Delft University of Technology
Faculty Electrical Engineering, Mathematics and Computer Science
Delft Institute of Applied Mathematics

Dependence Measures in Citation Analysis

The application of parametric copulas to capture the dependence structure between the publications of a reseacher and the citations of those publications.

(Dutch title: Afhankelijkheidsmaten in de Citatie Analyse

De afhankelijkheidsstructuur tussen het aantal publicaties van een onderzoeker en het aantal citaten van de desbetreffende publicaties vastleggen via parametrische copulas.)

A thesis submitted to the
Delft Institute of Applied Mathematics
as part to obtain

the degree of

BACHELOR OF SCIENCE
in
Applied Mathematics

by

Ashni Bachasingh

Delft, the Netherlands
December 2018

BSc thesis APPLIED MATHEMATICS

"Dependence Measures in Citation Analysis

The application of parametric copulas to capture the dependence structure between the publications of a researcher and the citations of those publications."

(Dutch title: "Afhankelijkheidsmaten in de Citatie Analyse

De afhankelijkheidsstructuur tussen het aantal publicaties van een onderzoeker en het aantal citaten van de desbetreffende publicaties vastleggen via parametrische copulas.")

Ashni Bachasingh

Delft University of Technology

Supervisor

Dr.Ir. G.F. Nane

Other members of the committee

Dr. D.C. Gijswijt

Drs. E.M. van Elderen

December, 2018

Delft

Abstract

In this thesis we try to capture the dependence structure of the publications of a scholar and the citations of those publications via copulas. To do so, we will use a sample of Quebec researchers for who their publication amount as well as their citation amounts are known. We are provided with multiple variables concerning citation. We study the dependence structure between these variables, with the aim of fitting copulas to this structure, by calculating correlation scores and visualising the structure. Copulas are functions that "join together" one-dimensional distribution functions with a dependence structure, in order to represent joint distributions. The correlation scores are calculated across various ranges of the variables to provide us with a deeper understanding of the dependence structure between the variables.
Using Sklar's theorem and some helpful functions in various packages in the software program **R**, parametric copulas fit the dependence structures of the various pairs of variables. Based on a *Goodness-of-fit* test, certain parametric copula models are rejected at a 5% significance level. Unsurprisingly, there are also dependence structures that can be well captured with a parametric copula.
Parametric copula families are not only used for fitting the data, but also for prediction. Since a good fitting model does not necessarily imply a good predictive model, we have also performed a validation analysis. The parametric copula models that are not rejected by the test at a 5% significance level are validated via k-fold cross validation. Part of the data have been used to fit the model and the remaining has been validated using a k-fold cross validation. It turns out that the best fitting copula model does not always perform well in term of prediction. That is, these copulas do not always perform best during the cross-validation.

Contents

1 Introduction

The literary idea of a copula arose in the 19th century. This was based on the multivariate cases of non-normality. In 1959 Abe Sklar first employed the word copula in a mathematical or statistical sense in the theorem which now bears his name. The theorem describes how the joint distribution can be specified in terms of the marginal distribution and the copula function, as we will see in the next chapter. The notion of copulas became increasingly popular at the end of the nineties. At this time, researchers in the applied field of finance discovered the notion of the copula. This lead to a wealth of investigations about copulas, especially the applications of the copulas.

1.1 Copulas in finance and insurance

In the financial world, financial risk management is currently a hot topic. When practicing financial risk management risks are measured and managed across a diverse range of activities used in, e.g., banking, securities and insurance sectors. The dependencies between random variables, such as risks, credit scores, etc., play an important role here.
The correlation coefficient is a popular and often used dependence measure within Financial Risk Management. This is a good measure when the random variables are multivariate normally distributed. It is a reasonable measure when the random variables are elliptically distributed. However, in practice most of the data, thus the variables, are not multivariate normally distributed, nor elliptically distributed.
In order to obtain an optimal portfolio selection, the Capital Asset Pricing Model (CAPM) and the Arbitrage Pricing Theory (APT), Embrechts(1999), employ an elegant theory which is essentially founded on the assumption of normally distributed returns. The assumption is applied in many risk management applications, because this distribution is easy to implement. However, research has shown that the distribution does not account for the thickness of the tails of the marginals and their dependence structure. These assumptions are more problematic in insurance because of the typically skewness and heavy-tailedness of insurance claim data.
As a solution to this problem copulas are used. Copulas have become an appropriate dependence measure when the random variables are not multivariate normally distributed nor elliptically distributed. The copula manoeuvres around the pitfalls of correlation. This has resulted in it's popularity in Financial Risk Management to model dependencies between risks. Some of the areas of applications are credit risk modelling, portfolio Value at Risk calculations, default and credit risk dependence, and tail dependence.
A copula accurately describes the dependence relationship only if the right copula is applied. During the late 1990's the CDOs appeared on the financial market. This was a new financial derivative called Collateralized Debt Obligations. Banks were allowed to form securities out of different types of debts, e.g. mortgages, via these derivatives. The correlation between defaults needed to be modelled in order to price these securities. David X. Li's Gaussian copula approach was used to model just that. The Gaussian copula, which will be introduced in paragraph 2.3, is a helpful tool and relatively easy to fit. However, the Gaussian copula does not capture tail dependencies. Risks in the tail are underestimated, but cases where the risks are simultaneously in the tails of each distribution, are seen as highly improbable. In 2008 the crisis hit Wall street and the CDO market collapsed. Li's Gaussian copula model has been accused of increasing the severity of the financial crisis (Felix Salmon, 2009).
The flaws of the Gaussian copula, as well as many other copula models, are documented, see for example Salmon (2009), Forsland (2012) and Embrechts (2009). The successes and failures in the financial field on copulas are of great help in other fields where an increasing effort is made

in accurately depicting the overall dependencies, as well as tail dependencies. Some examples include Hydrology, Medicine, Biology and Epidemiology.

1.2 Citation Analysis

Part of the evaluation of researchers consists of quantifying the research output, by the size, impact and quality of their research output, as well as the citation impact of the research output. Academic institutions use publications counts as well as the subjective opinions of peers in order to evaluate researchers. Committees for hiring, promoting or evaluating tenure trackers also rely on citation analysis to obtain a more objective assessment of a researchers work.
Current citation analysis is most often used to couple a quantitative indicator to an evaluation of research performance. Most of the current research in citation analysis assumes a linear relationship between bibliometric indicators. Indicators reflect number of publications, citations, international collaboration, etc. But citation patterns vary greatly between disciplines, publication types, authors, etc.
The current models are regression-based models. Methods used so far include ordinary least squares linear regression, logistic regression, a distribution-free regression method, multinomial logistic regression and negative binomial regression, see Thelwall. In absence of alternatives, citation counts have been investigated with negative binomial regression. Just as in the financial field, the variables in citation analysis are highly skewed. The distribution of citations, for example, is highly skewed, Thelwall. So tests based on the normal distribution, that is ordinary least squares regression are not appropriate. The negative binomial regression can cope with the skewed data, but as said before this selection is based in the absence of alternatives.

1.3 Goal of the Research

The focus in this thesis lies in modelling the dependence structure between the publications of a researcher and the citations of those publications. We will attempt to capture this dependence structure with the help of parametric copulas. Our data, as can be seen in chapter 4, is highly skewed. Copulas have been used in the financial field to cope with highly skewed data, as mentioned before. Therefore, the copula seemed like a promising tool to capture the dependence structure. This leads to the main question throughout this thesis, which is:

'How well do parametric copulas capture the dependence structure between the publications of a researcher and the citations of those publications?'.

We try to obtain the answer to this question by answering the following sub questions: What is a copula function and what are its properties? Are copulas a better representation of the dependence structure in citation analysis? And how well do these copula models perform in terms of prediction?

1.4 Outline of the Thesis

To answer the questions mentioned in the previous paragraph, we will need to take appropriate steps. First, we will discuss some well known dependence measures, such as correlation. Precise definitions are given and some useful theorems are noted. Followed by the notion of the copula. Subsequently, definitions are given and useful properties and theorems are discussed. We end chapter 2 with a list of non-parametric and parametric copulas.
In chapter 3 we introduce our dataset. The variables used in this thesis are defined and some

background information on the data set is given.
In chapter 4 we analyse our variables by looking at various descriptive statistics and various plots concerning the structure of the variables, like density- and boxplots. The rank correlation values between all the pairs of variables are calculated. This is also done for the various fields in our data set. To create a deeper analysis, the data is split into bins and rank correlation values for the bins are calculated as well.
Chapter 5 discusses the copula fitting. The software model used to fit parametric copula families is discussed followed by an analysis on the outcome of the fitting. The fitted copulas are tested with a *Goodness of fit* test, more on this in paragraph 5.1. The results of the fitting and the test are then analysed. The copulas that are rejected by the test, at a 5% significance level, are compared with the empirical copula, which is why we end this chapter by evaluating some of these empirical copulas.
In chapter 6 the copula models that seem to capture the dependence structure between the variables rather nicely, according to the test, are validated via a k-fold cross validation. This chapter will also elaborate on the notion of model validation.
Lastly, we consider our results and discuss the fitting and performance of copulas. In the Appendices a table with the main descriptive values of the data set can be found. Furthermore, it includes the output of the copula fitting when the ties in our data set are unaccounted for and the R codes used during the fitting.

2 Dependence measures

An important aspect of many statistical investigations is the stochastic dependence between random measurements. Why? Firstly, suppose (X_i, Y_i), $i = 1, \cdots, n$ is a random sample from a bivariate population with joint distribution function $F(X, Y)$ and marginal distribution functions F_1, F_2. Let $X_1 \leq \cdots \leq X_n$. The first motivation behind the construction of the dependence measure is to ascertain the degree of conformity of this ascending order with respect to variable Y. These measures are insensitive to monotone transformations of X and Y and non-parametric. Most of these measures are used as test statistics for testing the hypothesis of independence. The second motivation behind dependence measures is prediction. Dependence measures are used to predict one variable from another. The thought and desire behind this is that if X is closer to a function of Y, then the measure should be higher. In this chapter, we will discuss several dependence measures.
Correlation is by far the best known dependence measure. In fact, in financial theory the notion of correlation is central. Take modern portfolio theory for example. Correlation coefficients are used as a measure between the returns of different assets. The assets that are less likely to lose value at the same time are selected. Thus, we begin by discussing correlation. Here we distinguish between linear correlation and rank correlation. Before introducing the copula, an alternative, more recent and less well known way to describe dependence compared to correlation, it is sensible to discuss some basic notions, such as the definition of a distribution function and the quasi-inverse of a function. These play a fundamental role when working with copulas. Followed by the notion of 2-increasing function, which is needed to define the copula. Then the copula is defined and some useful theorems concerning the copula are given. The proofs of the theorems in this chapter can be found in Schweizer and Sklar (1983), Nelson (2006) and Joe (2015). Almost all of our definitions, properties and theorems, especially the ones directly referring to copulas, consider two dimensions. This is because we confine ourselves to the two-dimensional copulas in this thesis. Evidently, all definitions, properties and theorems in this chapter can be extended to n dimensions, see Nelson (2006). Finally, a list of copula families is given, which serves as an aid in copula selection.

2.1 Correlation

Linear correlation, also known as Pearson's correlation, is most frequently used in practice as a measure of dependence. There are several reasons behind the popularity of linear correlation. Firstly, linear correlation is often straightforward to calculate. Especially compared to for example the calculation of comonotonicity and rank correlation. A second motivation behind the use of linear correlation is the fact that it is easy to manipulate under linear operations. This fact is commonly exploited in portfolio theory. Another reason is that linear correlation is a natural measure of dependence in multivariate normal distributions, since the correlation coefficient completely defines the dependence structure of said distribution. However, correlation has some shortcomings. We repeat two of the several mentioned fallacies by Embrechts et al. (2002), according to Joe (2014), concerning linear correlation.
Fallacy 1. Marginal distributions and correlation determine the joint distribution.
Fallacy 2. Given marginal distributions F_1 and F_2 for X_1 and X_2 respectively, all linear correlations between -1 and 1 can be attained through suitable specification of the joint distribution.
Furthermore, linear correlation is not preserved by copulas. Which means that two pairs of correlated variables with the same copula can have different correlation coefficients.
Kendall's τ and Spearman's ρ are the two most common measures of association, when working

with copulas. The reason rank correlation is chosen over linear correlation is because they are invariant under monotonic transformations and it captures monotonic rather than linear dependence.
In this paragraph we define *Spearman's rank correlation* and *Kendall's rank correlation.* We close this paragraph with the definition of the *coefficient of tail dependence*, which is also an important notion when working with copulas.

Recall that a *random variable* is a function from a sample space to the real numbers. The number produced by the function is random because the outcome of the experiment with said sample space is random. For example, we call X random variable where X defines the total number of heads observed during a sequence of coin tosses. Consider a pair of real-valued random variables (X, Y) with finite variances, which are not constant.

Definition 2.1.1. ***Pearson's linear correlation coefficient*** *between X and Y is*

$$\rho(X,Y) = \frac{Cov[X,Y]}{\sqrt{\sigma^2[X]\sigma^2[Y]}},$$

where Cov[X,Y] is the covariance between X and Y and $\sigma^2[X]$, $\sigma^2[Y]$ denote the variances of X and Y.

Definition 2.1.2. *Let X and Y be random variables with distribution functions F and G and a joint distribution function H.* ***Spearman's rank correlation*** *is given by*

$$\rho_S(X,Y) = \rho(F(X), G(Y))$$

where ρ is Pearson's linear correlation.

Definition 2.1.3. *Let (x_i, y_i) and (x_j, y_j) denote two oberservations from a vector (X, Y) of continuous random variables. We say that (x_i, y_i) and (x_j, y_j), $i, j = 1, \ldots, n$, are* ***concordant*** *if $x_i < x_j$ and $y_i < y_j$ or if $x_i > x_j$ and $y_i > y_j$.*

Similarly we say that (x_i, y_i) and (x_j, y_j) are **discordant** if $x_i < x_j$ and $y_i > y_j$ or if $x_i > x_j$ and $y_i < y_j$.

Definition 2.1.4. *Assume (X, Y) are continuous random variable with a joint distribution H. Let (X_1, Y_1) and (X_2, Y_2) be two independent pairs of random variables from joint distribution function H, then* ***Kendall's rank correlation*** *is given by*

$$\rho_\tau(X,Y) = \mathbb{P}[(X_1 - X_2)(Y_1 - Y_2) > 0] - \mathbb{P}[(X_1 - X_2)(Y_1 - Y_2) < 0].$$

Definition 2.1.5. *Let X and Y be random variables with distribution functions F and G. The* ***coefficient of upper tail dependence*** *of (X, Y) is*

$$\lim_{\alpha \to 1} \mathbb{P}[Y > G^{-1}(\alpha) | X > F^{-1}(\alpha)] = \lambda$$

provided a limit $\lambda \in [0, 1]$ exists. If $\lambda \in (0, 1]$ X and Y are said to be asymptotically dependent (in the upper tail). If $\lambda = 0$ they are asymptotically independent. Similarly, the ***coefficient of lower tail dependence*** *of X and Y is*

$$\lim_{\alpha \to 0} \mathbb{P}[Y < G^{-1}(\alpha) | X < F^{-1}(\alpha)] = \lambda.$$

2.2 Basic notions

As said before, distribution functions and the quasi-inverse of a function plays a fundamental role when working with copulas, as will be seen in paragraph 2.4. Therefore, it seems sensible to begin by introducing the definition of a *distribution function*.
But before citing the first definition, we introduce some notation. These notations are the same notations Nelsen (2006) used in *"An introduction to copulas"*. Just like Nelsen, we let $\mathbb{R}$ denote the ordinary real line $(-\infty, \infty)$. $\bar{\mathbb{R}}$ denotes the extended real line $[-\infty, \infty]$. So $\bar{\mathbb{R}}^2$ denotes the extended real plane $\bar{\mathbb{R}} \times \bar{\mathbb{R}}$. A rectangle in $\bar{\mathbb{R}}^2$ is the Cartesian product A of two closed intervals: $A = [x_1, x_2] \times [y_1, y_2]$. The *vertices* of a rectangle A are the points $(x_1, y_1), (x_1, y_2), (x_2, y_1)$ and (x_2, y_2). The unit square $\mathbb{I}^2$ is the product $\mathbb{I} \times \mathbb{I}$ where $\mathbb{I} = [0, 1]$. A *2-place real function* H is a function whose domain, denoted as Dom H, is a subset of $\bar{\mathbb{R}}^2$ and whose range, denoted as Ran H, is a subset of $\mathbb{R}$.

Definition 2.2.1. *We say that F is the* ***cumulative distribution function*** *of the random variable X when for all $x \in \bar{\mathbb{R}}$: $F(x) = \mathbb{P}[X \leq x]$.*

Note that some properties of the cumulative distribution function are.

1. F is nondecreasing, i.e. $F(a) \leq F(b)$ for all $a \leq b$ where $a, b \in \mathbb{R}$,
2. $F(-\infty) = 0$ and $F(\infty) = 1$

The notation $X \sim F$ mean that the random variable X has distribution function F.

Theorem 2.2.1. *Let X be a random variable with distribution function F. Let F^{-1} be the quasi inverse function of F, i.e.*

$$F^{-1}(\alpha) = inf\{x | F(x) \geq \alpha\},$$

$\alpha \in (0, 1)$. Then

1. *For any for any uniformly distributed random variable $U \sim U[0, 1]$, we have that $F^{-1}(U)$ has distribution function F. This gives a simple method for simulating random variates with distribution function F.*
2. *If F is continuous, then $F(X) \sim U[0, 1]$.*

2.3 Preliminaries

A *2-increasing function* is a two-dimensional analog of a nondecreasing function of one variable. Naturally, a precise definition is desired. The definition builds on the definition of the *H-volume of a rectangle.*

Definition 2.3.1. *Let S_1 and S_2 be nonempty subsets of $\bar{\mathbb{R}}$, and let H be a two-place real function such that Dom $H = S_1 \times S_2$. Let $A = [x_1, x_2] \times [y_1, y_2]$ be a rectangle all of whose vertices are in Dom H. Then the* ***H-volume of A*** *is given by*

$$V_H(A) = H(x_2, y_2) - H(x_2, y_1) - H(x_1, y_2) + H(x_1, y_1).$$

Definition 2.3.2. *A 2-place real function H is* ***2-increasing*** *if $V_H(A) \geq 0$ for all rectangles A whose vertices lie in Dom H.*

Next to the definition of a *2-increasing function*, the definition of a *grounded function* is also needed before the copula can be defined. Furthermore, 2-increasing and grounded functions also help us define *joint distribution functions*, which is what a 2-dimensional copula is.

Definition 2.3.3. *Suppose that S_1 has a least element a_1 and that S_2 has a least element a_2. We say that a function H from $S_1 \times S_2$ into $\mathbb{R}$ is* ***grounded*** *if*

$$H(x, a_2) = 0 = H(a_1, y) \text{ for all } (x, y) \in S_1 \times S_2.$$

Hence, we have

Definition 2.3.4. *A* ***joint distribution function*** *is a function H with domain $\bar{\mathbb{R}}^2$ such that*

1. *H is 2-increasing*
2. *$H(x, -\infty) = H(-\infty, y) = 0$ and $H(\infty, \infty) = 1$*

Thus H is grounded. Because Dom $H = \bar{\mathbb{R}}^2$, H has margins F and G given by $F(x) = H(x, \infty)$ and $G(y) = H(\infty, y)$.

2.4 Copulas

Definition 2.4.1. *An* ***two-dimensional copula*** *(or 2-copula) is a function C whose domain is $\mathbb{I}^2$ with the following properties:*

- *C is grounded and 2-increasing*
- *For every $(u, v) \in \mathbb{I}^2$,*

$$C(u, 1) = u \text{ and } C(1, v) = v.$$

Now that we've defined the copula, we'll go through some useful properties.

Theorem 2.4.1. *Let C be a copula. Then for every $(u_1, u_2), (v_1, v_2))$ in Dom C:*

$$|C(u_2, v_2) - C(u_1, v_1)| \leq |u_2 - u_1|_|v_2 - v_1|$$

Hence C is uniformly continuous on its domain.

Theorem 2.4.2. *Let X and Y be continuous random variables. Then X and Y are independent if and only if $C_{XY} = \Pi$.*

Here Π denotes the **product copula** $\Pi(u, v) = uv$.

Theorem 2.4.3. *Let X and Y be random variables with continuous distribution functions F and G, joint distribution function H and copula C. Then the following are true:*

1. *$\rho_S(X, Y) = \rho_S(Y, X)$, $\rho_\tau(X, Y) = \rho_\tau(Y, X)$*
2. *If X and Y are independent then $\rho_S(X, Y) = \rho_\tau(X, Y) = 0$*
3. *$-1 \leq \rho_S(X, Y), \rho_\tau(X, Y) \leq 1$*
4. *$\rho_S(X, Y) = 12 \int_0^1 \int_0^1 \{C(x, y) - xy\} dx dy$*
5. *$\rho_\tau(X, Y) = 4 \int_0^1 \int_0^1 C(u, v) dC(u, v) - 1$*
6. *For $T : \mathbb{R} \longrightarrow \mathbb{R}$ strictly monotonic on the range of X, both ρ_S and ρ_τ satisfy property 4*
7. *$\rho_S(X, Y) = \rho_\tau(X, Y) = 1 \Leftrightarrow C = C_u \Leftrightarrow Y = T(X)$ a.s. with T increasing*
8. *$\rho_S(X, Y) = \rho_\tau(X, Y) = -1 \Leftrightarrow C = C_l \Leftrightarrow Y = T(X)$ a.s. with T decreasing*

Sklar's theorem seems like a sensible follow up. This is perhaps the most important result regarding copulas. The theorem is used in essentially all applications of copulas.

Theorem 2.4.4. ***Sklar's theorem*** *Let H be a joint distribution function of the random variables X and Y with margins F and G. Then there exists a copula C such that for all x, y in $\bar{\mathbb{R}}$,*

$$H(x, y) = C(F(x), G(y)). \tag{1}$$

The copula C is uniquely defined on Ran $F\times$ Ran G and is therefore unique if all the marginals are continuous. Conversely, if C is a copula and F and G are joint distribution functions, then the function H defined through equation (1) is a joint distribution function with margins F and G.

Corollary 2.4.1. *Let H be a joint distribution function with continuous margins F and G and let C be a copula. Let $F^{(-1)}$ and $G^{(-1)}$ be quasi-inverses of F and G, respectively. Then for any $(u, v) \in \mathbb{I}^2$:*

$$C(u, v) = H(F^{(-1)}(u), G^{(-1)}(v)).$$

Without the continuity assumption, care has to be taken; see Nelsen (1999).

Example 2.3.1 Let H be a joint distribution function:

$$H(x, y) = \begin{cases} \frac{(x+1)(e^y-1)}{x+2e^y-1}, & (x, y) \in [-1, 1] \times [0, \infty], \\ 1 - e^y, & (x, y) \in (1, \infty] \times [0, \infty], \\ 0 & \text{elsewhere.} \end{cases}$$

With margins F and G given by

$$F(x) = \begin{cases} 0, & x < -1 \\ \frac{(x+1}{2}, & x \in [-1, 1], \\ 1, & x > 1 \end{cases} \qquad G(y) = \begin{cases} 0, & y < 0 \\ 1 - e^{-y}, & y \geq 0 \end{cases}$$

The quasi-inverses of F and G are given by $F^{(-1)}(u) = 2u - 1$ and $G^{(-1)}(v) = -ln(1 - v)$ for $(u, v) \in \mathbb{I}$. Via corollary 2.4.1 we obtain the copula C given by

$$C(u, v) = \frac{uv}{u + v - uv}.$$

There are many parametric copula families that describe the dependence between random variables. Some important families will be included in the next paragraph. We end this paragraph with another useful copula property, the *The Fréchet-Hoeffding Bounds.*

Theorem 2.4.5. ***The Fréchet-Hoeffding Bounds*** *Let C be a copula. Then for every $(u, v) \in \mathbb{I}^2$:*

$$W(u, v) = max(u + v - 1, 0) \leq C(u, v) \leq min(u, v) = M(u, v).$$

We refer to M as the ***Fréchet-Hoeffding upper bound*** *and W als the* ***Fréchet-Hoeffding lower bound****.*

These bounds help with the compution of the copula, as can be seen in Sempi (2011, p86).

2.5 Non-parametric copulas and parametric copula families

In this chapter we create a list of the most common bivariate parametric copula families. We include useful properties of parametric copula families which serve as an aid in model selection.

Name	Generator function	Parameter range	Kendall's τ	Tail dependence (lower, upper)
Gaussian		$\rho \in (-1,1)$	$\frac{2}{\pi}\arcsin(\rho)$	0
Student-t		$\rho \in (-1,1)$, $\nu > 2$	$\frac{2}{\pi}\arcsin(\rho)$	$2t_{\nu+1}(-\sqrt{\nu+1}\sqrt{\frac{1-\rho}{1+\rho}})$
Clayton	$\frac{1}{\theta}(t^{-\theta}-1)$	$\theta > 0$	$\frac{\theta}{\theta+2}$	$(2^{-\frac{1}{\theta}}, 0)$
Gumbel	$(-\log t)^{\theta}$	$\theta \geq 1$	$1-\frac{1}{\theta}$	$(0, 2-2^{\frac{1}{\theta}})$
Frank	$-\log[\frac{e^{-t\theta}-1}{e^{-\theta}-1}]$	$\theta \in \mathbb{R}\setminus\{0\}$	$1-\frac{4}{\theta}+4\frac{\int_0^{\theta}\frac{c/\theta}{e^x-1}dx}{\theta}$	$(0,0)$
Joe	$-\log[1-(1-t)^{\theta}]$	$\theta > 1$	$1+\frac{4}{\theta^2}\int_0^1 t\log(t)(1-t)^{2(1-\theta)/\theta}dt$	$(0, 2-2^{\frac{1}{\theta}})$
BB1	$(t^{-\theta}-1)^{\delta}$	$\theta > 0, \delta \geq 1$	$1-\frac{2}{\delta(\theta+2)}$	$(2^{-\frac{1}{\theta\delta}}, 2-2^{\frac{1}{\delta}})$
BB6	$(-\log[1-(1-t)^{\theta}])^{\delta}$	$\theta \geq 1, \delta \geq 1$	$1+\frac{4}{\delta\theta}\int_0^1(-\log(1-(1-t)^{\theta})$ $\times(1-t)(1-(1-t)^{-\theta}))dt$	$(0, 2-2^{\frac{1}{\delta\theta}})$
BB7	$(1-(1-t)^{\theta})^{-\delta}-1$	$\theta \geq 1, \delta > 0$	$1+\frac{4}{\delta\theta}\int_0^1(-(1-(1-t)^{\theta})^{\delta+1}$ $\times\frac{(1-(1-t)^{\theta})^{-\delta}-1}{(1-t)^{\theta-1}})dt$	$(2^{-\frac{1}{\delta}}, 2-2^{\frac{1}{\theta}})$
BB8	$-\log[\frac{1-(1-t\delta)^{\theta}}{1-(1-\delta)^{\theta}}]$	$\theta \geq 1, \delta \in (0,1]$	$1+\frac{4}{\delta\theta}\int_0^1(-\log(\frac{(1-t\delta)^{\theta}-1}{(1-\delta)^{\theta}-1})$ $\times(1-t\delta)(1-(1-t\delta)^{-\theta}))dt$	$(0,0)$

Table 1: List and properties of most common bivariate copula families.

Let Φ denote the standard univariate normal distribution function and let Φ_R^n denote the standard multivariate normal distribution function with linear correlation matrix R. Then

$$C(u_1, \cdots, un) = \Phi_R^n(\Phi^{-1}(u_1), \cdots, \Phi^{-1}(u_n))$$

is the Gaussian or normal n-copula.

