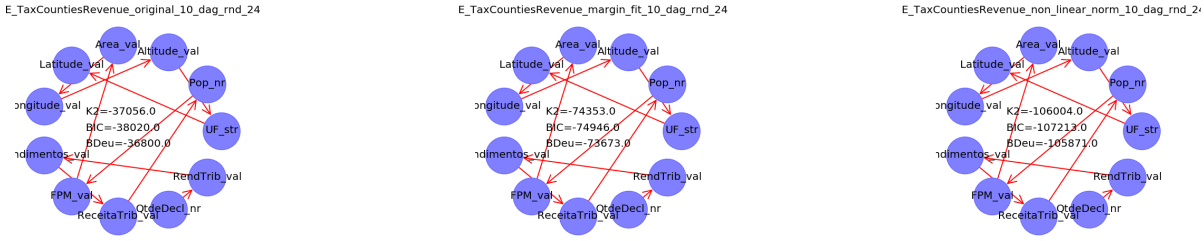


Figure 72 – Comparison of evidenced Bayesian network when all the three normalizations applied to the Brazilian counties tax revenue real case are considered together. Each column corresponds to a different normalization, from left to right: none, fitted marginals, and non-linear normalization. The figures show also the score measures for comparison.

Figure 73 shows score ranking for each possible network structure and none of the normalization methods has produced relevant perturbation in the tendency of the score ranking.

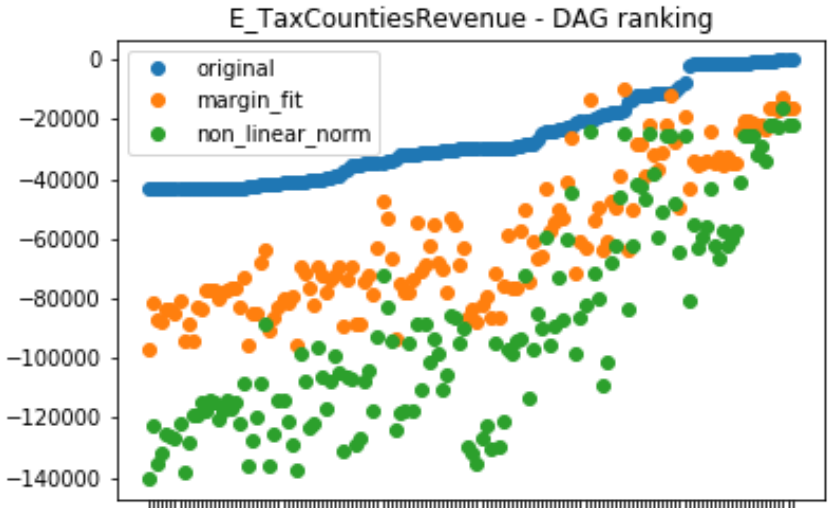


Figure 73 – Comparison of original, marginal distribution fitting and normalized Bayesian networks score ranking for the bivariate normal distribution cases. In the figures, each Bayesian network structure is a colored dot in the graphs where the horizontal axis represents different network structures and the vertical stands for scores. Different dot colors represent different normalization techniques: blue is for no normalization (original dataset), orange for marginal distribution fitting normalization and green for the proposed non-linear normalization.

4.6 Results Summary

Analyzing the results ensemble, we can say that the more direct result is that the non-linear normalization by sample reducing method showed strong viability as it pointed to similar tendencies and classification in the Bayesian network structure search.

In a more abstract sense, a Bayesian networks established method was successfully applied for a copula modeling instead of a joint distribution, in the sense that no disturbance or specific problem arise during the modeling caused by its copula subject.

One interesting point for the simulated datasets is that the non-linear normalization by sample reducing practically matched the scores computed for the known marginal distribution fitting sample transformation, suggesting that the non-linear normalization could be a good approximation for the real marginal distribution in a dataset where other fitting methods present poor performance. Of course, that might depend on the number of samples available, reinforcing our large enough sample premise.

Another important result was the tendency alignment presented by the non-linear normalization by sample reducing with all other methods. Although score ordering was not preserved in a point-wise sense, an overall tendency was preserved leading to coherence in structure selection.

Finally, compared to the direct use of original data with discretization as input for the Bayesian network modeling, the non-linear normalization by sample reducing showed an interesting dispersion through a broader range of scores which can be further studied for possible benefits in processing cost and speed for search algorithms.

5 CONCLUSIONS

Machine learning is an exponentially growing field in the last decade and many techniques have evolved as others have emerged from the fusion of previously separated ones, but it is still not usual to observe a common knowledge or use of copulas among those specialists.

Our purpose with this research was to perform a relevant step to promote more discussion on the possibilities that the concept of copula could offer for Machine Learning modeling. As already mentioned, we did not focus on copula tools for machine learning or Bayesian networks as the previous works on CBN or PCC existing in literature, but in the very concept of copula as a modeling approach, by decoupling individual variables behavior from its associations, aiming to avoid that one specific part of the modeling disturbs or conditions the other, as is the case, for example, of traditional initial discretization stages for modeling discrete Bayesian networks.

The results have proved the utility of using this technique of non-linear normalization as a stage for modeling Bayesian networks (BNs) and focusing on modeling the copula of a given phenomenon by that BN instead of its joint distribution, without loss of generality.

Therefore, one important contribution of this research to the scientific community can be described as one more push for a multidisciplinary approach both in Bayesian networks and copula fields regarding the introduction of the copula concept and Sklar's decoupling theorem in the BN modeling paradigm by the results here presented, indicating that copulas can be modeled by Bayesian networks just as joint distributions are and that probability decoupling may be a tool to be used in BN modeling. Hence, this multidisciplinary approach can benefit both areas, copulas and BN.

Results also showed that copulas can be modeled by Bayesian networks just as any generic joint distributions are.

A relevant point to be observed is the dispersion showed by the structures scoring after non-linear normalization in relation to all other techniques which may improve performance in some of the score-based structure search algorithms already developed in literature.

Although not treated within this research, the copula modeling can be improved by adopting continuous/hybrid Bayesian networks for the parameter fitting after the structural modeling search stage using the best scored structures selected by that stage.

From this point, many additional research can be suggested, directly or indirectly related with the aspects and results treated on this text.

First, the sample reduction itself can be subject to improvements. The one used here is based on an empirical calibration but we also suggest an approach based on sample moving average. And so, any other similar technique can be tried.

Second, as already mentioned, the dispersion showed by the structures scoring after non-linear normalization claims for further research to see how it would improve performance in structure search algorithms, for example by testing the copula modeling with structure searching algorithms with known performances to compare both results and performance itself.

Another possible and interesting line of research is to use the empirical copula as an initial map on the dependencies behavior, mainly by observing non-linear local information on it instead of restricting only to global indexes, and, from that analysis, to guide the structure search by its global profile and its tail-body dependence profile distinctions.

Continuous/hybrid Bayesian networks can be tested as a further modeling stage for better approximation after the structural search using the best scored structures detected by the proposed methodology.

Finally, more research could be done aiming deeper studies on the theoretical foundations for the empirical results to orientate further improving, for example in terms of probabilistic decoupling effect on entropy.

Concluding this text, we can declare that our research proved itself extremely successful while capable of producing results very promising for an open number of areas within Machine Learning and Artificial Intelligence, and even beyond those areas, and also showed how powerful a multidisciplinary approach can be just by mixing different conceptual frameworks, even not in its most sophisticated versions.

REFERENCES

ACHCAR, J. A.; MARTINEZ EDSON ZANGIACOMI ANDTOVAR CUEVAS, J. R. Bivariate lifetime modelling using copula functions in presence of mixture and non-mixture cure fraction models, censored data and covariates. **Model Assisted Statistics and Applications**, 2016.

ANDRIEU, C. *et al.* An introduction to mcmc for machine learning. **Machine Learning**, 2003.

AVRAHAM, W. C. U. **Density and contour plot of a Bivariate Gaussian Distribution. The density of the join distribution is obtained by joining a Gaussian Copula (rho=0.5) with two identical Gaussian univariate distributions (mean=0, sd=1).** 2008. Disponível em: https://commons.wikimedia.org/wiki/File:Gaussian_copula_gaussian_marginals.png.

AVRAHAM, W. C. U. **An example of the bivariate Gaussian (normal), Student-t, Gumbel, and Clayton copulæ**. 2015. Disponível em: https://commons.wikimedia.org/wiki/File:Four_Correlations.png.

BAUER, A.; CZADO, C. Pair-copula bayesian networks. **Journal of Computational and Graphical Statistics**, 2016.

BAUER, A.; CZADO, C.; KLEIN, T. Pair-copula constructions for non-gaussian. **The Canadian Journal of Statistics**, 2012.

BEDFORD, T.; COOKE, R. M. Vines - a new graphical model for dependent random variables. **Annals of Statistics**, 2002.

BIELZA, C.; NAGA, P. L. Bayesian networks in neuroscience: a survey. **Frontiers in Computational Neuroscience**, v. 8, p. 131, 2014.

BONDÁR, I. *et al.* Statistical analysis of wind-generated infrasound noise. 2005. Disponível em: https://www.researchgate.net/publication/324653884_Statistical_Analysis_of_Wind-Generated_Infrasound_Noise.

BOUZEBDA, S. On the strong approximation of bootstrapped empirical copula processes with applications. **Mathematical Methods of Statistics**, 2012.

BOUZEBDA, S. Kac's representation for empirical copula process from an asymptotic viewpoint. **Statistics and Probability Letters**, 2017.

BRASIL, M. d. S. **Banco de dados do SistemaÚnico de Saúde-DATASUS**. 2020. Disponível em: http://www.datasus.gov.br.

CAI, B. *et al.* Remaining useful life estimation of structure systems under the influence of multiple causes: Subsea pipelines as a case study. **IEEE Transactions on Industrial Electronics**, 2020.

CHATRABGOUN, O. *et al.* Copula-based probabilistic assessment of intensity and duration of cold episodes: A case study of malayer vineyard region. **Agricultural and Forest Meteorology**, 2020.

CHEN, E. Y. J.; DARWICHE, A.; CHOI, A. On pruning with the mdl score. **International Journal of Approximate Reasoning**, 2018.

CHICKERING, D. M.; GEIGER, D.; HECKERMAN, D. **Learning Bayesian Networks is NP-hard**. [*S.l.*], 1994.

CHOI, K. K.; NOH, Y.; DU, L. Reliability based design optimization with correlated input variables. **SAE Technical Papers**, 2007.

COOPER, G. F. The computational complexity of probabilistic inference using bayesian belief networks. **Artificial Intelligence**, 1990.

CUNNANE, C. Unbiased plotting positions - a review - comments. **Journal of Hydrology**, 1979.

DEHEUVELS, P. La fonction de dépendance empirique et ses propriétés. un test non paramétrique d'indépendance. **Bulletin de la Classe des sciences**, 1979.

DIKS, C. *et al.* Comparing the accuracy of multivariate density forecasts in selected regions of the copula support. **Journal of Economic Dynamics and Control**, 2014.

DJANGO SOFTWARE FOUNDATION. **Django**. Django Software Foundation, 2013. Disponível em: https://djangoproject.com.

DURANTE, F.; FERNÁNDEZ-SÁNCHEZ, J.; TRUTSCHNIG, W. On the singular components of a copula. **Journal of Applied Probability**, 2015.

ECHAUST, K. Asymmetric tail dependence between stock market returns and implied volatility. **The Journal of Economic Asymmetries**, 2021.

ELIDAN, G. Copula bayesian networks. **Advances in Neural Information Processing Systems**, 2010.

ELIDAN, G. Inference-less density estimation using copula bayesian networks. *In*: **Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence, UAI 2010**. [*S.l.: s.n.*], 2010.

ELIDAN, G. Copula network classifiers (cncs). **Journal of Machine Learning Research**, 2012.

ELIDAN, G. Copulae in mathematical and quantitative finance. *In*: **Lecture Notes in Statistics-Proceedings.. . . .** [*S.l.: s.n.*], 2013.

FERMANIAN, J. D. Goodness-of-fit tests for copulas. **Journal of Multivariate Analysis**, 2005.

GELMAN, A.; RUBIN, D. B. Inference from iterative simulation using multiple sequences. **Statistical Science**, 1992.

GORDON, A. D. *et al.* **Probabilistic programming. In Proceedings of the on Future of Software Engineering**. [*S.l.: s.n.*]: ACM, New York, NY, USA, 2014. v. 82. 167-181 p.

GOSWAMI, M.; DAULTANI, Y.; DE, A. Decision modeling and analysis in new product development considering supply chain uncertainties: A multi-functional expert based approach. **Expert Systems with Applications**, 2021.

GROSS, T. J. *et al.* An analytical threshold for combining bayesian networks. **Knowledge-Based Systems**, 2019.

HAIR, J. F. J. *et al.* **Multivariate data analysis**. [*S.l.: s.n.*]: Prentice-Hall, Inc., 1998.

HAND, D. J. Finite mixture and markov switching models by sylvia frühwirth-schnatter. **International Statistical Review**, 2007.

HANEA, A. M.; KUROWICKA, D.; COOKE, R. M. Hybrid method for quantifying and analyzing bayesian belief nets. **Quality and Reliability Engineering International**, 2006.

HEKERMAN, D.; GEIGER, D.; CHICKERING, D. M. Learning bayesian networks: The combination of knowledge and statistical data. **Machine Learning**, v. 20, n. 3, p. 197–243, 1995.

HOFFMAN, M. D.; GELMAN, A. The no-u-turn sampler : Adaptively setting path lengths. **Journal of Machine Learning Research**, 2011.

HU, M.; LIANG, H. A copula approach to assessing granger causality. **NeuroImage**, 2014.

IPEA. **IPEADATA**. 2017. Disponível em: http://www.ipeadata.gov.br/.

JI, Q. *et al.* Probabilistic assessment of remote sensing-based terrestrial vegetation vulnerability to drought stress of the loess plateau in china. **Energy**, 2020.

JIANG, W.; CAO, Y.; DENG, X. A novel z-network model based on bayesian network and z-number. **IEEE Transactions on Fuzzy Systems**, 2020.

JOVANOVIC, S. *et al.* Copula as a dynamic measure of cardiovascular signal interactions. **Biomedical Signal Processing and Control**, 2018.

KARRA, K.; MILI, L. Hybrid copula bayesian networks. **Journal of Machine Learning Research**, 2016.

KIRSHNER, S. Learning with tree-averaged densities and distributions. **Advances in Neural Information Processing Systems 20 - Proceedings of the 2007 Conference**, 2009.

KOLLER, D.; FRIEDMAN, N. **Probabilistic graphical models: principles and techniques.** [*S.l.: s.n.*]: MIT press, 2009.

KUROWICKA, D. Conditionalization of copula-based models. **Decision Analysis**, 2012.

LARSON, H. J. **Introduction to Probability Theory and Statistical Inference**. [*S.l.: s.n.*]: John Wiley & Sons, Inc., 1982.

LIMA, E. L. **Curso de Análise, Volume 1**. [*S.l.: s.n.*]: Livros Técnicos e Científicos Editora S.A., 1976.

LIMA, E. L. **Curso de Análise, Volume 2**. [*S.l.: s.n.*]: Livros Técnicos e Científicos Editora S.A., 1981.

LIU, Y. *et al.* Comparison of support vector machine and copula-based nonlinear quantile regression for estimating the daily diffuse solar radiation: A case study in china. **Renewable Energy**, 2020.

LOPES, R. D. *et al.* Safety and efficacy of antithrombotic strategies in patients with atrial fibrillation undergoing percutaneous coronary intervention: A network meta-analysis of randomized controlled trials. **JAMA Cardiology**, 2019.

MÄCHLER, M. H. andMartin. Nested archimedean copulas meet r: The nacopula package. **Journal of Statistical Software**, 2011.

MATHWORKS. **Optimizing Market Risk using Copula Simulation**. 2013. Disponível em: https://www.mathworks.com/matlabcentral/mlc-downloads/downloads/submissions/42514/versions/4/previews/2_Analyze/html/copulaVaR.html.

NEAPOLITAN, R. E. **Learning Bayesian Networks**. [*S.l.: s.n.*]: Prentice-Hall, Inc., 2003.

NELSEN, R. B. **An Introduction to Copulas**. [*S.l.: s.n.*]: Springer-Verlag New York, Inc., 2006.

NOH, Y.; CHOI, K. K.; DU, L. New transformation of dependent input variables using copula for rbdo. *In*: **7th World Congresses of Structural and Multidisciplinary Optimization**. [*S.l.: s.n.*], 2007.

OWZAR, K.; SEN, P. K. Copulas: Concepts and novel applications. **Metron**, 2003.

PAPROTNY, D. *et al.* Banshee–a matlab toolbox for non-parametric bayesian networks. **SoftwareX**, 2020.

PATKI, N.; WEDGE, R.; VEERAMACHANENI, K. The synthetic data vault. **Proceedings - 3rd IEEE International Conference on Data Science and Advanced Analytics, DSAA 2016**, 2016.

PEARL, J. **Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference**. [*S.l.: s.n.*]: Morgan Kaufmann Publishers Inc., 1988.

PETERLE, V. C. U. *et al.* Indicators of morbidity and mortality by femur fractures in older people: A decade-long study in brazilian hospitals. **Acta Ortopedica Brasileira**, 2020.

RFB. **Receita Federal do Brasil - RFB Dados Abertos. Dados Econômico-Tributários e Aduaneiros da Receita Federal**. 2017. Disponível em: http://idg.receita.fazenda.gov.br/dados/receitadata/estudos-e-tributarios-e-aduaneiros/estudos-e-estatisticas/11-08-2014-grandes-numeros-dirpf/grandes-numeros-dirpf-capa.

ROBINSON, R. W. Counting unlabeled acyclic digraphs. *In*: **Combinatorial mathematics V**. [*S.l.: s.n.*]: Springer, 1977.

ROSENBLATT, M. Remarks on a multivariate transformation. **The Annals of Mathematical Statistics**, 1952.

SALVATIER, J.; WIECKI, T. V.; FONNESBECK, C. Probabilistic programming in python using pymc3. **PeerJ Computer Science**, 2016.

SCHMIDT, T. **Coping with Copulas**. 2006. 1–23 p.

SCHWEIZER, B.; WOLFF, E. F. On nonparametric measures of dependence for random variables. **Annals of Statistics**, 1981.

SCUTARI, M.; GRAAFLAND, C. E.; GUTIÉRREZ, J. M. Who learns better bayesian network structures: Accuracy and speed of structure learning algorithms. **International Journal of Approximate Reasoning**, 2019.

SEGERS, J.; SIBUYA, M.; TSUKAHARA, H. The empirical beta copula. **Journal of Multivariate Analysis**, 2017.

SEN, M. K.; DUTTA, S.; LASKAR, J. I. A hierarchical bayesian network model for flood resilience quantification of housing infrastructure systems. **ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineering**, 2021.

SHIRYAEV, A. N. **Probability**. [*S.l.: s.n.*]: Springer-Verlag New York, Inc., 1996.

SILVA, R.; GRAMACY, R. B. Mcmc methods for bayesian mixtures of copulas. **Journal of Machine Learning Research**, 2009.

SKLAR, M. n -dimensional distribution functions and their marginals. **Publications de l'Institut de Statistique de l'Université de Paris**, 1959.

STRAUB, D.; PAPAIOANNOU, I.; BETZ, W. Bayesian analysis of rare events. **Journal of Computational Physics**, 2016.

STRELEN, J. C. Tools for dependent simulation input with copulas. *In*: **SIMUTools 2009 - 2nd International ICST Conference on Simulation Tools and Techniques**. [*S.l.: s.n.*], 2009.

THE PANDAS DEVELOPMENT TEAM. **pandas-dev/pandas: Pandas**. Zenodo, 2020. Disponível em: https://doi.org/10.5281/zenodo.3509134.

TRAN, M. N. *et al.* Copula-type estimators for flexible multivariate density modeling using mixtures. **Journal of Computational and Graphical Statistics**, 2014.

VAART, A. van der. **Asymptotic Statistics**. [*S.l.: s.n.*]: Cambridge University Press, 1998.

VERMA, T.; PEARL, J. Equivalence and synthesis of causal models. *In*: **Proceedings of Sixth Conference on Uncertainty in Artificial Intelligence**. [*S.l.: s.n.*], 1991.

VILLANUEVA, E.; MACIEL, C. D. Modeling associations between genetic markers using bayesian networks. **Bioinformatics**, 2011.

VOSE, D. **Risk analysis: a quantitative guide**. [*S.l.: s.n.*]: John Wiley & Sons, 2009.

YUAN, Z.; HU, T. pyvine: The python package for regular vine copula modeling, sampling and testing. **Communications in Mathematics and Statistics**, 2019.

ZANDI, M. **Density and contour plot of a Bivariate Distribution, the density of the join distribution is obtained by joining a Gumbel Copula (param=2) with two identical Gaussian univariate distributions (mean=0, sd=1).** 2010. Disponível em: https://commons.wikimedia.org/wiki/File:Biv_gumbel_dist.png.

ZANDI, M. **Graph of the Fréchet-Hoeffding copula limits**. 2010. Disponível em: https://commons.wikimedia.org/wiki/File:Copule_ord.svg.

ZHANG, Q.; SHI, X. A mixture copula bayesian network model for multimodal genomic data. **bioRxiv**, 2017.

ZHAO, X.; SHANG, P.; LIN, A. Universal and non-universal properties of recurrence intervals of rare events. **Physica A: Statistical Mechanics and its Applications**, 2016.

