Delft University of Technology

Bachelor thesis

Modelling finite mixture joint distributions

Modelleren van eindige gemengde simultane verdelingen

Written by
Bastiaan Bakker (4605004)
to obtain the degree of Bachelor of Science
at the Delft University of Technology

Supervised by
Dorota Kurowicka

July 20, 2020

Preface

The thesis *Modelling finite mixture models* you are reading at this moment is written in order to obtain the degree of Bachelor of Science. The writing and research has been conducted under supervision of D. Kurowicka from the department of Applied Probability of the faculty of EEMCS at the TU Delft.

In this thesis I will attempt to find a good model for a transportation data set. The modeling will be done with copula functions, which will be discussed in the first chapter. In order to fit a good model, we first define copulas and explain their usage, and after that we show how we estimate parameters for models together with how we compare different copula models with each other.

After we have analyzed our original data set, we will cluster our data set to fit possibly better copula models.

The modelling will be done with regular vine copula models and other multivariate copula models such as the Normal, Student-t or the Clayton copula.

Finally, I would like to give special thanks to D. Kurowicka for the supervision, support and help this semester. I am grateful for the help I have received in the form of weekly Zoom meetings during these difficult times of COVID-19 and the help in how to write scientifically.

Furthermore, I want to thank my roommates in Delft who supported me during these COVID-19 times and were always available to talk about my struggles and process during the writing of this thesis.

Moreover, I would also like to thank M. Keijzer and J. Birkens for taking place in the thesis committee.

In this thesis we will work a lot with the statistical software R. If you, the reader, wants to know the implementation of things in this thesis in R, you can contact me on bastiaan.bakker@upcmail.nl.

B. Bakker
Delft, July 2020

Contents

1 Introduction

Many people travel every day, and due to climate change, electric transportation devices are going to get more popular and more widely used in the future. According to the report on the growth of electric vehicles presented by the International Energy Agency [29], electric car deployment has been growing rapidly over the past ten years, with the global stock of electric passenger cars passing 5 million in 2018, an increase of 63 % of the previous year. With more people using ways of transportation powered by electricity, the demand for electricity rises as well. This demand could add to the existing picks in electricity use as electric vehicles are plugged in to charge when people come home, using electricity for charging at the same time as they turn on the electrical heating at home, start electrical cooking and turn on their lights. Hence it is useful to model the behaviour of people with electric transportation devices. Such models can be used later to design charging strategies that will not overwhelm the electricity network when many electric vehicles are charging their batteries together. More information about this topic can be found in [28] and [29].

My far going interest in sustainability, statistics and probability will settle the goal of this thesis, which is modelling the behaviour of the people with electric vehicles as well as possible. The type of modelling we will do is often used in many areas of applications, for instance in finance, biology or in the area of health to create virtual large population of patients suffering from a disease. These created models can be used to simulate patients from this population and study different treatment strategies.

In this thesis, we will model a data set containing transportation data provided by the Dutch Ministry of Transportation. The data set consists of information on trips made by drivers in the Netherlands who travel by car. We will use part of the data for drivers who make just one trip per day and drive with electric vehicles.

Different models will be considered for this data set. We will first introduce the theory of copula functions which we will use in this thesis. Copulas are so called dependence functions which allow us to model the distribution of random vectors by separating the marginal distributions and the dependence in the joint distribution. This leads to much richer and much more flexible sest of models for the data than the usually used multivariate distributions as the Gaussian distribution.

In this thesis we will consider elliptical, Archimedean and vine copulas. We will try to answer the following questions: "What are copula functions?", "How do copula functions look like?", “What are properties of copulas?” and "What can we do with copula functions?".

After studying the theory of copula functions we move to relevant statistical questions, namely: "How do we estimate parameters of copula models?" and "How do we compare copula models?". We use the statistical software R to illustrate all concepts and perform computations.

After having explained all these concepts, we will look at the data set described in this introduction. Even though the data set we will model is just 3-dimensional, it is very challenging indeed. Variables in this data set have complicated dependencies and it is very difficult to find a probabilistic model that fits this data well. One of the reasons could be that the data set is not homogeneous. There are possibly two or more groups of people in our data set with very different behaviour. This is very intuitive as people use cars for different purposes. The purpose of trips in the data set is not provided so to improve performance of the model we apply clustering techniques to split this data set on groups with similar behaviour. After the different groups are created we build a copula model for each group and join them into a mixture model. This allows to improve performance of the model for this data set.

In the Chapter 2 we will look at copula functions, explain the theory and visualize these functions in 2- and 3-dimensions. Chapter 3 is concerned with statistics of copula models.

In this chapter we will also present an example of modelling a simulated data set which will allow us to test all presented theoretical concepts. In the fourth chapter we will build copula models for the transportation data set treating it as homogeneous data. Then in Chapter 5 this data will be clustered with different methods and the best grouping of the data will be used in Chapter 6 to find a mixture of copulas model. In the last two chapters we will present the conclusions we have made and discuss our findings.

2 Copulas

In this thesis we are required to find a parametric distribution describing a multivariate data set. An elegant way of doing this is to model the marginal distributions and the dependence between them separately, in the form of a dependence function, called a copula. The copula is simply a distribution on the unit hyper-cube with uniform marginal distribution. In this chapter the theory of copulas will be presented. We will discuss different parametric families of copulas. This section is based on [1]

2.1 Copula theory

A d-dimensional copula is a joint distribution with uniform marginal distributions on the interval (0,1) marginal distributions, denoted as $U(0,1)$.

Definition 2.1. A copula $C : (0,1)^d \rightarrow [0,1]$ is a joint distribution function of standard uniform random variables. That is,

$$C(u_1, ..., u_d) = P(U_1 \leq u_1, ..., U_d \leq u_d), \tag{1}$$

where $U_i \sim U(0,1)$ for i = 1,....,d (with $\sim$ denoting that it has the given distribution).

The relationship between a joint distribution function and a corresponding copula is given in the famous Sklar's theorem presented next.

Theorem 1. (Sklar's theorem). Let F be a joint cumulative distribution function (cdf) of random vector $(X_1, ..., X_d)$ with marginal cdfs $F_1, F_2, ..., F_d$. Then there exists a copula C such that

$$F(x_1, x_2, ..., x_d) = C(F_1(x_1), F_2(x_2), ..., F_d(x_d)), \tag{2}$$

If $(X_1, ..., X_d)$ is a continuous random vector, then C is unique.

Sklar's theorem follows naturally from the simple result:

Proposition 1. If X is a continuous random variable with cdf F then $Y = F(X) \sim U(0,1)$.

Proof. To prove the above we have to show that it holds for 0 < y < 1 that:

$$P(Y \leq y) = P(F(X) \leq y) = y. \tag{3}$$

If F is strictly increasing then it is inevitable, otherwise the quasi-inverse defined as $F^{[-1]}(p) = inf\{x \in X; F(x) \geq p\}$ can be used instead of the inverse. If not, then we get

$$\begin{aligned} P(F(X) \leq y) &= P(X \leq F^{-1}(y)) \\ &= F(F^{-1}(y)) \\ &= y \end{aligned}$$

This gives us the result. □

Lemma 1. For every d-dimensional distribution function F with continuous marginals $F_1, ..., F_d$ there exists a unique copula C such that, for all $x = (x_1, ..., x_d) \in \mathbb{R}^d$,

$$F(x) = C(F_1(x_1), ...F_d(x_d)).$$

Such a C is determined, for all $(u_1, .., u_d) \in [0,1]^d$, via the formula

$$C(u) = F\Big(F_1^{[-1]}(u_1), ..., F_d^{[-1]}(u_d)\Big),$$

where, for $i \in \{1, ..., d\}, F_i^{[-1]}$ is the quasi-inverse of F_i.

If $(X_1, ..., X_d)$ is an absolutely continuous random vector with density function f then differentiating equation 2 with respect to $x_1, .., x_d$ we get the density of $(x_1, .., x_d)$ is:

$$f(x_1, x_2, ..., x_d) = f_1(x_1)...f_d(x_d)c(F_1(x_1), ..., F_d(x_d)), \tag{4}$$

where f_i denotes the density of X_i and c is a copula density.

We can observe from the formula above that when the copula density is equal to 1 on the whole unit hyper-cube than the density f is equal to the product of marginal densities and the variables $X_i, i = 1, ..., d$ are mutually independent. Otherwise the copula density provides a weight containing information about how much the dependence between random variables contributes to the product of marginals.

A copula describes fully dependence between random variables. In the next subsection we show how this information can be summarized into just one number, a dependence measure.

2.2 Measures of dependence

The most popular dependence measures for two dimensional random vectors are Spearman's ρ and Kendall's τ. Spearman's rho and Kendalls tau are both rank correlation measures.

Spearman's rho is simply the Pearson correlation of transformed to uniform variables. It is defined as follows:

Definition 2.2. Let X and Y be random variables with respective distribution functions F_X and F_Y, then Spearman's rho is:

$$\rho_s(X, Y) = \rho(F_X(X), F_Y(Y)),$$

where ρ is the Pearson product moment correlation coefficient which can be computed as follows:

$$\rho(X, Y) = \frac{cov(X, Y)}{\sigma_X \sigma_Y},$$

where *cov* is the covariance and σ is the standard deviation.

Kendall's tau is defined using concept of concordance. We consider two independent copies of random vector (X, Y) denoted as (X_1, Y_1) and (X_2, Y_2). The pair is called concordant if $(X_1 - X_2)(Y_1 - Y_2) > 0$ and discordant when $(X_1 - X_2)(Y_1 - Y_2) < 0$. Kendall's tau is defined as a difference of probabilities of concordance and discordance [1]:

Definition 2.3.

$$\tau_K(X, Y) = P((X_1 - X_2)(Y_1 - Y_2) > 0) - P((X_1 - X_2)(Y_1 - Y_2) < 0)$$

There are relationships between copula and Kendall's tau or Spearman's rho.
If X and Y are random variables with continuous marginal distributions and C is a unique copula corresponding to cdf of (X, Y), then Kendall's tau and Spearman's rho are [10]:

$$\tau_K = 4 \int_0^1 \int_0^1 C(u, v) dC(u, v) - 1 \tag{5}$$

$$\rho_s = 12 \int_0^1 \int_0^1 C(u, v) du dv - 3 \tag{6}$$

Note that in the above we write to shorten the notation τ_K instead $\tau_K(X,Y)$ and ρ_s instead of $\rho_s(X,Y)$.

Kendall's tau and Spearman's rho can also be estimated from data. Given a matrix of data containing n observations with columns $X = (X_1, ..., X_n)$ and $Y = (Y_1, ..., Y_n)$. Let Con be the number of concordant pairs and Dis be the number of discordant pairs, then Kendall's tau can be estimated from the data by the following equation [10] [12]:

$$\tau_K = \frac{(Con - Dis)}{n(n-1)/2}$$

To estimate Spearman's rho from the data we define R_X^i and R_Y^i as the rank of the i-th element of the random vector X and the rank of the i-th element of the random vector Y, respectively. The theory of ranking is explained in 3.1.2. Spearmann's rho is then computed as follows [12]:

$$\rho_s = 1 - \frac{6\sum(R_X^i - R_Y^i)^2}{n(n^2-1)}.$$

In case when ties are present in the data, they would have to be resolved to obtain ranks of the observations. Similarly the formula for τ_k will have to be adjusted when ties are present.

Both Spearman's ρ and Kendall's τ measure average dependence between random variables over their domain. In applications it might also be of interest whether the dependence is stronger in certain regions of the domain. The upper (UTD) and lower (LTD) tail dependence coefficients (λ_U, λ_L) are designed to describe how the variables interact in their tails.

The UTD and LTD are defined as follows, where X, Y are continuous random variables with distribution function F_X, F_Y, respectively [1].

Definition 2.4.

$$\lambda_U = \lim_{t \to 1^-} P\Big\{F_Y(Y) > t | F_X(X) > t\Big\} \tag{7}$$

The definition of the LTD is as follows:

Definition 2.5.

$$\lambda_L = \lim_{t \to 0^+} P\Big\{F_Y(Y) \leq t | F_X(X) \leq t\Big\} \tag{8}$$

If $\lambda_U > 0$, then we can say that (X,Y) is upper tail dependent. If $\lambda_U = 0$, then we say that (X,Y) is upper tail independent. Similarly $\lambda_L = 0$ means lower tail independence and $\lambda_L > 0$ lower tail dependence.

As in case of Spearman's rho and Kendall's tau the tail dependence coefficients can be computed using the copula corresponding to the distribution of random vector (X,Y) [1].

$$\lambda_U = \lim_{t \to 1^-} \frac{1 - 2t + C(t,t)}{1-t} \tag{9}$$

$$\lambda_L = \lim_{t \to 0^+} \frac{C(t,t)}{t} \tag{10}$$

2.3 Bivariate copulas

In this subsection the most popular families of bivariate copulas are presented. We show scatter plots and density plots for each of the presented copulas. To be able to appreciate differences between these families they are plotted for the same values of Kendall's tau: 0.3, 0.5, 0.7 and 0.9. Moreover we show what the tail dependencies of all the copulas are.

2.3.1 Elliptical

The first two copulas we will look into are called the elliptical copulas, as they are copulas corresponding to elliptical distribution, namely Gaussian/ Normal and Student-t distributions [6].

Normal copula The Normal copula with parameter $\rho \in (-1, 1)$ is:

$$C_\rho(u_1, u_2) = \int_{-\infty}^{\Phi^{-1}(u_1)} \int_{-\infty}^{\Phi^{-1}(u_2)} \frac{1}{2\pi(1-\rho^2)^{\frac{1}{2}}} \exp\Big\{ -\frac{u^2 - 2\rho uv + v^2}{2(1-\rho^2)} \Big\} du dv,$$

where Φ^{-1} is the inverse of the univariate standard Normal distribution function and ρ is the linear correlation coefficient of normal variables.

The relationships between the parameter of normal copula ρ and Kendall's tau and Spearman's rho are shown below:

$$\rho = \sin\Big(\frac{\pi}{2}\tau_k\Big),$$
$$\rho = 2\sin\Big(\frac{\pi}{6}\rho_S\Big).$$

Where τ_k is Kendall's tau and ρ_S is the Spearman's rho value.

The tail dependence of the Normal copula is as follows (calculations are given in the appendix A.1):

$$\lambda_L(X,Y) = \lambda_U(X,Y) = 2 \lim_{x \to -\infty} \phi\Big(x \frac{\sqrt{1-\rho}}{\sqrt{1+\rho}}\Big) = 0. \tag{11}$$

Hence the tail dependence does not depend on the value of ρ and is always equal to 0 when ρ is not equal to 0. [1]

In figures below we show density plots of the Normal copula with with parameter ρ equal to 0.4539905, 0.7071068, 0.8910065, 0.9876883 which corresponds to Kendall's tau equal to 0.3, 0.5, 0.7 and 0.9, respectively.

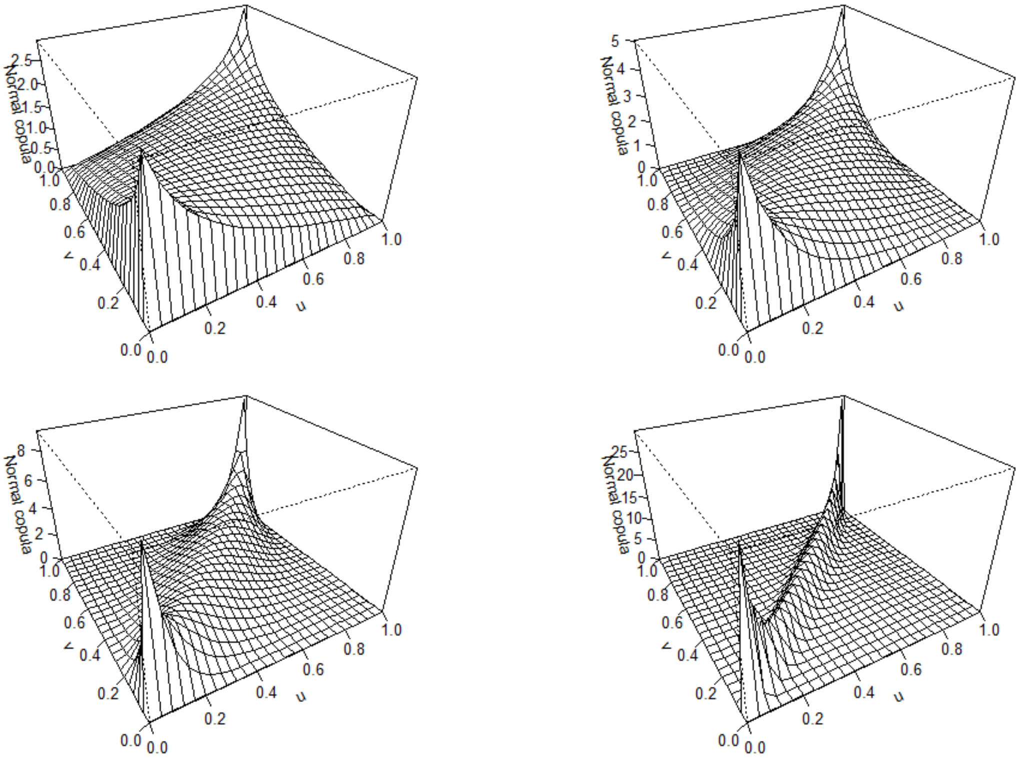

Figure 1: Density plots of normal copulas with parameters corresponding to Kendall's tau of 0.3, 0.5, 0.7 and 0.9.

The figures below show scatter plots generated from normal copulas with parameters corresponding to Kendall's tau of 0.3, 0.5, 0.7 and 0.9. We take n = 8579 in our simulations, this is because our own dataset which we will use later also contains 8579 datapoints.

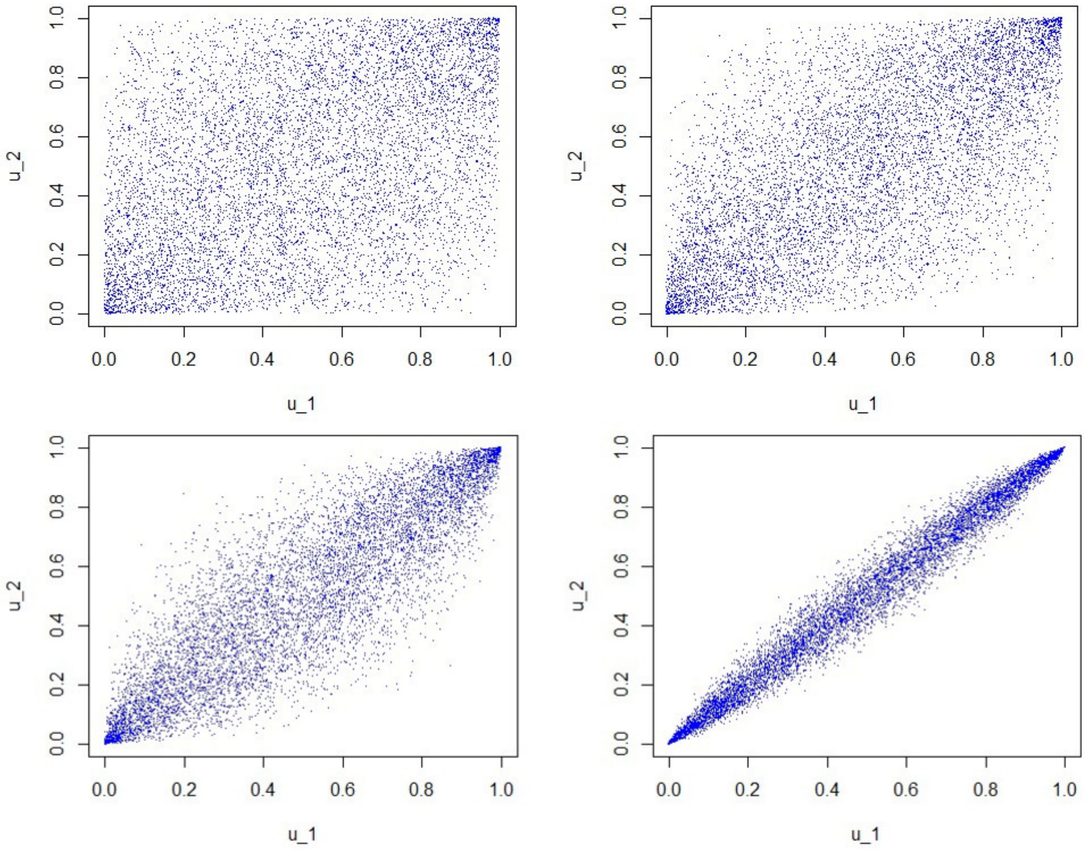

Figure 2: Scatter plots of n=8579 samples from normal copulas with parameters corresponding to Kendall's tau of 0.3, 0.5, 0.7 and 0.9.

As we can see in our scatter plots there is a larger concentration of points in upper and lower corners when correlation increases, however according to the result above the tail dependence coefficients stay equal to zero. The upper and lower tail dependence coefficients are limiting quantities and measure how fast the cdf of copula converges to 1 when both variables converge to 1 (converge to zero when both converge to zero). This is also the reason that they are not easily estimated from data. For more about estimation of tail dependence coefficients from data we refer to [31] and [1].

Student t-copula Another copula of the elliptical family is the Student t-copula. The Student t-copula with parameters $\rho \in (-1, 1)$ and $\nu \geq 2$ is [6]:

$$C_{\rho,\nu}(u_1, u_2) = \int_{-\infty}^{t_\nu^{-1}(u_1)} \int_{-\infty}^{t_\nu^{-1}(u_2)} \frac{1}{2\pi(1-\rho^2)^{\frac{1}{2}}} \Big\{ 1 + \frac{u_1^2 - 2\rho uv + v^2}{\nu(1-\rho^2)} \Big\}^{-(\nu+2)/2} dudv$$

where the parameter ρ is the linear correlation coefficient, ν is the degrees of freedom of the copula and t_v^{-1} is the inverse of the standard univariate Student t-distribution with ν degrees of freedom, expectation 0 and variance equal to $\frac{\nu}{\nu-2}$ [1].

If the parameter ν converges to ∞ then Student-t copula converges to the Normal copula. Hence the Student-t and Normal copula can be considered as nested models. The relationship between parameters of the Student-t copula and Kendall's tau is.

$$\rho = \sin\left(\frac{\pi}{2}\tau\right)$$

There is unfortunately no simple relationship between the parameters of bivariate Student t-copula and Spearman's rho.

The tail dependence coefficient of the Student t-copula is (the calculation of the tail dependence is given in the appendix A.2):

$$\lambda_U = \lambda_L = 2t_{\nu+1}\Big(-\sqrt{\nu+1}\sqrt{\frac{1-\rho}{1+\rho}}\Big) \tag{12}$$

In the figures below density plots of Student-t copulas with ν equal to 3 and ρ equal to 0.4539905, 0.7071068, 0.8910065, 0.9876883 corresponding to Kendall's tau of 0.3, 0.5, 0.7 and 0.9 are shown.

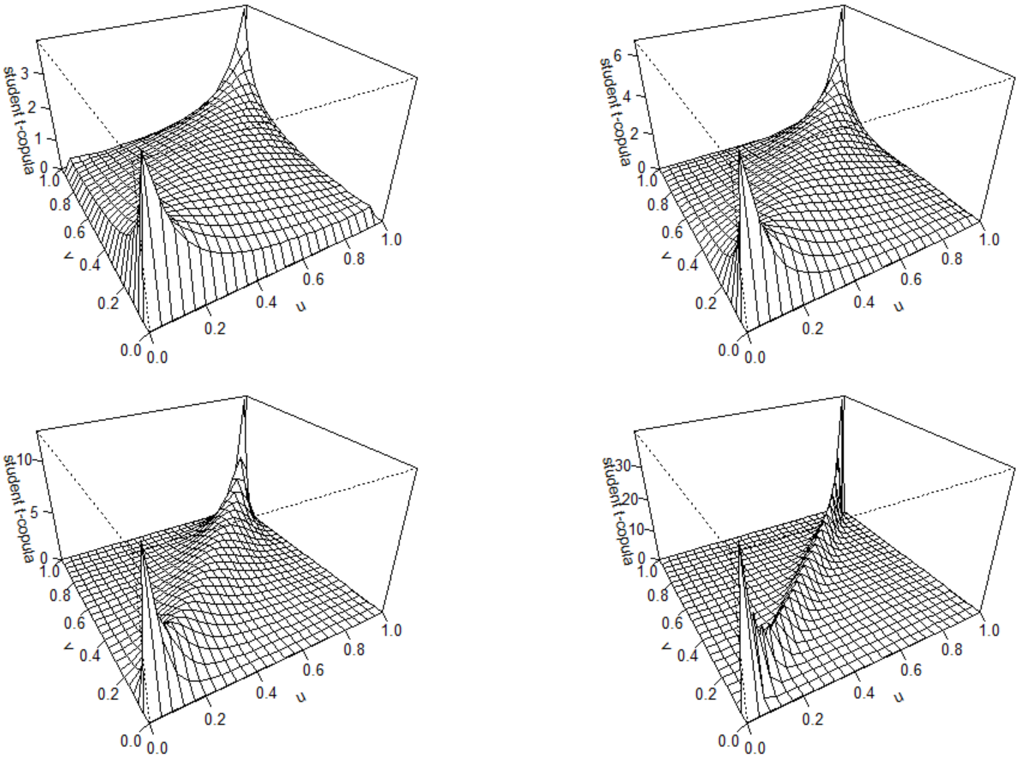

Figure 3: Density plots of Student t-copulas with degree of freedom ν=3 and ρ corresponding to Kendall's tau of 0.3, 0.5, 0.7 and 0.9.

Below we show scatter plots of samples simulated from Student-t copulas with densities presented above.

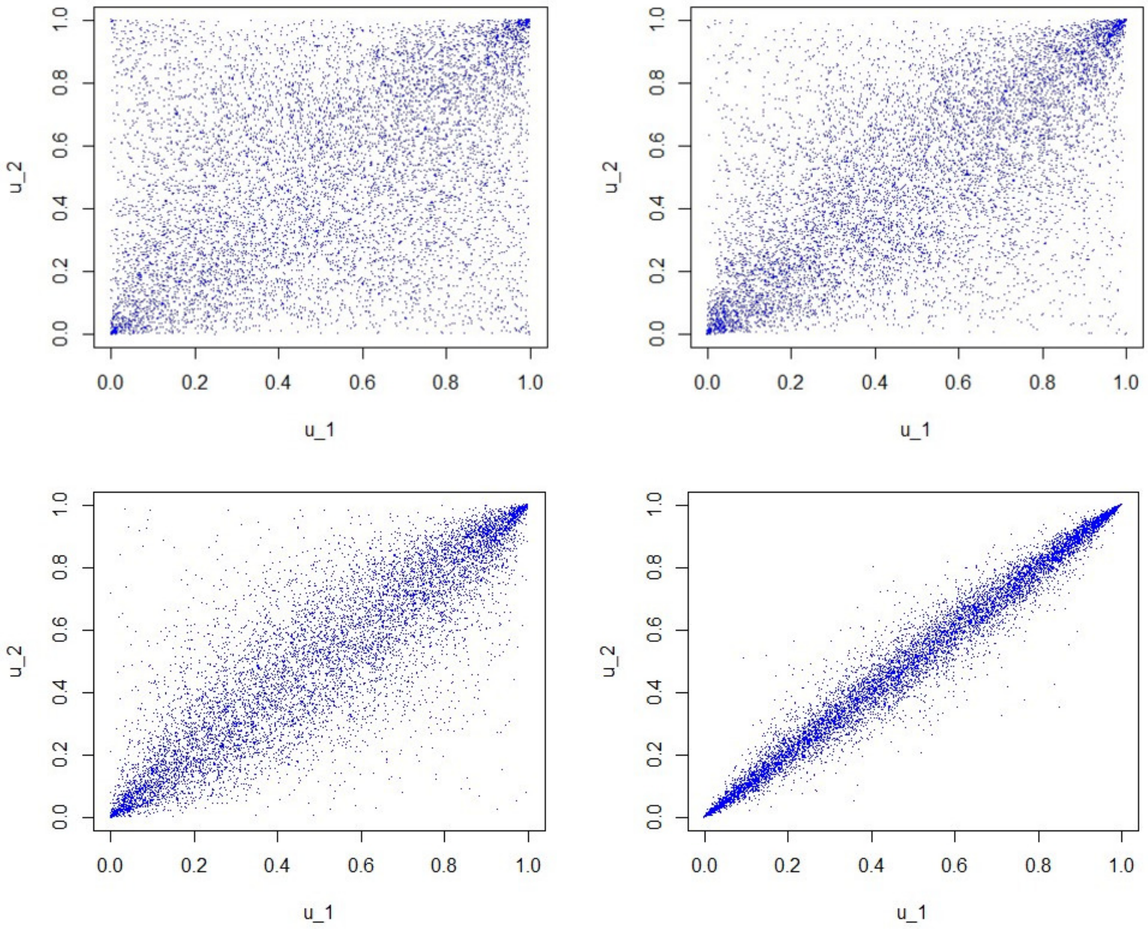

Figure 4: Scatter plots of n=8579 samples from Student t-copulas with degree of freedom ν=3 and ρ corresponding to Kendall's tau of 0.3, 0.5, 0.7 and 0.9.

