RESERACH SEMINAR

Speed check: copula selection and parameter estimation using non-parametric moment inverses and maximum likelihood

Author:
Niklas PAULIG

Supervisor:
Prof. Dr. Bernhard SCHIPP
M.Sc. Paul Felix REITER

A research seminar submitted in fulfillment of the requirements for enrolling a Master thesis

at the

Chair of Quantitative Methods, esp. Econometrics

November 22, 2020

Abstract

Faculty of Business and Economics
Chair of Quantitative Methods, esp. Econometrics

Research seminar

Speed check: copula selection and parameter estimation using non-parametric moment inverses and maximum likelihood

by Niklas PAULIG

The purpose of this article is to conduct a Monte-Carlo simulation study that compares the efficiency, accuracy and empirical computation time of three different estimators in terms of selecting copulae and estimating their parameters. I will compare two moment-based estimators, the inverse of Kendall's Tau and Blomqvist's Beta, to the classic canonical maximum likelihood estimator. I find that, in terms of estimating parameters for a given copula, the moment-based estimators are inferior to maximum likelihood, especially for small sample sizes. For selecting copulae from a set of data, however, all three estimators are equally accurate while the moment-based ones are at least 16 times faster computationally.

Contents

List of Figures

List of Tables

List of Abbreviations

i.i.d	**i**ndependent and **i**dentically **d**istributed
CDF	**C**umulative **D**istribution **F**unction

1 Introduction into copulae

1.1 Motivation

A copula is a bi- or multivariate distribution function, that allows to model dependency between variables with arbitrary marginal distributions. It has, therefore, become a handy tool in risk and portfolio analysis but also in geology and hydrology as the joint behavior of arbitrarily distributed variables can be modeled precisely. Apart from the empirical copula, all copulae are parameterized by either one, or two parameters that allows them to be fit to practically any desired dependence structure. The efficient estimation of a copula's parameter or the selection of a suitable copula itself is subject to current research. For this study I use a novel approach for selecting a copula given a set of data, using two computationally efficient moment-based estimators in combination with Akaike's Information Criterion, to achieve maximum-likelihood-like accuracy with only $1/16^{th}$ the computational effort.

The rest of this article is organized as follows. The first chapter gives an introduction into copulae itself, and how they are defined. Chapter two details the estimation methods used in this Monte-Carlo study while chapter three carries out the study and concludes.

1.2 Copula existence and Sklar's theorem

At the very heart of copula theory is the theorem of Sklar (1959) stating that any multivariate distribution can be decomposed into its univariate margins and a copula, which captures the dependence between variables. Therefore, let $\mathbf{X} = (X_1, \ldots, X_d)^\top$ be a d-dimensional random vector with joint distribution F and marginal distributions $F_{X_1}, \ldots, F_{X_d}$. Now, for every realization $\mathbf{x} = (x_1, \ldots, x_d)^\top \in [-\infty, \infty]^d$,

$$\begin{aligned} F(\mathbf{x}) &= C\left(F_{X_1}(x_1), \ldots, F_{X_d}(x_d)\right) \\ &= P(X_1 < x_1, \ldots, X_d < x_d), \end{aligned} \tag{1.1}$$

where the copula C can be interpreted as a d-dimensional distribution function defined on the unit hypercube $C : [0,1]^d \to [0,1]$ with uniformly distributed marginals (see Angus (1994) for proof). If all $F_{\mathbf{X}}$ are continuous C is unique.

The inverse theorem also holds: If $\mathbf{u} \equiv (u_1, \ldots, u_d)^\top \in [0,1]^d$ and $u_i = F_{X_i}(x_i)$ for $i = 1, \ldots, d$ the copula C can be expressed as

$$C(\mathbf{u}) = F\left(F_{X_1}^{-1}(u_1), \ldots, F_{X_d}^{-1}(u_d)\right), \tag{1.2}$$

with the corresponding multivariate distribution F, marginal distributions $F_{\mathbf{X}}$ and inverses F^{-1}, given their existence (Durante, Fernandez-Sanchez, and Sempi, 2013). Furthermore, using the relationship between density functions and distribution functions and their partial derivatives, the

multivariate joint density can be expressed as the product of the copula density and the marginal densities

$$f(\mathbf{x}) = c\left(F_{X_1}(x_1), \ldots, F_{X_d}(x_d)\right) f_{X_1}(x_1) \ldots f_{X_d}(x_d), \tag{1.3}$$

with the copula density c being derived as $c = \frac{\partial C(\mathbf{u})}{\partial \mathbf{u}}$.

1.3 Selected copulae and properties

Since their discovery a variety of copula functions have been constructed for different purposes. There are only few Archimedean copulae defined for $d \geq 2$, therefore we will stick to bivariate representations for this simulation. This is sufficient as there is a concept - the Pair-Copula-Construction (PCC) - that allows to model multivariate problems by splitting them into several tow-dimensional ones and solving them consecutively.

There are two major classes of copulae distinguished by how they are generated:

1.3.1 Elliptical copulae

Following equation 1.2 a copula C is elliptical if F is elliptical, meaning its density can be represented by

$$f(\mathbf{x}; \boldsymbol{\mu}, \Sigma) = k_d |\Sigma|^{-\frac{1}{2}} g\left((\mathbf{x}-\boldsymbol{\mu})^\top \Sigma^{-1} (\mathbf{x}-\boldsymbol{\mu})\right), \tag{1.4}$$

with some constant $k_d \in \mathbb{R}$ dependent on the dimension, a vector of means $\boldsymbol{\mu} \in \mathbb{R}^d$, a symmetric positive definite matrix $\Sigma \in \mathbb{R}^{d \times d}$ and some function $g : \mathbb{R}_0^+ \to \mathbb{R}_0^+$ which is independent of the dimension d. One prominent example of a copula derived from elliptical distributions is the d-variate Gaussian copula,

$$C(\mathbf{u}; \Sigma) = \Phi_\Sigma^d\left(\Phi^{-1}(u_1), \ldots \Phi^{-1}(u_d)\right), \tag{1.5}$$

where $\Phi(\cdot)$ is the standard normal distribution function $N(0,1)$ and $\Phi_\Sigma^d(\cdot, \ldots, \cdot)$ being the d-variate normal distribution function with zero means, unit variances and correlation matrix Σ.

The Gaussian copula had been widely used in finance to model the joint default risk of portfolios containing different asset classes. This, however, led to an underestimation of defaulting risks as the phenomenon of dependent extreme values, which is often observed in financial return data cannot be modeled using this copula (see Salmon (2012) for a popular science article about this misconception and its effects on the financial crisis of 2007).

Many recent paper (Mashal, Naldi, and Zeevi (2003), Breymann, Dias, and Embrechts (2003) or Lucas, Schwaab, and Zhang (2014)) therefore use the t-copula which resembles the dependence structure implicit in a multivariate t-distribution as it is able to better capture extreme value dependence due to its heavy tails. The d-variate t-copula is defined as

$$C(\mathbf{u}; R, \nu) = T_{R,\nu}\left(T_\nu^{-1}(u_1), \ldots, T_\nu^{-1}(u_d)\right), \tag{1.6}$$

where $T_{R,\nu}$ denotes the multivariate students t-distribution with density

$$f(\mathbf{x}) = \frac{\Gamma\left(\frac{\nu+d}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)(\pi\nu)^{\frac{d}{2}}} |R|^{-\frac{1}{2}} \left(1 + \frac{(\mathbf{x}-\mu)^\top R^{-1} (\mathbf{x}-\mu)}{\nu}\right)^{\frac{-\nu+d}{2}}, \tag{1.7}$$

scale matrix $R \in [-1,1]^{d\times d}$ and the degrees of freedom $\nu > 0$. T_ν^{-1} represents the quantile function of the univariate students t- distribution with the same $\nu > 0$ degrees of freedom (see Demarta and McNeil (2005) for a more detailed description). Figure 1.1 shows the two just presented copulae for some visual intuition.