ZHAO, Y. *et al.* Learning bayesian network structures under incremental construction curricula. **Neurocomputing**, 2017.

ZILKO, A. A. *et al.* The copula bayesian network with mixed discrete and continuous nodes to forecast railway disruption lengths. **6th International Conference on Railway Operations Modelling and Analysis**, 2015.

APPENDIX

APPENDIX A – THEORETICAL REFERENCES

This chapter has the sole intent of consolidating in a same place a minimal coherent block of theoretical content we have found most connected to our research for readability for those who want more detailed knowledge of one or more of the subjects treated in the main text. Therefore, everything here is just a reproduction from an external reference, even when expressed in different words.

Those main above mentioned references, along with any other specifically mentioned throughout the following text, are: Shiryaev (1996), Lima (1976), Lima (1981), Larson (1982), Neapolitan (2003), Nelsen (2006), Vose (2009), Elidan (2013), Vaart (1998), Gelman e Rubin (1992), Hand (2007), Andrieu *et al.* (2003), Zhao, Shang e Lin (2016), Hoffman e Gelman (2011), Salvatier, Wiecki e Fonnesbeck (2016), Avraham (2008), Zandi (2010a), Zandi (2010b), Avraham (2015), Strelen (2009), Gordon *et al.* (2014), Kirshner (2009), Silva e Gramacy (2009), Schweizer e Wolff (1981), Robinson (1977), Gross *et al.* (2019), Bielza e Naga (2014), and Hekerman, Geiger e Chickering (1995).

A.1 Statistics Basic Concepts

Statistics is a field of Science which deals with entities called sample spaces and random variables to make it possible to study non-deterministic phenomena and an essential part of its basis comes from probability theory. A formal approach of probability theory is the one based on set and measure theories where a probability model or a probability space is defined upon three elements: a sample space Ω, a σ-algebra $\mathcal{F}$ of subsets of Ω, and a probability P on $\mathcal{F}$. Good reference on probability theory by this approach is found in (SHIRYAEV, 1996), from where all following definitions, lemmas and theorems were extracted (unless otherwise noted).

A.1.1 Random Variables and Distributions

Definition: **Sample Space (Ω)** is the set of all possible elementary outcomes ω that might be observed from an experiment.

Definition: An **event** is any subset $A \subset \Omega$.

Definition: A system $\mathcal{F}$ of subsets of Ω is a **σ-algebra** if:

(a) $\Omega \in \mathcal{F}$,

(b) $A_n \in \mathcal{F} \implies \cup A_n \in \mathcal{F}, \cap A_n \in \mathcal{F}$,

(c) $A \in \mathcal{F} \implies \bar{A} \in \mathcal{F}$

Definition: The pair of a space Ω together with a σ-algebra $\mathcal{F}$ of its subsets is a **measurable space** $(\Omega, \mathcal{F})$.

Definition: Let $(\Omega, \mathcal{F})$ be a measurable space. A set function $P = P(A), A \in \mathcal{F}$, taking values in $[0, \infty]$, with $P(\Omega) = 1$, is a **probability measure** or a **probability** if, for all pairwise disjoint subsets $A_1, A_2, ...$ of $\mathcal{F}$ with $\sum A_n \in \mathcal{F}$:

$$P\left(\sum_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} P(A_n) \tag{A.1}$$

Definition: An ordered triple $(\Omega, \mathcal{F}, P)$ is called a **probabilistic model** or a **probabilistic space** when:

(a) Ω is a set of points ω,

(b) $\mathcal{F}$ is a σ-algebra of subsets of Ω,

(c) P is a probability on $\mathcal{F}$

Here Ω is the sample space or space of elementary events, the sets A in $\mathcal{F}$ are events, and $P(A)$ is the probability of the event A.

It is important to notice that for such a probabilistic space:

(a) $P(\emptyset) = 0$,

(b) $A, B \in \mathcal{F}, P(A \cup B) = P(A) + P(B) - P(A \cap B)$,

(c) $A, B \in \mathcal{F}, B \subseteq A \implies P(A) \leq P(B)$

(d) $A_n \in \mathcal{F}, n = 1, 2, ..., \cup A_n \in \mathcal{F}, \implies P(A_1 \cup A_2 \cup ...) \leq P(A_1) + P(A_2) + ...$

Definition: Let Ω be a sample space. $\mathcal{F}_* = \{\emptyset, \Omega\}$ and $\mathcal{F}^* = \{A : A \subseteq \Omega\}$ are called the "poorest" σ-algebra and the "richest" σ-algebra of Ω.

Lemma: Let $\mathcal{E}$ be a collection of subsets of Ω. Then there are a smallest algebra $\alpha(\mathcal{E})$ and a smallest σ-algebra $\sigma(\mathcal{E})$ containing all the sets that are in $\mathcal{E}$.

Definition: Let $R = (-\infty, \infty)$ be the real line and let $\mathcal{A}$ be the system of subsets of R which are finite sums of disjoint intervals of the form $(a, b]$, with $\infty \leq a < b < \infty$ and $(a, \infty]$ taken as (a, ∞):

$$A = \sum_{i=1}^{n} (a_i, b_i], n < \infty \implies A \in \mathcal{A} \tag{A.2}$$

$\mathcal{A}$ is an algebra but not a σ-algebra, since $\cup (0, 1 - 1/n] = (0, 1) \notin \mathcal{A}$. The **Borel algebra** $\mathcal{B}(R)$ is the smallest σ-algebra containing $\mathcal{A}$ and its sets are called **Borel sets**.

Definition: A **distribution function** $F = F(x)$ on the real line R is a function satisfying:

1. $F(x)$ is nondecreasing;
2. $F(-\infty) = \lim_{x \to -\inf} F(x) = 0$;
3. $F(\infty) = \lim_{x \to \inf} F(x) = 1$;
4. $F(x)$ is continuous on the right and has a limit on the left at each $x \in R$.

Let $(R, \mathcal{B}(R))$ be the measurable space defined by the real line R and the Borel σ-algebra $\mathcal{B}(R)$. Let $P = P(A)$ be a probability measure defined on the Borel subsets A of the real line. Take $A = (-\infty, x]$ and put $F(x) = P((-\infty, x]),\ x \in R$. This function F is the unique distribution function corresponding to the probability measure P and it can be proven that the converse is also true, and the probability measure P, called *Lebesgue-Stieltjes* probability measure, is constructed from the corresponding distribution F by taking $P((a, b]) = F(b) - F(a)$. This concept is intimately associated with measure theory and the Lebesgue measure.

Definition: (LIMA, 1976) The Lebesgue measure is the standard way of assigning a length, area or volume to subsets of Euclidean space. A subset A of R has **null Lebesgue measure** and is considered to be a null set in R if and only if:

$$\forall \epsilon, \exists \{I_n\},\ A \subset\ \cup_{i=1}^{\infty} I_i,\ \sum_{i=1}^{\infty} vol(I_i) < \epsilon \tag{A.3}$$

Definition: **Discrete probability measures** are probability measures P for which the corresponding distributions $F = F(x)$ are piecewise constant changing their values at the points $x_1, x2, ...$ In this case the measure is concentrated at those points with $P(\{x_k\}) > 0$, and $\sum_k P(\{x_k\}) = 1$. $F = F(x)$ is called a **discrete distribution**.

Definition: **Absolutely continuous probability measures** are probability measures P for which the corresponding distributions $F = F(x)$ are such that

$$F(x) = \int_{-\infty}^{x} f(t).dt \tag{A.4}$$

where $f(t)$ is a nonnegative function called **density** of the distribution function $F = F(x)$ or the **density of the probability distribution** and $F = F(x)$ is called an **absolutely continuous distribution**.

Definition: **Singular probability measures** are probability measures P for which the corresponding distributions $F = F(x)$ are continuous but have all their points of increases on sets of zero Lebesgue measure. $F = F(x)$ is called a **singular distribution**.

Theorem: Every distribution function $F = F(x)$ can be represented in the form $F = p_1.F_1 + p_2.F_2 + p_3.F_3$, where F_1 is discrete, F_2 is absolutely continuous and F_3 is singular, and p_1, p_2, p_3 are non-negative numbers with $p_1 + p_2 + p_3 = 1$.

Next we introduce the concept of random variable which will allow to avoid dealing with extremely nature diverse sample spaces by focusing on their corresponding random variable representations.

Definition: Let $(\Omega, \mathcal{F})$ be a measurable space and let $(R, \mathcal{B}(R))$ be the measurable space defined by the real line R and the Borel σ-algebra $\mathcal{B}(R)$. A real function $\xi = \xi(\omega)$ defined on $(\Omega, \mathcal{F})$ is an $\mathcal{F}$-measurable function, or a **random variable**, if

$$\{\omega : \xi(\omega) \in B\} \in \mathcal{F}, \forall B \in \mathcal{B}(R) \tag{A.5}$$

or, equivalently, if the inverse image $\xi^{-1}(B) = \{\omega : \xi(\omega) \in B\}$ is a measurable set in Ω.

The theory so-developed can be extended from the real line R to the n-dimensional real space R^n with some generalization in the concept of non-decreasing function.

Definition: An ordered set $(\eta_1(\omega), ..., \eta_n(\omega))$ of random variables is called an **n-dimensional random vector**.

Definition: The **difference operator** $\Delta_{a_i,b_i} : R^n \to R$, where $a_i \leq b_i$, is defined by the expression

$$\Delta_{a_i,b_i} F_n(x_1, ..., x_n) = F_n(x_1, ..., xi - 1, b_i, xi + 1, ..., x_n) - F_n(x_1, ..., xi - 1, a_i, xi + 1, ..., x_n) \tag{A.6}$$

Definition: An **n-dimensional distribution function** $F = F(\mathbf{x})$ on R^n is a function satisfying:

1. $F(\mathbf{x})$ is n-nondecreasing or quasi-monotone in the sense that $\Delta_{a_1,b_1}...\Delta_{a_n,b_n}F_n(x_1,...,x_n) \geq 0$ for arbitrary $a = (a_1,...,a_n), b = (b_1,...,b_n)$;
2. $\lim_{\mathbf{x}\to\mathbf{y}} F_n(\mathbf{x}) = 0$, where at least one coordinate of $\mathbf{y}$ is $-\infty$;
3. $F_n(+\infty,...,+\infty) = 1$;
4. $F(\mathbf{x})$ is continuous on the right with respect to the variables collectively, i.e. if $x^{(k)} \downarrow x,\ x^{(k)} = (x_1^{(k)},...,x_n^{(k)})$, then $F_n(x^{(k)}) \downarrow F_n(x), k \to \infty$.

Let $(R^n, \mathcal{B}(R^n))$ be the measurable space defined by the real n-dimensional space R^n and the Borel σ-algebra $\mathcal{B}(R^n)$. Let $P = P(A)$ be a probability measure defined on the Borel subsets A of R^n. Put $F_n(x_1,...,x_1) = P((-\infty,x_1])X...X\,(-\infty,x_n])$. This function F_n is the unique distribution function corresponding to the probability measure P and the converse is also true, and the probability measure P is constructed from the corresponding distribution F_n by taking $P((a,b]) = \Delta_{a_1,b_1}...\Delta_{a_n,b_n}F_n(x_1,...,x_n)$.

Definition: (LIMA, 1981) An **n-dimensional open block** is the Cartesian product $B = \prod_{i=1}^n (a_i,b_i) = (a_1,b_1)x...x(a_n,b_n) \subset R^n$. The **n-dimensional volume** is defined as $vol(C) = \prod_{i=1}^n (b_i - a_i)$. An **n-dimensional open cube** is an n-dimensional open block where all intervals have the same length $b_i - a_i = a$.

Definition: (LIMA, 1981) From the theory of Lebesgue for integration, the following definition can be derived: a set $A \subset \Omega$ has **null measure** $m(A) = 0$ when $\forall\epsilon > 0$, $\exists\{C_i\} = C_1, C_2, ..., C_i, ...$, a sequence of n-dimensional open cubes where

$$A \subset \cup_{i=1}^{\infty} C_i,\ \sum_{i=1}^{\infty} vol(C_i) < \epsilon \tag{A.7}$$

Remark: Just as adopted by (NELSEN, 2006), whenever a propriety applies to a given sample space or function domain except for subsets of Lebesgue measure zero the terms **"almost surely"** or **"almost everywhere"** will be used in this text.

Definition: Let $\xi = \xi(\omega)$ be a nonnegative random variable and $\{\xi_n\}_{n\geq 1}$ a constructed sequence of simple nonnegative random variables such that $\xi_n(\omega) \uparrow \xi(\omega),\ n \to \infty,\ \forall\xi \in \Omega$. The **Lebesgue integral** or the **expectation** of ξ is $E\xi = \lim_n E\xi_n$.

Definition: Let $\xi = \xi(\omega)$ be a random variable and $\{\xi_n\}_{n\geq 1}$ and let $\xi^+ = max(\xi, 0)$ and $\xi^+ = -min(\xi, 0)$ be two nonngeative random variables so defined. The **Lebesgue**

integral or the **expectation** of ξ, $E\xi$, **exists** of **is defined**, if $min(E\xi^+, E\xi^-) < \infty$, and, if so, $E\xi = E\xi^+ - E\xi^-$.

Definition: Let $\xi = \xi(\omega)$ be a random variable. The **variance** of ξ is $V\xi = E(\xi - E\xi)^2$, and the **standard deviation** of ξ is the number $\sigma = +\sqrt{V\xi}$.

Definition: Let $(\Omega, \mathcal{F}, P)$ be a probability space, and $A \in \mathcal{F}$ be an event such that $P(A) > 0$. Let $\mathcal{D} = \{D_1, D_2, ...\}$ be a countable decomposition with $P(D_i) > 0, i \geq 1$ and $\mathcal{G} = \sigma\{\mathcal{D}\}$ the decomposition σ-algebra. The **conditional probability** of B with respect to $\mathcal{D}$ is

$$P(B|\mathcal{D}) = \sum_{i \geq 1} P(B|D_i).I_{D_i}(w),\ P(B|D_i) = P(B.D_{i]})/P(D_i) \tag{A.8}$$

Definition: Let ξ be a nonnegative random variable with respect to the σ-algebra $\mathcal{G}$ previously defined. The **conditional expectation** of ξ is denoted by $E(\xi|\mathcal{G})(\omega)$ such that

1. $E(\xi|\mathcal{G})(\omega)$ is $\mathcal{G}$ measurable;
2. $\forall A \in \mathcal{G},\ \int_A \xi dP = \int_A E(\xi|\mathcal{G})dP$

Definition: Let $B \in \mathcal{F}$ and the σ-algebra $\mathcal{G} \subseteq \mathcal{F}$. Then $E(I_B|\mathcal{G})$ is called **conditional probability of B with respect to $\mathcal{G}$**.

Definition: A function $P(\omega; B)$, defined for all $\omega \in \Omega$ and $B \in \mathcal{F}$ is a **regular conditional probability** with respect to $\mathcal{G}$ if

1. $P(\omega; .)$ is a probability measure on $\mathcal{F}$ for every $\omega \in \Omega$;
2. $\forall B \in \mathcal{F},\ P(\omega; B) = P(B|\mathcal{G})(\omega)(a.s.)$

Theorem: Let $P(\omega; B)$ be a regular conditional probability with respect to $\mathcal{G}$ and let ξ be an integrable random variable. Then

$$E(\xi|\mathcal{G})(\omega) = \int_\Omega \xi(\tilde{\omega}).P(\omega; d\tilde{\omega})(a.s.) \tag{A.9}$$

Definition: Let $(E, \mathcal{E})$ be a measurable space, $X = X(\omega)$ a random element with values in E, and $\mathcal{G}$ a σ-algebra of $\mathcal{F}$. A function $Q(\omega; B)$, defined for $\omega \in \Omega$ and $B \in \mathcal{E}$ is a regular conditional distribution of X with respect to $\mathcal{G}$ if

1. $Q(\omega; B)$ is a probability measure on $(E, \mathcal{E}) \forall \omega \in \Omega$;

2. $\forall B \in \mathcal{E},\ Q(\omega; B) = P(X \in B|\mathcal{G})(\omega)(a.s.)$

Definition: Let ξ be a random variable. A function $F = F(\omega; x), \omega \in \Omega, x \in R$ is a **regular distribution function** for ω with respect to $\mathcal{G}$ if

1. $F(\omega; x)$ is a distribution function on $R,\ \forall\omega \in \Omega$;

2. $F(\omega; x) = P(\xi \leq x|\mathcal{G}(\omega)), \forall x \in R,\ (a.s.)$

Theorem: A regular distribution function and a regular conditional distribution function **always exist** for the random variable ξ with respect to $\mathcal{G}$.

Compatible to all that more formal construction are the usual following definitions of random variable, probability function and distribution function commonly present in the applied literature. Here we adopt the usual notation, where a random variable is noted by X instead of ξ. This will be adopted throughout this text.

Definition: (LARSON, 1982) **Random Variable (X)** is a function which associates a real number to each one of all subsets from a sample space S corresponding to a given experiment.

Definition: (LARSON, 1982) **Probability Function (P)** is a real-valued set function defined on the class of all subsets of a sample space S satisfying the following rules:

$$\begin{aligned} &1.P(S) = 1; \\ &2.P(A) \geq 0, \forall A \subset S; \\ &3.A \cap B = \emptyset \implies P(A \cup B) = P(A) + P(B) \end{aligned} \tag{A.10}$$

Definition: (LARSON, 1982) The **Distribution Function ($F_X(x)$)** for a random variable X is a function which gives the value of $P(X \leq x)$ for any real x.

Here are defined important basic measures for a random variable (LARSON, 1982).

Definition: The **mean** or **expected value** of a random variable X is:

1. if X is discrete: $\mu_X = E[X] = \sum_{x \in R_X} x.p_x(x)$

2. if X is continuous: $\mu_X = E[X] = \int_{-\infty}^{\infty} x.f_X(x)dx$

Definition: The **variance** of a random variable X is $\sigma_X^2 = E[(X_1 - E[X_1])^2]$ and the **standard deviation** is $\sigma_X = \sqrt{\sigma_X^2}$.

All those concepts can be extended to consider a set of variables instead of a single one (LARSON, 1982):

Definition: (LARSON, 1982) Let $\mathbf{X} = (X_1, X_2, ..., X_n)$ be a rule associating an n-tuple with each element ω of a sample space S. Then $\mathbf{X}$ is called an **n-dimensional random vector**. Probability function $p_{X_1,..,X_n}(x_1, ..., x_n)$ and distribution function $F_{X_1,..,X_n}(x_1, ..., x_n)$ of $\mathbf{X}$ are also called **joint probability function** and **joint distribution function** in this multivariate case.

Definition: (LARSON, 1982) The **joint distribution function ($F_X(x)$)** for a random vector X is a function which gives the value of $P(X_1 \leq x_1, ...X_n \leq x_n)$ for any real vector x.

Definition: (LARSON, 1982) (adapted) Let $\mathbf{X} = (X_1, X_2, ..., X_n)$ be an n-dimensional random vector. Then the **marginal probability function** for X_k, $k = 1, ..., n$, is:

$$p_{X_k}(x_k) = P(X_k = x_k) = \sum_{x_{i_1}} \sum_{x_{i_2}} ... \sum_{x_{i_{n-1}}} p_{X_1,..,X_n}(x_{i_1}, ..., x_{i_{k-1}}, x_k, x_{i_k}, ..., x_{i_{n-1}}), \quad i_j = 1, ..., k-1, k+1, ..., n \tag{A.11}$$

Definition: (LARSON, 1982) The **marginal distribution function ($F_{X_i}(x)$)** for a random variable X_i of a random vector $\mathbf{X}$ is a function which gives the value of $P(X_i \leq x_i)$ for any real value x_i where P is the probability function for $\mathbf{X}$.

Definition: (LARSON, 1982) Let $\mathbf{X} = (X_1, X_2, ..., X_n)$ be an n-dimensional <u>discrete</u> random vector with probability function $p_{X_1,..,X_n}(x_1, ..., x_n)$. Then the **conditional probability function** for X_k, $k = 1, ..., n$, , given $X_k = x$, is:

$$p_{X_{i_1},...,X_{i_{k-1}},X_{i_k},...,X_{i_{n-1}}|X_k}(x_{i_1}, ..., x_{i_{k-1}}, x_{i_k}, ..., x_{i_{n-1}}|x) = \frac{p_{X_1,..,X_n}(x_{i_1}, ..., x_{i_{k-1}}, x_k, x_{i_k}, ..., x_{i_{n-1}})}{p_{X_k}(x)}, \quad i_j = 1, ..., k-1, k+1, ..., n \tag{A.12}$$

Definition: (LARSON, 1982) Let $\mathbf{X} = (X_1, X_2, ..., X_n)$ be an n-dimensional <u>continuous</u> random vector with density function $f_{X_1,..,X_n}(x_1, ..., x_n)$. Then the **conditional proba-**

bility function for X_k, $k = 1, ..., n$, given $X_k = x$, is:

$$\begin{aligned} &f_{X_{i_1},...,X_{i_{k-1}},X_{i_k},...,X_{i_{n-1}}|X_k}(x_{i_1}, ..., x_{i_{k-1}}, x_{i_k}, ..., x_{i_{n-1}}|x) = \\ &\frac{f_{X_1,..,X_n}(x_{i_1}, ..., x_{i_{k-1}}, x_k, x_{i_k}, ..., x_{i_{n-1}})}{f_{X_k}(x)}, \\ &i_j = 1, ..., k-1, k+1, ..., n \end{aligned} \tag{A.13}$$

Definition: (LARSON, 1982) The random variables in the vector $\mathbf{X} = (X_1, X_2, ..., X_n)$ are **independent** if and only if, $\forall x_1, x2, ..., x_n$:

- (a) if X is discrete, $p_{X_1,X_2,...,X_n}(x_1, x_2, ..., x_n) = p_{X_1}(x_1).p_{X_2}(x_2)...p_{X_n}(x_n)$

- (b) if X is continuous, $f_{X_1,X_2,...,X_n}(x_1, x_2, ..., x_n) = f_{X_1}(x_1).p_{X_2}(x_2)...f_{X_n}(x_n)$

A.1.2 Discrete and Continuous Random Variables

On a more pure basis (SHIRYAEV, 1996) there are three important types of random variables to be considered when observing what characterizes its probability measure: discrete random variables, continuous random variables and its subset of absolutely continuous random variables.