Similarly, let t denote t the student t distribution function and let $t_{\nu,\Sigma}$ denote the multivariate student t distribution function with Σ the covariance matrix and ν the degrees of freedom. Then

$$C_{\nu,\Sigma}^t(u_1, \cdots, un) = t_{\nu,\Sigma}(t_{\nu,\Sigma}^{-1}(u_1), \cdots, t_{\nu,\Sigma}^{-1}(u_n))$$

is the student-t n-copula.
The remaining copulas in table one can take the following form,

$$C(u_1, \cdots, u_n|\theta) = \psi^{-1}(\psi(u_1|\theta) + \cdots + \psi(u_n|\theta))$$

where $\psi(u|\theta)$ is called the generator function and θ represents the parameters of the copula. Copulas with this form are called Archimedean Copulas. Note that these copulas usually cover the bivariate cases. When working with more than two variables, the Gaussian or Student t copulas are most frequently used.
Another useful copula is the Tawn copula. In 1988, Tawn added two additional parameters to the Gumbel Copula. This was the solution to the not so reasonable assumption of symmetry. In some applications the Gumbel Copula did not satisfy $C(u_1, u_2) = C(u_2, u_1)$. So random variables, say X_1 and X_2, modelled by C were not exchangeable. Tawn copulas belong to the *extreme values copulas.*

Definition 2.5.1. *Let $A : [0,1] \to [\frac{1}{2}, 1]$ be a convex function satisfying $max(w, 1-w) \leq A(w) \leq 1$ for all $w \in [0,1]$. The extreme value copula due to Pickands (1981) is defined by*

$$C(u_1, u_2) = \exp\left[\log(u_1 u_2) A(\frac{\log u_2}{\log(u_1 u_2)})\right].$$

Independence corresponds to $A(w) = 1$ for all $w \in [0,1]$. Complete dependence corresponds to $A(w) = max(w, 1-w)$.

The Tawn Copula, also known as the asymmetric logistic model, is generated by

$$A(w) = 1 - (\theta + \phi)w + \theta w^2 + \phi w^3$$

where $w \in [0,1], 0 \leq \phi_1, \phi_2 \leq 1$ and $\theta \in [0, \infty]$. This is also known als the Tawn Type 1 Copula. The Tawn Type 2 Copula is generated by

$$A(w) = (1-\phi_1)(1-w) + (1-\phi_2)w + [(\phi_1 w)^{\frac{1}{\theta}} + (\phi_2(1-w))^{\frac{1}{\theta}}]^{\theta}$$

where $w \in [0,1], 0 \leq \phi_1, \phi_2 \leq 1$ and $\theta \in [0, \infty]$. More on the Tawn Copula, it's properties and extreme-value copulas can be found in (Eschenburg, 2013).
In addition to these families, rotated versions of the last 8 copulas exist. The survival copulas refure to the copulas rotated by 180 degrees.

Definition 2.5.2. *We define the **survival copula** as a function $\hat{C} : \mathbb{I}^2 \longrightarrow \mathbb{I}$ defined by*

$$\hat{C}(u, v) = u + v - 1 + C(1-u, 1-v) = \mathbb{P}[U_1 > u_1, U_2 > u_2]$$

for all $(u, v) \in \mathbb{I}^2$.

We can also rotate the copulas by 90 and 270 degrees. The distribution functions of the rotated copulas are given as follows:

$$\begin{aligned} C_{90}(u_1, u_2) &= u_2 - C(1-u_1, u_2), \\ C_{270}(u_1, u_2) &= u_1 - C(u_1, 1-u_2). \end{aligned}$$

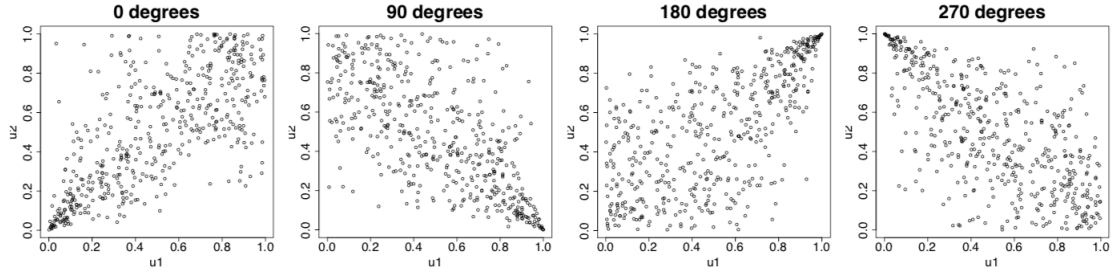


Figure 1: Samples from Clayton copulas rotated by 0, 90, 180 and 270 degrees with parameters corresponding to Kendall's τ values of 0.5 for positive dependence and -0.5 for negative dependence.

We end this paragraph with the *empirical copula*.

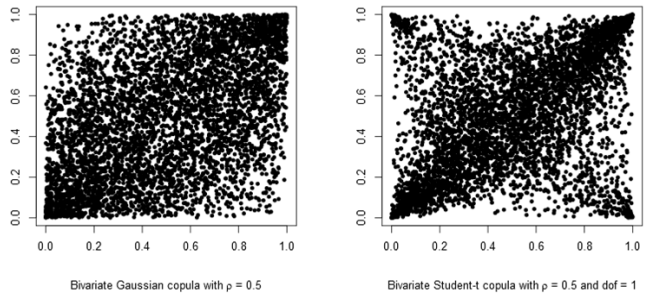
Definition 2.5.3. *Let* $\{(x_k, y_k)\}_{k=1}^m$ *denote a sample of size* m *from a continuous bivariate distribution. The* ***empirical copula*** *is a function* EC_m *give by*

$$EC_m(\frac{i}{m}, \frac{j}{m}) = \frac{\#\{(x,y)|x \leq x_{(i)}, y \leq y_{(j)}\}}{m}$$

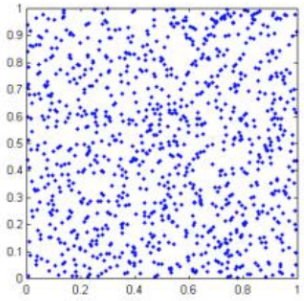
where the pair (x, y) *counted are from the sample and where* $x_{(i)}$ *and* $y_{(j)}$ *for* $1 \leq i, j \leq m$ *denote order statistics from the sample.*

The empirical copula is a non-parametric copula. The definition is a special case of the definition of an empirical distribution function. They should not be mistaken for a distribution function, as they represent a dependency structure and are only defined on $\mathbb{I}$. We rely on empirical copulas when parametric copulas fail to fit the data well enough. It is an important building block of the *Goodness-of-fit* test for copulas, which will be explained in chapter 5.

A comparison of the shape of some copulas can be found in figure 2.



(a) Structure of some parametric copulas. Left: Bivariate Gaussian copula with $\rho = 0.5$. Right: Bivariate Student t copula with $\rho = 0.5$ and df $= 1$.



(b) Structure of an independent copula.

Figure 2: Source: David Gold. (2017). *An introduction to Copulas.* Water programming: A collaborative Research Blog. Retrieved from `https://waterprogramming.wordpress.com/2017/11/11/an-introduction-to-copulas/`.

3 Data set

In this thesis we use a data set that is part of a larger data set. The original data set is considered one of the largest data about bibliometric information of researchers. This original data set has been used in other studies, e.g. Gingras et al., 2008; Larivière et al., 2001; Costas et al., 2015. The data set is composed by 13626 scholars from Quebec. Each professor has published at least one article between 1980-2012.
As mentioned before, in this thesis we use part of the above mentioned data set. Our data set consists of 3574 of the 13626 researchers from Quebec. These researchers obtained a PhD after 1980. Thus we can assume that the first publication of each researcher in our set, was published between 1980 and 2012 on the Web of Science. All the citation data comes from the Web of Science. Our data set provides us the following variables:

- p: The total number of publications of a researcher published between 1980 and 2012
- *mcs*: Mean citations of all p publications
- *mncs*: The nomalized mean of all citations of all p publications
- *pp_top_prop*: The percentage of p publications which are in the top 10% most cited papers in their field per publication year
- *pp_int_collab*: The percentage of p publications which were international collaborations

Table 2 summarises some main descriptive values of the data set.

	p	mcs	mncs	pp_top_prop	pp_int_collab
N	3574	3574	3574	3574	3574
Mean	26.989	17.537	1.351	0.135	0.295
Std. Deviation	36.890	35.129	1.678	0.160	0.267
Minimum	1	0	0	0	0
Maximum	777	1550.5	47.333	1	1

Table 2: Main descriptive values of the data set.

N corresponds with the amount of observations. The table also shows the mean, standard deviation, minimum and maximum per variable.
Each researcher belongs to a certain division. A division is one of nine disciplinary fields of activity of the scholar, which is based on the 2000 revision of the U.S. Classification of Instructional Programs (CIP) developed by the U.S. Department of Education's National Center for Education Statistics (NCES). A table with the main descriptive values per division can be found in Appendix A. In the next chapter, we'll analyse these values and more.

4 Publication & Citation Analysis

Before copulas are selected and fitted, it is important to analyse the data. Descriptive statistics and plots help with the interpretation of the data. Furthermore, it helps us to evaluate the output of a model selection with a critical eye.
In this chapter, a visual analysis of each variable is given. We will be confronted with some outliers as a result of the visual analysis. To create a deeper understanding with regard to the outliers, we will use extra bibliometric indicators to analyse these outliers. Next to the bibliometric indicators introduced in chapter 3, we will also analyse the birth year of the researchers who are regarded as outliers, the amount of publications the researcher published in the first year of his/her first publication, the year when the researcher published for the first time and the average number of authors of all publications for the researcher.
Furthermore, we analyse the correlation between each pair of variables. For a deeper analysis, the correlation coefficients are also computed by binning our data. We create bins based on the quartiles, the 90th percentile and the 95th percentile of the data. The correlation coefficients of the bins are calculated. This should provide more information about the tails of the joint distribution. Finally, the correlation coefficients are computed per division.

4.1 Visual analysis

We will visually inspect the variables *p, mcs, mncs, pp_top_prop* and *pp_int_collab*. This visual analysis will give us an insight on possible outliers and the characteristics of the distribution of the variables.

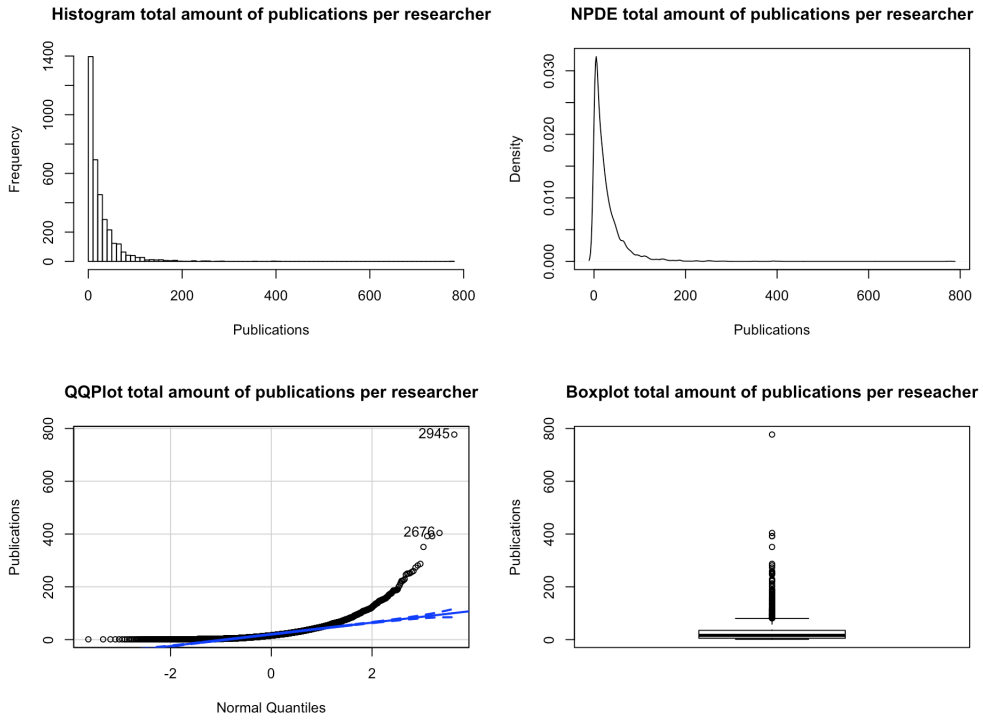


Figure 3: Visualising the variable p via a histogram, non-parametric density estimate, QQplot and boxplot.

Publications

By visually inspecting the variable p, we can conclude that the distribution of the variable is positively skewed. This follows from the histogram and the non-parametric density estimate, since it has a long tail on the right. From the same figures we can conclude that the distribution has one mode. Furthermore, we see a high peak in both figures around the zero. This suggests that most of the researchers have a very low number of publications. In fact, 7% of researchers have 1 publication and 25.9% of researchers have at most 5 publications.
The QQplot compares the quantiles of the empirical distribution based on data with the quantile function of the normal distribution. The quantile- comparison plot in figure 3 tells us that we can speak of non-normality with respect to the distribution. There are a lot of data points outside of the CI, especially in the right tail. Most of the data points huddle together between the 1 and 287, but there are also about 5 outliers. A short summary on the outliers is presented in table 3.

p	Birthyear	Phd year	pfy	fpy	Division	authors	mcs	mncs	pp_top_prop	pp_int_collab
351	1962	1987	1	1990	Eng	3.69	7.84	1.25	0.14	0.39
392	1960	1990	1	1988	BMS	9.88	84.37	3.22	0.35	0.42
392	1960	1990	1	1988	HS	9.88	84.37	3.22	0.35	0.42
404	1974	2002	1	1999	Sc	1869.88	17.34	3.34	0.24	0.99
777	1968	1998	6	1993	Sc	1214.5	21.51	2.65	0.24	0.99

Table 3: Summary on the outliers of variable p.

Here, *pfy* denotes how many publications the researcher published in the first year of his/her first publication. *fpy* denotes the year when the researcher published for the first time. *authors* denotes the average number of authors of all publications for the researcher. The abbreviations used for the divisions are *BMS* for *Basic Medical Sciences*, *Eng* for Engineering, *HS* for Health Sciences and *Sc* for Sciences.
Note that the second and third outlier are identical. This can be interpreted as a typo. However, after evaluating the complete dataset, this overlap between these two divisions is not uncommon. Which leads to believe that this isn't due to some typo. A more logical explanation would be that the researcher is active in both disciplinary fields. Especially since the division *Health Sciences* and *Basic Medical Sciences* have overlapsing branches.
Furthermore, note that the last two outliers both concern the division *Sciences* and nearly all publications were international collaborations. The average number of authors of all publications is higher than that of the other three outliers. The last author published 6 times in the year of his first publication. These numbers do not necessarily imply anything. The second and third researcher published almost as much as the fourth, and have a lower *pp_int_collab* value. The *mncs* values of these three researchers are around the same value, where the last researcher has a lower *mncs* value. The high publication value in the year of his first publication might imply a high publication value on a yearly basis, which in turn might explain the high value for p.

Mean citations of all publications

The distribution of the variable *mcs*, see figure 4, is positively skewed. This follows from the histogram and the non-parametric density estimate, since it has a long tail on the right. From the same figures we can conclude that the distribution has one mode. Again, we see a high peak in both figures around the zero. This suggests that most of the researchers have an averaged citation score around zero. In fact, 3.2% of researchers have an averaged citation score equal

to zero. 12.7% of researchers have an averaged citation score in the range of [0,1]. 27% of researchers have an averaged citation score which at most equals 5.
The quantile-comparison plot tells us that we can speak of non-normality with respect to the distribution. There is a lot of data outside of the CI, more so in the right tail then in the left tail. There are also about 4 outliers. A short summary on the outliers is presented in table 4.

mcs	Birthyear	Phd year	pfy	fpy	Division	authors	p	mncs	pp_top_prop	pp_int_collab
402.86	1962	1990	1	1991	Eng	3.21	14	22.85	0.57	0.57
402.86	1962	1990	1	1991	Sc	3.21	14	22.85	0.57	0.57
562.2	1966	1991	1	1998	BMS	32.33	5	24.69	0.8	0.6
1550.5	1961	1992	1	1993	Sc	156.92	8	47.33	0.5	0.75

Table 4: Summary on the outliers of variable *mcs*.

As with the outliers of the variable p, here the first two outliers are identical. They belong to different divisions, which again leads to the conclusion that there are a lot of researchers who are active in multiple disciplinary fields.
Again the division *Sciences* pops up between the outliers and contributes the highest outlier. However, with three variables left to analyse, lets refrain from concluding anything yet.
Notice how these high averaged citation scores do not relate to high p scores. However the *pp_top_prop* and *pp_int_collab* scores are very high. This might be related to the high *mcs* scores.

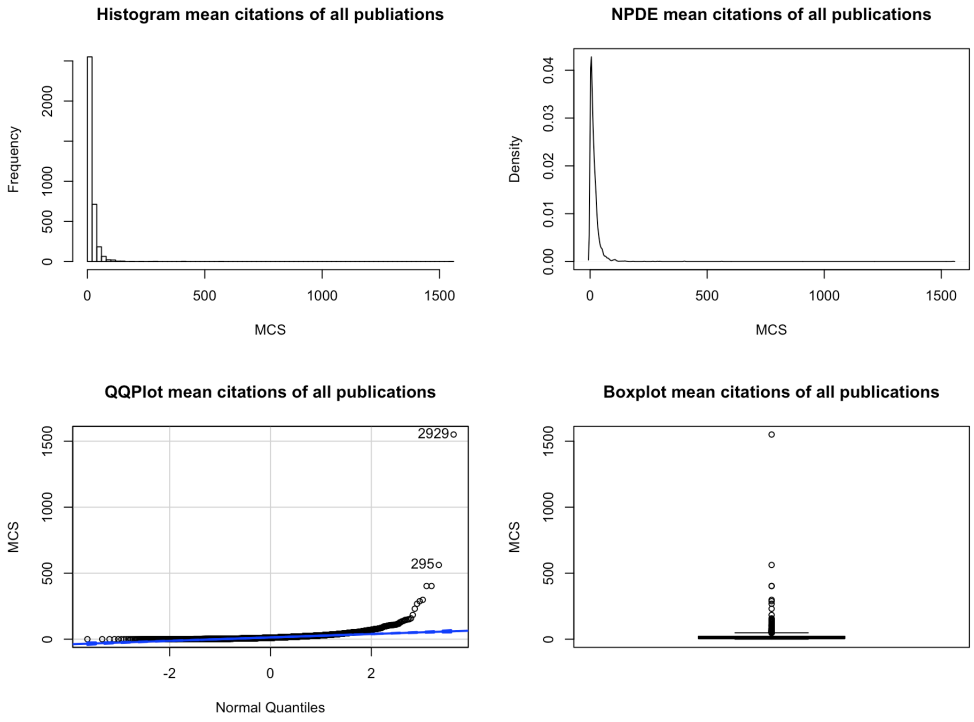


Figure 4: Visualising the variable *mcs* via a histogram, non-parametric density estimate, QQplot and boxplot.

Normalized mean citations of all publications

Just as with the previous two variables, we can conclude that the distribution of the variable *mncs* is positively skewed. The histogram and the non-parametric density estimate show us a long tail on the right. From the same figures we can conclude that the distribution has one mode. Unlike with the variables p and *mcs*, we see a high peak in both figures around the three, not around the zero. Which would imply that most researchers have a normalized mean citation score around the 3. In fact, 2.9% of researchers have a normalized averaged citation score equal to 0. 48% of researchers have a normalized averaged citation score in the range of [0,1]. Which implies that the peak in figure 5 represents a normalized mean citation score around the 0.5 rather than 3. 75% of researchers have a normalized averaged citation score in the range of [0,1.62].

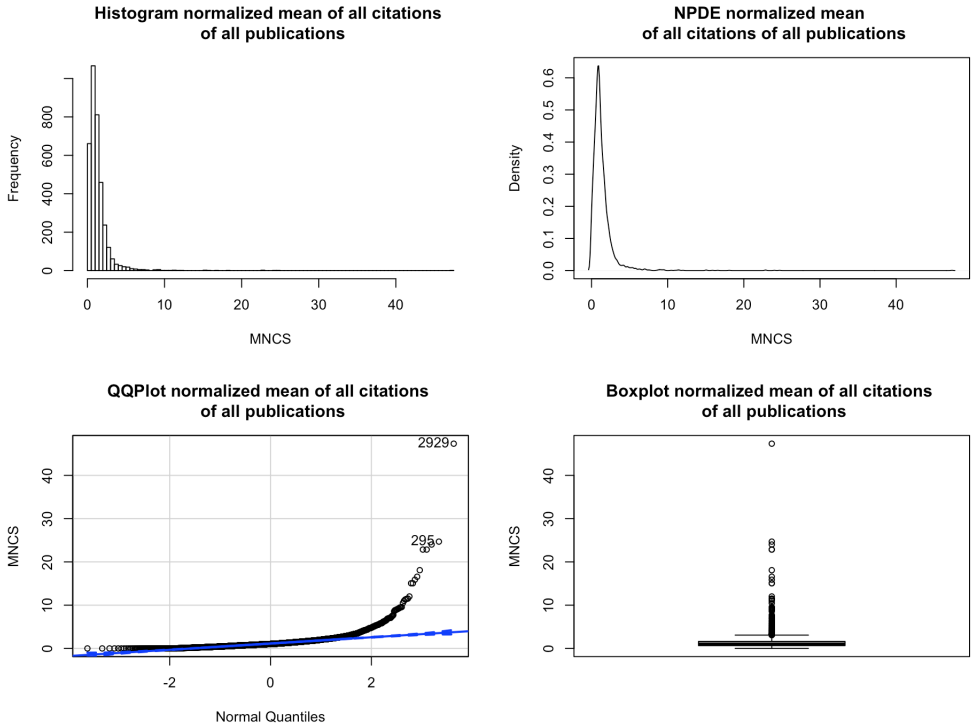


Figure 5: Visualising the variable *mncs* via a histogram, non-parametric density estimate, QQ-plot and boxplot.

The quantile- comparison plot tells us that we can speak of non-normality with respect to the distribution. There is a lot of data outside of the CI, more so than with the averaged mean citation scores. Again, most of these data points are in the right tail.
There are two gaps between the huddles of data points. The data points after these gaps are mainly huddled together, except for one.

mncs	Birthyear	Phd year	pfy	fpy	Division	authors	p	mcs	pp_top_prop	pp_int_collab
47.33	1961	1992	1	1993	Sc	156.92	8	1550.5	0.5	0.75

Table 5: Summary on the outlier of variable *mncs*.

As we can see in table 5, the outlier in this devision is the same is the last outlier in the

last division. This is not very surprising, since the normalized averaged mean citation score is calculated based on the averaged mean citation score. In this case, the noticeably high averaged mean citation score causes the outlier in *mncs*.

Percentage of publications which are in the top 10% most cited papers in their field

By visually inspecting the variable pp_top_prop, we can conclude that the distribution of the variable is positively skewed. This follows from the histogram and the non-parametric density estimate, since it has a long tail on the right. From the same figures we can conclude that the distribution has one mode. Though there is only one mode, we see two peaks in the histogram. The first peak is the most obvious peak at the zero. The second peak is a much smaller peak, but compared to all the data points a peak, around the 0.1. In fact, 27.8% of researchers have 0 publications which are in the top 10% most cited papers in their field in the publication year. 47.4% of researchers have at most 20% of their publications belonging to the top 10% most cited papers in their field. Furthermore, about 24 researchers have publications which are all in the top 10% most cited papers in their field. However, 19 of these researchers have only published once. 3 researchers published twice, one published thrice and one researcher published four times. Most researchers belong to the division *Humanities*. All publications were published after 2000.

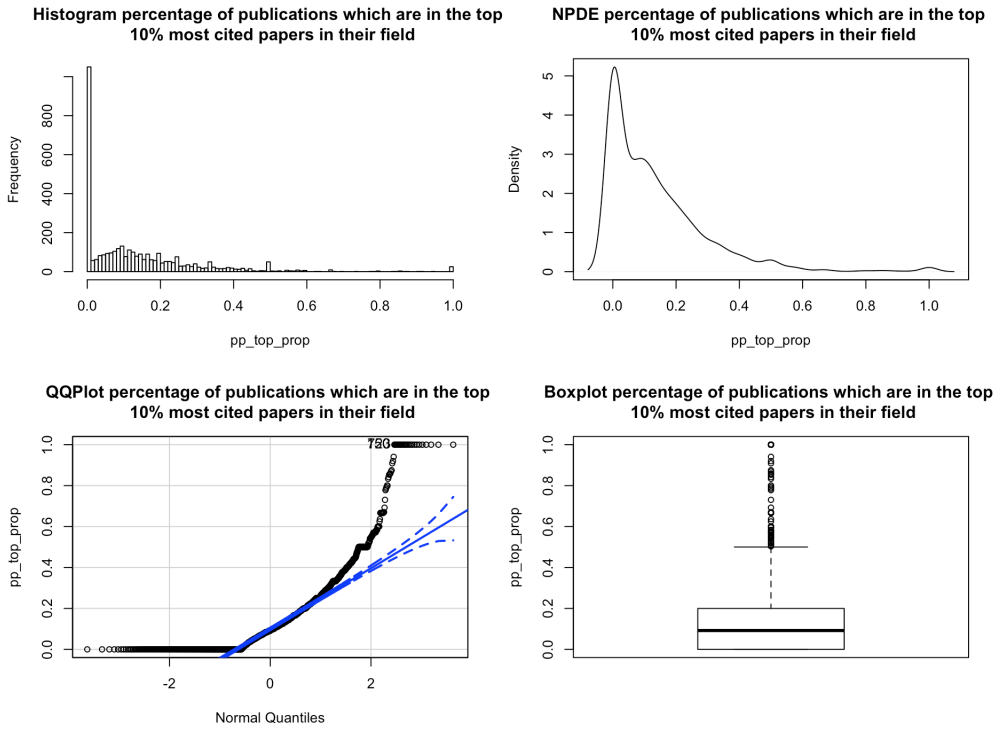


Figure 6: Visualising the variable *pp_top_prop* via a histogram, non-parametric density estimate, QQplot and boxplot.

The quantile- comparison plot tells us that we can definitely speak of non-normality with respect to the distribution. There is a lot of data outside of the CI, more so in the right tail than in the left tail. The amount of data in the left tail is however very noticeable, especially compared to previous variables. We see that there are a lot of outliers. A lot of the outliers are

at the zero. The boxplot shows us that the mean of this variable is around the 0.1, but there are still a lot of data points between the 0.5 and 1.

Percentage of publications which were international collaborations

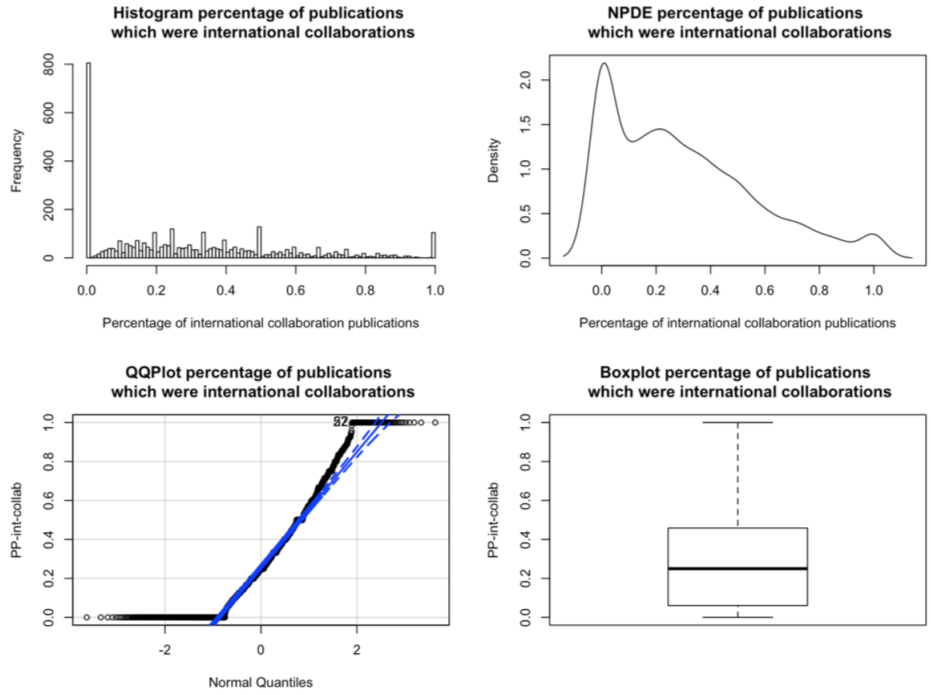
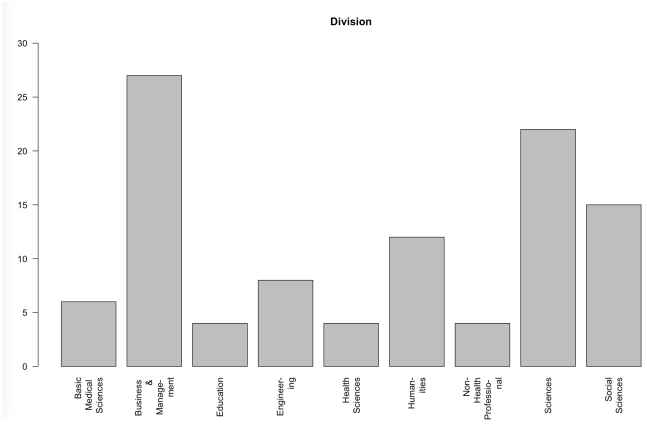
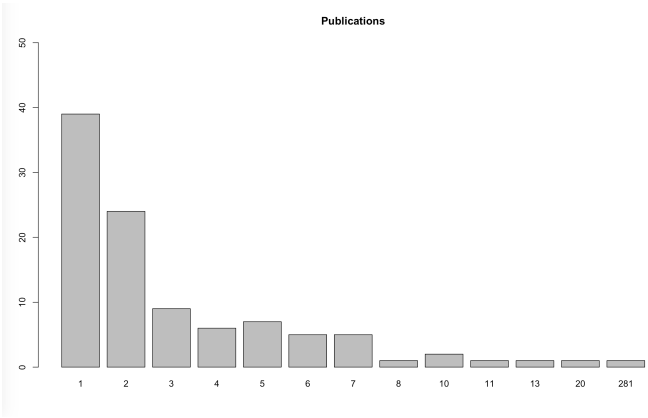


Figure 7: Visualising the variable *pp_int_collab* via a histogram, non-parametric density estimate, QQplot and boxplot.

Our final variable is also positively skewed, though much less skewed than the previous distribution. The histogram and the non-parametric density estimate of *pp_int_collab* show a long tail on the right. Like all the previous variables, the distribution has one mode. Though there is only one mode, we see quite a few peaks in the histogram. The first peak is again at the zero. We have five other peaks, all about the same height but much smaller than the first peak. We see these small peaks at the 0.2, 0.3, 0.5 and 1.0. In fact,
The quantile- comparison plot tells us that we can speak of non-normality with respect to the distribution. There are a lot of data outside of the CI. Mostly at the zero and around the 1. About 102 researchers have publications which were all international collaborations. Again all of these publications were published after 2000. In figure 8 we see that most of these publications are published by researchers active in the field *Business & Management*, with the division *Sciences* following suite. The value of p for these publications has a rather wide range. Most researchers have only published once or twice. But there is also a researcher who has published 281 times and collaborated internationally for every publication.



(a) Barplot *divisions*.



(b) Barplot p.

Figure 8: Barplots of the variables p and *division* which correspond with a *pp_int_collab* score equal to 1.

The high peaks around the zero, this can also be seen in the histograms, and the data points outside of the confidence intervals in the QQ-plots, suggest the possibility of outliers. However, after some evaluation, none of the researchers are excluded. All of the scholars published at least once. Deleting scholars who are not cited, would not give us a realistic representation of the publication world. To confirm this, two restrictions were applied. The first restriction required the scholars to have at least four publications and at mean normalised citation score of 0.85. This excluded 368 scholars. The correlation scores of each pair of variables decreased, which is not what we desired. The second restriction excluded scholars with less than four publications and a mean normalised citation score equal to zero. This excluded 100 scholars. Again the correlation scores dropped. Both restrictions had the opposite effect of what we desired. Which led to the conclusion that no scholars should be excluded.
Now let's see how these variables are correlated. Are the correlation coefficients higher in the tails because of the positively skewed distributions? In the next paragraph, we will compute the correlation coefficients of the variables pairwise. Furthermore, we will compute the correlation scores per quantile and compute the correlation score of the 90th percentile and the 95th percentile. This should give us more information about the tail dependency. Which will be a tremendous help with the fitting of the copulas.

4.2 Correlation

Chapter 2 clarifies why rank correlation is favoured over ordinary correlation with our data set. To compute Spearman's rank correlation and Kendall's rank correlation, R is used. Since the data set consists of a lot of ties, it is sensible to evaluate how the software handles these ties. R uses a variation of the Kendall correlation coefficient in order to deal with the tied ranks. This variation is known as **Kendall's tau-b coefficient**. The *Kendall's tau-b coefficient* is defined as

$$\begin{aligned} \tau_B &= \frac{\text{P - Q}}{\sqrt{N_1} \times \sqrt{N_2}} \\ \text{where } P &= \text{number of concordant pairs} \\ Q &= \text{number of discordant pairs} \\ N_1 &= \text{number of data pairs not tied in the first variable} \\ N_2 &= \text{number of data pairs not tied in the second variable} \end{aligned}$$

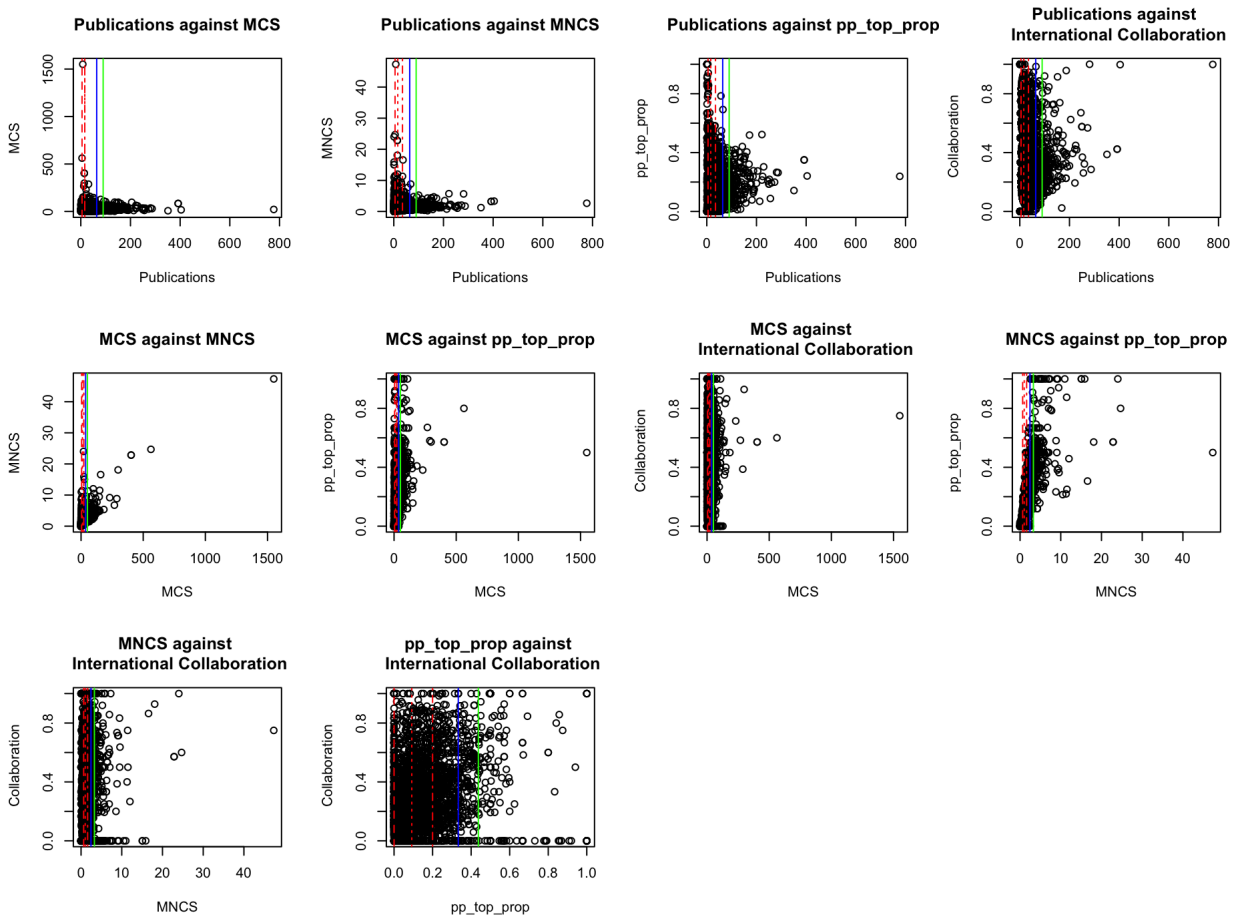
R uses the same formula for Spearman's correlation as described by definition 2.4.3 in paragraph 2.4. R fails in the computation of the correlation coefficient when the rank of one of the variables or both has a standard deviation equal to zero. This can be the case when, for example, the correlation between the first ten researchers are computed with regard to the ranked variables *pp_top_prop* and *mcs*. The value of *pp_top_prop* for these ten scholars equal zero. So the standard deviation equals zero.
Though the bins created by splitting the data set according to their quartiles contain ties, the standard deviation of these ranked bins do not equal zero. So Spearman's rank correlation is computed without difficulty. Kendall's correlation coefficient is in fact Kendall's tau-b coefficient.

Variable one	Variable two	Method	All data	Q1 $\leq\vert$	Q1 - Q2	Q2 -Q3	Q3 $\geq\vert$	90th percentile $\geq\vert$	95th percentile $\geq\vert$
p	mcs	S	0.5	016	0.17	0.14	0.20	0.25	0.19
		K	0.36	0.12	0.12	0.1	0.14	0.17	0.13
p	mncs	S	0.31	0.13	0.09	0.10	0.15	0.28	0.16
		K	0.22	0.1	0.06	0.07	0.1	0.19	0.11
p	pp_top_prop	S	0.38	0.19	0.08	0.11	0.13	0.28	0.15
		K	0.27	0.15	0.05	0.07	0.09	0.19	0.1
p	pp_int_collab	S	0.33	0.22	0.05	0.04	0.06	0.16	0.17
		K	0.23	0.18	0.03	0.02	0.04	0.11	0.12
mcs	mncs	S	0.71	0.53	0.22	0.24	0.58	0.59	0.67
		K	0.55	0.39	0.15	0.16	0.42	0.43	0.49
mcs	pp_top_prop	S	0.65	0.29	0.15	0.19	0.44	0.34	0.24
		K	0.49	0.22	0.11	0.13	0.3	0.23	0.16
mcs	pp_int_collab	S	0.36	0.33	0.11	0.02	0.19	0.14	0.16
		K	0.26	0.24	0.08	0.01	0.13	0.1	0.11
mncs	pp_top_prop	S	0.87	0.42	0.43	0.47	0.60	0.44	0.19
		K	0.72	0.33	0.31	0.33	0.43	0.31	0.13
mncs	pp_int_collab	S	0.3	0.29	0.08	0.05	0.09	0.06	-0.06
		K	0.21	0.2	0.05	0.03	0.06	0.04	-0.04
pp_top_ prop	pp_int_collab	S	0.29	0.23	0.11	0.11	-0.02	-0.26	-0.41
		K	0.22	0.21	0.07	0.08	-0.01	-0.19	-0.33

Table 6: Correlation coefficients.

Table 6 reports the correlation coefficients of the data set. The S and K listed under method denote Spearman and Kendall respectively. This notation applies for the rest of this thesis.
Now going back to table 6, the correlation coefficients are quite dispersed. Notice how all the Kendal correlation coefficients are lower than the Spearman correlation coefficients. Where Kendal is based on the concordance and discordance of data pairs, Spearman is based on deviations. This leads to smaller correlation coefficients when Kendall's method is used. This is also why Kendall's method is usually preferred over Spearman's method in Statistics. It is said that Kendall's tau is less sensitive to outliers and the p-values, calculated when testing the null hypothesis that Kendall's tau equals 0, are more accurate with smaller sample sizes, where Spearman's rho is more sensitive to outliers and discrepancies in data. However, in most situations, the interpretations of Kendall's tau and Spearman's rank correlation coefficient lead to the same inferences because they are very similar.
The correlation coefficients of the bins help with the copula selection. Take the pair *mcs, mncs* as an example. Spearman's rank correlation coefficient for this pair equals 0.71. Furthermore, the tails have higher correlation coefficient than the middle. In general, the random correlations show a weak dependence in the middle of the data. An observation that could not have been made with merely the correlation score of 0.71, nor with the scatterplot below. The pair *mncs, pp_top_prop* has a completely different structure. The pair has higher correlation coefficients than the previous pair, but they are not similar in structure at all. The correlation coefficient in the first three bins are quite similar. The last bin has a higher correlation coefficient. However, looking at the bins on the right side of the 90th percentile and the 95th percentile, does not suggest a thicker tail. The correlation coefficient in the tail is rather light.
Now let's look at a pair with a low correlation score. Take the pair *p,pp_int_collab*. Though the correlation coefficients of this pair is very low, we can still say a lot about the dependence structure because of the correlation coefficients of the bins. It is clear that this pair is much thicker in the tails than in the middle.
In this paragraph, we have seen that just looking at one correlation coefficient can put us on the wrong path when it comes to copula selection. A high correlation coefficient can give us just as much information on the dependence structure as a low correlation coefficient, if the analysis runs deeper. Just as looking at one correlation coefficient is not enough, looking at the complete data set can be a pitfall as well. Furthermore, in the beginning of this paragraph we saw that the first few outliers were often linked to the division *Sciences*. We also saw that a high international collaboration score is linked to the divisions *Business & Management* and *Sciences*. This implies that every division has a different publication and citation behaviour. These are the reasons why in the next paragraph the data set is split by its divisions.

Scatterplots of the different pairs of variables. The first three red lines denote Q1,Q2 and Q3 respectively. The blue line denotes the 90th percentile and the green line denotes the 95th percentile.



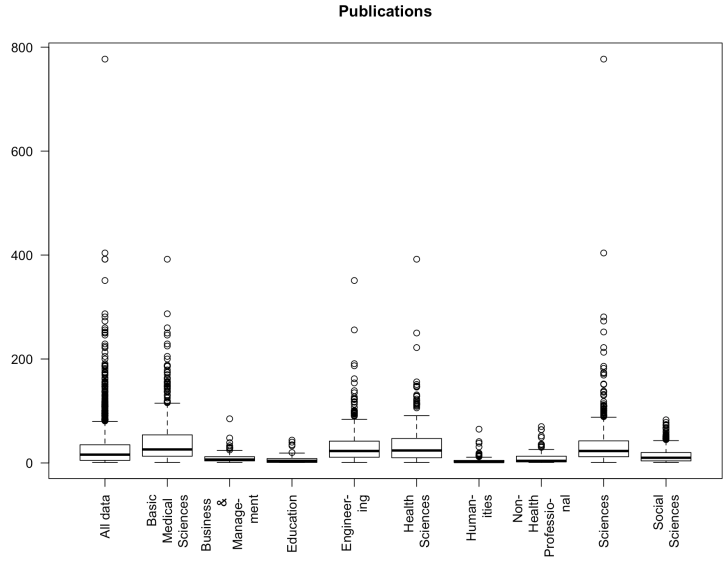
4.3 Data by division

Every division has its own specific publication and citation behaviour. For example, it is well known that papers are regularly published in the field of medicine. More so than in other fields. That is probably why this is the division with the largest averaged publication score, see Appendix A. According to current assumptions, that might mean that the division *Basic Medical Sciences* publishes more than the other divisions. Assume that the citation values do not increase with this higher publication rate. This leads to a dependence structure which differs from the dependence structure between the same variables in different divisions. The dependence structures analysed in the previous paragraph will also differ from the dependence structures of the division *Basic Medical Sciences*. Which leads to different copula parameters and maybe even a completely different copula.

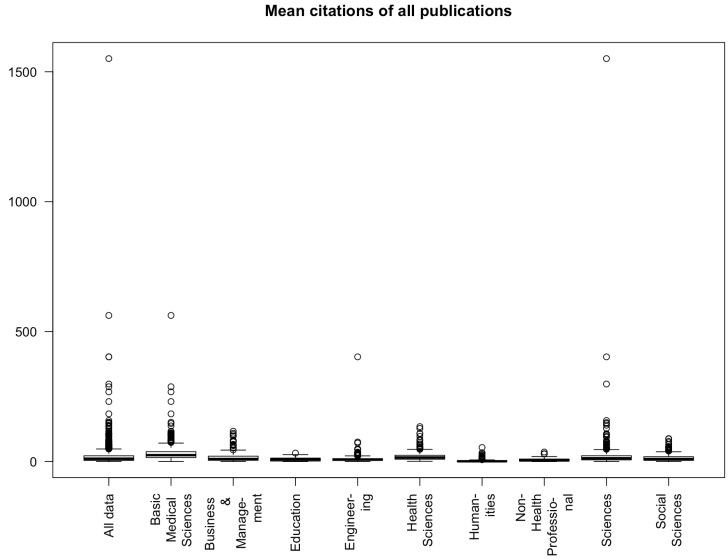
Division	Amount of observations
Basic Medical Sciences	711
Business & Management	238
Education	47
Engineering	512
Health Sciences	288
Humanities	324
Non-Health Professionals	108
Sciences	824
Social Sciences	500

Table 7: Amount of observations per division.

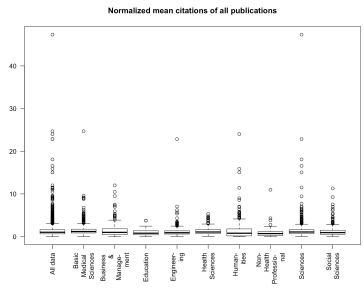
Figure 9 gives us a quick overview of the diverse distributions of the different divisions. Applying the same copula for each division would result in an unrealistic model. So let's analyse our data per division. In table 7 we see that the divisions *Basic Medical Sciences* and *Sciences* provide the most observations. Which means that most of our researchers are active in these disciplinary fields. Appendix A summarizes some main descriptive values per division per variable. As said before, the researchers active in the divisions *Basic Medical Sciences* publishes the most. This can be concluded from the highest average publication score. The division *Humanities* has the lowest averaged amount of publications. Notice that the mean scores for the variable p in Appendix A differ a lot from the mean score for the same variable of the complete data set. This actually goes for all the mean values of the variables. The division *Basic Medical Sciences* also has the highest mean value for the variables *mcs* and *mncs*. Thus, the researchers in active in this field publish publications which are cited more often than in other fields. However, *Business & Management* and *Sciences* are the divisions with the highest mean values for *pp_top_prop* and *pp_int_collab*. So the researchers in these fields tend to collaborate more internationally for their publications and have 10% of their publications as highly cited. The divisions *Education* and *Non-Health Professional* have the lowest mean values for the variable *mncs*. *Humanities* has the lowest mean value for *mcs*, but has a high *mncs* mean value. This division also has the lowest *pp_int_collab* mean value. Finally, the divisions *Education* and *Non-Health Professional* also have the lowest mean value for the variable *pp_top_prop*.
In the next part of the thesis, the same analysis method as in the previous paragraph is applied. Some tables contain a '-' instead of a correlation coefficient. These correlation coefficients could not be computed due to very low number of observations. For Spearman, this results in a standard deviation equal to zero. For Kendall this means that all the data points are tied. So according to the Kendall tau-b coefficient formula, the correlation coefficient can not be computed.



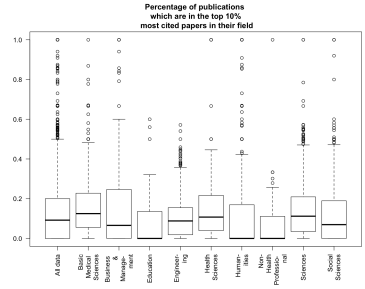
(a) Boxplot variable publications



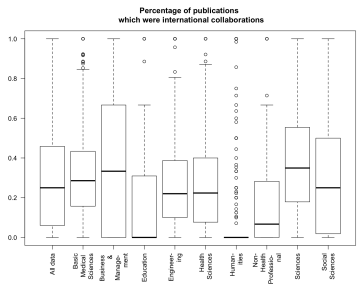
(b) Boxplot variable MCS



(c) Boxplot variable MNCS



(d) Boxplot variable pp_top_prop



(e) Boxplot variable International Collaboration

Figure 9: Boxplot of each variable. The divisions are compared with each other and with the distribution of the complete data set.

Basic Medical Sciences

The complete data set consists of information about 3570 scholars. 711 of these scholars belong to the division *Basic Medical Sciences*. Table 8 reports the correlation coefficients of the division *Basic Medical Sciences*. Just as with the complete data set, the Kendall correlation coefficients are lower than the Spearman correlation coefficients. Compared to the correlation coefficients of the complete data set, most correlation coefficients have decreased. Furthermore, most pairs in this division have a dependence structure where the middle has a light density and the tails have a high(er) density. The tails are not strongly correlated. The density in the tails is higher than the density in the middle. The data points in the middle are completely scattered. This causes the light density. This differs highly from the structures given by table 6.
Also notice the negative correlation coefficients. Especially in the 95th percentile. The negative correlation coefficients imply that an increase in variable 1 is associated with a decrease in variable 2. For the pair *(p,pp_int_collab)* the structure between Q1 and Q2 is rather monotonic. This is implied by the correlation coefficients for this pair, which is 0 (spearman) and -0.04 (kendall). The pair *(pp_top_prop,pp_int_collab)* has higher negative correlation coefficients in area above the 95th percentile. In this case, this means that researchers active in the field *Basic Medical Sciences* with at least 40.8% of their publications in the top 10% most cited papers in said field, tend to internationally collaborate less. This means that the *pp_int_collab* score for these researchers was mostly below 0.4
The pairs *(mcs, mncs),(mcs, pp_top_prop)* and *(mncs,pp_top_prop)* have increased correlation coefficients and a different structure. The first pair has a dense middle with strongly correlated tails. The second pair has a similar structure, but without the strongly correlated tails. The tails are denser than the middle. Finally, the last pair has a strongly correlated left tail.

Variable one	Variable two	Method	Complete data set	All data in division	$\leq$ Q1	Q1 - Q2	Q2 -Q3	$\geq$ Q3	$\geq$ 90th percentile	$\geq$ 95th percentile
p	mcs	S	0.5	0.31	0.21	0.01	0.03	0.16	0.08	0.04
		K	0.36	0.22	0.15	0.005	0.02	0.11	0.06	0.03
p	mncs	S	0.31	0.31	0.28	0	0.07	0.23	0.18	0.06
		K	0.22	0.22	0.20	-0.004	0.05	0.16	0.12	0.06
p	pp_top_prop	S	0.38	0.31	0.35	0.02	0.03	0.21	0.19	0.02
		K	0.27	0.22	0.25	0.01	0.01	0.15	0.12	0.03
p	pp_int_collab	S	0.33	0.27	0.37	-0.08	0.12	0.11	0.23	0.09
		K	0.23	0.19	0.26	-0.06	0.08	0.07	0.17	0.07
mcs	mncs	S	0.71	0.84	0.63	0.39	0.32	0.61	0.75	0.68
		K	0.55	0.66	0.47	0.27	0.21	0.45	0.57	0.52
mcs	pp_top_prop	S	0.65	0.75	0.45	0.26	0.27	0.36	0.34	0.33
		K	0.49	0.56	0.34	0.18	0.18	0.25	0.24	0.22
mcs	pp_int_collab	S	0.36	0.31	0.18	0.01	-0.01	0.20	0.24	0.09
		K	0.26	0.22	0.12	0.01	-0.01	0.14	0.16	0.07
mncs	pp_top_prop	S	0.87	0.89	0.64	0.48	0.50	0.44	0.27	-0.05
		K	0.72	0.72	0.49	0.33	0.35	0.31	0.18	-0.05
mncs	pp_int_collab	S	0.3	0.35	0.31	0.08	0.08	0.18	0.23	0.32
		K	0.21	0.25	0.22	0.05	0.05	0.12	0.16	0.22
pp_top_ prop	pp_int_collab	S	0.29	0.33	0.22	0.09	0.12	0.04	-0.05	-0.29
		K	0.22	0.24	0.17	0.07	0.08	0.03	-0.03	-0.22

Table 8: Correlation coefficients Basic Medical Sciences.

Business & Management

Table 9 computes the correlation coefficients of 238 scholars. Notice how the correlation coefficients of the last six pairs have increased compared to the correlation coefficients of the complete data set. A noticeable amount of bins are negatively correlated in this division. For some pairs, this negative correlation coefficient represents only one bin. However some pairs have negative correlation coefficients for multiple bins. Let's take a look at the first pair. The tails of this pair are negatively correlation. Which means that researchers with a very low amount of publications, and researchers with a very high amount of publications tend to be cited less in this field. In other words, say that a researcher in this field publishes once, according to the negative correlation coefficient chances are that the publication does not get cited. For a researcher with let say 83 publications, in this field, chances are that the averaged amount of citations of all publications is lower than 83. This reasoning also holds for researchers with a publication amount in Q2-Q3. This kind of structure applies to the first four pairs. The other distribution functions in this division have a denser middle and light to no tails. The pair *(mcs,mncs)* is an exception to this, which holds the same structure as discussed for the previous division. The pair *(mncs,pp_int_collab)* has a negative correlated tail. So researchers with a very high normalized averaged mean citation score tend to collaborate less on an international basis.

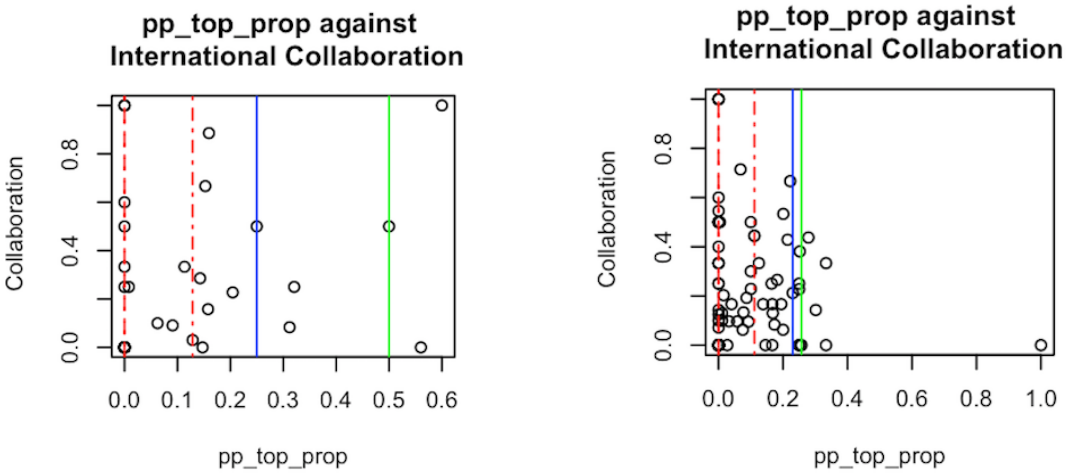
Variable one	Variable two	Method	Complete data set	All data in division	$\leq$ Q1	Q1 - Q2	Q2 -Q3	$\geq$ Q3	$\geq$ 90th percentile	$\geq$ 95th percentile
p	mcs	S	0.5	0.19	-0.19	0.33	-0.11	0.03	-0.12	-0.14
		K	0.36	0.13	-0.14	0.26	-0.08	0.02	-0.09	-0.11
p	mncs	S	0.31	0.26	-0.10	0.25	-0.11	0.11	0.03	-0.27
		K	0.22	0.18	-0.08	0.19	-0.08	0.07	0.03	-0.16
p	pp_top_prop	S	0.38	0.3	-0.11	0.27	-0.15	0.17	0.04	-0.08
		K	0.27	0.22	-0.10	0.22	-0.12	0.11	0.03	-0.05
p	pp_int_collab	S	0.33	0.07	-0.14	0.35	-0.10	-0.03	-0.06	-0.31
		K	0.23	0.04	-0.13	0.29	-0.09	-0.03	-0.05	-0.21
mcs	mncs	S	0.71	0.92	0.69	0.58	0.36	0.76	0.71	0.34
		K	0.55	0.75	0.51	0.41	0.24	0.57	0.51	0.23
mcs	pp_top_prop	S	0.65	0.8	0.18	0.44	0.23	0.58	0.43	0.12
		K	0.49	0.63	0.15	0.33	0.17	0.43	0.34	0.06
mcs	pp_int_collab	S	0.36	0.45	-0.06	0.39	0.03	0.23	-0.17	0.33
		K	0.26	0.33	-0.05	0.27	0.02	0.15	-0.12	0.23
mncs	pp_top_prop	S	0.87	0.89	0.06	0.35	0.64	0.73	0.32	0.01
		K	0.72	0.75	0.05	0.26	0.46	0.56	0.26	0.06
mncs	pp_int_collab	S	0.3	0.39	0.13	0.15	-0.09	0	-0.09	-0.30
		K	0.21	0.28	0.11	0.11	-0.06	0.002	-0.04	-0.13
pp_top_ prop	pp_int_collab	S	0.29	0.36	-	0.05	0.15	0.09	0.02	0.5
		K	0.22	0.28	-	0.04	0.11	0.06	0.01	0.41

Table 9: Correlation coefficients Business & Management.