We also calculated the tail dependence coefficients:

Kendall's tau	parameter ρ, $\nu = 3$	λ_U	λ_L
0.3	0.4539905	0.2875744	0.2875744
0.5	0.7071068	0.4539962	0.4539962
0.7	0.8910065	0.6561929	0.6561929
0.9	0.9876883	0.8825527	0.8825527

So far the densities and scatter plots as well as the computed tail dependence coefficients were shown for ν equal to 3. When we compute the tail dependence for a specific ρ, we see that as we take a bigger ν, the tail dependence gets smaller. There is a restriction on ν that it has to be bigger then 2, so if we take a value for ν close to 2, we get a high tail dependence. If we take a ν much bigger than 2, we get a smaller tail dependence. So the tail dependence gets bigger for larger ρ and smaller ν, and the tail dependence gets lower for smaller ρ and bigger ν. When ν gets extremely large, the Student-t copula converges to the Normal copula and the tail dependence goes to zero.

When comparing the scatter plots of the Normal and the student t copula, the most explicit distinction are the tails. In the student t-copula scatter plots we see a significantly larger concentration of data points in the corners. We visualize this more in the figure below where probability densities of both copulas are plotted next to each other.

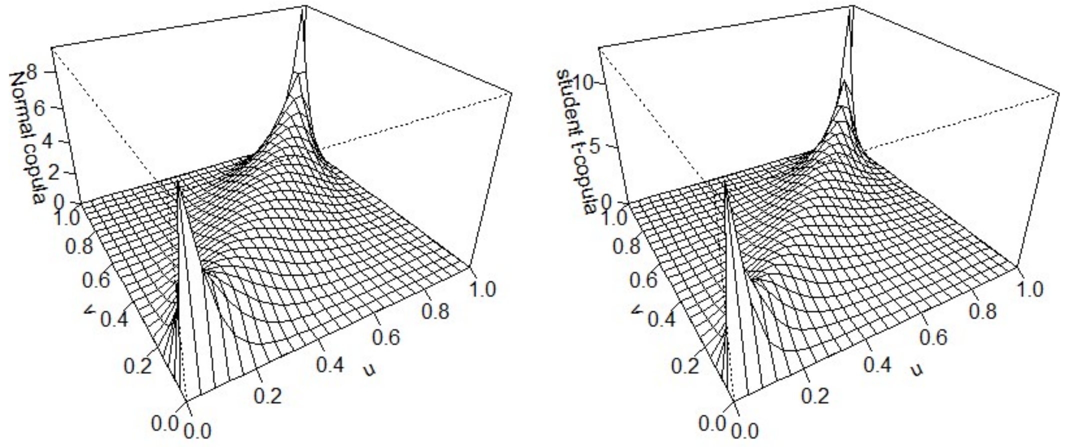

Figure 5: Density plots of the Normal copula and the Student t-copula with degree of freedom ν=3 (for student t-copula) and ρ corresponding to Kendall's tau of 0.7

2.3.2 Archimedean

Another popular type of copulas are the Archimedean copulas. The bivariate Archimedean copulas are defined as follows:

$$C(u_1, u_2) = \phi^{-1}(\phi(u_1) + \phi(u_2))$$

where ϕ is the generator function of the copula. For $C(u_1, u_2)$ to be a copula function, the generator function ϕ has to be 2-monotone on $(0, 1)$. This means ϕ has to satisfy:

$$(-1)^k \phi^{(k)}(x) \geq 0 \text{ for } k = 1, 2$$

Where $x \in (0, 1)$ and $\phi^{(k)}$ is the k^{th} derivative of ϕ. Equivalently, ϕ has to be a continuous, convex and decreasing function and has to satisfy $\phi(1) = 0$ [16].

Clayton copula The first Archimedean copula we will look at is the Clayton copula [9]. The Clayton copula with parameter $\alpha \in \big[-1, \infty\big) \setminus \{0\}$ is [6]:

$$C_\alpha(u_1, u_2) = \max\left(\left[u_1^{-\alpha} + u_2^{-\alpha} - 1\right]^{-1/\alpha}, 0\right)$$

Where the generator function of the Clayton copula is defined as follows:

$$\phi_\alpha(t) = \frac{1}{\alpha}(t^{-\alpha} - 1)$$

The relationship between the parameter of the Clayton copula and Kendall's tau is:

$$\alpha = \frac{2\tau}{1 - \tau}$$

Unfortunately for the Clayton copula, as well as for the Student t-copula, there is no simple formula describing the relationship between the Spearman's rho and its parameter.

The upper tail dependence coefficient of the Clayton copula is zero and the lower tail dependence of the Clayton copula is given by (the calculations of the tail dependencies of the Clayton copula can be found in the appendix A.3):

$$\lambda_L = 2^{-\frac{1}{\alpha}}.$$

In the figures below density plots of Clayton copulas with α equal to 0.8571429, 2, 4.666667 and 18 corresponding to Kendall's tau of 0.3, 0.5, 0.7 and 0.9 are shown.

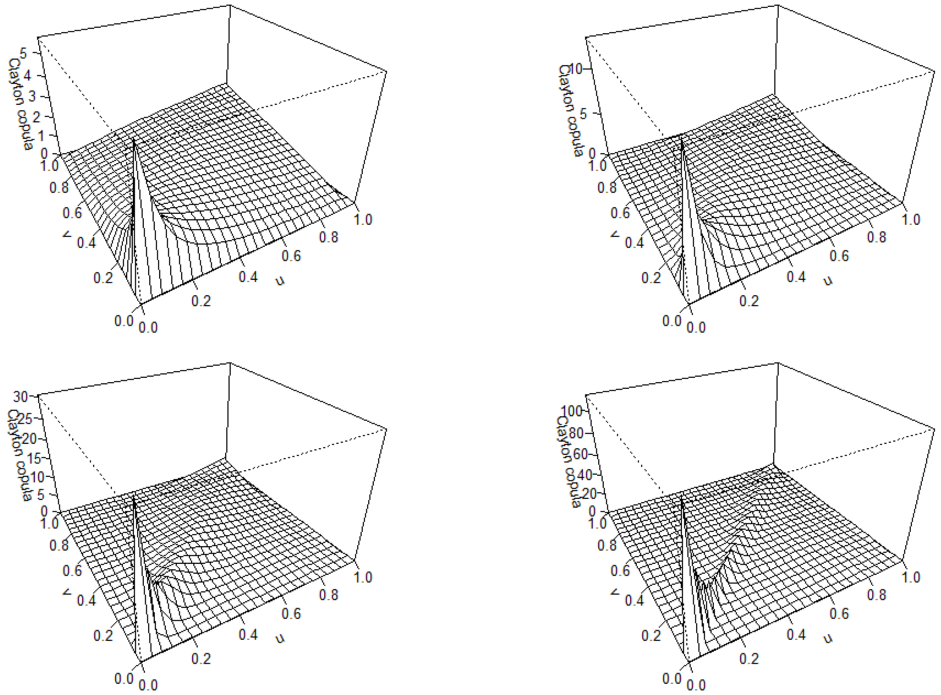

Figure 6: Density plots of Clayton copula with α corresponding to Kendall's tau of 0.3, 0.5, 0.7 and 0.9.

Below we show scatter plots of samples simulated from Clayton copulas with densities presented above.

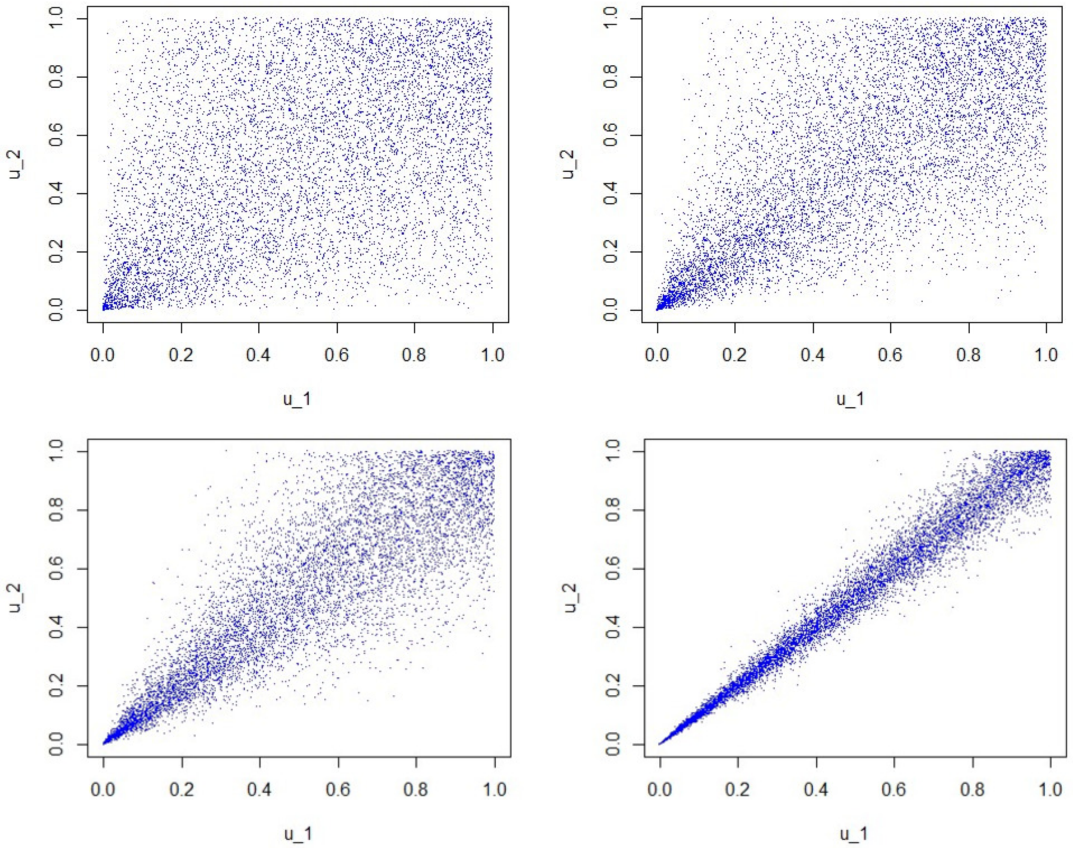

Figure 7: Scatter plots of n = 8579 samples from Clayton copulas with α corresponding to Kendall's tau of 0.3, 0.5, 0.7 and 0.9

We also calculated the tail dependence coefficients:

Kendall's tau	parameter α	λ_L
0.3	0.8571429	0.4454494
0.5	2	0.7071068
0.7	4.666667	0.8619728
0.9	18	0.9622238

The Clayton copula shows really high lower tail dependence as α increases, so the Clayton copula could model data where the data is heavily centered around the lower tail.

Gumbel Copula The second Archimedean copula we will look at is the Gumbel copula. The Gumbel copula with parameter $\alpha \in [1, \infty]$ is [6] [3]:

$$C_\alpha(u,v) = \exp\left\{ -\left[(-\ln u)^\alpha + (-\ln v)^\alpha\right]^{1/\alpha}\right\}$$

Where the generator function of the Gumbel copula is defined as follows:

$$\phi_\alpha(t) = (-\ln t)^\alpha$$

The relationship between parameter α and Kendall's tau is:

$$\alpha = \frac{1}{1-\tau}$$

Unfortunately, there is no closed form for the relation ship between the Clayton copula parameter and Spearman's rho.

The lower tail dependency of the Gumbel copula is equal to zero, the upper tail dependency is given by (calculations of the tail dependencies are given in the appendix A.4):

$$\lambda_U = 2 - 2^{\frac{1}{\alpha}}.$$

In the figures below density plots of Gumbel copulas with α equal to 1.428571, 2, 3.33333 and 10 corresponding to Kendall's tau of 0.3, 0.5, 0.7 and 0.9 are shown.

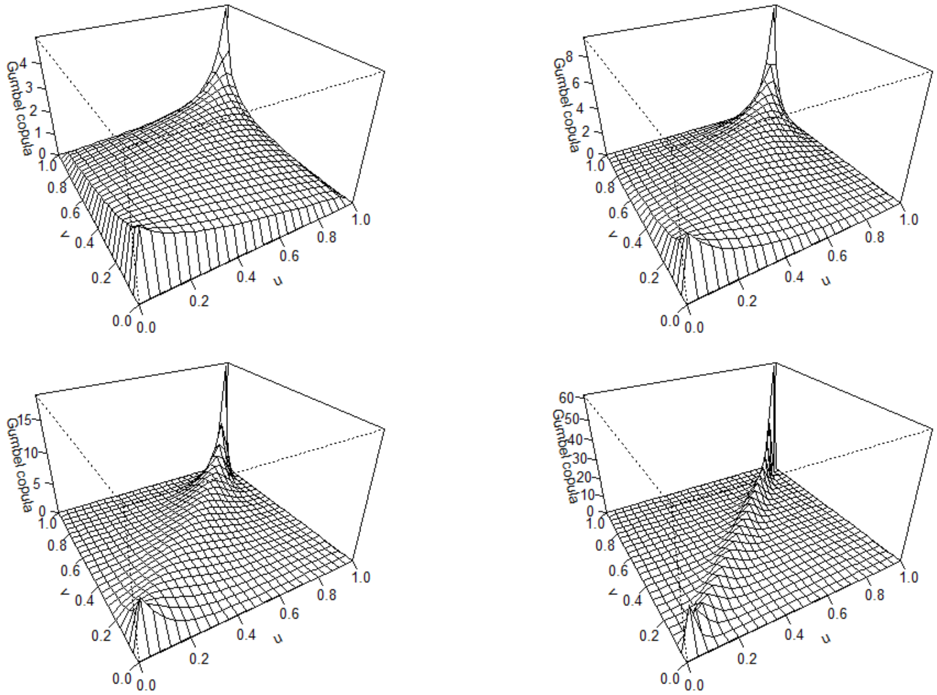

Figure 8: Density plots of Gumbel copula with α corresponding to Kendall's tau of 0.3, 0.5, 0.7 and 0.9.

Below we show scatter plots of samples simulated from Gumbel copulas with densities presented above.

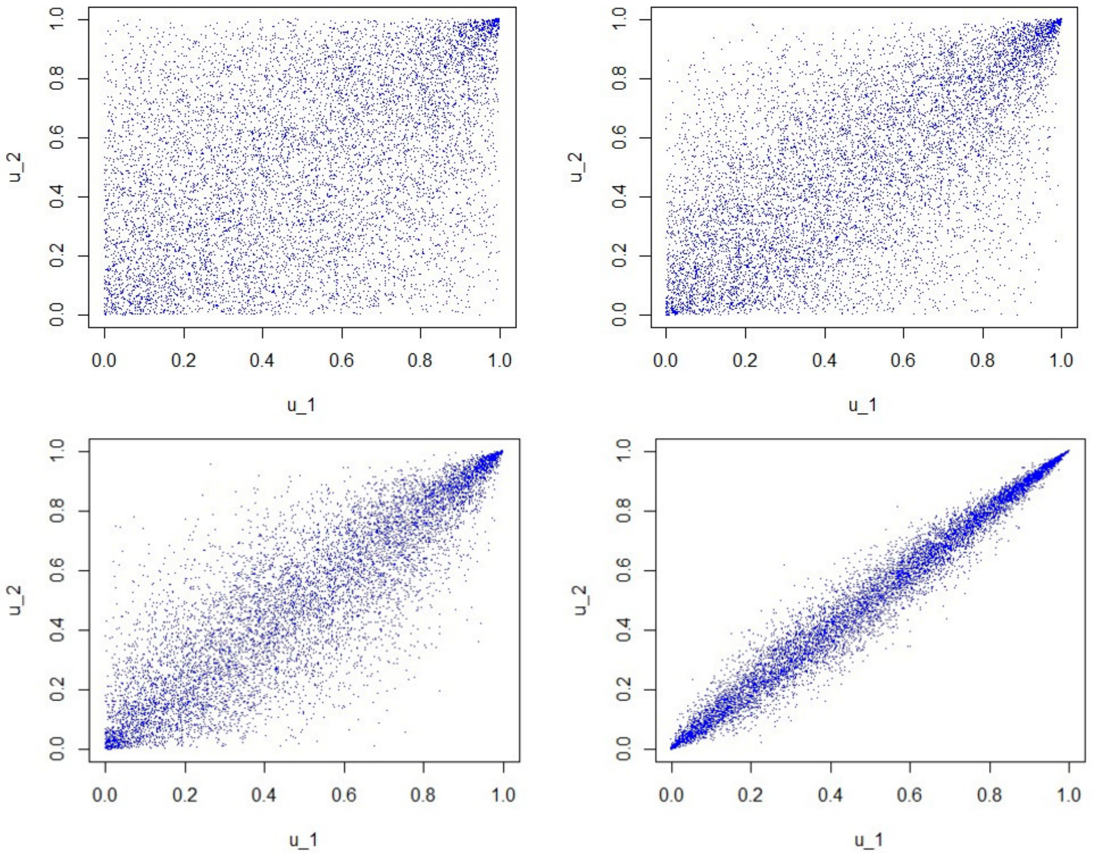

Figure 9: Scatter plots of n = 8579 samples from Gumbel copulas with α corresponding to Kendall's tau of 0.3, 0.5, 0.7 and 0.9

We also calculated the tail dependence coefficients:

Kendall's tau	parameter α	λ_U
0.3	1.428571	0.375495
0.5	2	0.5857864
0.7	3.333333	0.7688556
0.9	10	0.9282265

The Gumbel copula is an asymmetric copula, and in our plots we can see high dependence in the upper tail and low dependence in the lower tail. Comparing this with our tail dependence coefficients, we also see that there is no tail dependence in the lower tail, but a lot of tail dependence in the upper tail. In fact, the Gumbel copula is an extreme value copula, the unique extreme value copula in the Archimedes family.

Frank copula The third Archimedean copula we will look at is the Frank copula. The Frank copula with parameter $\alpha \in \big(-\infty, \infty\big) \setminus \{0\}$ is defined as follows [6]:

$$C_\alpha(u,v) = -\frac{1}{\alpha}\ln\Big(1 + \frac{(e^{-\alpha u}-1)(e^{-\alpha v}-1)}{e^{-\alpha}-1}\Big)$$

Where the generator function of the Frank copula is defined as follows:

$$\phi_\alpha(t) = -\ln\Big(\frac{\exp(-\alpha t)-1}{\exp(-\alpha)-1}\Big)$$

The relationship between the parameter α of the Frank copula and Kendall's tau is:

$$\tau = 1 + \frac{4}{\alpha}(D_1(\alpha) - 1)$$

The relationship between the parameter α of the Frank copula and Spearman's rho is:

$$\rho_S = 1 - \frac{12}{\alpha}(D_2(-\alpha) - D_1(-\alpha))$$

Here we use the Debye function where we define D_k as such:

$$D_k(\alpha) = \frac{k}{\alpha^k}\int_0^\alpha \frac{t^k}{e^t - 1}dt$$

The tail dependencies of the Frank copula are both equal to zero. Calculations of these tail dependencies are given in the appendix A.5.

In the figures below density plots of Frank copulas with α equal to 2.933163, 5.817526, 11.43621 and 38.30123 corresponding to Kendall's tau of 0.3, 0.5, 0.7 and 0.9.

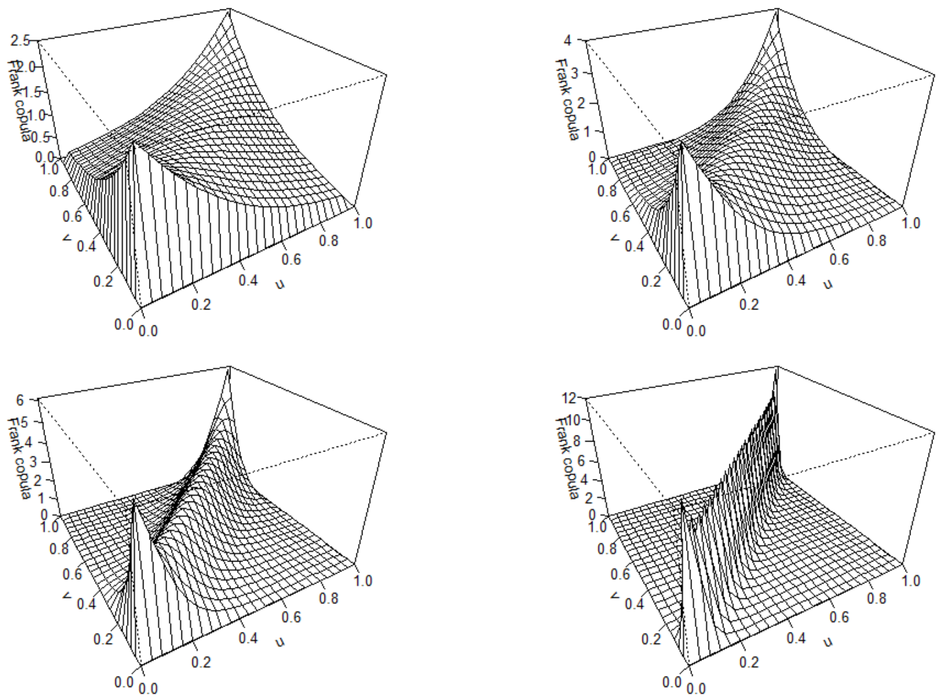

Figure 10: Density plots of Frank copula with α corresponding to Kendall's tau of 0.3, 0.5, 0.7 and 0.9

Below we show scatter plots of samples simulated from Frank copulas with densities above.

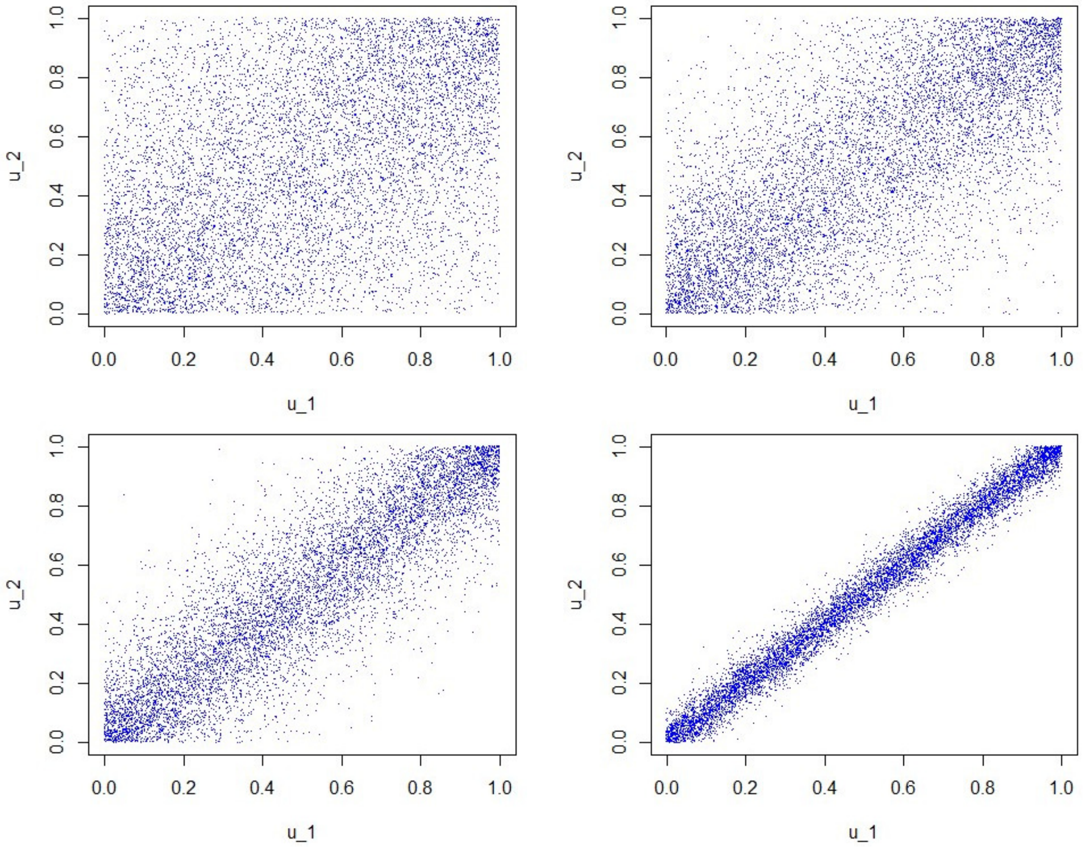

Figure 11: Scatter plots of n = 8579 samples from Frank copulas with α corresponding to Kendall's tau of 0.3, 0.5, 0.7 and 0.9.

As we can see in these probability density plots and scatter plots, there is no upper and lower tail dependence due to the absence of clear tails. Since the parameter of the Frank copula can be any value possible except for zero, the copula is a very flexible copula which could be a good model for symmetric data without tail dependence.

To appreciate the differences between copulas from the Archimedean family, we plot the probability density functions of them next to each other with Kendall's tau equal to 0.7.

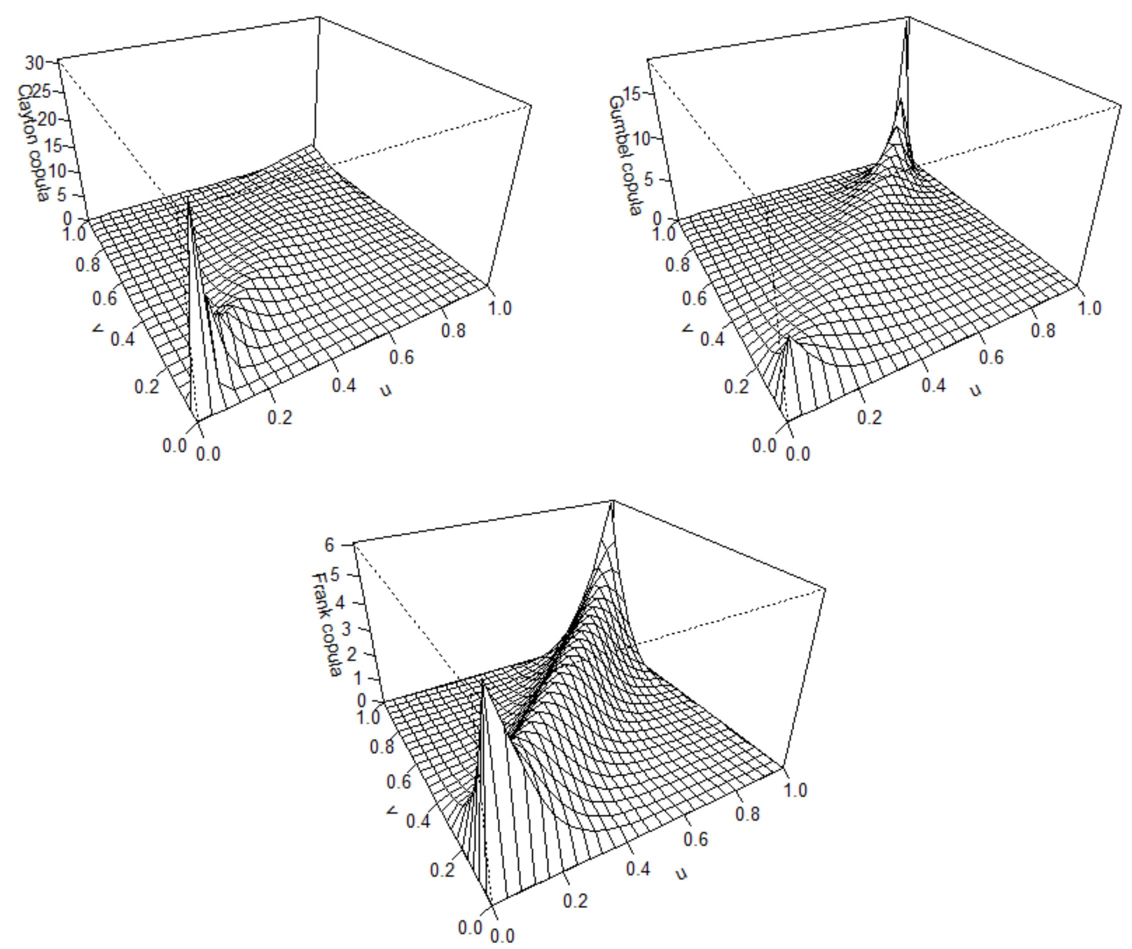

Figure 12: Density plots of the Clayton, Gumbel and Frank copula with ρ corresponding to Kendall's tau of 0.7

They all behave differently, although all describing the same correlation.