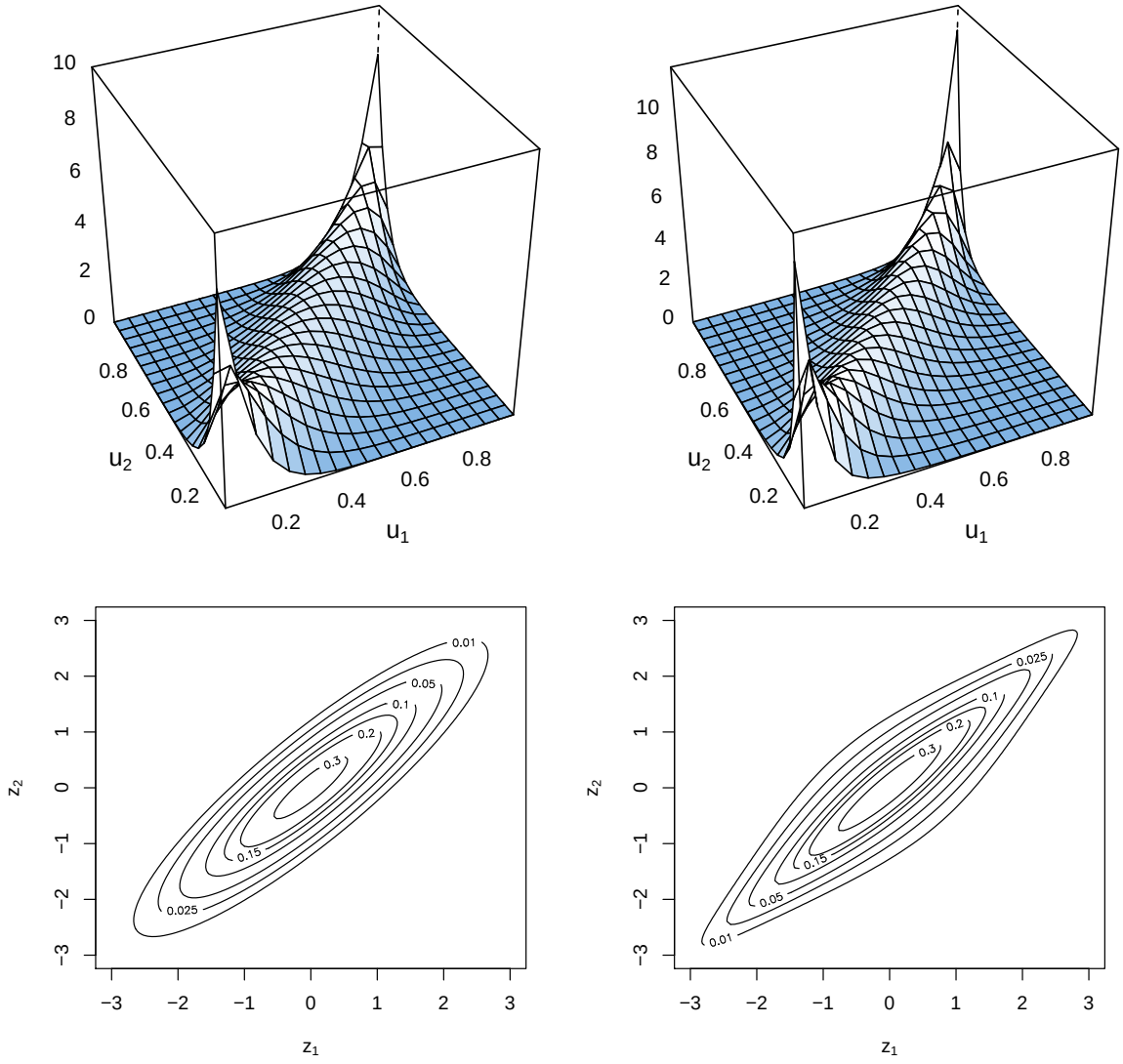

FIGURE 1.1: BIVARIATE ELLIPTICAL COPULA DENSITIES: Top: three-dimensional density plots of a theoretical Gaussian copula (left) and a t-copula with $\nu = 4$ degrees of freedom (right). The parameters have been chosen in such a way that Kendall's $\tau = 0.7$. Below are the respective contour plots belonging to the densities. The plots were generated using the BiCop function from the VineCopula package in R.

1.3.2 Archimedean copulae

Archimedean copulae and their detailed properties have already been discussed by for example Nelsen (2007) or McNeil, Nešlehová, et al. (2009). Therefore, we now derive just the fundamental building blocks for bivariate Archimedean copulae.

Following the work of Genest and Rivest (1993) a copula is called *Archimedean* if it can be expressed as

$$C_\varphi(\mathbf{u}) = \varphi^{[-1]}\left(\varphi(u_1) + \cdots + \varphi(u_d)\right), \tag{1.8}$$

for some continuous, strictly monotone decreasing, and convex function $\varphi : \mathbf{I} \to [0, \infty]$, satisfying $\varphi(1) = 0$ and $\mathbf{I}$ being the unit interval. This function is called the *generator function* for the copula and is furthermore called *strict* if $\varphi(0) = \infty$. $\varphi^{[-1]}$ is the pseudo-inverse satisfying the following conditions

$$\varphi^{[-1]}(t) := \begin{cases} \varphi^{-1}(t) & , 0 \le t \le \varphi(0) \\ 0 & , \varphi(0) \le t \le \infty. \end{cases} \tag{1.9}$$

We will now present the Archimedean copulae with its generators and properties that this simulation study will use. To date there has been defined a wide variety of Archimedean copulae with contrasting properties (see Nelsen (2007) for a comprehensive list).

For $\varphi_\theta(u) = (-\log(u))^\theta$ with $\theta \in [1, \infty)$ we attain the Gumbel copula. We include it in this simulation because it is representative for the subclass of asymmetric tail dependent copulae (upper tail dependence only) while the Gaussian and t-copulae represent either tail independence or symmetric tail dependence respectively. We will formally introduce this property in the next section. The Gumbel distribution function is defined as

$$C_\theta^G(\mathbf{u}; \theta) = \exp\left[-\left\{\sum_{j=1}^{d} (\log u_j)^\theta\right\}^{\frac{1}{\theta}}\right]. \tag{1.10}$$

Just like the Gaussian copula from the elliptical class there exists an Archimedean one that is symmetrically tail independent, the Frank copula. The generator is defined by $\varphi_\theta(u) = -\log\left\{\frac{\exp(-\theta u)-1}{\exp(-\theta)-1}\right\}$ with $\theta \in (-\infty, \infty) \setminus \{0\}$. The accompanying cdf can now easily be derived as

$$C_\theta^F(\mathbf{u}; \theta) = -\frac{1}{\theta}\log\left[1 + \frac{\prod_{j=1}^{d}\{\exp(-\theta u_j) - 1\}}{\{\exp(-\theta) - 1\}^{d-1}}\right]. \tag{1.11}$$

The last copula that we will include in the simulation is the Clayton copula, named after its discoverer in Clayton (1978). It somewhat resembles the counterpart to the Gumbel copula as it is only lower tail dependent. Its generator is defined by $\varphi_\theta(u) = \frac{1}{\theta}\left(u^{-\theta} - 1\right)$ for $\theta \in \left[-\frac{1}{d-1}, \infty\right) \setminus \{0\}$, which lets us define its distribution as

$$C_\theta^{Cl}(\mathbf{u}; \theta) = \left\{\left(\sum_{j=1}^{d} u_j^{-\theta}\right) - d + 1\right\}^{-\frac{1}{\theta}} \tag{1.12}$$

To get a graphical intuition about the different presented copulae, their three-dimensional bivariate densities and the respective contour plots are given in Figure 1.2.

1.4 Dependence measures

We will now, first, introduce the concept of the aforementioned tail dependence and, second, describe two ways to quantify the existence, direction and strength of dependence between two random variables and their connection to the dependence structure of a copula. The first of which will be *Kendall's Tau* and the second *Blomqvist's Beta*. There also exists a variety of other dependence measures; for the sake of simplicity, however, we will only present those needed for our simulation study.