Definition: A random variable ξ that has a representation

$$\xi(\omega) = \sum_{i=1}^{\infty} x_i.I_{A_i}(\omega),\ \sum A_i = \Omega,\ A_i \in \mathcal{F} \tag{A.14}$$

is called **discrete**. If the sum if finite, the random variable is called **simple**.

Definition: A probability measure P_ξ on $(R, \mathcal{B}(R))$ with

$$P_\xi(B) = P\{\omega : \xi(\omega) \in B\},\ B \in \mathcal{B}(R), \tag{A.15}$$

is called **probability distribution** of ξ on $\mathcal{B}(R)$.

Definition: The **distribution function** of ξ is the function:

$$F_\xi(x) = P(\omega : \xi(\omega) \leq x),\ x \in R \tag{A.16}$$

For a discrete random variable the measure P_ξ is concentrated on an at most countable set and can be represented in the form $P_\xi(B) = \sum_{\{k:x_k \in B\}} p(x_k)$, where $p(x_k) = P\{\xi : \xi = x_k\} = \Delta F_\xi(x_k)$.

Definition: A random variable ξ is called **continuous** if its distribution $F_\xi(x)$ is continuous for $x \in R$.

Definition: A random variable ξ is called **absolutely continuous** if there is a non-negative function $f = f_\xi(x)$, called its **density**, such that

$$F_\xi(x) = \int_{-\infty}^{x} f_\xi(y).dy,\ x \in R \tag{A.17}$$

It is also common to focus on the two main disjoint groups, discrete and continuous, and to make definitions more intuitive as follows (LARSON, 1982):

Definition: A random variable X is called **discrete** if its range R_x is a discrete set.

Definition: A random variable X is called **continuous** if its distribution function $F_x = P(x \leq t)$ is a continuous function of $t, t \in R$.

While distribution functions remain useful for both cases, probability functions of continuous random variables give information only about intervals but are always forced to zero in any individual point to respect its essential properties, so it is of great help to define another function in that case that gives useful information about the probability profile in each possible value for the variable (LARSON, 1982):

Definition: The **density function** (also called **probability density function** or **pdf**) for a continuous random variable X is

$$f_X(t) = \frac{dF_X(t)}{dt} \tag{A.18}$$

A.2 Dependence, Association, Correlation and Concordance

A phenomenon with more than one component or feature will necessarily leads to a model based on a random vector with as many random variables. When analyzing such a case, beyond each component individual behavior, it is also important to consider interdependencies between its uncertain components (VOSE, 2009). Detected correlations between observed data may represent a real logical relationship between variables, an external factor affecting both variables or a matter of pure chance where no real correlation actually exists and statistical confidence tests mus be run to help determine their nature (VOSE, 2009). Also, variables interdependence can be defined in some different forms.

The first concept to be visited in terms of interdependence between random variables is the one of dependence associated with the definition of conditional probability and with

Bayes' formula and theorem (SHIRYAEV, 1996):

Definition: The **conditional probability** of event B assuming event A is:

$$P(B|A) = P(A \cap B)/P(A), P(A) > 0 \tag{A.19}$$

From the definition of conditional probability it can be derived the following theorem for relating conditional probability between two random variables in both logical directions:

Bayes's Theorem: If the events $A_1, ..., A_n$ form a decomposition of S, then:

$$P(A_i|B) = \frac{P(A_i).P(B|A_i))}{\sum_{j=1}^{n} P(A_j.P(B|A_j))} \tag{A.20}$$

Definition: Events A and B are called **independent** if $P(A \cap B) = P(A).P(B)$.

Bayes's Formula: Let A and B events with $P(A) > 0$ and $P(B) > 0$, then:

$$P(A|B) = \frac{P(A).P(A|B)}{P(B)} \tag{A.21}$$

The definition of independence is naturally extended from events to their representative random variables (LARSON, 1982):

$$P(A|B) = \frac{P(A).P(A|B)}{P(B)} \tag{A.22}$$

When a set of variables is dependent, this dependence can assume many profiles and there are also many conceptual tools for eliciting and measuring that profile. We shall adopt here the terminology proposed by (NELSEN, 2006):

1. **association**: for any general dependence profile between random variables;
2. **concordance**: for non-linear dependence profile between random variables and its measuring; informally, a pair of random variables are concordant if "large" values of one tend to be associated with "large" values of the other and vice-versa;
3. **correlation**: for linear dependence profile between random variables and its measuring;

Definition: The **co-variance** of two random variables X_1 and X_2 is:

$$Cov[X_1, X_2] = E[(X_1 - E[X_1]).(X_2 - E[X_2])] \tag{A.23}$$

Definition: The **correlation** between two random variables X_1 and X_2 is:

$$\rho_{X_1X_2} = \frac{Cov[X_1, X_2]}{\sqrt{E[(X_1 - E[X_1])^2]}.\sqrt{E[(X_2 - E[X_2])^2]}} \tag{A.24}$$

A.3 Marginal Distribution Fitting

After selecting the relevant variables, one ends up with a corresponding set of marginal distributions as prior defined, and the task is to fit a distribution model for each variable based on the sample values.

(VOSE, 2009) (pp. 263-300) dedicates a full chapter on the subject of "Fitting distributions to data" and its importance for the risk analyst. He explains that this task can be done from two sources, available data and experts opinion, but, considering that the subject is to study a methodology applicable in various fields and not to treat any specific dataset, this study will restrain itself to fitting distributions to available data. The referenced author also mentions that data can be fit to empirical (non-parametric) or parametric distributions and the fitting can consider many approaches concerning its complexity, like a first-order distribution based only on variability or a second-order distribution taking into account both variability and uncertainty.

A.3.1 Empirical Distribution Fitting

For the purposes of starting this line of research, the empirical fitting (VAART, 1998) was taken as a first approach, as follows:

$$F_k^n(x) = \frac{1}{n}.\sum_{i=1}^{n} \mathbf{1}(X_k^i \leq x) \tag{A.25}$$

where **1** is the **indicator function** which is 1 if the argument expression is true and 0 otherwise.

The fitting of the obtained empirical distribution to the fitted sample can be checked by comparing a detailed sampling of that distribution and the sample itself. For better checking it is wise to use a remarkably superior sampling rate than the one associated to the original sample, signifying the sampling to have a much superior number of instances. This leads to a much smaller granularity in the distribution curve graph than in the sample and will reflect in the corresponding graphs as a horizontal line pattern for the more rarefied part of the probability distribution (the tail concerning higher values) while the distribution curve remains decreasing. For example, if the sample has $5,000$ instances and the margin distribution is discretized by a $50,000$ sampling, the distribution histogram will have unit steps about one-tenth the sample steps.

Of course, there are some limitations in using empirical fitting. This kind of non-parametric approach is disadvantageous for estimating probabilities very close to zero or one, when a very large sample size would be needed for avoiding excluding extreme values

not represented in the particular sample used.

A.3.2 Bayesian Inference based on MCMC

Another more sophisticated method for fitting a univariate distribution to a sample is by using Bayesian inference and sampling that distribution by an asymptotic method. For that we are going to need to dive into three theoretical concepts: the Monte Carlo method, the Markov Chain process and sampling methods such as Gibbs' or No-U-Turn-Sampler (NUTS).

A.3.2.1 Monte Carlo Method

The Monte Carlo method consists basically in repeating an experiment a great number of times for computing numeric properties from the complete sample and assume the computed values represent a good approximation of the real population values based on the probabilistic-statistical regularities assured by the theory. (SHIRYAEV, 1996) curiously mentions that one of its first uses was by R. Wolf in 1850 to compute the value of the number π by throwing a needle $5,000$ times between two parallel lines and counting how many times it remained in-between, coming up with a value of 3.1596.

Therefore, (ANDRIEU *et al.*, 2003) a Monte Carlo simulation is an technique for drawing an i.i.d. (independent and identically distributed) set of samples $\{x^{(i)}\}_{i=1}^{N}$ from a target probability density $p(x)$ defined on a high-dimensional space $\mathcal{X}$ so that it can be used to approximate the density with the empirical point-mass function

$$p_N(x) = \frac{1}{N}.\sum_{i=1}^{N}\delta_{X^{(i)}}(x), \tag{A.26}$$

where $\delta_{X^{(i)}}(x)$ denotes the delta-Dirac mass located at $x^{(i)}$.

The viability of using that approximation is granted by the following result (ANDRIEU *et al.*, 2003):

$$I_N(f) = \frac{1}{N}\sum_{i=1}^{N} f(x^{(i)}) \xrightarrow[N\to\infty]{a.s.} I(f) = \int_{\mathcal{X}} f(x)p(x)dx \tag{A.27}$$

which means the estimate $I_N(f)$ is unbiased and it will almost surely converge to $I(f)$; and if the variance (considering a univariate case) is limited, then the variance of the estimator is inversely proportional to N and the central limit theorem yields convergence

in distribution of the error, i.e.:

$$\begin{aligned} &\sigma_f^2 \triangleq E_{p(x)}(f^2(x)) - I^2(f) < \infty \implies var(I_N(f)) = \frac{\sigma_f^2}{N} \\ &\sqrt{N}.(I_N(f) - I(f)) \impliedby [N \to \infty]\mathcal{N}(0, \sigma_f^2) \end{aligned} \tag{A.28}$$

This method associated with Markov chains and Bayesian inference is used to fit and sample a probability density when, among other reasons, it is too hard to compute it by deterministic means, as will be seen in the next subsections.

A.3.2.2 Markov Chain Processes

(SHIRYAEV, 1996) establishes formally the main definitions and results on Markov chain processes:

Definition: Let $(\Omega, \mathcal{F}, P)$ be a probability space with a non-decreasing family $(\mathcal{F}_n)$ of σ-algebras, $\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq ... \subseteq \mathcal{F}$. A sequence $X = (X_n, \mathcal{F}_n)$ where each X_n is $\mathcal{F}_n$-measurable is called a **stochastic sequence**.

Definition: Let $(\Omega, \mathcal{F}, P)$ be a probability space with a non-decreasing family $(\mathcal{F}_n)$ of σ-algebras, $\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq ... \subseteq \mathcal{F}$. A stochastic sequence $X = (X_n, \mathcal{F}_n)$ is called a **Markov chain** if

$$P\{X \in B|\mathcal{F}_m\} = P\{X_n \in B|X_m\}(P - a.s.), \forall n \geq m \geq 0, \forall B \in \mathcal{B}(R) \tag{A.29}$$

*"a.s." stands for "almost surely".
In the special case when $\mathcal{F}_n = \mathcal{F}_n^X = \sigma\{\omega : X_0, ..., X_n\}$ the sequence (X_n) is called itself a Markov chain.

Definition: The functions $P_n = P_n(x, B)$ such that $P_n(X_{n-1}; B) = P(X_n \in B|X_{n-1})$ are called **transition functions**. When they coincide, i.e. $P_1 = P_2 = ...$, the corresponding Markov chain is said to be **homogeneous**.

Definition: Let $(\Omega, \mathcal{F}, P)$ be a probability space and $\xi = (\xi_1, \xi_2, ...)$ a sequence of random variables, also said a random sequence. Let $\theta_k\xi$ denote the sequence $(\xi_{k+1}, \xi_{k+2}, ...$ A random sequence is said to be **stationary (in the strict sense)** if the probability distributions of $\theta_k\xi$ and ξ are the same for every $k \geq 1$.

Definition: A set $A \in \mathcal{F}$ is **invariant** with respect to a sequence ξ if there is a set $B \in \mathcal{B}(R\infty)$ such that $A = \{\omega : (\xi_1, \xi_2, ...) \in B\}$. The collection of all such invariant sets

is a σ-algebra denoted by $\mathcal{I}_\xi$.

Definition: A stationary sequence ξ is **ergodic** if the measure of every invariant set is either 0 or 1.

Ergodic Theorem: Let $\xi = (\xi_1, \xi_2, ...)$ be a stationary (strict sense) random sequence with $E|\xi_1| < \infty$. Then

$$\lim \frac{1}{n} \sum_{k=1}^{n} \xi_k(\omega) = E(\xi_1 | \mathcal{I}_\xi) \tag{A.30}$$

If ξ is also an ergodic sequence, then

$$\lim \frac{1}{n} \sum_{k=1}^{n} \xi_k(\omega) = E\xi_1 \tag{A.31}$$

A.3.2.3 MCMC Algorithm

Basically, Monte Carlo Markov Chain (MCMC) algorithms are based in assuming a group of conditional distributions whose composition results in the joint distribution of all the random variables involved, there included both independent variables, covariates and distribution parameters, and then using a sampling methodology for walking in steps on a Markov chain until enough convergence is achieved for some parameters, each step consisting of sampling from marginal distributions parameters through conditional distributions and obtaining a new posterior joint distribution from priors and likelihood at the gives sampled parameters (GELMAN; RUBIN, 1992) (HAND, 2007) (ANDRIEU *et al.*, 2003) (ZHAO; SHANG; LIN, 2016). Many different sampling methodologies are available, such as Gibbs sampling, Metropolis-Hastings sampling, a family based in Hamiltonian algorithms and so on (ANDRIEU *et al.*, 2003).

Algorithm convergence derives from Monte Carlo and Markov chain convergence results and the sampling methodology responds essentially for the convergence speed and computational costs (ANDRIEU *et al.*, 2003).

One of the most efficient and used in recent literature is the No-U-Turn Sampler (NUTS) (HOFFMAN; GELMAN, 2011) and that is the one we used in our research. This is one of the algorithms implemented in the Python package "pymc3" that we used in our tests (SALVATIER; WIECKI; FONNESBECK, 2016).

A.4 Copulas

The copula theory is one of the modeling techniques in Statistics. Its remarkable advantage is the possibility to deal separately with each random variable isolated behavior and the associations among those variables (NELSEN, 2006).

A.4.1 Concept of Copula

According to (NELSEN, 2006), an n-dimensional copula or n-copula is a function with the following definition:

Definition: an **n-dimensional copula** or **n-copula** is a function C from $I^n = [0,1]^n$ to $I = [0,1]$, for which:

1. if u in I^n has at least one coordinate equal to zero, then $C(u) = 0$ (grounded);
2. if $\boldsymbol{u}$ in I^n has all but u_k equal to one, then $C(\boldsymbol{u}) = u_k$;
3. for every $\boldsymbol{a}, \boldsymbol{b}$ in I^n such that $a_k \leq b_k$ for all k, then $V_c([\boldsymbol{a}, \boldsymbol{b}]) \geq 0$ (n-increasing);

where $B = [\boldsymbol{a}, \boldsymbol{b}]$ is the n-box $[a_1, b_1] \times ... \times [a_n, b_n]$, $V_c(B)$ is the $C - volume$ given by $V_{\boldsymbol{c}}(B) = sgn(\boldsymbol{c}).C(\boldsymbol{c})$ over all vertices $\boldsymbol{c}$ of B, where $sgn(\boldsymbol{c})$ is 1 for $\boldsymbol{c}$ having an even number of coordinates taken from $\boldsymbol{a}$ or -1 otherwise.

(NELSEN, 2006) also presents a fundamental result in copula theory named **Sklar's theorem in n-dimensions**, stating that:

Sklar's theorem in n-dimensions: for every n-dimension distribution function H with margins $F_1, ..., F_n$ there exists a n-copula C such that for all x in R^n:

$$H(x_1, ..., x_n) = C(F_1(x_1), ..., F_n(x_n)) \tag{A.32}$$

If all the margins are continuous, then C is unique; otherwise, C is uniquely determined on $RanF_1 \times ... \times RanF_n$, where $RanF_i$ stands for the range of F_i in its image set. Conversely, if C is a n-copula and $F_1, ..., F_n$ are distribution functions, then the function H defined by A.32 is a n-dimension distribution function with margins $F_1, ..., F_n$.

According to (NELSEN, 2006) for the bivariate case, one thing to remark is that, unlike bivariate distributions in general, a copula has no "atoms" (individual points in I^2

with non-zero C-measure) because it has always continuous margins. This leads to the following natural decomposition for any copula in two parts (not necessarily copulas):

$$C(u,v) = A_C(u,v) + S_C(u,v),\ A_C(u,v) = \int_0^u \int_0^v \frac{\partial^2}{\partial s \partial t} C(s,t),\ S_c = C - A_C \quad \text{(A.33)}$$

where A_C is the absolute continuous component and S_C is the singular component. Obviously, if $C \equiv A_C$, then C has a joint density $\frac{\partial^2 C(u,v)}{\partial u \partial v}$ and is absolutely continuous in I^2, whereas if $C \equiv S_C$, $\frac{\partial^2 C(u,v)}{\partial u \partial v} = 0$ almost everywhere in I^2 and C is singular.

Definition: The **support** of a copula $C(u,v)$ is the complement of the union of all open subsets of I^2 with C-measure zero. C has full support when its support is I^2 itself and when C is singular its support has Lebesgue measure zero.

Definition: (ELIDAN, 2013) Let C be an n-copula with marginal distributions $F_1, ..., F_n$ on a random vector X and corresponding marginal densities $f_1, ..., f_n$, which means its joint distribution is defined by $F_X(x_1, ..., x_n) = C(F_1(x_1), ..., F_n(x_n))$. If C has n'th order partial derivatives, then the **copula density** is defined by

$$c(F_1(x_1), ..., F_n(x_n)) = \frac{\partial^n C(F_1(x_1), ..., F_n(x_n))}{\partial F_1(x_1), ..., \partial F_n(x_n)} \quad \text{(A.34)}$$

and the joint density can be derived from the copula density using the derivative chain rule

$$f_X(x) = c. \prod_i f_i(x_i) \quad \text{(A.35)}$$

Along the time since their introduction, many copula families were constructed based on copula definition hypothesis. Just for illustration, the following copula families appear in (NELSEN, 2006):

1. Marshall-Olkin or Generalized Cuadras-Auge;
2. Farlie-Gumbel-Morgenstern (FGM);
3. t-copula;
4. Archimedian families:
    a) Frechet;
    b) Frank;
    c) Cook and Johnson or Pareto;
    d) Ali-Mikhail-Haq;
    e) Gumbel-Hougaard;
    f) Gumbel-Barnett;

5. Empirical;

Each family tends to represent some specific characteristic mapping for the correlation among the random variables. For example, Figures 74, 75 and 76 show graphs for different families and it is very remarkable the differences between them, mainly regarding their extremities and Figure 77 presents scatterplots from various copulas for emphasizing the different correlation profiles.

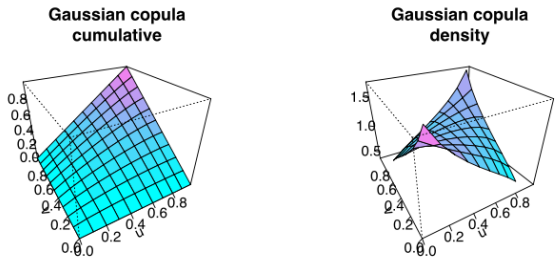


Figure 74 – Density and contour plot of a Bivariate Gaussian Distribution. The density of the join distribution is obtained by joining a Gaussian Copula (rho=0.5) with two identical Gaussian univariate distributions (mean=0, sd=1). (AVRAHAM, 2008)

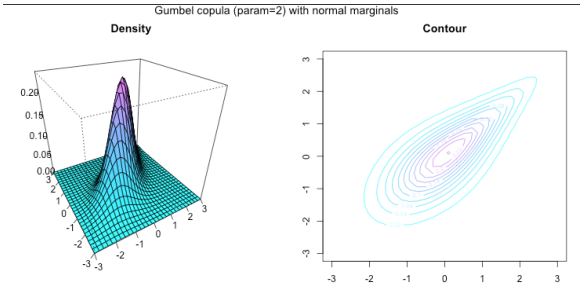


Figure 75 – Density and contour plot of a Bivariate Distribution, the density of the join distribution is obtained by joining a Gumbel Copula (param=2) with two identical Gaussian univariate distributions (mean=0, sd=1). (ZANDI, 2010a)

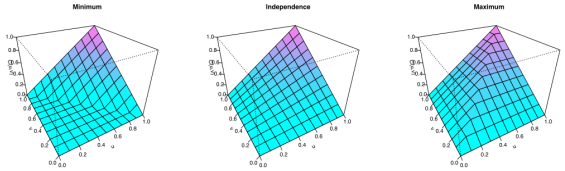


Figure 76 – Graph of the Frechet-Hoeffding copula limits and of the independence copula (middle). (ZANDI, 2010b)

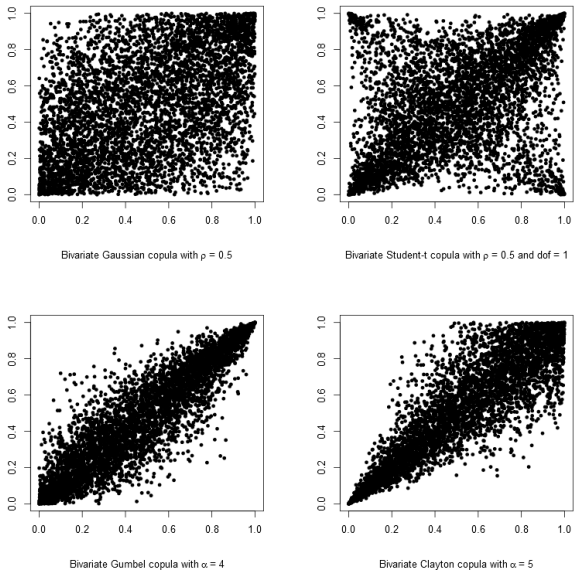


Figure 77 – Comparison between examples of the bivariate Gaussian (normal), Student-t, Gumbel, and Clayton copula scatterplots. (AVRAHAM, 2015)

From another perspective, Figure 78 presents the differences between pairwise correlation among random variables in its original sample values and after modeling by an adequate parametrized copula.