Some copulas e.g Gumbel and Clayton allow modelling of only positive correlation. This is however not a restriction since copulas can be rotated. Copula C(u_1, u_2) rotated by 90°, 180°, and 270°can be computed

$$\begin{aligned} C_{90}(u_1, u_2) &= C(1 - u_1, u_2) \\ C_{180}(u_1, u_2) &= C(1 - u_1, 1 - u_2) \\ C_{270}(u_1, u_2) &= C(u_1, 1 - u_2). \end{aligned}$$

Below we see the scatter plots of a bivariate Clayton, Clayton 90°, Clayton 180°and Clayton 270°copula with parameter corresponding to correlation 0.7.

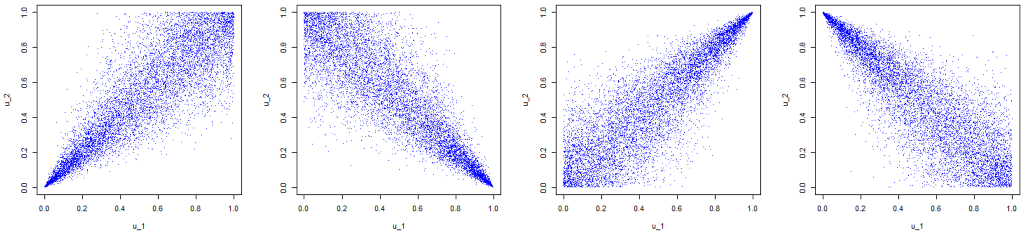

Figure 13: scatter plot of a bivariate Clayton, Clayton 90°, Clayton 180°and Clayton 270°copula with Kendall's tau equal to 0.7

As we can see, the rotated Clayton copula can model different correlations and tail dependencies.

2.4 Multivariate copulas

In this subsection the multivariate copulas are presented. We show the scatter plots of the 3-dimensional simulated data sets of each one of the presented copulas. To be able to appreciate differences between these families they are plotted for the same value of Kendall's tau, 0.7 for all three two dimensional margins.

2.4.1 Elliptical

The first two multivariate copulas we will look into are the elliptical copulas, as they are copulas corresponding to elliptical distribution, namely the Gaussian/Normal and Student-t distributions [16].

Multivariate Normal copula The multivariate Normal copula is defined as follows:

$$C(u_1, ..., u_d; R) = \phi_R^d(\phi^{-1}(u_1), ..., \phi^{-1}(u_d)),$$

where ϕ_R^d denotes the standardizes d-variate normal distribution with correlation matrix R and ϕ^{-1} denotes the quantile function of a univariate standard normal distribution.

The multivariate Normal copula has density function:

$$c(u_1, ..., u_d) = \frac{e^{-(1/2)x^{'}(R^{-1}-I_d)x}}{|R|^{1/2}}$$

Where x $= (\phi^{-1}(u_1), ..., \phi^{-1}(u_d))^{'}$.

If each bivariate margin of the normal distribution is also normal then each bivariate margin of normal copula is the normal copula with its parameter corresponding to the appropriate entry of R. Hence each bivariate margin has zero upper and lower tail dependence index. Below we show a scatter plot of the bivariate margin (u_1, u_2). Remaining margins are the same as this data is simulated from the exchangeable model.

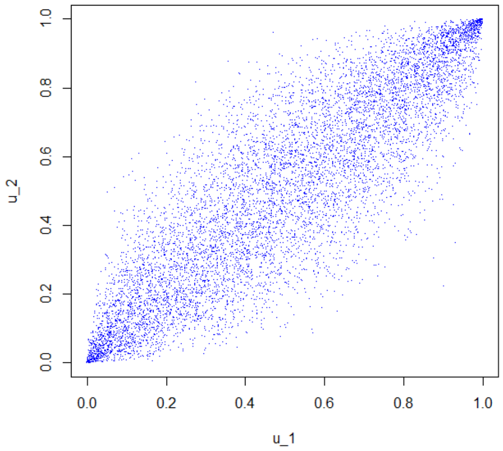

Figure 14: Scatter plot of n = 8579 samples of bivariate margin (u_1, u_2) of exchangeable 3-dimensional Normal copula with parameter corresponding to Kendall's tau of 0.7.

Multivariate Student t-copula The copula of the multivariate student t-copula is defined as [16]:

$$C(u_1, ..., u_d; \rho, \nu) = t^d_{\rho,\nu}(t^{-1}_\nu(u_1), ..., t^{-1}_\nu(u_d)),$$

where t^{-1}_ν denotes the inverse function of the classical univariate t-distribution.

The density function of the student t-copula is defined as follows:

$$c(u_1, ..., u_d; \rho, \nu) = \frac{1}{\sqrt{\rho}} \frac{\Gamma(\frac{\nu+d}{2})}{\Gamma(\frac{\nu}{2})} \left(\frac{\Gamma(\frac{\nu}{2})}{\Gamma(\frac{\nu+1}{2})} \right)^d \frac{\prod_{j=1}^d \left(1 + \frac{t^{-1}_\nu(u_j)^2}{\nu}\right)^{\frac{\nu+1}{2}}}{\left(1 + \frac{x'\rho^{-1}x}{\nu}\right)^{\frac{\nu+d}{2}}}$$

Where x = $(t^{-1}_\nu(u_1), ..., t^{-1}_\nu(u_d))$, ν is the degrees of freedom, Γ is the Gamma function and ρ is defined as:

$$\rho = (\rho_{ij})_{1 \leq i,j \leq d}$$

Below we show the scatter plots simulated from the Student t multivariate copula.

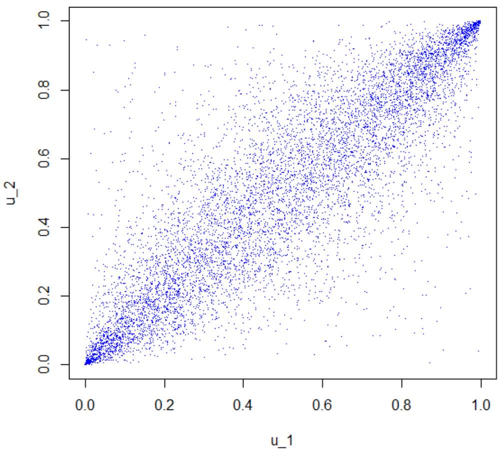

Figure 15: Scatter plot of n = 8579 samples of bivariate margin (u_1, u_2) of exchangeable 3-dimensional Student-t copula with parameter corresponding to Kendall's tau of 0.7.

2.4.2 Archimedean

Given a strict generator $\phi : [0,1] \longrightarrow [0, \infty]$, we can define the multivariate Archimedean copulas as follows:

$$C(u_1, ..., u_d) = \phi^{-1}(\phi(u_1) + ... + \phi(u_d))$$

Where C is a copula if and only if the function ϕ^{-1} is completely monotonic on $\mathbb{R}_+$ [16].

Clayton copula The multivariate Clayton copula can be given by [9]:

$$C(u_1, ..., u_d) = \left(\sum_{j=1}^{d} u_j^{-\delta} - \delta + 1 \right)^{-1/\delta}$$

The density function of the Clayton copula is defined as follows:

$$c(u_1, ..., u_d; \delta) = \delta^d \frac{\Gamma(\frac{1}{\delta} + d)}{\Gamma(\frac{1}{\delta})} \left(\prod_{j=1}^{d} u_j^{-\delta-1} \right) \left[\sum_{j=1}^{d} u_j^{-\delta} - d + 1 \right]^{-1/\delta - d}$$

Below we show the scatter plots simulated from the Clayton multivariate copula.

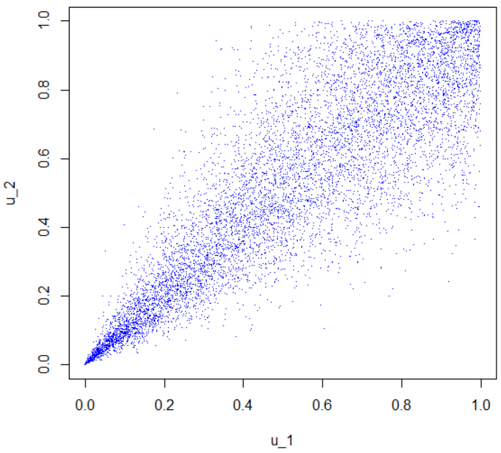

Figure 16: Scatter plot of n = 8579 samples of bivariate margin (u_1, u_2) of exchangeable 3-dimensional Clayton copula with parameter corresponding to Kendall's tau of 0.7.

Gumbel copula The multivariate Gumbel copula can be given by [5]:

$$C(u_1, ..., u_d) = \exp\left[\left(-\sum_{j=1}^{d}(-\log u_j)^{\delta}\right)^{-1/\delta}\right]$$

Below we show the scatter plots simulated from the Gumbel multivariate copula.

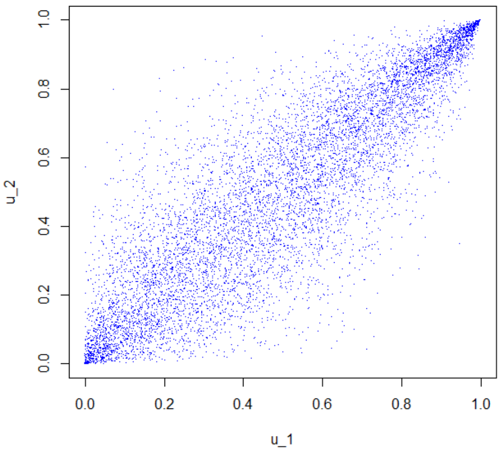

Figure 17: Scatter plot of n = 8579 samples of bivariate margin (u_1, u_2) of exchangeable 3-dimensional Gumbel copula with parameter corresponding to Kendall's tau of 0.7.

Frank copula The multivariate Frank copula can be given by [5]:

$$C(u_1, ..., u_d) = -\frac{1}{\alpha}\left(1 - (1 - e^{-\alpha})^{-d+1}\prod_{i=1}^{d}(1 - \exp(-\alpha u_i))\right)$$

Below we show the scatter plots simulated from the Frank multivariate copula.

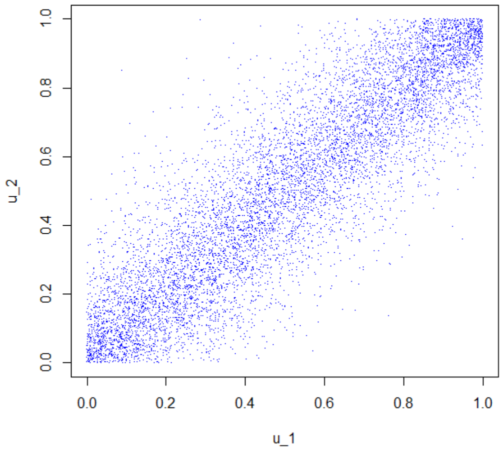

Figure 18: Scatter plot of n = 8579 samples of bivariate margin (u_1, u_2) of exchangeable 3-dimensional Frank copula with parameter corresponding to Kendall's tau of 0.7.

2.4.3 Regular vine copulas

There are many flexible copula models in the 2-dimensional case, as seen in the beginning of this chapter. Their extensions to multivariate case are unfortunately not sufficiently flexible as we have seen in the scatter plots of the multivariate copulas, where every scatter plots shows similar behaviour with its corresponding 2-dimensional copula. The Normal as well as the Student-t copula allow different correlation values for bivariate margins. These correlation values have to specify a positive definite matrix. Each bivariate margin is of the same type, hence all bivariate margins have the same tail dependence. Extensions of Archimedaen copulas presented above allow for only one generator function, hence they lead to unchangeable models.

In this section vine copulas are presented. The presentation is based on [18]. The main idea of vine construction is to build a multivariate copula by combining bivariate and conditional-bivariate copulas using a graphical structure, called the regular vine. This construction allows any family of bivariate copulas to be used and does not impose any constraints on parameters of these copulas allowed in the construction of multivariate copulas. The density function for the regular vine copula can be given by [18]:

$$f(u_1, ..., u_d) = \prod_{j=1}^{d-1} \prod_{i=1}^{d-j} c_{i,i+j|i+1,....,i+j-1} * \prod_{k=1}^{d} f_k(u_k) \tag{13}$$

with

$$c_{i,j|i_1,....,i_k} := c_{i,j|i_1,....,i_k}(F(u_i|u_{i,1}, ..., u_{i,k}), (F(u_j|u_{i,1}, ..., u_{i,k})) \tag{14}$$

for $i, j, i_1, ..., i_k$ with $i < j$ and $i_1 < ... < i_k$.

With the formula above, we are able to represent a density $f(x_1, ..., x_d)$ as a product of pair copula densities and marginal densities. To make the formula above more understandable, we present it for a 3-dimensional density. Such density's can be decomposed in three different ways shown below:

$$\begin{aligned} f_{123}(x_1, x_2, x_3) &= f_{3|12}(x_3|x_1, x_2) \times f_{2|1}(x_2|x_1) \times f_1(x_1) \\ &= f_{2|31}(x_2|x_3, x_1) \times f_{1|3}(x_1|x_3) \times f_3(x_3) \\ &= f_{1|23}(x_1|x_2, x_3) \times f_{3|2}(x_3|x_2) \times f_2(x_2). \end{aligned}$$

Where the following equations hold

$$\begin{aligned} f_{3|12}(x_3|x_1, x_2) &= c_{13|2}(F_{1|2}(x_1|x_2), F_{3|2}(x_3|x_2))f_{3|2}(x_3|x_2) \\ f_{3|2}(x_3|x_2) &= c_{23}(F_2(x_2), F_3(x_3))f_3(x_3) \\ f_{2|1}(x_2|x_1) &= c_{12}(F_1(x_1), F_2(x_2))f_2(x_2). \end{aligned}$$

Hence we obtain the following three expressions for $f_{123}(x_1, x_2, x_3)$

$$\begin{aligned} f_{123}(x_1, x_2, x_3) &= c_{13|2}(F_{1|2}(x_1|x_2), F_{3|2}(x_3|x_2)) \\ &\times c_{12}(F_1(x_1), F_2(x_2))c_{23}(F_2(x_2), F_3(x_3)) \\ &\times f_3(x_3)f_2(x_2)f_1(x_1), \end{aligned}$$

$$\begin{aligned} f_{123}(x_1, x_2, x_3) &= c_{23|1}(F_{2|1}(x_2|x_1), F_{3|1}(x_3|x_1)) \\ &\times c_{12}(F_1(x_1), F_2(x_2))c_{13}(F_1(x_1), F_3(x_3)) \\ &\times f_2(x_2)f_3(x_3)f_1(x_1), \end{aligned}$$

$$\begin{aligned} f_{123}(x_1, x_2, x_3) &= c_{12|3}(F_{1|3}(x_1|x_3), F_{2|3}(x_2|x_3)) \\ &\times c_{13}(F_1(x_1), F_3(x_3))c_{23}(F_2(x_2), F_3(x_3)) \\ &\times f_1(x_1)f_2(x_2)f_3(x_3). \end{aligned}$$

Each one of the three representations contains the product of margins and a 3-dimensional copula which is build as a product of two unconditional copulas and one conditional copula. If we take a look at the first expression for our density given above as an example, $f_3(x_3)f_2(x_2)f_1(x_1)$ are the marginals, $c_{12}(F_1(x_1), F_2(x_2))$ and $c_{23}(F_2(x_2), F_3(x_3))$ are the unconditional copulas and $c_{13|2}(F_{1|2}(x_1|x_2), F_{3|2}(x_3|x_2))$ is the conditional copula. The notation used for the conditional copula should actually be $c_{13|2}(F_{1|2}(x_1|x_2), F_{3|2}(x_3|x_2), x_2)$ to indicate that this conditional copula can depend directly on the conditional variable (its parameter can depend directly on the conditioning variable or even its form can change depending on the conditioning variable). If the conditional copula is assumed not to depend directly on the conditioning variable then we deal with the so called simplifying assumption. In this rapport only simplified vine copulas will be discussed. Due to this assumption, we get an approximation of the density and it leads to the difference of performance of different decompositions of the density.

The three decompositions of the density can be matched to the three graphical representations, called regular vines, plotted below. The graphical structure where circles represent random variables and edges between nodes 1,2 and 2,3 will be assigned with unconditional copulas c_{12} and c_{23}. The extra edge denoted as 13|2 that joints edge (1,2) and (2,3) will be assigned with the conditional copula $c_{13|2}$

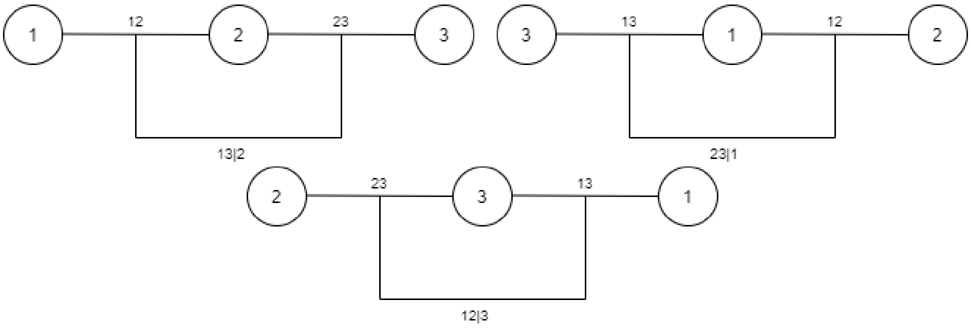

Figure 19: 3D Vine copula constructions

In the 3-dimensional case there are only 3 possible regular vine structures and 3 density decomposition's. The number of possible regular vine structures corresponds to the number of density decomposition's in every dimension. Below we can see the structure of a 3-dimensional vine copula with different copula families with parameter values included in () assigned to the edges of vine structure.

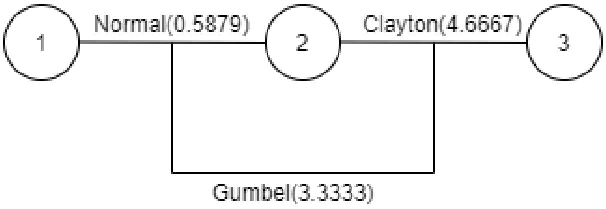

Figure 20: 3D Vine copula structure

The scatter plot of n = 8579 samples from the vine copula shown in figure 20 is presented below. We visualize these in the form of a matrix with the three scatter plots with the regression line given in red below the diagonal, histograms on the diagonal and Kendall's tau for every bivariate pair above the diagonal. In this thesis we will refer to these type of matrix visualizations as 'Matrix visualization of ....'.

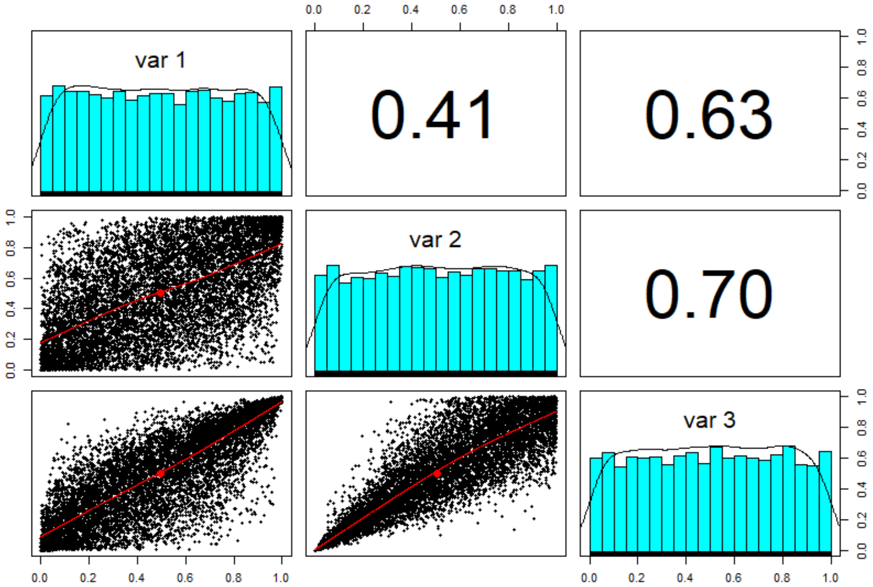

Figure 21: Regular vine copula, from left to right: histogram of u_1, Kendall's tau of (u_2, u_1), Kendall's tau of (u_3, u_1), scatter plot of (u_1, u_2), histogram of u_2, Kendall's tau of (u_3, u_2), scatter plot of (u_1, u_3), scatter plot of (u_2, u_3), histogram of u_3.

We see that different bivariate copulas are 2 dimensional margins of the vine copula presented in figure 20. The copula of variable 1 and 2 is Normal as specified, variables 2 and 3 are joint by the Clayton copula (we clearly see lower tail dependence in the scatterplot of (var1, var2)). The bivariate copula of (var1, var3) is specified via conditional copula $c_{13|2}$. There is no general result to figure out what parametric copula this could be, when $c_{13|2}$ is Gumbel and copulas c_{12}, c_{23} Normal and Clayton, respectively.

3 Estimation and comparison of models

In this chapter we present how to estimate parameters of copula based models. Moreover, different methods of comparing performance of parametric models will be discussed as well as a goodness of fit test that we intend to use in this thesis. All theoretical concepts are then used to model a data set simulated in 2.4.3 from the vine copula. We estimate parameters and compare performance of Normal, Student-t and vine copula models on this data.

3.1 Parameter estimation

In this section we explain how parameters of copulas are estimated from data. We explain this for the situation where we work with a three dimensional data set.

3.1.1 Likelihood

Lets assume that we have a data set in the form of a matrix with $X_{i,j}$, $i = 1, ..., n$, $j = 1, 2, 3$ with n independent observations from parametric density f with parameters θ and ω, then the log likelihood function of parameters θ and ω in this model is:

$$l(\theta, \omega) = \sum_{i=1}^{n} \log c(F_1(X_{1i}; \omega_1), F_2(X_{2i}; \omega_2), F_3(X_{3i}; \omega_3); \theta) + \sum_{i=1}^{n} \sum_{k=1}^{3} \log f_k(X_{ki}; \omega_k) \tag{15}$$

With c being the copula density with parameter θ and f_k being the marginal densities with parameters ω_k.

Hence the log likelihood can be separated as follows:

$$\begin{aligned} l(\omega, \theta; u) &= \sum_{k=1}^{n} \log(f_{1,2,3}(u_{1k}, u_{2k}, u_{3k}; \omega, \theta)) \\ &= l_M(\omega; u) + l_C(\omega, \theta; u) \end{aligned}$$

where u = $(u_1, ..., u_n)$, l_M is the likelihood of the margins and l_C the likelihood of the copula.

In principle the parameters of a copula depend on the values of the parameters of the margins. However, often estimation of θ and ω is done sequentially, where the maximum likelihood estimate of ω_k is used to transform data to uniforms (called pseudo observations) and then copula parameter θ are estimated on the pseudo observations.

Alternatively, margins are modeled with the empirical distributions and data is transformed to uniforms via ranking explained in 3.1.2.

For vine copula models l_C can be rewritten as (assuming the structure in figure 20):

$$l_C(\omega, \theta; u) = \sum_{i=k}^{n} \Big[\log c_{12}(F_1(u_{1i}), F_2(u_{2i})) + \log c_{23}(F_2(u_{2i}), F_3(u_{3i})) + \log c_{13|2}(F_{1|2}(u_{1i}|u_{2i}), F_{3|2}(u_{3i}|u_{2i})) \Big]$$

In this case estimations of parameters of copula is done sequentially by first estimating parameters of unconditional copulas and transforming data trough conditional cdf's of estimated unconditional copulas and then estimating parameter of conditional copulas.

3.1.2 Ranking data

Ranking corresponds to applying empirical cdf's to data. A simple example is given below, where the upper part of the table are the values we observe in our random vector, and the lower part of the table are the new assigned values for these data points, the ranks.

Data point value	2	4	7	12	5	8	9	13	1	10
Rank	2	3	5	8	4	6	7	9	1	10

We have got 10 data points in this random vector, so the biggest data point gets rank 10 and the smallest data point gets rank 1. When there are equal data points in your data set, so called ties, their ranks will be equal to the average of ranks that these values would get if they were different. An example of this is shown below:

Data point value	2	4	7	2	5	2	9	5	1	10
Rank	3	5	8	3	6.5	3	9	6.5	1	10

If we divide every rank given in the data set by the number of data points in total in the random vector plus one, our new ranked data set values will be in the [0,1] interval. The ranking of a data set is done in R by the rank() function.

3.2 Model comparison

If two models have the same number of parameters, the likelihood function is sufficient to decide which model is more appropriate for the data. In this section we present methods to compare models with different number of parameters.

3.2.1 Copula selection

Due to restrictions in some functions which we will use in R only the following copulas are selected for modelling in this thesis: Normal, Student-t, Clayton, Gumbel, Frank, Joe, Clayton 180 °(SC), Gumbel 180 °(SG), Joe 180 °(SJ), Clayton 90 °(C90), Gumbel 90 °(G90), Joe 90 °(J90), Clayton 270 °(C270), Gumbel 270 °(G270) and Joe 270 °(J270). Some copulas will be abbreviated further in this thesis by the abbreviation in the brackets behind the copula names.

3.2.2 AIC, BIC

Let k be the number of estimated parameters in the model and n be the amount of data points of the model. The AIC (Akaike information criterion) is computed as follows [7]:

$$AIC = 2k - 2 * \log L(\theta)$$

The BIC (Bayesian Information Criterion) is computed as follows:

$$BIC = \ln(n)k - 2 * \log L(\theta)$$

The AIC and BIC value are simply penalized log likelihood measures. The lower they are the better the model.

3.2.3 Likelihood Ratio Test

Assume that we have two parametric nested models. The set of possible parameters for model 1 is Θ_0 and model 2 is Θ, where Θ_0 is a subset of Θ. We will test the null hypothesis H_0 that model 1 is better than model 2 against the alternative hypothesis H_1 that model 2 is better than model 1.

The likelihood ratio test statistic is:

$$\lambda_{LR} = -2\log\frac{\max\limits_{\theta\in\theta_0} L(\theta)}{\max\limits_{\theta\in\Theta} L(\theta)} \tag{16}$$

With λ_{LR} being the likelihood ratio value. For nested models, the ratio of the likelihoods is always smaller than one, so λ_{LR} is always positive. It is shown in [11] that:

$$\lambda_{LR} \sim \chi^2_{df}$$

With *df* equal to the amount of parameters of the bigger model minus the amount of parameters of the smaller model.

3.2.4 Vuong closeness test

When models are not nested the Vuong test can be used to test the null hypothesis (H_0) that model 1 and 2 are equally good against the alternative hypothesis (H_1) that model 1 is a better model than model 2. If the models are not nested then the likelihood ratio statistic (16) might become negative. Hence the Vuong statistic is simply the normalized log of likelihoods of both models which is shown in [19] to be approximately standard normally distributed. In this thesis we will use the RVineVuongTest() function in R to perform the Vuong test for regular vine models.

3.2.5 Goodness of fit test

The Goodness of fit test we will use in this thesis is implemented in the VineCopula package in R as the function RVineGOFTest(). This test is performed for parametric vine copula models. For the copula density with parameter vector θ the expected Hessian matrix (matrix of second order derivatives of log likelihood with respect to θ) is compared with the expected outer-product of gradients of the log likelihood. Under the null hypothesis that the data is generated for this copula they should be equal (shown by the theorem of White (1982) [27]). In [35] the function of eigenvalues of both matrices is used as the test statistics which is shown to be asymptotically chi-square distributed with p(p+1)/2 degrees fo freedom, where p is the amount of parameters of the tested model.