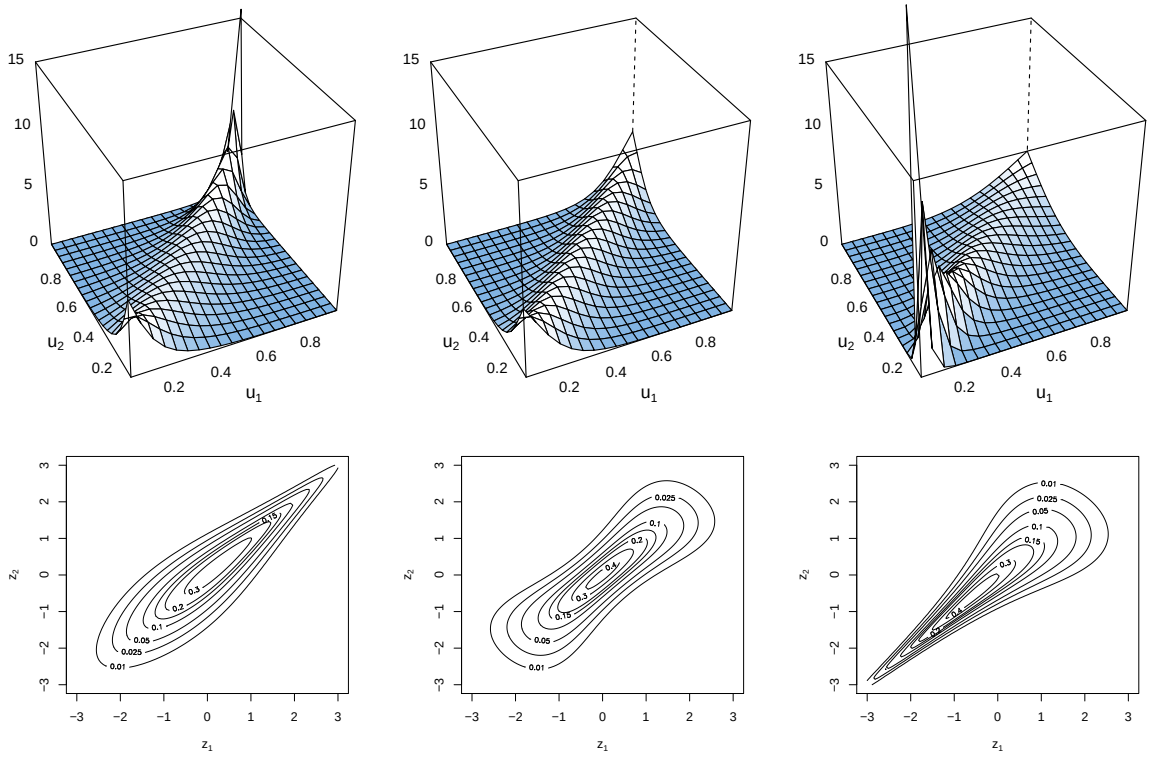

FIGURE 1.2: BIVARIATE ARCHIMEDEAN COPULA DENSITIES: Top: three-dimensional density plots for theoretical Gumbel (left), Frank (middle) and Clayton (right) copulae. The parameters have been chosen in such a way that Kendall's $\tau = 0.7$. Below are the respective contour plots belonging to the densities.

Tail dependence

The concept of tail dependence arises from extreme value theory and quantifies the joint presence of either very small or very large values or both via a tail-dependence-coefficient. Given two random variables X_1 and X_2 it is defined by the limit

$$\lambda^{\mathrm{u}} = \lim_{q \to 1^-} P\left(X_2 > F_2^{-1}(q) \mid X_1 > F_1^{-1}(q)\right) = \lim_{q \to 1^-} \frac{1 - 2q + C(q,q)}{1 - q} \tag{1.13}$$

for upper tail dependence and

$$\lambda^{\mathrm{l}} = \lim_{q \to 0^+} P\left(X_2 \leq F_2^{-1}(q) \mid X_1 \leq F_1^{-1}(q)\right) = \lim_{q \to 0^+} \frac{C(q,q)}{q} \tag{1.14}$$

for lower tail dependence. The coefficient's support is on $0 \leq \lambda^{\mathrm{u}}, \lambda^{\mathrm{l}} \leq 1$ and can be understood as a measure of strength, getting closer to one, the closer the link between X reaching large values and Y reaching large values likewise. A more detailed introduction into extreme value theory can be found in Chapter 7 of McNeil, Frey, and Embrechts (2015).

Kendall's Tau

Kendall's Tau (Kendall, 1938) is a rank based correlation coefficient that splits the present data into concordant or discordant pairs and calculates its difference. Suppose there is a set of x-sorted observations (x_i, y_i) for $i \in \{1, \ldots, n\}$ and $x_1 < x_2 < \cdots < x_n$, then two pairs of variables (x_i, y_i)

and (x_j, y_j) for $i < j$ are called *concordant* if either both $x_i > x_j$ and $y_i > y_j$ or $x_i < x_j$ and $y_i < y_j$, otherwise they are called *discordant*.

The product of their differences is summed up and divided by $\binom{n}{2} = \frac{n(n-1)}{2}$ to build the empirical τ-coefficient defined as

$$\hat{\tau} := \frac{2}{n(n-1)} \sum_{i<j} \operatorname{sgn}(x_i - x_j) \operatorname{sgn}(y_i - y_j) \tag{1.15}$$

with $\operatorname{sgn}(\cdot)$ being the signum function. $\hat{\tau}$ can take any value in $[-1, 1]$ where -1 and $+1$ imply perfect negative or positive dependence respectively. The provided formula above does not allow for ties where $\operatorname{sgn}(x_i - x_j) \operatorname{sgn}(y_i - y_j) = 0$. Kendall's Tau is directly connected to the bivariate copula distribution via (see Appendix A for proof)

$$\tau = -1 + 4 \int_{[0,1]^2} C(u_1, u_2) \, \mathrm{d}C(u_1, u_2). \tag{1.16}$$

The fact that Kendall's Tau is solely dependent on the copula allows us to estimate the copula parameter θ via the inverse of 1.16 such that $\tau(C_\theta) = \hat{\tau}$. This will be discussed in more detail when we introduce the copula estimators for our simulation in Chapter 2.

Blomqvist's Beta

An even simpler way to measure bivariate dependence is the use of *Blomqvist's Beta*, first described by Blomqvist (1950). The idea is to divide the data plane into four quadrants and count all respective observations via a 2×2 contingency matrix. From this Blomqvist's β is constructed as follows:

Let $\tilde{X}$ and $\tilde{Y}$ be the respective medians of two random variables $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$. To construct a useful measure of dependence, the $x - y$ plane is then divided into four parts by splitting it at $x = \tilde{X}$ and $y = \tilde{Y}$ and counting all points that lie in diagonally opposing quadrants. We then define n_1 as the number of points in the lower left and upper right region and n_2 as the number of points in the upper left and lower right region to construct the empirical Blomqvist's Beta

$$\hat{\beta} = \frac{n_1 - n_2}{n_1 + n_2} = \frac{2n_1}{n_1 + n_2} - 1, \tag{1.17}$$

where $-1 \leq \hat{\beta} \leq 1$. For an odd sample size, either one or two points must lie on one or both segmenting lines $x = \tilde{X}$ and $y = \tilde{Y}$ respectively, which raises the question to which quadrant the point will be counted. Blomqvist suggests that for one point lying on a median line, it shall be ignored, while in the case of two points lying on both segmenting lines, one is ignored and the second is counted to the quadrant both points adjoin.

From the population analogue of Blomqvist's Beta

$$\beta = P\{(X - \tilde{X})(Y - \tilde{Y}) > 0\} - P\{(X - \tilde{X})(Y - \tilde{Y}) < 0\}, \tag{1.18}$$

one can derive the Copula representation (done in Appendix A) which is given by

$$\beta = 4C\left(F_X(\tilde{X}), F_Y(\tilde{Y})\right) - 1 = 4C\left(\frac{1}{2}, \frac{1}{2}\right) - 1 \tag{1.19}$$

This representation allows us estimate the copula parameter θ by solving the equation $\beta(C_\theta) = \hat{\beta}$. For both inversion methods it is assumed that $C \in C_\theta$ and $\theta \in \mathbb{R}$.

2 Copula selection and parameter estimation

2.1 Estimation methods

The simulation in this article uses three different estimators to select and parameterize bivariate copulae. The first one will be the classical maximum likelihood estimator, while the last two are the inverses of the just presented moment-based functions, Kendall's Tau and Blomqvist's Beta. A similar study just for the parameter estimation process has already been done by Genest, Carabarín-Aguirre, and Harvey (2013), finding that the ML-estimator is the most precise in every application, yet it is slower. This is congruent to other studies comparing maximum likelihood to several other estimators, for example Weiß (2011).

The following section details how the just presented estimators operate and how the estimation process will take place. In a first step I will present how the parameter describing a copula distribution is estimated given that the true copula is known, and in a second step I will use a selection criterion in combination with the parameter to select the best copula out of a range. For this step I suppose that the true copula is not known.

2.1.1 Maximum likelihood

At the beginning of a bivariate estimation process one needs to specify the marginal distributions of the available data. Suppose an i.i.d sample of bivariate data $\{x_{i1}, x_{i2}\}$ for $i = 1, \ldots, n$. If the marginal distribution is known the copula data can be directly obtained using the probability integral transform

$$(u_{i1}, u_{i2}) := (F_1(x_{i1}), F_2(x_{i2})) \qquad \text{for } i = 1, \ldots, n. \tag{2.1}$$

In real-world data applications, however, margins are rarely known, so Oakes (1994) suggested replacing them with their empirical versions

$$\hat{F}_j(x_j) = \frac{1}{n+1} \sum_{i=1}^{n} \mathbf{1}_{(x_{ij} \leq x_j)} \qquad \text{for } j = \{1, 2\}, \tag{2.2}$$

with $\mathbf{1}_{(E)}$ being the indicator function for event E. To avoid boundary problems in further calculations as $\hat{F}(x_{nj}) \approx 1$, it is convenient to use the re-scaled version of Genest, Ghoudi, and Rivest (1995) which you see in the equation above. Here, the factor $\frac{1}{n+1}$ replaces the problematic $\frac{1}{n}$ in the old version.