Education

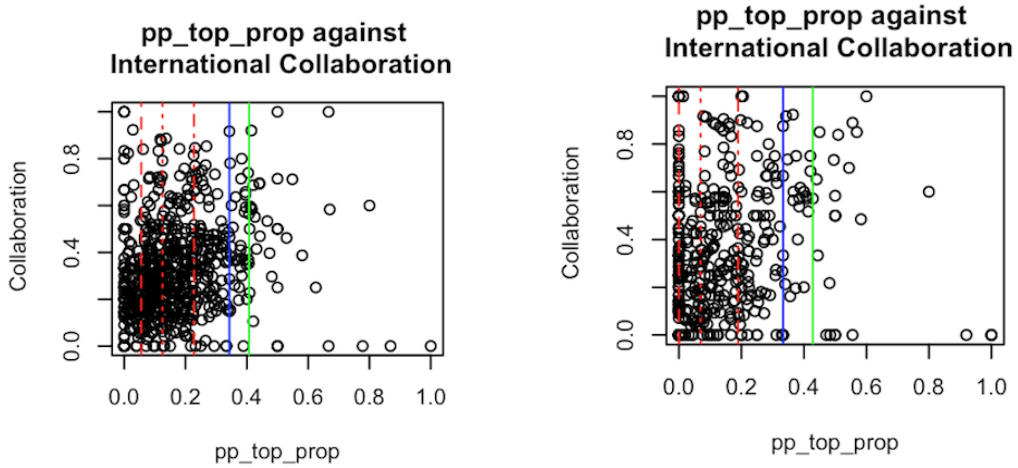
Education is the smallest division in our data set. With only 47 observations, the correlation coefficients and the dependence structures differ a lot from the other divisions and the complete dataset. For one, the ranked variables are quickly tied below Q1 because the dataset is so small.

For the last pair of variables, this goes for the bin up until Q1 and the bin between Q1 and the median. Furthermore, the data points are very dispersed for such a small data set. The division *Non-Health Professional* has 108 observations, which is a little more over twice as many observations as the current division, but has a completely different structure. The data points in the division *Non-Health Professional* have a similar structure compared to for example the divisions *Basic Medical Sciences* and *Social Sciences*.



(a) Education.

(b) Non-Health Professional.



(c) Basic Medical Sciences.

(d) Social Sciences.

Figure 10: Scatterplots pp_top_prop vs pp_int_collab of different divisions. The first three red lines denote Q1,Q2 and Q3 respectively. The blue line denotes the 90th percentile and the green line denotes the 95th percentile.

As figure 10 shows, the data points of the divisions *Non-Health Professional*, *Basic Medical Sciences* and *Social Sciences* mainly huddle together. The structure of figure 10(a) is completely different.

Three correlation coefficients increased compared to the correlation coefficients of the complete data set. The most important observation, when inspecting table 10, are the (strong) upper tails. Compared to the lower tail and the middle, the upper tail continuously stands out in

this division. Especially for the pairs *(mcs, pp_top_prop)* and *(mcs,pp_int_collab)*, where the area above the 95th percentile tends to be perfectly positively correlated. This implies that researchers with a high averaged mean citation score, in this field that means around the 30, have published publications which practically all are in the top 10% most cited papers in their field and were practically all international collaborations.
As with the previous two divisions, the pair *(mcs,mncs)* has a strong upper and lower tail with a dense middle. Unlike the previous division, there aren't that many negative correlation coefficients. Notice that the pair *(mcs, pp_int_collab)* is negatively correlated throughout the first half of the structure. This implies that researchers with low averaged mean citation scores publish less publications which were international collaborations. Furthermore, notice how again, *(mncs,pp_int_collab)* is negatively correlated in the upper tail. Coincidently, the pair *(p,pp_int_collab)* has a negatively correlated upper tail as well. So researchers with a high amount of publications, in this field that means about 40 publications, tend to collaborate less on an international basis.

Variable one	Variable two	Method	Complete data set	All data in division	Q1 $\vee\vert$	Q1 - Q2	Q2 -Q3	Q3 $\wedge\vert$	$\geq\vert$ 90th percentile	$\geq\vert$ 95th percentile
p	mcs	S	0.5	0.44	-	0.09	-0.06	0.60	0.43	0.5
		K	0.36	0.35	-	0.05	-0.02	0.5	0.33	0.33
p	mncs	S	0.31	0.26	-	-0.22	-0.10	0.38	0.43	0.5
		K	0.22	0.21	-	-0.19	-0.02	0.29	0.33	0.33
p	pp_top_prop	S	0.38	0.52	-	0.03	0.18	0.45	0.43	0.5
		K	0.27	0.40	-	0.05	0.15	0.37	0.33	0.33
p	pp_int_collab	S	0.33	0.31	-	0.13	0.35	0.65	0.03	-0.5
		K	0.23	0.23	-	0.1	0.24	0.47	0.07	-0.33
mcs	mncs	S	0.71	0.81	0.68	0.63	0.35	0.80	0.26	0.5
		K	0.55	0.68	0.52	0.5	0.29	0.61	0.2	0.33
mcs	pp_top_prop	S	0.65	0.58	0.04	0.31	0.47	0.30	0.23	1
		K	0.49	0.47	0.04	0.27	0.34	0.23	0.14	1
mcs	pp_int_collab	S	0.36	0.23	-0.25	-0.33	-0.06	0.27	0.35	1
		K	0.26	0.17	-0.24	-0.28	0.02	0.22	0.28	1
mncs	pp_top_prop	S	0.87	0.69	-	0.39	0.52	0.63	0.87	0.5
		K	0.72	0.59	-	0.35	0.38	0.45	0.69	0.33
mncs	pp_int_collab	S	0.3	0.16	0.12	0.22	0.10	-0.14	0.33	-0.5
		K	0.21	0.13	0.1	0.16	0.06	-0.1	0.3	-0.33
pp_top_ prop	pp_int_collab	S	0.29	0.48	-	-	0.32	0.12	0.17	0.5
		K	0.22	0.41	-	-	0.21	0.09	0.14	0.33

Table 10: Correlation coefficients Education.

Engineering

The correlation coefficients are based on 512 observations. The output is similar to that of the division *Basic Medical Sciences*. The dependence structure between most pairs consist of a light density middle and a high(er) density tail. All pairs have decreased correlation coefficients with respect to the complete data set, except for the pairs *(mcs, mncs),(mcs, pp_top_prop)* and *(mncs,pp_top_prop)*. The first pair has a dense middle with strongly correlated tails. The second pair has a constant structure according to the correlation coefficients. The last pair has a structure similar to the first pair. Notice how the last pair in table 11 has a dense(r) upper tail.

Finally, notice how there are only a few negative correlation coefficients in this division. All of these coefficients are so low, that they rather imply a monotonic structure than a decreasing structure.

Variable one	Variable two	Method	Complete data set	All data in division	$\leq$ Q1	Q1 - Q2	Q2 -Q3	$\geq$ Q3	$\geq$ 90th percentile	$\geq$ 95th percentile
p	mcs	S	0.5	0.39	0.10	-0.01	0.14	0.04	0.11	0.33
		K	0.36	0.27	0.06	-0.01	0.1	0.03	0.06	0.16
p	mncs	S	0.31	0.34	0.16	0.01	0.10	0.10	0.09	0.08
		K	0.22	0.23	0.1	0.01	0.07	0.08	0.05	0.02
p	pp_top_prop	S	0.38	0.31	0.24	0.05	0.07	0.05	0.12	0.16
		K	0.27	0.22	0.17	0.03	0.04	0.04	0.08	0.1
p	pp_int_collab	S	0.33	0.25	0.12	-0.03	0.09	0.02	0.25	-0.06
		K	0.23	0.18	0.07	-0.03	0.05	0.01	0.16	-0.07
mcs	mncs	S	0.71	0.8	0.66	0.47	0.27	0.56	0.51	0.42
		K	0.55	0.62	0.48	0.33	0.18	0.4	0.36	0.26
mcs	pp_top_prop	S	0.65	0.72	0.49	0.42	0.19	0.54	0.50	0.24
		K	0.49	0.54	0.37	0.28	0.13	0.38	0.37	0.15
mcs	pp_int_collab	S	0.36	0.16	0.08	0	0.05	-0.03	-0.01	0.10
		K	0.26	0.11	0.07	0.003	0.04	-0.02	-0.01	0.08
mncs	pp_top_prop	S	0.87	0.9	0.66	0.39	0.37	0.69	0.33	0.44
		K	0.72	0.74	0.52	0.28	0.26	0.51	0.24	0.35
mncs	pp_int_collab	S	0.3	0.23	0.03	0.19	-0.03	0.04	0.29	0.09
		K	0.21	0.16	0.02	0.14	-0.02	0.02	0.2	0.07
pp_top_ prop	pp_int_collab	S	0.29	0.22	0.07	0.05	0.06	0.07	0.33	0.15
		K	0.22	0.16	0.06	0.03	0.04	0.12	0.22	0.11

Table 11: Correlation coefficients Engineering.

Health Sciences

The correlation coefficients of the division *Health Sciences* are very similar to those of the complete data set. Six pairs have an increased correlation coefficient with respect to the complete data set. 288 scholars belong to this division. The first three pairs and pair seven and nine seem to have a joint distribution function with a denser upper tail. Pair four, six and ten seem to have a denser lower tail. Pair eight has a rather constant structure according to its correlation coefficients. There are few negative correlation coefficients. Notice how the last pair is negatively correlated throughout most of the structure and has a rather strong negative correlation coefficient in the upper tail. This means that a lot of the researchers active in the field *Health Sciences* with at least 37.2% of their publications in the top 10% most cited papers in said field, tend to internationally collaborate less. This means that the *pp_int_collab* score for these researchers was mostly below 0.37.

Variable one	Variable two	Method	Complete data set	All data in division	$\leq$ Q1	Q1 - Q2	Q2 -Q3	$\geq$ Q3	$\geq$ 90th percentile	$\geq$ 95th percentile
p	mcs	S	0.5	0.43	0.26	-0.04	0.06	0.37	0.21	0.33
		K	0.36	0.3	0.19	-0.03	0.04	0.25	0.15	0.23
p	mncs	S	0.31	0.35	0.09	0.09	0.18	0.31	0.23	0.26
		K	0.22	0.25	0.08	0.06	0.13	0.21	0.16	0.23
p	pp_top_prop	S	0.38	0.35	0.15	0.17	0.10	0.27	0.19	0.21
		K	0.27	0.25	0.12	0.11	0.07	0.19	0.13	0.13
p	pp_int_collab	S	0.33	0.34	0.27	0.07	0.16	0.19	0.36	0.14
		K	0.23	0.25	0.19	0.04	0.11	0.13	0.28	0.13
mcs	mncs	S	0.71	0.75	0.52	0.35	0.31	0.75	0.86	0.85
		K	0.55	0.58	0.38	0.24	0.22	0.57	0.7	0.7
mcs	pp_top_prop	S	0.65	0.68	0.42	0.32	0.18	0.47	0.42	0.26
		K	0.49	0.52	0.31	0.21	0.13	0.33	0.32	0.17
mcs	pp_int_collab	S	0.36	0.28	0.27	-0.05	-0.08	0.12	-0.08	0.33
		K	0.26	0.2	0.19	-0.03	-0.05	0.08	-0.05	0.25
mncs	pp_top_prop	S	0.87	0.9	0.65	0.43	0.52	0.56	0.02	0.44
		K	0.72	0.74	0.5	0.32	0.37	0.41	0.01	0.29
mncs	pp_int_collab	S	0.3	0.25	0.22	0.13	0.01	0	0.11	0.47
		K	0.21	0.18	0.16	0.08	0.01	0.003	0.07	0.31
pp_top_ prop	pp_int_collab	S	0.29	0.33	0.16	-0.06	0.003	-0.13	-0.03	-0.57
		K	0.22	0.16	0.13	-0.09	-0.0004	-0.08	0.02	-0.46

Table 12: Correlation coefficients Health Sciences.

Humanities

The lack of correlation coefficients in table 13 is not a result of a small number of observations. In the division *Education*, there were only 47 very dispersed observations to work with. However, this division provides 324 observations. It is however very common in this division to publish once, as can be deducted from figure 11 as well. This causes ties in certain parts of the ranked variables. Which explains the lack of correlation coefficients in table 13.

Variable one	Variable two	Method	Complete data set	All data in division	$\leq$ Q1	Q1 - Q2	Q2 -Q3	$\geq$ Q3	$\geq$ 90th percentile	$\geq$ 95th percentile
p	mcs	S	0.5	0.16	-	0.05	0.01	0.14	0.31	0.46
		K	0.36	0.12	-	0.05	0.01	0.1	0.23	0.34
p	mncs	S	0.31	0.06	-	0.09	-0.09	0.02	0.18	0.39
		K	0.22	0.05	-	0.07	-0.07	0.02	0.15	0.33
p	pp_top_prop	S	0.38	0.23	-	0.17	-0.03	0.17	0.22	0.34
		K	0.27	0.18	-	0.16	-0.02	0.13	0.18	0.32
p	pp_int_collab	S	0.33	0.23	-	0.02	0.09	0.38	0.31	0.48
		K	0.23	0.19	-	0.02	0.08	0.31	0.24	0.38
mcs	mncs	S	0.71	0.8	0.99	0.32	0.20	0.14	0.36	0.20
		K	0.55	0.63	0.95	0.24	0.16	0.09	0.23	0.1
mcs	pp_top_prop	S	0.65	0.5	0.37	-0.08	0.26	0.17	0.29	0.20
		K	0.49	0.39	0.34	-0.07	0.21	0.12	0.2	0.16
mcs	pp_int_collab	S	0.36	0.33	0.14	-0.03	-0.01	0.33	0.44	0.33
		K	0.26	0.27	0.14	-0.03	-0.01	0.26	0.32	0.23
mncs	pp_top_prop	S	0.87	0.73	-0.07	0.33	0.36	0.69	0.31	0.25
		K	0.72	0.6	-0.07	0.26	0.26	0.51	0.22	0.18
mncs	pp_int_collab	S	0.3	0.14	0.12	0.08	0.03	0.09	0.17	0.47
		K	0.21	0.11	0.12	0.06	0.03	0.07	0.14	0.38
pp_top_ prop	pp_int_collab	S	0.29	0.04	-	-	0.28	-0.13	-0.05	0.02
		K	0.22	0.04	-	-	0.23	-0.1	-0.04	0.02

Table 13: Correlation coefficients Humanities.

Notice how all the correlation coefficients have decreased except for the correlation coefficients of the pair *(mcs,mncs)*. A lot of the correlation coefficients have dropped significantly, for example the correlation coefficient of the pair *(p,mcs)* dropped from 0.5 to 0.16. Figure 11 sheds more light on these decreases. The data in this division is quite dispersed, more so than in the complete data set. According to table 13, most of joint distributions have a dense upper tail. The exceptions in this division concern the joint distribution functions of the pairs *(mcs,mncs)* and *(pp_top_prop,pp_int_collab)*. The last pair has a denser middle and no tails.

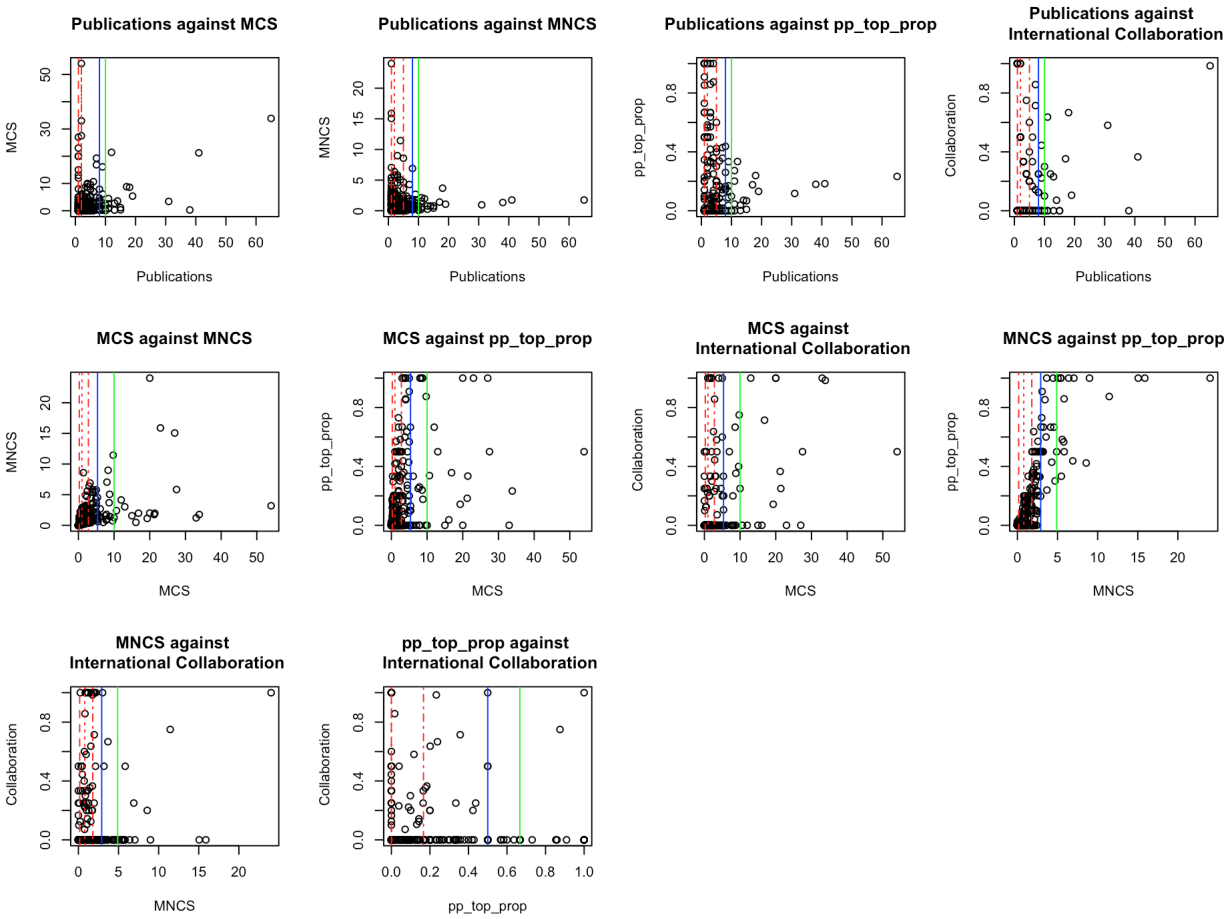


Figure 11: Scatterplots of the different pairs of variables in the division Humanities. The first three red lines denote Q1,Q2 and Q3 respectively. The blue line denotes the 90th percentile and the green line denotes the 95th percentile.

Non-Health Professional

The second smallest division with 108 observations. Most of the correlation coefficients below the first quartile are negative. As in all previous division the pair *(mcs, mncs)* is the exception. The joint distribution of the pairs *(mcs,pp_top_prop)* and *(mncs,pp_top_prop)* seem to have a dense upper tail. The pair *(mncs,pp_int_collab)* seems to have a dense lower tail. The other pairs seem to have a denser middle and light to no tails. The middle might be denser, but is not strongly correlated. Notice how almost all pairs have a negatively correlated lower tail. Except for the pair *(mncs,pp_int_collab)*, which again has a negatively correlated upper tail, just like in the divisions *Education* and *Business & Management*. The last pair is also negatively correlated in the upper tail, just like in the division *Health Sciences*.

Variable one	Variable two	Method	Complete data set	All data in division	$\leq$ Q1	Q1 - Q2	Q2 -Q3	$\geq$ Q3	$\geq$ 90th percentile	$\geq$ 95th percentile
p	mcs	S	0.5	0.45	-0.24	0.16	0.02	0.02	0.09	-0.23
		K	0.36	0.33	-0.21	0.14	-0.01	0.02	0.03	-0.28
p	mncs	S	0.31	0.48	-0.24	0.24	0.01	-0.14	-0.11	0
		K	0.22	0.34	-0.21	0.18	0.01	-0.12	-0.09	0
p	pp_top_prop	S	0.38	0.54	-0.06	0.26	-0.06	-0.19	-0.11	0.12
		K	0.27	0.39	-0.05	0.24	-0.07	-0.17	-0.09	0.14
p	pp_int_collab	S	0.33	0.43	-0.08	0.21	0.22	0.06	0.04	0
		K	0.23	0.31	-0.08	0.19	0.17	0.04	0	0
mcs	mncs	S	0.71	0.74	0.90	0.15	0.42	0.27	0.55	0.83
		K	0.55	0.58	0.8	0.13	0.3	0.2	0.42	0.73
mcs	pp_top_prop	S	0.65	0.54	-0.20	0.20	0.20	0.15	0.40	0.64
		K	0.49	0.42	-0.18	0.17	0.14	0.11	0.3	0.55
mcs	pp_int_collab	S	0.36	0.33	0.39	0.14	0.10	0.29	0.05	0.26
		K	0.26	0.25	0.37	0.13	0.06	0.23	0.09	0.21
mncs	pp_top_prop	S	0.87	0.76	-0.20	0.30	0.51	0.67	0.41	0.33
		K	0.72	0.63	-0.17	0.24	0.38	0.48	0.31	0.3
mncs	pp_int_collab	S	0.3	0.3	0.42	0.24	0.45	-0.01	-0.41	-0.20
		K	0.21	0.22	0.36	0.19	0.32	-0.02	-0.29	-0.14
pp_top_ prop	pp_int_collab	S	0.29	0.32	-	-	0.57	-0.13	-0.07	-0.25
		K	0.22	0.27	-	-	0.44	-0.1	-0.03	-0.23

Table 14: Correlation coefficients Non-Health Professional.

Sciences

This is the biggest division in the data set. It contains 824 observations. Notice how the first four pairs and *(*(mcs, pp_int_collab)) contain a denser upper tail. Whilst the last two pairs have a denser middle and basically no tails. *(mcs,mncs)* and *(mncs,pp_top_prop)* have a dense middle with strongly correlated tails. *(mcs,pp_top_prop)* has a dense lower tail. Finally, the last pair is negatively correlated in the upper tail. Which means that researchers active in this field who have most of their publications in the top 10% of the most cited publications in their field tend to collaborate less on an international basis.

Variable one	Variable two	Method	Complete data set	All data in division	$\leq$ Q1	Q1 - Q2	Q2 -Q3	$\geq$ Q3	$\geq$ 90th percentile	$\geq$ 95th percentile
p	mcs	S	0.5	0.4	0.14	0.04	0.10	0.01	0.37	0.35
		K	0.36	0.28	0.1	0.03	0.07	0.01	0.25	0.23
p	mncs	S	0.31	0.27	0.09	0	0.03	-0.03	0.35	0.40
		K	0.22	0.19	0.06	-0.002	0.02	-0.02	0.25	0.28
p	pp_top_prop	S	0.38	0.27	0.10	0.05	0.02	-0.05	0.33	0.30
		K	0.27	0.19	0.07	0.04	0.01	-0.03	0.22	0.19
p	pp_int_collab	S	0.33	0.02	0.10	0	0.04	-0.02	0.12	0.31
		K	0.23	0.01	0.07	-0.01	0.03	-0.02	0.08	0.19
mcs	mncs	S	0.71	0.77	0.66	0.16	0.25	0.65	0.61	0.69
		K	0.55	0.59	0.48	0.11	0.17	0.49	0.47	0.52
mcs	pp_top_prop	S	0.65	0.69	0.53	0.14	0.21	0.44	0.29	0.13
		K	0.49	0.51	0.39	0.1	0.14	0.32	0.19	0.09
mcs	pp_int_collab	S	0.36	0.05	-0.06	0.04	0.04	0.18	0.19	0.15
		K	0.26	0.03	-0.04	0.03	0.03	0.12	0.13	0.1
mncs	pp_top_prop	S	0.87	0.91	0.54	0.56	0.46	0.62	0.42	0.16
		K	0.72	0.76	0.42	0.39	0.33	0.45	0.3	0.1
mncs	pp_int_collab	S	0.3	0.19	-0.02	0.08	0.10	0.11	-0.02	0.07
		K	0.21	0.13	-0.01	0.05	0.07	0.07	-0.01	0.04
pp_top_ prop	pp_int_collab	S	0.29	0.18	0.1	0.05	0.19	0.1	-0.06	-0.24
		K	0.22	0.13	0.07	0.03	0.12	0.07	-0.03	-0.16

Table 15: Correlation coefficients Sciences.

Social Sciences

The last division contains 500 scholars. The correlation coefficients are close to that of the complete data set. The dependence structures in these divisions are very divergent. The pair *(mcs, mncs)* has the same structure as in the previous divisions. Where the pair *(p,pp_top_prop)* has a slight lower tail and a negatively correlated upper tail, *(mncs,pp_prop_top)* has a dense upper tail. Notice how quite a few pairs have a negatively correlated upper tail. The first four and last two pairs have light lower tail, the middle is quite dispersed and the tails are negatively correlated. For these pairs it holds that high values for variable 1 go hand in hand with lower values for variable 2. The negative correlation coefficients aren't very high, so this implies a rather slight decrease. Pair six and eight have dense upper tails with a light lower tail and a denser middle than the other pairs. Finally, pair seven has light tails and a dispersed middle.

Variable one	Variable two	Method	Complete data set	All data in division	$\leq$ Q1	Q1 - Q2	Q2 -Q3	$\geq$ Q3	$\geq$ 90th percentile	$\geq$ 95th percentile
p	mcs	S	0.5	0.49	0.17	0.11	0	0.22	0.17	-0.31
		K	0.36	0.35	0.13	0.08	0.002	0.15	0.12	-0.21
p	mncs	S	0.31	0.3	0.22	0.08	-0.05	0.08	0.02	-0.34
		K	0.22	0.21	0.17	0.05	-0.04	0.05	0.01	-0.23
p	pp_top_prop	S	0.38	0.39	0.23	0.06	-0.01	0.11	0.05	-0.32
		K	0.27	0.28	0.2	0.04	-0.02	0.07	0.02	-0.22
p	pp_int_collab	S	0.33	0.3	0.32	0.10	0.10	0.07	-0.14	-0.02
		K	0.23	0.21	0.26	0.07	0.06	0.05	-0.09	0.01
mcs	mncs	S	0.71	0.77	0.59	0.27	0.47	0.55	0.69	0.58
		K	0.55	0.6	0.44	0.19	0.33	0.4	0.51	0.43
mcs	pp_top_prop	S	0.65	0.71	0.21	0.18	0.37	0.49	0.56	0.29
		K	0.49	0.54	0.17	0.13	0.26	0.34	0.41	0.22
mcs	pp_int_collab	S	0.36	0.37	0.37	-0.04	0.13	0.20	0.25	0.24
		K	0.26	0.26	0.28	-0.02	0.1	0.14	0.19	0.14
mncs	pp_top_prop	S	0.87	0.84	0.30	0.39	0.48	0.71	0.52	0.55
		K	0.72	0.7	0.24	0.28	0.34	0.54	0.38	0.44
mncs	pp_int_collab	S	0.3	0.36	0.34	0.02	0.09	0.13	-0.04	-0.33
		K	0.21	0.26	0.23	0.02	0.07	0.08	-0.04	-0.25
pp_top_ prop	pp_int_collab	S	0.29	0.33	-	0.18	0.12	0.14	-0.19	-0.12
		K	0.22	0.25	-	0.13	0.08	0.11	-0.13	-0.07

Table 16: Correlation coefficients Social Sciences.

5 Fitting copulas

We aim to find the parametric copula families that fit the data best. In this chapther we analyse the selection of the copula. Since the copulas are selected with the help of R, we start of by discussing the model used by R. Followed by some comments on the ties. Then the selected copulas are analysed using the *Goodness-of-Fit* test (GOF test) for copulas. The test compares the empirical copula with a given parametric copula derived under the null hypothesis. This leads to the rejection of some of the selected copulas. These rejected copulas are replaced by the empirical copula, since this should be a bitter fit according to the output the GOF test. Finally, we will analyse a few of these empirical copulas.

5.1 Model

The bivariate copulas are selected via the function *BiCopSelect* in R. This function can be found in the package *VineCopula*. The function selects the best fitting bivariate copula family for given pseudo-observations according to the AIC. The pseudo-observations are computed via the function *pobs*, which can also be found in the VineCopula package in R. The function transforms a random variable into normalized ranked data.
BiCopSelect uses pseudo-observations rather than the original dataset, due to Sklar's theorem, which states that the copula is a function of uniform margins. Therefore, the variables need to be transformed into normalized ranked data points.
The function selects the copula in a few steps. First, it fits all available parametric families of copulas. In R, more than 35 have been implemented. The corresponding parameters are obtained by pseudo-maximum likelihood estimation. That is, the margins are replaced by their empirical cumulative distribution functions (cdfs). Then the empirical cdfs are plugged into the copula density to calculate the estimate via

$$l(\theta) = \sum_{i=1}^{n} \log[c_\theta(\hat{F}_1(x_{i,1}), \hat{F}_2(x_{i,2})|\theta)],$$

where $\hat{F}$ denotes the marginal empirical cdf. $l(\theta)$ is then maximised, which leads to the desired parameter. The parameters for each available copula are computed. The available copulas are the copulas mentioned in paragraph 2.5, this includes the rotated versions of these copulas. Then Akaike Information Criteria, AIC, is computed for each copula as well as the rotated versions of these copulas.

Definition 5.1.1. *For observations $u_{i,j}, i = 1, \cdots, N, j = 1, 2$, the AIC of a bivariate copula family C with parameter(s) θ is defined as*

$$AIC := -2 \sum_{i=1}^{N} l(\theta) + 2k,$$

where $k = 1$ for one parameter copulas and $k = 2$ for two parameter copulas.

The copula with the lowest AIC-value is chosen as the best fitting bivariate copula.
Note that the function chooses a parametric copula from the available parametric families. The empirical copula is not considered in the fitting. The *Goodness-Of-Fit Tests for Copulas*, via the function *gofCopula* in the package *Copula*, accounts for the empirical copula. The tests are based on the empirical process

$$\mathbb{C}_n(\mathbf{u}) = \sqrt{n}(C_n(\mathbf{u}) - C_{\hat{\theta}}(\mathbf{u})), \quad (\mathbf{u}) \in [0, 1]^2.$$

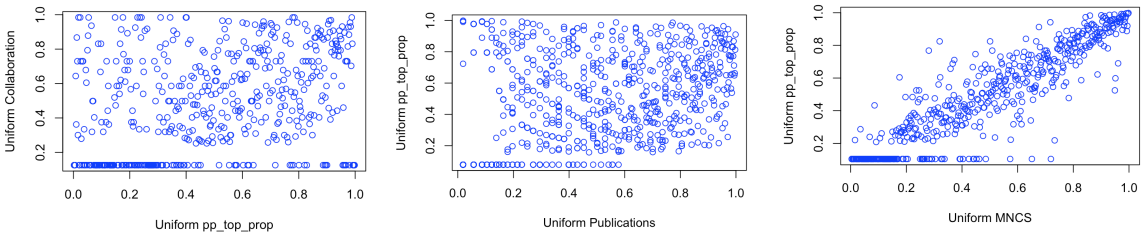
where $C_{\hat{\theta}}$ is an estimator of C under the hypothesis that $H_0 : C \in \{C_\theta\}$ holds and C_n denotes the empirical copula which R defines as

$$C_n(\mathbf{u}) = \frac{1}{n} \sum_{i=1}^{n} 1(U_i \leq u), (\mathbf{u}) \in [0,1]^2.$$

where U_i denote the the pseudo-observations for $i = 1, \cdots, n$ and sample size n. The test compares the empirical copula with a given parametric copula derived under the null hypothesis. This because the empirical copula process converges uniformly to the true copula.

5.2 Ties: To account or not to account for?

The *Goodness-Of-Fit tests* were derived under the assumption of continuous marginals. In other words, the assumption made is that ties occur with probability zero. In the previous chapter, we saw that in our data set this probability does not equal zero. For some divisions, in some ranges, the correlation values could not be computed because of these ties as mentioned in the previous chapter. In our data set, there is a non-negligible number of ties in every division between a lot of the variables, as can be seen quite clearly in figure 12.



(a) Division: Social Sciences. (b) Division: Basic Medical Sciences. (c) Division: Engineering

Figure 12: Ties in different divisions for different pairs of variables. Figure (a): *pp_top_prop* versus *pp_int_collab*. Figure (b): *p* versus *pp_top_prop*. Figure (c): *mncs* versus *pp_top_prop*.

Appendix B summarises the output of the copula fitting and the *Goodness of Fit (GOF) test* when the ties are unaccounted for. Lets look at the summary for the division *Education*

Variable 1	Variable 2	Copula	Parameter 1	Parameter 2 (or df)	Kendalls Tau	AIC	Satistic	p-value
p	mcs	Tawn T2	2.76	0.51	0.38	-17.01	0.014	0.9688
p	mncs	Survival Tawn T1	5.27	0.25	0.23	-13.72	0.015	0.8854
p	pp_top_prop	Clayton	1.71	-	0.46	-17.1	0.054	0.8854
p	pp_int_collab	Survival Joe	2.13	-	0.38	-10.18	0.0312	0.9688
mcs	mncs	Survival Tawn T1	4.78	0.86	0.7	-77.74	0.014	0.6354
mcs	pp_top_prop	Gumbel	1.88	-	0.47	-19.49	0.035	0.8438
mcs	pp_int_collab	Student t	0.3	2.04	0.29	-1.79	0.034	0.8021
mncs	pp_top_prop	Tawn T1	3.3	0.74	0.55	-38.75	0.025	0.8438
mncs	pp_int_collab	Student t	0.26	2	0.17	-1.98	0.027	0.8646
pp_top_ prop	pp_int_collab	Survival Tawn T1	5.01	0.68	0.57	-41.84	0.017	0.9896

Output Copula selection via R for the division Education.

Notice how most of the p-values are abnormally high. The presence of the ties in the data substantially affects the approximate p-values for all divisions. Luckily, this is a known problem. Kojadinovic and Yan (2010) suggest a way of dealing with these ties. According to them, the pseudo-observations should be constructed by randomly breaking the ties. The randomization does not change the results qualitatively, that is the parameter estimate is not effected by the randomization. They stress that ignoring the ties int the computation of the pseudo-observations, by using the default ranking method "average", leads to the rejection of a lot of well fitting copulas. This is caused by the assumption of continuous margins. The proof can be found in Kojadinovic and Yan (2010).

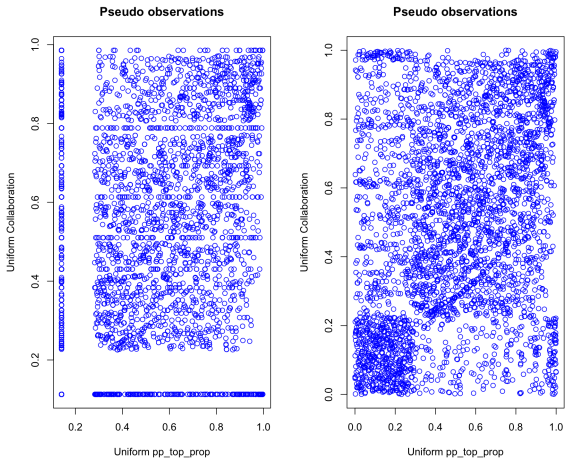


Figure 13: Not accounting for the ties (top) versus randomizing the ties (bottom) and thus accounting for the ties. Pseudo-observations of *pp_top_prop* versus *pp_int_collab*.

5.3 Copula Analysis

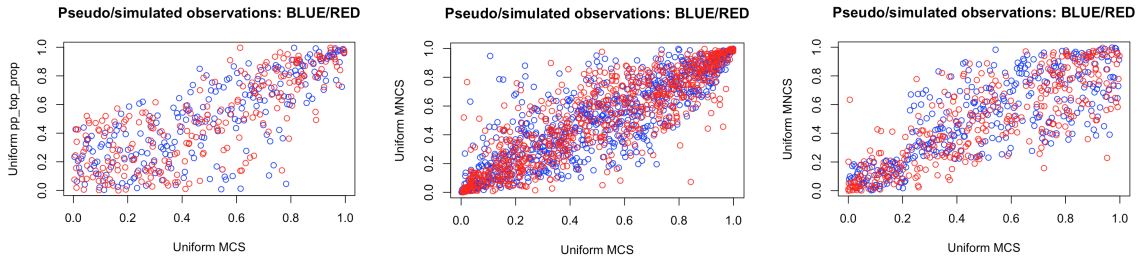
Now that the ties are accounted for, the p-values seem plausible. So let's look at the fitting. We will not evaluate every fitting of every bivariate copula. Instead, we will discuss main results of the fitting given that a lot of the results are similar to each other. We set the significance level at 5%. The output of the copula selection is summarized in tables 17-25. These tables can be found at the end of this paragraph.
Firstly, an important note on the values for Kendalls τ. After evaluating the values in tables 17-25 and comparing these with the values in 8-16 in paragraph 4.3., we can conclude that in most of the divisions these values are practically the same. There are a few exception. Most of these exceptions concern the division *Education*. Lets summarize these differences.

Variable 1	Variable 2	Kendalls τ data	Kendalls τ copula ties randomized	Kendall τ copula ties averaged
p	mcs	0.35	0.36	0.38
p	mncs	0.21	0.19	0.23
p	pp_top_prop	0.40	0.22	0.46
p	pp_int_collab	0.23	0.18	0.38
mcs	mncs	0.68	0.67	0.7
mcs	pp_top_prop	0.47	0.34	0.47
mcs	pp_int_collab	0.17	0.2	0.29
mncs	pp_top_prop	0.59	0.42	0.55
mncs	pp_int_collab	0.13	0.12	0.17
pp_top_prop	pp_int_collab	0.41	0.37	0.57

Different Kendall τ values for the division Education.

The third column represents the Kendall τ correlation coefficients which can also be found in paragraph 4.3. The fourth column represents the Kendall τ of the copulas when the ties are randomized. The fifth column represents the Kendall τ of the copulas when the ties are averaged, thus unaccounted for. Recall that the division *Education* has the smallest amount of observations. So randomizing the ties might have had a notable effect in this division, which leads to bigger differences between the values for Kendalls τ. However, not accounting for the ties created bigger differences between more variables, as we can see in the last column.
On that note, not accounting for the ties created (notable) differences between some of the *Kendall's τ correlation coefficients* in all the divisions. When the ties were randomized, all these differences decreased significantly.
There are three more notable differences in the Kentall τ values. One of these can be found in the division *Business & Management* between the variables *p* and *pp_int_collab*. Where we calculated a value of 0.04 in paragraph 4.3, it now is 0.15, see table 18. Two more can be fond in the division *Humanities* regarding the pairs *(mcs,pp_top_prop)* and *(mcs,pp_int_collab)*. In paragraph 4.3, we computed 0.39 and 0.27 respectively for these pairs. For the copulas these values are 0.27 and 0.17, see table 22. However, the copulas of these pairs are all rejected at the 5% significance level.
To continue our analysis, a visual aid is created. To create a visual aid, the pseudo-observations are be plotted. Via the *rCopula* function in the package *Copula*, pseudo-observations can be simulated for a given copula and coefficients. By plotting the pseudo-observations and the simulations in one figure, we obtain an insight into the structure of the selected copulas and simultaneously compare it to the structure of the pseudo-observations.
Unsurprisingly, some copulas are not rejected at the 5% significance level. In figure 14 such copulas are shown. The first is the BB8 copula with coefficients 4.34 and 0.91 and p-value equal to 0.07671. The second copula is the Student-t copula with a parameter value equal to 0.86 and

4 degrees of freedom. The p-value equals 0.2286. The last copula is the Tawn Type 2 copula with coefficients 3.35 and 0.78 and a p-value equal to 0.3471.



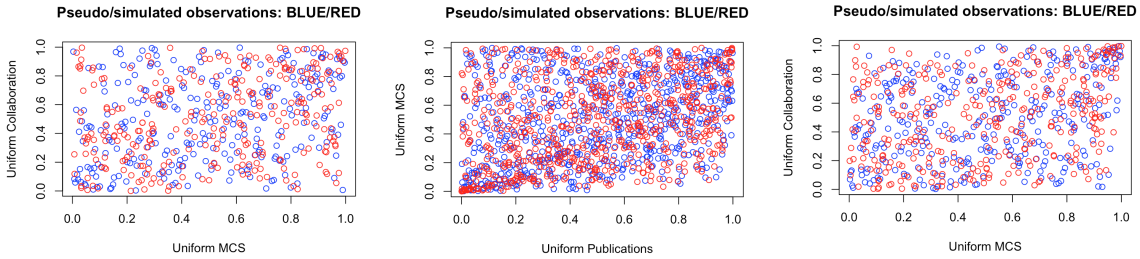
(a) Division: Business & Management.

(b) Division: Basic Medical Sciences.

(c) Division: Humanities.

Figure 14: Copulas where the pseudo-observations and simulation match. Figure (a): *mcs* versus *pp_top_prop*. Figure (b): *mcs* versus *mncs*. Figure (c): *mcs* versus *mncs*.

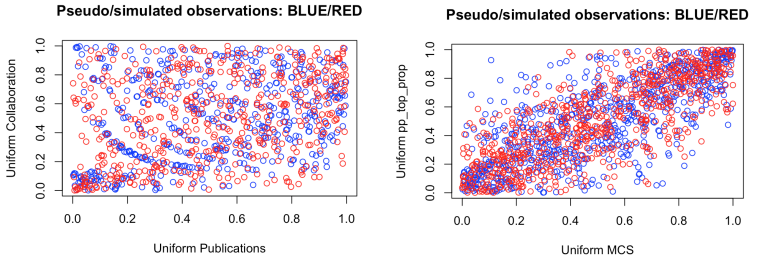
Figure 14 suggests that the pseudo-observations of the variables match the simulated pseudo-observations quite well. Note that the variables of the pseudo-observations were highly correlated according to the findings in the previous chapter. Let us compare the Kendall tau correlation coefficients between the pseudo-observations and the simulated observations. The Kendall tau correlation coefficients of the pseudo-observations were 0.63, 0.66 and 0.63 respectively. For the simulated observations these are 0.57, 0.66 and 0.6 respectively. So the Kendall tau correlation coefficients for both sets of observations are quite similar, as mentioned before.
More importantly, in the previous chapter we analysed the correlation coefficients and thus the dependence structure between variates via bins. This gave us a better understanding of the structure of the two variates. The question is, are these structures preserved? According to table 9, figure 14(a) should have a dispersed lower tail due to a low correlation coefficient, followed by a denser structure till the middle. A quarter of the structure after the middle was supposed to have more dispersed structure followed by yet again a denser quarter with a somewhat light upper tail. The copula does a nice job at capturing this structure, though the dense structure up until the middle isn't as dense as it should be. This could be a result of the randomness of the simulation. With every simulation, the data points change. The over al structure however, is quite what we want it to be. This definitely holds for figure 14(b) and 14(c).
Though the visual aids are quite informative, some copulas and structures are rather chaotic, as it can be seen in figure 15, that the p-value is the only indication as to wether the copula fits the data well or not. Note that the copulas in figure 15 have low correlation coefficients, 0.2, 0.35, 0.27 respectively for Kendall's tau. Which are related to the dispersed structure. The correlation coefficients are actually quite low in every bin, as seen in the previous chapter. Especially the middle of the structures are characterised in the previous chapter either by negative or positive correlation coefficients which are very close to zero.
Though the correlation coefficients are low, a parametric copula can still capture the dependence structure. Traditionally, these values are excluded from the model. But with the copula, even low correlated variates can be included. In citation analysis this is very helpful, since the publication variable is weakly correlated to every other variable. Naturally, not every low correlated pair of variables can be modelled with the help of a parametric copula. The copula in figure 15(c) is rejected with a 5% significance level.



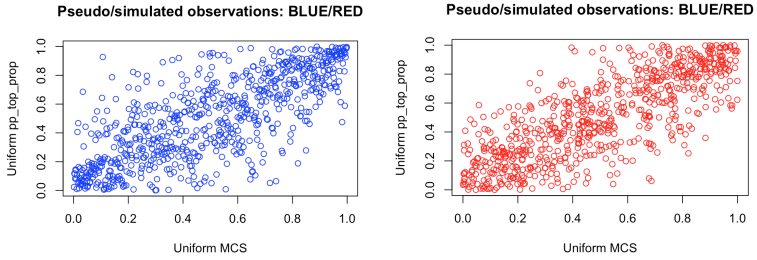
(a) Division: Health Sciences. (b) Division: Basic Medical Sciences. (c) Division: Humanities.

Figure 15: Copulas and pseudo-observations with a very dispersed structure. Figure (a): *mcs* versus *pp_int_collab*. Figure (b): *p* versus *mcs*. Figure (c): *mcs* versus *pp_int_collab*.

Furthermore, there are those copulas which obviously do not fit data well, figure 16(a) , and those which seem to fit the data well, but using hypothesis testing lead to the conclusion that the copula did not fit the data well. Figure 16(b) is a good example of this last situation. This is the Frank copula with coefficient 6.83. Only using the visual aid might lead to the conclusion that the copula is a good fit. However, the p-value of 0.000249 contradicts this conclusion. Now taking a closer look at the simulated observations and the pseudo-observations, see figure 16(c) and 16(d), it is clear that the tails are captured by the copula, but the middle of the pseudo-observations has a different structure all together.



(a) Division: Engineering. (b) Division: Basic Medical Sciences.

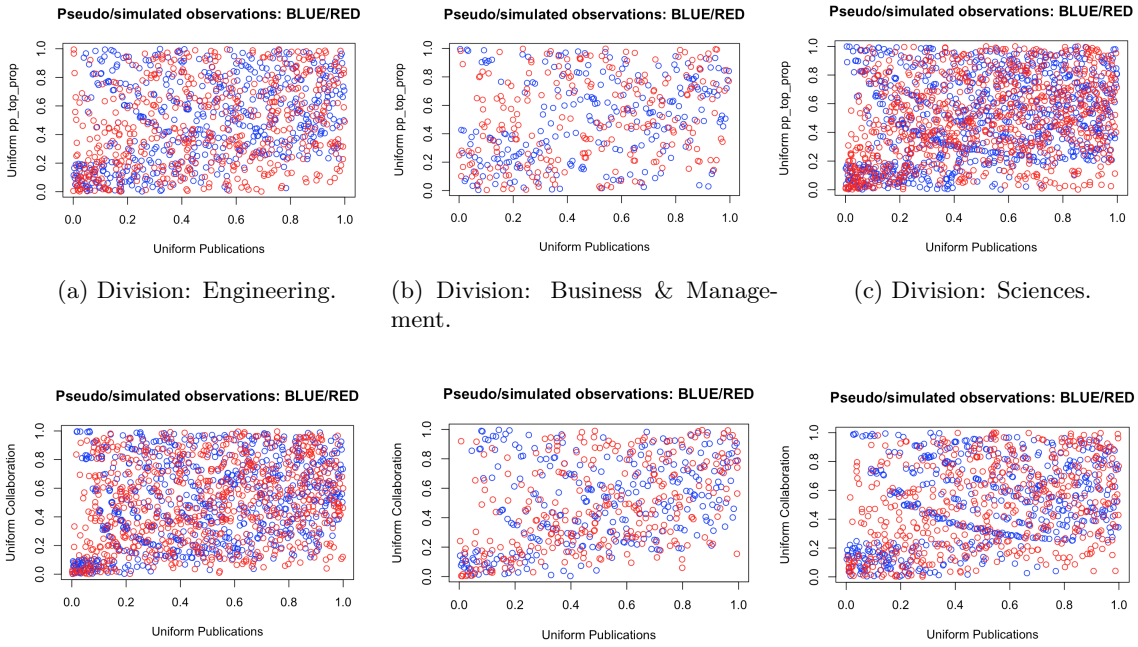


(c) Pseudo-observations of (b). (d) Simulated observations of (b).

Figure 16: Copulas rejected at the 5% significance level. Figure (a): *p* versus *pp_int_collab*. Figure (b), (c) and (d): *mcsversuspp_top_prop*.

Notice how in almost every division, except for the divisions *Education* and *Non-Health Professional*, the fitted parametric copulas for the pairs *(p,pp_top_prop)* and *(p,pp_int_collab)*

are rejected at the 5% significance level. According to the analysis in the previous chapter, the structures of these pairs differ per division. These pairs of variables are weakly correlated in every division. However, we've seen that the structure of weakly correlated variates can still be captured via a copula. Figure 17 sheds some light on the rejection of the copulas. Clearly, the variables in the divisions have a similar sort of structure, despite the different correlation coefficients. The differences in correlation coefficients is most likely created by the different amount of observations per division and by the slight differences in the structures. All with all, the overall structure of the variates is quite similar in every division. Which explains why the selected copulas are rejected at the 5% significance level in almost every division. It appears that the kind of structure we see in figure 17 is hard to capture with a parametric copula.



(a) Division: Engineering.

(b) Division: Business & Management.

(c) Division: Sciences.

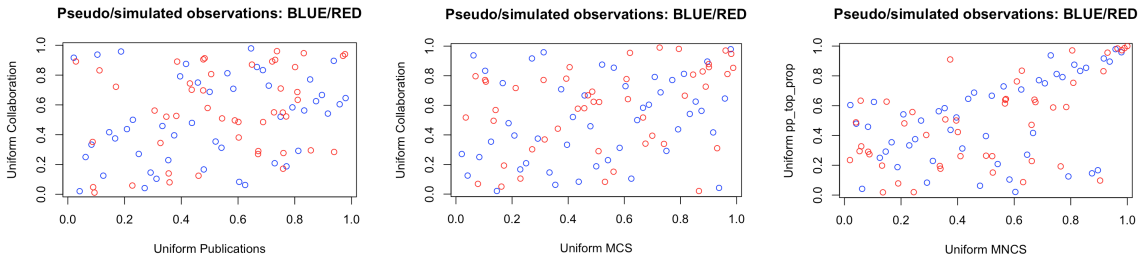
(d) Division: Basic Medical Sciences.

(e) Division: Health Sciences.

(f) Division: Social Sciences.

Figure 17: Copulas rejected at the 5% significance level for p versus *pp_top_prop* (a,b,c) and p versus *pp_int_collab* (d,e,f).

When zooming into the divisions, we see that Education and Non-Health Professional, the divisions with the lowest amount of observations, have only one copula rejected at the 5% significance level. From figure 18 and figure 19 no logical explanation can be given to explain the rejection of the Joe copula and the Survival BB8 copula when comparing it with the copulas which were not rejected at the 5% significance level. A possible drawback from the selected copulas in these divisions is that due to limited size of data, the selected copula can be unreliable.

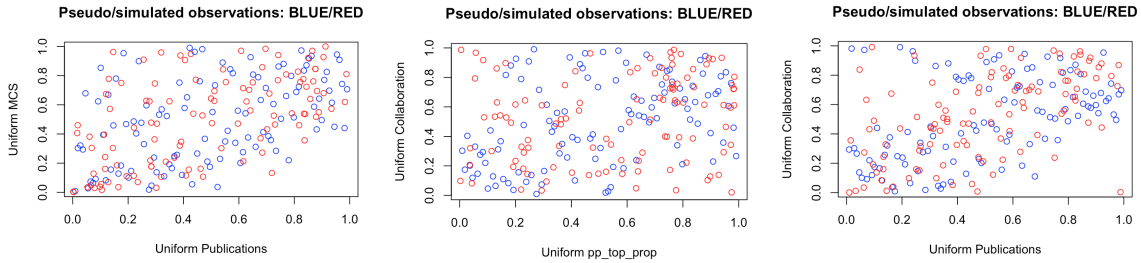


(a) Survival Tawn Type 1 copula (1.98, 0.26;0.2491).

(b) Gumbel copula (1.25;0.5365).

(c) Joe copula (2.34;0.04873).

Figure 18: Scatterplots of the pseudo- and simulated observations of the division Education for p versus *pp_int_collab* (a), *mcs* versus *pp_int_collab* (b) and *mncs* versus *pp_top_prop* (c). The numbers between the parentheses denote the parameters and the p-values.



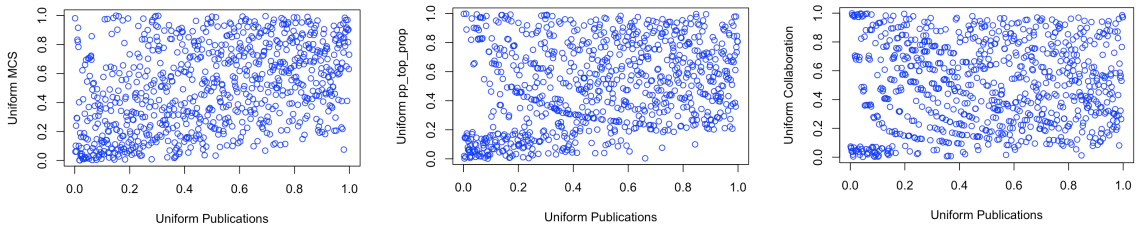
(a) Survival Tawn Type 1 copula (2.16, 0.46;0.553).

(b) Survival BB8 copula (3.06, 0.68;0.06472).

(c) Survival BB8 copula (3.34, 0.7;0.0002499).

Figure 19: Scatterplots of the pseudo- and simulated observations of the division Non-Health Professional for p versus *mcs* (a), *pp_top_prop* versus *pp_int_collab* (b) and p versus *pp_int_collab* (c). The numbers between the parentheses denote the parameters and the p-values.

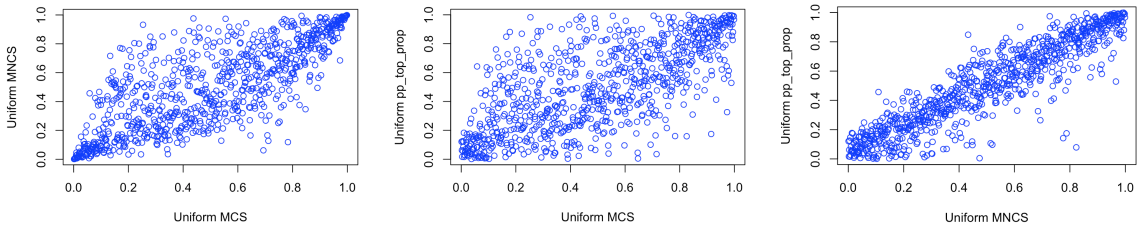
Sciences and Social Sciences have the most copulas rejected at the 5% significance level, 6 and 7 copulas respectively. Inspecting the dependence structure of these divisions, figure 20 and figure 21, give us more insight on this phenomenon. From figure 20, we can derive that most of the dependence structures between the variates with a rejected copula are rather hard to capture. The first three structures in figure 20 show a sort of exponential decrease in the first half of the structure. Followed by a very dispersed structure. It is therefore understandable that the fitted parametric copulas are rejected by the GOF test. Just like with the copula in figure 16(a). The last three structures however, seem very dense. A quick glance at the structures might imply that a parametric copula should be able to capture this structure. However, the selected copulas are rejected at the 5% significance level via the GOF test. So why is that? A glance at table 15, see the previous chapter, tell us that the middle part of the copulas are weakly correlated. The last structure has a negatively correlated weak lower tail and a weak upper tail with a weak correlated middle. However, figure 20(f) shows us quite a dense structure. So the low correlation coefficients might be a result of the data points constantly jumping from high correlated points to lower correlated points. It seems that the parametric copulas can not capture these kinds of structures. The same difficult structures can be seen in figure 21 for the division *Social Sciences*.



(a) Survival Tawn Type 1 copula (1.71, 0.47;0.0002249).

(b) Survival BB8 copula (1.57, 0.97;0.0002249).

(c) Student t copula (0.01;0.0002499).

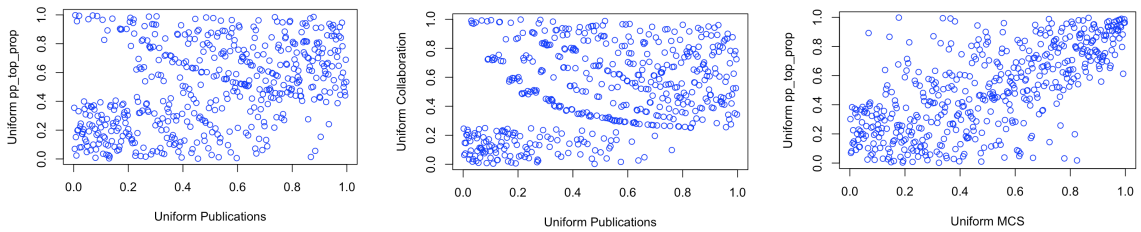


(d) BB7 copula (2.34, 1.75;0.01324).

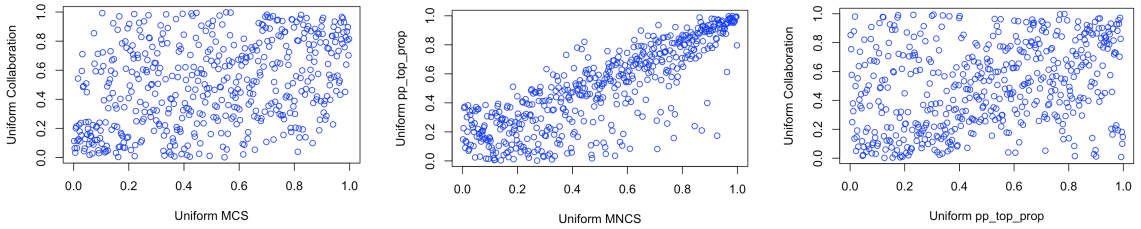
(e) Student t copula (0.69;0.01174).

(f) Frank copula (14.05;0.01774).

Figure 20: Scatterplots of the pseudo-observations of the division Sciences for p versus mcs (a), p versus pp_top_prop (b), p versus pp_int_collab (c), mcs versus $mncs$ (d), mcs versus pp_top_prop (e) and $mncs$ versus pp_top_prop (f). The selected copula is denoted underneath each structure. The numbers between the parenthesis denote the parameters and the p-values.



(a) Survival Tawn Type 1 copula (1.88,0.55;0.07221). (b) Survival BB8 copula (2.71,0.73;0.0002499). (c) Survival BB8 copula (1.75, 0.95;0.0002499).



(d) BB8 copula (5.21, 0.72;0.002249). (e) BB8 copula (5.37, 0.92;0.003748). (f) Frank copula (2.15;0.01074).

Figure 21: Scatterplots of the pseudo-observations of the division Social Sciences for p versus *pp_top_prop* (a), p versus *pp_int_collab* (b), *mcs* versus *pp_top_prop* (c), *mcs* versus *pp_int_collab* (d), *mncs* versus *pp_top_prop* (e) and *pp_top_prop* versus *pp_int_collab* (f). The selected copula is denoted underneath each structure. The numbers between the parenthesis denote the parameters and the p-values.

Variable 1	Variable 2	Copula	Parameter 1	Parameter 2 (or df)	Kendalls Tau	AIC	Satistic	p-value
p	mcs	Survival Tawn T1	1.74	0.4	0.23	-126.1522	0.016911	0.576
p	mncs	Survival Tawn T1	1.88	0.39	0.24	-155.046	0.026466	0.1767
p	pp_top_prop	Survival Tawn T1	1.87	0.42	0.25	-164.6077	0.10963	0.0002499
p	pp_int_collab	Survival Joe	1.5	-	0.22	-148.0063	0.090333	0.0002499
mcs	mncs	Student t	0.86	4.23	0.66	-990.2998	0.018266	0.2286
mcs	pp_top_prop	Frank	6.83	-	0.55	-564.8477	0.038988	0.004748
mcs	pp_int_collab	Student t	0.33	4.86	0.21	-93.67568	0.047312	0.003248
mncs	pp_top_prop	Frank	12.14	-	0.72	-1070.007	0.021769	0.05022
mncs	pp_int_collab	Student t	0.37	5.29	0.24	-116.7953	0.024184	0.1977
pp_top_ prop	pp_int_collab	Student t	0.37	3.56	0.24	-128.1458	0.027469	0.1277

Table 17: Output Copula selection via R for the division Basic Medical Sciences when ties are randomized.

Variable 1	Variable 2	Copula	Parameter 1	Parameter 2 (or df)	Kendalls Tau	AIC	Satistic	p-value
p	mcs	Survival Tawn T1	1.75	0.25	0.16	-21.41723	0.045643	0.03673
p	mncs	Survival Tawn T1	1.76	0.28	0.17	-26.08477	0.031067	0.1487
p	pp_top_prop	Survival BB8	2.17	0.75	0.2	-17.94222	0.063986	0.0002499
p	pp_int_collab	Survival Tawn T1	1.64	0.25	0.15	-17.02133	0.13419	0.0002499
mcs	mncs	Gaussian	0.93	-	0.75	-453.9142	0.0081444	0.7599
mcs	pp_top_prop	BB8	4.34	0.91	0.57	-225.8427	0.024564	0.07671
mcs	pp_int_collab	Frank	3.36	-	0.34	-58.73813	0.017671	0.485
mncs	pp_top_prop	BB8	6	0.94	0.69	-369.587	0.036092	0.003248
mncs	pp_int_collab	Survival Gumbel	1.37	-	0.27	-44.32918	0.019068	0.4515
pp_top_ prop	pp_int_collab	Student t	0.4	5.6	0.26	-38.07874	0.029902	0.08171

Table 18: Output Copula selection via R for the division Business & Mangement when ties are randomized.

Variable 1	Variable 2	Copula	Parameter 1	Parameter 2 (or df)	Kendalls Tau	AIC	Satistic	p-value
p	mcs	Tawn T2	2.76	0.48	0.36	-16.18246	0.017919	0.4775
p	mncs	Survival Tawn T1	5.23	0.21	0.19	-9.769757	0.024359	0.456
p	pp_top_prop	Tawn T2	3.26	0.25	0.22	-8.400577	0.032008	0.2631
p	pp_int_collab	Survival Tawn T1	1.98	0.26	0.18	-2.221652	0.021185	0.2491
mcs	mncs	Survival Tawn T1	4.19	0.86	0.67	-69.1505	0.01774	0.2701
mcs	pp_top_prop	Joe	1.92	-	0.34	-13.32335	0.02868	0.3696
mcs	pp_int_collab	Gumbel	1.24	-	0.2	-2.445932	0.020686	0.5365
mncs	pp_top_prop	Joe	2.34	-	0.42	-24.49909	0.052711	0.04873
mncs	pp_int_collab	Student t	0.18	2	0.12	-5.891713	0.021053	0.539
pp_top_ prop	pp_int_collab	Frank	3.73	-	0.37	-12.20947	0.030446	0.2791

Table 19: Output Copula selection via R for the division Education when ties are randomized.

Variable 1	Variable 2	Copula	Parameter 1	Parameter 2 (or df)	Kendalls Tau	AIC	Satistic	p-value
p	mcs	Survival Tawn T1	1.81	0.39	0.23	-99.81123	0.040872	0.02174
p	mncs	Survival Tawn T1	1.78	0.35	0.21	-88.73386	0.021432	0.3611
p	pp_top_prop	Survival BB8	1.74	0.94	0.24	-84.02995	0.10569	0.0002499
p	pp_int_collab	Survival Tawn T1	1.65	0.37	0.2	-69.00649	0.06236	0.002249
mcs	mncs	BB1	0.96	1.79	0.62	-628.7733	0.0072446	0.9758
mcs	pp_top_prop	BB1	0.14	1.89	0.5	-374.6893	0.036506	0.01024
mcs	pp_int_collab	Student t	0.16	6.12	0.1	-14.70836	0.027136	0.1147
mncs	pp_top_prop	BB8	6	0.92	0.67	-801.8934	0.089092	0.0002499
mncs	pp_int_collab	Student t	0.23	5.59	0.15	-30.8334	0.020154	0.3731
pp_top_ prop	pp_int_collab	Student t	0.25	4.04	0.16	-46.86273	0.02062	0.3481

Table 20: Output Copula selection via R for the division Engineering when ties are randomized.

Variable 1	Variable 2	Copula	Parameter 1	Parameter 2 (or df)	Kendalls Tau	AIC	Satistic	p-value
p	mcs	Survival Tawn T1	1.88	0.5	0.29	-78.03466	0.012726	0.8223
p	mncs	Survival Tawn T1	2.18	0.37	0.26	-78.13603	0.010354	0.9583
p	pp_top_prop	Survival Tawn T1	1.79	0.45	0.25	-59.00139	0.045102	0.01574
p	pp_int_collab	Survival Tawn T1	1.84	0.44	0.26	-61.6409	0.041164	0.02874
mcs	mncs	Tawn T2	3.35	0.78	0.58	-335.7209	0.016658	0.3471
mcs	pp_top_prop	Student t	0.7	3.59	0.49	-191.64	0.028969	0.07721
mcs	pp_int_collab	Student t	0.3	6.08	0.19	-24.42149	0.020347	0.3751
mncs	pp_top_prop	Frank	12.62	-	0.72	-447.7892	0.020366	0.7771
mncs	pp_int_collab	Student t	0.25	4.82	0.16	-20.37149	0.019847	0.4125
pp_top_ prop	pp_int_collab	Student t	0.24	3.2	0.16	-23.18173	0.020395	0.403

Table 21: Output Copula selection via R for the division Health Sciences when ties are randomized.

Variable 1	Variable 2	Copula	Parameter 1	Parameter 2 (or df)	Kendalls Tau	AIC	Satistic	p-value
p	mcs	Tawn T2	1.38	0.28	0.12	-17.76798	0.037807	0.1047
p	mncs	Clayton	0.19	-	0.09	-6.82957	0.10636	0.0002499
p	pp_top_prop	Frank	0.95	-	0.11	-6743335	0.093069	0.0002499
p	pp_int_collab	Tawn T2	1.42	0.3	0.14	-17.9313	0.024802	0.2211
mcs	mncs	Survival BB8	5.56	0.84	0.6	-357.2335	0.045178	0.006747
mcs	pp_top_prop	Tawn T1	2.23	0.38	0.27	-95.75013	0.065881	0.005247
mcs	pp_int_collab	Joe	1.37	-	0.17	-39.44886	0.018035	0.6034
mncs	pp_top_prop	BB8	3.13	0.99	0.52	-305.8094	0.0778	0.0002499
mncs	pp_int_collab	Survival Tawn T2	1.37	0.26	0.11	-11.76515	0.014817	0.7794
pp_top_ prop	pp_int_collab	Indep	0	-	0	0	-	-

Table 22: Output Copula selection via R for the division Humanities when ties are randomized.

Variable 1	Variable 2	Copula	Parameter 1	Parameter 2 (or df)	Kendalls Tau	AIC	Satistic	p-value
p	mcs	Survival Tawn T1	2.16	0.46	0.31	-28.78789	0.017547	0.553
p	mncs	Survival Tawn T1	2.04	0.54	0.33	-32.22176	0.020281	0.407
p	pp_top_prop	Survival BB8	6	0.44	0.31	-21.74635	0.025323	0.2201
p	pp_int_collab	Survival BB8	3.34	0.7	0.32	-22.7634	0.070271	0.0002499
mcs	mncs	Survival Joe	3.62	-	0.58	-123.1726	0.015383	0.4645
mcs	pp_top_prop	Frank	3.66	-	0.36	-20.96846	0.023719	0.2251
mcs	pp_int_collab	Frank	1.93	-	0.21	-8.26	0.018818	0.4655
mncs	pp_top_prop	Joe	2.93	-	0.51	-91.62874	0.038625	0.1027
mncs	pp_int_collab	Frank	1.77	-	0.19	-6.66	0.027042	0.1477
pp_top_ prop	pp_int_collab	Survival BB8	3.06	0.68	0.27	- 16.2187	0.026859	0.06472

Table 23: Output Copula selection via R for the division Non-Health Professional when ties are randomized.

Variable 1	Variable 2	Copula	Parameter 1	Parameter 2 (or df)	Kendalls Tau	AIC	Satistic	p-value
p	mcs	Survival Tawn T1	1.71	0.47	0.25	-173.5842	0.052475	0.0002249
p	mncs	Survival Tawn T1	1.7	0.32	0.19	-118.3297	0.033738	0.06622
p	pp_top_prop	Survival BB8	1.57	0.97	0.21	-115.5426	0.12114	0.0002499
p	pp_int_collab	Student t	0.01	3.29	0.01	-55.29731	0.12149	0.0002499
mcs	mncs	BB7	2.34	1.75	0.59	-960.8376	0.035963	0.01324
mcs	pp_top_prop	Student t	0.69	12.09	0.49	-524.7775	0.036798	0.01174
mcs	pp_int_collab	Survival Tawn T1	1.42	0.1	0.06	-13.91832	0.033245	0.1172
mncs	pp_top_prop	Frank	14.05	-	0.75	-1422.191	0.024147	0.01774
mncs	pp_int_collab	Student t	0.21	5.41	0.13	-52.29152	0.03157	0.05522
pp_top_ prop	pp_int_collab	Student t	0.19	4.56	0.12	-53.36074	0.017931	0.5015

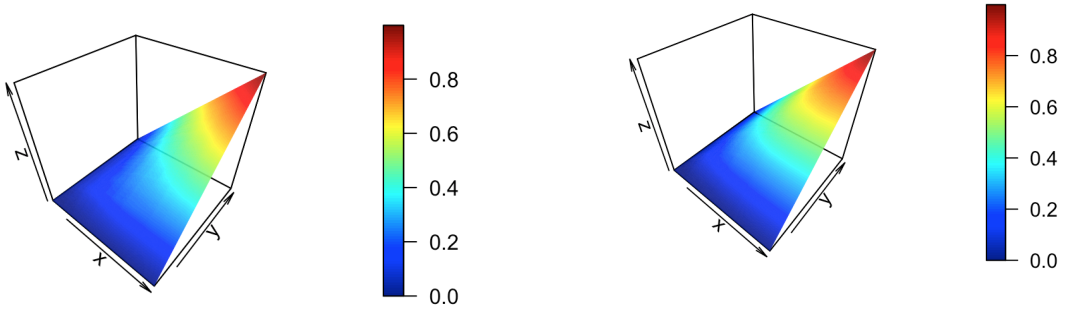
Table 24: Output Copula selection via R for the division Sciences when ties are randomized.

Variable 1	Variable 2	Copula	Parameter 1	Parameter 2 (or df)	Kendalls Tau	AIC	Satistic	p-value
p	mcs	Survival Tawn T1	1.88	0.55	0.31	-160.94	0.031601	0.07221
p	mncs	Survival Tawn T1	1.75	0.4	0.23	-92.46801	0.024877	0.2076
p	pp_top_prop	Survival BB8	2.71	0.73	0.26	-79.58566	0.11406	0.0002499
p	pp_int_collab	Survival BB8	1.75	0.95	0.24	-87.28426	0.14966	0.0002499
mcs	mncs	Survival Tawn T1	3.36	0.8	0.59	-609.3139	0.017337	0.2786
mcs	pp_top_prop	BB8	5.21	0.72	0.49	-306.9911	0.038543	0.002249
mcs	pp_int_collab	Survival BB8	2.01	0.88	0.25	-81.57794	0.035704	0.004248
mncs	pp_top_prop	BB8	5.37	0.92	0.65	-648.6196	0.037727	0.003748
mncs	pp_int_collab	Survival BB8	2.29	0.81	0.25	-77.27636	0.029159	0.03573
pp_top_ prop	pp_int_collab	Frank	2.15	-	0.23	-54.33529	0.042072	0.01074

Table 25: Output Copula selection via R for the division Social Sciences when ties are randomized.

5.4 The empirical copula

Since more than half of the selected parametric copulas are rejected at the 5% significance level, it might be interesting evaluate the empirical copula of these variables. As said before, the empirical copula converges to the true copula. In other words, the empirical copula converges to the real underlying dependence structure.
In the previous paragraph we evaluated the Survival Tawn T1 copula with coefficients 1.65 and 0.37. This copula captured the dependence structure between variables publication and international collaboration of the division Engineering. The Frank copula with coefficient 6.83 was also evaluated. This captured the dependence structure between variables *mcs* and *pp_top_prop* of the division Basic Medical Sciences. Both copulas are rejected at the 5% significance level. It seems that no parametric copula in the package fits the data well enough. That is why we rely on the empirical copulas.

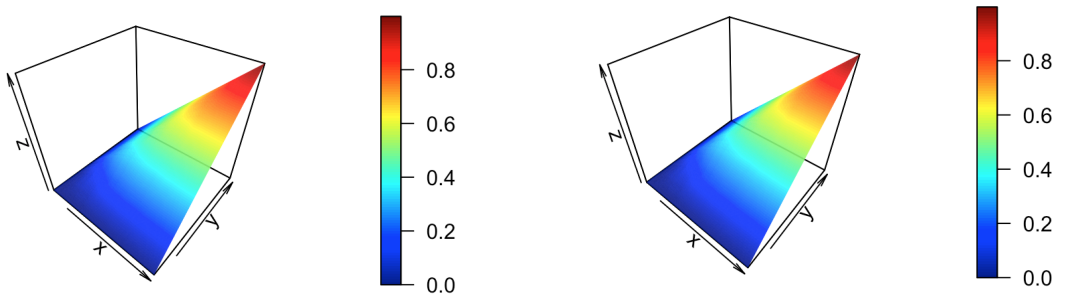


(a) Variates: *p* and *pp_int_collab*. Division: Engineering.

(b) Variates: *mcs* and *pp_top_prop*. Division: Basic Medical Sciences.

Figure 22: 3D-plot of the empirical copulas.

In figure 15(a) we saw a certain decrease in the first part of the figure. Low publication amounts had high collaboration scores. As the publication amount increased, the collaboration scores decreased. This structure is captured by the empirical copula in figure 22(a). The dark blue color is higher for low x values and high y values and lower for high x values and low y values. Figure 15(b) had a more dense structure, with some linear decreases in the middle of the structure. According to the correlation coefficients, the tails are slightly more correlated. This can also be seen in figure 22(b), where the red area at the top is bigger than in figure 22(a). Let's also plot the empirical copula for two copulas in the division Social Sciences, since this was one of the divisions with the most rejected copulas.

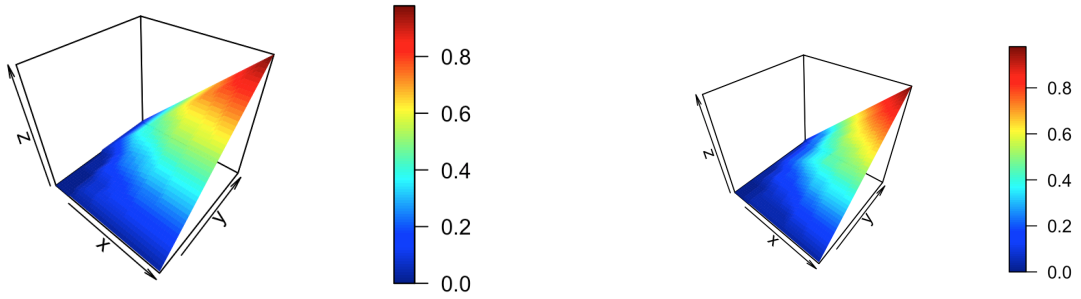


(a) Variates: p and *pp_top_prop*. Division: Social Sciences.

(b) Variates: *mcs* and *pp_int_collab*. Division: Social Sciences.

Figure 23: 3D-plot of the empirical copulas.

Figure 23(a) emphasises this quite nice with the bigger blue-green color. The decrease Engineering show in the first half of the structure in figure 15(a), Social Sciences shows in the middle of the structure in figure 23(a). Again the empirical copula captures this quite nicely.
With these results, it seems sensible to evaluate the dependence structure in the division *Education* and *Non-Health Professional*. Figure 24, shows us that the dependence structure for two pairs of variables in the division *Education* is also nicely captured. The negative correlation in the lower tail is nicely captured as we can see in figure 24(b). Furthermore, stronger correlation in the middle and upper tail is emphasised in figure 24(a).



(a) Variates: *mncs* and *pp_top_prop*. Division: Education.

(b) Variates: *MCS* and *pp_int_collab*. Division: Education.

Figure 24: 3D-plot of the empirical copulas.

6 Cross-validation

In chapter 5 we have created various copula models. Since it is not possible to collect new data with which we can asses these models, we use model validation, specifically k-fold cross-validation. Model validation is a general strategy which is used to evaluate the accuracy of the fitted models. Model validation uses part of the data, called the training subsample, to specify a statistical model. The statistical model is then evaluated using the remaining part of the data, called the validation subsample. Cross-validation is an application of this simple idea. The idea might be simple, but it is very powerful. In cross-validation the roles of training and validation subsamples are rotated.
We consider two variants on cross-validation. Suppose that model M is one of the models under consideration. Suppose that we have n observations. The data set is shuffled randomly. *Leave-one-out cross-validation* fits the model n times, omitting the ith observation at step i. The resulting fitted model is then used to obtain a predicted value for the omitted observation.
k-fold cross validation partitions the original data into k equal sized subsamples. One subsample is now the validation subsample and the remaining k-1 subsamples are considered the training sample. The validation process is now repeated k times. For example, a 5-fold cross validation partitions the data into 5 equally sized subsamples and uses 1 for validation and 4 for training. We therefore use 80% of the data for training and 20% of data for validation. This is the standard approach for validation. The validation and training subsamples are rotated 5 times, such that the validation process is repeated 5 times.
To summarize the skill of a model, the Cross-Validation Copula Information Criterion (CIC) is used. In R, the CIC is computed with the help of the function *xvCopula* in the package *Copula*. The function computes the cross-validation criterion for a given parametric copula family, which serves as the model, and a given data set. The criterion is a crossvalidated log likelihood which is denoted as

$$xv_n = n^{-1} \sum_{i=1}^{n} \log f_{\hat{\alpha}(i)}(X_i),$$

where $\hat{\alpha}(i)$ is the maximum likelihood estimate based on the sample without the ith observation, or without the validation subsample in case of the k-fold cross-validation. A parametric copula model is selected when said CIC value is higher for that parametric copula model compared to other parametric copula models.
We applied k-fold cross validation to the selected copula models in chapter 5, which were not rejected at a 5% significance level. We evaluate about 14 parametric copula families to see if the selected copulas in chapter 5 performs better than the other 14 parametric copulas with respect to cross-validation. The 14 parametric copula families are selected based on the results in chapter 5. These are the copulas that were selected as the best fitting copulas in chapter 5. We apply the leave-one-out cross-validation, the 5-fold cross-validation and the 10-fold cross-validation. By evaluating the results of the three different methods, we try to decrease the bias and variability in the output. A lower value for k tends to be produce more biased outputs. A higher value of k tends to be less biased, but suffers from large variability. We discuss the main results of the validation.

Table 26 summarizes the CIC values for the division *Basic Medical Sciences*. The left table summarizes the CIC values of the copulas based on the pair *(p,mcs)*. In chapter 5 we concluded that the Survival Tawn Type 1 copula captures the dependence structure of the data best. In table 26, this copula has the highest CIC value. Therefore the best fitting copula is also the best performing copula in terms of prediction. The same can be said for the right table. Here the

pair *(mncs, pp_top_prop)* is used. In chapter 5, we concluded that the Frank copula captured the dependence structure of the data best. The CIC values in table 26 lead to the same conclusion. This pair had a lower p-value compared to the first pair. However, we see that this has no influence on the output of the validation.

Copula	k=NULL	k=5	k=10
Joe	11.46857	11.54671	10.5126
Gumbel	24.68732	23.33006	25.18468
Frank	36.07373	35.73548	35.9391
Gaussian	35.51154	36.50794	35.65344
Student t	39.05237	40.3817	39.17311
BB1	46.42834	47.25591	47.5371
BB7	46.33887	47.58433	45.92197
BB8	33.99079	34.56649	34.46198
Tawn T1	11.20764	12.45613	9.89789
Tawn T2	37.28916	37.22415	37.3777
Survival	62.72214	62.04656	61.07137
Tawn T1			
Survival	24.39661	25.12961	24.50947
Tawn T2			
Survival Joe	45.64724	45.75801	45.55043
Survival BB8	46.51687	43.6569	46.48376

Copula	k=NULL	k=5	k=10
Joe	342.0062	342.137	341.9459
Gumbel	422.9221	423.0179	424.0997
Frank	519.7074	520.1073	518.7188
Gaussian	443.1522	441.774	443.0874
Student t	448.9554	448.9692	448.4317
BB1	428.5745	425.2616	428.8958
BB7	371.8547	376.4652	370.3673
BB8	476.6065	477.5237	476.0359
Tawn T1	432.6156	434.9178	434.8342
Tawn T2	404.112	407.8196	407.7448
Survival	392.5516	392.3746	391.1691
Tawn T1			
Survival	409.2426	410.5376	406.9165
Tawn T2			
Survival Joe	303.0652	298.306	301.5701
Survival BB8	452.363	451.3884	452.5869

Table 26: Cross validation copula information criterion (CIC) for the division *Basic Medical Sciences*. Left, the parametric copula family of the pair *p,mcs* is tested. Right, the parametric copula family of the pair *mncs,pp_prop_top* is tested. **k=NULL** corresponds to leave-one-out cross-validation.

Basic Medical Sciences provides us with the second highest amount of observations. Could the fact that the division *Basic Medical Sciences* has a lot of observations, influence the outcome of the validation? Let's look at a division with a lower amount of observations. The division *Health Sciences* has a lower amount of observations and coincidently has very high p-values compared to the other divisions. After testing the selected copula with the highest p-value, the copula of the pair *(p,mncs)*, the CIC values in table 27 concludes that the Survival Tawn Type 1 copula performs best by far. However, the copula of the pair *(mcs,pp_top_prop)* also has a high p-value. In table 21 we see that the Student t copula was selected based on a AIC value of -191.64 and a p-value of 0.771. However, the CIC values in table 27 implies that Frank copula performs better. The Frank copula with parameter value 6.01, $\tau = 0.51$ and AIC = -184.09 performs better according to cross-validation. So apparently, the best fitting copula does not always perform the best during validation.

Copula	k=NULL	k=5	k=10
Joe	8.694232	10.72067	7.616167
Gumbel	15.34071	14.85883	15.38766
Frank	18.54779	17.6464	18.22014
Gaussian	19.50865	20.78941	18.73398
Student t	21.62299	21.21289	21.29085
BB1	23.63973	23.49915	24.17109
BB7	23.13344	22.63188	24.03625
BB8	17.57571	18.01827	17.93214
Tawn T1	7.721054	8.845138	8.336312
Tawn T2	21.47128	20.85684	20.38332
Survival	38.42674	39.12844	38.808
Tawn T1			
Survival	11.3148	11.884	11.43523
Tawn T2			
Survival Joe	22.56224	20.89371	22.50804
Survival BB8	23.56824	21.88846	23.38785

Copula	k=NULL	k=5	k=10
Joe	62.99776	63.61573	65.00216
Gumbel	76.8275	73.9176	77.82877
Frank	86.67132	83.18489	88.30059
Gaussian	65.17572	69.62349	68.23418
Student t	84.03488	84.27122	86.63768
BB1	78.31184	79.43975	76.03016
BB7	70.2493	72.49398	72.09928
BB8	86.99812	86.3268	87.52485
Tawn T1	63.04824	61.47155	62.66749
Tawn T2	82.56082	82.54906	82.39562
Survival	69.28454	68.33479	69.5405
Tawn T1			
Survival	62.17672	60.73671	61.31719
Tawn T2			
Survival Joe	49.45072	48.00337	49.4406
Survival BB8	82.09392	82.47017	81.8114

Table 27: Cross validation copula information criterion (CIC) for the division *Health Sciences*. Left, the parametric copula family of the pair *p,mncs* is tested. Right, the parametric copula family of the pair *mcs,pp_top_prop* is tested. **k=NULL** corresponds to leave-one-out cross-validation.

Now another interesting division to discuss is the division *Education*. The smallest division with regard to the amount of observations. Also the division with the highest amount of parametric copulas which were not rejected at a 5% significance level. The same conclusion can be drawn for this division as for the division *Health Sciences*. We took the pair of variables for which the selected copula had the highest p-value, and a random pair with a lower p-value (but which was not rejected by the 5% significance level). The left table summarizes the CIC values of the pair of which the copula had the highest significance level and of course the highest AIC value. The copula selected in chapter 5 was the Student t copula with an AIC value of -5.89. However, table 28 suggests that the Survival Tawn Type 1 copula performs better according to its CIC value of 0.17. The Survival Tawn Type 1 copula with parameters 1.63 and 0.33, $\tau = 0.18$ and AIC = -0.13 performs better. Significantly better when the compare the CIC value of the Student t copula with that of the Survival Tawn Type 1 copula. The second pair, *(mncs,pp_int_collab)*, has a similar conclusion. Here we selected the Frank copula in chapter 5 with a AIC equal to -12.2. However the CIC values in table 28 suggest that the Tawn Type 2 copula performs better. The Frank copula was selected over the Tawn Type 2 (parameters 2.15 and 0.53, $\tau = 0.34$) copula because the Tawn Type 2 copula has an AIC equal to -11.61. The results of cross-validation applied to this division lead to the conclusion that the copulas fitted in chapter 5 for this division are indeed unreliable due to the low amount of observations.

Copula	k=NULL	k=5	k=10
Joe	-2.780996	-4.093931	-4.299342
Gumbel	-2.824813	-0.09783391	-0.8871302
Frank	-1.88421	-2.403815	0.06381934
Gaussian	-1.584622	-3.323312	-1.014183
Student t	0.03907034	0.6989769	-1.672064
BB1	-1.954051	-0.9108673	-0.9446276
BB7	-2.405286	-3.212366	-1.853746
BB8	-2.934236	-1.466141	-1.216325
Tawn T1	-2.380646	-5.368764	-0.5944863
Tawn T2	-0.6672554	-2.565152	-0.9322962
Survival	-0.9463774	1.649495	0.1702685
Tawn T1			
Survival	-1.034189	-1.411548	-1.885412
Tawn T2			
Survival BB8	-1.099357	-0.260566	-0.412445

Copula	k=NULL	k=5	k=10
Joe	2.270717	1.417477	3.99659
Gumbel	5.809938	7.382876	7.410219
Frank	7.550381	7.960839	8.258357
Gaussian	6.466112	6.358307	6.455025
Student t	5.417686	4.980025	4.686118
BB1	5.915197	4.85575	4.852217
BB7	2.492363	0.748625	7.05661
BB8	6.857144	7.324729	5.66423
Tawn T1	1.815204	2.95713	3.043191
Tawn T2	9.486488	9.482405	9.002856
Survival	7.355279	8.232456	7.995707
Tawn T1			
Survival	3.982362	4.652753	3.940191
Tawn T2			
Survival BB8	8.040412	9.167627	8.285998

Table 28: Cross validation copula information criterion (CIC) for the division *Education*. Left, the parametric copula family of the pair *mncs,pp_int_collab*. Right, the parametric copula family of the pair *pp_top_prop,pp_int_collab* is tested. **k=NULL** corresponds to leave-one-out cross-validation.

Up until now, the division *Basic Medical Sciences* is the only division where the outcome of the CIC values corresponds with the output of chapter 5. Lowering the amount of observations by testing divisions with lower amounts of observations leads to the conclusion that the best fitting copulas do not perform the best when cross-validation is applied. Do the amounts of observations really influence the outcome? To answer this question, let's discuss the output of the biggest data set. *Sciences* is the division with the largest amount of observations. It is also the set with the highest amount of rejected copulas (at a 5% significance level). Since only 3 parametric copulas seem to fit, let's evaluate two of these. The left table summarizes the results when we use the data set of the pair *(mcs,pp_int_collab)*. Chapter 5 concluded that the Survival Tawn Type 1 copula with AIC -13.92 captured the dependence structure of the data best. According to table 29, the Tawn Type 2(parameters 1.44 and 0.07, $\tau = 0.05$ and AIC -9.35) copula has the highest CIC, followed by the Survival Tawn Type 1 copula. However, the Student t copula captured the dependence structure of the pair *(mncs, pp_int_collab)* best, with AIC equal to -52.29. The CIC values in table 29 correspond with this.
So the amount of observations plays no role in performance of the fitted copula during cross-validation. Of course higher amounts of observations are always preferred over low amounts of observations to create better models and predict more accurately. In this case however, the conclusion is simply that fitted copula models do not always perform best during cross-validation.

Copula	k=NULL	k=5	k=10
Joe	-2.47885	-2.564808	-2.601394
Gumbel	-2.334628	-3.107021	-2.597303
Frank	-0.3889486	0.7480375	-0.008019788
Gaussian	-1.341627	-1.270874	-1.531591
Student t	0.8579793	-0.4716252	2.316804
BB1	-1.069603	-2.227069	-0.6496148
BB7	-1.007821	-0.5504916	-0.3457765
BB8	-0.2542654	0.3214533	-1.267824
Tawn T1	0.01317877	-184.4099	-0.001829289
Tawn T2	4.473062	5.603481	3.48789
Survival	3.018965	3.425729	2.5542
Tawn T1			
Survival	-1.016091	0.019811	-0.2631417
Tawn T2			
Survival Joe	1.501441	2.176645	0.5183629
Survival BB8	1.400032	2.035923	1.487525

Copula	k=NULL	k=5	k=10
Joe	0.9113788	-0.04247094	2.088985
Gumbel	6.433094	6.481505	6.239449
Frank	15.1116	14.32442	14.36364
Gaussian	9.957936	10.71551	10.69402
Student t	18.40011	19.32587	18.18832
BB1	8.657472	11.04789	8.216903
BB7	6.894428	5.391992	6.318198
BB8	17.25678	17.80591	17.45637
Tawn T1	-1.007846	-2.707016	-1.9213
Tawn T2	11.25452	10.31695	12.38562
Survival	15.52498	13.84742	15.41771
Tawn T1			
Survival	10.06837	9.292335	10.66625
Tawn T2			
Survival Joe	11.48832	11.57316	11.57578
Survival BB8	14.69565	13.67579	14.45242

Table 29: Cross validation copula information criterion (CIC) for the division *Sciences*. Left, the parametric copula family of the pair *mcs,pp_int_collab*. Right, the parametric copula family of the pair *mncs,pp_int_collab* is tested. **k=NULL** corresponds to leave-one-out cross-validation.

7 Conclusions

In general, the dependence structure between variables cannot be distinguished on the grounds of correlation coefficients alone. This also holds for the dependence structure between the publications of a researcher and the citations of those publications. Copulas are a useful tool to model the dependence between random variables. Especially when coping with highly skewed data. From the visual analysis and the correlation coefficients of the five variables used in this thesis, it becomes clear that the variables are highly skewed and contain a lot of ties. We can speak of non-normality with respect to the distribution. The correlation coefficients between most of the variables are rather low and do not elaborate on the dependence structures. A deeper analysis is created by binning the data and calculating the correlation coefficients of each bin. This is done for each of the nine fields. An image of the dependence structure between the pairs of variables is created via these correlation coefficients. This serves as a good reference when analysing the fitted copula models.

Before fitting copulas to the data, it is very important to inspect the data on ties. We saw that in our data set there were a non-negligible number of ties in every division between a lot of the variables. Since we assume that ties occur with probability zero, we constructed the pseudo-observations by randomly breaking these ties. Not breaking the ties leads to either the rejection of a lot of well fitting copulas or abnormally high p-values as a result of the GOF test. The fitted copula models have Kendall τ correlation coefficients which are similar the Kendall τ correlation coefficients of the corresponding variables. The fitted copula models in the division *Education* are the only exception to this. There are notable differences between the Kendall τ correlation coefficients of some of the fitted copula models and the Kendall τ correlation coefficients of the corresponding variables. The division Education has the smallest amount of observations. So randomising the ties might have had a notable effect in this division, which leads to bigger differences between the values for Kendalls τ. However, not accounting for the ties created bigger differences between more variables. The last statement holds for all the divisions. There are three more notable differences in the Kendall τ values in different divisions. However, the copulas of these pairs are all rejected at the 5% significance level according to the GOF test.

Some dependence structures are easily captured via a copula throughout most of the divisions. The dependence structure between the variables *(mcs,mncs)* is a good example of this. The dependence structures between the variables *(p,pp top prop)* and *(p,pp int collab)* are a lot harder to capture with a parametric copula model. In fact, the GOF test rejects these models at a 5% significance level. When zooming in on the divisions, we see that Education and Non-Health Professional, the divisions with the lowest amount of observations, have only one copula rejected at the 5% significance level. A possible drawback from the selected copulas in these divisions is that due to limited size of data, the selected copula can be unreliable. Sciences and Social Sciences have the most copulas rejected at the 5% significance level, 6 and 7 copulas respectively. It seems that the parametric copulas can not capture these kinds of structures.

A big advantage of copula models is that they can capture the dependence structure between variables with low correlation coefficients. Furthermore, dependence structures which seem to be able to be modelled with a parametric copula, for example models with strong tails and a dense middle, can be rejected at a 5% significance level after fitting a parametric copula. The copula models try to capture the overall dependence structure between variables as well as the specific dependence structure between the data points. This lead to very reliable models, when working with data sets which contain a large amount of observations.

However, the best fitting copula model does not always perform the best in terms of prediction. The division with the highest amount of parametric copulas which were not rejected at a 5%

significance level, *Education*, depicts this perfectly. We applied leave-one-out, 5-fold and 10-fold cross-validation on the best fitting parametric copula models. The CIC values of most of these models turned out to be lower than that of other parametric copula models. The CIC values of part of the parametric copulas models which were not rejected at a 5% significance level throughout most the divisions turned out to be lower than that of other parametric copula models. The division *Basic Medical Sciences* was the only exception to this. The parametric copulas models which were not rejected at a 5% significance level in this divisions also had the highest CIC values. The amount of observations, height of the p-values derived under the GOF-test and correlation coefficients play no role in performance of the fitted copula during cross- validation. So our only conclusion is simply that fitted copula models do not always perform best during cross-validation.

To summarise, except for the division *Humanities*, parametric copulas are able to capture the dependence structure between the publications of a researchers and the citations of those publications. However, when we consider more bibliometric indicators, parametric copulas are not always able to capture the dependence structure between the various variables. Almost every division has 5 out of 10 parametric copula models which are not rejected by a 5% significance level. The divisions *Sciences* and *Social Sciences* are the only exceptions to this. These fitted parametric copula models do not always perform best during cross-validation.

References

[1] Harry Joe. (2014). *Dependence Modeling with Copulas*. Chapman & Hall/CRC Monographs on Statistics and Applied Probability. 1st edition.

[2] David Gold. (2017). *An introduction to Copulas*. Water programming: A collaborative Research Blog. Retrieved from `https://waterprogramming.wordpress.com/2017/11/11/an-introduction-to-copulas/`.

[3] Schweizer, B., & Wolff, E. F.. (1981). *On nonparametric measures of dependence for random variables*. The Annals of Statistics.

[4] Roger B. Nelson. (2006). *An introduction to copulas*. New York: Springer. 2nd edition.

[5] Paul Embrechts, Alexander McNeil, Daniel Straumann. (1999). *Correlation and dependance in risk management: properties and pitfalls*.

[6] Carlo Sempi. (2011) *An introduction to Copulas*. The 33rd Finnish Summer School on Probability Theory and Statistics. June 6th - 10th 2011.

[7] Erik Forslund, Daniel Johansson. (2012). *Gaussian Copula. What happens when models fail?*. Retrieved from `http://www.math.chalmers.se/Stat/Grundutb/CTH/mve220/1213/gr15_forslund_johansson_gaussian_cop.pdf`.

[8] Mike Thelwall, Paul Wilson. (n.d.) *Regression for Citation Data: An Evaluation of Different Methods*. Retrieved from `http://scitsc.wlv.ac.uk/~cm1993/papers/RegressionForCitationDataPreprint.pdf`.

[9] Felix Salmon. (2009). *Recipe for disaster: The formula that killed wall street.*. Retrieved from `https://www.wired.com/2009/02/wp-quant/`.

[10] Patrick Eschenburg. (2013). *Properties of extreme-value copulas*. Retrieved from `https://mediatum.ub.tum.de/doc/1145695/1145695.pdf`.

[11] Rodrigo Costas, Tina Nane, Vincent Larivière. (2015). *Is the Year of First Publication a Good Proxy of Scholars' Academic Age?*. Retrieved from `https://pdfs.semanticscholar.org/d2b8/1e6ff7a47799e0cd6f6c5baff3690885c739.pdf`

[12] Eike Brechmann, Ulf Schepsmeier. (2013). *Modeling Dependence with C- and D-Vine Copulas: The R Package CDVine*. Journal of Statistical Software. Volume 52. Issue 3.

[13] Farzia Habiboellah. (2007). *Copulas. Modeling dependencies in Financial Risk Management*. Master Thesis.

[14] Martin Haugh. (2016). *An introduction to Copulas*. Springer.

[15] Paul Embrechts. (2009). *Copulas: A personal view*. Department of Mathematics, ETH Zurich. Switzerland.

[16] Paul Embrechts, Filip Lindskog, Alexander McNeil. (2001). *Modelling Dependence with Copulas and Applications to Risk Mangement*. Department of Mathematics, ETH Zurich. Switzerland. Retrieved from `www.math.ethz.ch/finance`

[17] Michael Sherris, John van der Hoek. (2008). *Fitting and Estimating Risk Dependence using Copulas for Multivariate Data*. Research Conference, Amorta Hotel, Sydney. Institute of Actuaries of Australia.

[18] Alexander. (2017) *What is an Empirical Copula?*. Retrieved from `http://www.deep-mind.org/2017/09/24/empirical_copula/`.

[19] Gregor Weiss. (2009). *Copula Parameter Estimation by Maximum- Likelihood and Minimum Distance Estimators ? A Simulation Study*. Retrieved from `https://www.minet.uni-jena.de/Marie-Curie-ITN/Workshop/talks/WS_Weiß.pdf`.

[20] Ivan Kojadinovic, Jun Yan. (2010). *Modeling Multivariate Distributions with Continuous Margins Using the copula R Package*. Journal of Statistical Software. Volume 34. Issue 9.

[21] Kumar Joag-Dev. (1984). *Handbook of Statistics: 4 Measures of dependence*. Retrieved from `https://www.sciencedirect.com/science/article/pii/S0169716184040062`.

[22] Lars Arne Jordanger, Dag Tjostheim. (2014). *Model selection of copulas: AIC versus a cross validation copula information criterion*. Elsevier.

[23] John Fox. (2016). *Applied regression analysis & generalized linear models*. Los Angeles: Sage. 3rd edition.

[24] Geoffrey Grimmett, Dominic Welsh. (2014). *Probability, an introduction.* Oxford: Oxford university press. 2nd edition.

[25] John A. Rice. (2007). *Mathematical Statistics and Data Analysis.* Canada: Brooks/Cole, Cengage Learning. 3rd edition.

Appendices

A Main descriptive values

Division		P	MCS	MNCS	pp_top_prop	pp_int_collab
Basic Medical Sciences	N	711	711	711	711	711
	Mean	40.879	31.295	1.507	0.156	0.313
	Std. Deviation	44.599	33.405	1.428	0.139	0.218
	Minimum	1	0	0	0	0
	Maximum	392	562.2	24.692	1	1
		P	MCS	MNCS	pp_top_prop	pp_int_collab
Business & Management	N	238	238	238	238	238
	Mean	9.088	17.259	1.496	0.158	0.378
	Std. Deviation	8.973	20.621	1.680	0.222	0.342
	Minimum	1	0	0	0	0
	Maximum	85	116.667	12.009	1	1
		P	MCS	MNCS	pp_top_prop	pp_int_collab
Education	N	47	47	47	47	47
	Mean	7.596	9.018	0.979	0.083	0.214
	Std. Deviation	10.623	8.112	0.874	0.152	0.323
	Minimum	1	0	0	0	0
	Maximum	44	32.6	3.775	0.6	1
		P	MCS	MNCS	pp_top_prop	pp_int_collab
Engineering	N	512	512	512	512	512
	Mean	31.799	10.055	1.121	0.109	0.263
	Std. Deviation	33.134	19.319	1.236	0.109	0.219
	Minimum	1	0	0	0	0
	Maximum	351	402.857	22.853	0.571	1
		P	MCS	MNCS	pp_top_prop	pp_int_collab
Health Sciences	N	288	288	288	288	288
	Mean	36.747	20.561	1.295	0.143	0.263
	Std. Deviation	42.622	18.960	0.849	0.149	0.229
	Minimum	1	0.833	0.075	0	0
	Maximum	392	134.432	5.324	1	1
		P	MCS	MNCS	pp_top_prop	pp_int_collab
Humanities	N	342	342	342	342	342
	Mean	3.906	2.730	1.381	0.133	0.086
	Std. Deviation	5.459	5.453	2.255	0.242	0.229
	Minimum	1	0	0	0	0
	Maximum	65	54	24.015	1	1
		P	MCS	MNCS	pp_top_prop	pp_int_collab
Non-Health Professional	N	108	108	108	108	108
	Mean	10.278	6.192	0.923	0.070	0.177
	Std. Deviation	14.163	6.492	1.251	0.131	0.248
	Minimum	1	0	0	0	0
	Maximum	70	36.667	10.987	1	1
		P	MCS	MNCS	pp_top_prop	pp_int_collab
Sciences	N	824	824	824	824	824
	Mean	34.049	19.946	1.498	0.143	0.384
	Std. Deviation	44.019	58.775	2.282	0.142	0.262
	Minimum	1	0	0	0	0
	Maximum	777	1550.5	47.333	1	1
		P	MCS	MNCS	pp_top_prop	pp_int_collab
Social Sciences	N	500	500	500	500	500
	Mean	14.794	13.433	1.194	0.121	0.312
	Std. Deviation	15.083	13.069	1.124	0.161	0.287
	Minimum	1	0	0	0	0
	Maximum	83	88.25	11.302	1	1

Table 30: Statistics per division of the data set. N corresponds with the amount of observations. The table also shows the mean, standard deviation, minimum and maximum of each division per variable.

B Copula selection: average ties

Variable 1	Variable 2	Copula	Parameter 1	Parameter 2 (or df)	Kendalls Tau	AIC	Satistic	p-value
p	mcs	Survival Tawn T1	1.74	0.4	0.23	-127.26	0.017	0.644
p	mncs	Survival Tawn T1	1.9	0.39	0.24	-156.94	0.026	0.1833
p	pp_top_prop	Survival Tawn T1	2.03	0.42	0.27	-171.89	0.186	0.0358
p	pp_int_collab	Survival Joe	1.58	-	0.24	-155.82	0.123	0.0007
mcs	mncs	Student t	0.86	4.25	0.66	-989.67	0.018	0.2409
mcs	pp_top_prop	Student t	0.76	9.61	0.55	-576.42	0.044	0.3715
mcs	pp_int_collab	Student t	0.34	3.95	0.22	-99.54	0.046	0.0625
mncs	pp_top_prop	Frank	12.31	-	0.72	-1080.3	0.0262	0.4993
mncs	pp_int_collab	Survival Gumbel	1.35	-	0.26	-121.67	0.036	0.1552
pp_top_ prop	pp_int_collab	Student t	0.38	2.35	0.25	-147.67	0.068	0.2704

Table 31: Output Copula selection via R for the division Basic Medical Sciences.

Variable 1	Variable 2	Copula	Parameter 1	Parameter 2 (or df)	Kendalls Tau	AIC	Satistic	p-value
p	mcs	Survival Tawn T1	1.77	0.25	0.16	-21.25	0.049	0.0607
p	mncs	Survival Tawn T1	1.87	0.27	0.18	-29.43	0.0303	0.3285
p	pp_top_prop	Survival Joe	1.66	-	0.27	-31.76	0.09	0.7092
p	pp_int_collab	Survival Joe	1.37	-	0.17	-15.26	0.152	0.1402
mcs	mncs	Gaussian	0.93	-	0.75	-454.91	0.008	0.818
mcs	pp_top_prop	BB8	5.24	0.85	0.59	-234.2	0.024	0.931
mcs	pp_int_collab	Frank	3.23	-	0.33	-53.59	0.02	0.8682
mncs	pp_top_prop	BB8	6	0.95	0.69	-391.04	0.036	0.8347
mncs	pp_int_collab	Student t	0.43	3.41	0.28	-40.67	0.025	0.8891
pp_top_ prop	pp_int_collab	BB7	1.9	0.08	0.35	-50.78	0.017	0.9435

Table 32: Output Copula selection via R for the division Business & Management.

Variable 1	Variable 2	Copula	Parameter 1	Parameter 2 (or df)	Kendalls Tau	AIC	Satistic	p-value
p	mcs	Tawn T2	2.76	0.51	0.38	-17.01	0.014	0.9688
p	mncs	Survival Tawn T1	5.27	0.25	0.23	-13.72	0.015	0.8854
p	pp_top_prop	Clayton	1.71	-	0.46	-17.1	0.054	0.8854
p	pp_int_collab	Survival Joe	2.13	-	0.38	-10.18	0.0312	0.9688
mcs	mncs	Survival Tawn T1	4.78	0.86	0.7	-77.74	0.014	0.6354
mcs	pp_top_prop	Gumbel	1.88	-	0.47	-19.49	0.035	0.8438
mcs	pp_int_collab	Student t	0.3	2.04	0.29	-1.79	0.034	0.8021
mncs	pp_top_prop	Tawn T1	3.3	0.74	0.55	-38.75	0.025	0.8438
mncs	pp_int_collab	Student t	0.26	2	0.17	-1.98	0.027	0.8646
pp_top_ prop	pp_int_collab	Survival Tawn T1	5.01	0.68	0.57	-41.84	0.017	0.9896

Table 33: Output Copula selection via R for the division Education.

Variable 1	Variable 2	Copula	Parameter 1	Parameter 2 (or df)	Kendalls Tau	AIC	Satistic	p-value
p	mcs	Rotated Tawn type 1 180 degrees	1.8	0.39	0.23	-98.12	0.042	0.03216
p	mncs	Rotated Tawn type 1 180 degrees	1.82	0.34	0.21	-91.84	0.022	0.3499
p	pp_top_prop	Survival Joe	1.6	-	0.25	-98.24	0.122	0.2154
p	pp_int_collab	Rotated Tawn type 1 180 degrees	1.87	0.38	0.24	-83.33	0.088	0.0166
mcs	mncs	BB1	0.99	1.78	0.52	-630.78	0.007	0.9717
mcs	pp_top_prop	Student t	0.75	14.33	0.16	-397.01	0.021	0.8021
mcs	pp_int_collab	Survival Joe	1.24	-	0.12	-19.78	0.0214	0.7261
mncs	pp_top_prop	Frank	13	-	0.73	-811.83	0.04	0.4883
mncs	pp_int_collab	Student t	0.24	4.26	0.16	-37.33	0.022	0.6248
pp_top_ prop	pp_int_collab	Survival Joe	1.45	-	0.2	-50.56	0.049	0.6248

Table 34: Output Copula selection via R for the division Engineering.

Variable 1	Variable 2	Copula	Parameter 1	Parameter 2 (or df)	Kendalls Tau	AIC	Satistic	p-value
p	mcs	Survival Tawn T1	1.89	0.5	0.29	-79.06	0.012	0.8945
p	mncs	Survival Tawn T1	2.17	0.38	0.26	-80.09	0.01	0.974
p	pp_top_prop	Survival Tawn T1	2.08	0.46	0.3	-72.17	0.077	0.1505
p	pp_int_collab	Survival Tawn T1	2.05	0.45	0.29	-71.13	0.064	0.1678
mcs	mncs	Tawn T2	3.35	0.78	0.58	-335.88	0.0168	0.3097
mcs	pp_top_prop	Student t	0.72	3.63	0.51	-198.19	0.017	0.8979
mcs	pp_int_collab	Survival BB1	0.001	1.27	0.21	-25.71	0.0216	0.7318
mncs	pp_top_prop	Frank	13.11	-	0.73	-460.52	0.022	0.6142
mncs	pp_int_collab	Survival Gumbel	1.23	-	0.19	-22.23	0.018	0.8529
pp_top_ prop	pp_int_collab	Student t	0.29	2	0.19	-35.7	0.322	0.9221

Table 35: Output Copula selection via R for the division Health Sciences.

Variable 1	Variable 2	Copula	Parameter 1	Parameter 2 (or df)	Kendalls Tau	AIC	Satistic	p-value
p	mcs	Survival BB7	1.32	0.14	0.2	-20.56	0.0626	0.844
p	mncs	Survival Tawn T1	2	0.25	0.18	-25.17	0.1761	0.3542
p	pp_top_prop	Survival Joe	1.79	-	0.3	-45.99	0.09183	0.949
p	pp_int_collab	Clayton	0.91	-	0.31	-30.95	0.065574	0.9956
mcs	mncs	BB7	1.42	5.11	0.71	-541.53	0.148	0.2551
mcs	pp_top_prop	Survival Tawn T2	5.9	0.33	0.31	-171.63	0.224	0.7303
mcs	pp_int_collab	Survival Tawn T2	9.81	0.21	0.2	-118.43	0.179	0.8761
mncs	pp_top_prop	Survival BB7	1.55	2.21	0.56	-322.59	0.053	0.984
mncs	pp_int_collab	Survival Tawn T2	8.85	0.21	0.2	-107.83	0.048	0.981
pp_top_ prop	pp_int_collab	Survival Joe	2.26	-	0.41	-60.51	0.0255	0.9373

Table 36: Output Copula selection via R for the division Humanities.

Variable 1	Variable 2	Copula	Parameter 1	Parameter 2 (or df)	Kendalls Tau	AIC	Satistic	p-value
p	mcs	BB1	0.92	1	0.31	-21.43	0.039	0.289
p	mncs	Survival Tawn T1	2.35	0.43	0.31	-31.14	0.0295	0.4266
p	pp_top_prop	Clayton	1.49	-	0.43	-38.54	0.0647	0.8486
p	pp_int_collab	Survival Joe	2.2	-	0.4	-34.58	0.0833	0.7477
mcs	mncs	Survival Joe	4.44	-	0.64	-144.98	0.018	0.5917
mcs	pp_top_prop	Frank	4.27	-	0.41	-33.86	0.023	0.922
mcs	pp_int_collab	Clayton	0.84	-	0.3	-15.47	0.018	0.9954
mncs	pp_top_prop	Gumbel	2.49	-	0.6	-102.16	0.033	0.922
mncs	pp_int_collab	Clayton	0.8	-	0.29	-14.15	0.199	0.922
pp_top_ prop	pp_int_collab	Survival Joe	2.8	-	0.49	-43.91	0.039	0.9587

Table 37: Output Copula selection via R for the division Non-Health Professional.

Variable 1	Variable 2	Copula	Parameter 1	Parameter 2 (or df)	Kendalls Tau	AIC	Satistic	p-value
p	mcs	Survival Tawn T1	1.7	0.48	0.25	-175.52	0.052	0.005455
p	mncs	Survival Tawn T1	1.69	0.33	0.19	-118.56	0.035	0.06485
p	pp_top_prop	Survival Joe	1.52	-	0.23	-129.61	0.16083	0.1303
p	pp_int_collab	Student t	0.01	2.82	0.01	-57.69	0.14397	0.0006061
mcs	mncs	BB7	2.34	1.75	0.59	-960.44	0.036147	0.01273
mcs	pp_top_prop	Student t	0.71	12.22	0.5	-540.22	0.047	0.5024
mcs	pp_int_collab	Survival Tawn T1	1.34	0.17	0.08	-11.46	0.0598	0.001818
mncs	pp_top_prop	Frank	14.28	-	0.75	-1436.55	0.0223	0.6685
mncs	pp_int_collab	Student t	0.21	4.6	0.13	-52.81	0.0315	0.1109
pp_top_ prop	pp_int_collab	Student t	0.2	3.18	0.13	-63.95	0.023	0.9388

Table 38: Output Copula selection via R for the division Sciences.

Variable 1	Variable 2	Copula	Parameter 1	Parameter 2 (or df)	Kendalls Tau	AIC	Satistic	p-value
p	mcs	Survival Tawn T1	1.93	0.56	0.32	-167.84	0.024	0.3004
p	mncs	Survival Tawn T1	1.89	0.38	0.24	-100.34	0.023	0.4162
p	pp_top_prop	Clayton	0.93	-	0.32	-110.29	0.1525	0.6198
p	pp_int_collab	Survival Joe	1.71	-	0.28	-113.51	0.19263	0.2665
mcs	mncs	Survival Tawn T1	3.54	0.79	0.6	-630.52	0.016	0.3822
mcs	pp_top_prop	Frank	6.06	-	0.52	-325.88	0.2064	0.9132
mcs	pp_int_collab	Clayton	0.7	-	0.26	-84.68	0.041	0.7136
mncs	pp_top_prop	Tawn type 1	3.4	0.89	0.65	-671.49	0.052	0.8373
mncs	pp_int_collab	Survival Tawn T2	1.6	0.46	0.22	-78.18	0.052	0.6158
pp_top_ prop	pp_int_collab	Survival Tawn T1	2.05	0.52	0.32	-108.65	0.162	0.9491

Table 39: Output Copula Selection via R for the division Social Sciences.