3.2.6 Constant Conditional Correlation test

The Constant Conditional Correlation test is used to test the simplifying assumption for regular vine copula models as already discussed in 2.4.3 and is implemented in R as the pacotest() function. The null hypothesis is that conditional copulas do not directly depend on the conditioning variables and the alternative hypothesis is that there is at least one copula that does depend on the conditioning variable(s). If one wants to test whether the vine with copulas c_{12}, c_{23} and conditional copula $c_{13|2}$ is simplified, the parameters of copulas c_{12}, c_{23} are estimated, the data is transformed to pseudo observations $U_{1|2}$, $U_{3|2}$ and by considering different partitions of samples depending on values of variable U_2 it is checked if rank correlations on these partitions stay relatively constant. The test statistic is chosen to be the difference of correlations on different partitions which is shown to be asymptotically normally distributed. Obviously different partitions can be taken and in the pacotest implementations, the partions leading to maximizing such difference are searched using decision trees. For more information about this test we refer to [25].

3.3 Example

To explain all concepts presented above we will give an example. The regular vine shown in figure 20 is used to sample the data set which we will use in this example. Moreover we have transformed margins of this data set to N(0,1) for the first variable, Exponential distribution with the rate equal to two for the second variable and the Student-t distribution with degrees of freedom equal to five for the third variable. The matrix visualisation of this data set is given in the figure below.

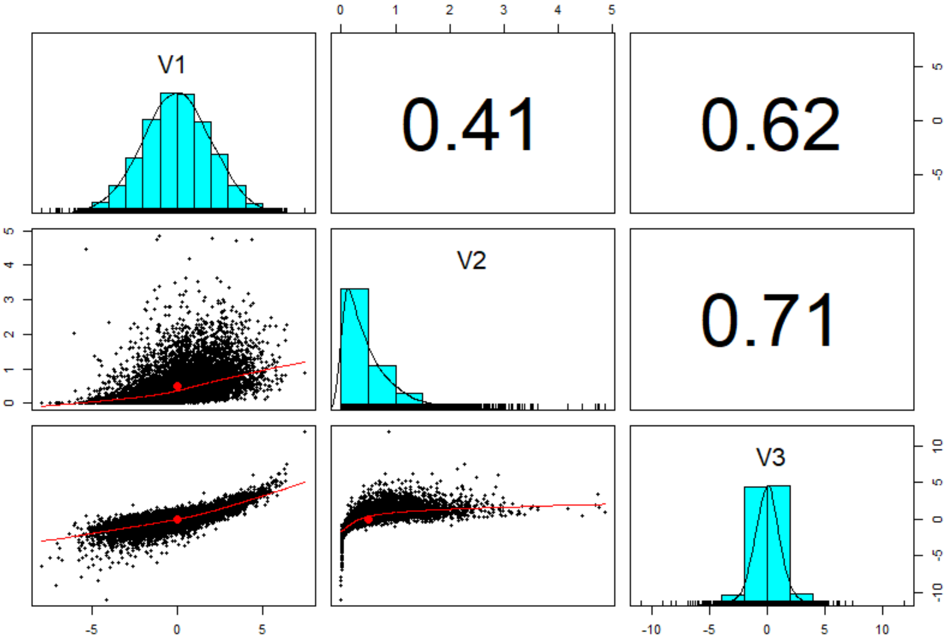

Figure 22: Matrix visualization of the Vine copula data set

If we want to model the data, we have 2 approaches to choose from. We can model the margins and simulate pseudo observations, and model these pseudo observations with copula functions or we can rank the whole data set and divide every data point by the amount of data points plus one. We have done both methods and the results are presented below.

The estimated distributions and parameters of the marginal distributions are:

Marginal distribution	Estimated distribution	parameter(s)
u_1	Normal	0, 2
u_2	Exponential	2
u_3	Student-t	df = 5

After applying the cumulative distribution functions to the data we get the following matrix visualization:

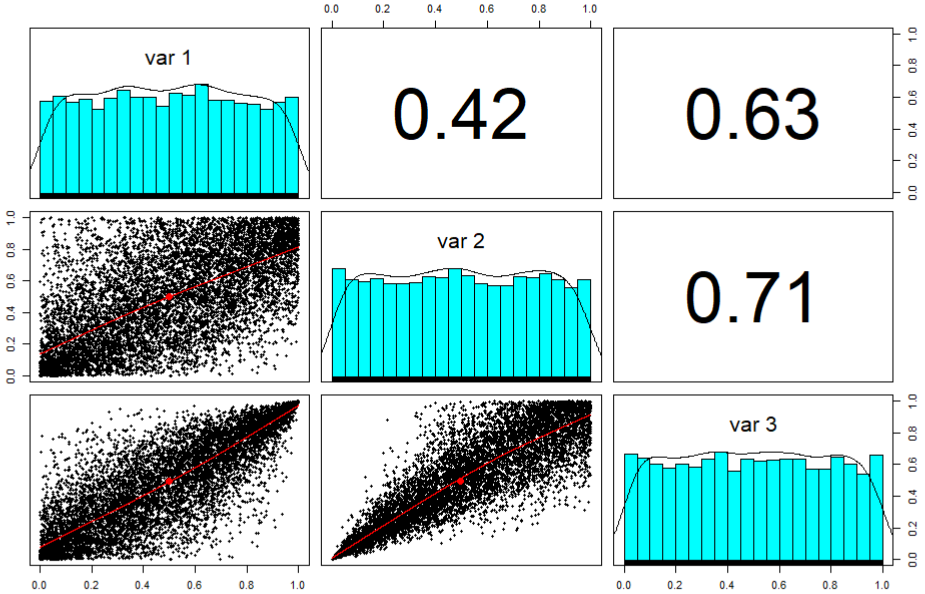

Figure 23: Matrix visualization of the data set after modeling the margins and simulating pseudo observations.

We can recognize the Normal, Gumbel and Clayton copula in the scatter plots. These were the original copulas of the data set, so we have successfully modeled the margins.

When we rank the data we get the following matrix visualization:

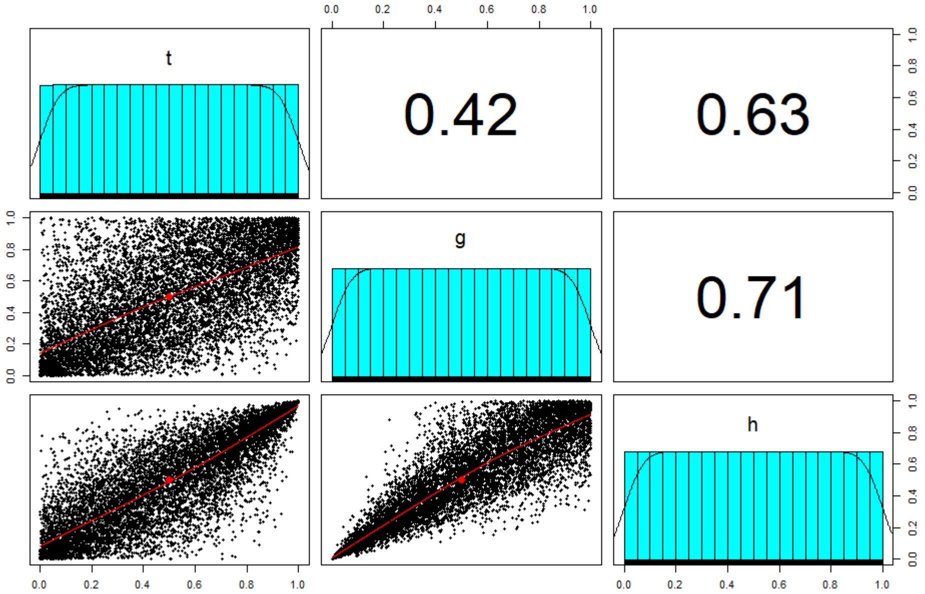

Figure 24: Matrix visualization of the ranked data set

In this plot we can also recognize the original copulas, so also the ranking has given us the right copulas which we started with. We will now model the original data set, the one without the added marginal distributions.

We will first fit a Normal copula and a Student-t copula model to the data set. In the table below we have given the maximum likelihood estimates of parameters of the Normal and Student t models. The Copulafit() function in R is used to compute maximum likelihood estimates for the estimation of parameters of these models. The results are given below.

Model	max log likelihood	parameter(s)
Normal	11327	$\rho_1 = 0.5979$, $\rho_2 = 0.8293$, $\rho_3 = 0.8555$
Student t	11931	$\rho_1 = 0.5709$, $\rho_2 = 0.8085$, $\rho_3 = 0.8743$, $\nu = 4.7619$

The Student-t model has a higher likelihood than the Normal model, but it also contains one more parameter. The higher likelihood suggests that the Student t model will be the better model for this data set. In the table below the AIC, BIC and amount of parameters of the models are given.

Model	AIC	BIC	#parameters
Normal	-22647.16	-22625.99	3
Student t	-23854.21	-23825.98	4

The AIC and BIC of the Student-t model is lower, so this suggests that the Student-t model is a better model for our data. To see if the difference in the likelihoods of both models is significant we will use the likelihood ratio test, which gives us the following results:

$$\lambda_{LR} = -2\log(\frac{11327}{11932}) = 0.10407$$

The likelihood ratio test in R is performed with the function lrtest(). This function gives us a p-value of $2.2 * 10^{-16}$, which is much smaller than our significance level ($\alpha = 0.05$). This leads us to reject the null hypothesis H_0 in favour of the alternative hypothesis H_1. We can then conclude that the Student t model is a better model than the Normal model for the data set.

When we look at the scatter plots of the data set and the Student-t model, we see that the model is not the best fit for our data. The scatter plot below shows that it is clear that the model does not fit the data set very well. We only show the (u_2, u_3) margin for comparison. Obviously this has been expected as the data we are using has been generated with the distribution where this margin's copula is the Clayton copula.

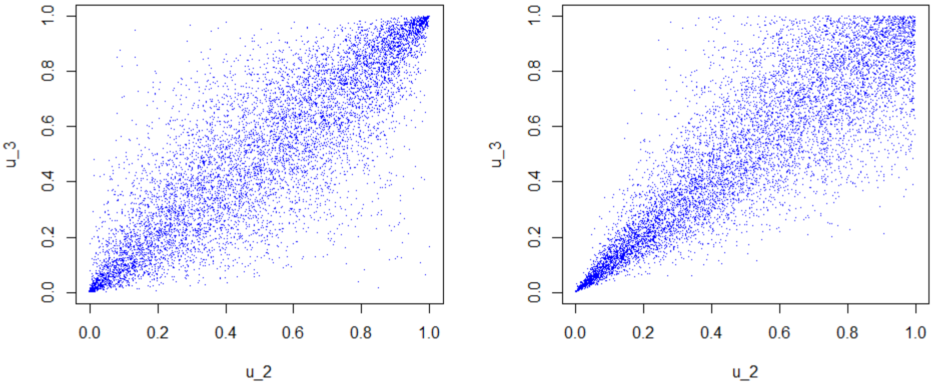

Figure 25: Left: scatter plot Student-t model. Right: corresponding scatter plot of the original data set

Now we have fitted a Normal and Student-t copula model to the data, we will fit regular vine copula models to the data set. The parameters for the regular vine copula models are estimated in R with the function RVineCopSelect(). The structure of the vine is fixed and the copulas are chosen from the set of available copulas. Three copula models called Rvine1, Rvine2 and Rvine3 with chosen copula families are estimated. Parameters are shown in figure 26. Moreover we show in the table below the likelihood, AIC, BIC and the number of parameter for each of these vines.

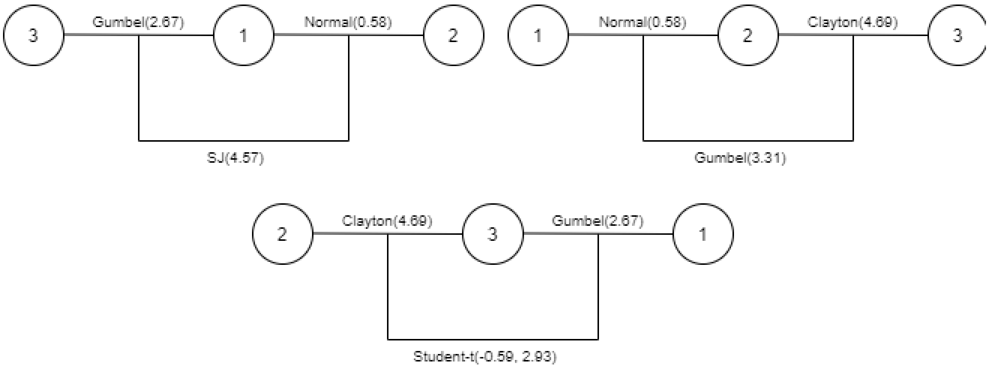

Figure 26: From left to right: regular vine structure of Rvine1, Rvine2 and Rvine3.

Model	Likelihood	AIC	BIC	#parameters
Rvine1	13482.5	-26959	-26937.83	3
Rvine2	16492.57	-32979.13	-32957.96	3
Rvine3	15241.97	-30475.94	-30447.71	4

From the table above we see that Rvine2 has the highest likelihood and the lowest AIC and BIC. Therefore we can conclude that Rvine2 is the best fit for our data set.

To see if differences between these vine models are statistically significant we use the Vuong test. Results are shown in the table below:

Model 1 - Model 2	Vuong statistic	p-value
Rvine1 - Rvine2	-30.71357	$3.750136 * 10^{-207}$
Rvine1 - Rvine3	-22.40461	$3.548729 * 10^{-111}$
Rvine2 - Rvine3	25.32448	$1.7174 * 10^{-141}$

We can conclude that Rvine2 is a better model than Rvine1, Rvine3 is a better model than Rvine1 and Rvine2 is a better model than Rvine3. Rvine2 is the best model for the data and it is significantly better than the other other two competitors.

To see if Rvine2 model is a good model for the data we apply a goodness of fit test.

Model	Test statistic	p-value
Rvine2	13.9786	0.8559459

From the p-value we can not reject H_0 and conclude that Rvine2 is a good fit for the data.

The results of the constant conditional correlation test are shown below:

Model	CCC statistic	p-value
Rvine1	1.079615	0.2987838
Rvine2	0.7420465	0.3890061
Rvine3	224.4055	0

For both Rvine1 and Rvine2 the simplified assumption can not be rejected at the significance level of 0.05. This assumption was however strongly rejected for the Rvine3 model.

3.3.1 Conclusion

After looking at the likelihood, AIC, BIC and the Vuong test we can conclude that the Rvine2 model is the best model for our data set. The goodness of fit test gave us also the result that the Rvine2 model is a good fit for our data. To further show this, we look at the scatter plots of our model and the example data set, which can be found in the appendix A.6. Here we can see that our model fits the data set very well. It was expected that Rvine2 would be the best as the vine used to simulate the example data is almost as the vine used for Rvine2. Below we can see the regular vine structure of the original data set on the left, and the Rvine2 regular vine structure on the right together with the the chosen copulas as well as the estimated and original parameters.

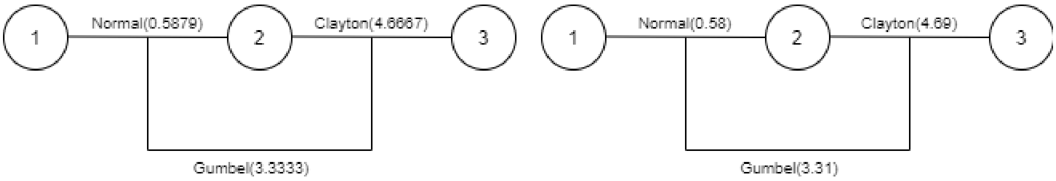

Figure 27: The regular vine structure of the original data set on the left, Rvine2 on the right.

4 Analysis of 3-dimensional data set

Now we have learned how to estimate parameters and compare different models, we will apply this knowledge to a real data set.

4.1 The data set

In this thesis we are working with a 3-dimensional data set. This data set contains transportation data, provided by the Dutch Ministry of Transportation. The data set consists of information on trips made with a car in the Netherlands by persons taking part in a transportation survey. In the data set we have only taken the data of people who made just one trip from home to home in a day. Each respondent was eligible to provide the details of his/her commuting activities for one day: time of departure from home (td), time of arrival at home (ta) and the distance of the trip (d). Below we see the matrix visualization of the data.

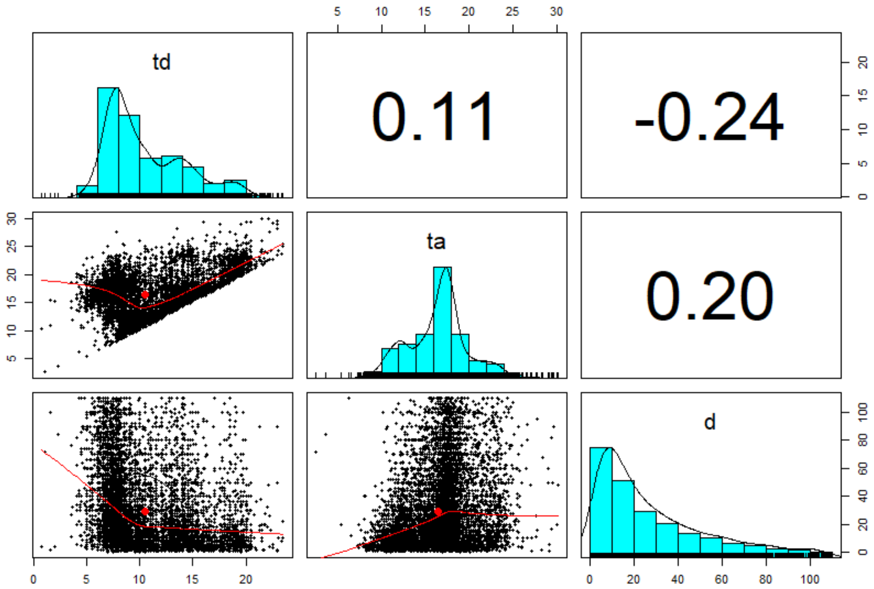

Figure 28: Matrix visualization of the data set

We see that the histograms of the marginal distributions all have different shapes. The correlations between the variables are not high, but they do give us valuable information, the sign of correlations makes intuitive sense. The positive correlation of ta and td makes sense since if you leave home later, you will probably arrive home later as well. Negative correlation of d and ta is reasonable as if you travel further, you will probably arrive later at home etc.

We observe quite interesting behaviour in the scatter plot of (td, ta). All points are lying above a line. This is due to the fact that in practice someones time of arrival is always later than someones time of departure. If we want to model this data with probabilistic methods it might happen that ta will be smaller than td. To avoid this problem we change ta in the data set into $ta - td$. $ta - td$ represents the time spend between the departure and arrival, hence the time of being away from home. The matrix visualization of this new data set is given below.

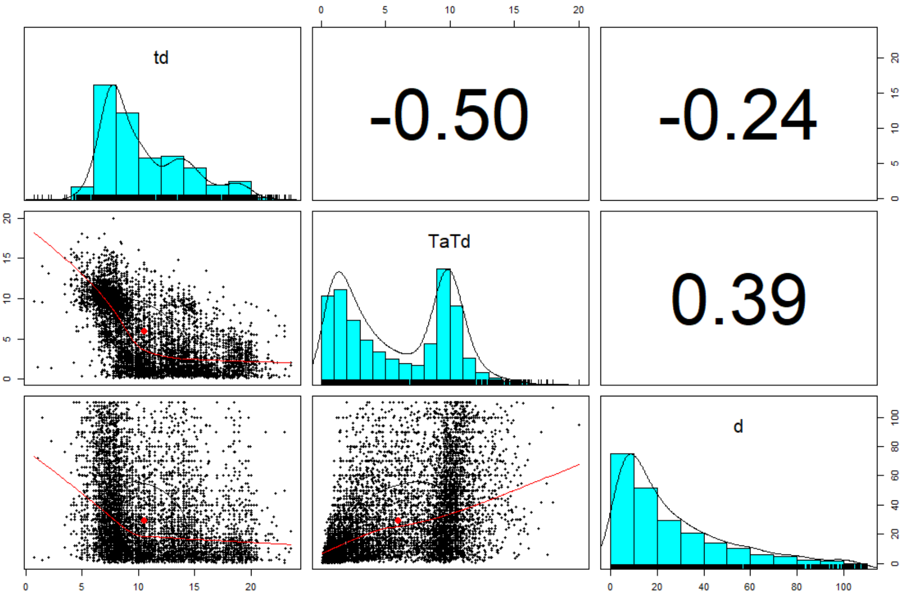

Figure 29: Matrix visualization of the data set where ta is changed to $ta - td$

The only histogram that changed is the histogram for $ta - td$. This histogram shows interesting behaviour. It appears that it contains two different peaks. This could suggest that we maybe have two groups in our data set, people who are away for a short time (the first peak in the $ta - td$ histogram) and people who are away for a long time (the second peak in the $ta - td$ histogram). We will discuss this possibility of 'two different groups of people in our data set' in the last two chapters of this thesis. The correlation between $ta - td$ and td is large and negative, because when people are away for a longer time, their time of departure will be most likely earlier. There is a positive correlation between d and $ta - td$, because when people travel further, the duration of their travel might be longer.

4.2 Ranking the data set

The ranking method, as explained in 3.1.2, is used for this data set. The matrix visualization of the new ranked data set is given below.

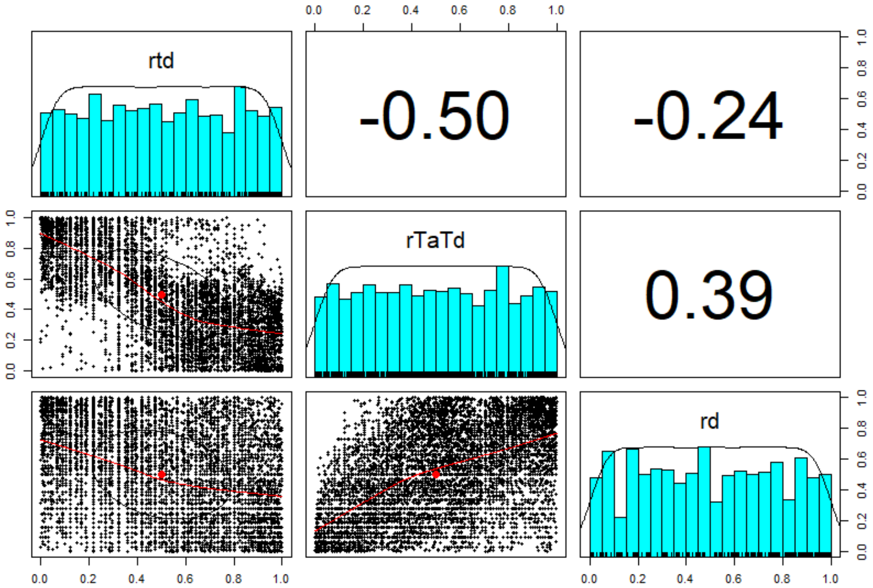

Figure 30: Matrix visualization of the ranked data set where ta is changed to $ta - td$

In the scatter plots above, we can observe 'white lines'. They are caused by existence of repeated values (due to rounding of) in the data, so called ties, which are also explained in 3.1.2. Existence of ties affects estimation of parametric models as well as performance of statistical tests [26]. There have been attempts in the literature to deal with tied data when dealing with copula models [38], but the solutions are not yet satisfactory and are computationally very expensive. In this thesis we choose to transform the data to one without ties by jittering, hence resolving ties at random. This procedure adds small noise to the data and leads to biased estimator of copula parameters [26]. We do the jittering in R with the jitter(x, factor = 0.05), where x is the data vector you want to jitter and factor = 0.05 ensures us that only a small amount of noise will be added to our data set. To further demonstrate how jittering works, we will show an example of jittering where we use the same data set with ties as in the ranking example 3.1.2.

Data point value	2	4	7	2	5
Jittered data point value	1.9951946	3.9932421	7.0092693	1.9944074	5.0011469
Data point value	2	9	5	1	10
Jittered data point value	1.9930611	8.9910950	4.9937137	0.9957343	10.0089346

4.3 Jittering the data set

Below we can see the matrix visualisation for the jittered data set without being ranked.

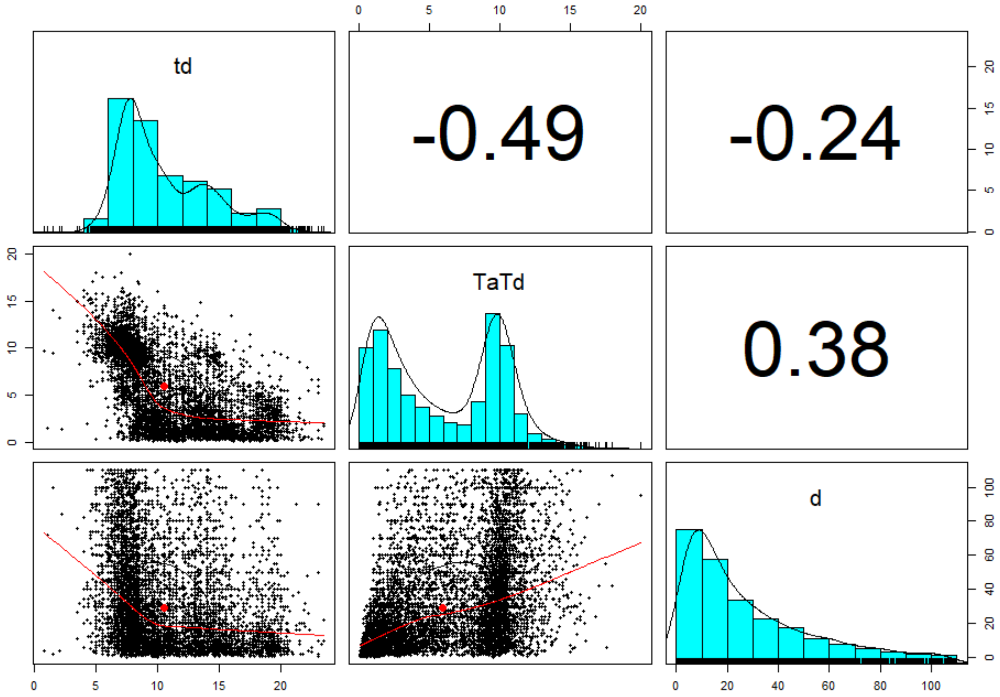

Figure 31: Matrix visualisation of the data set after jittering.

If we compare this matrix visualization with figure 29 we see that correlations did change slightly. The correlations become a bit smaller as we have introduced noise to the data by jittering. However the scatter plots as well as histograms did not change as compared to figure 29. The jittered data after ranking is given below.

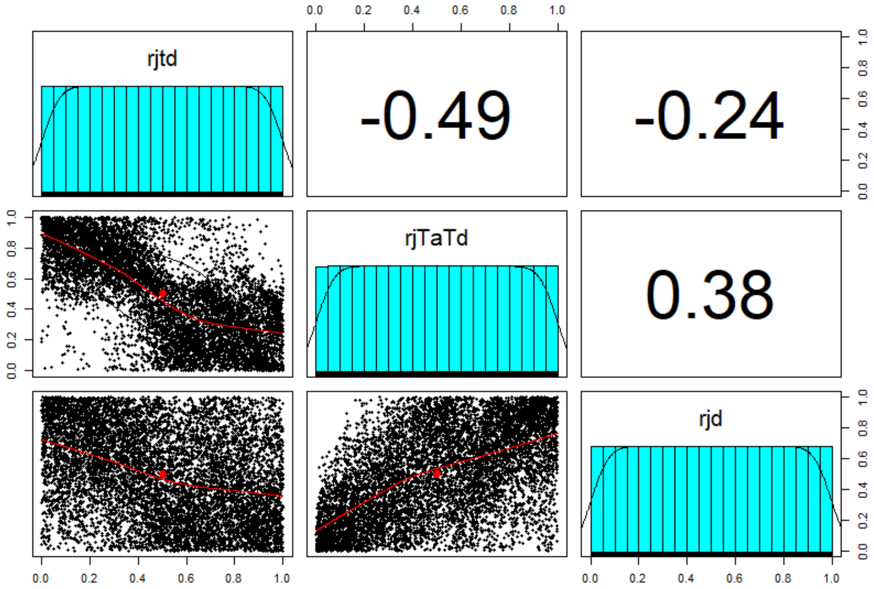

Figure 32: Matrix visualisation of the data set after ranking, jittering, and changing ta into $ta - td$

In comparison to figure 30, the 'white lines' have vanished in the scatter plots, and the histograms are all of equal height. The jittering has thus got rid of the ties in the data set. This jittered data set is the data set which we will use from now on in this thesis. From here on we will refer to the variables of this data set (td, $ta - td$, d) and to pseudo observations (rtd, $rTaTd$, rd)

4.4 Modelling and analysis

The modeling of the data set has been done in the same way as in the previous chapter. The first models we want to fit are again the Normal and Student-t copulas. The results of the fitted Normal and Student-t model is given below:

Model	likelihood	parameters
Normal	3676	ρ_1 = -0.6364, ρ_2 = -0.3296, ρ_3 = 0.5365
Student t	3697.38	ρ_1 = -0.6436, ρ_2 = -0.3367, ρ_3 = 0.5426, ν = 24.4579

The AIC and BIC are given below:

Model	AIC	BIC	#parameters
Normal	-7346.772	-7325.601	3
Student t	-7419.637	-7391.409	4

We can observe that the Student- t model is a better model for the data than the Normal model. The likelihood ratio test results in a p-value of $2.2 * 10^{-16}$ which gives us that we can reject the null hypothesis H_0 and conclude that the Student-t model is a significantly better model for the data than the Normal model. To show how well the Student-t model fits our data, the $(rtd, rTaTd)$ scatter plot of the fitted Student-t copula model and the data set have been scatter plotted below. Here we can see that the Student-t model does not fit the data set well.

