Quantifying the Non-Linear Coupling Between Species: Copula and Rank Based Statistics

Lledo Esquerra-Ortells
University of Colorado at Boulder, lledo.esquerra@colorado.edu

Follow this and additional works at: https://scholar.colorado.edu/ecen_gradetds
Part of the Biology Commons, and the Electrical and Computer Engineering Commons

Recommended Citation
https://scholar.colorado.edu/ecen_gradetds/67

This Thesis is brought to you for free and open access by Electrical, Computer & Energy Engineering at CU Scholar. It has been accepted for inclusion in Electrical, Computer & Energy Engineering Graduate Theses & Dissertations by an authorized administrator of CU Scholar. For more information, please contact cuscholaradmin@colorado.edu.
Quantifying the Non-Linear Coupling Between Species: 
Copula and Rank Based Statistics

by

Lledó Esquerra-Ortells
B.S. and M.Sc., Polytechnic University of Catalonia,
Barcelona, Spain, 2011

A thesis submitted to the
Faculty of the Graduate School of the
University of Colorado in partial fulfillment
of the requirements for the degree of
Master of Science
Department of Electrical, Computer and Energy Engineering
2012
This thesis entitled:
Quantifying the Non-Linear Coupling Between Species: Copula and Rank Based Statistics
written by Lledó Esquerra-Ortells
has been approved for the Department of Electrical, Computer and Energy Engineering

______________________________
François G. Meyer

______________________________
Prof. Samuel M. Flaxman

______________________________
Prof. Manuel Lladser

Date _______________

The final copy of this thesis has been examined by the signatories, and we find that both the content and the form meet acceptable presentation standards of scholarly work in the above mentioned discipline.
Understanding the process of coevolution, the evolution of interacting species, is a major endeavor of evolutionary biology. Coevolution is a potent source of adaptive evolution, since it can sustain selection pressures indefinitely even in the absence of changes in the physical environment. In this work, we want to explain eco-evolutionary patterns in terms of adaptive behavior. Pursuant to this goal our methodology and the accompanying metrics are aimed at capturing different aspects of the dependence that has been shaped by evolution to respond in dynamic, adaptive ways to relevant features of the organisms environment. Previous studies, Nuismer et al. 2010[14], have shown that adaptive behavior does not fit in a linear model. Consequently, we must go beyond correlation analysis to track the population dynamics and the evolutionary changes over time.

A variety of ways to measure dependence exist. We are interested in those which are capable of tracking non-linear dependence, since a vast amount of literature has been written about linear correlation meanwhile dealing with non-linear dependence seems almost an uncharted field. We studied different concepts that would allow us to find the coupling, if it exists, among the different species statistics. Among them, copula and rank based statistics turned out to be very promising. Hence, this work describes this copula concept, and how it is related to other rank statistic tools, giving some toy examples to explain how it helps in exploring the underlying structure of the data. Finally, we apply those concepts to analyze
the dependence structure among the species of a simulated tritrophic system (predators-prey-resources or parasite-host-resources). We use the simulated data to explore two general questions. First, does coupling exist? And in the affirmative case, how is this dependence described? Second, does the underlying dependence structure of the predator-prey model describe the dependence between parasites and host; in other words, would we expect the two models to behave similarly in terms of adaptive behavior?.
Dedication

To my mother, my brother and Albert, for their unconditional encouragement.

And always remembering my father.
I would like to acknowledge Prof. Samuel M. Flaxman, Department of Ecology and Evolutionary Biology, for giving us access to his data and all the discussions, information and help he gave to me. I am acknowledging Prof. Manuel Lladser, Department of Applied Mathematics, for being part the committee. I want to thank my advisor Prof. François G. Meyer, Department of Electrical Engineering, for his advice, knowledge and support.

And I also wish to thank the Balsells Graduate Fellowship program, which involves Mr. Pete Balsells, the College of Engineering and Applied Science at CU Boulder, and la Generalitat de Catalunya, for funding my research.
Contents

Chapter

1 Introduction 1
   1.1 Motivation ................................................. 1
   1.2 Simulated Data ............................................. 3

2 Copula’s Mathematical Background 6
   2.1 Introductory Example ................................. 6
   2.2 Definition of Copula .................................. 12
   2.3 Rank Based Statistics ................................. 14
      2.3.1 Spearman’s $\rho$ .................................. 14
      2.3.2 Kendall’s $\tau$ .................................. 16
   2.4 Well-Known Copulas .................................. 16
   2.5 Empirical Copula ....................................... 18
      2.5.1 Parametrical Estimation of Copulas ............ 18
      2.5.2 Empirical Copula Definition .................... 19
      2.5.3 Goodness of fit .................................. 20

3 Results 23
   3.1 Graphical Diagnosis .................................... 23
   3.2 Rank Based Statistics ................................. 25
      3.2.1 Predator-Prey scenario ......................... 26
3.2.2 Parasite-Host scenario ........................................ 28
3.2.3 Extreme Cases .................................................... 30
3.3 Copula ............................................................... 32
3.4 Goodness of Fit ...................................................... 36

4 Conclusions and Future Work ....................................... 41
  4.1 Conclusions ......................................................... 41
  4.2 Future Work ....................................................... 42

Bibliography .................................................................. 44
Tables

3.1 KS statistic as a function of $k_p$, for $\theta_V = \theta_P = 0.01$. Parasite-host case.
Figures

Figure

2.1 The left panel shows the scatter plot of two independent lognormal random variables, while the right panel represents two dependent lognormal random variables. .............................................................. 7

2.2 The left panel shows the conventional scatter plot of the pairs \((Z_1, Z_2)\). The right panel represents the pair \((e^{Z_1}, e^{Z_2})\). In both cases \(r = 0.8\). ...................... 8

2.3 The left panel corresponds to the scatter plot of the pairs of ranks from the data set \((Z_1, Z_2)\). The right panel represents the scatter plot of the pairs of ranks from the data set \((X_1, X_2)\). In both cases \(r = 0.8\)............. 8

2.4 Bivariate Beta/t distribution based on a Gaussian copula. ................... 11

2.5 Bivariate Beta/t distribution based on a \(t\) copula with \(\nu = 1\). ........ 11

3.1 Joint distribution function of traits \(u\) and \(v\), host-parasite case. In the case where there is no plot is because the parasites went extinct. ................. 24

3.2 \(\chi\)-plot for \(\theta_V = \theta_P = 0.01, k_p = 1, \rho = 0.7996\). Host-Parasite case. .......... 25

3.3 Spearman’s \(\rho\) for the Predator-Prey case ............................................. 27

3.4 Spearman’s \(\rho\) as a function of \(k_p, \theta_p = \theta_p = 0.01\) for the Host-Parasite case.

The norm of the residuals for the linear regression is 0.4821. ................. 29

3.5 Spearman’s \(\rho\) for the Host-Parasite case ............................................. 30

3.6 Spearman’s \(\rho\) as a function of \(k_p\) for \(\theta_p = 1\). Red solid lines represent extinct cases for the Predator-Prey case. .................................................. 31
3.7 Spearman’s $\rho$ as a function of $k_p$ for $\theta_p = 1$. Red solid lines represent extinct cases for the Host-Parasite case.

3.8 The left panel represents the estimated joint distribution between the parasite $v$ trait and the host $u$ trait for $k_p = 1$ with $\theta_V = \theta_P = 0.01$, while the right plot corresponds to parasite $v$ trait and the host $u$ trait for $k_p = 1$ with $\theta_V = \theta_P = 1$.

3.9 The top-left plot corresponds to the 2D copula of the host $u$ trait for $k_p = 1$, $\theta_V = \theta_P = 0.01$, $\rho = 0.79956$, and the top-right shows the 2D copula of the host $u$ trait for $k_p = 1$, $\theta_V = \theta_P = 1$, $\rho = 0.6397$. The bottom pictures display the 3D versions.