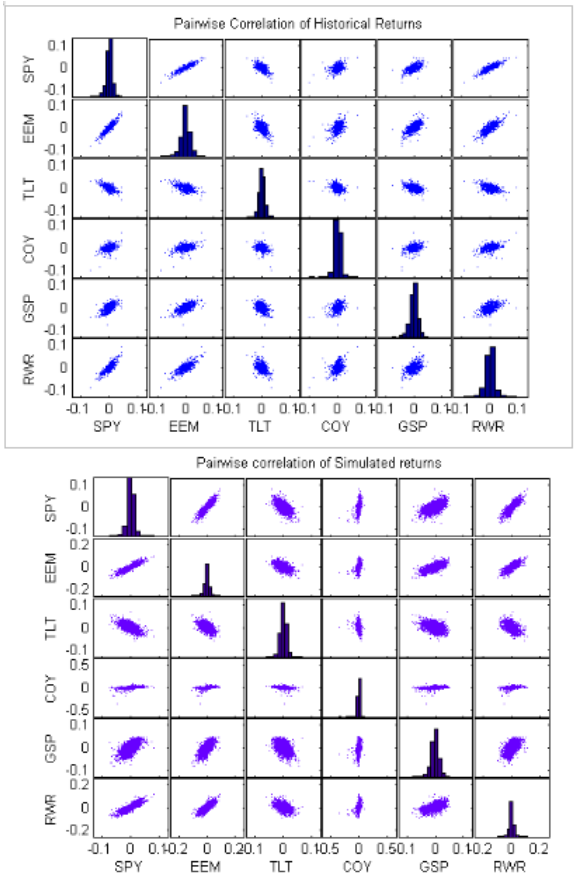


Figure 78 – Comparison between original sample and simulated sample after modeling by parametrization of a copula from a t-copula family (MATHWORKS, 2013).

A.4.2 Referential Copulas and Correlated Results

We adopt in this text the term "referential copulas" for those copulas which are conceptually constructed as reference for canonical dependence relations: complete dependence or complete independence. Those reference copulas derive from the following expressed results.

Theorem: Let C be a copula. Then:

$$\forall(u,v) \in DomC,\ max(u+v-1,\ 0) \leq C(u,v) \leq min(u,v) \tag{A.36}$$

Indeed, both the inferior and superior bounds in that theorem are themselves copulas and, along with a third important copula, the product one, complete the referential copulas defined as follows.

Definition: The **Fréchet-Hoeffding lower bound copula** W is the copula defined as $W(u,v) = max(u+v-1,\ 0), \forall(u,v) \in I^2$.

Definition: The **Fréchet-Hoeffding lower bound copula** M is the copula defined as $W(u,v) = min(u,v), \forall(u,v) \in I^2$.

Definition: The **product copula** Π is defined as $\Pi(u,v) = u.v, \forall(u,v) \in I^2$.

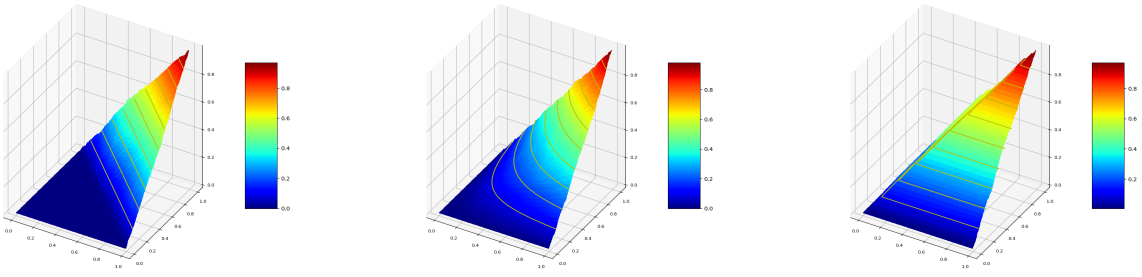
For the complete dependence case, in two dimensions, the Fréchet-Hoeffding copulas derived from the corresponding inequalities are those references, while the product copula is the reference for the complete independence case.

In a intuitive sense a complete independent copula means that all random variables are independent and given any set of fixed values for some, the probabilities for the others remain homogeneously distributed among all possible values. In contrast, complete positive dependent or *comonotonic* copula represents variables that grows altogether; and, in the inverse perspective, complete negative dependent or *countermonotonic* copula represents variables that decrease altogether. When limited to two variables, it is also possible to define the complete negative dependent copula for that case in which every growth in one variable corresponds to a decrease in the other and vice-versa, but this concept is not trivially extendable for more than two variables, although there also is a corresponding lower bound for $n \geq 3$ but which is not a copula nor can be associate to that simple negative dependence intuition.

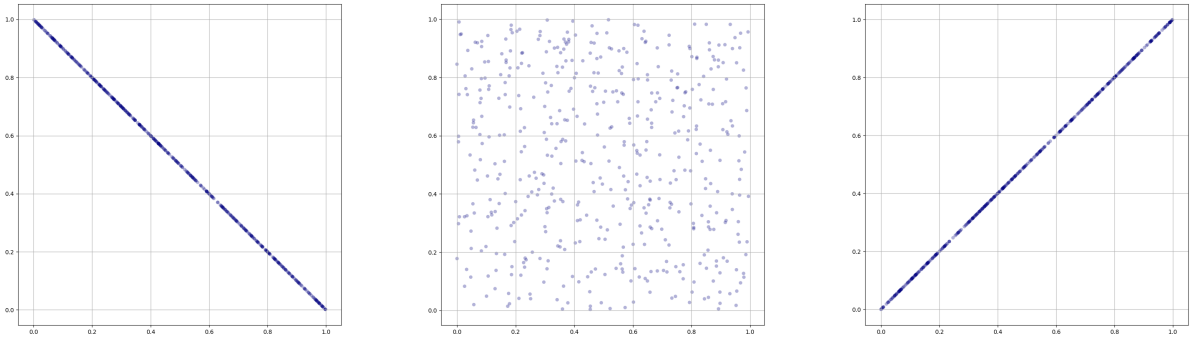
For the bivariate case, (NELSEN, 2006) remarks that the Fréchet-Hoeffding bounds suggests a partial order on the set of copulas, which can be extended under certain adaptation to the multivariate case, and the presents a definition for this order:

Definition: If C_1 and C_2 are copulas, it is said that C_1 is **smaller than** C_2 (or C_2 is **larger than** C_1), $C_1 \prec C_2$ ($C_1 \succ C_2$), if $\forall u, vinI,\ C_1(u,v) \leq C_2(u,v)$.

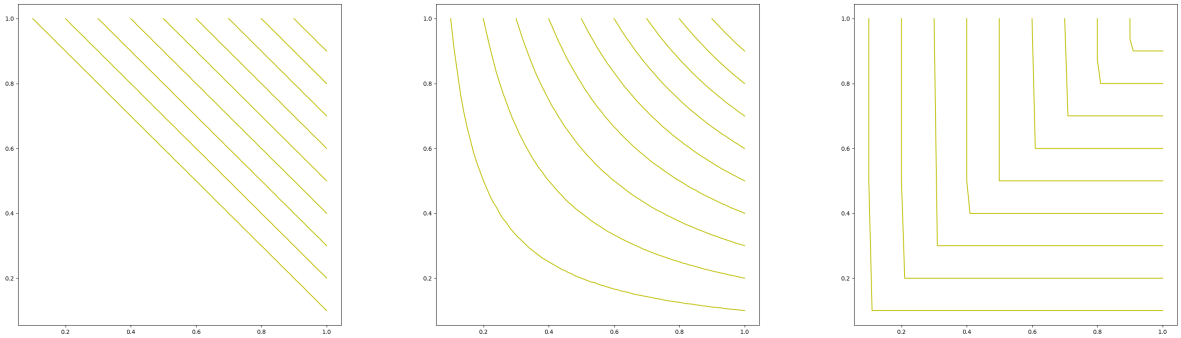
The three reference copulas are more graphically represented in Figure 79.



(a1) W copula 3D graph (a2) Π copula 3D graph (a3) M copula 3D graph



(b1) W copula scatterplot (b2) Π copula scatterplot (b3) M copula scatterplot



(b1) W copula level curves (b2) Π copula level curves (b3) M copula level curves

Figure 79 – Comparison of the reference copulas: W Copula, the product Π Copula and the M Copula. (a) shows the 3D graphs with the 0.1 to 0.9 probability level curves in grey. (b) shows the scatterplot of each curve. (c) presents the level curves itselves projected in the base plan.

An example of a commonly used n-copula is the Gaussian n-copula (ELIDAN, 2013) defined by

$$C_\Sigma(\{F_i(x_i)\}) = \Phi_\Sigma(\Phi^{-1}(F_1(x_1)), ..., \Phi^{-1}(F_1(x_n))) \tag{A.37}$$

where Φ is the standard normal distribution and Φ_Σ is a zero mean normal distribution

with correlation matrix Σ.

A.4.3 Inversion Copulas

A first group of copula families can be obtained from the direct inversion method: given a bivariate distribution function H with continuous margins F and G, the copula C is obtained from $C(u,v) = H(F^{(-1)}(u), G^{(-1)}(v))$.

A.4.3.1 Marshall-Olkin Copula Family

The Marshall-Olkin copula family, also known as the generalized Cuadras-Augé family, is a two-parameter family obtained from the bivariate exponential distribution with the same name and has the following equation

$$C_{\alpha,\beta}(u,v) = min(u^{1-\alpha}.v, u.v^{1-\beta}) = \begin{cases} u^{1-\alpha}.v, u^{\alpha} \geq v^{\beta} \\ u^{1-\alpha}.v, u^{\alpha} \geq v^{\beta} \end{cases} , 0 \leq \alpha, \beta \leq 1 \tag{A.38}$$

For $0 \leq \alpha, \beta \leq 1$ this family have full support but have both absolutely continuous and singular components with the mass of the singular component concentrated at the curve $u^{\alpha} = v^{\beta}$.

A.4.3.2 Circular Uniform Copula

This is a single copula (not a family) constructed from the experiment of choosing at random a point on the unit circle and it is represented by:

$$C(u,v) = \begin{cases} M(u,v), \ |u-v| > \frac{1}{2} \\ W(u,v), \ |u+v-1| > \frac{1}{2} \\ \frac{u+v}{2} - \frac{1}{4}, \ otherwise \end{cases} \tag{A.39}$$

This copula is singular, because it has $\partial^2 C / \partial u \partial v = 0$ almost everywhere.

A.4.4 Geometric Copulas

There are some geometric methods for constructing copulas from its supports, from its sections or by fitting in parts of other copulas as pieces of a puzzle. We are going to focus in this text on those methods which may lead to approximations of a given copula of interest.

A.4.4.1 Ordinal Sums Copula

Definition: Let $J_i = [a_i, b_i]$ denote a partition of I (a collection of closed, non-overlapping, except for common endpoints, non-degenerate intervals whose union is I) and let C_i be a corresponding collection of copulas. Then the **ordinal sum of** C_i is the copula C given by:

$$C(u,v) = \begin{cases} a_i + (b_i - a_i).C_i(\frac{u-a_i}{b_i-a_i}, \frac{v-a_i}{b_i-a_i}),\ (u,v) \in J_i^2, \\ M(u,v),\ otherwise \end{cases} \tag{A.40}$$

A.4.4.2 Shuffles of a Reference Copula

Definition: Let C be a given copula (usually the M copula). Let n be a positive integer, J_i a finite partition of I into n closed subintervals, π a permutation on $S_n = 1, 2, ..., n$ and $\omega : S_n \implies -1, 1$ a function where -1 stands for flipping the strip $J_i X I$ and 1 for not flipping it. Then $C(n, J_i, \pi, \omega)$ is called a **shuffle** of C.

Theorem: For any $\epsilon > 0$, there exists a shuffle C_ϵ of M such that:

$$\sup_{u,v \in I} |C_\epsilon(u,v) - \Pi(u,v)| < \epsilon \tag{A.41}$$

Theorem: Let C be a copula and $(a,b) \in (0,1)^2$ with $C(a,b) = \theta$ and $max(0, a+b-1) \leq \theta \leq min(a,b)$. Then C has best-possible bounds represented by $C_L(u,v) \leq C(u,v) \leq C_U(u,v)$, where C_U and C_L are shuffles copulas of M given by:

$$C_U = M(4, [0,\theta], [\theta, a], [a, a+b-\theta], [a+b-\theta, 1], (1,3,2,4), 1) \tag{A.42}$$

$$C_L = M(4, [0, a-\theta], [a-\theta, a], [a, 1-b+\theta], [1-b+\theta, 1], (4,2,3,1), -1) \tag{A.43}$$

A.4.4.3 Convex Sum Copula

Theorem: Let C_θ be a finite collection of copulas. Then any convex linear combination $C = \sum \lambda_\theta C_\theta,\ lambda_\theta \geq 0,\ \sum \lambda_\theta = 1$ of the copulas in C_θ is also a copula.

Theorem: Let C_θ be a collection of copulas based on a continuous parameter θ. Let the value of θ be the resulting observation from a continuous random variable Θ with distribution function Λ. Then the function C defined as below is a copula

$$C(u,v) = \int_R C_\theta(u,v) d\Lambda(\theta) \tag{A.44}$$

A.4.4.4 Horizontal and Vertical Sections Copulas

First thing to notice is that horizontal and vertical sections of a copula are proportional to conditional distribution functions:

$$C(u_0, v)/u_0 = P[V \leq v | U \leq u_0] \tag{A.45}$$

Second, the only copula with both horizontal and vertical linear section is the Π copula.

Theorem: C is a copula with quadratic section in u and it has the form $C(u, v) = u.v + \psi(v).u.(1-u)$ if, and only if, ψ is absolutely continuous on I, $|\psi^{'}(v)| \leq 1$ almost everywhere on I and $|\psi(v)| \leq min(v, 1-v)$ for all v in I. In this case C is absolutely continuous.

Definition: The **Farlie-Gumbel-Morgenstern (FGM)** family of copulas is composed by all copulas with quadratic sections on both u and v, and they happen also to be symmetric. The FGM familiy has the analytic form

$$C_\theta(u, v) = u.v + \theta.u.v.(1-u).(1-v),\ \theta \in [-1, 1] \tag{A.46}$$

Theorem: C is a copula with cubic section in u and it has the form $C(u, v) = u.v + [\alpha(v).(1-u) + \beta(v).u].u.(1-u)$ if, and only if, α and β are absolutely continuous on I, and $1 + \alpha^{'}(v).(1 - 4u + 3u^2) + \beta^{'}(v).(2u - 3u^2)$ for all u in I and almost all v in I. In this case C is absolutely continuous.

Theorem: C is a copula with both cubic sections in u and v, then it has the form $C(u, v) = u.v + u.v.(1-u).(1-v).[A_1 v.(1-u) + A_2(1-v).(1-u) + B_1 u.v + B_1 u.(1-v)]$, where A_1,A_2,B_1, B_2 are real constants such that the points (A_2, A_1),(B_1, B_2),(B_1, A_1),(A_2, B_2) all lie in $S = ([-1, 2]x[-2, 1]) \cup \epsilon$, with ϵ the set of points in and on the ellipse with equation $x^2 - xy + y^2 - 3x + 3y = 0$.

A.4.4.5 Diagonal Copulas

Definition: A **diagonal** is a function $\delta : I \to I$ with $\delta(1) = 1$, $\delta(t) \leq t \forall t \in I$, and $0 \leq \delta(t_2 - \delta(t_1)) \leq 2.(t_2 - t_1),\ \forall t_1, t_2 \in I$ and $t_1 \leq t_2$.

Theorem: Let δ be any diagonal. Then $C(u, v) = min(u, v, [\delta(u) + \delta(v)]/2)$ is a copula with δ for diagonal section ($C(t, t) = \delta(t)$). Those copulas are called **diagonal copulas**.

A.4.5 Algebraic Constructed Copulas

A.4.5.1 Placket Copulas

Definition: The **Plancket** family of copulas is the parametric family given by

$$C_\theta(u,v) = \begin{cases} \frac{[1+(\theta-1)(u+v)]-\sqrt{[1+(\theta-1)(u+v)]^2-4uv\theta(\theta-1)}}{2(\theta-1)},\ \theta \neq 1 \\ u.v,\ \theta = 1 \end{cases} \tag{A.47}$$

A.4.5.2 Ali-Mikhail-Haq Copulas

Definition: The **Ali-Mikhail-Haq** family of copulas is the parametric family given by

$$C_\theta(u,v) = \frac{u.v}{1-\theta(1-u)(1-v)},\ \theta \in [-1,1] \tag{A.48}$$

A.4.6 Transformation Constructed Copulas

Theorem: If C is a copula and n a positive integer, then the function $C_{(n)}$ as defined by the following equation is a copula and represents the copula associated to $X_{(n)} = maxX_i$ and $Y_{(n)} = maxY_i$, where X_i and Y_i are independent and identically distributed pairs of random variables with copula C.

Theorem: Let C be an arbitrary copula and $\gamma : [0,1] \leftarrow [0,1]$ a continuous and strictly increasing function with $\gamma(0) = 0$ and $\gamma(1) = 1$, with γ^{-1} its inverse. Then, $C_\gamma(u,v) = \gamma^{-1}(C(\gamma(u),\gamma(v)))$, for $u,v \in [0,1]$ is a copula if and only if γ is concave or, equivalently, γ^{-1} is convex.

Definition: A copula C_* is an **extreme value copula** if there exists a copula C, said to be in the domain of attraction of C_* such that

$$C_*(u,v) = \lim_{n \to \inf} C^n(u^{1/n}, v^{1/n}) \tag{A.49}$$

A.4.7 Archimedian Copulas

Definition: Let φ be a continuous, strictly decreasing convex function from I to $[0,\infty]$ such that $\varphi(1) = 0$, and let $\varphi^[-1]$ be the pseudo-inverse of φ, defined as $\varphi^{[-1]} : [0,\infty] \leftarrow I$ given by

$$\varphi^{[-1]}(t) = \begin{cases} \varphi^{[-1]}(t),\ 0 \leq t/leq\varphi(0)0,\ \varphi(0) \leq t/leq\infty \end{cases} \tag{A.50}$$

Then the function $C : I^2 \leftarrow I$ given by the following equation is a copula called **Archimedian** and the function φ is called its generator.

$$C(u,v) = \varphi^{[-1]}(\varphi(u) + \varphi(v)) \tag{A.51}$$

A.4.7.1 One-parameter Archimedian Copulas

Definition: **Clayton** copula family

$$C_\theta(u,v) = [max(0, u^{-\theta} + v^{-\theta} - 1)]^{-1/\theta},\ \varphi_\theta(t) = \frac{1}{\theta}(t^{-\theta} - 1),\ \theta \in [-1, \infty)\backslash 0 \tag{A.52}$$

Definition: **Ali-Mikhail-Haq** copula family (also equation A.48)

$$C_\theta(u,v) = \frac{u.v}{1 - \theta(1-u)(1-v)},\ \varphi_\theta(t) = ln\frac{1 - \theta(1-t)}{t},\ \theta \in [-1, 1] \tag{A.53}$$

Definition: **Gumbel-Hougaard** or simply **Gumbel** copula family

$$C_\theta(u,v) = exp(-[(\ln u)^\theta + (\ln v)^\theta]^{1/\theta}),\ \varphi_\theta(t) = (-\ln t)^\theta,\ \theta \in [1, \infty) \tag{A.54}$$

Definition: **Frank** copula family

$$C_\theta(u,v) = -\frac{1}{\theta}\ln(1 + \frac{(e^{-\theta.u} - 1)(e^{-\theta.v} - 1)}{e^{-\theta} - 1)},\ \varphi_\theta(t) = -\ln\frac{e^{-\theta.t} - 1}{e^{-\theta} - 1},\ \theta \in (-\infty, \infty)\backslash 0 \tag{A.55}$$

Definition: **Frank-Joe** copula family

$$C_\theta(u,v) = 1 - [(1-u)^\theta + (1-v)^\theta - (1-u)^\theta(1-v)^\theta]^{1/\theta},\ \varphi_\theta(t) = -\ln[1-(1-t)^\theta],\ \theta \in [1, \infty) \tag{A.56}$$

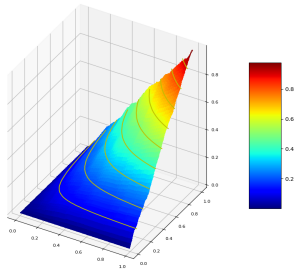
Definition: **Gumbel-Barnett** copula family

$$C_\theta(u,v) = u.v.\exp(-\theta \ln u \ln v),\ \varphi_\theta(t) = \ln(1 - \theta \ln t),\ \theta \in (0, 1] \tag{A.57}$$

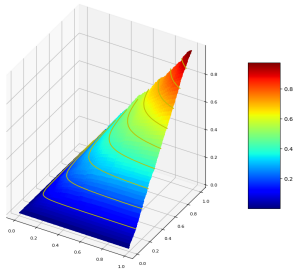
Definition: **Genest-Ghoudi** copula family

$$C_\theta(u,v) = max(0, 1 - [(1 - u^{1/\theta})^\theta + (1 - v^{1/\theta})^\theta]^{1/\theta})^\theta,\ \varphi_\theta(t) = (1 - t^{1/\theta})^\theta,\ \theta \in [1, \infty) \tag{A.58}$$

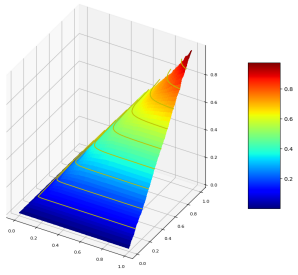
Some of those copula graphics are presented in the next figures for different values of τ for comparison on types and degrees of concordance.