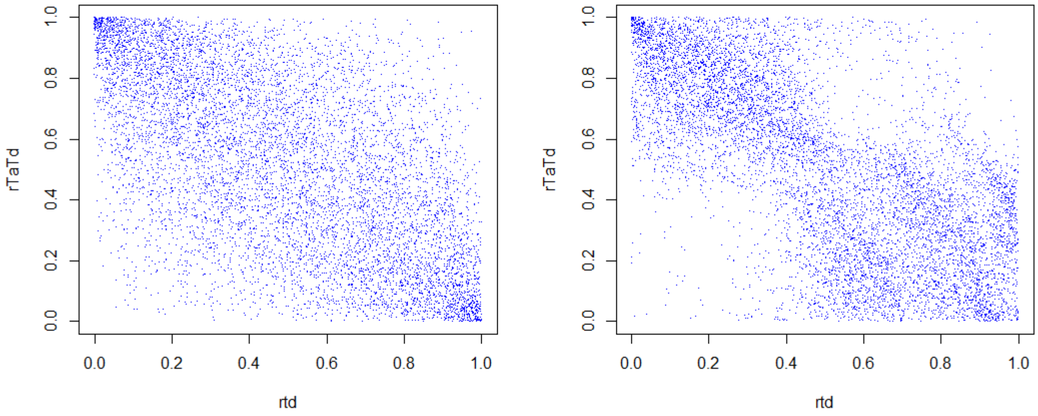

Figure 33: Scatter plot of the Student-t copula model on the left, and the corresponding original data scatter plot on the right.

The fitting of the regular vine copula models has been performed with the RVineCopSelect() function in R. The copula structure of the regular vine models together with copula families and estimated parameters are shown in the figure below:

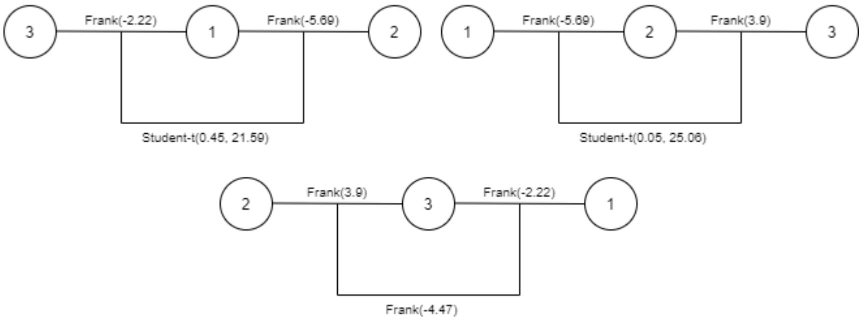

Figure 34: From left to right: Rvine1, Rvine2 and Rvine3

The likelihood, AIC, BIC and the amount of parameters are given in the table below for all the models:

Model	Likelihood	AIC	BIC	#parameters
Rvine1	4174.25	-8340.49	-8312.26	4
Rvine2	4183.69	-8359.38	-8331.15	4
Rvine3	3906.15	-7806.31	-7785.14	3

All three models have likelihood, AIC and BIC values which are close to each other. Rvine2 has the highest likelihood and lowest AIC and BIC values, so this suggests for now that Rvine2 is the best model for our data. To see if Rvine2 is significantly better than Rvine 1 and Rvine 3 we perform the Vuong test.

Model 1 - Model 2	Vuong statistic	p-value
Rvine1 - Rvine2	-0.6518112	0.514523
Rvine1 - Rvine3	13.03082	$8.172939 * 10^{-39}$
Rvine2 - Rvine3	15.82511	$2.088253 * 10^{-56}$

From the table above we conclude that Rvine1 and Rvine2 are of the same quality and are both better than Rvine3.

The goodness of fit test gives us the following result:

Model	Test statistic	p-value
Rvine1	23.74253	0.00831392
Rvine2	20.40614	0.02563704
Rvine3	337.6462	$6.918913 * 10^{-70}$

In all three cases, we can reject the null hypothesis H_0 and conclude that all three models do not fit our data well.

Moreover we performed the Constant Conditional Correlation test. The results are shown below:

Model	CCC statistic	p-value
Rvine1	0.2265649	0.6340827
Rvine2	2.181629	0.1396666
Rvine3	2.121392	0.1452545

All the three p-values are above significance level of ($\alpha = 0.05$), so we can conclude that they are all simplified vines.

From the analysis above we can conclude that Rvine2 model is the best model for the data, but according to the Vuong test it is not significantly better than Rvine1. The goodness of fit test gives us that both Rvine1 and Rvine2 do not fit the data well. Below we compare the scatter plots of the bivariate margin rtd and $rTaTd$ from simulated pseudo observations from Rvine2 with the original data set.

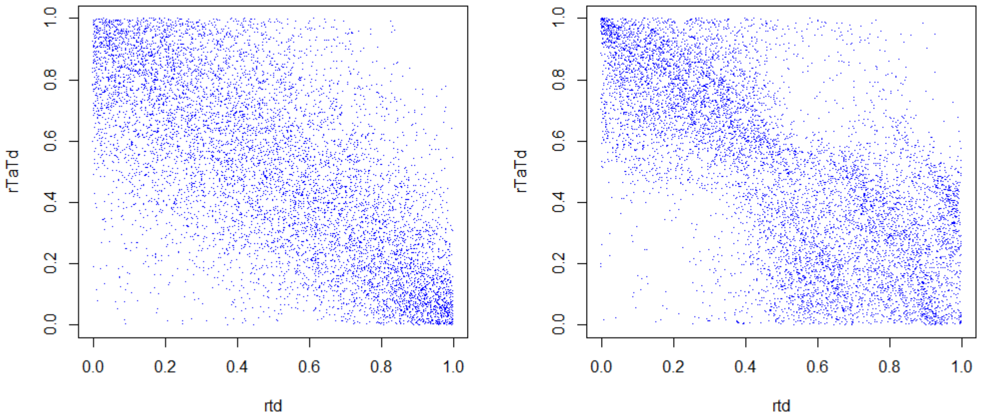

Figure 35: Scatter plot of Rvine2 model on the left, and the corresponding original data scatter plots on the right.

When looking at the scatter plots above, we could see that our model doesn't fit the data well, which we also concluded from the goodness of fit test for Rvine2.

If we look back at the matrix visualization in figure 29, we could see that this data set might be consist out of different groups. If we can model these different groups separately, we could possibly get a better model for the data set. To separate different groups from the data set, we use data clustering techniques which will be explained and applied in the next chapters.

5 Data clustering

The models we fitted to our data in the previous chapter were not up to our satisfaction. This could be because there are different types of 'groups' in the data set. In figure 29 we already had seen and discussed the possibility of the existence of two different 'families' in our data set. In this figure we can observe that the marginal distribution of $ta - td$ is bimodal (a distribution with two different modes). Hence at least two groups could be in our data. One group is away from home for a short period of time and the other group is away for a longer period of time (8-10 hours). We could also observe that there are people who are away from home for a long time but did not travel a long distance. Such distinct 'groups' within our data set portray different behaviour and might make the data difficult to model. Hence in this chapter we will try to separate these groups using clustering methods and build mixture model which we hope will be a better fit for our data set.

5.1 Distance measure

Clustering algorithms divide data sets in groups such that data points in one group are close to each other and are far away to points in other groups. To measure how 'close' a data point is to another data point, distance measures are used. There are multiple distances too choose from, but we will look at two distance measures. First we will use the Euclidean distance, which is defined as:

$$d_{euc}(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}, \tag{17}$$

Where x and y are two groups of data points and x_i and y_i are two data points.

The second distance measure we will consider is Kendall's correlation distance. Kendall's correlation distance measures correspondence of the ranked variables x and y. Kendall's correlation distance is rank based, and thus non parametric. The Kendall's correlation distance is defined as follows [8]:

$$K(x, y) = |\{(i, j) : i < j, (x(i) < x(j) \wedge y(i) > y(j)) \vee (x(i) > x(j) \wedge y(i) < y(j))\}| \tag{18}$$

where x(i) and y(i) are the rankings of element i in the list x and y respectively. An example is given below, where x and y are ranked lists:

List	A	B	C	D
x	1	4	3	2
y	4	2	1	3

The Kendall's distance is calculated as follows:

Pair	x	y	Count
(A,B)	$1 < 4$	$4 > 2$	Yes
(A,C)	$1 < 3$	$4 > 1$	Yes
(A,D)	$1 < 2$	$4 > 3$	Yes
(B,C)	$4 > 3$	$2 > 1$	No
(B,D)	$4 > 2$	$2 < 3$	Yes
(C,D)	$3 > 2$	$1 < 3$	Yes

This gives us that the Kendall's distance is equal to 5, since we have five pairs who compile with equation 18.

In this chapter we will look at two methods for data clustering: k-means clustering and hierarchical clustering, both will be explained later on in this chapter. We will first use the Euclidean distance for the k-means method, and after that we will do the k-means method again but then with the Kendall's distance measure and compare them. After having used both distances for the k-means method, we will decide which distance works the best for our clusters. We will work with this distance measures for the hierarchical method.

5.2 Scaling our data

The clustering methods will be performed on the data set which is visualized in figure 29. In this data set we have variables td and $ta - td$ measured in hours and variable d measured in kilometers. Hence td and $ta - td$ are between 0 and 24, and d has values between 0 and 110. When a certain variable in the data set, in our case d, has much higher values than the other 2 variables, the clustering algorithm might be affected as the variable with larger values might influence the value of distance much more. To avoid this we scale our data set with linear transformation using the scale() function in R.

5.3 K-means clustering

The k-means clustering method partitions the data set into k clusters. This method classifies data points in k different clusters, such that the data points within the same cluster are as similar as possible to each other and data points from another cluster are as dissimilar as possible to this data point. Every cluster has its centre, which is the mean of the points assigned to the cluster. The method divides all the data points of the data set into the different clusters such that the total sum of squares from the data points to the centre of their respective cluster is minimized, which is the same as minimizing the following: [22] [23]:

$$W(C_k) = \sum_{x_i \in C_k} (x_i - \mu_k)^2 \tag{19}$$

Where x_i is the data point belonging to cluster C_k and μ_k being the mean value of the points assigned to cluster C_k.

Equation 19 measures the goodness of a single cluster. The following equation computes the total within cluster variation, which measures the goodness of the total amount of clusters. The k-means method performs well if the total within cluster variation is small. The total within cluster variation is defined as follows:

$$\sum_{k=1}^{k} W(C_k) = \sum_{k=1}^{k} \sum_{x_i \in C_k} (x_i - \mu_k)^2 \tag{20}$$

The algorithm of the k-means can be done in R with the kmeans(). The algorithm of the k-means method is as follows:

1) Choose the amount of clusters k.

2) Choose k random data points as the centre of the starting k clusters.

3) Compute the sum of the squared distances between all the data points which are not yet inside a cluster and the centres of the clusters.

4) Assign the data point which is the closest to a specific cluster to that cluster (the cluster for which distance between this point and the center of this cluster is the smallest).

5) Compute the new centre of each cluster as the average of all points assigned to this cluster.

6) Repeat step 3, 4 and 5 until the centre of clusters do not change anymore and thus every data point is assigned to a cluster.

As explained in the beginning of this chapter, we suggest that there are at least 2 different groups in our data set. Therefore we will first cluster our data set into 2 clusters. In R there is a function which shows you what the 'optimal' amount of clusters for your data set is based on the total within sum of squares. This function calculates the total within sum of squares of the data set as a function of the number of clusters k, where k ranges from 1 to 10. The results for our data set are plotted below:

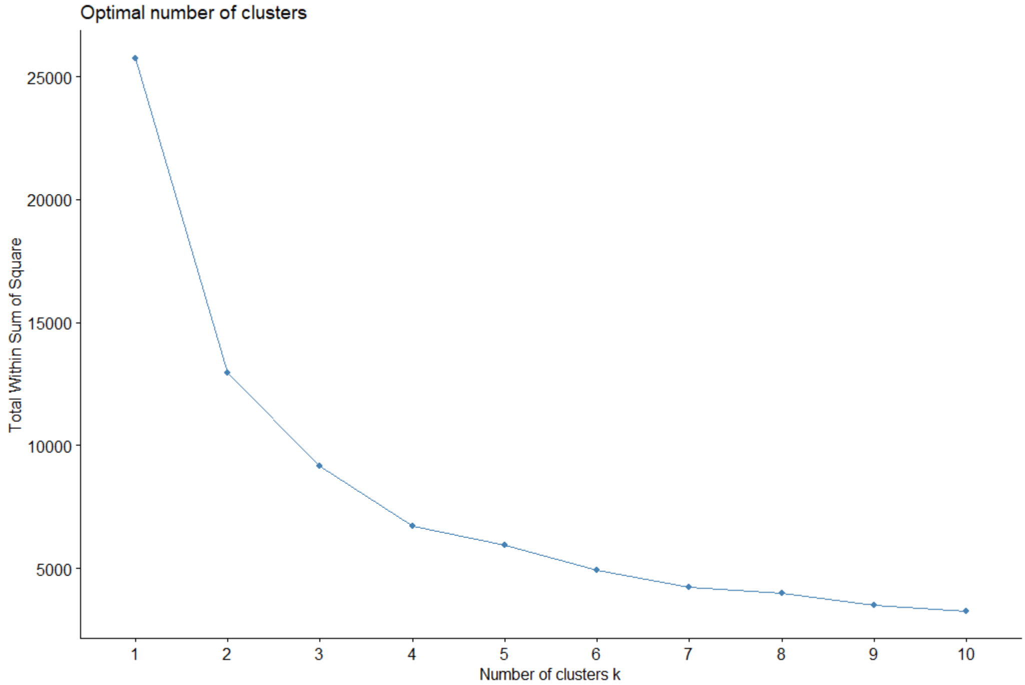

Figure 36: Graph of the within sum of squares of the cluster against the amount of clusters for our data set.

When the number of clusters is equal to 4 we see an 'elbow' shape in the figure. This indicates that from $k = 4$ onwards the total within sum of squares will not become much smaller if we have more clusters. Therefore we conclude that, based on the within sum of squares, our optimal amount of clusters is 4. However we will also explore results for 2 and 3 clusters for comparison

5.3.1 Visualizations

In the following of this chapter (so also for the hierarchical clustering), we will look at different choices of amount of clusters. We will visualize scatter plots, histograms and silhouette plots of the data for each cluster method and amount of clusters. The clusters in each figure will be represented by their color, where the cluster number and respective color is given in the table below:

Cluster	Color
Cluster 1	Red
Cluster 2	Blue
Cluster 3	Yellow
Cluster 4	Green
Cluster 5	Black

5.3.2 2 Clusters

Clustering with two clusters gives us that our data set is separated into two data sets. Below we see how many data points are in each cluster.

Data set	Data points in cluster
Cluster 1	4313
Cluster 2	4266

In the figure below we see the scatter plots of the data set with the clusters and its marginal distributions.

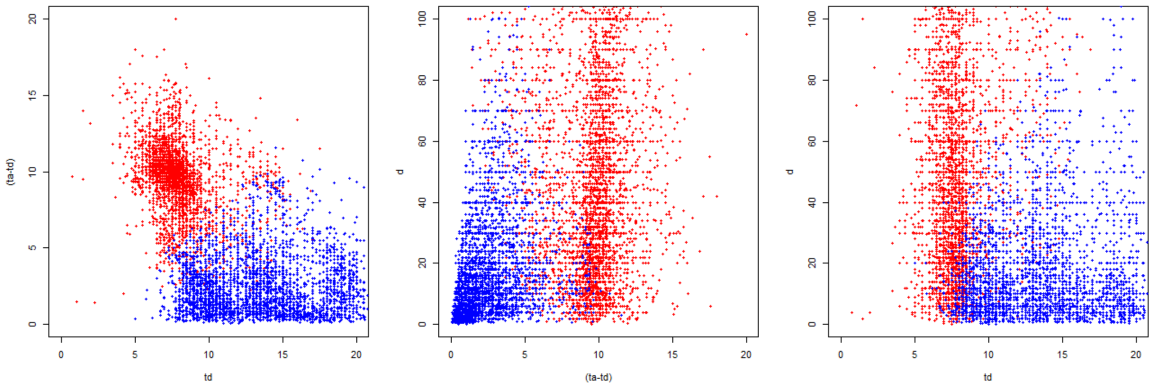

Figure 37: Scatter plot of the data set with 2 clusters

We will also plot the marginal distributions of each cluster. We will use the same colors for the clusters as in the scatter plot.

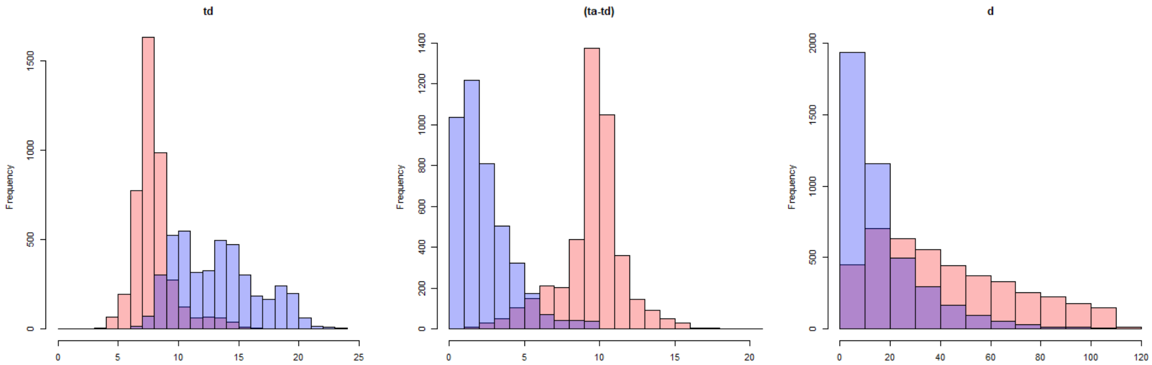

Figure 38: Marginal distributions of the data set with 2 clusters

We observe that the first cluster contains people who leave home relatively early, stay away for quite a long time and travel different distances, they could correspond to the working population. The second cluster is built out of data points where distance traveled is relatively short and $ta - td$ is short. Hence this cluster corresponds to people who go for short trips in different times of the day. These people are probably not working or part time working people.

To see how well the clusters perform, we will look at silhouette plots for every cluster choice. A silhouette plot shows how similar data points of a specific cluster are to its own cluster and how dissimilar this data point is compared to the other clusters. This can indicate the goodness of each cluster. The silhouette width is defined as [24]:

$$s(i) = \left\{ \begin{array}{ll} 1 - a(i)/b(i), & \text{if } a(i) < b(i) \\ 0, & \text{if } a(i) = b(i) \\ b(i)/a(i) - 1, & \text{if } a(i) > b(i) \end{array} \right\}$$

with

$$a(i) = \frac{1}{|C_i| - 1} \sum_{j \in C_i, i \neq j} d(i,j)$$

$$b(i) = \min_{k \neq i} \frac{1}{|C_i|} \sum_{j \in C_k} d(i,j)$$

where i is a data point in the cluster C_i, $|C_i|$ is the amount of data points of the cluster, and d the Euclidean distance.

The silhouette width is between -1 and 1, where values close to 1 indicate that the data point is similar to its cluster, and a value closer to -1 means it is not similar to its own cluster. In a silhouette plot, we see the clusters and their respective silhouette width. We also see the total silhouette width, which is the average silhouette value from all the data points. The higher the average silhouette of a cluster the better the cluster and the higher the overall average silhouette value, the better the total clustering. Below we find the silhouette plot of the k-means method with two clusters.

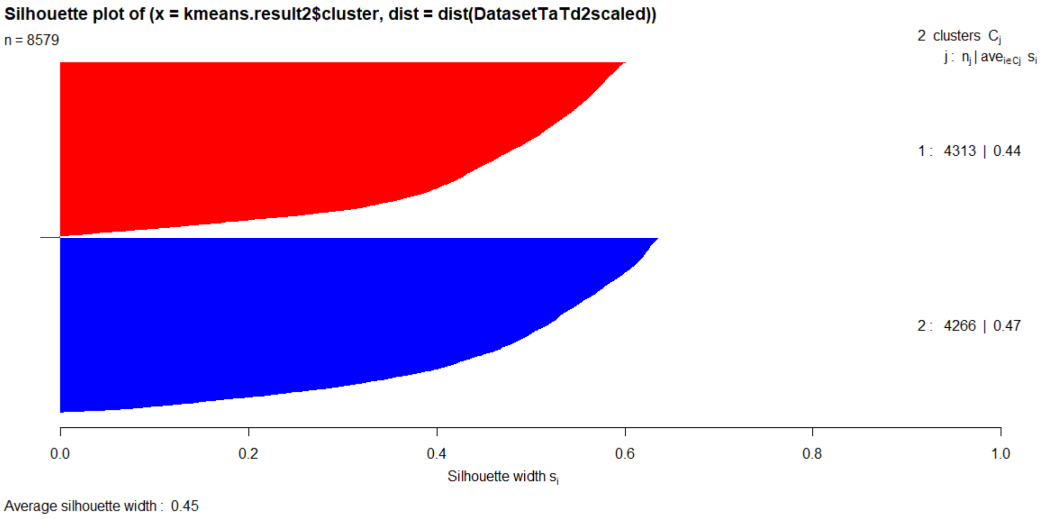

Figure 39: Silhouette plot for 2 clusters with the k-means method.

We see that both clusters have an average silhouette width of around 0.45, and therefore so is our overall silhouette width. A silhouette width of 0.45 gives us that we have a reasonable clustering.

5.3.3 3 Clusters

When we cluster the data with three clusters, we get three data sets with 1672, 4060 and 2847 data points in cluster 1, 2 and 3 respectively. In the figure below we see the scatter plots of the data set with the clusters and its marginal distributions.

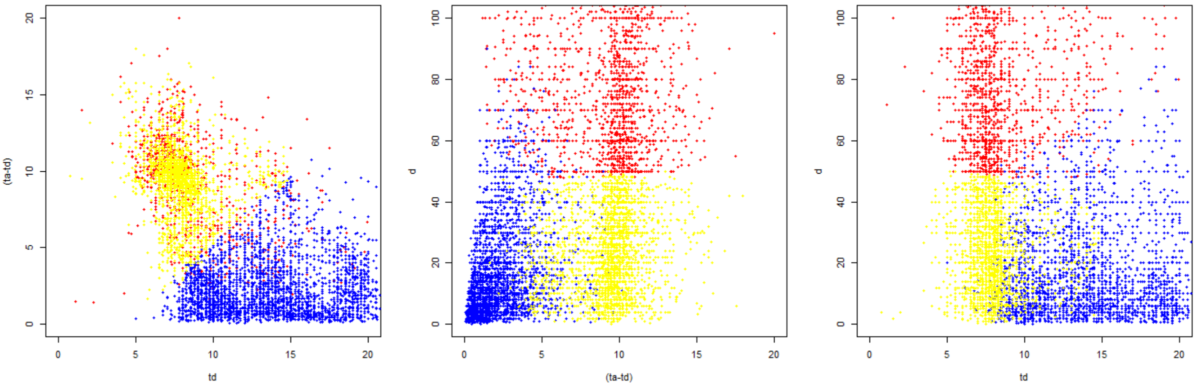

Figure 40: Scatter plot of the data set with 3 clusters

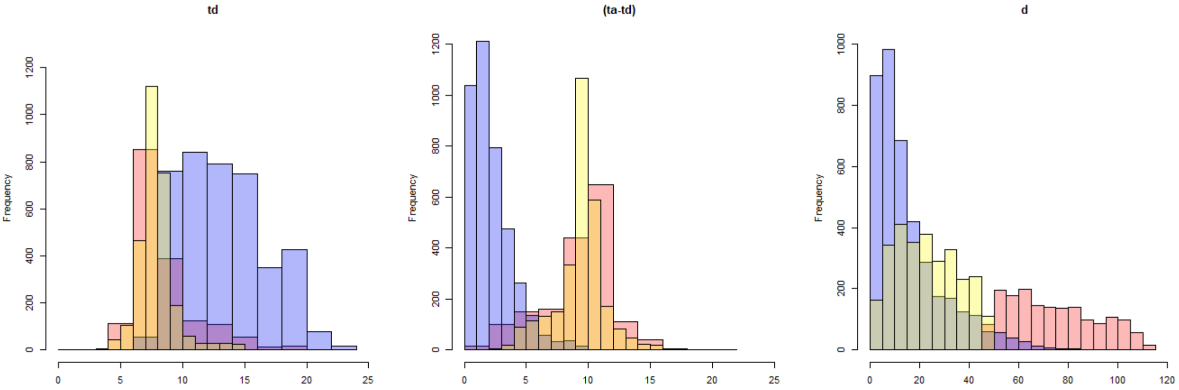

Figure 41: Marginal distribution of the data set with 3 clusters

Cluster 1 now contains people who traveled relatively long distances, departed early and were away from home for most likely 8-10 hours. Cluster 2 contains people who take short trips in different times of the day and cluster 3 contains people who were away for around 10 hours, left early and traveled relatively short distances. Below we show the silhouette plot of the k-means method with three clusters.

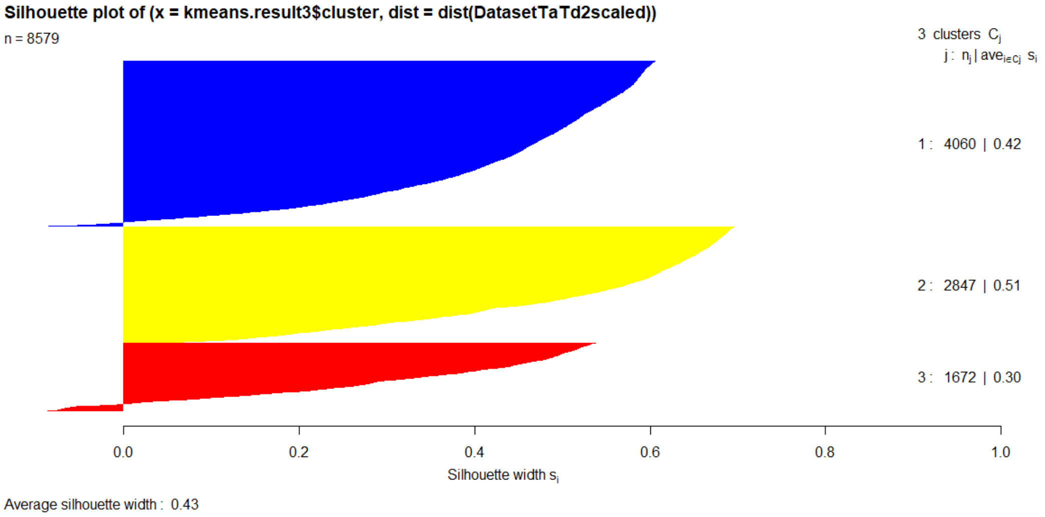

Figure 42: Silhouette plot for 3 clusters with the k-means method.