The generated observations using the empirical margins are called *pseudo observations* and for $n \to \infty$ the empirical margin estimator converges towards the real marginal distribution. However, even if misspecified marginal distributions are used, the overall effect on model validity is insignificant as long as the margin of error does not grow too large (Kim, Silvapulle, and Silvapulle, 2007).

To estimate parameters via maximum likelihood we require the copula to be an element of all parameterized copulae $C \in (C_\theta)$ and the parameter θ to be real valued. If the copula for the observations $\{x_{i1}, x_{i2}\}$ has a density for all $u_1, u_2 \in (0,1)$ given by

$$c_\theta(u_1, u_2) = \frac{\partial^2}{\partial u_1 \partial u_2} C_\theta(u_1, u_2),$$

the log-likelihood function for θ, given known margins is defined as

$$\ell(\theta) = \sum_{i=1}^{n} \log \left[c_\theta \left\{ F_1 \left(x_{i1} \right), F_2 \left(x_{i2} \right) \right\} \right]. \tag{2.3}$$

In the more practical case of unknown margins, they are, as mentioned above, replaced by their empirical versions to yield

$$\ell(\theta) = \sum_{i=1}^{n} \log \left[c_\theta \left\{ \hat{F}_1 \left(x_{i1} \right), \hat{F}_2 \left(x_{i2} \right) \right\} \right], \tag{2.4}$$

where

$$\hat{\theta}_n^{ML}(u_1, u_2) \equiv \arg\max_{\theta \in \Theta} \ell(\theta). \tag{2.5}$$

As Genest, Ghoudi, and Rivest (1995) showed, the maximum likelihood estimator is consistent and asymptotically normal under regularity conditions.

2.1.2 Inverse of Kendall's Tau

Following equation 1.16 there is a direct connection between the copula distribution and the Kendall rank correlation coefficient. For the inversion method we choose a copula family that has an explicit relationship between its parameter θ (or ρ for elliptical copulae) and Kendall's τ such that

$$\tau = \delta(\theta),$$

with $\delta(\cdot)$ being the respective copula-tau relationship function calculated from 1.16. From this we can directly derive an estimator for θ by inverting the relationship function and using the empirical version of τ as its argument, yielding

$$\hat{\theta}_\tau := \delta^{-1}(\hat{\tau}). \tag{2.6}$$

For the copulae introduced in Chapter 1 the relationship function and their inverses can be found in Table 2.1.

2.1.3 Inverse of Blomqvist's Beta

As with the inversion of Kendall's Tau there also exists a direct connection between the copula parameter θ or ρ and Blomqvist's Beta via equation 1.19. The major difference between β-inversion and τ-inversion is the algorithmic complexity. While the first has a linear complexity of $\mathcal{O}(n)$, the latter has quadratic complexity $\mathcal{O}(n^2)$ (see Genest, Carabarín-Aguirre, and Harvey, 2013) and is thus computationally more challenging.

To derive an estimator for Blomqvist's Beta, the same process is used as for the inversion of Kendall's Tau. At first, a one to one relation of the form

Copula family	$\delta(\theta)$ or $\delta(\rho)$	$\delta^{-1}(\hat{\tau})$
Gaussian	$\tau = \frac{2}{\pi}\arcsin(\rho)$	$\hat{\theta}_\tau = \sin(\hat{\tau}\frac{\pi}{2})$
Students t	$\tau = \frac{2}{\pi}\arcsin(\rho)$	$\hat{\theta}_\tau = \sin(\hat{\tau}\frac{\pi}{2})$
Clayton	$\tau = \frac{\theta}{\theta+2}$	$\hat{\theta}_\tau = 2\frac{\hat{\tau}}{1-\hat{\tau}}$
Gumbel	$\tau = 1 - \frac{1}{\theta}$	$\hat{\theta}_\tau = \frac{1}{1-\hat{\tau}}$
Frank	$\tau = 1 - \frac{4}{\theta} + 4\frac{D_1(\theta)}{\theta}$ with $D_1(\theta) = \int_0^\theta \frac{x/\theta}{e^x-1}dx$	no closed form expression

TABLE 2.1: Relationship functions for selected copulae and their inverses. For the Gaussian and t-copula, the dependence parameter is called ρ and will, therefore be the argument of its function.

Copula family	$\varrho(\theta)$ or $\varrho(\rho)$	$\varrho^{-1}(\hat{\beta})$
Gaussian	$\beta = \frac{2}{\pi}\arcsin(\rho)$	$\hat{\theta}_\beta = \sin(\hat{\beta}\frac{\pi}{2})$
Students t	$\beta = \frac{2}{\pi}\arcsin(\rho)$	$\hat{\theta}_\beta = \sin(\hat{\beta}\frac{\pi}{2})$
Clayton	$\beta = -1 + 4(2^{\theta+1}-1)^{-1/\theta}$	no closed form expression
Gumbel	$\beta = -1 + 4\exp\left[-\log(2)\cdot 2^{1/\theta}\right]$	$\hat{\theta}_\beta = -\frac{\log(2)}{\log\left(-\frac{\log(2)}{\log\left(\frac{\hat{\beta}+1}{4}\right)}\right)}$ only for $\hat{\beta} \in [0,1)$
Frank	$\beta = \frac{4}{\theta}\log\cosh\frac{\theta}{4}$	no closed form expression

TABLE 2.2: Relationship functions for selected copulae and their inverses for the inversion of Blomqvist's Beta. For the Gaussian and t-copula, the dependence parameter is called ρ and will, therefore be the argument of its function.

$$\beta = \varrho(\theta),$$

is used to then find an inverse that acts as an estimator for the copula parameter given an estimate of Blomqvist's Beta

$$\hat{\theta}_\beta := \varrho^{-1}(\hat{\beta}). \tag{2.7}$$

As above I will provide a table with all relationship functions and their respective inverses used in this simulation study i.e. Table 2.2. Notice that for the Gumbel estimator $\hat{\theta}_\beta$ there is no analytic inverse for the complete range of $\hat{\beta} \in [-1,1]$ so I use a pseudo-inverse which is only well defined for $\hat{\beta} \in [0,1)$. For the implementation of this simulation the pseudo-inverse can be used without problems as only positive dependencies and, thus, parameters are estimated. The inverse relationship functions which do not have closed form expressions will be inverted numerically.

2.2 Copula selection strategy

So far I described the process of finding a suitable parameter for known data and a known copula. In real-world applications, however, the underlying or true copula is rarely, if ever, known so one needs a process that selects an appropriate copula from a given set of data $\{x_{i1}, x_{i2}\}$ for $i = 1, \ldots, n$.

In a first step the data needs to be transformed to the copula scale $u_{i1}, u_{i2} \in [0,1]$ via the probability integral transform. This can be done either via equation 2.1 for known margins or via equation 2.2 for unknowns. In my simulation this step is skipped as I will derive the observations directly from a copula.

In the next step, the transformed data is fitted to every available copula as described in section 2.1. This yields one parameter per copula per estimator. To now find the model with the best fit based upon the estimates, I use the Akaike Information Criterion (AIC) (Akaike, 1973) defined as:

$$AIC := -2 \sum_{i=1}^{n} \log \left[c \left(u_{i1}, u_{i2} \mid \hat{\boldsymbol{\theta}} \right) \right] + 2p, \tag{2.8}$$

with p being the number of parameters; for this study $p = 1$ as only one-parameter copulae are estimated.

Since both inversion methods do not depend on likelihood functions when estimating parameters, the AIC as a comparing selection criterion can only be used by plugging the inversion-estimates back into the log-likelihood function found for the maximum likelihood procedure. Thus, an AIC for every estimator can be obtained by setting:

$$\hat{\boldsymbol{\theta}} := \begin{cases} \hat{\theta}_n^{ML} & \text{, for ML} \\ \delta^{-1}(\hat{\tau}) & \text{, for Kendall's Tau inversion} \\ \varrho^{-1}(\hat{\beta}) & \text{, for Blomqvist's Beta inversion.} \end{cases} \tag{2.9}$$

In a last step the copula with the smallest AIC value per estimator is selected. By means of the AIC, the so chosen model is the most parsimonious.

In this simulation I will record the relative computation time it takes for every estimator to select a copula from the data on the basis of AIC. I expect the inversion methods to be significantly faster than ML as they skip the optimization process (equation 2.5) for ML, which is computationally challenging, especially with large sample sizes.