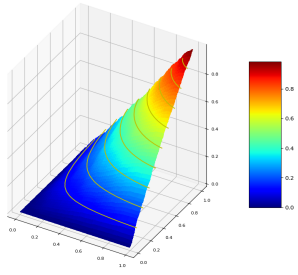
(a1) Clayton τ 0.2 copula



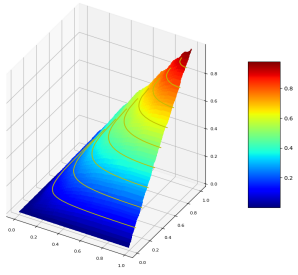
(a2) Clayton τ 0.5 copula



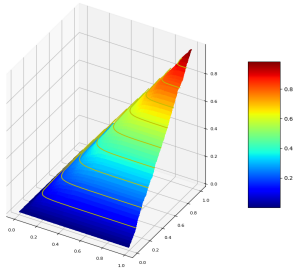
(a3) Clayton τ 0.8 copula



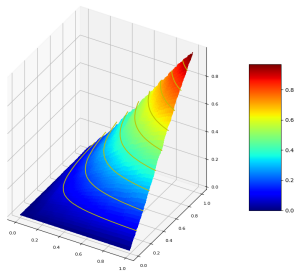
(b1) Gumbel τ 0.2 copula



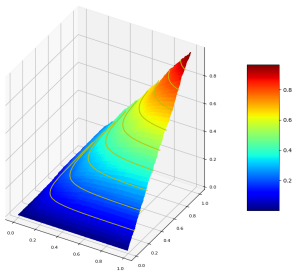
(b2) Gumbel τ 0.5 copula



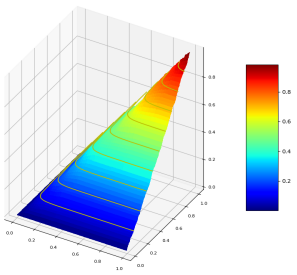
(b3) Gumbel τ 0.8 copula



(c1) Frank τ 0.2 copula

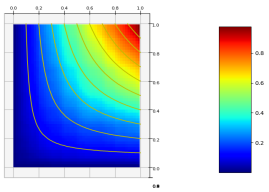


(c2) Frank τ 0.5 copula

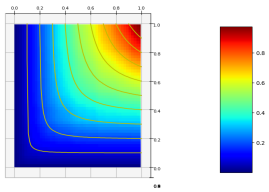


(c3) Frank τ 0.8 copula

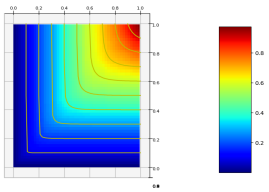
Figure 80 – Comparison of the 3D graphs of archimedian copulas: (a) the Clayton Copula, (b) the Gumbel Copula and (c) the Frank Copula, progressively with a τ of 0.2, 0.5 and 0.8, from left to right. (a) shows the 3D graphs with the 0.1 to 0.9 probability level curves in grey. (b) shows the scatterplot of each curve. (c) presents the level curves itselves projected in the base plan. (MATHWORKS, 2013)



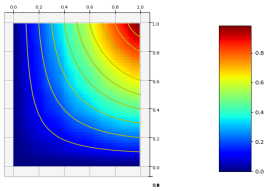
(a1) Clayton τ 0.2 copula



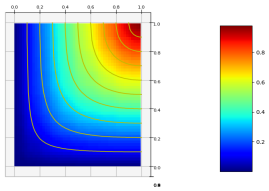
(a2) Clayton τ 0.5 copula



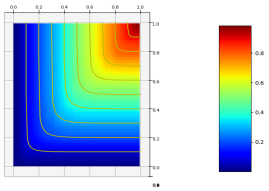
(a3) Clayton τ 0.8 copula



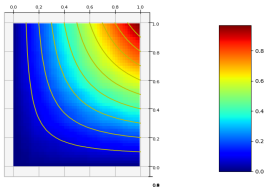
(b1) Gumbel τ 0.2 copula



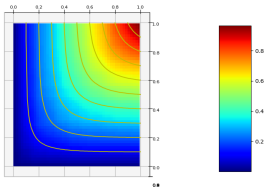
(b2) Gumbel τ 0.5 copula



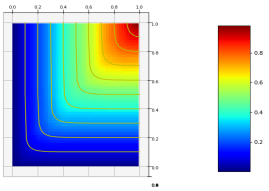
(b3) Gumbel τ 0.8 copula



(c1) Frank τ 0.2 copula

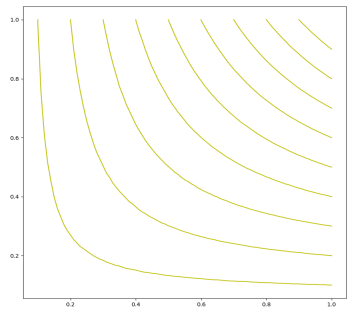


(c2) Frank τ 0.5 copula

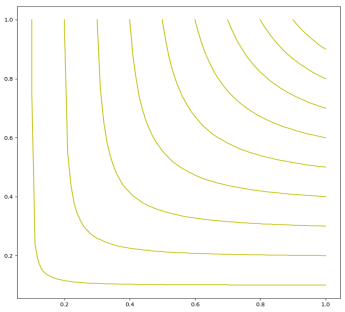


(c3) Frank τ 0.8 copula

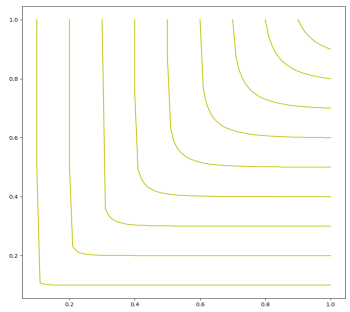
Figure 81 – Comparison of the 3D level curves projection of archimedian copulas: (a) the Clayton Copula, (b) the Gumbel Copula and (c) the Frank Copula, progressively with a τ of 0.2, 0.5 and 0.8, from left to right.



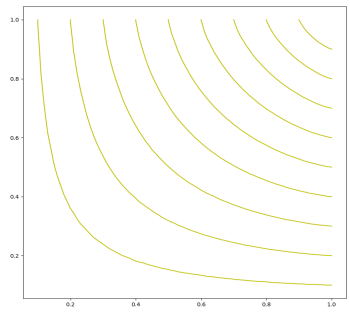
(a1) Clayton τ 0.2 copula



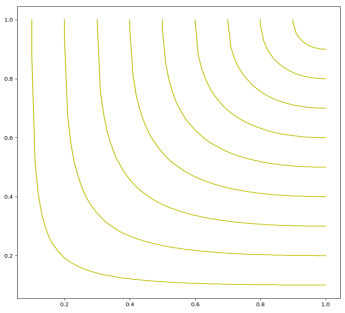
(a2) Clayton τ 0.5 copula



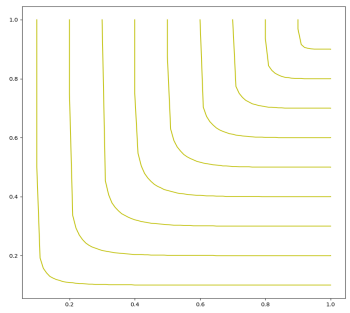
(a3) Clayton τ 0.8 copula



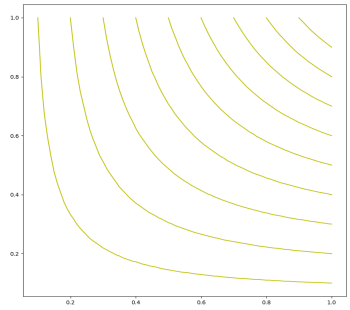
(b1) Gumbel τ 0.2 copula



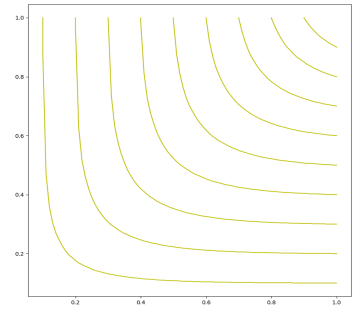
(b2) Gumbel τ 0.5 copula



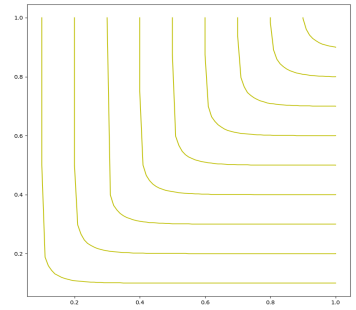
(b3) Gumbel τ 0.8 copula



(c1) Frank τ 0.2 copula

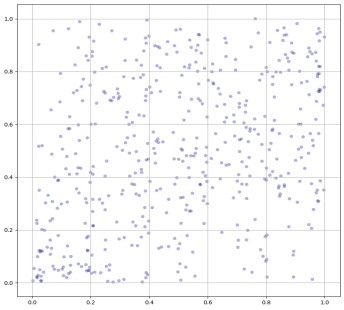


(c2) Frank τ 0.5 copula

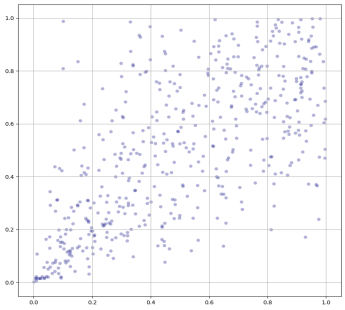


(c3) Frank τ 0.8 copula

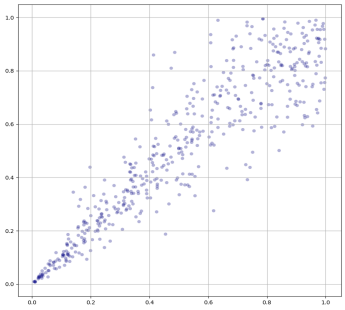
Figure 82 – Comparison of the curve levels of archimedian copulas: (a) the Clayton Copula, (b) the Gumbel Copula and (c) the Frank Copula, progressively with a τ of 0.2, 0.5 and 0.8, from left to right.



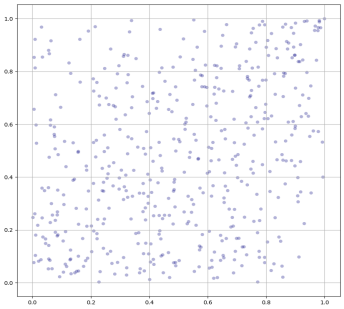
(a1) Clayton τ 0.2 copula



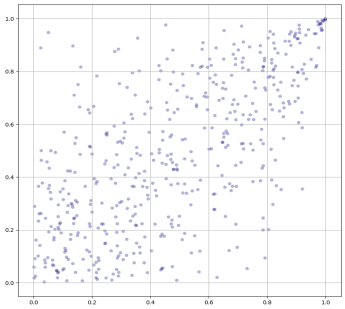
(a2) Clayton τ 0.5 copula



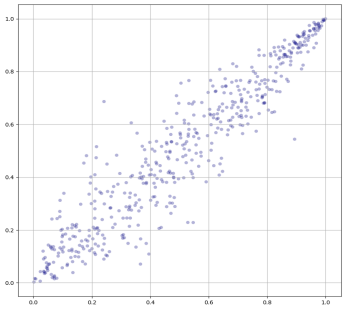
(a3) Clayton τ 0.8 copula



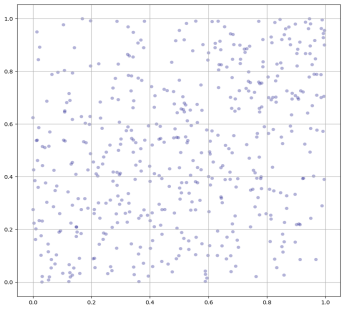
(b1) Gumbel τ 0.2 copula



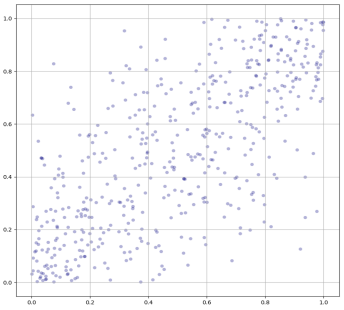
(b2) Gumbel τ 0.5 copula



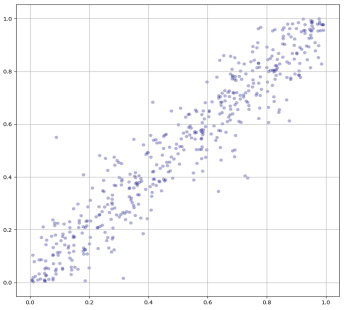
(b3) Gumbel τ 0.8 copula



(c1) Frank τ 0.2 copula



(c2) Frank τ 0.5 copula



(c3) Frank τ 0.8 copula

Figure 83 – Comparison of the scatterplots of samples with 1,000 units of archimedian copulas: (a) the Clayton Copula, (b) the Gumbel Copula and (c) the Frank Copula, progressively with a τ of 0.2, 0.5 and 0.8, from left to right.

Other one-parameter Archimedian copula families:

$$C_\theta(u,v) = max(0, 1 - [(1-u)^\theta + (1-v)^\theta]^{1/\theta}),\ \varphi_\theta(t) = (1-t)^\theta,\ \theta \in [-1, \infty) \quad \text{(A.59)}$$

$$C_\theta(u,v) = max(0, \theta.u.v + (1-\theta).(u+v-1),\ \varphi_\theta(t) = -\ln[\theta.t + (1-\theta)],\ \theta \in (0,1] \quad \text{(A.60)}$$

$$C_\theta(u,v) = max(0, \frac{\theta^2 uv - (1-u)(1-v)}{\theta^2 - (\theta-1)^2(1-u)(1-v)},\ \varphi_\theta(t) = \frac{1-t}{1+(\theta-1)t},\ \theta \in [1,\infty) \quad \text{(A.61)}$$

$$C_\theta(u,v) = \frac{uv}{[1+(1-u^\theta)(1-v^\theta)]^{1/\theta}},\ \varphi_\theta(t) = \ln(2t^{-\theta}-1),\ \theta \in (0,1] \quad \text{(A.62)}$$

$$C_\theta(u,v) = [max(0, u^\theta v^\theta - 2(1-u^\theta)(1-v^\theta))]^{1/\theta},\ \varphi_\theta(t) = \ln(2-t^\theta),\ \theta \in (0,1/2] \quad \text{(A.63)}$$

$$C_\theta(u,v) = (1+[(u^{-1}-1)+(v^{-1}-1)]^{1/\theta})^{-1},\ \varphi_\theta(t) = (\frac{1}{t}-1)^\theta,\ \theta \in [1,\infty) \quad \text{(A.64)}$$

$$C_\theta(u,v) = \exp(1-[(1-\ln u)^\theta + (1-\ln v)^\theta - 1]^{1/\theta}),\ \varphi_\theta(t) = (1-\ln t)^\theta - 1,\ \theta \in (0,\infty) \quad \text{(A.65)}$$

$$C_\theta(u,v) = (1+[(u^{-1/\theta}-1)^\theta (v^{-1/\theta}-1)^\theta]^{1/\theta})^{-\theta},\ \varphi_\theta(t) = (t^{-1/\theta}-1)^\theta,\ \theta \in [1,\infty) \quad \text{(A.66)}$$

$$\begin{aligned} C_\theta(u,v) =& \frac{1}{2}(S + \sqrt{S^2+4\theta}),\ S = u+v-1-\theta(\frac{1}{u}+\frac{1}{v}-1), \\ & \varphi_\theta(t) = (\frac{\theta}{t}+1)(1-t),\ \theta \in [0,\infty) \end{aligned} \quad \text{(A.67)}$$

$$\begin{aligned} C_\theta(u,v) =& (1+\frac{[(1+u)^{-\theta}-1][(1+v)^{-\theta}-1]}{2^{-\theta}-1})^{-1/\theta} - 1, \\ & \varphi_\theta(t) = -\ln\frac{(1+t)^{-\theta}-1}{2^{-\theta}-1},\ \theta \in (-\infty,\infty)\backslash 0 \end{aligned} \quad \text{(A.68)}$$

$$C_\theta(u,v) = max(0, 1+\frac{\theta}{\ln[e^{\theta/(u-1)}+e^{\theta/(v-1)}]}),\ \varphi_\theta(t) = e^{\theta/(t-1)},\ \theta \in [2,\infty) \quad \text{(A.69)}$$

$$C_\theta(u,v) = \frac{\theta}{\ln(e^{\theta/u}+e^{\theta/v}-e^\theta)},\ \varphi_\theta(t) = e^{\theta/t}-e^\theta,\ \theta \in (0,\infty) \quad \text{(A.70)}$$

$$C_\theta(u,v) = [\ln(\exp(u^{-\theta})+\exp(v^{-\theta})-e)]^{-1/\theta},\ \varphi_\theta(t) = \exp(t^{-\theta}-e),\ \theta \in (0,\infty) \quad \text{(A.71)}$$

$$\begin{aligned} C_\theta(u,v) =& 1-(1-max(0,[1-(1-u)^\theta]^{1/\theta}+[1-(1-v)^\theta]^{1/\theta}-1)^\theta)^{1/\theta}, \\ & \varphi_\theta(t) = 1-[]1-(1-t)^\theta]^{1/\theta},\ \theta \in [1,\infty) \end{aligned} \quad \text{(A.72)}$$

$$\begin{aligned} C_\theta(u,v) =& max(0,[1-(1-u^\theta)\sqrt{1-(1-v^\theta)^2}-(1-v^\theta)\sqrt{1-(1-u^\theta)^2}]^{1/\theta}), \\ & \varphi_\theta(t) = \arcsin(1-t^\theta),\ \theta \in (0,1] \end{aligned} \quad \text{(A.73)}$$

A.4.7.2 Two-parameter Archimedian Copulas

Theorem: Let φ be a function which is a generator for a given copula family, so φ is a continuous, strictly decreasing convex function on $[0, 1]$. Then this same function can be applied to generate a two-parameter copula family by using as generators the composites given by:

$$\varphi_{\alpha,\beta}(t) = [\varphi(t^{\alpha})]^{\beta},\ \alpha in(0, 1],\ \beta \in [1, \infty) \tag{A.74}$$

If φ is twice differentiable and $t.\varphi^{'}$ is nondecreasing on $(0, 1)$, this result extends to $\alpha \in (0, \infty)$.

Theorem: The function $C_{\alpha,\beta}$ defined on I^2 by

$$C_{\alpha,\beta}(u, v) = max(0, \frac{u.v - \beta(1-u).(1-v)}{1 - \alpha(1-u).(1-v)}) \tag{A.75}$$

is a copula, called **rational Archimedian copula**, if and only if $0 \leq \beta \leq 1 - |\alpha|$.

A.4.8 Empirical Copula

As copula traditional parametric approaches have restrictions that would not allow the generality intended for the model, such as forcing similar correlation orders midst all variables as a reflex of the used function family characteristics, the more general concept of empirical copula (Nelsen) was adopted, which is defined for the bivariate case and a sample of size n as the function C_n given by:

$$C_n(\frac{i}{n}, \frac{j}{n}) = \frac{\#\{(x, y) \in sample | x \leq x_{(i)}, y \leq y_{(j)}\}}{n}, \tag{A.76}$$

where $x_{(i)}, y_{(j)}$ denote order statistics from the sample.

This can be generalized to the multivariate case (STRELEN, 2009) as:

$$C^n(u_1, .., u_d) = \frac{1}{n}.\sum_{i=1}^{n} \mathbf{1}(\tilde{U}_1^i \leq u_1, .., \tilde{U}_d^i \leq u_d), \tag{A.77}$$

where each $\tilde{U}_j^i$ are the pseudo copula observations defined by:

$$(\tilde{U}_1^i, ..., \tilde{U}_d^i) = (F_1(X_1^i), ..., F_n(X_n^i)) \tag{A.78}$$

Again, just as in the case of margin fitting, it is worth remarking that the empirical approach has its limitations in terms of performance, but it was still valid enough for keeping the modeling simple. allowing to focus on the complete modeling instead of in

its precision. Further development in the research line implies trying any other copula modeling which can improve the modeling, including graphical approaches, machine learning structures or even parametric copula families fitting if viable.

A.4.9 Copula and Concordance Measures

Copulas and concordance measures are related, as we shall see.