We see that the first cluster has an average silhouette value of 0.3, the second cluster an average silhouette width of around 0.4 and the third cluster an average silhouette width of around 0.5. This gives us that the third cluster is a good cluster, and the second and first clusters are not as good as the third cluster, but they are reasonable clusters. The total overall silhouette width is 0.43, which is close to the total silhouette width of k-means with 2 clusters. Therefore the silhouette plot does not give a clear answer whether using 3 clusters is worse than 2 clusters.

5.3.4 4 clusters

When we cluster the data with four clusters, we get four data sets with 2336, 1944, 2721 and 1578 data points in cluster 1, 2, 3 and 4 respectively. In the figure below we see the scatter plots of the data set with the clusters and its marginal distributions.

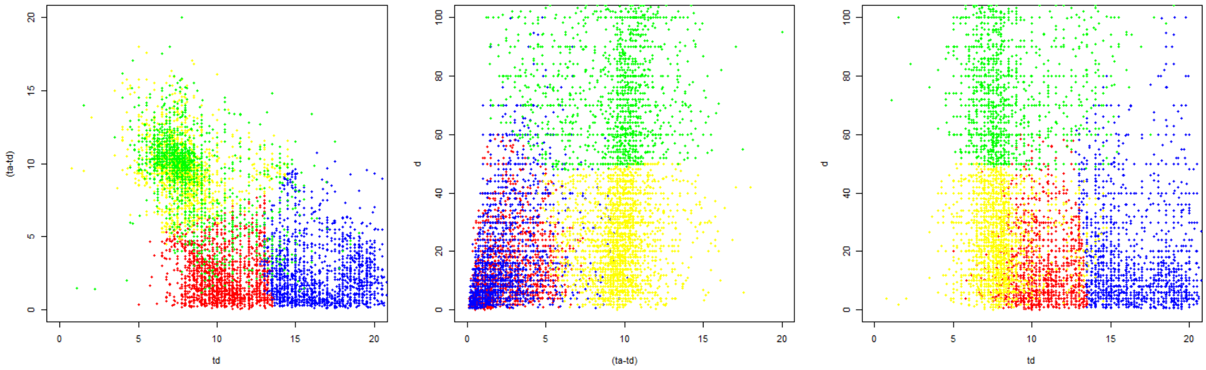

Figure 43: Scatter plots of the data set with 4 clusters

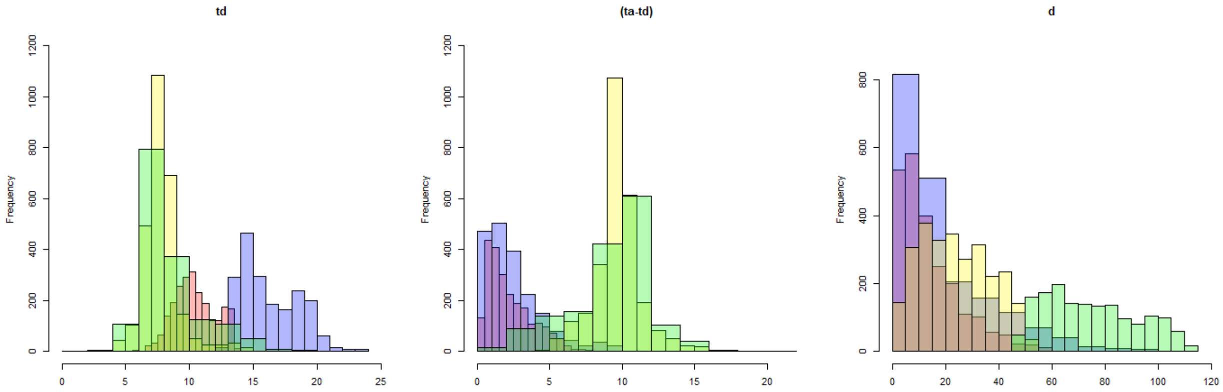

Figure 44: Marginal distribution of the data set with 4 clusters

We can summarize the obtained results as follows:

Cluster 1 contains people who take short trips, departure between 10 and 14, and are therefore probably not working but are going shopping or something similar.

Cluster 2 contains people who take short trips and depart after 14, so they are people who are also probably not working.

Cluster 3 contains people who departure early, are away from home for around 10 hours and travel short distances. They are probably working people who work close to home.

Cluster 4 contains people who depart early, are away from home for around 10 hours and drive long distances. They are probably working people who work far away from home.

Below we find the silhouette plot of the k-means method with 4 clusters.

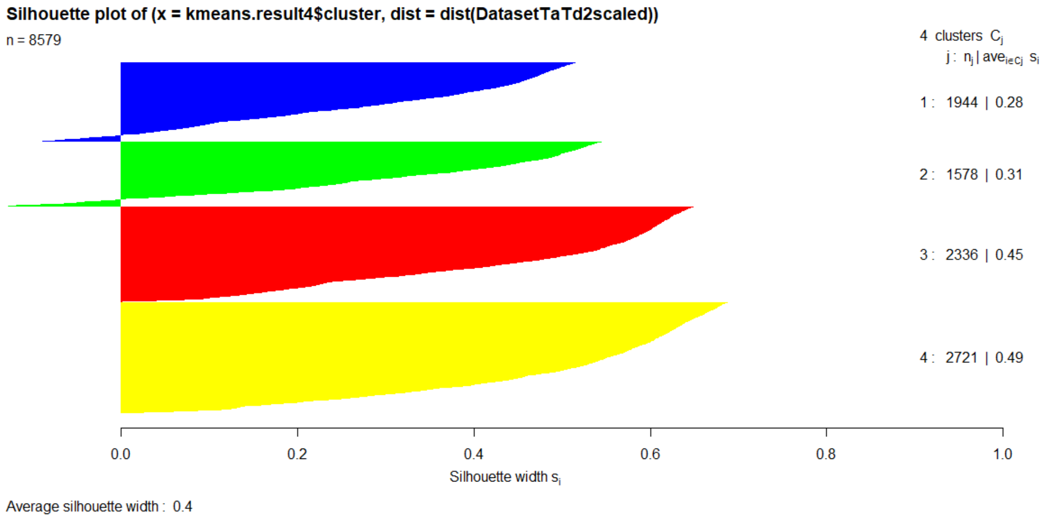

Figure 45: Silhouette plot for 4 clusters with the k-means method.

The results of the silhouette plot indicate that the clusters are reasonably good.

We were interested in if 5 or 6 clusters gave us a better clustering based on the silhouette width. Their silhouette plots can be found in the appendix at A.7 and A.8. The silhouette width of the 5 and 6 clusters is smaller then the silhouette width for 2, 3 and 4 clusters and therefore we prefer to use 2, 3 or 4 clusters based on the silhouette width.

We have analyzed silhouette plots for 2, 3, 4, 5 and 6 clusters. As seen in figure 36 we found that, based on the total within sum of squares, 4 clusters will be the optimal amount of clusters. Below we find the same figure but then based on the silhouette width, where the optimal amount of clusters is given:

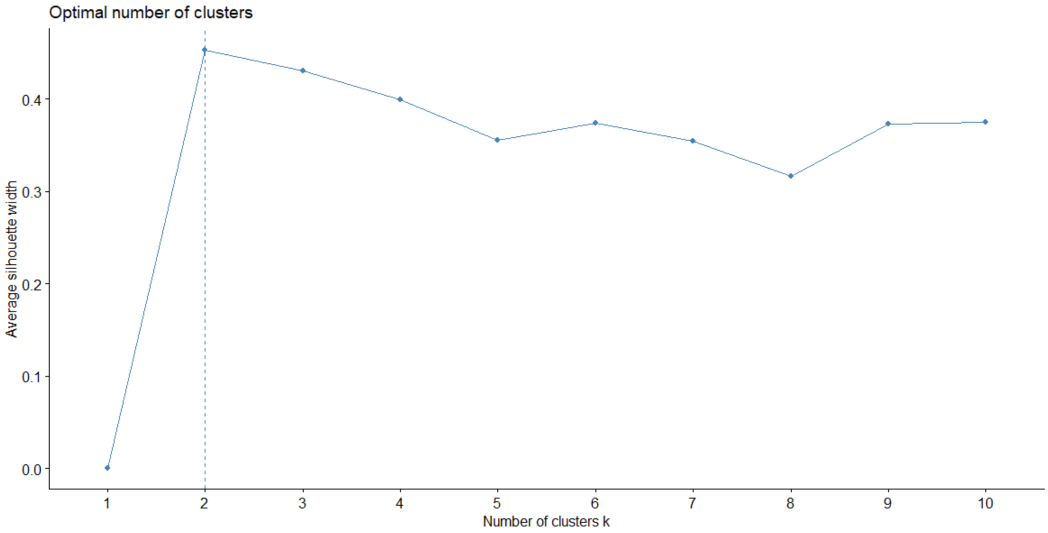

Figure 46: Average silhouette plot with on the y axis the total average silhouette width and on the x axis the chosen amount of clusters

We see that, based on the silhouette width, 2 clusters is the optimal amount of clusters.

5.3.5 Kendall's distance measure

In the cluster analysis performed above, we have used the Euclidean distance measure. We also want to know how well our clustering method performs if we use another distance measure. Therefore we will do the same k-means method for 2, 3 and 4 clusters and use Kendall's distance measure. For clustering data with the Kendall correlation distance we can not use the same kmeans() function in R which we used for the Euclidean distance, but we have to use the Kmeans() method from the amap package in R. The amount of data points in each cluster is given below, where we looked at the amount of clusters equal to 2, 3 and 4:

Data points in cluster	2 Clusters	3 Clusters	4 Clusters
Cluster 1	4222	1847	1847
Cluster 2	4357	3737	1339
Cluster 3	-	2995	3018
Cluster 4	-	-	2375

We see that the distribution in amount of data points in each cluster for 2, 3 and 4 clusters is similar to the distribution of data points in clusters with these amount of clusters for the Euclidean distance. Therefore we want to look at the scatter plots to see if there are big differences. For all the three amount of clusters, they are given below:

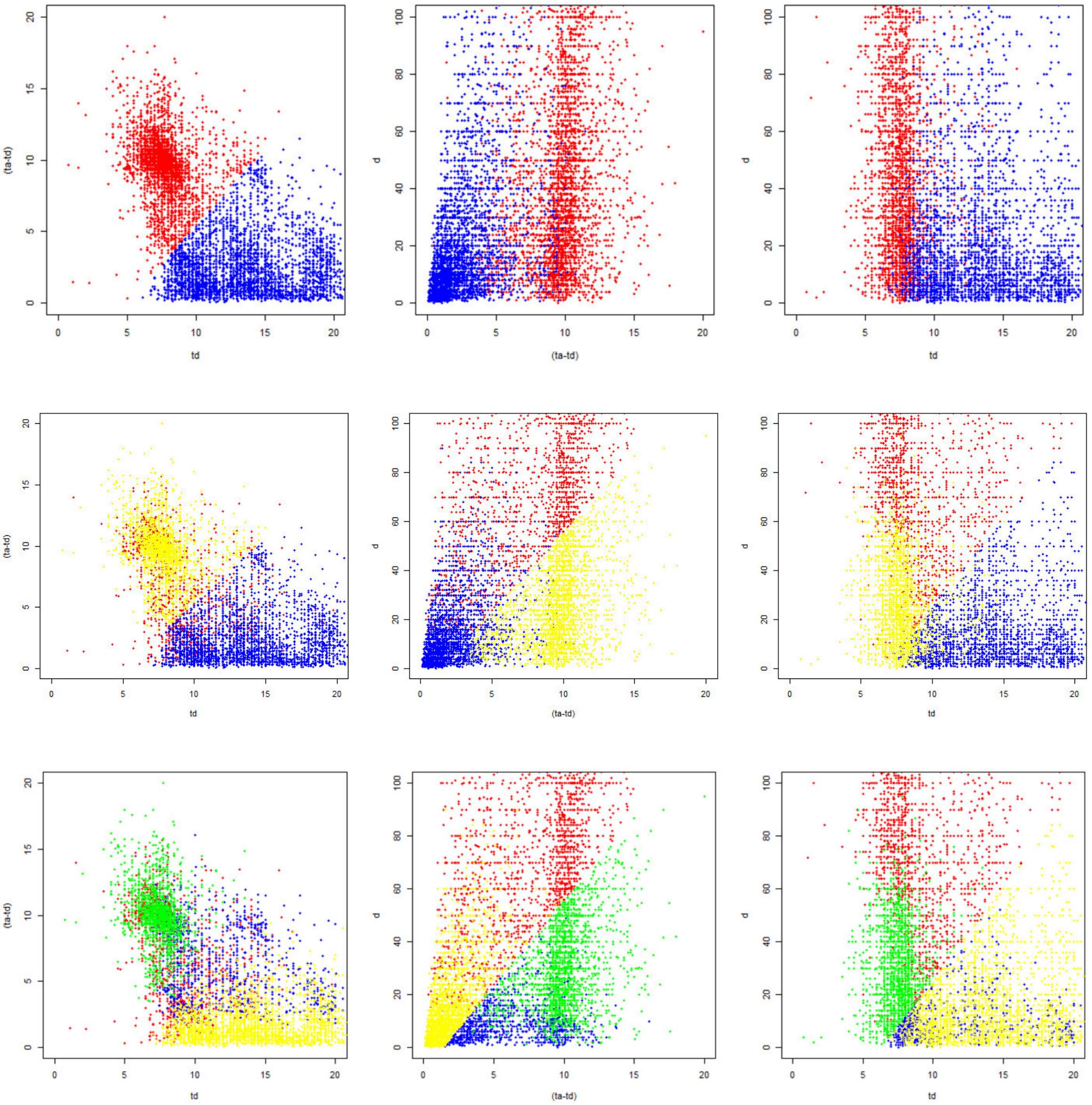

Figure 47: Scatter plots of the clusters with 2 clusters above, 3 clusters in the middle and 4 clusters below.

We see different behaviour in the scatter plots with 3 and 4 clusters compared to the scatter plots of the Euclidean distance. We see a diagonal line in the division between the scatter plots for the $(ta - td, d)$ and (td, d) scatter plot. Due to this diagonal line, there is no clear division in traveled distance for the full time working people as we have seen for the Euclidean distance. For further comparison, the marginal distributions of the 3 and 4 clusters case can be found in the appendix A.12. In the scatter plots we also see that the not full time working people clusters, especially in the four clusters case, overlap with the full time working people clusters. This is illustrated in the 4 clusters scatter plot at the $(ta - td, d)$ and (td, d) figure, were we see that the full time working people who don't travel far are in the same place in the scatter plot of the not full time working people (blue and green clusters). We can see that the change of distance measure has given different behaviour of the clusters. To see how different the clusters from both distance measures are, we will look at the Jaccard index of the clusters.

The Jaccard index measures how similar clusters are to each other. The Jaccard index for two vectors X and Y can be computed by [36] [37]:

$$J(X,Y) = \frac{|X \cap Y|}{|X| + |Y| - |X \cap Y|} \tag{21}$$

With J the Jaccard index and |.| defining the length of the vector. We can see that J is always between 0 and 1, with an index of 1 giving that the vectors are very similar and 0 giving that the vectors are not similar.

Since our data set is jittered, it contains no ties. Therefore, we can check our clusters by only looking at data for one variable. Therefore we use the *td* vector of the clusters for computing the Jaccard indexes. The results for 2, 3 and 4 clusters are given below. CKi and CKKi correspond to the i-th cluster of the k-means method for the Euclidean and Kendall's correlation distance respectively. We will start with the 2 cluster case.

J(,)	CK1	CK2
CKK1	0.01762379	0.9136771
CKK2	0.9145204	0.02822581

We can see that the first cluster for the Euclidean distance corresponds with the second cluster for the Kendall's correlation distance, and the second cluster for the Euclidean distance corresponds with the first cluster for the Kendall's correlation distance. These two Jaccard indexes are around 0.91, which gives us that the clusters are almost similar to each other. The results for three clusters are given below:

J(,)	CK1	CK2	CK3
CKK1	0.8353754	0.07361399	0.00227305
CKK2	0.006566997	0.04724683	0.9221283
CKK3	0.0358505	0.5937328	0.001136148

We can see that CK1 corresponds to CKK1, CK2 corresponds to CKK3 and CK3 corresponds to CKK2. The Jaccard index for (CK1, CKK1) and (CK3, CKK2) are really high, so these clusters pairs are almost similar to each other. The (CK2,CKK3) Jaccard index is around 0.6, which gives that they are averagely similar. The results for 4 clusters is given below.

J(,)	CK1	CK2	CK3	CK4
CKK1	0.002061147	0.1159075	0.1590066	0.1342593
CKK2	0.04844221	0.402865	0.0001572822	0.4288448
CKK3	0.6232639	0	0.02306297	0.01741862
CKK4	0.06866721	0	0.6863005	0.01007719

We can see that CK1 corresponds to CKK3, CK2 corresponds to CKK2, CK3 corresponds to CKK4 and CK4 corresponds to CKK2 as well. These results are interesting, because we have two clusters from the Euclidean distance corresponding to one cluster of the Kendall's distance. We conclude that the clusters from both distances for the 4 cluster case are not similar to each other.

For the Kendall's distance we see a diagonal line in the scatter plots, which we don't see with the Euclidean distance, where there is a more clear distinction between the clusters based on the travelled amount of distance. For the Euclidean distance, there is also a more clear division between the clusters for the full time working people and the not full time working people. If we take this into account, together with the fact that for both distance measures (based on the Jaccard index) we get different clusters, we conclude that we prefer to use the Euclidean distance. Therefore we will use the Euclidean distance from now on in this thesis.

5.3.6 Conclusion of the k-means method

We have found that 2, 3 and 4 clusters with the k-means method and the Euclidean distance gave us a good clustering, based on their silhouette width, scatter plots and histogram plots. Figure 36 gave us that, based on the total within sum of squares of all the clusters, that 4 clusters would be optimal for our data set. If we interpret all the silhouette widths with paper [24], we get that for 2, 3 and 4 clusters we have reasonable good clusters. Therefore we conclude that 2, 3 and 4 clusters with the Euclidean measure we get good a clustering of our data set.

5.4 Hierarchical clustering

The second data clustering method we will use is the hierarchical method. The hierarchical clustering builds an 'hierarchy' from the bottom up. The algorithm is as follows:

1) Assign every data point in the data set to its own unique cluster.

2) Combine the two closest clusters into one new cluster.

3) Repeat step 2 until we end up with one cluster.

The algorithm gives a dendogram as output where we can see all the steps taken in the algorithm. The hierarchical method can use different ways to determine how to measure the distance between clusters and decide which one is the closest. These different methods for measuring all use the Euclidean distance. We will use two different ways of determining how close two clusters are to each other: complete linkage and mean linkage. In complete linkage the distance between two clusters is measured by evaluating the distances between all the data points from two clusters, and the maximal distance of all these distances is chosen as the distance between the two clusters. With the mean linkage we measure the distance between two clusters the same as for the complete linkage, but instead of taking the maximal distance of all the distances, we take the average distance of all these distances [23].

5.4.1 Complete linkage clustering

When we use the complete linkage clustering method, our dendrogram structure is as follows:

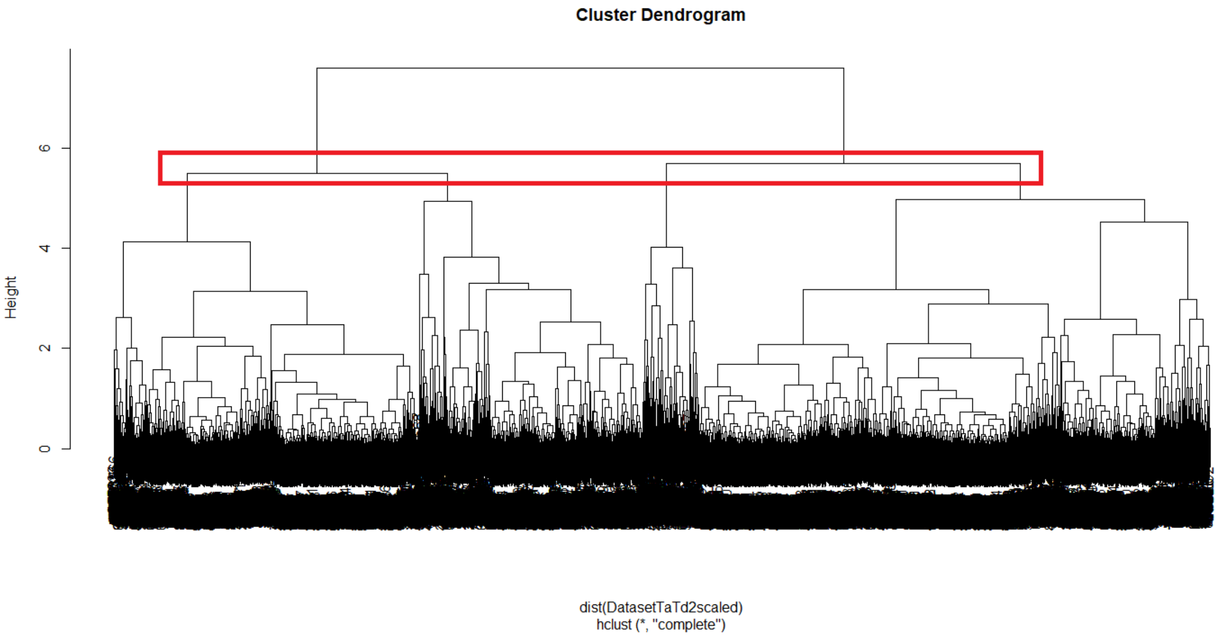

Figure 48: Dendrogram generated by the complete linkage clustering with the optimal amount of clusters encircled in red.

The bottom part of the dendrogram is where all our data points are shown. We have 8579 data points, so they become one black mass as we can see due to them all trying to fit on this stroke. Above this black stroke, the clustering begins. As already explained, the hierarchical method starts clustering from the bottom up. The clusters are represented by the lines drawn from the data points. Every data point is in the beginning assigned to its own cluster, which can be seen as the second black stroke. This is because every data point gets a 'line' representing their cluster, so with 8579 lines it creates a black stroke. The lines represent the chosen clusters, and we can use this to decide what amount of clusters we want to use. This decision on what amount of clusters to use is made visually, so it is subjective. To make this decision, we try to find the height in the dendrogram for which the difference in height between the next clustering (so going down in the dendrogram) is still sufficiently big. As you can read, this procedure of visually checking for the optimal amount of clusters is very subjective. If we do this in our case, visualized by the red encircling, we see that 4 clusters is our optimal amount of clusters. The amount of data points per cluster is given below:

Data set	Data points in cluster
Cluster 1	1751
Cluster 2	2390
Cluster 3	4002
Cluster 4	436

In the figure below we see the scatter plots of the data set with the clusters, where every cluster has its own color as given in 5.3.1.

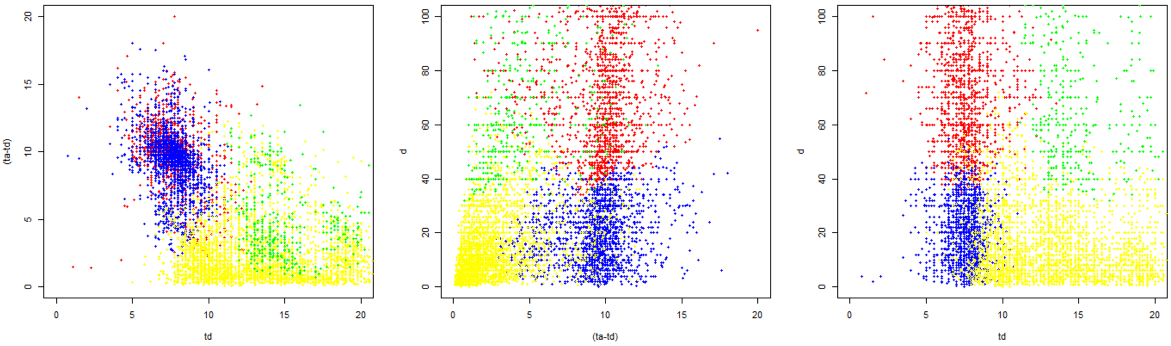

Figure 49: Scatter plot of the data set with 4 clusters from the complete linkage

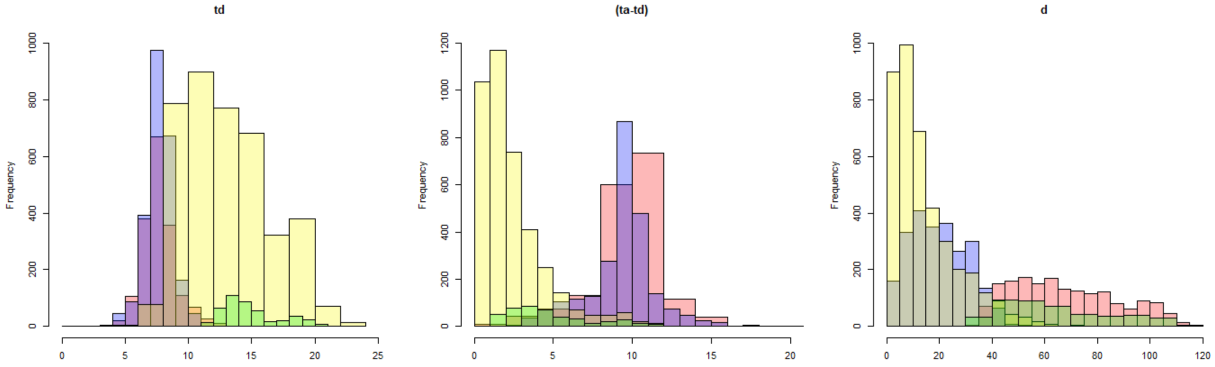

Figure 50: Marginal distribution of the data set with 4 clusters from the complete linkage

If we visually analyze the scatter plots, we see that the first and the second cluster represent the full time working people, since they departure early on the day and are away for a long time. From the second scatter and marginal distribution plot we find that the first cluster represents the full time working people who work far away, and the second cluster represents people who work full time closer to home. The third and fourth cluster represent the not full time working people, where the difference between them is that the third cluster represents the people who do not cover a lot of distance, while the fourth cluster represents the people who cover more distance.

The silhouette widths of the clusters are as follows:

Cluster	Silhouette width
Cluster 1	0.48
Cluster 2	0.33
Cluster 3	0.14
Cluster 4	0.15
Average total	0.26

This gives us that the data points of the first cluster are very similar to each other, but the data points in the third and fourth cluster are not very similar to each other. The average silhouette width of 0.26 is very low compared to the average silhouette widths found for the k-means method.

5.4.2 Mean linkage clustering

When we use the mean linkage clustering method, our dendrogram structure is as follows:

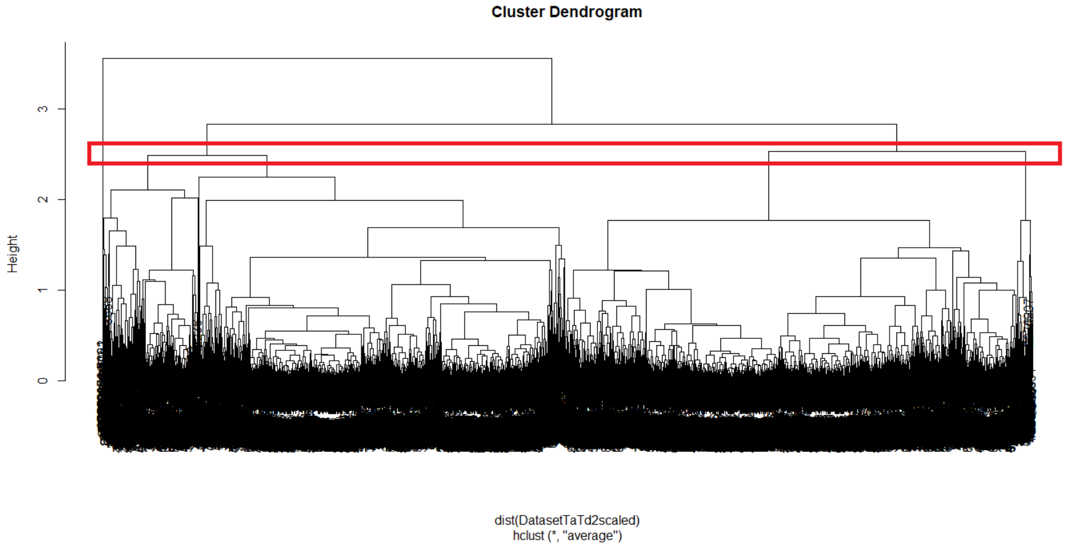

Figure 51: Dendrogram generated by the mean linkage clustering with the optimal amount of clusters in red.

This indicates us that the best choice for the amount of clusters is 5. The amount of data points per cluster is given below:

Data set	Data points in cluster
Cluster 1	3382
Cluster 2	883
Cluster 3	4175
Cluster 4	137
Cluster 5	2

As we can see, this clustering technique has given us a cluster which contains 2 data points. This is not useful for us if we want to model the clusters. Therefore we decide to not further analyze this hierarchical clustering technique and conclude that the complete linkage method gives us a better clustering than the mean linkage method.

5.5 Comparing the k-means and hierarchical method clusters

We will compare the clusters for the k-means method and the clusters of the hierarchical method with the complete linkage. We will compare the case with 2 and 3 clusters. The scatter plots of the 2 and 3 cluster complete linkage method can be found in the appendix A.9 A.10. Let CK1, CK2, CH1, CH2 be clusters 1 and 2 of the k-means method and clusters 1 and 2 of the hierarchical method, respectively. This gives us the following results:

J(,)	CK1	CK2
CH1	0.6551501	0.2259731
CH2	0.03807107	0.5275382

We can see that CK1 corresponds with CH1 and CK2 corresponds with CH2. The two Jaccard indexes are around 0.65 and 0.53 respectively, which gives us that the clusters from the k-means method are different from the clusters of the hierarchical method in the 2 clusters case.

We have done the same for the case with three clusters. The results are given below:

J(,)	CK1	CK2	CK3
CH1	0.4769939	0	0.2287949
CH2	0.02704236	0.2858083	0.3972081
CH3	0.07077229	0.5303792	0.006050605

As we can see, CK1 corresponds with CH1, CK2 corresponds with CH3 and CK3 corresponds with CH2. All the three Jaccard indexes of these clusters are around 0.4 or 0.5, which gives us that the k-means and the hierarchical method give different clusters for 3 clusters as well.

5.6 Conclusion

The k-means and the hierarchical method both create clusters where the working and not working people get separated. The k-means method is a better method for clustering our data set since its clusters are better than the clusters from the hierarchical clustering based on the silhouette width. The Euclidean distance provides a better clustering then the Kendall's correlation distance. Based on the scatter plots, histogram plots and figures 36 and 46, we find that the optimal amount of clusters for our data set is 2, 3 or 4 clusters. We will use all three amount of clusters for the analysis of the data set in the next chapter.

6 Analysis of the 3-dimensional clustered data sets

In this chapter we will model the data set which has been clustered with the k-means method. We have fitted and analyzed the best model for each cluster in the case of 2, 3 and 4 clusters and build mixtures models with them. We will analyze the best mixture model from these three amount of clusters. A mixture model is a model made up of different models. Since we work with clusters, the model for every clusters will be combined into one mixture model. If we want to compare different mixture models, we need to define hoe to compute the likelihood and AIC of a mixture model.

Lets say we model k clusters and get a best model (with a likelihood and AIC value) for each cluster. Then:

$$l_t = \sum_{i=1}^{k} l_i$$

With l_t being the total log likelihood of the mixture model and l_i being the log likelihood of the model of the i-th cluster. The total AIC of the model is then:

$$AIC_t = -2 * l_t + 2 * \left((\#mc - 1) + \sum_{i=1}^{k} \#pi \right)$$

Where AIC_t is the total AIC of the mixture model, #mc the amount of modeled clusters and #pi the amount of parameters of the model for cluster i.

The AIC_t value in the next part of this thesis will be computed by ranking data for each cluster. Strictly speaking the AIC_t's computed on ranked data for different clusters should not be compared. This comparison would have been correct if the margins of each cluster were modeled parametrically.

We have fitted and compared models for 2, 3 and 4 clusters. Only looking at the likelihoods of every model of each cluster, we decided that the 3 clusters case gave us the best modelling. Since there is no clear test for comparing mixture models with each other and deciding which one is better, we could have also chosen for 2 or 4 clusters as our best amount of clusters. Therefore only the results for 3 clusters are presented in this chapter. The results for 2 and 4 clusters can be found in appendix A.13 and A.14.

The best model for all three clusters were regular vine copula models. The best possible model for each cluster is given below:

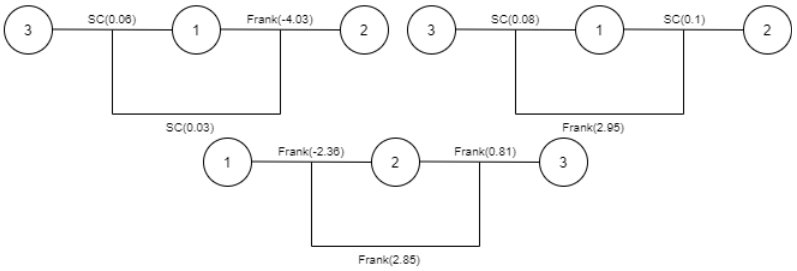

Figure 52: From left to right: best copula model for cluster 1, 2 and 3.

The table below show us the results for each individual model:

Model	Likelihood	AIC	#parameters
Cluster 1	300.4	-594.8	3
Cluster 2	476.15	-946.31	3
Cluster 3	217.24	-428.49	3

As we can see, the likelihoods are very low. If we look at the fitted models, we see that the copulas are getting close to being independent. This could explain the low likelihood values.

The goodness of fit test for each cluster model gave us the following results:

Model	Test statistic	p-value
Cluster 1	31.17661	$2.345499 * 10^{-5}$
Cluster 2	10.49074	0.1054495
Cluster 3	55.11865	$4.386669 * 10^{-10}$

This gives us that the model for cluster 2 fits the data well, but the other 2 models do not fit the data well. Next to the goodness of fit test, we want to compare the scatter plots of the simulated models and their respective cluster. Below we find the three comparisons between the simulated models per cluster and the respective cluster.

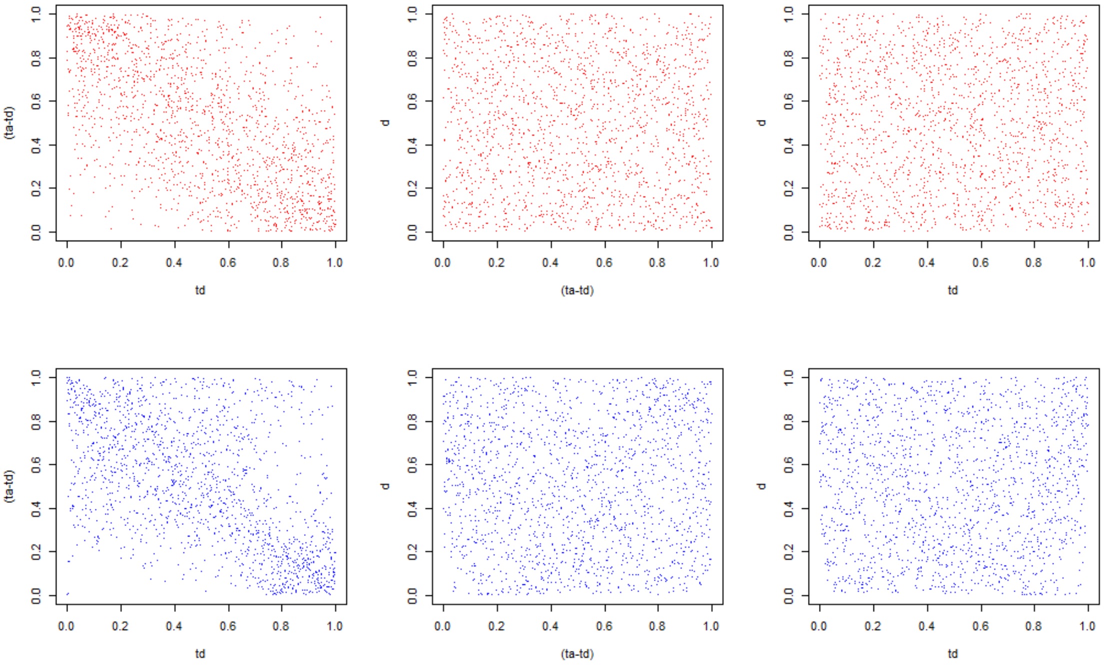

Figure 53: Above: scatter plots of the simulated model for cluster 1. Below: the scatter plots of cluster 1.

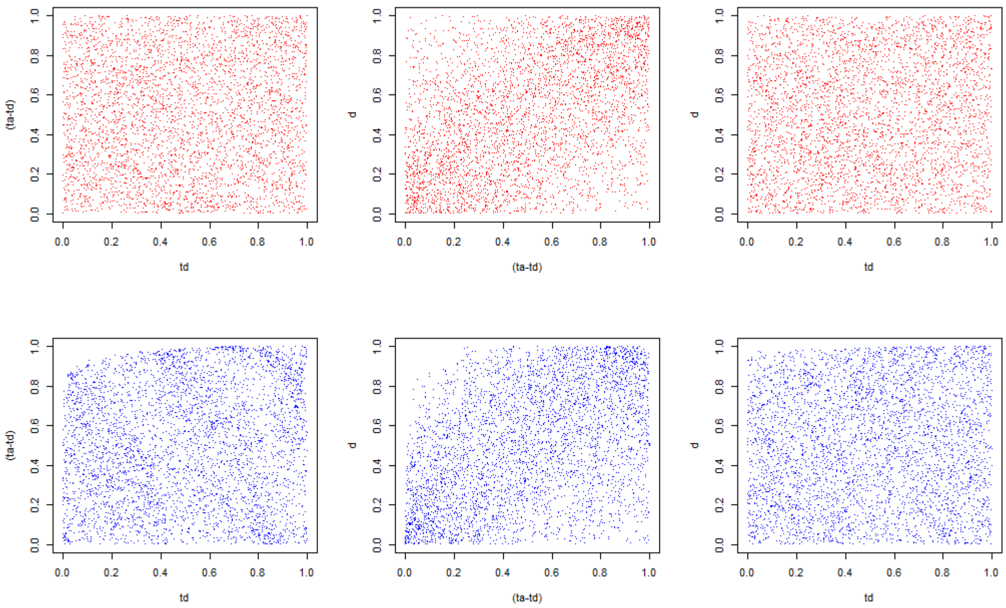

Figure 54: Above: scatter plots of the simulated model for cluster 2. Below: the scatter plots of cluster 2.

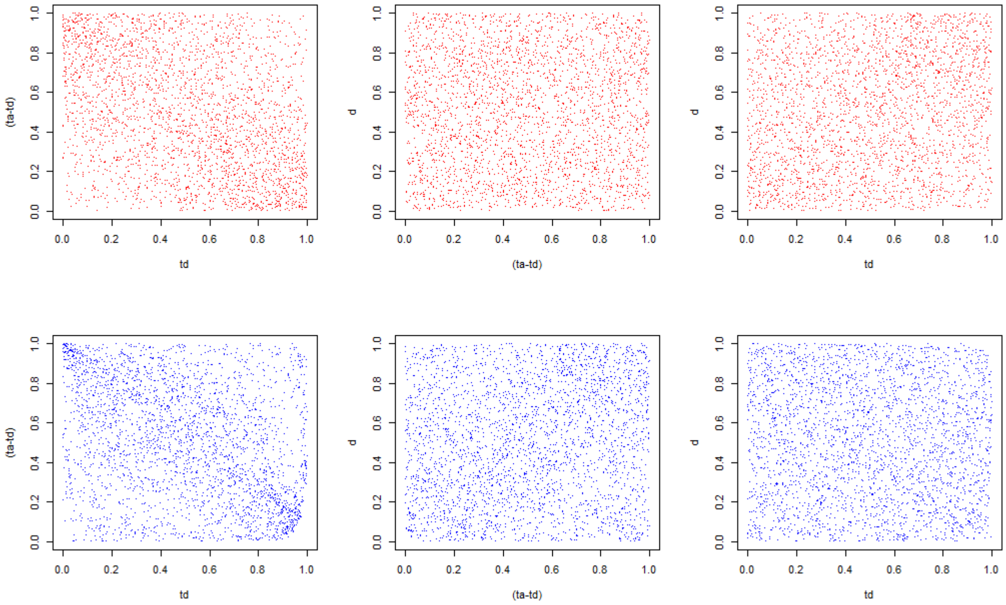

Figure 55: Above: scatter plots of the simulated model for cluster 3. Below: the scatter plots of cluster 3.

As we can see in these figures, the models are not a perfect fit. The scatter plots above are the models from the ranked clusters, and the ranked clusters itself. Therefore we transform the simulated data from each cluster model with the inverse empirical cdf of its respective cluster, which allows us to compare the original data from each cluster with their respective transformed simulated data from the cluster model. The results are given below, starting with cluster 1:

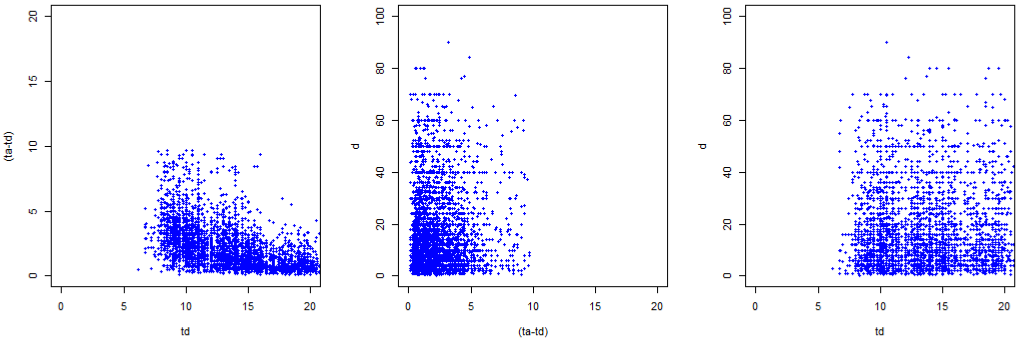

Figure 56: Scatter plots of the transformed simulated data from the copula model for cluster 1

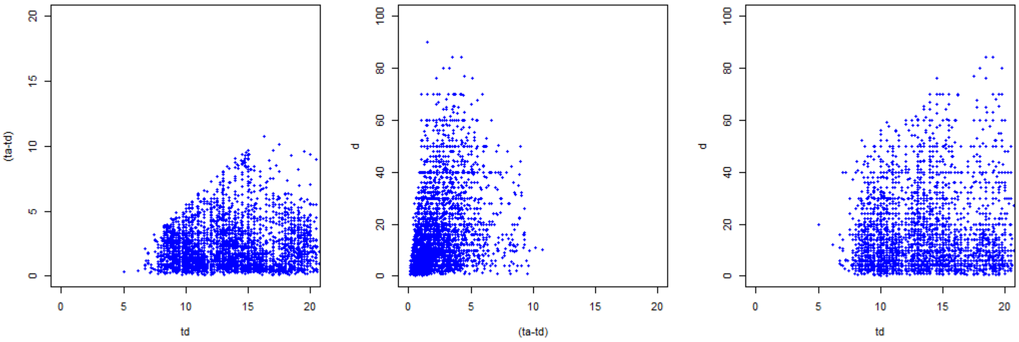

Figure 57: Scatter plots of cluster 1.

We see that the $(td, ta - td)$ scatter plot of the simulated data differs a lot from the respective scatter plot of the original cluster. The other two scatter plots of the simulated data show similarity with their respective original cluster. The result for cluster 2 is given below:

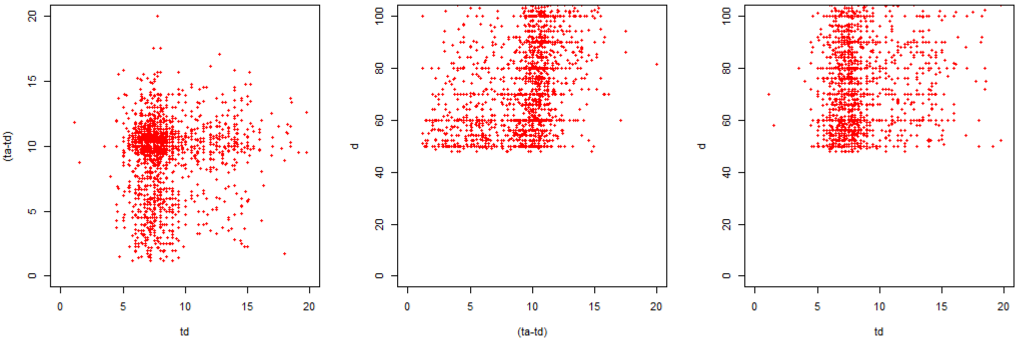

Figure 58: Scatter plots of the transformed simulated data from the copula model for cluster 2

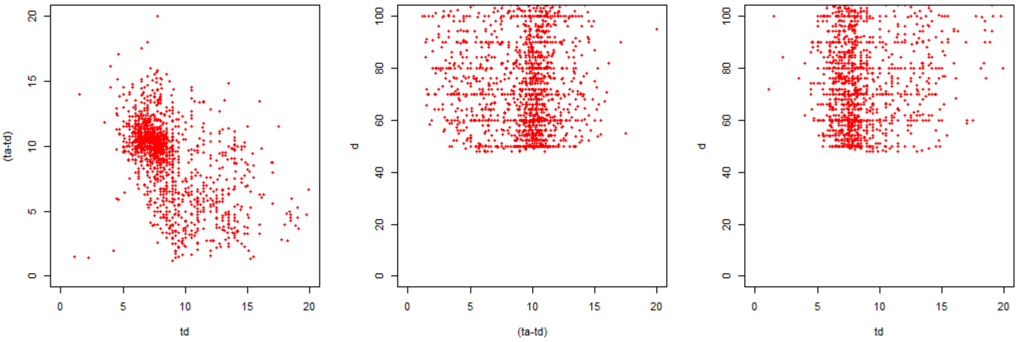

Figure 59: Scatter plots of cluster 2.

The $(td,\ ta - td)$ and $(ta - td,\ d)$ scatter plots differ from from their respective scatter plot of the original cluster. The (td,d) scatter plot of the simulated data show similarities with its respective original cluster scatter plot. The result for cluster 3 is given below:

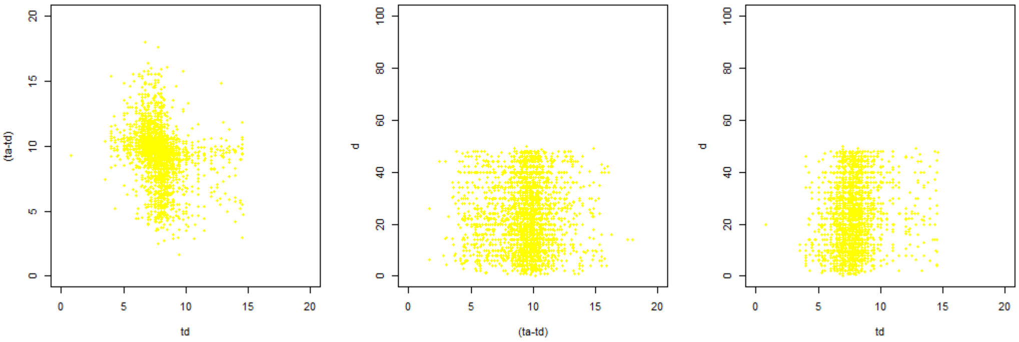

Figure 60: Scatter plots of the transformed simulated data from the copula model for cluster 3

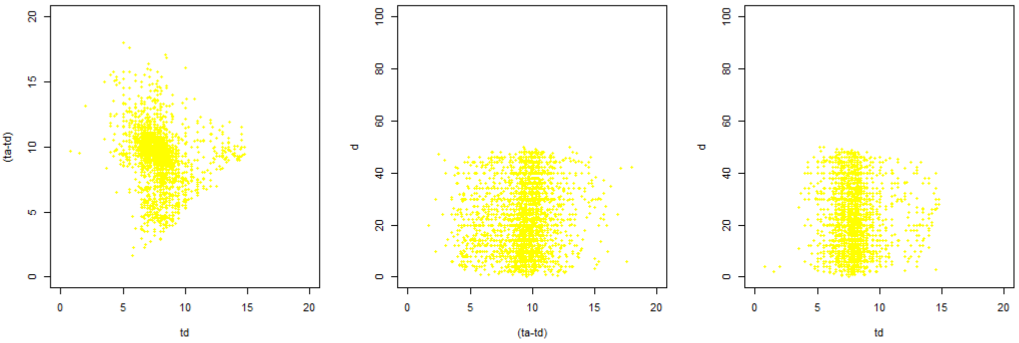

Figure 61: Scatter plots of cluster 3.

We see that again the $(td, ta - td)$ scatter plot from the simulated data differs from its original cluster scatter plot. The other two scatter plots show much similarities with their respective original cluster scatter plot.

We compared every cluster and its corresponding simulated data separately. Here we saw that the simulated data is not the same as the original data, but all simulated data sets are quite similar to their respective original cluster. Therefore we would like to compare the original clusters and the simulated data when we combine them. This lets us compare the 3 original clusters together with the 3 simulated data sets together. The results are given below, where the colors of the data correspond to their respective cluster:

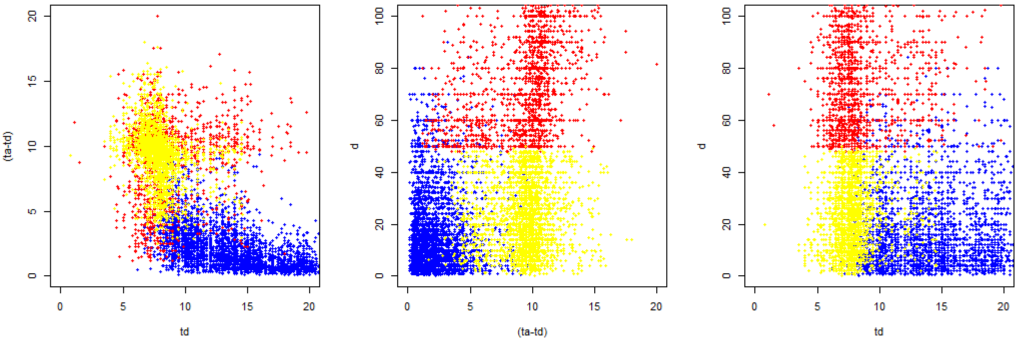

Figure 62: Scatter plots of the simulated data combined.

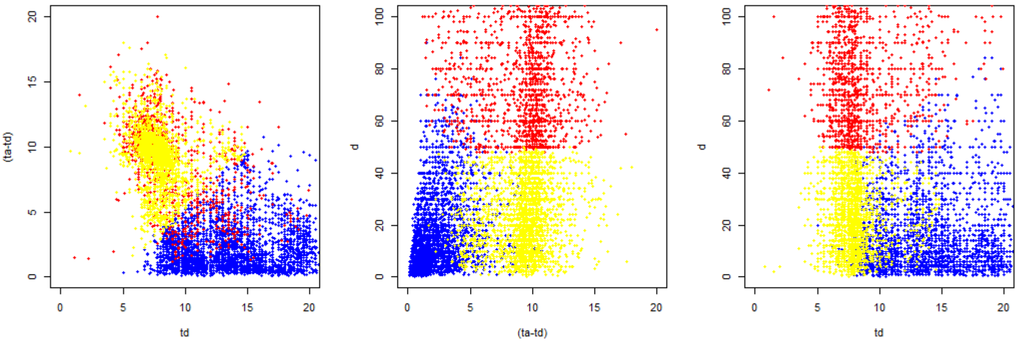

Figure 63: Scatter plots of the combined data of the three original clusters.

We see that the $(td, ta - td)$ scatter plot of the simulated model data differs from the original data, but the other two scatter plots show much similarity between the simulated data and the original data. This lets us to conclude that the models we found for each cluster are not a perfect fit for the data, but the models did provide a good simulation of data which resembled the original data quite reasonable.

Now we have analyzed the 2, 3 and 4 clusters case, we want to know if we get a better mixture model when we cluster the data with more than 4 clusters. Therefore we have analyzed the data set with 5 and 6 clusters to see if their respective mixture models were better, but these amount of clusters did not give us better models. The analysis of these amount of clusters is given in the appendix A.15. We will compare all the amount of clusters (2 to 6) based on their goodness of fit test p-values. The results are given below:

Model	Cluster 1	Cluster 2	Cluster 3	Clusters 4	Cluster 5	Cluster 6
2 clusters	0.0042509	0	-	-	-	-
3 clusters	$2.3455 * 10^{-5}$	0.10545	$4.3867 * 10^{-10}$	-	-	-
4 clusters	0.00092966	0.001728	$8.1864 * 10^{-5}$	0	-	-
5 clusters	0	0.00026633	0.00026379	$2.302 * 10^{-11}$	0.17635	-
6 clusters	0.014104	0.22901	$1.3323 * 10^{-15}$	$4.3516 * 10^{-5}$	$6.9901 * 10^{-9}$	$1.6764 * 10^{-14}$

Here we can see that there is no clear amount of clusters which gives us the best result, since no amount of clusters gives us models which all fit their cluster well. Therefore we conclude that clustering our data has, for now, not given us a better model for our data set.

7 Conclusion

The goal of this thesis was to find a good copula model for the transportation data set. If a good model for the behaviour of people using electric vehicles as a mean of transportation is built then this model can be used to simulate a virtual population of electric vehicle users. This virtual population can be used to study the influence of the increased penetration of electric vehicles on the electricity network.

We have introduced the theory of copula models which has been shown to be the most promising for this purpose. Since we have observed quite complicated dependencies between variables in this data set, the best performance was achieved by the vine copula model. Still even this flexible dependence model was not able to fit the data very well.

The reason for this underwhelming performance of the vine copula model was that there were multiple groups in the data set. Behaviour of people in different groups varies significantly and when all are analysed together there is no simple model that can capture their behaviour. The information of the purposes of trips that the people included in the data set was not provided. Hence to separate people with different behaviour we have used unsupervised classification techniques called clustering.

The k-mean clustering algorithm divided the data in groups that were easily interpretable. However, it was difficult to decide how many groups are actually present in the data set. Finally, we have built mixtures of vine copula models for different number of clusters starting from 2 and we showed that we can not clearly compare mixture models. We could compare them based on the goodness of fit test p-value of the model for each cluster for each chosen amount of clusters. This gave us that there was no clear best model for a given amount of clusters, but we choose the 3 clusters case as the best mixture model.

These 3 clusters contain 2 clusters representing the working people and 1 cluster representing the not working people. The 2 clusters of the working people are separated based on the traveled distance of the people in the clusters.

From the goodness of fit tests we concluded that in the 3 clusters case 2 models did not perform good, and one cluster performed good. When we compare the simulated cluster models and their respective cluster scatter plots, we see that the models are reasonable fits for the data. Some models do fit the data well, and some models fit the data less good. We also compared the whole clustered data set with the transformed simulated data from the cluster models in figure 63. In this figure we saw that in the 3 clusters case, the simulated data from the cluster models did not simulate the original data set exactly, but did simulate data which was reasonably similar to the original data set.

We have used clustering techniques on our data set in hope to find a better model for the data set, but this didn't unfortunately happen. The model we found in Chapter 4 was the best model for the original data set in terms of the AIC and likelihood value. For this model, we saw that it does not fit the data very well, and therefore we conclude that we have found a best model, but this model is not an optimal model. Our best model can not yet be used for simulating a virtual population of electric vehicle users. Further research has to be done in order to find such a model. The research and techniques presented in article [33] can be tested for our models to improve our results.

8 Discussion

The goal of this thesis was to find a good copula model for a transportation data set. We found a best copula model by modelling the original data set. Since we had a suspicion of different groups being present in the data set, we clustered our data set. We build mixture models by modelling the clusters, in hope to find a new best model for the data set. Unfortunately, there was not an definitive way of comparing these mixture models with each other. The p-values of the separate cluster models also gave us that these mixture models do not fit the data very well. Therefore we want to discuss why these mixture models performed worse than the original model and which further research can be done in order to obtain a possible optimal model for the data set.

8.1 Mixture models

To be able to fit models to the clusters, we had to rank every cluster separately. Due to every cluster getting ranked separately, we treated the marginal distributions of every cluster differently. Because of this, we could not compare the likelihood and AIC value of the different models for each amount of clusters with each other. Therefore the only way to compare the clusters is to analyze the goodness of fit test results of each separate model for its cluster. No good comparisons between the models of each cluster can therefore be made, and therefore a Vuong test could be implemented for comparing these cluster models with each other as further research. This would allow us to compare different cluster models better, instead of only looking at the goodness of fit p-values. This Vuong test could test if a mixture model with more parameters (so more clusters) is better then a model with less parameters (so less clusters). This is useful for our thesis, since we use different amount of clusters. Using a lot of clusters could also result in over fitting the data set. As already mentioned in the conclusion, the techniques from paper [33] could be used for modeling our clustered data set better.

8.2 Copulas

As explained in Chapter 3, we only use a certain set of copula functions. We did this due to restrictions of functions in R. For further research, more copulas could be used in the model fitting which could allow for better models. Therefore the functions in R should be extended for the use of more copula functions, in order to perform goodness of fit and Vuong tests for these new copula models.

8.3 Goodness of fit test

We perform the goodness of fit test in R with the RVineGOFTest() function. This function gives the test statistic with a corresponding bootstrapped or non bootstrapped p-value (which one can be chosen beforehand). From [35] we know that this test statistic is chi squared distributed with p(p+1)/2 (where p are the amount of parameters of the model) degrees of freedom. Due to this we have chosen to use the non bootstrapped p-value in this thesis, which gave us that every model, except for the model for example data set, in this thesis did not fit their respective data set very well. Therefore, bootstrapped values could be used to further investigate the goodness of fit of all our model. In order to get reasonable bootstrapped p-values, large amount of bootstrap steps have to be chosen. This resulted in very large computational times in R, so we have not analyzed them in this thesis. In further research, these bootstrapped p-values could be further researched and computed.

A Appendix

A.1 Calculation of the Normal copula tail dependence

Consider a Gaussian/Normal copula C. Since the Gaussian/Normal copula is symmetric, the upper and lower tail dependence is the same. So we can calculate the lower tail dependence only, to also know the upper tail dependence. The lower tail dependence of the Normal copula is as follows [13]:

$$\lambda = \lim_{q \to 0^+} \frac{\partial C(q,q)}{\partial q} \tag{22}$$

$$= \lim_{q \to 0^+} \mathrm{P}(V \leq q \,|\, U = q) + \lim_{q \to 0^+} \mathrm{P}(V \leq q \,|\, U = q) \tag{23}$$

This gives

$$\lambda = 2 \lim_{q \to 0^+} \mathrm{P}(V \leq q \,|\, U = q)$$

Now, let:

$$(Z_1, Z_2) := \Big(\Phi^{-1}(U),\, \Phi^{-1}(V)\Big)$$

This means that (Z_1, Z_2) has a bivariate normal distribution with standard marginals and correlation ρ. Now:

$$\lambda = 2 \lim_{q \to 0^+} \mathrm{P}(\Phi^{-1}(V) \leq \Phi^{-1}(q) \,|\, \Phi^{-1}(U) = \Phi^{-1}(q)) \tag{24}$$

$$= 2 \lim_{x \to -\infty} \mathrm{P}(Z_2 \leq x \,|\, Z_1 = x) \tag{25}$$

Finally, we know that $Z_2 \,|\, Z_1 \sim N(\rho x, 1 - \rho^2)$, so:

$$\lambda = 2 \lim_{x \to -\infty} \Phi\left(x \sqrt{\frac{(1-\rho)}{(1+\rho)}} \right) = 0$$

A.2 Calculation of the Student-t copula tail dependence

We apply l'Hopital's rule on 10 for the expression $\lambda = \lambda_l$, and we obtain [30]:

$$\lambda = \lim_{q \to 0^+} \frac{\partial C(q,q)}{\partial q} \tag{26}$$

$$= \lim_{q \to 0^+} \mathrm{P}(U_2 \leq q \,|\, U_1 = q) + \lim_{q \to 0^+} \mathrm{P}(U_1 \leq q \,|\, U_2 = q), \tag{27}$$

where (U_1, U_2) is a random pair whose df is C and the second equality follows from an easily established property of the derivative of copulas (see [1]). Suppose we now define $Y_1 = t_\nu^{-1}(U_1)$ and $Y_2 = t_\nu^{-1}(U_2)$ so that $(Y_1, Y_2) \sim t_2(\nu, 0, P)$. We have, using the exchangeability of (Y_1, Y_2), that

$$\lambda = 2 \lim_{y \to -\infty^+} \mathrm{P}(Y_2 \leq y \,|\, Y_2 = y) \tag{28}$$

Since, conditionally on $Y_1 = y$ we have

$$\left(\frac{\nu+1}{\nu+y^2}\right)^{1/2} \frac{Y_2 - \rho y}{\sqrt{1-\rho^2}} \sim t_1(\nu+1, 0, 1) \tag{29}$$

this limit may now be easily evaluated and shown to be 12

A.3 Calculation of the Clayton copula tail dependence

$$\begin{aligned}
\lambda_U &= \lim_{q\to 1^+} \frac{1-2q+C(q,q)}{1-q} \\
&= \lim_{q\to 1^+} \frac{1-2q+(q^{-\alpha}+q^{-\alpha}-1)^{-1/\alpha}}{1-q} \\
&= \lim_{q\to 1^+} \frac{1-2q+(2q^{-\alpha}-1)^{-1/\alpha}}{1-q} \\
&= \lim_{q\to 1^+} \frac{-2+2(2q^{-\alpha}-1)^{-\frac{1}{\alpha}-1}q^{-\alpha-1}}{-1} \\
&= \lim_{q\to 1^+} 2-2(2q^{-\alpha}-1)^{-\frac{1}{\alpha}-1}q^{-\alpha-1} \\
&= 2-2*(1)*1 \\
&= 0
\end{aligned}$$

$$\begin{aligned}
\lambda_L &= \lim_{q\to 0^+} \frac{C(q,q)}{q} \\
&= \lim_{q\to 0^+} \frac{(q^{-\alpha}+q^{-\alpha}-1)^{-1/\alpha}}{q} \\
&= \lim_{q\to 0^+} \frac{(2q^{-\alpha}-1)^{-1/\alpha}}{q} \\
&= \lim_{q\to 0^+} \frac{(2-q^{\alpha})^{-1/\alpha}(q^{-\alpha})^{-1/\alpha}}{q} \\
&= \lim_{q\to 0^+} \frac{(2-q^{\alpha})^{-1/\alpha}q}{q} \\
&= \lim_{q\to 0^+} \frac{1}{(2-q^{\alpha})^{1/\alpha}} \\
&= \frac{1}{2^{1/\alpha}} \\
&= 2^{-\frac{1}{\alpha}}
\end{aligned}$$

A.4 Calculation of the Gumbel copula tail dependence

$$\begin{aligned}
\lambda_L &= \lim_{q\to 0^+} \frac{C(q,q)}{q} \\
&= \lim_{q\to 0^+} \frac{\exp -[(-\ln q)^\alpha + (-\ln q)^\alpha]^{1/\alpha}}{q} \\
&= \lim_{q\to 0^+} \frac{\exp\,[2(\ln q)^\alpha]^{1/\alpha}}{q} \\
&= \lim_{q\to 0^+} \frac{\exp 2^{1/\alpha} \ln q}{q} \\
&= \lim_{q\to 0^+} \frac{q^{2^{1/\alpha}}}{q} \\
&= \lim_{q\to 0^+} q^{2^{1/\alpha}-1} \\
&= 0
\end{aligned}$$

$$\begin{aligned}
\lambda_U &= \lim_{q\to 1^+} \frac{1-2q+C(q,q)}{1-q} \\
&= \lim_{q\to 1^+} \frac{1-2q+\exp -[(-\ln q)^\alpha + (-\ln q)^\alpha]^{1/\alpha}}{1-q} \\
&= \lim_{q\to 1^+} \frac{1-2q+\exp\,[2(\ln q)^\alpha]^{1/\alpha}}{1-q} \\
&= \lim_{q\to 1^+} \frac{1-2q+\exp 2^{1/\alpha} \ln q}{1-q} \\
&= \lim_{q\to 1^+} \frac{1-2q+q^{2^{1/\alpha}}}{1-q} \\
&= \lim_{q\to 1^+} \frac{-2+2^{1/\alpha} q^{2^{1/\alpha}-1}}{-1} \\
&= \frac{-2+2^{1/\alpha}}{-1} \\
&= 2-2^{\frac{1}{\alpha}}
\end{aligned}$$

A.5 Calculation of the Frank copula tail dependence

The Frank copula is a symmetric copula, so the upper tail dependence is equal to the lower tail dependence.

$$
\begin{aligned}
\lambda_L &= \lim_{q \to 0^+} \frac{C(q,q)}{q} \\
&= \lim_{q \to 0^+} \frac{-\frac{1}{\alpha} \ln\left(1 + \frac{(e^{-\alpha q}-1)^2}{(e^{-\alpha}-1)}\right)}{q} \\
&= \lim_{q \to 0^+} -\frac{1}{\alpha} * \frac{1}{1 + \frac{(e^{-\alpha q}-1)^2}{(e^{-\alpha}-1)}} * \frac{2(e^{-\alpha q}-1)}{(e^{-\alpha}-1)} * -\alpha e^{-\alpha q} \\
&= -\frac{1}{\alpha} * \frac{1}{1} * 2 * 0 * -\alpha \\
&= 0 \\
&= \lambda_U
\end{aligned}
$$

A.6 Scatter plots of our model and the original example data set

Below we show the corresponding scatter plots between the bivariate random vectors of both data sets next to each other.

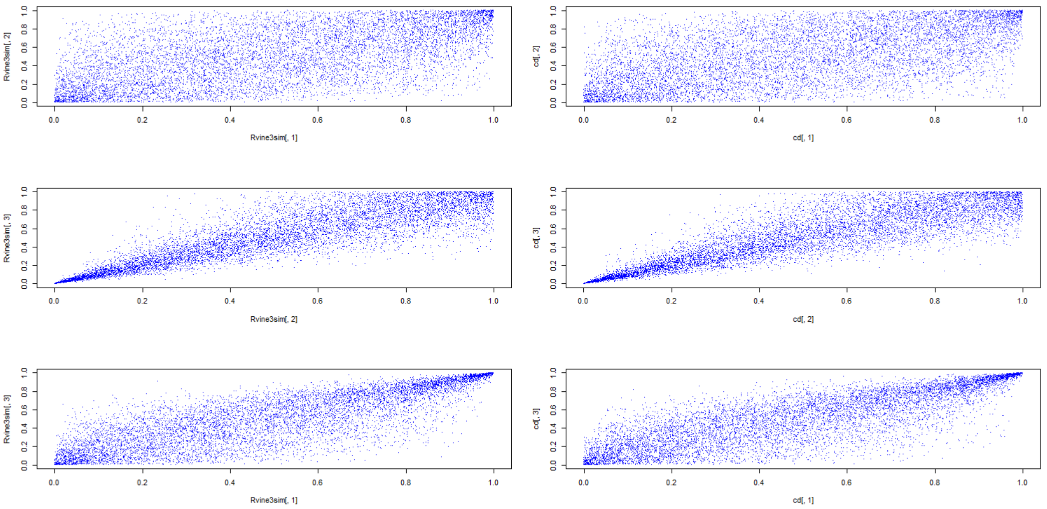

Figure 64: Left: scatter plots Rvine2 model. Right: corresponding scatter plots of the original data set

A.7 Silhouette plot of the 5 clusters case with the k means method

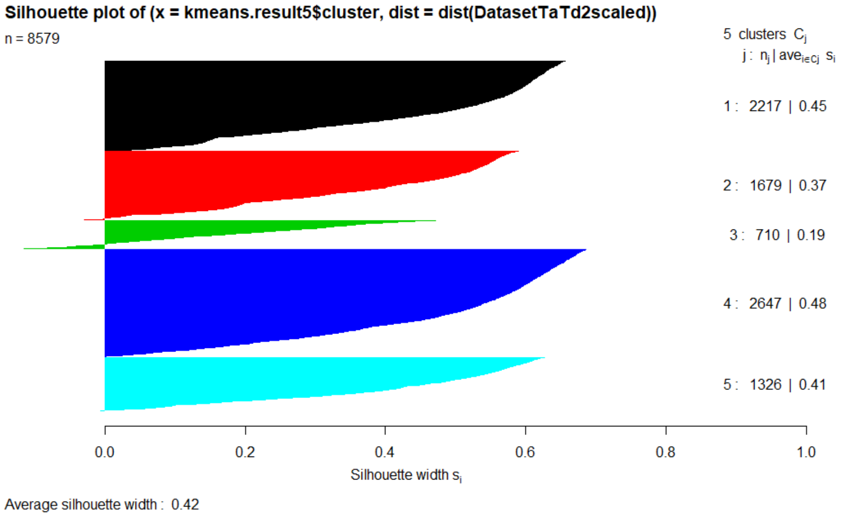

Figure 65: Silhouette plot for 5 clusters with the k means method.

A.8 Silhouette plot of the 6 clusters case with the k means method

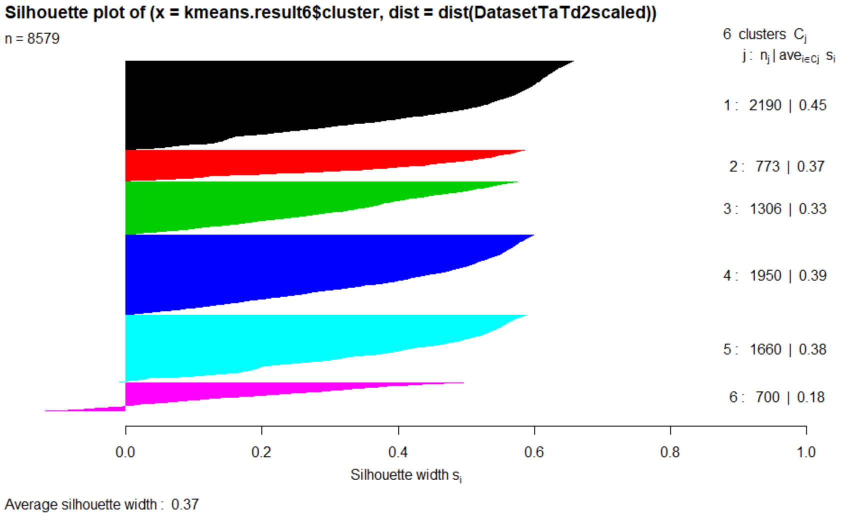

Figure 66: Silhouette plot for 6 clusters with the k means method.

A.9 Hierarchical clustering with the Euclidean space and the complete linkage method for 2 clusters scatter plot.

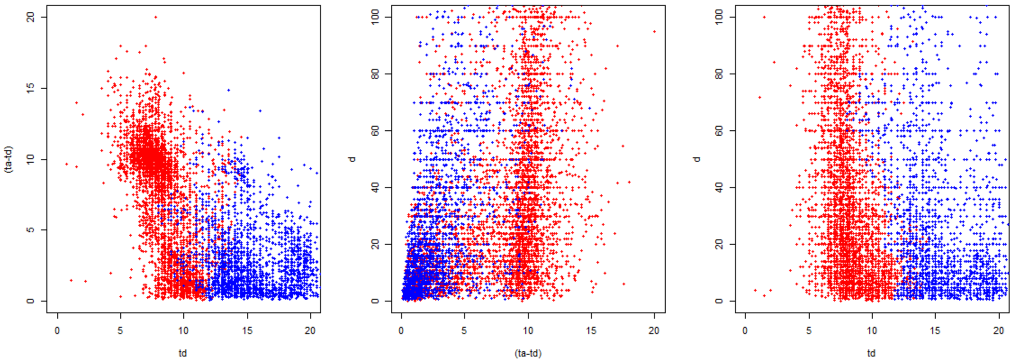

Figure 67: Silhouette plot for 6 clusters with the k means method.

A.10 Hierarchical clustering with the Euclidean space and the complete linkage method for 3 clusters scatter plot.

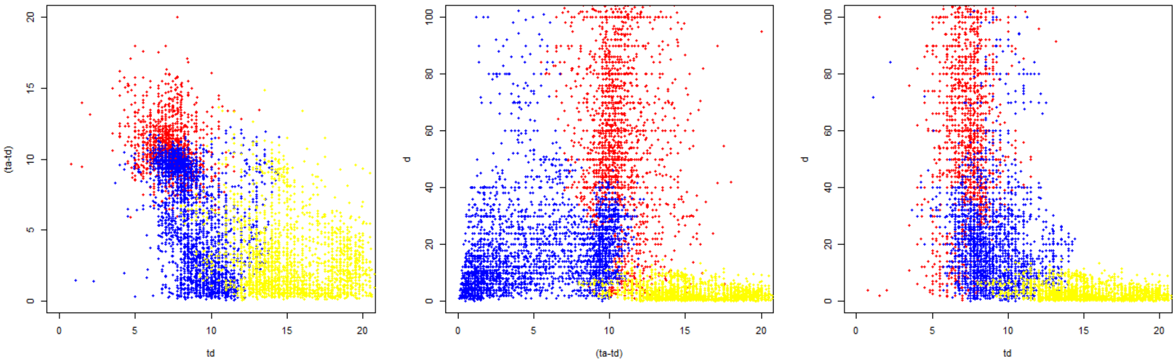

Figure 68: Silhouette plot for 6 clusters with the k means method.

A.11 Marginal plots for the k means method with Kendall correlation distance for 3 clusters

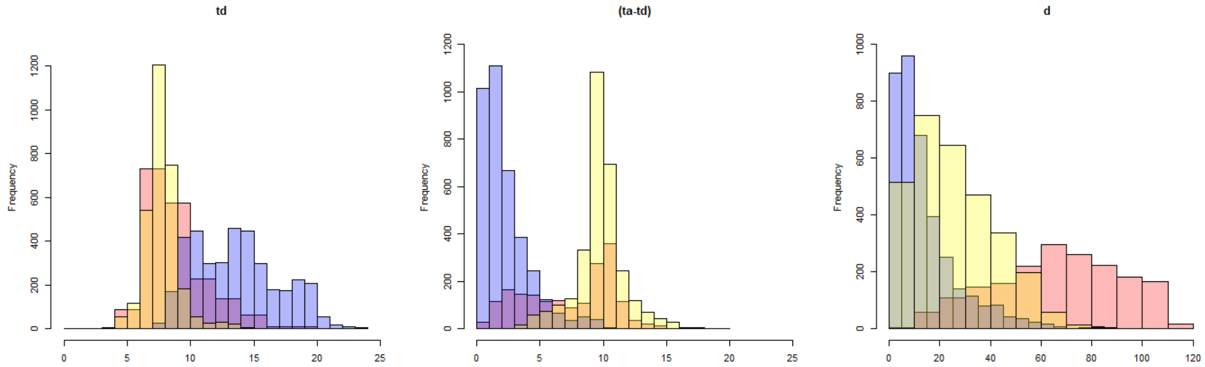

Figure 69: Silhouette plot for 6 clusters with the k means method.

A.12 Marginal plots for the k means method with Kendall correlation distance for 4 clusters

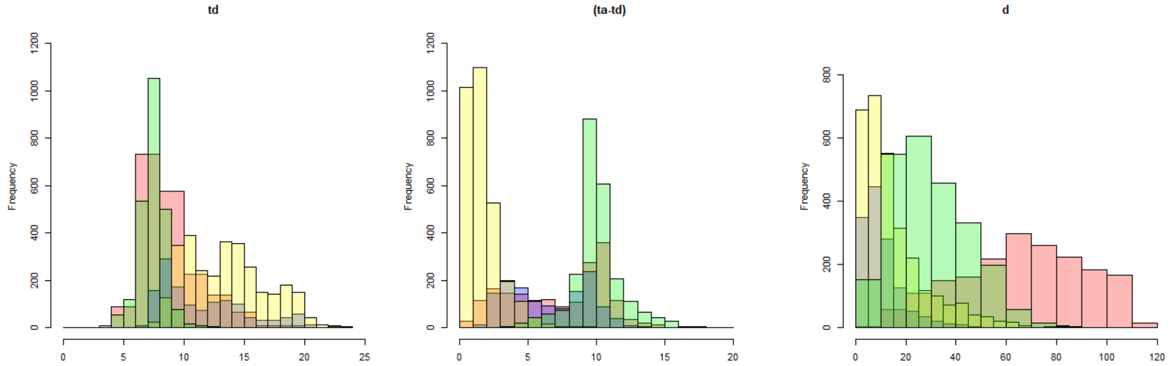

Figure 70: Silhouette plot for 6 clusters with the k means method.

A.13 Analysis of the k means method with 2 clusters

The best model for both the clusters were a regular vine copula model. The best possible model for each cluster are given below:

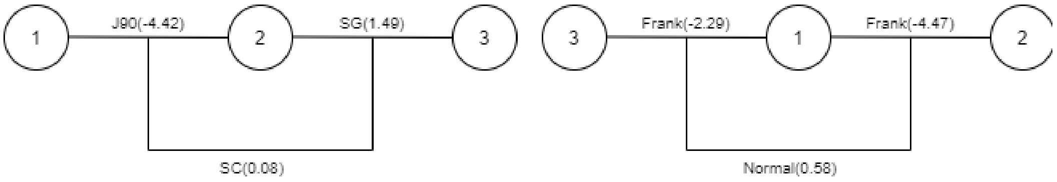

Figure 71: On the left the best copula model for cluster 1, on the right the best copula model for cluster 2.

The table below show us the results for each individual model:

Model	Likelihood	AIC	#parameters
Cluster 1	465.24	-924.49	3
Cluster 2	511.67	-1015.34	4

A.14 Analysis of the k means method with 4 clusters

The best mixture model for the clustered data set with 4 clusters was composed of the following four models:.

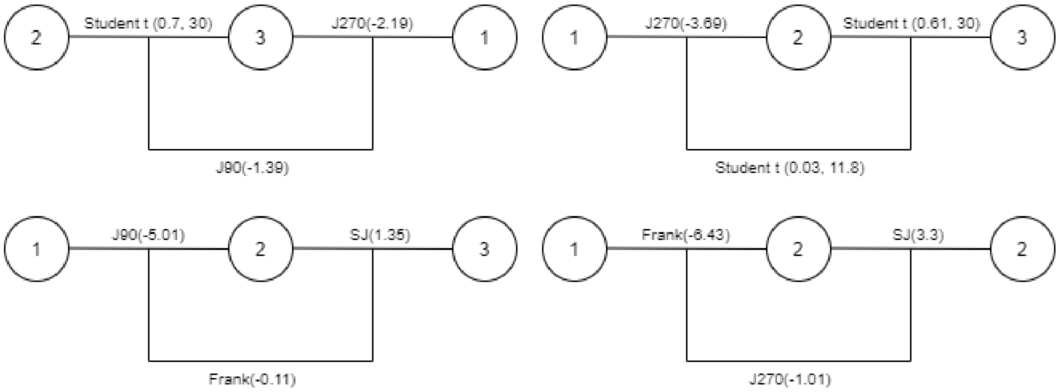

Figure 72: From left to right: the best vine copula model for cluster 1, cluster 2, cluster 3 and cluster 4.

The table below shows us the results for each individual model:

Model	Likelihood	AIC	#parameters
Cluster 1	264.39	-522.78	3
Cluster 2	273.31	-540.63	3
Cluster 3	230.73	-455.47	3
Cluster 4	218.19	-428.38	4

A.15 Analysis of the mixture model for 6 clusters

The table below show us the results for each individual model:

Model	Likelihood	AIC	#parameters
Cluster 1	137.18	-268.36	3
Cluster 2	44.48	-80.97	4
Cluster 3	116.81	-225.62	4
Cluster 4	194.59	-383.17	3
Cluster 5	100.05	-192.11	4
Cluster 6	11.56	-17.11	3

A.16 Analysis of the k means method with 5 clusters

The table below show us the results for each individual model:

Model	Likelihood	AIC	#parameters
Cluster 1	9.74	-13.48	3
Cluster 2	192.69	-379.39	3
Cluster 3	140.5	-274.99	3
Cluster 4	189.71	-373.42	3
Cluster 5	65.95	-123.89	4

References

[1] Roger B. Nelsen, An introduction to Copulas, second edition, 2006

[2] Kruskal, W. H., Ordinal Measures of Association, 1958

[3] Gumbel, E., Bivariate Exponential Distributions. Journal of the American Statistical Association, 1960

[4] D. Clayton, A Model for Association in Bivariate Life Tables and Its Application in Epidemiological Studies of Familial Tendency in Chronic Disease Incidence, 1978

[5] Harry Joe, Multivariate Models and Multivariate Dependence Concepts, 1997

[6] Pravin K. Trivedi, David M. Zimmer, Copula Modeling: An Introduction for Practitioners, 2007

[7] H. Akaike, A new look at the statistical model identification, 1974

[8] M. Kendall, A New Measure of Rank Correlation, 1938

[9] Clayton, D., A Model for Association in Bivariate Life Tables and its Application in Epidemiological Studies of Family Tendency in Chronic Disease Incidence, 1978

[10] Cherubini, W., Luciano, W. and Vecchiato, W., Copula methods in finance, 2004

[11] Wilks, S. S., The large-sample distribution of the likelihood ratio for testing composite hypotheses, 1938

[12] Embrechts, P., McNeil, A. and Straumann, D., Correlation and Dependence in Risk Management: Properties and Pitfalls, Risk Management: Value at Risk and Beyond, 2002

[13] StackExchange, Why is Gaussian Copula's Tail Dependence Zero?, 2017 `https://stats.stackexchange.com/questions/245638/why-is-gaussian-copulas-tail-dependence-zero`

[14] Eike Christian Brechmann, Ulf Schepsmeier, Modeling Dependence with C- and D- Vine Copulas: The R package CDVine, 2013

[15] Alexander J. McNeil, Johanna Neslehova, Multivariate Archimedean copulas, D-monotone functions and λ_1-norm symmetric distributions, 2009

[16] Harry Joe, Dorota Kurowicka, Dependence Modeleing: Vine Copula Handbook, 2010

[17] Czado, Claudia, Analyzing Dependent Data with Vine Copulas, 2019

[18] Tim Bedford, Roger M. Cooke, Probability Density Decomposition for Conditionally Dependent Random Variables Modeled by Vines, 2001

[19] Quang H. Vuong, Likelihood ratio tests for model selection and non-nested hypotheses, 1986

[20] Ingrid Hobaekhaff, Parameter estimation for pair-copula constructions, 2013

[21] A.K. Jain, M.N. Murty, P.J. Flynn, Data clustering: A review, 1999

[22] Anil K. Jain, Data clustering: 50 years beyond K-means, 2009

[23] Jiawei Han, Micheline Kamber, Jian Pei, Data mining concepts and techniques, 2011

[24] Peter J.Rousseeuw, Silhouettes: A graphical aid to the interpretation and validation of cluster analysis, 1987

[25] Kurz, M. S. and F. Spanhel, Testing the simplifying assumption in high-dimensional vine copulas, 2017

[26] Ivan Kojadinovic, Some copula inference procedures adapted to the presence of ties, 2017

[27] White, H., Maximum likelihood estimation of misspecified models, 1982

[28] Alicja Lojowska, Dorota Kurowicka, Georgios Papaefthymiou, Lou van der Sluis, From Transportation Patterns to Power Demand: Stochastic Modeling of Uncontrolled Domestic Charging of Electric Vehicles , 2011

[29] Nicolò Daina, Aruna Sivakumar, John W. Polak, Modelling electric vehicles use: a survey on the methods, 2017

[30] Stefano Demarta, Alexander J. McNeil, The t Copula and Related Copulas, 2004

[31] Gabriel Frahm, Markus Junker, Rafael Schmidt,, Estimating the tail-dependence coefficient: Properties and pitfalls, 2005

[32] IEA, Global EV Outlook 2019, 2019 `https://www.iea.org/reports/global-ev-outlook-2020`

[33] Mingyang Sun, Ioannis Konstantelos, G. Strbac, C-Vine Copula Mixture Model for Clustering of Residential Electrical Load Pattern Data, 2017

[34] Amerise, Ilaria L, Correction methods for ties in rank correlations, 2015

[35] Ulf Schepsmeier, Efficient goodness-of-fit tests in multi-dimensional vine copula models, 2013

[36] Jaccard P., Nouvelles recherches sur la distribution florale. Bull., 1908

[37] Real R., Vargas J.M., The Probabilistic Basis of Jaccard's Index of Similarity Systematic Biology, 1996

[38] Yan Li, Yang Li, Yichen Qin, and Jun Yan, Copula modeling for data with ties, 2016