3 Simulation study

A sizable Monte-Carlo simulation study has been carried out to test the performance, accuracy and computational efficiency of the three presented estimators (maximum likelihood, Kendall's Tau inversion and Blomqvist's Beta inversion) in terms of copulae parameter estimation and selecting an appropriate copula from data. To measure parameter-estimation-accuracy, the mean-square-error (MSE) for every parameter estimate per estimator is calculated and averaged over a large number of experiments. The performance is determined via relative computation time it takes for my algorithm to converge. I do not use absolute values, because they carry no additional information about computational performance but much rather about the specific hardware I am using.

In terms of selecting one copula for a given set of data, relative computation time for my algorithm to converge to one copula is measured per estimator. The validity of its choice is also denoted.

3.1 Procedure

The design of this simulation study will base on several other studies that measure performance of different parameter estimators for copulae in a Monte-Carlo manner such as Kim, Silvapulle, and Silvapulle (2007), Kojadinovic and Yan (2010), Weiß (2011) or Genest, Carabarín-Aguirre, and Harvey (2013).

For every parametric copula presented above the following steps are repeated k times in sequence fr the parameter estimation:

1. A sample of size n is simulated from a theoretical copula $C \in (C_\theta)$ with true parameter θ.
2. Compute the parameter estimates with the maximum-likelihood and each of the moment-based inverses under the premise that the parametric form of the copula is known.
3. Calculate the average MSE per parameter by $MSE(\hat{\boldsymbol{\theta}}) \equiv k^{-1} \sum_{i=1}^{k} E\left[\theta - \hat{\boldsymbol{\theta}}_i\right]^2$, with θ being the true parameter and $\hat{\boldsymbol{\theta}}$ being the estimated parameter vector from equation 2.9 for every iteration.
4. Provide relative computation time per estimator per copula, where median maximum-likelihood time is set to 100 ($\tilde{t}_{ML} = 100$). All benchmarking exercises are done using the `microbenchmark` package.

For this study I will set $k = 1000$ and $n = \{30, 50, 100\}$ similar to other studies provided above. The optimization process for maximum-likelihood uses a bounded, limited-memory BFGS algorithm, which has been proven to be an optimal choice for maximum entropy problems (Malouf, 2002). I will test the whole process for a range of parameter values. For the elliptical copulae I use $\rho \in \{-0.9, -0.8 \ldots 0.8, 0.9\}$ while for the archimedean copulae I use $\theta \in \{1, 1.5 \ldots 9.5, 10\}$. I will set the degrees of freedom for the t-copula fix to $\nu = 4$ as it drastically increases computational speed since the estimation process for the degrees of freedom can be circumvented (Lucas, Schwaab, and Zhang, 2014).

The copula selection process will follow a similar fashion as the parameter estimation. The following process is repeated k times:

1. A sample of size n is simulated from a known copula $C \in (C_\theta)$ with known parameter θ.
2. The sample is fitted to every copula available for each estimator and respective AIC values are computed.
3. Copula with lowest AIC per estimator is chosen. Correct selections over k are returned ($\frac{\#\text{of correct selections}}{k}$).
4. Relative computation times per copula and estimator are computed. Again done with the `microbenchmark` package.

During the copula selection procedure, I use $k = 100$ and $n = \{30, 100\}$. As with the parameter estimation, the range of parameters per copula and estimator will remain the same.

The simulation is performed on version `4.0.3` of `R`. The copula observations are simulated using the `BiCopSim` function from the `VineCopula` package.

3.2 Results

The results of the parameter estimation study are shown in Figures 3.1, 3.2 and 3.3. Relative computation time can be found in table 3.1. The figures show the average mean-square-error for every estimator and parameter for 1000 repetitions per parameter step. The sample size $n = 50$ is omitted in the figures as the result is congruent with the evidence showing that accuracy increases with sample size for all estimators.

3.2.1 Estimator choice for parameter estimation

Interestingly, the results show a relatively uniform behavior for the inversion of Kendall's Tau and maximum likelihood estimators. Regardless of the magnitude of dependence, both estimators show very small MSEs across all parameters with Tau's inversion being slightly inferior in high dependence scenarios for archimedean copula where $\theta > 6$.

The relative computation times from Table 3.1 reveal that, although being approximately equally accurate, Kendall's Tau inversion can be more than ten times faster than maximum likelihood. This outcome occurs when the density functions of the respective copulae are computationally challenging as for example with the t-copula. Vice versa, in scenarios where there is no closed form relationship function between the copula and its dependence measure, computation time is almost equal, i.e Tau inversion for the Frank copula still needs 92.3% of the time it takes ML to converge.

Blomqvist's Beta inversion is not recommended to use for parameter estimation under any circumstance. It may be faster to compute in some situations, however, the large margin of error can make predictions error-prone and unpredictable. This is especially evident in extreme value copulae such as the Gumbel copula, with MSE values exceeding ten.

3.2.2 Estimator choice for copula selection

Figures 3.4 and 3.5 depict the accuracy of each estimator copula and parameter for the selection process based the AIC for $n = \{30, 100\}$ and $k = 100$. Surprisingly, all estimators seem to perform equally in terms of selecting a copula, given a set of data. The difference in computation time between the estimators from Table 3.2, however, is noteworthy. Even for the smallest sample size $n = 30$, there is no notable difference in accuracy across all copulae and parameters, yet the inversion methods are

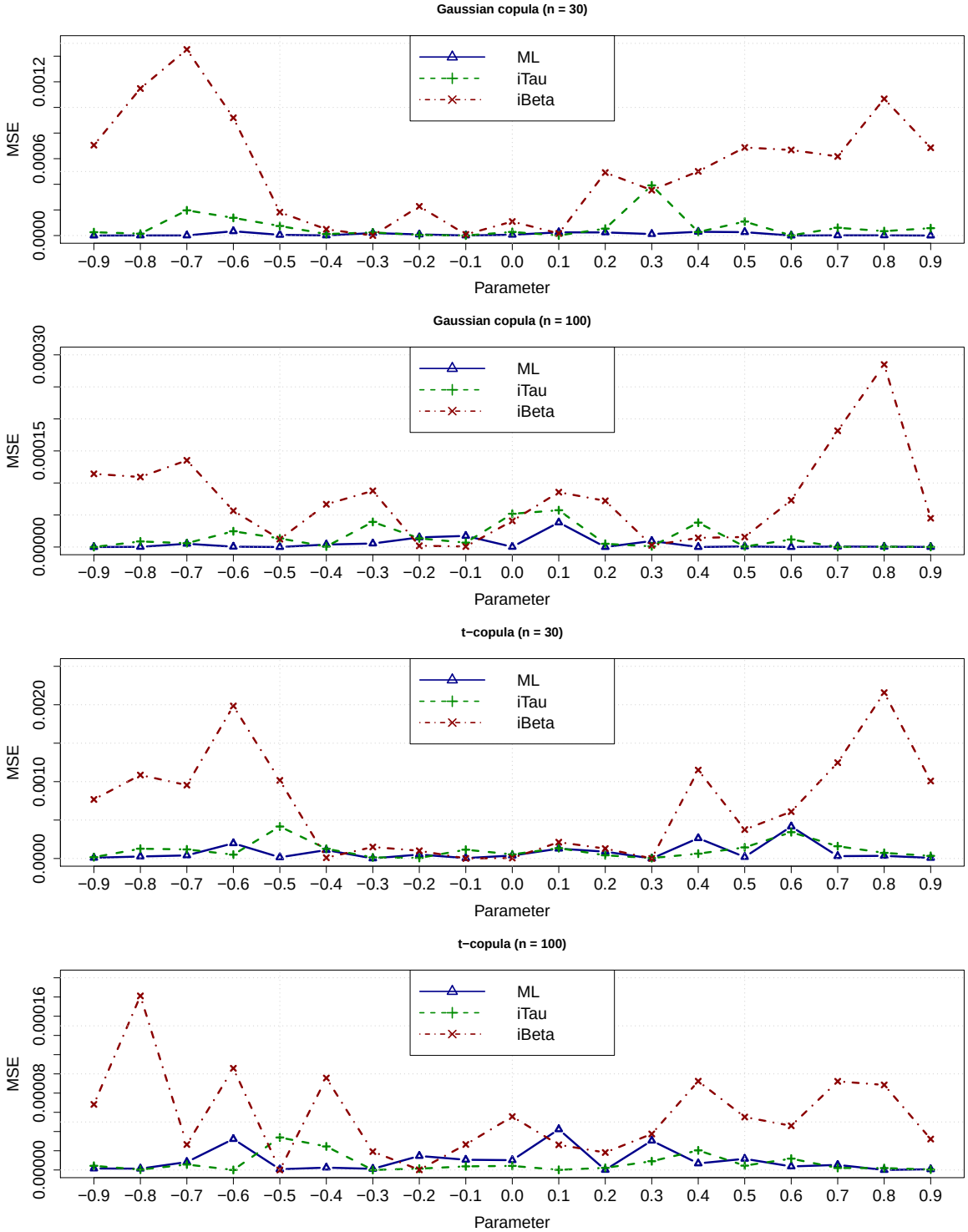

FIGURE 3.1: Mean square error for the maximum-likelihood (ML) and both inversion methods for different copulae and parameters. Sample size alternates between $n = \{30, 100\}$ and $k = 1000$. It is assumed here that the simulated data comes from a true known copula.

at least 16 times faster than maximum-likelihood. This is most likely due to skipped optimization process for the inversion methods, as no first- or second-order derivatives need to be calculated.