3.10 Theoretical copulas. From left to right: independent, upper Fréchet bound and lower Fréchet bound.

3.11 Empirical copulas. From left to right: independent, upper Fréchet bound and lower Fréchet bound.

3.12 Norm 2 distance between the independent, upper and lower Fréchet bounds copula and the empirical copula as a function of $k_p$, for $\theta_P = \theta_V = 0.01$. Host-parasite case.

3.13 Original data vs. simulated data coming from a Gaussian copula with $\rho = 0.7411$, corresponding to the host-parasite case with $\theta_V = \theta_P = 0.01$ and $k_p = 1.75$.

3.14 Original data vs. simulated data coming from a $t$ (left) or Clayton (right) copula with $\rho = 0.7411$, corresponding to the host-parasite case with $\theta_V = \theta_P = 0.01$ and $k_p = 1.75$. 
Chapter 1

Introduction

1.1 Motivation

The goal of this work is to quantify the non-linear coupling that characterizes two species. In fact, studying the linear correlation between traits of interacting species has long been used for identifying cases of coevolution. But it has been shown that coevolution is neither a necessary nor a sufficient condition for the evolution of linear correlated traits between two species, Nuismer et al 2010[14]. Consequently, we must go beyond linear correlation analysis to track this coupling.

We studied different concepts that would allow us to find the coupling, if it exists, among different random variables. Among them, copula and rank based statistics turn out to be promising.

Dependence properties are closely related to measures of association. Important types of measures of associations are rank based statistics, and from those we work with Spearman’s $\rho$ and Kendall’s $\tau$. Both of them measure a form of dependence, Nelsen 2006[13], called concordance: pairs of random variables are concordant if “large” (small) values of one tend to be associated with “large” (small) values of the other.

From another point of view, copula, which is closely related to rank based statistics, has been widely used in financial and credit risk studies (it is very useful for analyzing and pric-
ing volatility smile/skew of exotic baskets, for example). But after doing some research we realized that we could not find literature referring the use of copula for the aim of tracking coevolution.

In general terms, copula can be understood as a way of formalizing dependence structures between random variables. It has been shown that the dependence between two random variables is uniquely characterized by a copula (we are going to show evidence of that later in this work). What is more, if the random variables come from continuous marginal distribution functions this copula is unique.

But what is important is to state the advantages of using copula. In other words, the reasons why we would like to study and use copula instead of other tools are also able to detect dependence.

On the one hand, in contrast with Pearson’s $r$, copula is able to track non-linear correlation. On the other hand, rank based statistics, which are able to detect non-linear association, only provide a one dimensional scalar, while copula describes the complete dependence structure. In addition, copula allows a multidimensional framework giving up the Gaussian framework, therefore, it avoids the main restrictions of the most common approaches: both the individual and joint behavior of the two variables must be described by certain distributions and under restrictive conditions to fit into a particular joint distribution. Furthermore, comparing with joint distribution functions, copula allows the modeling of the marginals separately from their dependence structure, this greatly simplifies the estimation problem of a joint stochastic problem since not always a good estimation of the margins leads to a good estimation of the joint; and by construction, the graph of any copula is a continuous surface within the unit cube whose boundary is the skew quadrilateral with vertices $(0,0,0), (1,0,0), (1,1,1), (0,1,0)$, this fact allows to compare the dependence structures underlying different pairs of random variables using copulas, a task difficult to achieve using
the joint distribution functions

1.2 Simulated Data

The model represents a tritrophic system and considers two scenarios: predator-prey-resources and parasite-host-resources. For instance, we can imagine a system involving fava bean plants (Vicia faba), pea aphids (Acyrthosiphon pisum), and sevenspotted lady beetles (Coccinella septempunctata) in experimental microcosms. Our objective is to investigate the details of predator-prey-resources (parasite-host-resources) interactions in appropriate spatial contexts, since these interactions are the links that form the structure of food webs and ecological communities. Furthermore, we wish to know about the eco-evolutionary consequences of these interactions. We would like to answer questions about how predators and prey (parasite and host) evolve in their environments, relative to resources and relative to each other: can predators benefit by using the distribution of the preys resource as a guide in choosing habitat? How should prey respond adaptively to predators?

In this model population dynamics occur at all three levels. It is described in Flaxman et al. 2011[6]. This model incorporated: a) variability, interactions, and stochasticity within populations, b) conditional movement based on local information individuals have about densities, c) movement mechanisms evolve (mutation, selection, drift). The habitat is arranged in $N$ discrete patches. During a single step, movements can only occur between adjacent patches. Suppose that the $i$th predator is located in patch $n$ at time $t$. The probability that this predator (parasite) will move to adjacent patch $m$ ($= n - 1$ or $n + 1$) is defined by

$$
\mu_{i,m} = \Delta t (d_P + \alpha_{P,i} [\tau_P g(R_{m,t}, R_{n,t})] + (1 - \tau_P) g(V_{m,t}, V_{n,t})).
$$
A similar equation defines the behavior of the prey (host). However, the new model used in this work has some additional traits. The suitability of \( j \)th prey (host) to predator (parasite) \( i \)th is modeled by

\[
exp(-\theta_P(v_i - u_j)),\tag{1.1}
\]

and the consumption of resources by \( j \)th prey (host) is constrained by

\[
exp(-\theta_V(u_j - u_{opt})).\tag{1.2}
\]

For large values of \( \theta_P \), even a small difference between \( v_i \) and \( u_j \) can make the prey unsuitable, by contrast for small values of \( \theta_P \) (i.e., close to zero), nearly all prey can be suitable for a predator, regardless of \( v_i \) and \( u_j \). Note that Eq. 1.1 not only determines how well suited predator \( i \) is to consuming prey \( j \), but also thus determines how much risk of death prey \( j \) faces from predator \( i \). Similarly to the role played by \( \theta_P \) for predators, for prey, if \( \theta_V \) is large, then prey can only consume resources when \( u_j \approx u_{opt} \). As \( \theta_V \) decreases, the prey is less and less evolutionarily constrained in the values of \( u_j \) that can evolve, and in the limiting case of \( \theta_V = 0 \), values of \( u_j \) are not constrained. Naturally, in the parasite-host scenario the effects are the same. Thus, it is easy to infer the important role that \( u \) and \( v \) play in this model.

Among the parameters that shape the behavior of the system, the ones under study are:

1) abundances, 2) \( \alpha_P, \alpha_V \) that describe the predator/prey or host/parasite sensitivity to inter-patch differences, 3) \( \tau_P, \tau_V \) defining predator/prey or host/parasite sensitivity in making movements, 4) predators intrinsic mortality \( k_p \), 5) \( v \) that plays the role of predator’s or parasite’s trait value, 6) prey’s or host’s trait value \( u \). To give a more intuitive idea about the role of \( u \) and \( v \) we can understand as any phenotype characteristic that may facilitate or make more difficult the adaptation to the environment. For example their sizes, color, speed, etc.
We consider the entire system without differentiating the distinct patches; differen-
tiation between patches is left as future work. Note that the equations only describe the
individual behavior, i.e., there is not any deterministic description of the system conduct
from top. Hence, the aim of our experiments is to track and describe codependence, if it
exists, in the overall system. In particular, between $u$ and $v$ under different environmental
circumstances and system scenarios, as a representative illustration of the whole system.

We simulate the data from the aforementioned model under two different scenarios: predator-
prey and parasite-host. The attributes that differentiate these scenarios are: first, the par-
asites population is much larger than the predators one; in addition, their conversion effi-
ciency, $e_p$, is greatly increased, which implies that a single successful parasite attack may
enable the parasite to produce many offspring; finally, the parasites intrinsic mortality rate,
$k_p$, is greatly increased, in combination with the change in $e_p$, this reflects a notion of a
parasites life history as being one which involves living “fast” and having a short lifespan.
Thence we expect the two models to behave differently in terms of adaptive behavior.