Definition: A numeric measure κ of association between two continuous random variables X and Y whose copula is C is a **measure of concordance** if it satisfies the following properties (NELSEN, 2006):

1. κ is defined for every pair X, Y of continuous random variables;
2. $-1 \leq \kappa_{X,Y} \leq 1, \kappa_{X,X} = 1, \kappa_{X,-X} = -1$
3. $\kappa_{X,Y} = \kappa_{Y,X}$
4. if X, Y are independent, then $\kappa_{X,Y} = \kappa_{\Pi} = 0$
5. $\kappa_{-X,Y} = \kappa_{X,-Y} = -\kappa_{X,Y}$
6. if C_1 and C_2 are copulas such that $C_1 \prec C_2$, then $\kappa_{C_1} \leq \kappa_{C_2}$
7. if (X_n, Y_n) is a sequence of continuous random variables with copulas C_n, and if C_n converges pointwise to C, then $\lim_{n \to \inf} \kappa_{C_n} = \kappa_C$

Definition: A numeric measure δ of association between two continuous random variables X and Y whose copula is C is a **measure of dependence** if it satisfies the following properties (NELSEN, 2006):

1. δ is defined for every pair X, Y of continuous random variables;
2. $\delta X, Y = \delta Y, X$;
3. $0 \leq \delta_{X,Y} \leq 1$;
4. $\delta_{X,Y} = 0$ if and only if X and Y are independent;
5. $\delta_{X,Y} = 1$ if and only if each of X and Y is almost surely a strictly monotone function of the other;
6. if α and β are almost surely strictly monotone functions on $RanX$ and $RanY$, respectively, then $\delta_{\alpha(X),\beta(Y)} = \delta_{X,Y}$;

7. if (X_n, Y_n) is a sequence of continuous random variables with copulas C_n, and if C_n converges pointwise to C, then $\lim_{n\to\inf} \delta_{C_n} = \delta_C$

In the following sections we are going to present four indices that can be proven to be measures of concordance: Kendall's tau τ, Spearman's rho ρ, Gini's measure of association γ and Blomqvist medial correlation coefficient β. In the same sense, some measures of dependence will be presented, like Schweizer and Wolff's σ, the Hoeffding dependence index and, more generally, any L_p distance between C and Π.

A.4.9.1 Kendall's Tau

Definition: Let $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ denote a random sample of n observations from a vector (X, Y) of continuous random variables. Considering all of the $\binom{n}{2}$ sample pairs $(x_i, y_i), (x_j, y_j)$, let c be the number of concordant pairs and d the number of discordant pairs. Then the **sample Kendall's tau** is defined as

$$t = \frac{c - d}{\binom{n}{2}} \tag{A.79}$$

Definition: Let (X_1, Y_1) and (X_2, Y_2) be independent and identically distributed continuous random vectors with the same joint distribution function H. Then the **population Kendall's tau** is defined as

$$\tau = \tau_{X,Y} = P[(X_1 - X_2).(Y_1 - Y_2) > 0] - P[(X_1 - X_2).(Y_1 - Y_2) < 0] \tag{A.80}$$

Theorem: Let (X_1, Y_1) and (X_2, Y_2) be independent vectors of continuous random variables with different joint distribution functions H_1 and H_2 but with common margins F for X_1, X_2 and G for Y_1, Y_2. Let

$$Q = P[(X_1 - X_2).(Y_1 - Y_2) > 0] - P[(X_1 - X_2).(Y_1 - Y_2) < 0] \tag{A.81}$$

Let C_1, C_2 be the corresponding copulas to (X_1, Y_1) and (X_2, Y_2) as in the previous definition; then

$$Q = Q(C_1, C_2) = 4. \int\int_{I^2} C_2(u, v).dC_1(u, v) - 1, \ and \tag{A.82}$$

Theorem: Let (X, Y) be a vector of continuous random variables and C its copula. Then

$$\tau_{X,Y} = \tau_C = Q(C, C) \tag{A.83}$$

A.4.9.2 Spearman's Rho

Definition: Let (X_1, Y_1), (X_2, Y_2) and (X_3, Y_3) be three independent continuous random vectors with the same joint distribution function H and same margins F and G. Then the **population Spearman's rho** is defined as

$$\rho = \rho_{X,Y} = 3.(P[(X_1 - X_2).(Y_1 - Y_3) > 0] - P[(X_1 - X_2).(Y_1 - Y_3) < 0]) \quad (A.84)$$

Just as Gini's measure, Blomqvist β also satisfies the properties for a **measure of concordance**.

Theorem: Let (X, Y) be a vector of continuous random variables and C its copula. Then

$$\rho_{X,Y} = \rho_C = 3.Q(C, \Pi) \quad (A.85)$$

Theorem: Let (X, Y) be a vector of continuous random variables. Then

$$-1 \leq 3\tau - 2\rho \leq 1 \begin{cases} \frac{3\tau-1}{2} \leq \rho \leq \frac{1+2\tau-\tau^2}{2}, \ \tau \geq 0 \\ \frac{\tau^2+2\tau-1}{2} \leq \rho \leq \frac{3\tau+1}{2}, \ \tau \leq 0 \end{cases} \quad (A.86)$$

A.4.9.3 Gini's Measure of Association

Definition: Let p_i and q_i denote the ranks in a sample of size n of two continuous random variables X and Y. The **sample Gini's measure of association** is defined by

$$g = \frac{1}{int(n^2/2)}.[\sum_{i=1}^{n} |p_i + q_i - n - 1| - \sum_{i=1}^{n} |p_i - q_i|] \quad (A.87)$$

Definition: Let (X, Y) be a vector of continuous random variables and C its copula. Then the **population Gini's measure of association** is given by

$$\gamma = 2. \int \int_{I^2} (|u + v - 1| - |u - v|).dC(u, v) \quad (A.88)$$

Theorem: Let (X, Y) be a vector of continuous random variables and C its copula. Then

$$\gamma_{X,Y} = \gamma_C = Q(C, M) + Q(C, W) = 4.[\int_0^1 C(u, 1-u)du - \int_0^1 [u - C(u, u)]du] \quad (A.89)$$

A.4.9.4 Blomqvist's Medial Correlation Coefficient

Definition: Let (X, Y) be a vector of continuous random variables. Then the **population Blomqvist's medial correlation coefficient** is given by

$$\beta = \beta_{X,Y} = P[(X - \tilde{x}).(Y - \tilde{y}) > 0] - P[(X - \tilde{x}).(Y - \tilde{y}) < 0] \quad (A.90)$$

Theorem: Let (X, Y) be a vector of continuous random variables and C its copula. Then

$$\beta = \beta_C = 4.C(\frac{1}{2}, \frac{1}{2}) - 1 \quad (A.91)$$

A.4.9.5 Quadrant Dependence

Definition: Let X and Y be random variables. X and Y are **positively quadrant dependent (PQD)** if

$$\forall (x,y) \in \Re^2, P[X \leq x, Y \leq y] \geq P[X \leq x].P[Y \leq y] or equivalently, C(u,v) \geq u.v,\ (u,v) \in I^2 \tag{A.92}$$

Analogously, X and Y are **negatively quadrant dependent (NQD)** if

$$\forall (x,y) \in \Re^2, P[X \leq x, Y \leq y] \leq P[X \leq x].P[Y \leq y] or equivalently, C(u,v) \leq u.v,\ (u,v) \in I^2 \tag{A.93}$$

Theorem: Let X and Y be continuous random variables with joint distribution H, margins F and G, respectively, and copula C. If X and Y are PQD, then

$$3.\tau_{X,Y} \geq \rho_{X,Y} \geq 0,\ \gamma_{X,Y} \geq 0,\ \beta_{X,Y} \geq 0 \tag{A.94}$$

Remark: Although PQD and NQD are global properties, they can also be used for local dependence profiling within specific regions.

A.4.9.6 Tail Monotonicity

Definition: Let X and Y be random variables. Then:

1. Y **is left tail decreasing in** X **(**$LTD(Y|X)$ **)** if
   $\forall y,\ P[Y \leq y | X \leq x]$ is a nonincreasing function of x

2. X **is left tail decreasing in** Y **(**$LTD(X|Y)$ **)** if
   $\forall x,\ P[X \leq x | Y \leq y]$ is a nonincreasing function of y

3. Y **is right tail increasing in** X **(**$RTI(Y|X)$ **)** if
   $\forall y,\ P[Y \leq y | X \leq x]$ is a nondecreasing function of x

4. X **is right tail increasing in** Y **(**$RTI(X|Y)$ **)** if
   $\forall x,\ P[X \leq x | Y \leq y]$ is a nondecreasing function of y

Theorem: Let X and Y be continuous random variables. If X and Y are LTD or RTI in any sense, then X and Y are PQD.

Theorem: Let X and Y be continuous random variables with copula C. Then

1. LTD$(Y|X)$ if and only if $\forall v$ in I, $C(u,v)/u$ is nonincreasing in u

2. LTD$(X|Y)$ if and only if $\forall u$ in I, $C(u,v)/v$ is nonincreasing in v

3. RTI$(Y|X)$ if and only if $\forall v$ in I, $[v - C(u,v)]/(1-u)$ is nondecreasing in u

4. RTI$(X|Y)$ if and only if $\forall u$ in I, $[v - C(u,v)]/(1-v)$ is nondecreasing in v

Theorem: Let X and Y be continuous random variables. If $LTD(Y|X)$ and $RTI(Y|X)$, then $\rho_{X,Y} \geq \tau_{X,Y} \geq 0$ (and similarly if $LTD(X|Y)$ and $RTI(X|Y)$)

A.4.9.7 General L_p Dependence Distances

(SCHWEIZER; WOLFF, 1981) observe that any L_p distance between the surfaces C and Π, as done by the following general equation, is a measure of dependence:

$$m_p = (k_p . \int \int_{I^2} |C(u,v) - u.v|^p du.dv)^{1/p}, 1 \leq p \leq \infty \tag{A.95}$$

with k_p a normalization constant to force $m_p = 1$ when $C = M$ or $C = W$.

A.4.9.8 Schweizer and Wolff's Dependence Index

Definition: Let X and Y be continuous random variables with copula C. Then, the **Schweizer and Wolff's dependence index** represents a measure based upon the L_1 distance between the graphs of C and Π (obtained by making $p = 1$ in the general L_p dependence distance equation) is defined by

$$\sigma_{X,Y} = \sigma_C = 12. \int \int_{I^2} |C(u,v) - u.v| du.dv \tag{A.96}$$

A.4.9.9 Hoeffding Dependence Index

Definition: Let X and Y be continuous random variables with copula C. Then, the **Hoeffding dependence index** represents a measure of dependence (obtained by making $p = 2$ in the general L_p dependence distance equation) is defined by

$$\Phi_{X,Y} = \Phi_C = (90. \int_{I^2} |C(u,v) - u.v| du.dv)^{1/2} \tag{A.97}$$

A.4.9.10 Tail Dependence

Definition: Let X and Y be continuous random variables with distribution functions F and G, respectively. Then, the **upper tail dependence parameter** λ_U and the **lower tail dependence parameter** λ_L is the limits, when it exists, defined, respectively, by the following equations

$$\lambda_U = \lim_{t \to 1^-} P[Y > G^{(-1)}(t) | X > F^{(-1)}(t)] \lambda_L = \lim_{t \to 0^+} P[Y > G^{(-1)}(t) | X > F^{(-1)}(t)] \tag{A.98}$$

Theorem: Let X and Y be continuous random variables with distribution functions F and G, respectively, with copula C and its diagonal δ_C. If the limits that define the tail dependence indexes exist, then

$$\lambda_U = 2 - \lim_{t \to 1^-} \frac{1 - C(t,t)}{1 - t} = 2 - \delta_C^{'}(1^-) \lambda_L = \lim_{t \to 0^+} \frac{C(t,t)}{t} = \delta_C^{'}()^+) \tag{A.99}$$

A.4.10 Copula Modeling

Based on the assumptions from Sklar's Theorem, the model construction consists in starting at a given dataset of samples from the system to be modeled to first determine the marginal distributions for each variable and then identifying the corresponding copula that compound with those functions gives a reasonable approximation of the joint distribution function of the former dataset.

A.5 Bayesian Networks

Another paramount theoretical element used in our research is the Bayesian network, which is a graphical approach to statistical modeling.

A Bayesian network is a model composed by a directed acyclic graph (DAG) $\mathcal{G}$ and a joint probability distribution P where $(\mathcal{G}, P)$ satisfies the Markov condition and so the joint distribution P can be decomposed in a product of conditional distributions defined by $\mathcal{G}$ as explained in the following paragraphs.

Definition 36 *(ELIDAN, 2013) A **Markov Network (MN)** is an undirected graphical model which uses an undirected graph $\mathcal{H}$ that encodes the independencies $I(\mathcal{H}) = \{(X_i \perp X \backslash \{X_i\} \cup Ne_i | Ne_i)\}$, where Ne_i are the neighbors of X_i in $\mathcal{H}$, which means that each node is independent of all others given its neighbors in $\mathcal{H}$, also known as the **Markov condition**.*

Theorem 9 (Hammersley-Clifford Theorem) *(ELIDAN, 2013) Let $\mathcal{C}$ be the set of cliques in $\mathcal{H}$, where a clique is a set of nodes such that each node is connected to all others*

in the set. For positive densities, if the independence statements encoded by $\mathcal{H}$ hold in $f_X(x)$, then ***the joint density decomposes according to the graph structure***

$$f_X(x) = \frac{1}{Z} \prod_{c \in \mathcal{C}} \phi_c(x_c) \tag{A.100}$$

where X_c is the set of nodes in the clique c, and $\phi_c : \Re^{|c|} \to \Re^+$ is any positive function over the values of these nodes. Z is a normalizing constant called the partition function. The converse composition theorem also holds.

Theorem 10 (Product induced by independence structure) *(ELIDAN, 2013) Let T be an undirected tree structured graph (i.e., a graph with no cycles) and let E denote the set of edges in T that connect two vertices. If the independencies $I(T)$ defined by T hold in $f_X(x)$, then*

$$f_X(x) = [\prod_i f_i(x_i)]. \prod_{(i,j) \in \mathcal{E}} \frac{f_{ij}(x_i, x_j)}{f_i(x_i).f_i(x_i)} \tag{A.101}$$

Theorem 11 (Copula Bivariate Decomposition) *(ELIDAN, 2013) Let T be an undirected tree structured graph and let E denote the set of edges in T that connect two vertices. If the independencies $I(T)$ defined by T hold in $f_X(x)$, then*

$$c_T(.) = \frac{f_X(x)}{\prod_i f_i(x_i)} = \prod_{(i,j) \in \mathcal{E}} \frac{f_{ij}(x_i, xj)}{f_i(x_i).f_i(x_i)} = \prod_{(i,j) \in \mathcal{E}} c_{ij}(F_i(X_i), F_j(X_j)) \tag{A.102}$$

where $c_T(.)$ is used to denote a copula density associated to the structure T and c_{ij} is used to denote the bivariate copula corresponding to the edge (i, j). The converse composition holds.

Definition 37 *(NEAPOLITAN, 2003) A* ***directed graph*** *is a pair (V, E), where V is a finite, nonempty set whose elements are called* ***nodes*** *(or vertices), and E is a set of ordered pairs of distinct elements of V whose elements are called* ***edges*** *(or arcs).*

Definition 38 *(NEAPOLITAN, 2003) A directed graph G is called a* ***directed acyclic graph (DAG)*** *if it contains no path from a node to itself (directed cycles).*

Definition 39 *(NEAPOLITAN, 2003) Given a DAG $\mathcal{G} = (V, E)$ and nodes X and Y in V, Y is called a* ***parent*** *of X if there is an edge from Y to X, Y is called a* ***descendent*** *of X and X is called an ancestor of Y if there is a path from X to Y.*

Definition 40 *(NEAPOLITAN, 2003) Suppose we have a joint probability distribution P of the random variables in some set V and a DAG $\mathcal{G} = (V, E)$. We say that $(\mathcal{G}, P)$ satisfies the* ***Markov condition*** *if for each variable $X \in V$, X is conditionally independent I_P of the set of all its nondescendents ND_X given the set of all its parents Pa_X,for which we adopt the notation $I_P(X, ND_X|Pa_X)$.*

Theorem 12 (Product of Conditional Distributions) *(NEAPOLITAN, 2003) If* $(\mathcal{G}, P)$ *satisfies the Markov condition, then* P *is equal to the* ***product of its conditional distributions*** *of all nodes given values of their parents, whenever these conditional distributions exist, that is*

$$P(x_n, x_{n-1}, ..., x_1) = P(x_n|pa_n).P(x_{n-1}|pa_{n-1})...P(x_1|pa_1),\ P(Pa_i) \neq 0, 1 \leq i \leq n \tag{A.103}$$

Theorem 13 (DAG Markov Condition) *(NEAPOLITAN, 2003) Let a DAG* $\mathcal{G}$ *be given in which each node is a random variable, and let a discrete conditional probability distribution of each node given values of its parents in* $\mathcal{G}$ *be specified. Then the product of these conditional distributions yields a joint probability distribution* P *of the variables, and* $(\mathcal{G}, P)$ *satisfies the Markov condition.*

Definition 41 *(NEAPOLITAN, 2003) Let* P *be a joint probability distribution of the random variables in some set* V*, and* $\mathcal{G} = (V, E)$ *be a DAG. We call* $(\mathcal{G}, P)$ *a* ***Bayesian network*** *if* $(\mathcal{G}, P)$ *satisfies the Markov condition. From Theorem 12,* P *is the product of its conditional distributions in* $\mathcal{G}$*, and this is the way* P *is always represented in a Bayesian network. Furthermore, from Theorem 13, if we specify a DAG* $\mathcal{G}$ *and any discrete conditional distributions (and many continuous ones), we obtain a Bayesian network. This is the way Bayesian networks are constructed in practice.*

Figure 84 shows an example of Bayesian network.

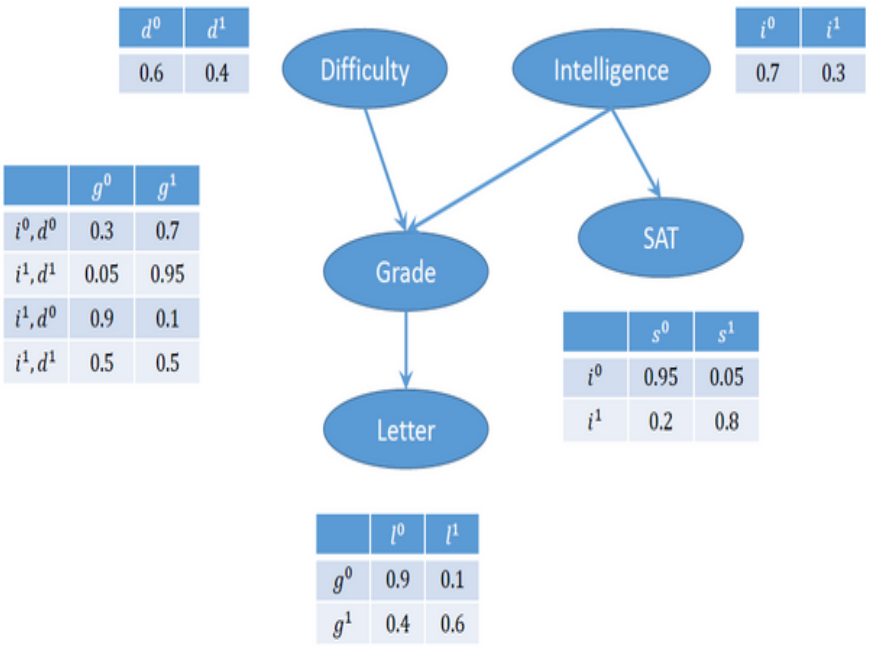


Figure 84 – Bayesian network example (GORDON *et al.*, 2014). It represents the probabilistic relations among variables in a student grading problem with the following features: student intelligence, discipline difficulty, discipline grade for that student, student SAT and a positive letter of recommendation for that student by the discipline teacher.

The contribution of this graphical decomposition is that estimation and learning are simplified by the compact representation, but at a trade-off of strong independence assumptions. Some further approaches to overcome those strong premises are (KIRSHNER, 2009) mixture of all copula trees model proposal, at the cost of some loss of flexibility by parameter sharing constraints, and (SILVA; GRAMACY, 2009) Bayesian approach of a mixture of some trees with flexible priors on all components of the model.

Theorem 14 (Product of conditional densities) *(ELIDAN, 2013) If the independences encoded by $\mathcal{G}$ hold in f_X, then*

$$f_X(x) = \prod_{i=1}^{n} f_{X_i|Pa_i}(x_i|pa_i) \tag{A.104}$$

and the converse composition theorem is valid, i.e., a product of any local conditional densities defines a valid joint density with the independences encoded by the DAG $\mathcal{G}$ associated to that product.

Although that decomposition and graphical representation simplifies the joint distribution modeling, it is still very far from a simple problem. Finding a Bayesian

network best structure is a NP-hard (non-deterministic polynomial-time hardness) problem, which can be taken as being super-exponentially time expensive for an exhaustive search algorithm. Figure 85 shows graphically the number of possible structures for Bayesian networks in terms of the number of variables. Observe that for 20 variables, this number is above 10^{70}.

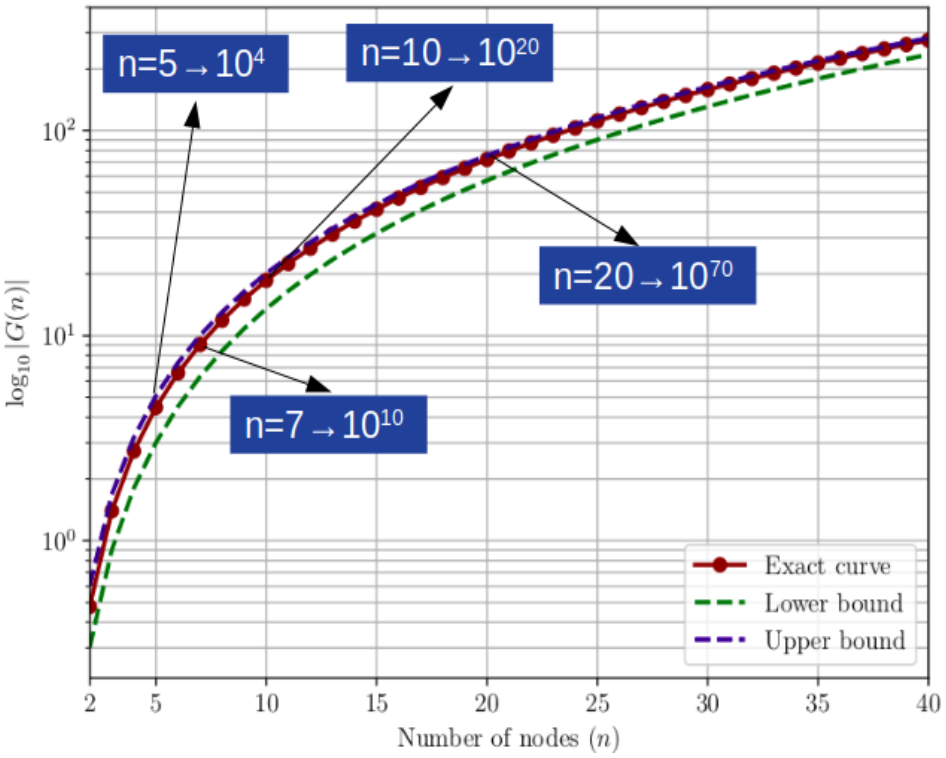


Figure 85 – Number of DAGs as the number of nodes increases according to Robinson's recurrence (ROBINSON, 1977). Graphics from Gross *et al.* (2019).

Therefore, one approach for Bayesian networks modeling is to treat it as a search for the optimal network structure and parameters among all possible structures for a given dataset. That problem is usually decomposed in a structure search followed by a parametric computation for the chosen structure, formulated as follows.

Definition 42 ***BN learning*** *is the optimization problem that, given a dataset D, find the BN $B = (\mathcal{G}, \Theta)$ that maximizes $P(B|D) = P(D|B).P(B) = P(D|\mathcal{G}, \Theta).P(\Theta|\mathcal{G}).P(\mathcal{G})$.*

Definition 43 ***Structure learning*** *is the part of BN learning focused on finding the network structure $\mathcal{G}$ that maximizes $P(\mathcal{G}|D)$*

- $P(\mathcal{G}|D) \propto P(D|\mathcal{G}).P(\mathcal{G})$
- $P(D|\mathcal{G}) = \int_{\Theta} P(D|\mathcal{G}, \Theta).P(\Theta|\mathcal{G}).d\Theta$

The optimization problem needs to be instrumented by a score function which associates to each possible structure a corresponding score measuring how good is that structure to represent the given dataset. In this research we choose a score, very used in literature, called Bayesian Dirichlet equivalence with uniform prior metric (BIELZA; NAGA, 2014) - BDeu as the referential score in our structure learning stages.

Definition 44 ***BDeu scoring*** *is a scoring measure for Bayesian network structures which assumes $P(\mathcal{G})$ to be a uniform distribution and $P(\Theta|\mathcal{G})$ to be a Dirichlet distribution resulting in the following equation (HEKERMAN; GEIGER; CHICKERING, 1995) for computing a network structure score for a given dataset:*

$$P(D|\mathcal{G}) = \prod_{i=1}^{n} \prod_{j=1}^{q_i} \frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij} + N_{ij})} . \prod_{k=1}^{r_i} \frac{\Gamma(\alpha_{ijk} + N_{ijk})}{\Gamma(\alpha_{ijk})} \tag{A.105}$$

where i stands for each structure node, j for each state of each node, and k for each node parents instance.

The CBN mentioned in Chapter 1 has the formal definition presented in Definifition 45, based on 2.

Lemma 2 (copula conditional density) *(ELIDAN, 2013) Let $f(x|\boldsymbol{y})$, with $\boldsymbol{y} = \{y_1, y_2, ..., y_n\}$, be a conditional density function. There exists a copula density function $c(F(x),$ $F_1(y_1), ..., F_k(y_k))$ such that*

$$f(x|\boldsymbol{y}) = R_c(F(x), F_1(y_1), ..., F_k(y_k)).f_X(x), \tag{A.106}$$

where R_c is the copula ratio

$$R_c(F(x), F_1(y_1), ..., F_k(y_k)) = \frac{c(F(x), F_1(y_1), ..., F_k(y_k))}{\frac{\partial^k C(1, F_1(y_1), ..., F_k(y_k))}{\partial F_1(y_1) ... \partial F_k(y_k)}} \tag{A.107}$$

and R_c is defined to be 1 when $\boldsymbol{Y} = \emptyset$. The converse is also true: for any copula, $R_c(F(x), F_1(y_1), ..., F_k(y_k)).f_X(x)$ defines a valid conditional density.

Definition 45 *(ELIDAN, 2013) A Copula Bayesian Network (CBN) is a triplet $\mathcal{C} = (\mathcal{I}, \Theta_C, \Theta_f)$ that defines $f_X(x)$. $\mathcal{I}$ encodes the independencies $\{(X_i \perp \boldsymbol{ND}_i|\boldsymbol{Pa}_i)\}$, assumed to hold in $f_X(x)$. Θ_C is a set of local copula functions $C_i(F(x_i).F(\boldsymbol{pa}_{i1})...F(\boldsymbol{pa}_{ik_i})$ that are associated with nodes of $\mathcal{I}$ that have at least one parent. In addition, Θ_f is the set of parameters representing the marginal densities $f_i(x_i)$ (and distributions $F_i(x_i)$). The joint density $f_X(x)$ then takes the form*

$$f_X(x) = \prod_{i=1}^{n} R_{c_i}(F(x_i).F(\boldsymbol{pa}_{i1})...F(\boldsymbol{pa}_{ik_i})).f_i(x_i). \tag{A.108}$$

Table 16 (ELIDAN, 2013) presents a summary of the different copula-based multivariate models and its application with the author's observations at the time (2013).

Table 16 – Summary of the different copula-based multivariate models extracted from (ELIDAN, 2013) with that author's observations.

Model	**Variables**	**Structure**	**Copula**	**Comments**
Vines	< 10 in practice	Conditional dependence	Any bivariate	well understood general purpose framework
Nonparametric BBN	100s 100s	vines vines	Gaussian in practice	mature application
Tree-averaged	10s	Mixture of trees	Any bivariate	requires only bivariate estimation
Nonparanormal	100-1000s	MN	Gaussian	high-dimensional estimation with theoretical guarantees
Copula networks	100s	BN	Any	flexible at the cost of partial control over marginals
Copula processes	∞ (replications)	-	Multivariate	Nonparametric generalization of Gaussian processes

APPENDIX B – DEVELOPED TOOLS

Visualization of each step in the process of modeling presented itself as an essential tool for granting good progress. For that matter, as many calibrations and trials were needed for feature fittings and copula modeling, it was soon realized that a graphical interactive interface which showed the modeling procedure step-by-step would be very helpful, leading to the idea of developing the LpsCopModel software, a graphical interface for the entire modeling process, starting from data set choosing and data acquisition, going through features marginal distribution fitting and finishing with the copula modeling itself for completion (only empirical copula in the current version).

LpsCopModel is developed over a 'django' (DJANGO SOFTWARE FOUNDATION, 2013) platform. Therefore the input for working with LpsCopModel is a directory containing one or more data files from a chosen phenomenon representing a matrix where each row stands for a sample and its features measures. Each file may be a CSV format data file in a rows/columns shape, a 'pandas' (THE PANDAS DEVELOPMENT TEAM, 2020) data frame saved in python/pickle format or a DBF pattern database as in Brazilian public healthcare data system (BRASIL, 2020). A few examples with small datasets are provided in directories placed under 'static/data/' in the server file tree structure with the directory name taken by LpsCopModel as the example name and included among embedded options, but any user dataset in CSV format can also be imported. After data acquisition/selection, users go sequentially through stages from slicing and filtering a set of interest to data analyzing and modeling by interactive and visualization software features.

Although the LpsCopModel software was used in our research for more specific purposes, it is expected to be considerably helpful in many other scientific researches which need to acquire simplified models from datasets before further analysis, especially whenever there is a main concern about dependence and concordance between features and variables. Material evidence of this software relevance are many other similar initiatives in copula modeling like Paprotny *et al.* (2020) and the Data to AI Lab at MIT projects "SDV" and "copulas" (PATKI; WEDGE; VEERAMACHANENI, 2016), recently released in its 0.3.3 version in Sep 18, 2020, each covering different aspects, while the ones emphasized here are visualization and panoramic analysis. Nevertheless, as the software has been built under an open-platform approach, it can be easily incremented for allowing originally not included parametric distributions or copula families.

B.1 Software Architecture

The LpsCopModel software covers the entire workflow from choosing among previously downloaded datasets to that data complete copula modeling (empirical copula, in this version). The process is split into five sequential stages: dataset selection and acquisition, data filtering and slicing, data description, marginal distributions fitting and copula modeling.

As this software proposal is to be a graphical interface for modeling, it is based on an architecture that combines an user-interface (web browser) and a high flexibility language (python). The chosen architecture resides on a Linux server running Django for controlling the interaction between a web interface, python script processing, and database access. The web interface is programmed basically in HTML/JavaScript with its usual tools (CSS, bootstrap, jQuery) and python was used along with many of its useful packages (django, numpy, scipy, pandas, pickle, pymc3 (SALVATIER; WIECKI; FONNESBECK, 2016), etc.).

The present version has been loaded with some previous datasets: a subset of Brazilian public healthcare system data and some simple canonical test data sets (ideally complete positive dependence, negative dependence and independence). For the healthcare data set, both DBF - which is the original format of publicized data - and pandas/pickle data files format were loaded and can be read by the software, while the test datasets are code generated. However, any data set can be loaded from CSV files or pandas data frame saved using pickle format files.

B.2 Software Functionalities

LpsCopModel is a web application to be deployed at an user web server service or it can be installed in local mode "django" for individual use. If in local mode, access is provided by opening any web browser and pointing to the "localhost" address including its port (usually 127.0.0.1:8000); for web server access, the web application address must be provided instead. The software home page is shown in Figure 86.

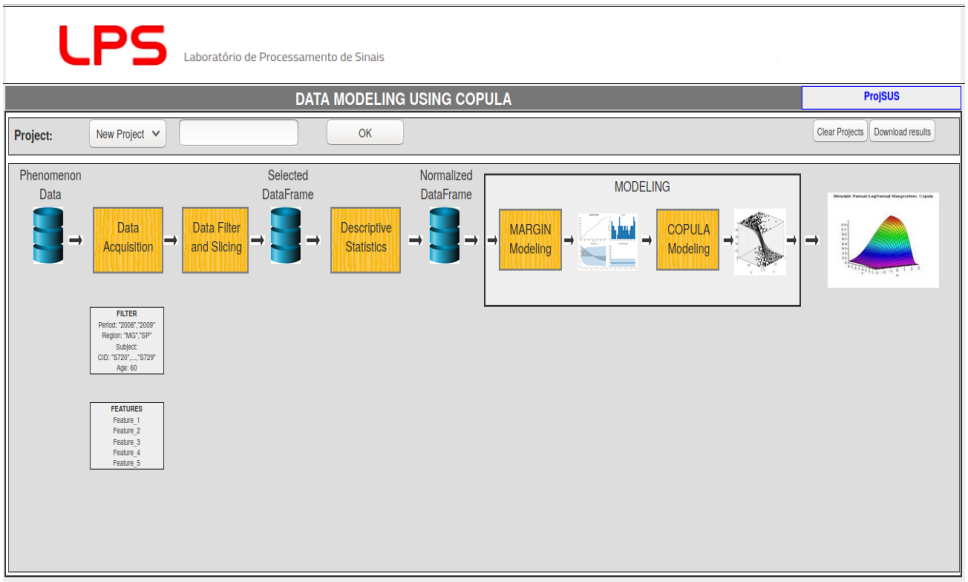


Figure 86 – LPSCopModel home page. A modeling methodology block diagram is presented for going sequentially from data acquisition to copula modeling. A simple project managing interface is provided in the top banner, where user can create, load and save projects.

LpsCopModel allows its users to define a project to save partial and complete analysis and this is accomplished by the user by creating a project or loading an existing one using the project banner at the home page top as a first step in operating the system.

Clicking on the corresponding block figure, users can proceed directly to any analysis stage in the software, provided that all previous analysis has proceeded or a corresponding project has been loaded. After completing any stage, users must return to the main page for the next stages by clicking on the corresponding block figure. At any stage, results can be downloaded in a python pickle format; also, any figure can be expanded in a pop-up window and saved.

For a fresh start, a dataset has to be chosen. If the user wants to analyze one of the previously installed data sets (DATASUS or test sets) that can be done by clicking on the "Data Acquisition" block and selecting the desired dataset in the drop-down button and then checking box values for each main feature by which data files are arranged. Any specific dataset in CSV format can be uploaded by the "generic" option in the dropbox (Figures 87 and 88).

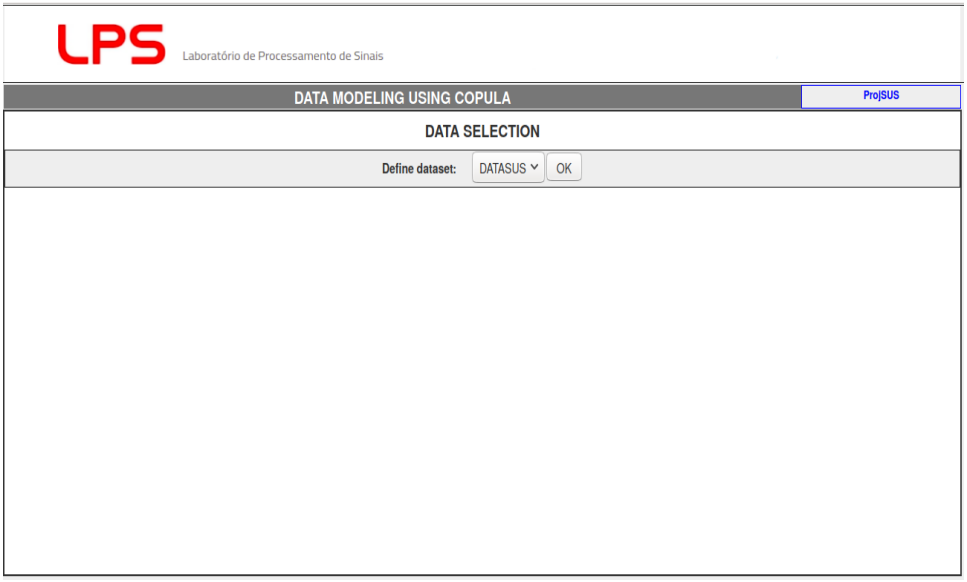


Figure 87 – Data set choosing page. User can choose among native datasets or insert a new one from CSV or other supported input files.

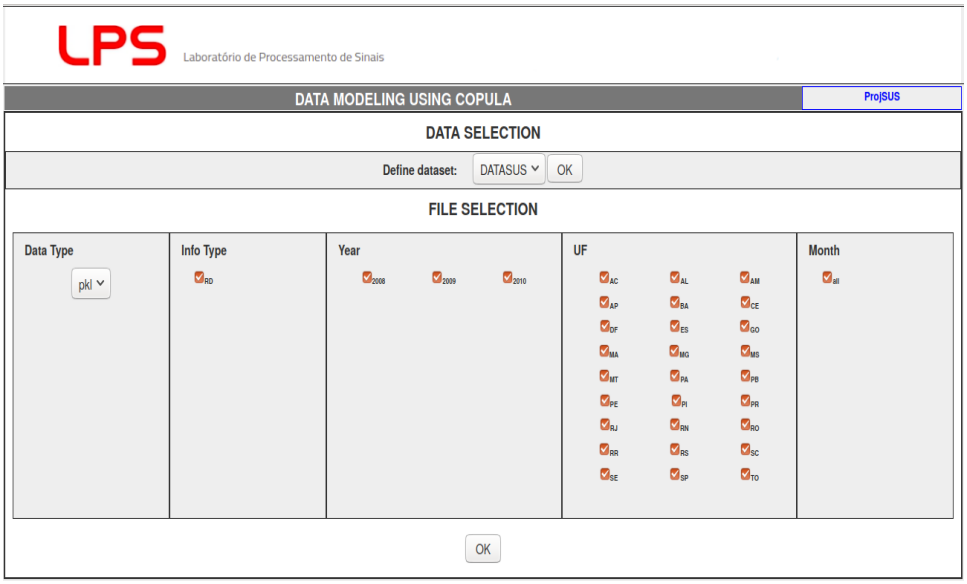


Figure 88 – Data set choosing and selecting page. Data type refers to file format input while the other sections allows data file selection by space, time and other features used to organize bigger data sets into separated files.

Next natural step after choosing a dataset, the filtering and slicing page allows users to select which data to work on. The page is divided vertically in three areas: original data description, filtering/slicing, and selected data visualization. Filtering and slicing are done in the middle section, where the upper region stands for filtering, and users can add any feature and respective values to be considered in filtering original data to a filter string to be sent to the system. Slicing selected data for including only predefined features

in the ongoing analysis occurs in the section below the filtering section by adding those features to a slicing string. For selections to be applied, the corresponding button at the bottom must be pressed. Figure 89 shows this stage screen.

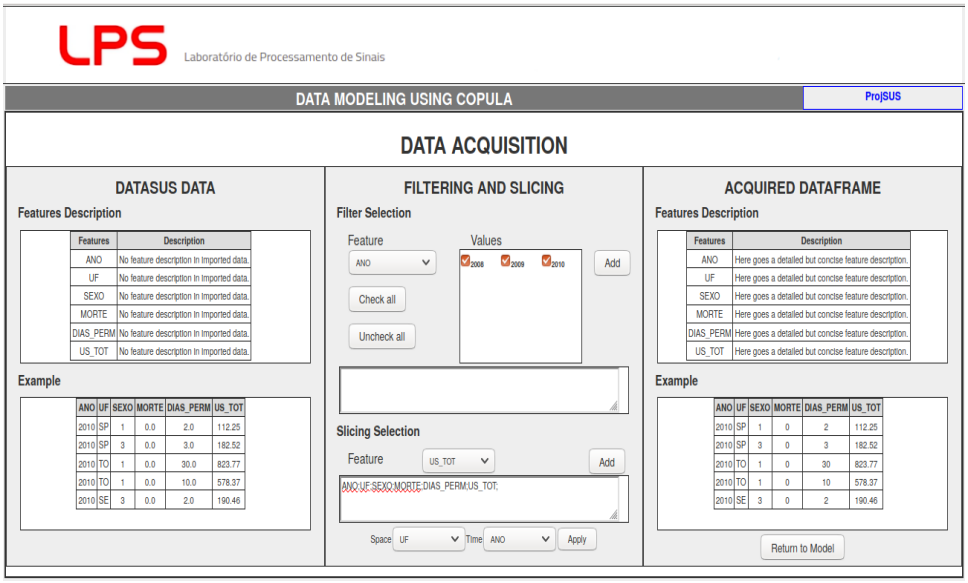


Figure 89 – Data filtering and slicing page.

Before any further modeling, a general descriptive statistical analysis is provided as a third stage in the system. This page is divided into a 2x3 matrix area with five feature descriptive table and figures plus one association description table. The feature to be described must be selected by users on a top frame. The first quadrant shows a simple table with main statistical measures over the chosen feature, depending on its nature; for numeric features, it shows numerical measures (count, mean, standard deviation, minimum, maximum, 25%, 50%, 75%), while for categorical features it shows categorical measures (count, unique, top and top frequency). Four figures are also displayed: a histogram, a box-plot (only for numeric), a box-plot timeline (only for numeric), and a spacial distribution (only for Brazilian "datasus"). The last square in that page is a table for showing associations among all features by two concordance indexes, Spearman's rho and Kendall's tau (NELSEN, 2006), which are presented in the same table by splitting it into upper-right (for tau) and bottom-left (for rho) parts in relation to its main diagonal. All that is shown in Figures 90 and 91.

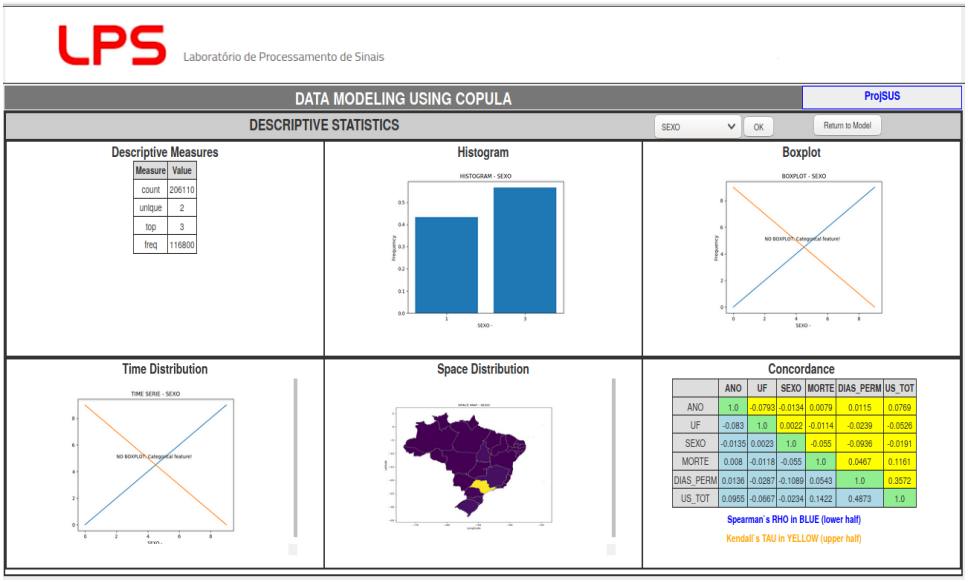


Figure 90 – Categorical feature descriptive example: hospital admissions by gender (1-male, 3-female). Measures (number of samples, gender with greater occurrences and its frequency), histogram, geographical distribution and concordance with other features. Box-plot based figures are not displayed for categorical features.

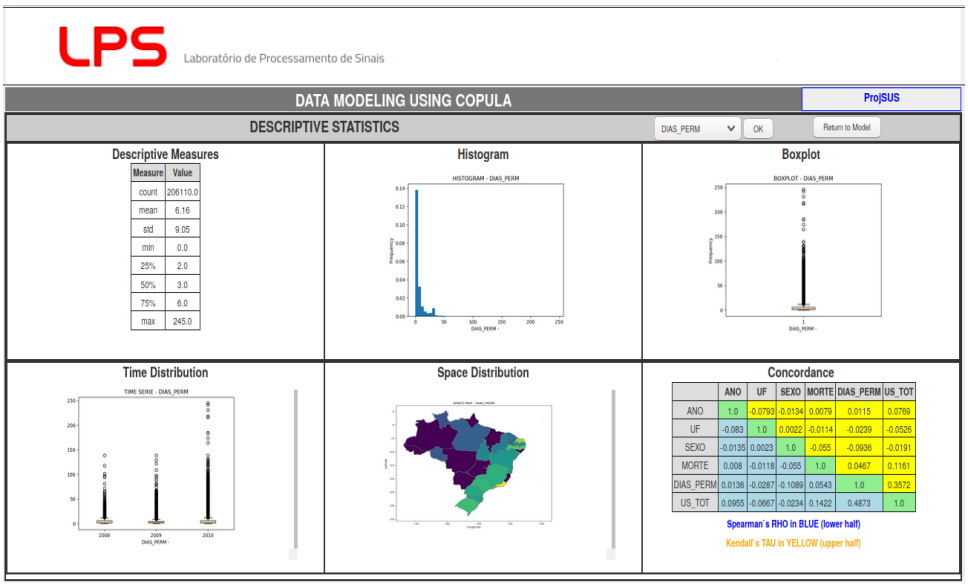


Figure 91 – Numeric feature descriptive example: days in hospital between admittance and discharge. Measures (number of samples, mean, standard deviation, minimum, maximum, quantiles), histogram, box-plots, box-plot time series, geographical distribution and concordance with other features.

A previous mandatory step for modeling copula is to model each feature marginal distribution. Here it is done by modeling categorical features by its frequency in a multinomial distribution, and numeric features by its probability density using Monte-Carlo

Markov Chain (MCMC) technique implemented in "pymc3" python package. Those so-acquired distributions are further used in modeling the copula. As a project choice, an initial set of parametric distributions were coded, but there is no restriction for further improvement including any other distribution compatible with the "scipy" and "pymc3" packages used in the system coding. The marginal distribution modeling page (Figures 92 and 93) consists of a left lateral banner with a button for modeling each feature and four squares on the right panel where fitting elements are displayed. The first square displays MCMC parameters for users' choice whenever a feature is numeric, the second shows MCMC distribution parameters convergence, and the bottom areas show original data histogram and marginal modeled distribution fitting referenced by frequencies in the sample. After each feature modeling, the corresponding button has its text turned red.

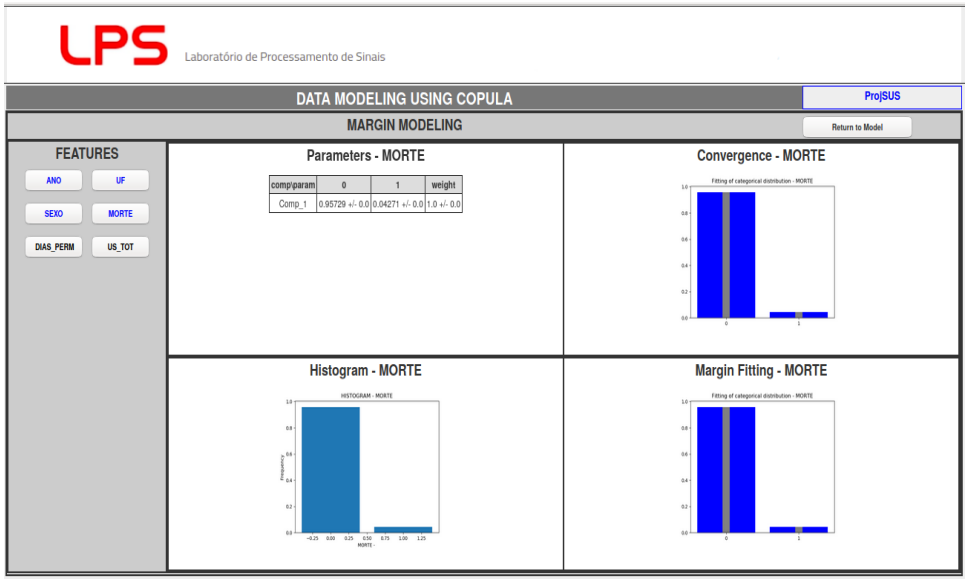


Figure 92 – Categorical feature fitting example: death in hospital. Parameters show probability by frequency estimation for each category (death in hospital or discharged alive).

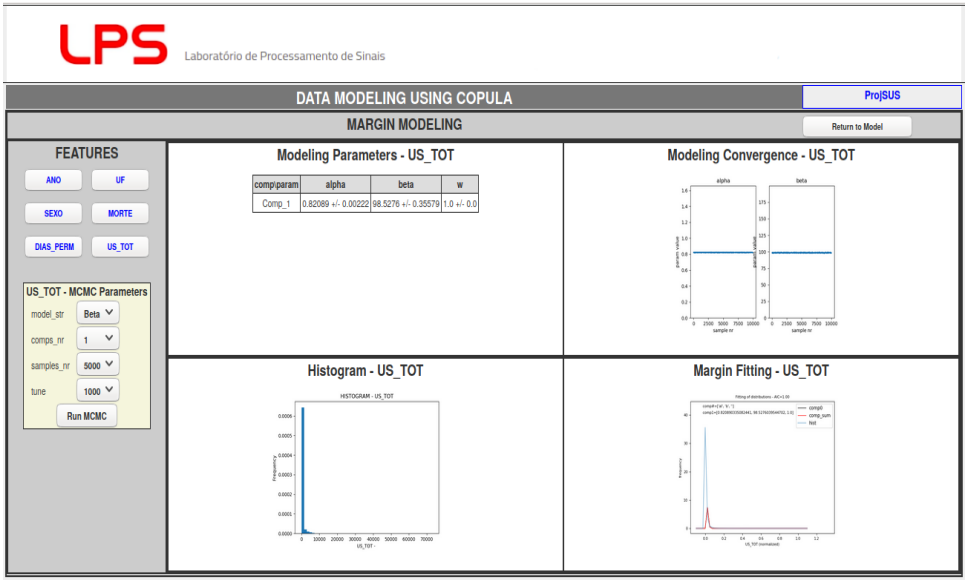


Figure 93 – Numeric feature fitting example: hospital treatment total costs in US dollars. Costs are very concentrated in low-cost area. In this case, beta fitting using MCMC (pymc3 package) resulted in a smoothed fitting for that number of samples and a spiky profile.

After all selected features are conveniently modeled, users can return to main page for starting copula modeling the data set. Copula modeling in this present version is based on empirical copula, and can be seen in Figure 94.

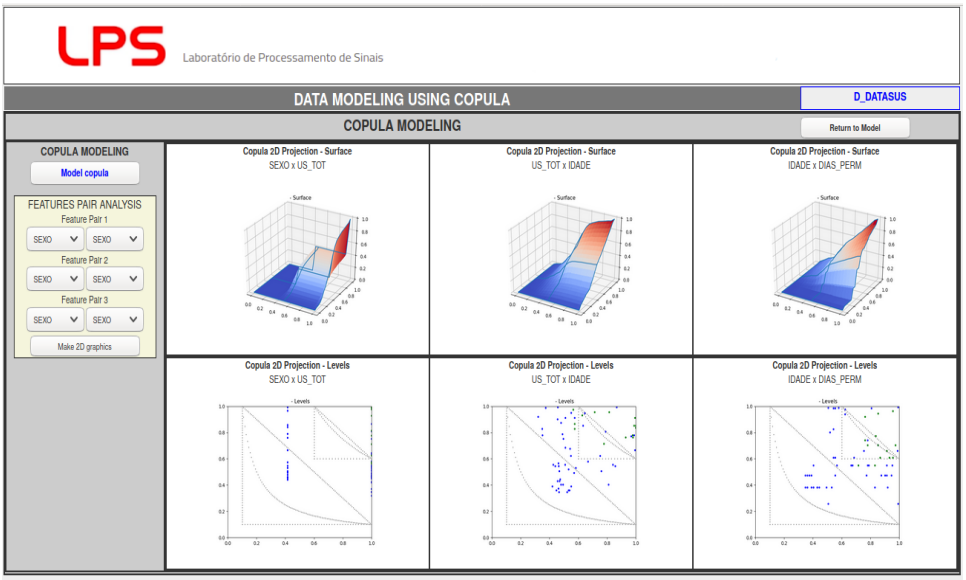


Figure 94 – Empirical copula modeling page showing features three pairs copula non smoothed flat projection surfaces and discrete footprints. Users can choose any three pairs for the corresponding copula projection to be displayed.

B.3 Illustrative Example

As a complete example of using the software, a DATASUS subset was taken considering a random proportional sample of all hospital admissions in Brazil through the years 2008 to 2010, which are exactly the results presented in all previous figures.

From the descriptive analysis (Figures 90 and 91), it can be seen that the subset contains 206,110 patients from all federations unities in Brazil. The most number of admissions corresponds to female patients (SEXO=3) and, from the map (in yellow), the maximum values for mean days in the hospital (DIAS_PERM) occur in Rio de Janeiro (RJ, Southeast) and Rio Grande do Norte (RN, Northeast). Also, the concordance table points towards strong concordance between days in hospital (DIAS_PERM) and cost (US_TOT).

Margin modeling of death (MORTE) and cost in US dollars (US_TOT) features resulted, respectively, in a multinomial with a 0.04271 probability of death and a Beta distribution with 0.82089 for alpha and 98.5276 for beta, with standard deviations of 0.00222 and 0.35579, as Figure 93 shows. It can also be noticed that those results for cost were obtained from an MCMC ran on 5,000 samples after a 1,000 tunning stage and a significant convergence was achieved.

Finally, Figure 94 presents three two-dimension projections of feature pairs and allows to identify a growing degree of positive dependence between cost and days in hospital as the surface initially follows a W-copula similar pattern but goes toward a more M-copula pattern at the top.

B.4 External Impact

The software can be a starting point tool for both existing and new research questions because it liberates researchers from more basic work on modeling by providing a good first whole model that is able to show each feature isolated behavior and features coupling general attributes by an empirical copula structure. Furthermore, the software was developed in a modular approach and can easily be improved by the inclusion of new marginal distribution and/or new copula parametric models by anyone in the scientific community.

The complete version of the software is very newly available even in the research group where it was developed but it has also been used in previous versions before the copula module was completed for descriptive health data analysis in a publication in early

2020 (PETERLE *et al.*, 2020) and also for helping in analyzing a multimodal distributed health feature regarding a disease study conducted by a multidisciplinary team. In both cases, it helped a lot by saving time and producing global and systematic phenomenon visualization for each research team.

As a general tool for modeling, this software is intended to be used in a widespread range of areas, wherever modeling is involved, especially when focusing concordance and dependence issues, with no previous restrictions.

LpsCopModel has a systemic approach and is intended to contribute in saving time and giving a panoramic visualization of a phenomenon by producing a first model based on MCMC parametric marginal distributions fitting and empirical copula modeling, where users can try some modeling options and grade complexity according to their needs.

It was developed for general use and therefore any dataset which can be registered in CSV format or converted to a pandas data frame can be input to be treated. In the same way, every figure and results can be exported, from parameters values and intervals to MCMC traces.

Simultaneously, by being open-source and modular, users can also upgrade by easily implementing by themselves new model options they happen to need or evolving the copula modeling to parametric models. In parallel, it is intended by the developer group to continue improving the software aggregating new modeling facilities, and enhancing the user interface.

APPENDIX C – REPRODUCIBILITY

Reproducibility has become a major concern in modern Science because methodologies are getting more complex and involving bigger datasets. Aware of this concern, we decided to dedicate an appendix chapter to describe in more practical details the proposed methodology deploy and to register where the interested reader could find data and code used in our research.

C.1 Data and Code Repositories

All data and code used in this research (except for other parts packages) are available at a GitHub repository at "https://github.com/wdarwinjr/lpscopmodel.git".

That repository follows a Django project directory structure. The project is the repository "lpscopmodel" itself which contains a configuration folder with the same name "lpscopmodel", an applications folder named "lpscopmodelapp", a "staticfiles" folder (empty and not used so far) and a folder for the HTML files named "templates" with the only HTML file in the project, "index.html".

Other than the Django structure and files, the complementary code and data to the lpscopmodel application used in this research is also there in a directory named "Jupyter" in a Jupyter Notebook file named "Thesis - Copulas and BNs.ipynb" and the corresponding input data and output directories in two folders named "data" and "figures".

C.2 Data Description

As mentioned in the main text, we have used five dataset groups in this research:

- **Experiment 01** - Bivariate (2D) Unimodal - 4 datasets:
  - independent
  - positive dependent
  - negative dependent
  - intermediate dependent
- **Experiment 02** - Bivariate (2D) Trimodal - 4 datasets:
  - independent
  - positive dependent
  - negative dependent
  - intermediate dependent
- **Experiment 03** - Multivariate (6D) Unimodal - 4 datasets:
  - independent
  - positive dependent
  - negative dependent
  - intermediate dependent
- **Real Case 01** - Brazilian Public Healthcare System (7D) - 1 dataset
- **Real Case 02** - Brazilian Counties Tax Revenue (11D) - 1 dataset

Each dataset in the first three groups was simulated by software using the first part of the code in "Thesis - Copulas and BNs.ipynb" jupyter notebook file, while the last two groups, containing only one dataset each, were colllected from their external sources. All resulting datasets from all five groups were saved each in one CSV file named after the dataset content.

- **Experiment 01** - A1
  - independent -> "A11_2_Indep_Uni_sample.csv"
  - positive dependent -> "A12_2_Posdep_Uni_sample.csv"
  - negative dependent -> "A13_2_Negdep_Uni_sample.csv"
  - intermediate dependent -> "A14_2_Rnddep_Uni_sample.csv"
- **Experiment 02** - A2
  - independent -> "A21_2_Indep_Tri_sample.csv"
  - positive dependent -> "A22_2_Posdep_Tri_sample.csv"
  - negative dependent -> "A23_2_Negdep_Tri_sample.csv"
  - intermediate dependent -> "A24_2_Rnddep_Tri_sample.csv"

- **Experiment 03** - B1
  - independent -> "B11_6_Indep_Uni_sample.csv"
  - positive dependent -> "B11_6_Posdep_Uni_sample.csv"
  - negative dependent -> "B11_6_Negdep_Uni_sample.csv"
  - intermediate dependent -> "B11_6_Rnddep_Uni_sample.csv"
- **Real Case 01** - D -> "D_DATASUS.csv"
- **Real Case 02** - E -> "E_TaxCountiesRevenue.csv"

Just as an explanation for the letters attributed to the groups, they originally followed a sequential alphabetic order ("A", "B", "C", an so on), but some of the dataset were not used in this research because it would not be worth it. The logic behind letters were to associate a different letter to different numbers of variables for the simulated datasets then followed by the real datasets. As we started with two groups of 2 and 10 variables, respectively, but later including a set with 6 variables, "A" stands for 2 variables simulated datasets, "B" for the 6 variables datasets and "C" for 10 variables datasets. Then, "D" and "E" were attached to the DATASUS and county tax administration datasets, respectively. As further we decided not to use the 10 variables dataset because the interested results were already achieved in a less comlpex way by the 6 variables dataset, it was excluded and then the jump from "B" to "D" in our registers.

For how the data was generated, it is detailed in the generation code itself, but as an overview, we can say that it was always produced from multivariated Gaussian distributions with the established number of variables for each dataset by attributing to it adequate mean vector and covariance matrix. The mean vector is formed by the simple unitary sequence starting at 1.0, $[1, 2, ..., n]$, where n is the number of variables in the dataset. The covariance matrix depends on the dependence to be attributed to the dataset, so it is constructed by setting each pairwise variable correlation to the desired dependence relation, which mneans a value from the set $[0.0, +0.999, -0.999, 0.3]$, corresponding to independence, strong positive dependence, strong negative dependence or intermediate dependence.

C.3 Code Description

The research was implemented by code in Python in four stages: simulated data generation and real data acquisition, LPSCopModel data treating (including MCMC marginal fitting), Bayesian networks structure scoring, and results analysis. For LPSCopModel, as it is a complete software application, we will let its description for its own repository documentation and restrain ourselves to describe here the Jupyter notebook code.

The Jupyter notebook summary is reproduced here:

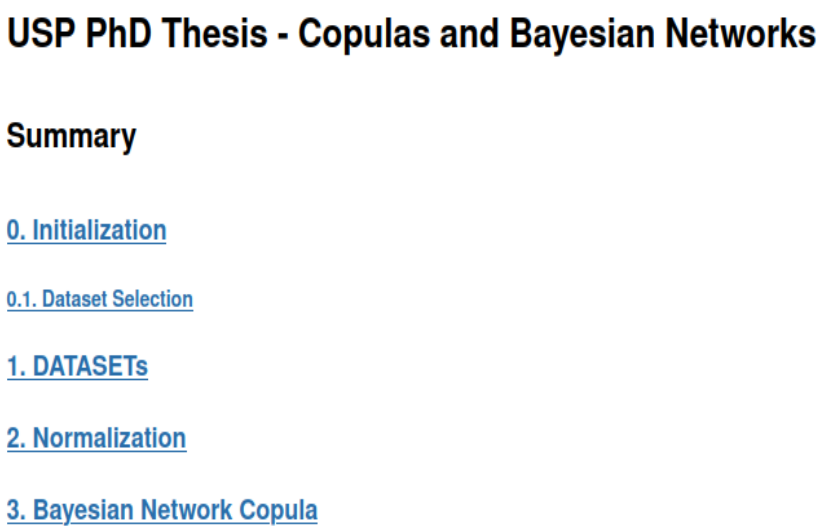


Figure 95 – Jupyter notebook summary.

First section is for initialization, therefore consisting in importing all modules, setting parameters and global variables, and constructing classes. Some variables and parameters are not necessarily used in the final version of the research, as parameters for 10 variables simulated datasets or some distribution fitting data parameters, which were later obtained directly from LPSCopModel by reading its output files, but it will be clear alongside the code analysis by the reader.

The main class for the research is the "Subject" class, which implements the object associated with each dataset in terms of its joint distribution and marginal distributions, with the corresponding probability computing and plotting functions. For the simulated data, a subclass ToySubject was also created, mainly for data generation from multivariate Gaussian distributions.

Section 1 is for simulated data generation and saving as CSV files, for further input

in LPSCopModel. It is very straightforward using classes and parameters. This section last part, the dataset analysis, is to be proceeded by using LPSCopModel having as input the dataset CSV file as a generic dataset. LPSCopModel will then generate an output in th form of a session file which will have distribution fitting parameters and concordances to be load in the last part of this section code.

Normalization and sample reducing for all methods (marginal distribution fitting, sample reducing and generative distribution mapping) is conducted in section 2, where also normalized joint distributions are plotted.

Section 3 stands for Bayesian network structure generating, searching, scoring, and plotting. First, a set of hundreds of structures is generated by two functions: one which uses referential basic structures, like naïve Bayes, sequential networks, binary trees, no connect, and so on, to generate many structures by variations of those elements, in some cases also inserting structures from what would be a rudimentary specialist guess. Before consolidating the possible structures set for each number of variables, it is cleaned from duplicated structures. After that, each dataset is taken as reference for scoring all networks structures in the set with the same number of variables as it, leading to a set of scored structures for each dataset. Then, two graphics are plotted showing in different colors each normalization method: one for the best score for each group of structures with the same number of edges for detecting the best number of edges and other for all scores altogether for tendency diagnosis. Finally, the structures which appear top ranked by all methods are analyzed and plotted.

Some collateral analysis and results are present along the described code and were not elements for the final version, but we have chosen not to delete them because some readers could be interested in those tests and figures. Nevertheless, if as it is a dynamic repository, if we further conclude that they are reducing code readability beyond any secondary benefit, we may exclude them, without any reproducibility loss.