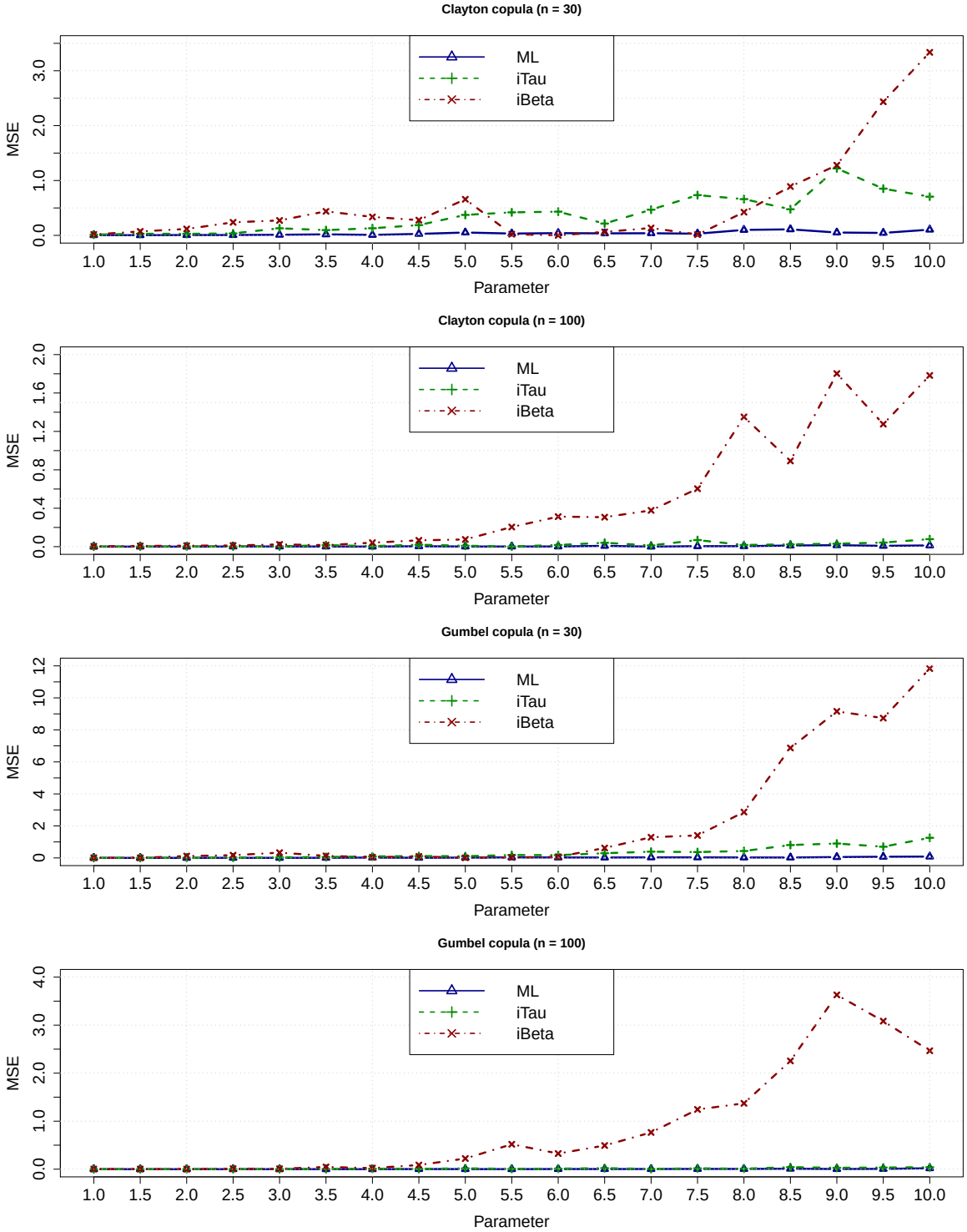

FIGURE 3.2: Continuation of Figure 3.1 for the Clayton and Gumbel copula.

3.3 Conclusion

This study compared the performance and accuracy of three different estimators regarding parameter estimation and selection of copulae. Concerning the first one, the results of previous studies could be reproduced, finding that for accurate estimation of a copula's parameter the maximum likelihood estimator remains the mean of choice. Both moment based estimators, especially Blomqvist's Beta are significantly worse in estimating, however, are considerably faster computationally. Hence,

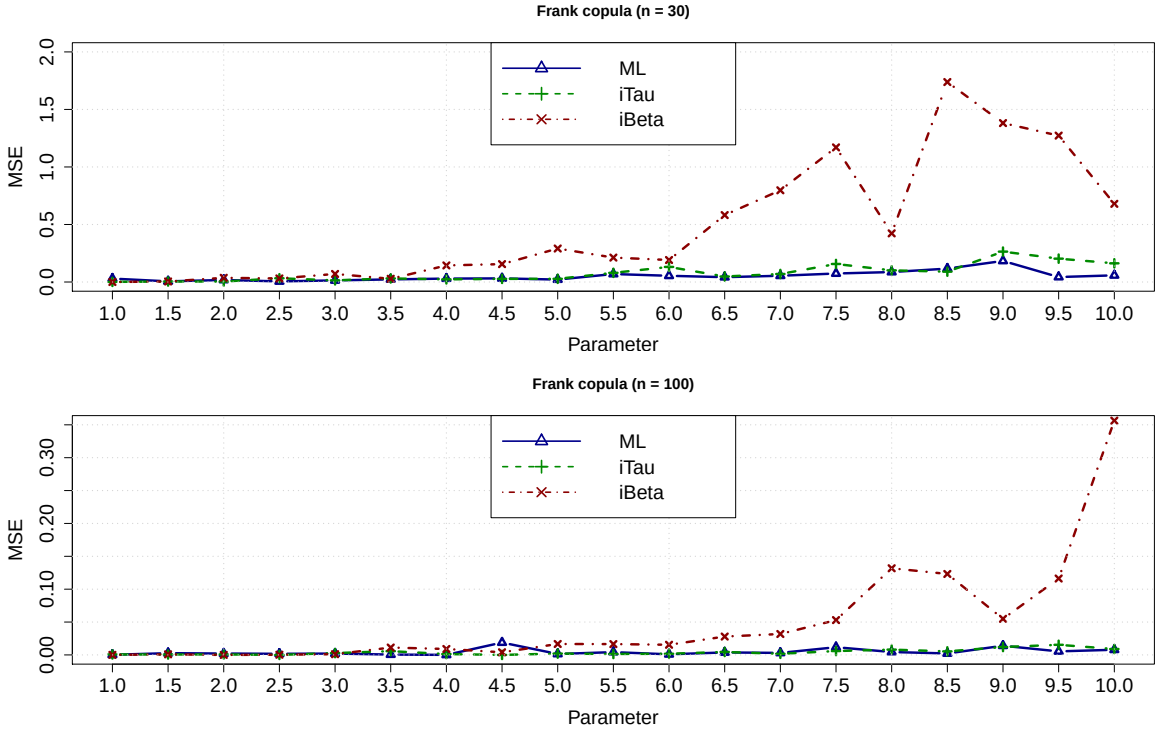

FIGURE 3.3: Continuation of 3.1 and 3.2 for the Frank copula.

Copula family	Estimator	$\min_t$	$\tilde{t}$	$\max_t$
Gaussian	ML	0.996	1.000	1.051
	iTau	0.146	0.147	0.148
	iBeta	0.182	0.183	0.185
t	ML	0.989	1.000	1.045
	iTau	0.032	0.032	0.033
	iBeta	0.035	0.035	0.036
Clayton	ML	0.992	1.000	1.006
	iTau	0.152	0.155	0.157
	iBeta	0.247	0.248	0.251
Gumbel	ML	0.999	1.000	1.002
	iTau	0.113	0.114	0.114
	iBeta	0.139	0.140	0.141
Frank	ML	0.997	1.000	1.006
	iTau	0.922	0.923	0.925
	iBeta	0.909	0.910	0.917

TABLE 3.1: Relative computation time per copula and estimator. For the benchmark I calculated the average time it takes the algorithm to converge to a parameter. I use $n = 100$ and $k = 5000$.

one could use those estimators to perform a quick-and-dirty estimation to find a starting value for a subsequent ML estimation. This poses the interesting question of how much more efficient such a procedure would be in comparison to only use ML.