C R-code: Correlation calculation

```
rm(list = ls()) #clear workspace
setwd("/Users/ashnibachasingh/Documents/BachelorColloqium/BEP") #set workdictory
library(readxl) #needed to read excelfile without having to transfer into txt file
Canadian_researchers_data <- read_excel("~/Documents/BachelorColloqium/BEP/
                                        Orginal Data/Canadian_researchers_data.xlsx")
data.per.division<-Canadian_researchers_data

#turns numeric data registered as non numeric into numeric data
data.per.division$authors_paper<-as.numeric(data.per.division$authors_paper)
data.per.division$institutes_paper<-as.numeric(data.per.division$institutes_paper)
data.per.division$countries_paper<-as.numeric(data.per.division$countries_paper)
data.per.division$pages_paper<-as.numeric(data.per.division$pages_paper)
data.per.division$refs_paper<-as.numeric(data.per.division$refs_paper)
data.per.division$p<-as.numeric(data.per.division$p)
data.per.division$mcs<-as.numeric(data.per.division$mcs)
data.per.division$mncs<-as.numeric (data.per.division$mncs)
data.per.division$pp_top_prop<-as.numeric(data.per.division$pp_top_prop)
data.per.division$mjs_mcs<-as.numeric(data.per.division$mjs_mcs)
data.per.division$mnjs_mncs<-as.numeric(data.per.division$mnjs_mncs)
data.per.division$mnjs_pp_top_prop<-as.numeric(data.per.division$mnjs_pp_top_prop)
data.per.division$pp_collab<-as.numeric(data.per.division$pp_collab)
data.per.division$pp_int_collab<-as.numeric(data.per.division$pp_int_collab)
data.per.division<-na.omit(data.per.division)

dataset<-readline(prompt="Which data/division would you like to use? ")
data.per.division<-subset(data.per.division,division==dataset)

#correlation per quartile (and tails)#switch method between kendall and spearman
#empty dataframes
Q1.p<-data.per.division[0,]
Q2.p<-data.per.division[0,]
Q3.p<-data.per.division[0,]
Q4.p<-data.per.division[0,]
Q90.p<-data.per.division[0,]
Q95.p<-data.per.division[0,]

Q1.mcs<-data.per.division[0,]
Q2.mcs<-data.per.division[0,]
Q3.mcs<-data.per.division[0,]
Q4.mcs<-data.per.division[0,]
Q90.mcs<-data.per.division[0,]
Q95.mcs<-data.per.division[0,]

Q1.mncs<-data.per.division[0,]
Q2.mncs<-data.per.division[0,]
Q3.mncs<-data.per.division[0,]
Q4.mncs<-data.per.division[0,]
Q90.mncs<-data.per.division[0,]
Q95.mncs<-data.per.division[0,]

Q1.collab<-data.per.division[0,]
Q2.collab<-data.per.division[0,]
Q3.collab<-data.per.division[0,]
Q4.collab<-data.per.division[0,]
Q90.collab<-data.per.division[0,]
Q95.collab<-data.per.division[0,]

Q1.top<-data.per.division[0,]
```

```
Q2.top<-data.per.division[0,]
Q3.top<-data.per.division[0,]
Q4.top<-data.per.division[0,]
Q90.top<-data.per.division[0,]
Q95.top<-data.per.division[0,]

sorted.p<-data.per.division[order(data.per.division$p),]
sorted.mcs<-data.per.division[order(data.per.division$mcs),]
sorted.mncs<-data.per.division[order(data.per.division$mncs),]
sorted.collab<-data.per.division[order(data.per.division$pp_int_collab),]
sorted.top<-data.per.division[order(data.per.division$pp_top_prop),]

elements<-nrow(data.per.division)
for(i in 1:elements){
  if(i<=round(0.25*elements)){
    Q1.p<-rbind(Q1.p,sorted.p[i,])
    Q1.mcs<-rbind(Q1.mcs,sorted.mcs[i,])
    Q1.mncs<-rbind(Q1.mncs,sorted.mncs[i,])
    Q1.collab<-rbind(Q1.collab,sorted.collab[i,])
    Q1.top<-rbind(Q1.top,sorted.top[i,])
  }
  else if(i>round(0.25*elements) & i<=round(0.5*elements)){
    Q2.p<-rbind(Q2.p,sorted.p[i,])
    Q2.mcs<-rbind(Q2.mcs,sorted.mcs[i,])
    Q2.mncs<-rbind(Q2.mncs,sorted.mncs[i,])
    Q2.collab<-rbind(Q2.collab,sorted.collab[i,])
    Q2.top<-rbind(Q2.top,sorted.top[i,])
  }
  else if(i>round(0.5*elements) & i<=round(0.75*elements)){
    Q3.p<-rbind(Q3.p,sorted.p[i,])
    Q3.mcs<-rbind(Q3.mcs,sorted.mcs[i,])
    Q3.mncs<-rbind(Q3.mncs,sorted.mncs[i,])
    Q3.collab<-rbind(Q3.collab,sorted.collab[i,])
    Q3.top<-rbind(Q3.top,sorted.top[i,])
  }
  else if(i>round(0.75*elements)){
    Q4.p<-rbind(Q4.p,sorted.p[i,])
    Q4.mcs<-rbind(Q4.mcs,sorted.mcs[i,])
    Q4.mncs<-rbind(Q4.mncs,sorted.mncs[i,])
    Q4.collab<-rbind(Q4.collab,sorted.collab[i,])
    Q4.top<-rbind(Q4.top,sorted.top[i,])
  }
}

#correlation values 90th percentile
for(i in 1:elements){
  if(i>=round(0.90*elements)){
    Q90.p<-rbind(Q90.p,sorted.p[i,])
    Q90.mcs<-rbind(Q90.mcs,sorted.mcs[i,])
    Q90.mncs<-rbind(Q90.mncs,sorted.mncs[i,])
    Q90.collab<-rbind(Q90.collab,sorted.collab[i,])
    Q90.top<-rbind(Q90.top,sorted.top[i,])
  }
}

#correlation values 95th percentile
for(i in 1:elements){
  if(i>=round(0.95*elements)){
    Q95.p<-rbind(Q95.p,sorted.p[i,])
    Q95.mcs<-rbind(Q95.mcs,sorted.mcs[i,])
```

```
    Q95.mncs<-rbind(Q95.mncs,sorted.mncs[i,])
    Q95.collab<-rbind(Q95.collab,sorted.collab[i,])
    Q95.top<-rbind(Q95.top,sorted.top[i,])
  }
}

#correlation values complete range
cor(data.per.division$p,data.per.division$mcs,method="kendall")
cor(data.per.division$p,data.per.division$mncs,method="kendall")
cor(data.per.division$p,data.per.division$pp_top_prop,method="kendall")
cor(data.per.division$p,data.per.division$pp_int_collab,method="kendall")
cor(data.per.division$mcs,data.per.division$mncs,method="kendall")
cor(data.per.division$mcs,data.per.division$pp_top_prop,method="kendall")
cor(data.per.division$mcs,data.per.division$pp_int_collab,method="kendall")
cor(data.per.division$mncs,data.per.division$pp_top_prop,method="kendall")
cor(data.per.division$mncs,data.per.division$pp_int_collab,method="kendall")
cor(data.per.division$pp_top_prop,data.per.division$pp_int_collab,method="kendall")

#correlation values per quartile
#correlation values Q1
cor(Q1.p$p,Q1.p$mcs,use = "complete.obs",method="kendall")
cor(Q1.p$p,Q1.p$mncs,use = "complete.obs",method="kendall")
cor(Q1.p$p,Q1.p$pp_top_prop,use = "complete.obs",method="kendall")
cor(Q1.p$p,Q1.p$pp_int_collab,use = "complete.obs",method="kendall")
cor(Q1.mcs$mcs,Q1.mcs$mncs,use = "complete.obs",method="kendall")
cor(Q1.mcs$mcs,Q1.mcs$pp_top_prop,use = "complete.obs",method="kendall")
cor(Q1.mcs$mcs,Q1.mcs$pp_int_collab,use = "complete.obs",method="kendall")
cor(Q1.mncs$mncs,Q1.mncs$pp_top_prop,use = "complete.obs",method="kendall")
cor(Q1.mncs$mncs,Q1.mncs$pp_int_collab,use = "complete.obs",method="kendall")
cor(Q1.top$pp_top_prop,Q1.top$pp_int_collab,use = "complete.obs",method="kendall")
#correlation values Q1–Q2
cor(Q2.p$p,Q2.p$mcs,use = "complete.obs",method="kendall")
cor(Q2.p$p,Q2.p$mncs,use = "complete.obs",method="kendall")
cor(Q2.p$p,Q2.p$pp_top_prop,use = "complete.obs",method="kendall")
cor(Q2.p$p,Q2.p$pp_int_collab,use = "complete.obs",method="kendall")
cor(Q2.mcs$mcs,Q2.mcs$mncs,use = "complete.obs",method="kendall")
cor(Q2.mcs$mcs,Q2.mcs$pp_top_prop,use = "complete.obs",method="kendall")
cor(Q2.mcs$mcs,Q2.mcs$pp_int_collab,use = "complete.obs",method="kendall")
cor(Q2.mncs$mncs,Q2.mncs$pp_top_prop,use = "complete.obs",method="kendall")
cor(Q2.mncs$mncs,Q2.mncs$pp_int_collab,use = "complete.obs",method="kendall")
cor(Q2.top$pp_top_prop,Q2.top$pp_int_collab,use = "complete.obs",method="kendall")
#correlation values Q2–Q3
cor(Q3.p$p,Q3.p$mcs,use = "complete.obs",method="kendall")
cor(Q3.p$p,Q3.p$mncs,use = "complete.obs",method="kendall")
cor(Q3.p$p,Q3.p$pp_top_prop,use = "complete.obs",method="kendall")
cor(Q3.p$p,Q3.p$pp_int_collab,use = "complete.obs",method="kendall")
cor(Q3.mcs$mcs,Q3.mcs$mncs,use = "complete.obs",method="kendall")
cor(Q3.mcs$mcs,Q3.mcs$pp_top_prop,use = "complete.obs",method="kendall")
cor(Q3.mcs$mcs,Q3.mcs$pp_int_collab,use = "complete.obs",method="kendall")
cor(Q3.mncs$mncs,Q3.mncs$pp_top_prop,use = "complete.obs",method="kendall")
cor(Q3.mncs$mncs,Q3.mncs$pp_int_collab,use = "complete.obs",method="kendall")
cor(Q3.top$pp_top_prop,Q3.top$pp_int_collab,use = "complete.obs",method="kendall")
#correlation values above Q3
cor(Q4.p$p,Q4.p$mcs,use = "complete.obs",method="kendall")
cor(Q4.p$p,Q4.p$mncs,use = "complete.obs",method="kendall")
cor(Q4.p$p,Q4.p$pp_top_prop,use = "complete.obs",method="kendall")
cor(Q4.p$p,Q4.p$pp_int_collab,use = "complete.obs",method="kendall")
cor(Q4.mcs$mcs,Q4.mcs$mncs,use = "complete.obs",method="kendall")
cor(Q4.mcs$mcs,Q4.mcs$pp_top_prop,use = "complete.obs",method="kendall")
cor(Q4.mcs$mcs,Q4.mcs$pp_int_collab,use = "complete.obs",method="kendall")
```

```
cor(Q4.mncs$mncs,Q4.mncs$pp_top_prop,use = "complete.obs",method="kendall")
cor(Q4.mncs$mncs,Q4.mncs$pp_int_collab,use = "complete.obs",method="kendall")
cor(Q4.top$pp_top_prop,Q4.top$pp_int_collab,use = "complete.obs",method="kendall")
#correlation values in the 90th percentile
cor(Q90.p$p,Q90.p$mcs,use = "complete.obs",method="kendall")
cor(Q90.p$p,Q90.p$mncs,use = "complete.obs",method="kendall")
cor(Q90.p$p,Q90.p$pp_top_prop,use = "complete.obs",method="kendall")
cor(Q90.p$p,Q90.p$pp_int_collab,use = "complete.obs",method="kendall")
cor(Q90.mcs$mcs,Q90.mcs$mncs,use = "complete.obs",method="kendall")
cor(Q90.mcs$mcs,Q90.mcs$pp_top_prop,use = "complete.obs",method="kendall")
cor(Q90.mcs$mcs,Q90.mcs$pp_int_collab,use = "complete.obs",method="kendall")
cor(Q90.mncs$mncs,Q90.mncs$pp_top_prop,use = "complete.obs",method="kendall")
cor(Q90.mncs$mncs,Q90.mncs$pp_int_collab,use = "complete.obs",method="kendall")
cor(Q90.top$pp_top_prop,Q90.top$pp_int_collab,use = "complete.obs",method="kendall")
#correlation values in the 95th percentile
cor(Q95.p$p,Q95.p$mcs,use = "complete.obs",method="kendall")
cor(Q95.p$p,Q95.p$mncs,use = "complete.obs",method="kendall")
cor(Q95.p$p,Q95.p$pp_top_prop,use = "complete.obs",method="kendall")
cor(Q95.p$p,Q95.p$pp_int_collab,use = "complete.obs",method="kendall")
cor(Q95.mcs$mcs,Q95.mcs$mncs,use = "complete.obs",method="kendall")
cor(Q95.mcs$mcs,Q95.mcs$pp_top_prop,use = "complete.obs",method="kendall")
cor(Q95.mcs$mcs,Q95.mcs$pp_int_collab,use = "complete.obs",method="kendall")
cor(Q95.mncs$mncs,Q95.mncs$pp_top_prop,use = "complete.obs",method="kendall")
cor(Q95.mncs$mncs,Q95.mncs$pp_int_collab,use = "complete.obs",method="kendall")
cor(Q95.top$pp_top_prop,Q95.top$pp_int_collab,use = "complete.obs",method="kendall")

uniformp<-pobs(data.per.division$p)
uniformmcs<-pobs(data.per.division$mcs)
uniformmncs<-pobs(data.per.division$mncs)
uniformtop<-pobs(data.per.division$pp_top_prop)
uniformcollab<-pobs(data.per.division$pp_int_collab)

#tail dependence coefficient lambda
#cut-off parameter p=0.05
p=0.05
fitLambda(as.matrix(data.frame(uniformp,uniformmcs)),p=p,lower.tail=TRUE)
fitLambda(as.matrix(data.frame(uniformp,uniformmcs)),p=p,lower.tail=FALSE)
fitLambda(as.matrix(data.frame(uniformp,uniformmncs)),p=p,lower.tail=TRUE)
fitLambda(as.matrix(data.frame(uniformp,uniformmncs)),p=p,lower.tail=FALSE)
fitLambda(as.matrix(data.frame(uniformp,uniformtop)),p=p,lower.tail=TRUE)
fitLambda(as.matrix(data.frame(uniformp,uniformtop)),p=p,lower.tail=FALSE)
fitLambda(as.matrix(data.frame(uniformp,uniformcollab)),p=p,lower.tail=TRUE)
fitLambda(as.matrix(data.frame(uniformp,uniformcollab)),p=p,lower.tail=FALSE)
fitLambda(as.matrix(data.frame(uniformmcs,uniformmncs)),p=p,lower.tail=TRUE)
fitLambda(as.matrix(data.frame(uniformmcs,uniformmncs)),p=p,lower.tail=FALSE)
fitLambda(as.matrix(data.frame(uniformmcs,uniformtop)),p=p,lower.tail=TRUE)
fitLambda(as.matrix(data.frame(uniformmcs,uniformtop)),p=p,lower.tail=FALSE)
fitLambda(as.matrix(data.frame(uniformmcs,uniformcollab)),p=p,lower.tail=TRUE)
fitLambda(as.matrix(data.frame(uniformmcs,uniformcollab)),p=p,lower.tail=FALSE)
fitLambda(as.matrix(data.frame(uniformmncs,uniformtop)),p=p,lower.tail=TRUE)
fitLambda(as.matrix(data.frame(uniformmncs,uniformtop)),p=p,lower.tail=FALSE)
fitLambda(as.matrix(data.frame(uniformmncs,uniformcollab)),p=p,lower.tail=TRUE)
fitLambda(as.matrix(data.frame(uniformmncs,uniformcollab)),p=p,lower.tail=FALSE)
fitLambda(as.matrix(data.frame(uniformtop,uniformcollab)),p=p,lower.tail=TRUE)
fitLambda(as.matrix(data.frame(uniformtop,uniformcollab)),p=p,lower.tail=FALSE)

#useful plots to help understand the correlation values and the bins
dev.off()
par(mfrow=c(2,3))
hist(data.per.division$p,100,main="Histogram Publications",xlab="Publications")
```

```
abline(v=c((max(Q1.p$p)),(max(Q2.p$p)),(max(Q3.p$p)),(min(Q90.p$p)),(min(Q95.p$p))),
       col=c("red","red","red","blue","green"),lty=c(2,3,4,1,1))
hist(data.per.division$mcs,100,main="Histogram Citations",xlab="MCS")
abline(v=c((max(Q1.mcs$mcs)),(max(Q2.mcs$mcs)),(max(Q3.mcs$mcs)),(min(Q90.mcs$mcs)),
           (min(Q95.mcs$mcs))),col=c("red","red","red","blue","green"),lty=c(2,3,4,1,1))
hist(data.per.division$mncs,100,main="Histogram Normalised Citation Values",xlab="MNCS")
abline(v=c((max(Q1.mncs$mncs)),(max(Q2.mncs$mncs)),(max(Q3.mncs$mncs)),(min(Q90.mncs$mncs)),
           (min(Q95.mncs$mncs))),col=c("red","red","red","blue","green"),lty=c(2,3,4,1,1))
hist(data.per.division$pp_top_prop,100,main="Histogram Percentage of publications
     \n which are in the top 10% most \n cited papers in their field",xlab="pp_top_prop")
abline(v=c((max(Q1.top$pp_top_prop)),(max(Q2.top$pp_top_prop)),(max(Q3.top$pp_top_prop)),
           (min(Q90.top$pp_top_prop)),(min(Q95.top$pp_top_prop))),
       col=c("red","red","red","blue","green"),lty=c(2,3,4,1,1))
hist(data.per.division$pp_int_collab,100,main="Histogram International Collaboration",
     xlab="International Collaboration")
abline(v=c((max(Q1.collab$pp_int_collab)),(max(Q2.collab$pp_int_collab)),
           (max(Q3.collab$pp_int_collab)),
           (min(Q90.collab$pp_int_collab)),(min(Q95.collab$pp_int_collab))),
       col=c("red","red","red","blue","green"),lty=c(2,3,4,1,1))
dev.off()
par(mfrow=c(3,4))
plot(data.per.division$p,data.per.division$mcs,main="Publications against MCS",
     xlab="Publications",ylab="MCS")
abline(v=c((max(Q1.p$p)),(max(Q2.p$p)),(min(Q3.p$p)),(min(Q90.p$p)),(min(Q95.p$p))),
       col=c("red","red","red","blue","green"),lty=c(2,3,4,1,1))
plot(data.per.division$p,data.per.division$mncs,main="Publications against MNCS",
     xlab="Publications",ylab="MNCS")
abline(v=c((max(Q1.p$p)),(max(Q2.p$p)),(max(Q3.p$p)),(min(Q90.p$p)),(min(Q95.p$p))),
       col=c("red","red","red","blue","green"),lty=c(2,3,4,1,1))
plot(data.per.division$p,data.per.division$pp_top_prop,
     main="Publications against pp_top_prop",
     xlab="Publications",ylab="pp_top_prop")
abline(v=c((max(Q1.p$p)),(max(Q2.p$p)),(max(Q3.p$p)),(min(Q90.p$p)),(min(Q95.p$p))),
       col=c("red","red","red","blue","green"),lty=c(2,3,4,1,1))
plot(data.per.division$p,data.per.division$pp_int_collab,
     main="Publications against \n International Collaboration",
     xlab="Publications",ylab="International \n Collaboration")
abline(v=c((max(Q1.p$p)),(max(Q2.p$p)),(max(Q3.p$p)),(min(Q90.p$p)),
           (min(Q95.p$p))),
       col=c("red","red","red","blue","green"),lty=c(2,3,4,1,1))
plot(data.per.division$mcs,data.per.division$mncs,main="MCS against MNCS",
     xlab="MCS",ylab="MNCS")
abline(v=c((max(Q1.mcs$mcs)),(max(Q2.mcs$mcs)),(max(Q3.mcs$mcs)),
           (min(Q90.mcs$mcs)),(min(Q95.mcs$mcs))),
       col=c("red","red","red","blue","green"),lty=c(2,3,4,1,1))
plot(data.per.division$mcs,data.per.division$pp_top_prop,
     main="MCS against pp_top_prop",
     xlab="MCS",ylab="pp_top_prop")
abline(v=c((max(Q1.mcs$mcs)),(max(Q2.mcs$mcs)),(max(Q3.mcs$mcs)),
           (min(Q90.mcs$mcs)),(min(Q95.mcs$mcs))),
       col=c("red","red","red","blue","green"),lty=c(2,3,4,1,1))
plot(data.per.division$mcs,data.per.division$pp_int_collab,
     main="MCS against \n International Collaboration",
     xlab="MCS",ylab="International \n Collaboration")
abline(v=c((max(Q1.mcs$mcs)),(max(Q2.mcs$mcs)),(max(Q3.mcs$mcs)),
           (min(Q90.mcs$mcs)),(min(Q95.mcs$mcs))),
       col=c("red","red","red","blue","green"),lty=c(2,3,4,1,1))
plot(data.per.division$mncs,data.per.division$pp_top_prop,
     main="MNCS against pp_top_prop",
     xlab="MNCS",ylab="pp_top_prop")
```

```
abline(v=c((max(Q1.mncs$mncs)),(max(Q2.mncs$mncs)),(max(Q3.mncs$mncs)),
            (min(Q90.mncs$mncs)),(min(Q95.mncs$mncs))),
       col=c("red","red","red","blue","green"),lty=c(2,3,4,1,1))
plot(data.per.division$mncs,data.per.division$pp_int_collab,
     main="MNCS against \n International Collaboration",
     xlab="MNCS",ylab="International \n Collaboration")
abline(v=c((max(Q1.mncs$mncs)),(max(Q2.mncs$mncs)),(max(Q3.mncs$mncs)),
            (min(Q90.mncs$mncs)),(min(Q95.mncs$mncs))),
       col=c("red","red","red","blue","green"),lty=c(2,3,4,1,1))
plot(data.per.division$pp_top_prop,data.per.division$pp_int_collab,
     main="pp_top_prop against \n International Collaboration",
     xlab="pp_top_prop",ylab="International \n Collaboration")
abline(v=c((max(Q1.top$pp_top_prop)),(max(Q2.top$pp_top_prop)),
            (max(Q3.top$pp_top_prop)),(min(Q90.top$pp_top_prop)),
            (min(Q95.top$pp_top_prop))),col=c("red","red","red","blue","green"),
       lty=c(2,3,4,1,1))
dev.off()
```

D R-code: Copula fitting

```
rm(list = ls()) #clear workspace
setwd("/Users/ashnibachasingh/Documents/BachelorColloqium/BEP") #set workdictory
library(readxl) #needed to read excelfile without having to transfer into txt file
Canadian_researchers_data <- read_excel("~/Documents/BachelorColloqium/BEP/
                                        Orginal Data/Canadian_researchers_data.xlsx")
data.per.division<-Canadian_researchers_data #to make sure that orginal
#data file stays in tact

#turns numeric data registered as non numeric into numeric data
data.per.division$authors_paper<-as.numeric(data.per.division$authors_paper)
data.per.division$institutes_paper<-as.numeric(data.per.division$institutes_paper)
data.per.division$countries_paper<-as.numeric(data.per.division$countries_paper)
data.per.division$pages_paper<-as.numeric(data.per.division$pages_paper)
data.per.division$refs_paper<-as.numeric(data.per.division$refs_paper)
data.per.division$p<-as.numeric(data.per.division$p)
data.per.division$mcs<-as.numeric(data.per.division$mcs)
data.per.division$mncs<-as.numeric (data.per.division$mncs)
data.per.division$pp_top_prop<-as.numeric(data.per.division$pp_top_prop)
data.per.division$mjs_mcs<-as.numeric(data.per.division$mjs_mcs)
data.per.division$mnjs_mncs<-as.numeric(data.per.division$mnjs_mncs)
data.per.division$mnjs_pp_top_prop<-as.numeric(data.per.division$mnjs_pp_top_prop)
data.per.division$pp_collab<-as.numeric(data.per.division$pp_collab)
data.per.division$pp_int_collab<-as.numeric(data.per.division$pp_int_collab)
data.per.division<-na.omit(data.per.division)

dataset<-readline(prompt="Which data/division would you like to use? ")
data.per.division<-subset(data.per.division,division==dataset)

#converts random variates to psuedo-observations
uniformp<-pobs(data.per.division$p,ties.method="random")
uniformmcs<-pobs(data.per.division$mcs,ties.method="random")
uniformmncs<-pobs(data.per.division$mncs,ties.method="random")
uniformtop<-pobs(data.per.division$pp_top_prop,ties.method="random")
uniformcollab<-pobs(data.per.division$pp_int_collab,ties.method="random")

#turns pseudo-observation vectors into matrix for gof test
a<-as.matrix(data.frame(uniformp,uniformmcs))
b<-as.matrix(data.frame(uniformp,uniformmncs))
c<-as.matrix(data.frame(uniformp,uniformtop))
d<-as.matrix(data.frame(uniformp,uniformcollab))
```

```
e<-as.matrix(data.frame(uniformmcs,uniformmncs))
f<-as.matrix(data.frame(uniformmcs,uniformtop))
g<-as.matrix(data.frame(uniformmcs,uniformcollab))
h<-as.matrix(data.frame(uniformmncs,uniformtop))
i<-as.matrix(data.frame(uniformmncs,uniformcollab))
j<-as.matrix(data.frame(uniformtop,uniformcollab))

#selection of copula based on AIC
selectedCopula<-BiCopSelect(uniformtop,uniformcollab,familyset=NA) #switch between variates
selectedCopula
selectedCopula$AIC

#goodness of fit test
fitcop<-frankCopula(6.83) #fill in copula with parameter
gofCopula(fitcop,f,N=2000,estim.method="mpl",simulation="mult")

random<-rCopula(711,fitcop) #sample observation
cor(random,method="spearman")
cor(random,method="kendall")

#By plotting the pseudo and simulated observations we can
#see how the simulation with the copula matches the pseudo observations
#PvsMCS
dev.off()
plot(uniformp,uniformmcs,main="Pseudo/simulated observations: BLUE/RED",
     xlab="Uniform Publications",ylab="Uniform MCS",col="blue")
points(random[,1],random[,2],col="red")

#PvsMNCS
dev.off()
plot(uniformp,uniformmncs,main="Pseudo/simulated observations: BLUE/RED",
     xlab="Uniform Publications",ylab="Uniform MNCS",col="blue")
points(random[,1],random[,2],col="red")

#PvsTop
dev.off()
plot(uniformp,uniformtop,main="Pseudo/simulated observations: BLUE/RED",
     xlab="Uniform Publications",ylab="Uniform pp_top_prop",col="blue")
points(random[,1],random[,2],col="red")

#PvsCollab
dev.off()
plot(uniformp,uniformcollab,main="Pseudo/simulated observations: BLUE/RED",
     xlab="Uniform Publications",ylab="Uniform Collaboration",col="blue")
points(random[,1],random[,2],col="red")

#MCSvsMNCS
dev.off()
plot(uniformmcs,uniformmncs,main="Pseudo/simulated observations: BLUE/RED",
     xlab="Uniform MCS",ylab="Uniform MNCS",col="blue")
points(random[,1],random[,2],col="red")

#MCSvsTop
dev.off()
plot(uniformmcs,uniformtop,main="Pseudo/simulated observations: BLUE/RED",
     xlab="Uniform MCS",ylab="Uniform pp_top_prop",col="blue")
plot(random[,1],random[,2],main="Pseudo/simulated observations: BLUE/RED",
     xlab="Uniform MCS",ylab="Uniform pp_top_prop",col="red")
points(random[,1],random[,2],col="red")
```

```
#MCSvsCollab
dev.off()
plot(uniformmcs,uniformcollab,main="Pseudo/simulated observations: BLUE/RED",
     xlab="Uniform MCS",ylab="Uniform Collaboration",col="blue")
points(random[,1],random[,2],col="red")

#MNCSvsTop
dev.off()
plot(uniformmncs,uniformtop,main="Pseudo/simulated observations: BLUE/RED",
     xlab="Uniform MNCS",ylab="Uniform pp_top_prop",col="blue")
points(random[,1],random[,2],col="red")

#MNCSvsCollab
dev.off()
plot(uniformmncs,uniformcollab,main="Pseudo/simulated observations: BLUE/RED",
     xlab="Uniform MNCS",ylab="Uniform Collaboration",col="blue")
points(random[,1],random[,2],col="red")

#TopvsCollab
dev.off()
plot(uniformtop,uniformcollab,main="Pseudo/simulated observations: BLUE/RED",
     xlab="Uniform pp_top_prop",ylab="Uniform Collaboration",col="blue")
points(random[,1],random[,2],col="red")

#emperical copula
n<-nrow(data.per.division)
X <- uniformmcs
Y <- uniformtop
Z <-data.frame(uniformmcs,uniformtop)
# sort sample
X.ascending <- sort(X)
Y.ascending <- sort(Y)
Z.ascending <- cbind(X.ascending, Y.ascending)
# prepare data structure
Cn <- as.data.frame(matrix(nrow = n, ncol = n), row.names = paste0("X", 1:n))
colnames(Cn) <- paste0("Y", 1:n)
# run through the indizes i (of X) and j (of Y)
for( i in 1:n){
  for( j in 1:n){
    Cn[i, j] <- sum(apply(X=Z, MARGIN = 1, FUN = function(Z.row, x.sorted, y.sorted)
      { sum(Z.row <= c(x.sorted, y.sorted), na.rm = TRUE) >=2 }, X.ascending[i], Y.ascending
      na.rm = TRUE)/n
  }
}
#3Dplot
dev.off()
x<-(1:n)/n
y<-x
persp3D(x,y,as.matrix(Cn),main="Emperical copula")
```