The simulated data is stored in groups of time series. The values of the parameters and
populations of the system within each group of time series represent the results from the sim-
ulations under different environmental circumstances imposed and modeled by the different
value combination of $\theta_P$, $\theta_V$ and $k_p$. Specifically, we are going to observe how $u$ and $v$ evolve
under all the possible combinations among $\theta_V = \{0.0001, 0.01, 1\}$, $\theta_P = \{0.0001, 0.01, 1\}$ and
$k_p$. We have already said that the mortality rate of the parasites is grater than the predi-
tors’; for instance, $k_p = \{0.04, 0.05, 0.06, 0.07, 0.08, 0.09, 0.10, 0.11, 0.12, 0.13\}$ for predators,
and $k_p = \{1, 1.5, 2, 2.5, 3\}$ for parasites. Finally, we want to highlight that $u$ and $v$ are chosen
from a Gaussian distribution with a fixed means and some standard deviation $\sigma_u$ and $\sigma_v$ that
change when mutation occurs.
2.1 Introductory Example

We are going to consider only samples given from a pair \((X_1, X_2)\) of continuous random variables. These variables could be, for instance, the prey and predator’s population sizes, or the host’s and parasite’s trait values.

We can define two random variables coming from a bivariate normal distribution with linear correlation parameter \(r\), and then transform them to get two lognormal random variables

\[
Z = [Z_1, Z_2] \sim N \left( [0, 0], \begin{bmatrix} 1 & r \\ r & 1 \end{bmatrix} \right)
\]

\[
X = [X_1, X_2] = [e^{Z_1}, e^{Z_2}]
\]

The left picture in Fig.2.1 shows the scatter plot for \(r = 0\), i.e. for \(Z_1\) and \(Z_2\) independent, and consequently, \(X_1\) and \(X_2\) independent too. While the picture in the right shows the scatter plot of two dependent lognormal variables. It is clear that there is a tendency in the second data set for large values of \(X_1\) to be associated with large values of \(X_2\), and similarly for small values. The correlation parameter \(r\) of the underlying bivariate normal determines this dependence. However, the correlation coefficient for \(X_1\) and \(X_2\) is not \(r\) anymore. That is because the linear correlation coefficient expresses the linear dependence between random variables, and when nonlinear transformations are applied to those random variables,
linear correlation is not preserved. Instead, a rank based statistic, such as Kendall’s \( \tau \) or Spearman’s \( \rho \), is more appropriate. Both of them are important measures of associations, which measure a form of dependence called concordance\(^1\), Nelsen 2006[13].

![Scatter plots](image)

Figure 2.1: The left panel shows the scatter plot of two independent lognormal random variables, while the right panel represents two dependent lognormal random variables.

Displayed in Fig. 2.2 is the scatter plot of pairs \((Z_1, Z_2)\) and \((X_1, X_2)\), with \(r = 0.8\), which look very different. However, Fig. 2.3 shows the scatter plot of the ranks associated with \((Z_1, Z_2)\) and \((X_1, X_2)\) respectively, which are exactly the same. That is explained by the fact that, unlike Pearson’s \(r\), rank based statistics are invariant by monotone transformations.

\(^1\) A pair of r.v. are concordant if “large” (small) values of one tend to be associated with “large” (small) values of the other.
Figure 2.2: The left panel shows the conventional scatter plot of the pairs \((Z_1, Z_2)\). The right panel represents the pair \((e^{Z_1}, e^{Z_2})\). In both cases \(r = 0.8\).

Figure 2.3: The left panel corresponds to the scatter plot of the pairs of ranks from the data set \((Z_1, Z_2)\). The right panel represents the scatter plot of the pairs of ranks from the data set \((X_1, X_2)\). In both cases \(r = 0.8\).
From these results we infer that it may be more interesting to find association in terms of ranks that, as it has been shown, are preserved under monotonic transformations. While there is nothing wrong with looking at the pattern of pairs \((Z_1, Z_2)\) or \((X_1, X_2)\), Fig. 2.2, we must be aware that these pictures not only incorporate information about the dependence between the pairs of random variables, but also about their marginal behavior. In effect, both plots are distortions of the same dependence structure. Thus the next question we should ask is: is there any “tool” which allows the modeling of the marginals separately from their dependence structure? And the answer is yes, and this tool is called copula.

Returning to our toy example, by definition, applying the normal cumulative distribution function (cdf)\(^2\), denoted here by \(\phi\), to a standard normal random variable results in a random variable that is uniform on the interval \([0, 1]\). To see this, if \(Z\) has a standard normal distribution, then the cdf of \(U = \phi(Z)\) is

\[
P(U \leq u) = P(\phi(Z) \leq u) = P(Z \leq \phi^{-1}(u)) = u
\]

and that is the cdf of a \(U(0, 1)\) random variable. The point in transforming our random variables in ones that are uniformly distributed is that borrowing from the theory of univariate random number generation, applying the inverse cdf of any distribution, \(F\), to a \(U(0, 1)\) random variable results in a random variable whose distribution is exactly \(F\) (Inverse Transform method). Hence, we can apply a two-step transformation to each variable of a standard bivariate normal, creating dependent random variables with arbitrary marginal distributions. Because the transformation works on each component separately, the two resulting random variables need not even have the same marginal distributions.

---

\(^2\) The cdf describes the probability that a real-valued random variable \(X\) with a given probability distribution will be found at a value less than or equal to \(x\).
The transformation is defined as:

\[ Z = [Z_1, Z_2] \sim N \left( [0, 0], \begin{bmatrix} 1 & r \\ r & 1 \end{bmatrix} \right) \]

\[ U = [U_1, U_2] = [\phi(Z_1), \phi(Z_2)] \]

\[ X = [X_1, X_2] = [G_1(U_1), G_2(U_2)] \]

where \( G_1 \) and \( G_2 \) are inverse cdfs of two possibly different distributions. It can be shown that for any \( G_1 \) and \( G_2 \) the pairs \((Z_1, Z_2)\) and \((X_1, X_2)\) have the same rank statistics. But, since \( \rho \) and \( \tau \) are already invariant by monotone transformations why should we want to use copula? Because different dependence structures could have the same rank based statistics. However, as we will see in Theorem 1, the dependence between the pairs of continuous random variables \((Z_1, Z_2)\) and \((X_1, X_2)\) is uniquely characterized by the same copula, whatever it may be.

Without going into technical details yet of how we can construct a copula, Fig. 2.4 shows a bivariate Beta/t distribution based on a Gaussian copula, while Fig. 2.5 corresponds to a bivariate Beta/t distribution but based on a \( t \) copula with one degree of freedom. Both have the same marginal distributions and the same rank correlation between variables but exhibit a very different dependence structure. This illustrates the fact that multivariate distributions are not uniquely defined by their marginal distributions, or by their correlations, and the necessity of using copula.
Figure 2.4: Bivariate Beta/t distribution based on a Gaussian copula.

Figure 2.5: Bivariate Beta/t distribution based on a $t$ copula with $\nu = 1$. 
Copula, by definition (see Theorem 1) can be interpreted as a joint distribution function between two uniform variables. But at least two main advantages can be found in the use of copula: 1) copula allows the modeling of the marginals separately from their dependence structure, this greatly simplifies the estimation problem of a joint stochastic problem (not always a good estimation of the margins leads to a good estimation of the joint); 2) by construction, the graph of any copula is a continuous surface within the unit cube whose boundary\(^3\) is the skew quadrilateral with vertices (0,0,0), (1,0,0), (1,1,1), (0,1,0), this fact allows to compare the dependence structures underlying different pairs of random variables using copulas, a task difficult to achieve using the joint distribution functions because they are usually constructed on different scales, giving us more reasons to further investigate copula.

2.2 Definition of Copula

The study of copulas and their applications in statistics is a rather new tool. It was not until 1959 that Sklar[16] introduced the word copula in mathematics and statistics with the following Theorem:

**Theorem 1 Sklar’s Theorem.** Let \( H \) be a two-dimensional distribution function with marginal distribution functions \( F \) and \( G \). Then there exists a copula such that

\[
H(x_1, x_2) = C(F(x_1), G(x_2))
\]

Conversely, for any univariate distribution functions and any copula, the function is a two-dimensional distribution function with marginals \( F \) and \( G \). Furthermore, if \( F \) and \( G \) are continuous, then \( C \) is unique.

\(^3\) Later in this section this boundary is justified.
Moreover, being $Z = \varphi(X)$ and $T = \psi(Y)$, where $\varphi$ and $\psi$ are monotone increasing transformations, define $\hat{H}(z, t) = \hat{C}\{\hat{F}(z), \hat{G}(t)\}$, with

$$\hat{F}(z) = P(Z \leq z) = P\{X \leq \varphi^{-1}(z)\} = F\{\varphi^{-1}(z)\}$$

$$\hat{G}(t) = P(T \leq t) = P\{Y \leq \varphi^{-1}(t)\} = G\{\varphi^{-1}(t)\}$$

then

$$\hat{H}(s, t) = P(Z \leq z, T \leq t) = P\{X \leq \varphi^{-1}(z), Y \leq \psi^{-1}(t)\} = \hat{C}\{\hat{F}(z), \hat{G}(t)\}$$

(2.1)

$$= H\{\varphi^{-1}(z), \psi^{-1}(t)\} = C\left[F\{\varphi^{-1}(z)\}, G\{\psi^{-1}(t)\}\right] =$$

$$= C\{\hat{F}(z), \hat{G}(t)\}$$

So, $\hat{C} = C$.

Eq. 2.1 shows that the unique copula associated with a random pair $(X, Y)$ is invariant by monotone increasing transformations of the marginals, Schweizer and Wolff 1981[15].

Based on Theorem 1, we can interpret copulas as functions that join or couple multivariate distributions to their one-dimensional margins, a way of formalizing dependence structures of random variables. In other words, a copula captures those properties of the joint distribution that are invariant under almost surely increasing transformation. Consequently, it turns out to be very useful to work together with rank based statistics (e.g. Spearman’s $\rho$ or Kendall’s $\tau$), insofar as it is the study of properties invariant under such transformations.

But what is very interesting, as Theorem 1 states, is the fact that since the dependence between two continuous random variables is uniquely characterized by a copula, an accurate graphical representation of dependence should exhibit the same invariance property, which in addition to understand the dependence structure underneath two random variables it may help us to extrapolate from known data what would be the codependence between them in the regions (situations) where no information is available.
Namely, having two datasets (two random vectors) coming from two unknown margins we can construct a tool which allow us to interpret the non-parametric nature of the dependence (or otherwise the lack of it) between them, as well as, to infer what would be the behavior of those in situations where we have no data.

2.3 Rank Based Statistics

One of the main focuses of this study is the exploration of the role that copulas play in the analysis of dependence. Dependence properties are closely related to measures of association. Two of the most important measures of associations are Spearman’s $\rho$ and Kendall’s $\tau$ (which are rank based statistics), both of them measure a form of dependence, Nelsen 2006[13] called concordance\(^4\). Later in this section we explain how these dependence measures are related to copulas.

2.3.1 Spearman’s $\rho$

An intuitive way to understand $\rho$ is defining the “common” linear correlation coefficient\(^5\) but using the ranks instead of directly the data. In a formal manner, this is to say that $\rho$ is proportional to the probability of concordance minus the probability of discordance. Thus, given a random sample of data $(X_i, Y_i)$, after calculating their ranks $(R_i, S_i)$, the empirical $\rho_n$ is estimated

$$\rho_n = \frac{12}{n(n + 1)(n - 1)} \sum_{i=1}^{n} R_i S_i - 3 \frac{n + 1}{n - 1} \in [-1, 1]$$

Moreover, it turns out that $\rho_n$ is an asymptotically unbiased estimator of

$$\rho = 12 \int_{[0,1]^{2}} C(u,v)dvdu - 3 = 12 \frac{1}{n} \sum_{i=1}^{n} \frac{R_i}{n + 1} \frac{S_i}{n + 1} - 3 = \frac{n - 1}{n + 1} \rho_n$$

\(^4\) A pair of r.v. are concordant if “large” (small) values of one tend to be associated with “large” (small) values of the other.

\(^5\) Pearson’s $r$
In fact, it has been shown, Genest et al.[8], that given a family $C_\theta$ of copulas indexed by a real parameter, the theoretical value of $\rho$ is, in general, a monotone increasing function of $\theta$. Of course, once a copula choice has been made different relations between $\theta$ and $\rho$ have already been described depending on the chosen copula.

In addition, using this fact an independence test for our data can be designed: under the null hypothesis $H_0 : C = \Pi$ (independent copula), the distribution of $\rho_n$ is close to a $N(0, \frac{1}{n-1})$, so that one may reject $H_0$, with a significance level $\alpha$, if

$$\sqrt{n-1}\left|\rho_n\right| > z_{\alpha/2} \quad (2.3)$$

When the data is roughly elliptically distributed and there are no prominent outliers, the results given by Spearman correlation and Pearson correlation are very similar.

At this point, the reader may wonder why not use Pearson’s $r$ instead. Although $r$ is widely used in the measurement of association between random variables, it was found to have some deficiencies compared to $\rho$: 1) if the expected value of the empirical $\rho_n$ is equal to $\pm 1$, then $X$ and $Y$ are functionally dependent, linearly or non-linearly (whenever their underlying copula is one of the two Fréchet-Hoeffding bounds, Nelsen 2006[13]); in contrast, when the expected value of the empirical $r_n$ is $\pm 1$ we can only state that $X$ and $Y$ are linear functions of one another, consequently having $r = 0$ does not imply independence, it means that there does not exist linear dependence, but other types of association may exist; 2) $\rho_n$ estimates a population parameter that is always well defined, whereas there are heavy-tailed distributions (such as the Cauchy, e.g.) for which a theoretical value of Pearson’s correlation does not exist; 3) the Spearman correlation is less sensitive than the Pearson correlation to strong outliers that are in the tails of both samples; 4) $r_n$ is not invariant under increasing transformations.
2.3.2 Kendall’s $\tau$

Kendall’s $\tau$ plays an essential role in estimating the parameters of certain copulas. The theoretical value of $\tau$ is a monotone increasing function of the copula real parameter $\theta$, Genest 2007[8] whenever a family $C_\theta$ of copulas is ordered by positive quadrant dependence$^6$.

The empirical version of $\tau$ is given by

$$\tau_n = \frac{4}{n(n-1)}P_n - 1$$

(2.4)

where $P_n$ is the number of concordant pairs$^7$. Note that $\tau_n$ can take the same range of values that $\rho$ does.

Under suitable conditions, $\tau_n$ is also an unbiased estimator of

$$\tau = 4 \int_{[0,1]^2} C(u,v)dC(u,v) - 1$$

Therefore, an alternative test of independence can be based on $\tau_n$, which turns out to be statistically close to a normal $N(0, \frac{2(2n+5)}{9n(n-1)})$, then reject $H_0$ if $\sqrt{\frac{9n(n-1)}{2(2n+5)}}|\tau_n| > z_{\alpha/2}$.

2.4 Well-Known Copulas

We want to present some of the well-known dependence structures defined by the different families of copulas, to see if they can fit our data. Since those copula families have been studied and characterized, once we can infer that our data has a specific copula structure several conclusions can be made.

For instance, it is easy to see that if the two variables under analysis are stochastically independent, then $C = F(x_1) \cdot G(x_2)$, Nelsen 2006[13].

$^6$ $P[X \leq x, Y \leq y] \geq P[X \leq x]P[Y \leq y]$.

$^7$ $(X_i - X_j)(Y_i - Y_j) > 0$. 
The other extreme case, one variable is a function of the other, corresponds to the Fréchet Bounds. When $X_2$ is almost surely an increasing function of $X_1$, i.e. they are comonotone variables they define an Upper Fréchet Bound, formally defined by

$$ C = M = \min(F(x_1), G(x_2)) $$

While the Lower Fréchet Bound represents countermonotone variables, that is to say, $X_2$ is almost surely a decreasing function of $X_1$, that can be technically translated into

$$ C = W = \max(F(x_1) + G(x_2) - 1, 0) $$

It has been shown that any copula $C$ that represents a model of dependence has to lie somewhere between this two extremes. More formally

$$ W(u, v) \leq C(u, v) \leq M(u, v) $$

where $u = F(x_1)$ and $v = G(x_2)$. This fact explains the copula boundary aforementioned.

The Fréchet Bounds are particularly interesting as a mechanism to compare how one variable influences the other; for example, how the prey’s (host’s) trait value shapes the predator’s (parasite’s) trait value.

Among all the different types of copulas the most famous family is the Archimedean, both because of the ease with they can be constructed and the great variety of subfamilies which belong to this class, which allows to model many different dependence structures. It is not the aim of this work to get into too much detail in how they are constructed, just remarking that the level curves of a Archimedean copula are convex, property that may be useful to understand the analysis of the data in the case that it fits that class of copula. Among the different subfamilies we want to highlight the Clayton family, defined by

$$ C(u, v) = \psi(\psi^{-1}(u), \psi^{-1}(v)), \quad \text{with } \psi(t) = (1 + \theta t)^{-1/\theta} $$
with $\theta$ being a parameter that can be estimated using some likelihood based method using the rank based statistics. Other interesting copula families that have interesting characteristics that may help to explain the dependence structure of our data are Gaussian copula, described by the following equation

$$C(u, v) = \Phi_{\Sigma}(\Phi_{\Sigma}^{-1}(u), \Phi_{\Sigma}^{-1}(v))$$

where $\Phi_{\Sigma}$ is the joint cumulative distribution function of a multivariate normal distribution with mean vector zero and covariance matrix equal to the correlation matrix $\Sigma$; and Student’s $t$ (we are not going to define it mathematically due to its complexity).

We mention this three families in particular because they are able to detect the presence, or not, of tail dependence and dependence asymmetry:

- A Gaussian copula describes a dependence structure where dependence occurs around the mean values (no tail dependence).

- When the data fitted by a Students t copula association the values of the variables are far from the mean (tail dependence).

- Clayton copula would be a suitable choice if correlation is located in the corner of the lower-left quadrant of a bivariate distribution.

2.5 Empirical Copula

2.5.1 Parametrical Estimation of Copulas

Notwithstanding, how can we choose a particular copula that will correctly formalize the dependence structure of our data? In other words, how we can model the dependence structure with consistency using a copula that is going to converge to the real underlying structure. Usually one takes a parametric family of copulas among many existing others
and fit it by estimating the family parameters, $\theta$, (maximum likelihood, information matrix, or estimation based on the dependence measure might be useful). There exist a variety of copulas at our disposal, besides several general methods of constructing them (inversion, geometric, algebraic, ...), Nelsen 2006[13]. Nevertheless, there does not exist a systematic rigorous method for the choice of the copula that will converge to the real structure dependence. This can provide biased results since according to the dependence structure selected the observed results might be very different. In other words, assuming an underlying dependence structure if we do not know what to expect from our data may be very risky and lead us to wrong results.

Despite that, there exist direct formulas that relate the empirical rank statistics, $\rho$ and $\tau$ with the parameter $\theta$ of the desired copula. Even though $\rho_n$ and $\tau_n$ are estimated non-parametrically, the relationship between them and $\theta$ depends on the chosen copula, therefore it becomes a parametric estimation. In any case, if the selected copula turns out to be an appropriate one, this is a simple and easy method. On of the most useful methods is the “maximum pseudo likelihood estimator” , formally defined by

$$l(\theta) = \sum_{i=1}^{n} \log[c_\theta\{F_n(X_i), G_n(Y_i)\}]$$ (2.5)

with $F_n(X_i) = \frac{R_i}{n+1}$, $G_n(Y_i) = \frac{S_i}{n+1}$ and $c_\theta$ is the density of the copula. This method may seem superficially less attractive than other methods, it is more generally applicable.

### 2.5.2 Empirical Copula Definition

However, we are interested in a full empirical method because it provides a robust and universal way for estimation purposes, besides it is a more flexible method. We compute what is called the empirical copula using the data that we have at our disposal. We have already stated that the pairs of ranks $((R_1, S_1), \ldots, (R_n, S_n))$ associated with a sample data are statistics which retain a great amount of information (apart from being invariant to
monotone increasing transformations). These ranks are unambiguously defined (ties occur with probability 0 assuming continuous margins). Plotting these pairs of ranks, Genest 2007[8] as we did in Fig. 2.3, upon rescaling of the axes by a factor of \( \frac{1}{n+1} \), we get a set of points in the unit square \([0,1]^2\), which form what is called the empirical copula domain, Deheuvels 1979[3] defined the rank based estimator of \( C(u,v) \)

\[
C_n(u,v) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I} \left( \frac{R_i}{n+1} \leq u, \frac{S_i}{n+1} \leq v \right),
\]

(2.6)

whose large-sample distribution is centered at \( C(u,v) \) and normal.

**2.5.3 Goodness of fit**

As in any modeling problem, when several choices can be made, the next step is to develop (use) a metric to determine how well different copulas (or whatever we are modeling) fit the observed data. In other words, we must find the copula that better describes the dependence distribution using different goodness of fit criterion.

A graphical diagnosis may be the most natural way of checking the adequacy of the model, though perhaps not the most accurate. We would compare the scatter plot of an artificial data set of the same size of the original one generated from the (parametrically or non-parametrically) estimated \( C_n \) with the scatter plot of the support of the empirical copula (i.e. \( R_i/(n+1), S_i/(n+1) \)). This procedure, may give us a first impression about the accuracy of our estimation.

In the general case, in order to have a more robust reason to accept or reject an estimated copula than a simple visual sensation, we could use a distance based on the \( p \)-norm between each considered theoretical copula and the empirical copula. It appears that any distance based on the \( p \)-norm would be proper. So, let us consider the 2-norm. Assume that we have
different empirical copulas \( \hat{C}_j \) with \( j = 1, \ldots, J \), and a finite subset of \( K \) copulas

\[
\tilde{C} = \{ C_k \}_{1 \leq k \leq K},
\]

\[
d_2(\hat{C}_j, C_k) = ||\hat{C}_j - C_k||_{L^2} := \sqrt{\sum_{i=1}^{N} |\hat{C}(u_i) - C_k(u_i)|^2}.
\]

Another possible goodness of fit test is the Kolmogorov-Smirnov test based on the Kolmogorov distance, defined as the supremum over the absolute difference between two cumulative density functions. We would like to contrast the results obtained from different criteria.

Note that we have already explained how one can compute the empirical copula. But we have not yet specified how to construct the theoretical copula that we want to compare with. For this purpose we are going to use the existent relationship between \( \rho \) and the parameter \( \theta \), which is going to be different depending on the copula.

When \( \theta \) is real, the rank based estimator is defined as \( \hat{\theta} = h(\rho_n) \). Gaenssler and Stute, 1987[7], showed that \( \rho_n \sim N(\rho, \sigma^2/n) \), where \( \sigma^2 \) depends on the underlying copula. Then the confidence interval for \( \theta \) is given by \( \hat{\theta}_n \pm x_{\alpha/2} \frac{1}{\sqrt{n}} \sigma_n |h'(\rho_n)| \). Once we have \( \hat{\theta} \), we are able to generate the theoretical copula, simulate new data using this \( C_{\text{the}} \) and compare the resulting data set with the original one.

For instance, if we consider the Normal bivariate case with correlation parameter \( r \), we can calculate an estimation of \( r \) from \( \rho_n \) using the fact that

\[
\rho = \frac{6}{\pi} \arcsin\left(\frac{r}{2}\right).
\]

We can also use \( \tau \) instead. For example, for the Clayton family, \( \tau_\theta = \theta / (\theta + 2) \)

\(^8\text{It is constructed from a multivariate normal distribution over by using the probability integral transform.}\)
On the other hand, there exist other ways to estimate $\theta$ without directly using $\rho$ or $\tau$, although they are also rank based. Using again the “maximum pseudo likelihood estimator”, from Eq. 2.5 one can estimate $\hat{\theta}$ solving Eq. 2.7

$$\hat{l}(\theta) = \frac{\delta}{\delta \theta} l(\theta) = \sum_{i=1}^{n} \frac{\hat{c}_\theta(R_{i/n+1}, S_{i/n+1})}{c_\theta(R_{i/n+1}, S_{i/n+1})} = 0. \quad (2.7)$$

Furthermore, $\hat{\theta}_n \sim N\left(\theta, \nu^2/n\right)$. A consistent estimate of $\nu^2$ is given by $\hat{\nu}^2_n = \hat{\sigma}^2_n/\hat{\beta}^2_n$ with

$$\hat{\sigma}^2_n = \frac{1}{n} \sum_{i=1}^{n} (M_i - \bar{M})^2 \quad \hat{\beta}^2_n = \frac{1}{n} \sum_{i=1}^{n} (N_i - \bar{N})^2,$$

with $M_i$ and $N_i$ being the pseudolikelihood observations.

Thus, to conclude this section we want to say that knowing the distribution of the estimated or empirically computed parameters will allow us to compute their confidence intervals (or p-values); and consequently, have an idea of how accurate are our conclusions inferred after computing the empirical copula or fitting a particular copula to our data.
3.1 Graphical Diagnosis

In the previous section we presented different methods to uncover the dependence structure among random variables. Now, we want to exploit this knowledge in order to show if what we infer is true: there exists some underlying dependence between the parameters $u$ and $v$ of our model.

Before any further analysis has been made, visual tools were used to check the presence of dependence. Firstly, after computing the histograms of a good portion of the model’s parameters stored in the different time series, we were able to infer the inherent distribution of each of them. In addition, we estimated (with some empirical procedure, which we are not going to describe since it escapes our purposes) the joint distribution functions between all the possible combinations of parameters. As we can appreciate in Fig. 3.1 a strong coupling between the traits $u$ and $v$ may exist as the value of the constrain parameters, $\theta_V$ and $\theta_P$, increases.
Figure 3.1: Joint distribution function of traits $u$ and $v$, host-parasite case. In the case where there is no plot is because the parasites went extinct.

Similarly, the scatter plot of the pairs $(u, v)$ and their corresponding ranks present in some occasions the presence of association. But that does not give enough evidence to declare the existence of correlation.

Besides scatter plots there exist other graphical tools to detect dependence, namely $\chi - plots$, Fisher et al.\cite{5}. They are based on the $\chi^2$ statistic for independence. Define

$$\chi_i = \frac{H_i - F_i G_i}{\sqrt{F_i(1 - F_i)G_i(1 - G_i)}}$$

and

$$\lambda_i = 4\text{sign}((F_i - 1/2)(G_i - 1/2))\max((F_i - 1/2)^2(G_i - 1/2)^2).$$

Values of $\chi_i$ that fall to far from 0 are indicative of departures from the independence
hypothesis. Looking at Fig 3.2 as a representation of the obtained plots we may say that they support our assumptions. Although not shown here, similar results were obtained in the predator-prey case. The heavy right tail of points which fall very far from the zero are indicative of departures from the independence hypothesis. In particular, they augur positive tail dependence, fact that supports the use of copula (which is good at detecting tail dependence). This tail dependence could mean that the strong dependence comes from extreme observations; thus, we may infer that when the traits start having values not close enough to their mean there exist a strong coupling.

![Figure 3.2: \(\chi\)-plot for \(\theta_V = \theta_P = 0.01, k_p = 1, \rho = 0.7996\). Host-Parasite case.](image)

Hence, all this graphs point to the existence of a relationship between these two variables.

### 3.2 Rank Based Statistics

To qualify the degree of dependence in the pair \((u, v)\), sample values of Spearman’s \(\rho\) and Kendall’s \(\tau\) were computed, using Eq. 2.2 and Eq. 2.4. Along with the \(p\)-values of the
associated tests of independence described by Eq. 2.3. In fact, we are only going to present the Spearman’s $\rho$ values since both measures lead to similar conclusions. For both models, predator-prey and host-parasite, in any of the situations defined by the different combinations of the parameters $\theta_V$, $\theta_P$ and $k_p$ the independence test rejected the null hypothesis with a confidence level of the 95% ($\alpha = 0.05$). This fact answers the first of the questions we wanted to explore using this model: it does exist coupling between predator-prey traits and also between host-parasite traits. To describe this dependence we should investigate in more detail the obtained $\rho$ values, as well as, the empirical copulas we constructed.

As we expected, and answering the second posed question, the results obtained for the predator-prey scenario differ from the ones obtained when dealing with the host-parasite model. Thus we are going to treat them separately.

### 3.2.1 Predator-Prey scenario

For the predator-prey case, first of all, mention the fact that when $\theta_V < \theta_P$ the predator’s population disappeared. Taking into account that predators’ population is rather small (compared to preys’ or parasites’ population) they must evolve slower; if we add to this situation the fact that the preys are less constrained, we may infer that the predators became extinct because prey’s $u$ values were far from $v$ values, making prey unsuitable for the predator).

Besides this connotation, from Fig. 3.3 (and much more we computed) it may be concluded: 1) for $\theta_V$ large and $\theta_P$ small, preys will be shaped by resource constraints rather than predation, thus predators and preys are somewhat independent; 2) for $\theta_V$ and $\theta_P$ equally constrained large negative values of $\rho$ are achieved when it becomes difficult for the predators to evolve (large $k_p$ value) and the preys’ $u$ trait evolve away from the $v$ trait; 3) for $\theta_V$ small and $\theta_P$ large, predators and preys are strongly dependent because predators are constrained by preys, but preys are not constrained by the resources, and again preys’ trait values keep
on evolving away from predators’ \( v \) value. However, predators are also likely to go extinct in this case because of a high \( k_p \) value and they become almost independent. Furthermore, when some kind of dependence is present in the predator-prey scenario, it is negative.

Figure 3.3: Spearman’s \( \rho \) for the Predator-Prey case
3.2.2 Parasite-Host scenario

On the other hand, like the predators, the parasites got extinct in some cases. In particular, when the hosts are less constrained than the parasites. We may infer that is because the parasites were not able to track properly the $u$ trait. In addition, if $k_p$ is too large (values above 2.75), the parasites disappeared whatever $\theta_V$ or $\theta_P$, inasmuch as parasites population is rather small and they evolve too slowly in order to survive.

In the host-parasite case we want to remark: first, the parasites population is much larger than the predators one; in addition, their conversion efficiency, $e_p$, is greatly increased, which implies that a single successful parasite attack may enable the parasite to produce many offspring; finally, the parasites intrinsic mortality rate, $k_p$, is greatly increased, in combination with the change in $e_p$, this reflects a notion of a parasites life history as being one which involves living “fast” and having a short lifespan. All that allows the parasites to evolve faster and with more success.

Fig. 3.4 reveals a positive dependence, decreasing in $k_p$ though, for the reason stated above. We already said that for the cases where $\theta_P > \theta_V$ the parasites got extinct, so we are going to focus on the cases with $\theta_P \leq \theta_V$. In particular we differentiate between $\frac{\partial \nu}{\partial \theta_P} = 1$ and $\frac{\partial \nu}{\partial \theta_P} \neq 1$. Results in Fig. 3.5 tell us that for $\frac{\partial \nu}{\partial \theta_P} \neq 1$ we may say that they are almost independent. On the other hand, if $\frac{\partial \nu}{\partial \theta_P} = 1$, when both the host and the parasite are almost unconstrained there is an attempt of the parasites to track the hosts but as $k_p$ increases $u$ evolves away from $v$. However, when both species are constraint there exist a high positive between them, although it slightly decreases with $k_p$. As we mentioned, parasites live “fast”, thus it is easier for the $v$ trait to evolve towards $u$. 
As mentioned above, one of the most interesting points in our results is that in the host-parasite, if it exists, the dependence is positive.

![Spearman's ρ as a function of kp, θ_p = θ_v = 0.01](image)

Figure 3.4: Spearman’s ρ as a function of kp, θ_p = θ_v = 0.01 for the Host-Parasite case. The norm of the residuals for the linear regression is 0.4821.
Figure 3.5: Spearman’s ρ for the Host-Parasite case

3.2.3 Extreme Cases

Because we were more curious, we were curious about what happened in the extinct cases, so we computed the rank statistics using the data we had before extinction, for both the predator-prey and host-parasite cases. The absolute value of $\rho_n$ turns out to be slightly higher than in the cases where there is no extinction, thus we can infer that for the extinct cases there exists a stronger dependence before the predators disappear, as Figs. 3.6 and 3.7 show, and also supports the tail dependence that the $\chi$-plots displayed. With that we
mean, that dependence, i.e. coevolution, is stronger when the traits have values far from the mean.

Figure 3.6: Spearman’s ρ as a function of $k_p$ for $\theta_p = 1$. Red solid lines represent extinct cases for the Predator-Prey case.

Figure 3.7: Spearman’s ρ as a function of $k_p$ for $\theta_p = 1$. Red solid lines represent extinct cases for the Host-Parasite case.
The possibility of the existence of this strong coupling increases the necessity of finding a copula that describes this correlation, in order to be able to sample data from those interesting cases.

3.3 Copula

After quantifying the dependence through Spearman’s $\rho$ the following step is to model the dependence structure between $u$ and $v$ using copula. First we are going to show one of the advantages of copula over a joint distribution function. The left picture of Fig. 3.8 represents the estimated joint distribution between the parasite $v$ trait and the host $u$ trait for $k_p = 1$ with $\theta_V = \theta_P = 0.01$, while the right plot corresponds to parasite $v$ trait and the host $u$ trait for $k_p = 1$ with $\theta_V = \theta_P = 1$. From this figure we could have inferred that the coupling in these two situations is completely different, since the joint distribution functions look quite different. The data from the joint distribution represented in the right panel is clearly concentrated around $(0, 0)$, while in the left panel seems to be more spreaded out. Nonetheless, Fig. 3.9 proves that the two empirical copulas are very similar (and the Spearman’s correlation coefficients have very close values too), and consequently they have a very similar dependence structures. Note that this result supports the ones we deduced from the rank based statistics.
Figure 3.8: The left panel represents the estimated joint distribution between the parasite $v$ trait and the host $u$ trait for $k_p = 1$ with $\theta_V = \theta_P = 0.01$, while the right plot corresponds to parasite $v$ trait and the host $u$ trait for $k_p = 1$ with $\theta_V = \theta_P = 1$.

Figure 3.9: The top-left plot corresponds to the 2D copula of the host $u$ trait for $k_p = 1$, $\theta_V = \theta_P = 0.01$, $\rho = 0.79956$, and the top-right shows the 2D copula of the host $u$ trait for $k_p = 1$, $\theta_V = \theta_P = 1$, $\rho = 0.6397$. The bottom pictures display the 3D versions.
In any case the visual analysis turns out to show that the computed empirical copulas resemble either the theoretical independent copula or one of the theoretical Fréchet bounds, Figs. 3.10 and 3.11.

In general, in the cases where \( \rho \) is almost zero (which indicates almost independence) the empirical copula resembles the left figure of Fig. 3.11, which looks like the theoretical independent copula, as we may infer looking at Fig. 3.10. While pairs of data sets with highly positive or highly negative Spearman’s \( \rho \) values have empirical copulas close to the middle and right plots of Fig. 3.11 respectively, which are close to the theoretical Fréchet copulas. From these results we may conclude that in those cases where we obtain an empirical copula close to the upper Fréchet Bound copula we can expect \( u \) to be almost surely an increasing function of \( v \) (or the other way round). On the contrary, if the empirical copula resembles the lower Fréchet Bound copula \( u \) is almost surely a decreasing function of \( v \). Moreover, we can use these empirical copulas to define the dependence structure between \( u \) and \( v \) under specific circumstances \( (k_p, \theta_V, \theta_P) \), and even being able to predict which is their coupling in situations where we do not have data. In addition, as we have already mentioned, since any copula is defined in the unit cube, we can compare the different dependence structures that \( u \) and \( v \) define depending on the different circumstances, as well as, different scenarios. This is to say, now we have a proper analysis tool to compare the predator-prey vs. parasite-host scenarios. Although in both scenarios we obtained empirical copulas that resembled either the independent copulas or one of the Fréchet Bounds, they did not have the exactly same shape, and consequently dependence structure, which supports the fact that predator-prey systems and host-parasite systems behave in a different manner. Even in the same case scenario we got different copula shapes, so we may infer that the same tritrophic system will evolve differently depending on the circumstances.
Figure 3.10: Theoretical copulas. From left to right: independent, upper Fréchet bound and lower Fréchet bound.

Figure 3.11: Empirical copulas. From left to right: independent, upper Fréchet bound and lower Fréchet bound.
3.4 Goodness of Fit

In order to give more evidence that supports our conclusions, we compute the norm 2 distance and the Kolmogorov-Smirnov tests between the empirical copulas and the theoretical independent, upper Fréchet Bound and lower Fréchet Bound copulas. Fig. 3.12 shows the results obtained in the cases represented in Fig. 3.4, where the Spearman’s $\rho$ is highly positive and decreases with $k_p$. As expected the norm 2 is smaller when comparing the empirical copula with the upper Fréchet Bound and increases with $k_p$ up to a certain point when the empirical copula is closer to the theoretical independent copula; in fact, for large $k_p$ $\rho$ is close to zero.

![Figure 3.12: Norm 2 distance between the independent, upper and lower Fréchet bounds copula and the empirical copula as a function of $k_p$, for $\theta_P = \theta_V = 0.01$. Host-parasite case.](image)

Figure 3.12: Norm 2 distance between the independent, upper and lower Fréchet bounds copula and the empirical copula as a function of $k_p$, for $\theta_P = \theta_V = 0.01$. Host-parasite case.
Next, we calculate the two samples Kolmogorov-Smirnov test for each of the three theoretical copulas, to compare the distributions of the values in the empirical data vectors and the data vector with data coming from one of the theoretical copulas. The null hypothesis is that both samples are from the same continuous distribution. The alternative hypothesis is that they are from different continuous distributions. The statistic it uses is $\max |C_{\text{emp}} - C_{\text{the}}|$. Nevertheless, the Kolmogorov-Smirnov test rejected the null hypothesis in all cases with a significance level of 0.05%. Table 3.1 show the results obtained for the data represented in Fig. 3.4. On one hand it reinforces that for small $k_p$ values the closest theoretical copula is the upper Fréchet Bound, and for large $k_p$ the independent copula is more appropriate. On the other hand the distance between the empirical copula and the theoretical ones is too large in order to say that they have the same dependence structure with an acceptable significance level ($p$-value). Thence, we may conclude that we should keep looking for other theoretical copulas that fit better our data, in order to be able to better define the underlying dependence structure.
| $k_p$ | $\text{max}|C_{\text{emp}} - C_{\text{Upp}}|$ | $\text{max}|C_{\text{emp}} - C_{\text{low}}|$ | $\text{max}|C_{\text{emp}} - C_{\text{Ind}}|$ |
|------|------------------|------------------|------------------|
| 1    | 0.1153           | 0.3848           | 0.1352           |
| 1.5  | 0.1122           | 0.3905           | 0.1422           |
| 1.7  | 0.1141           | 0.3881           | 0.1392           |
| 1.75 | 0.1143           | 0.3942           | 0.1455           |
| 1.8  | 0.1228           | 0.3868           | 0.1394           |
| 1.9  | 0.1328           | 0.3723           | 0.1231           |
| 2    | 0.1399           | 0.3603           | 0.1108           |
| 2.1  | 0.1537           | 0.3490           | 0.1039           |
| 2.2  | 0.1605           | 0.3397           | 0.0906           |
| 2.25 | 0.1743           | 0.3281           | 0.0815           |
| 2.3  | 0.1683           | 0.3330           | 0.0833           |
| 2.4  | 0.1619           | 0.3382           | 0.0883           |
| 2.5  | 0.1775           | 0.3280           | 0.0818           |
| 2.6  | 0.1801           | 0.3216           | 0.0728           |
| 2.7  | 0.1749           | 0.3275           | 0.0776           |
| 2.75 | 0.1960           | 0.3063           | 0.0622           |

Table 3.1: KS statistic as a function of $k_p$, for $\theta_V = \theta_P = 0.01$. Parasite-host case.
Since \( u \) and \( v \) were drawn from a Gaussian distribution, we may wonder if their dependence structure may fit a Gaussian copula (that does not imply a Gaussian joint distribution). Using the already computed values of the Spearman’s \( \rho \), we can generate new data coming from a Gaussian copula which has this particular \( \rho \) value and plot it against the original data. Fig. 3.13 displays the host-parasite case with \( \theta_V = \theta_P = 0.01, k_p = 1.75 \) and \( \rho = 0.7411 \). Observing the results, we may visually conclude that our data, at least in this particular case, may fit in a Gaussian copula. On the other hand, we also tested the Student \( t \) and the Clayton copula. The left picture of Fig. 3.14 shows the \( t \) copula against the original data, and the right plot corresponds to the Clayton copula vs. the original data, under the same circumstances as in the Gaussian case. From those plots we may infer that \( t \) and Clayton copula still fit quite well our data, but not as well as Gaussian does, since in both cases we can see some outliers from the simulated data. Of course, these visual observations do not give enough evidence to state that the underlying dependence structure corresponds to a Gaussian copula, so further test must be done in future work (KS test, for instance). But, in any case, what would this Gaussian copula structure imply? Or what would a Student \( t \) or Clayton copula structure mean? Without getting into very technical details we can summarize: 1) Gaussian copula implies no tail dependence, 2) \( t \) copula is tied to tail dependence, and 3) Clayton copula involves lower but not upper dependence (then coupling only would happen in the corner of the lower-left quadrant of a bivariate distribution. ). Then, accordingly to Fig.3.13 and 3.14 for the particular case of \( \theta_V = \theta_P = 0.01, k_p = 1.75 \) in the host-parasite scenario, there exists a positive codependence that may imply a comonotone association between \( u \) and \( v \) (due to the similarity with the upper Fréchet Bound copula), and this dependences happens for \( u \) and \( v \) values close to their mean (because of the no tail dependence implied by the Gaussian copula).
Figure 3.13: Original data vs. simulated data coming from a Gaussian copula with $\rho = 0.7411$, corresponding to the host-parasite case with $\theta_V = \theta_P = 0.01$ and $k_p = 1.75$.

Figure 3.14: Original data vs. simulated data coming from a $t$ (left) or Clayton (right) copula with $\rho = 0.7411$, corresponding to the host-parasite case with $\theta_V = \theta_P = 0.01$ and $k_p = 1.75$.

Although they are not showed here, in Chapter 2 we have already presented the Confidence Intervals that can be computed depending on the parameter we have estimated to compute those copulas.
Chapter 4

Conclusions and Future Work

4.1 Conclusions

There exists coupling between $u$ and $v$ in both scenarios. Consequently, there is co-evolution both between predators and preys and between hosts and parasites. We have been able to compute an independence test, supported by rank based statistics, which supported the dependence hypothesis with a 95% of confidence level. However, from the computed values of the Spearman’s $\rho$ we should expect a different behavior for each scenario. In particular, for the predator-prey system the dependence turns out to be negative. On the other hand, for the host-parasite case the dependence is positive.

An important aspect of this work is the fact that we can go one step beyond the $\rho$ coefficient, which only gives us a number. We can compute an empirical copula. Through that we can construct a dependence structure that allows us to interpret the codependence between two variables. In addition, we can try to fit one of the well-known copula to our data, so we can explain even better the codependence between the variables. For instance, with our data we could say that for high positive $\rho$ values there exists and underlying comonotone relationship; while for negative $\rho$ the dependence structure describes a countermonotone association, because of the similarity with the Fréchet Bounds copulas.
Visually these conclusions were clear and the norm 2 and the Kolmogorov-Smirnov tests supported our hypothesis. Notwithstanding the results did not provide enough evidence to get to some strong conclusion about the real underlying copula structure.

We also tried to fit other copula families. We tested the Gaussian, Student’s $t$ and Clayton copula with one of the particular cases of the parasite-host scenario. Visually, we may infer that the Gaussian copula is the one that better fits our data. In the future we will use some test (KS, for instance) to corroborate this assumption. Assuming that the Gaussian copula defines the underlying structure we may say that this comonotone dependence happens for $u$ and $v$ close to their means, since this kind of copula is associated with non-tail dependence. Nevertheless, it is possible that for other circumstances (different $\theta_P$, $\theta_V$ or $k_p$ values) different copulas will fit better the data.

Finally, we note that in both the rank based statistics and copula we can calculate the accuracy of the empirical estimations. The distribution of all the estimated coefficients and parameters are known, therefore we can establish a confidence interval for them.

## 4.2 Future Work

As future work we must be focused on designing an effective method to find the best copula fit; the reason we would do that is that being able to sample from the suitable theoretical copula would allow us to study extreme cases, for example cases where the predators (parasites) become extinct.
On the other hand, we are also very interested in tracking the dynamics of the copula. In this work we have assumed stationarity, and that the system was in a steady mode. In other words, it was in equilibrium. We would like to discover how the empirical copula evolves through time, or to identify the evolution of the parameters that describe the copula that fits our data as function of time. Probably at the dependence structure at the beginning differs from the structure defining the coupling at equilibrium. Thus, it would be interesting to track the dynamics of these dependence structures.
Bibliography


[8] C. Genest and A. C. Favre. “Everything you always wanted to know about copula modeling but were afraid to ask”. Dépt. de mathém. et de stat., Univ. Laval, Québec, Canada; Hydr. Stat., INRS, Québec, Canada; 2007.