Copula family	Estimator	$\min_t$	$\tilde{t}$	$\max_t$
Gaussian	ML	0.666	1	3.706
	iTau	0.051	0.053	0.115
	iBeta	0.050	0.051	0.112
t	ML	0.647	1	3.576
	iTau	0.050	0.051	0.159
	iBeta	0.048	0.049	.0125
Clayton	ML	0.606	1	2.936
	iTau	0.040	0.041	0.090
	iBeta	0.039	0.040	0.084
Gumbel	ML	0.643	1	2.638
	iTau	0.059	0.061	0.086
	iBeta	0.058	0.059	0.104
Frank	ML	0.670	1	3.288
	iTau	0.044	0.045	0.100
	iBeta	0.042	0.043	0.076

TABLE 3.2: Relative computation time per copula and estimator for the selection process. Time is measured relative to the median time it takes the maximum likelihood estimator in combination with the AIC to find an optimal copula. n has been set to 100 and $k = 1000$.

In a second step a fast procedure for selecting copulae could be found. The accuracy when selecting copulae remains almost equal across all copulae, parameters and sample sizes for every estimator. In terms of algorithmic complexity, the moment-based estimators stand out, as they only need a fraction of the time ML needs to select a copula from arbitrarily generated data. Therefore, it seems reasonable to prefer those over ML, especially in large sample size or big-data environments. The `R` workspace and functions can be provided upon request.

With over $855,000$ parameter estimations and around $57,000$ copula selections computed in this study, it counts to the more comprehensive ones carried out so far, yet there are also drawbacks. I only implemented the selection process for five copulae and three estimators, so it would be an interesting task to realize the selection algorithm for the remaining copulae and also test its performance. A real-world application putting the algorithm to use is also pending.

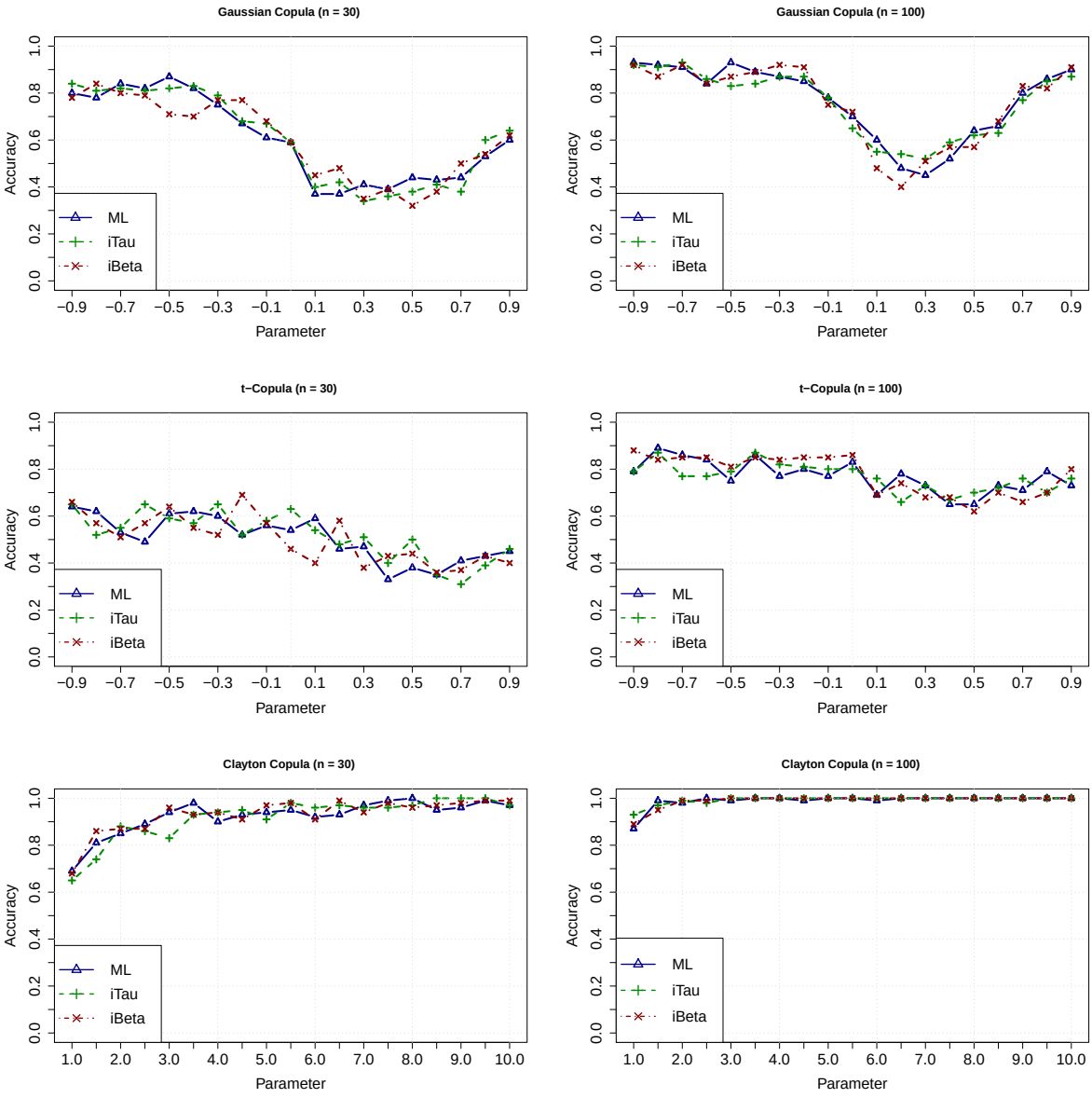

FIGURE 3.4: Accuracy of copula selection for different estimators with $n = \{30, 100\}$ and $k = 100$. Shown is the percentage of correct specifications per estimator, copula and parameter. This Figure images the behavior for the Gaussian, t, and Clayton copula.

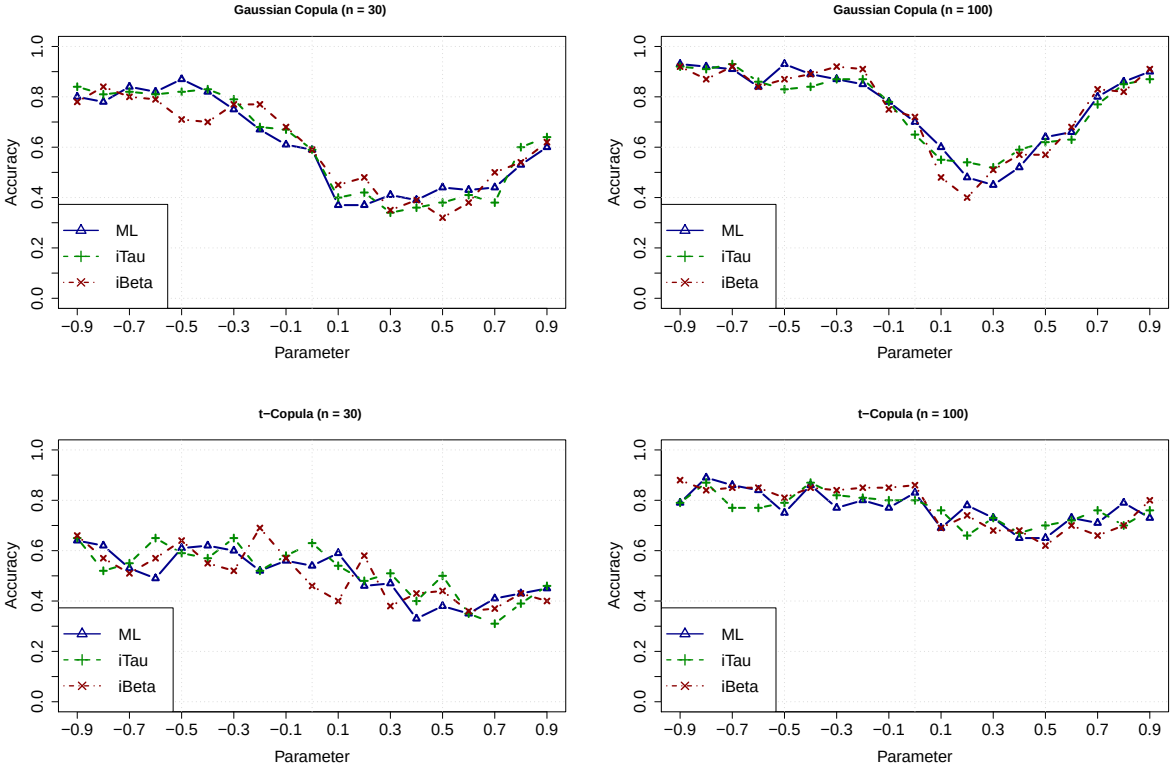

FIGURE 3.5: Continuation of Figure 3.4 for the Gumbel and Frank copula.

A Appendix

A.1 Proof of 1.16

In 1.15 we discussed an empirical version of Kendall's Tau. For this proof we use the population analogue, given by

$$\tau\left(X_1, X_2\right) = P\left[\left(X_{11} - X_{21}\right)\left(X_{12} - X_{22}\right) > 0\right] - P\left[\left(X_{11} - X_{21}\right)\left(X_{12} - X_{22}\right) < 0\right],$$

where (X_{11}, X_{12}) and (X_{21}, X_{22}) are independent copies of (X_1, X_2). Now, based on

$$P\left[\left(X_{11} - X_{21}\right)\left(X_{12} - X_{22}\right) > 0\right] = 1 - P\left[\left(X_{11} - X_{21}\right)\left(X_{12} - X_{22}\right) < 0\right],$$

one can write $\tau = 2P\left(\left(X_{11} - X_{21}\right)\left(X_{12} - X_{22}\right) > 0\right) - 1$. Furthermore, using

$$P[(X_{11}-X_{21})(X_{12}-X_{22})>0]=P(X_{11}>X_{21},X_{12}>X_{22})+P(X_{11}<X_{21},X_{12}<X_{22})$$

and the transformation $u_1 := F_1\left(x_1\right)$ and $u_2 := F_2\left(x_2\right)$ one obtains:

$$\begin{aligned} P\left(X_{11} > X_{21}, X_{12} > X_{22}\right) &= P\left(X_{21} < X_{11}, X_{22} < X_{12}\right) \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} P\left(X_{21} < x_1, X_{22} < x_2\right) dC\left(F_1\left(x_1\right), F_2\left(x_2\right)\right) \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} C\left(F_1\left(x_1\right), F_2\left(x_2\right)\right) dC\left(F_1\left(x_1\right), F_2\left(x_2\right)\right) \\ &= \int_0^1\int_0^1 C\left(u_1, u_2\right) dC\left(u_1, u_2\right). \end{aligned}$$

The same can be shown for

$$P\left(X_{11} < X_{21}, X_{12} < X_{22}\right) = \int_0^1\int_0^1 \left[1 - u_1 - v_1 + C\left(u_1, u_2\right)\right] dC\left(u_1, u_2\right),$$

and as C is the distribution function of $U_j := F_j(X_j)$ for $j = 1, 2$ with mean $1/2$ one obtains

$$\begin{aligned} P\left(X_{11} < X_{21}, X_{12} < X_{22}\right) &= 1 - \frac{1}{2} - \frac{1}{2} + \int_0^1\int_0^1 C\left(u_1, u_2\right) dC\left(u_1, u_2\right) \\ &= \int_0^1\int_0^1 C\left(u_1, u_2\right) dC\left(u_1, u_2\right) \end{aligned}.$$

Therefore,

$$\tau = -1 + 4\int_{[0,1]^2} C\left(u_1, u_2\right) \mathrm{d}C\left(u_1, u_2\right). \tag{A.1}$$

A.2 Proof of 1.19

This proof is taken from Genest, Carabarín-Aguirre, and Harvey (2013), but also repeated here for completeness. For two random variables X and Y with their respective medians $\tilde{X}$ and $\tilde{Y}$, Blomqvist's

Beta is given by

$$\beta = P\{(X - \tilde{X})(Y - \tilde{Y}) > 0\} - P\{(X - \tilde{X})(Y - \tilde{Y}) < 0\}.$$

From this we can directly show that

$$\begin{aligned} P\{(X - \tilde{X})(Y - \tilde{Y}) > 0\} &= P(X - \tilde{X} > 0, Y - \tilde{Y} > 0) + P(X - \tilde{X} < 0, Y - \tilde{Y} < 0) \\ &= P(X < \tilde{X}, Y < \tilde{Y}) + P(X > \tilde{X}, Y > \tilde{Y}) \end{aligned}$$

and by using the properties of the median

$$P(X > \tilde{X}, Y > \tilde{Y}) = P(X < \tilde{X}, Y < \tilde{Y})$$

and

$$P\{(X - \tilde{X})(Y - \tilde{Y}) > 0\} = 1 - P\{(X - \tilde{X})(Y - \tilde{Y}) < 0\}$$

it becomes evident that using Sklar's theorem in 1.1 we archive the Copula representation

$$\beta = 4C\left(\frac{1}{2}, \frac{1}{2}\right) - 1 \tag{A.2}$$

Bibliography

Akaike, H (1973). *Information theory and an extension of maximum likelihood principle*, pp. 267–281.

Angus, John E (1994). "The probability integral transform and related results". In: *SIAM review* 36.4, pp. 652–654.

Blomqvist, Nils (1950). "On a measure of dependence between two random variables". In: *The Annals of Mathematical Statistics*, pp. 593–600.

Breymann, Wolfgang, Alexandra Dias, and Paul Embrechts (2003). "Dependence structures for multivariate high-frequency data in finance". In.

Clayton, David G (1978). "A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence". In: *Biometrika* 65.1, pp. 141–151.

Demarta, Stefano and Alexander J McNeil (2005). "The t copula and related copulas". In: *International statistical review* 73.1, pp. 111–129.

Durante, Fabrizio, Juan Fernandez-Sanchez, and Carlo Sempi (2013). "A topological proof of Sklar's theorem". In: *Applied Mathematics Letters* 26.9, pp. 945–948.

Genest, Christian, Alberto Carabarín-Aguirre, and Fanny Harvey (2013). "Copula parameter estimation using Blomqvist's beta". In: *Journal de la Société Française de Statistique* 154.1, pp. 5–24.

Genest, Christian, Kilani Ghoudi, and L-P Rivest (1995). "A semiparametric estimation procedure of dependence parameters in multivariate families of distributions". In: *Biometrika* 82.3, pp. 543–552.

Genest, Christian and Louis-Paul Rivest (1993). "Statistical inference procedures for bivariate Archimedean copulas". In: *Journal of the American statistical Association* 88.423, pp. 1034–1043.

Kendall, Maurice G (1938). "A new measure of rank correlation". In: *Biometrika* 30.1/2, pp. 81–93.

Kim, Gunky, Mervyn J Silvapulle, and Paramsothy Silvapulle (2007). "Comparison of semiparametric and parametric methods for estimating copulas". In: *Computational Statistics & Data Analysis* 51.6, pp. 2836–2850.

Kojadinovic, Ivan and Jun Yan (2010). "Comparison of three semiparametric methods for estimating dependence parameters in copula models". In: *Insurance: Mathematics and Economics* 47.1, pp. 52–63.

Lucas, André, Bernd Schwaab, and Xin Zhang (2014). "Conditional euro area sovereign default risk". In: *Journal of Business & Economic Statistics* 32.2, pp. 271–284.

Malouf, Robert (2002). *A comparison of algorithms for maximum entropy parameter estimation*.

Mashal, Roy, Marco Naldi, and Assaf Zeevi (2003). "On the dependence of equity and asset returns". In: *RISK-LONDON-RISK MAGAZINE LIMITED-* 16.10, pp. 83–88.

McNeil, Alexander J, Rüdiger Frey, and Paul Embrechts (2015). *Quantitative risk management: concepts, techniques and tools-revised edition*. Princeton university press.

McNeil, Alexander J, Johanna Nešlehová, et al. (2009). "Multivariate Archimedean copulas, d-monotone functions and L1-norm symmetric distributions". In: *The Annals of Statistics* 37.5B, pp. 3059–3097.

Nelsen, Roger B (2007). *An introduction to copulas*. Springer Science & Business Media.

Oakes, David (1994). "Multivariate survival distributions". In: *Journaltitle of Nonparametric Statistics* 3.3-4, pp. 343–354.

Salmon, Felix (2012). "The formula that killed Wall Street". In: *Significance* 9.1, pp. 16–20.

Sklar, Abe (1959). "Fonctions de répartition à n dimensions et leurs marges". In: *Publ. inst. statist. univ. Paris* 8, pp. 229–231.

Weiß, Gregor (2011). "Copula parameter estimation by maximum-likelihood and minimum-distance estimators: a simulation study". In: *Computational Statistics* 26.1, pp. 31–54.

Declaration of Authorship

I, Niklas PAULIG, declare that this thesis titled, "Speed check: copula selection and parameter estimation using non-parametric moment inverses and maximum likelihood" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:

Date: