Technische Universität München

Zentrum Mathematik

Stepwise estimation of D-Vines with arbitrary specified copula pairs and EDA Tools

Diplomarbeit
von
Jiying Luo

Themensteller/in:	Prof. Dr. Claudia Czado
Betreuer/in:	Dr. Carlos Almeida
Abgabetermin:	01.03.2010

Hiermit erkläre ich, dass ich die Diplomarbeit selbstständig angefertigt und nur die angegebenen Quellen verwendet habe.

Garching, den 07. Januar 2011

Acknowledgments

This diploma thesis would not have been possible without the help of many individuals.

First and foremost, my utmost gratitude to Prof. Dr. Claudia Czado , the Chair of Mathematical Statistics whose sincerity and support I will never forget. Prof. Czado made it possible to write this diploma thesis in the first place and offered expert guidance and mentoring during the whole process.

Furthermore, I am heartily thankful to Dr. Carlos Almeida who was always accessible and willing to help me to solve my problems. Without his encouragement, valuable knowledge and advices from the initial to the final level this thesis would not have been completed.

I would also like to thank all of those who helped and inspired me in any respect during the completion of the thesis, especially Mr. Schepsmeier and Mr. Brechmann, for providing helpful programs and advices.

Last but not the least, my deepest gratitude goes to the my beloved families for their understanding and endless love through my years of studies. I am indebted to my mother for her everlasting care and love.

Contents

1 Introduction **1**

2 Foundations **3**
2.1 Measuring dependence 3
2.2 Copula 6
2.3 Bivariate copula 10
2.4 Pair-copula decomposition and vines 15
2.5 Statistical hypothesis testing 18
2.6 Determination of the p-value using bootstrapping 20

3 Kendall's process and the λ function **22**
3.1 Revisiting Kendall's τ 22
3.2 Kendall's process and λ-function 24

4 Graphical methods **28**
4.1 Scatter plot 28
4.2 Contour Plots 29
4.3 K-plots 36

5 GOF tests **40**
5.1 Copula goodness of fit 40
5.2 Two tests based on empirical copula process 41
5.3 Two tests based on Kendall's transform 43
5.4 Simulation studies 47

6 Application 1 **70**
6.1 Learning data set 70
6.2 Assessment of dependence 71
6.3 Selecting an appropriate D-vine decomposition 72
6.4 Model selection 73
6.5 Final model 82

7 Application 2 **86**
7.1 Learning data set 86
7.2 Copula selection 88
7.3 Final model and discussion 101

8 Summary and conclusions **106**

Bibliography **108**

Chapter 1

Introduction

In the financial world, the concept of dependencies among different assets is the foundation of modern portfolio management, asset allocation and risk management. Understanding dependence structure between assets is important not only for the potential comovements of various asset returns but also for risk diversification and portfolio construction. However, multivariate data usually exhibits a complex pattern of dependence. One popular approach to construct high dimensional dependence is based on copulas. In the past decade, a large number of applications of copula theory (Sklar [1959]) have been proposed for modeling dependence in finance. The popularity of copulas results from its capability of isolating the dependence structure from the joint and marginal distributions. Embrechts et al. [2001] demonstrated the efficiency of copulas in insurance and market risk management and illustrated the importance of the choice of the copula for expected losses and portfolio returns.

Yet multivariate copulas for high-dimensional (larger than 2) data are complex in comparison to the class of bivariate copulas. Bedford and Cooke [2001] and Bedford and Cooke [2002] developed a graphical model, called *regular vines,* to decompose high-dimensional copula into a group of bivariate copulas. Aas et al. [2009] first appreciated the general construction principle for deriving multivariate copulas. This principle is called as *pair-copula construction* (PCC).

The class of regular vines is still very general. In this paper, our main work is based on the PCC with respect to a specific vine, called *D-vine*. Our goal is to develop tools used in exploratory data analysis (EDA), such as contour plots, λ-function and Goodness-of-Fit (GOF) tests. With these tools we are able to find appropriate models in a D-vine decomposition. The D-vine structure will first be identified with the help of Kendall's τ and the choice of decomposition will be determined by selected models. In the entire process, finding models and identifying D-vine structure are interconnected in each step.

This thesis is organized as follows. In Chapter 2 we review fundamental material such as dependence measures, definitions for statistical hypothesis testing, copulas and PCC. In the last section we introduce a bootstrap procedure which is the building block for GOF tests. With this background we introduce the Kendell's process and the λ-functions in Chapter 3. Both definitions are based on the dependence measure Kendall's τ. Kendall's process is used in GOF tests. In Chapter 4 we develop graphical tools based on bivariate copulas. Chapter 5 provides two kinds GOF tests, tests based on empirical copula process

and tests based on Kendall's transform. All tests depend on the choice of the copula family. In this chapter, the bootstrap technique and Kendall's process will be used. In financial markets, bond and stock are the two most important assets. In Chapter 6 and 7 we illustrate two application with different kind financial data. All methods and tests from Chapter 4 and 5 will be applied here. Finally, Chapter 8 summarizes, discusses results and considers further research.

Chapter 2

Foundations

2.1 Measuring dependence

In statistics, dependence is one of the most important instrument to study statistical relationships between two or more random variables or observed data values. In this section two well-known measures of dependence will be introduced: correlation and Kendall's τ. Many alternative measures of dependence have been proposed in statistical literature. e.g. Spearman's ρ. See Genest and Favre [2007].

Correlation coefficient

The Correlation coefficient is the most familiar measure of dependence between two variables X and Y, giving a value between -1 and $+1$ inclusive. It is widely used as a measure of the strength of linear dependence between two variables.

Definition 2.1 (Correlation coefficient)**.** The correlation coefficient between two variables X and Y is defined as the covariance of the two variables $\mathrm{Cov}(X, Y)$ divided by the product of their standard deviations $\sigma(X) = \sqrt{\mathrm{Var}(X)}$, $\sigma(Y) = \sqrt{\mathrm{Var}(Y)}$:

$$\rho(X, Y) := \frac{\mathrm{Cov}(X, Y)}{\sigma(X)\sigma(Y)}.$$

Remark 2.2. Properties of correlation coefficient:

(i) $|\rho(X, Y)| \leq 1$.

(ii) X and Y independent $\Rightarrow \rho(X, Y) = 0$.

(iii) $\exists a, b$ with $b \neq 0$ such that $P(X = a + bY) = 1 \Leftrightarrow |\rho(X, Y)| = 1$.

(iv) If X and Y have a joint bivariate normal distribution with standard normal margins, then the correlation coefficient ρ is uniquely defined by the joint distribution.

Definition 2.3 (Estimation of the correlation coeffcient)**.** The estimation of the covariances and variances between X and Y is based on a sample (x_i, y_i), $\forall i = 1, \ldots, n$. It can be estimated by

$$\hat{\rho} := \frac{\widehat{\mathrm{Cov}}(X,Y)}{\hat{\sigma}(X)\hat{\sigma}(Y)} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}\sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}},$$

where

$$\widehat{\mathrm{Cov}}(X,Y) := \frac{1}{n-1}\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}),$$

$$\hat{\sigma}(X) := \frac{1}{n-1}\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}, \quad \hat{\sigma}(Y) := \frac{1}{n-1}\sqrt{\sum_{i=1}^n (y_i - \bar{y})^2},$$

$$\bar{x} := \frac{1}{n}\sum_{i=1}^n x_i, \quad \bar{y} := \frac{1}{n}\sum_{i=1}^n y_i.$$

Kendall's τ

A second well-known measure of dependence is the Kendall rank correlation coefficient (or simply Kendall's τ), developed by Kendall [1938]. A rank correlation is to study relationships between rankings on the same set of items and the rank correlation coefficient measures the correspondence between two rankings. See also Joe [1997], Nelsen [2006].

Definition 2.4 (Kendall's τ)**.** Let (X_1, Y_1) be a pair of random variables with a joint distribution H and marginal distributions F and G. (X_2, Y_2) is an independent copy of (X_1, Y_1). Then the Kendall's τ for (X_1, Y_1) is defined by

$$\begin{aligned}
\tau(X_1, Y_1) &:= P((X_1 - X_2)(Y_1 - Y_2) > 0) - P((X_1 - X_2)(Y_1 - Y_2) < 0) \\
&= E[\mathrm{sgn}(X_1 - X_2)\mathrm{sgn}(Y_1 - Y_2)] \\
&= P(X_1 < X_2, Y_1 > Y_2) - P(X_1 > X_2, Y_1 < Y_2),
\end{aligned}$$

where sgn is the sign-function.

Definition 2.5 (Concordance and discordance)**.** Two pairs (x_i, y_i) and (x_j, y_j) are said to be concordant, if $(x_i - y_i)(x_j - y_j) > 0$ and discordant if$(x_i - y_i)(x_j - y_j) > 0$.

Hence τ denotes the difference between the probability of concordance and discordance of (X, Y).

Definition 2.6 (Estimation of Kendall's τ)**.** The empirical Kendall's τ is defined by

$$\hat{\tau}_n := \frac{P_n - Q_n}{\binom{n}{2}} = \frac{4}{n(n-1)}P_n - 1, \tag{2.1}$$

where P_n and Q_n are the number of concordant and discordant pairs.

It is obvious that τ and $\hat{\tau}_n$ are functions of the ranks of the observations, since $(X_i - Y_i)(X_j - Y_j) > 0$ if and only if $(R_i - S_i)(R_j - S_j) > 0$, where R_i stands for the rank of X_i among $X_1, \ldots, X_n$and S_i stands for the rank of Y_i among $Y_1, \ldots, Y_n$.

Remark 2.7. Properties of Kendall's τ and $\hat{\tau}_n$ are as follows:

(i) $\tau(X, Y)$ is symmetric, i.e. $\tau(X, Y) = \tau(Y, X)$.

(ii) $\tau(-X, Y) = \tau(X, -Y) = -\tau(X, Y)$.

(iii) $-1 \leq \tau(X, Y) \leq 1$, $\tau(X, X) = 1$ and $\tau(X, -X) = -1$.

(iv) If X and Y are independent, $\tau(X, Y) = 0$.

(v) $\tau(T_x(X), T_y(Y)) = \tau(X, Y)$ for all increasing maps T_x and T_y.

(vi) $\tau = 4 \int_{[0,1]^2} F(x, y) \mathrm{d}F(x, y) - 1$.

(vii) If the vector (X, Y) is bivariate normal distributed with correlation coefficient ρ and has standard normal distributed margins, then $\rho = \sin(\frac{\pi}{2}\tau)$.

(viii) The denominator $\binom{n}{2}$ in Eq. (2.1) is the total number of pairs, so the coefficient must be in the range $-1 \leq \hat{\tau}_n \leq 1$.

(ix) If X and Y are perfect positive dependent (i.e. all pairs are concordant), then $\hat{\tau}_n = 1$. If X and Y are perfect negative dependent (i.e. all pairs are discordant), then $\hat{\tau}_n = -1$.

A test of independence based on Kendall's τ

As a measure of dependence, a larger value of $|\hat{\tau}_n|$ indicates a stronger dependence between two factors. A test of independence can be based on $\hat{\tau}_n$, since under H_0 of independence $(\tau_n = 0)$ the statistic is close to $\mathcal{N}\Big(0, \frac{2(2n+5)}{9n(n-1)}\Big)$. Thus, H_0 would be rejected at level $\alpha = 0.05$ if

$$\sqrt{\frac{9n(n-1)}{2(2n+5)}}|\hat{\tau}_n| > 1.96 \ . \tag{2.2}$$

See Genest and Favre [2007] for details. If Z is a standard normal random variable, the P-value of the test based on $\hat{\tau}_n$ is $2\Big(1 - P\big(Z \leq \sqrt{\frac{9n(n-1)}{2(2n+5)}}|\hat{\tau}_n|\big)\Big)$.

Tail dependence

The Correlation coefficient ρ and Kendall's τ measure the global dependency structure of the joint distribution function H in the bivariate case. Thus, one can distinguish between different models by comparing their global dependency structures or equally using these two measures. Alternatively, one can also use their local behavior to distinguish different models, especially in extreme values.

In contrast to correlation coefficient ρ and Kendall's τ , tail dependence investigates local dependencies in the upper- and lower-quadrant tails of a bivariate joint distribution function. It is relevant to measure dependence in extreme values. The definitions of tail dependence are the following, which can be found in Joe [1997] and Embrechts et al. [2001] as well.

Definition 2.8 (Upper tail dependence)**.** Let (X, Y) be a vector of continuous random variables with marginal distribution F and G. The coefficient of upper tail dependence of (X, Y) is

$$\Lambda_U := \lim_{u \nearrow 1} P(Y > G^{-1}(u) | X > F^{-1}(u))$$

provided that the limit $\Lambda_U \in [0, 1]$ exists. If $\Lambda_U \in (0, 1]$, X and Y are said to be asymptotically dependent in the upper tail; if $\Lambda_u = 0$, X and Y are said to be asymptotically independent in the upper tail.

Since $P(Y > G^{-1}(u) | X > F^{-1}(u)) =$

$$\frac{1 - P(X \leq F^{-1}(u)) - P(Y \leq G^{-1}(u)) + P(X \leq F^{-1}(u), Y \leq G^{-1}(u))}{1 - P(x \leq F^{-1}(u))},$$

an alternative and equivalent definition with the help of copula function, from which it is seen that the concept of tail dependence is a copula property, is given as follows

Definition 2.9. If a bivariate copula C is such that

$$\Lambda_U := \lim_{u \nearrow 1} \frac{1 - 2u + C(u, u)}{1 - u}$$

exists, then C has upper tail dependence if $\Lambda_U \in (0, 1]$, and upper tail independence if $\Lambda_U = 0$.

The coefficient of lower tail dependence is similarly defined as follows:

Definition 2.10 (Lower tail dependence for copulas)**.** If a bivariate copula C is such that

$$\Lambda_L = \lim_{u \searrow 0} \frac{C(u, u)}{u}$$

exists. Then C has lower tail dependence if $\Lambda_L \in (0, 1]$, and lower tail independence if $\Lambda_L = 0$.

Table 2.1 lists the tail dependence properties of different copulas which are studied in this thesis.

2.2 Copula

Copulas are functions that join multivariate distribution functions to their one-dimensional marginal distribution functions. Alternatively, a copula is a multivariate distribution function defined on $[0.1]^n$, whose one-dimensional margins are uniform on the unit interval. More details about copulas can be found in Joe [1997] and Nelsen [2006].

Table 2.1: Tail dependence of the different copula families

Copula	Upper tail dependence	Lower tail dependence
Normal	0	0
T	$2t_{\nu+1}\big(-\sqrt{\nu+1}\big)\cdot\sqrt{\frac{1-\theta}{1+\theta}}$	$2t_{\nu+1}\big(-\sqrt{\nu+1}\big)\cdot\sqrt{\frac{1-\theta}{1+\theta}}$
Clayton	-	$2^{-1/\theta}$
Gumbel	$2-2^{1/\theta}$	-
Frank	-	-
BB1	$2-2^{1/\delta}$	$2^{-1/\theta\delta}$
BB7	$2-2^{1/\theta}$	$2^{-1/\delta}$

where $t_{\nu+1}$ is the univariate t distribution function with $\nu+1$ degrees of freedom.

The copula approach to dependence modeling is rooted in a representation theorem due to Sklar [1959]. Sklar's theorem describes the connection between the marginal distributions and the joint distribution. For example, let (X, Y) be any pair of continuous random variables with marginal distributions $F(X)$ and $G(Y)$. The joint cumulative distribution function (cdf) $H(x, y)$ of (X, Y) can be written using a copula $C : [0, 1]^2 \to [0, 1]$ in the form

$$H(x, y) = C\big(F(x), G(y)\big),$$

since the joint distribution function H contains the description of the dependency structure between variables and the description of their marginal behavior. The copula function is the joint distribution function of $U(0, 1)$ distributed random variables $U = F(x)$ and $V = G(Y)$, which means C does not depend on the marginal distribution of X and Y. Therefore, copula provide a way to isolate the dependency structure. The general Sklar's theorem stated by Joe [1997] is given as follows:

Theorem 2.11 (Sklar's Theorem)**.** *Let $H : \bar{\mathbb{R}}^n \to [0, 1]$ with $\bar{\mathbb{R}} = \mathbb{R} \cup \{-\infty, +\infty\}$ be an n-dimensional distribution function with margins $F_1, \ldots, F_n : \bar{\mathbb{R}} \to [0, 1]$. Then there exists an n-copula C such that for all x in $\bar{\mathbb{R}}^n$,*

$$H(x_1, \ldots, x_n) = C(F_1(x_1), \ldots, F_n(x_n)). \tag{2.3}$$

If $F_1, ..., F_n$ are all continuous, then C is unique; otherwise C is uniquely determined on the range of margins, $RanF_1 \times \cdots \times RanF_n$. Conversely, if C is an n-copula and $F_1, \ldots, F_n$ are distribution functions, then the function H defined above is an n-dimensional distribution function with margins $F_1, \ldots, F_n$.

As can be seen in Eq. (2.3), for continuous multivariate distribution functions the univariate margins and the multivariate dependency structure can be separated. The dependency structure is represented by a copula.

Let F be a univariate distribution function. We define the generalized inverse of $F(\cdot)$ as $F^{-1}(t) = \inf\{x \in \mathbb{R} | F(x) \geq t\}$ $\forall t \in [0, 1]$ with the convention that $\inf \emptyset = -\infty$.

Corollary 2.12. *Let $F_1, \ldots, F_n$, H and C be defined as in Theorem 2.11. For any $\mathbf{u} := (u_1, \ldots, u_n) \in [0, 1]^n$ we have*

$$C(u_1, \ldots, u_n) = H(F_1^{-1}(u_1), \ldots, F_n^{-1}(u_n)).$$

Now we pay close attention to an important feature of copulas, which is copulas with strictly monotone transformations of random variables. For strictly monotone transformations of the random variables, copulas are either invariant or change in predictable ways. We treat the case of strictly increasing transformations first. The copula associated with a random pair (X, Y) is invariant by monotone increasing transformations of the marginals.

Theorem 2.13. *Let $X_1, \ldots, X_n$ be continuous random variables with copula C. If $\alpha_1(\cdot)$, $\ldots$, $\alpha_n(\cdot)$ are strictly increasing functions whose domains contain the range of $X_1, \ldots, X_n$, RanX,..., RanY, respectively, then $C_{\alpha_1(X_1),\ldots\alpha_n(X_n)} = C.$, i.e. C is invariant under strictly increasing transformations of $X_1, \ldots, X_n$.*

Proof. Let $F_1, \ldots, F_n$ denote the distribution functions of $X_1, \ldots, X_n$ and let $G_1, \ldots, G_n$ denote the distribution functions of $\alpha_1(X_1), \ldots \alpha_1(X_1)$, respectively. Since α_i is strictly increasing, one finds

$$F_i(x) = P(\alpha_i(X_i) \leq x) = P(X_i \leq \alpha_i^{-1}(x)) = F_i(\alpha_i^{-1}(x)) \; \forall i = 1, \ldots, n.$$

Thus, for any $x \in \bar{\mathbb{R}}$,

$$\begin{aligned} C_\alpha(G_1(x_1), \ldots, G_n(x_n)) &:= C_{\alpha_1(X_1),\ldots\alpha_n(X_n)}(G_1(x_1), \ldots, G_n(x_n)) \\ &= P(\alpha_1(X_1) \leq x_1, \ldots, \alpha_n(X_n) \leq x_n) \\ &= P(X_1 \leq \alpha_1^{-1}(x_1), \ldots, X_n \leq \alpha_n^{-1}(x_n)) \\ &= C(F_1(\alpha_1^{-1}(x_1)), \ldots, F_n(\alpha_n^{-1}(x_n))) \\ &= C(G_1(x_1), \ldots, G_n(x_n)). \end{aligned}$$

Since $X_1, \ldots, X_n$ are continuous and $\text{Ran}G_i = [0, 1] \; \forall i = 1, \ldots, n$, we have $C_\alpha = C$ on $[0, 1]^n$. □

Theorem 2.14. *Let $X_1, \ldots, X_n$ be random variables with copula C and $\alpha_1(\cdot), \ldots, \alpha_n(\cdot)$ be strictly monotone on RanX , ..., RanX$_n$, respectively, and have copula $C_{\alpha_1(X_1),\ldots\alpha_n(X_n)}$. Furthermore let α_k be strictly decreasing for some k. Without loss of generality let $k = 1$.Then*

$$C_{\alpha_1(X_1),\ldots\alpha_n(X_n)}(u_1, u_2, \ldots, u_n) = C_{\alpha_2(X_2),\ldots\alpha_n(X_n)}(u_2, \ldots, u_n) - C_{X_1,\ldots\alpha_n(X_n)}(1 - u_1, u_2, \ldots, u_n).$$

Proof. Let $F_1 \ldots, F_n$ and $G_1, \ldots, G_n$ be defined as in the proof of Theorem 2.13. Define $u_i := G_i(x_i) \quad \forall i = 1, \ldots, n$. If α_1 is strictly decreasing and $\alpha_2, \ldots, \alpha_n$ are strictly increasing, then $G_k(x) = F_k(\alpha_k^{-1}(x)) \; \forall k = 2, \ldots, n$ and

$$\begin{aligned} G_1(x) &= P(\alpha_1(X_1) \leq x) \overset{\alpha_1 \text{ decreasing}}{=} P(X_1 > \alpha_1^{-1}(x)) \\ &= 1 - P(X_1 \leq \alpha_1^{-1}(x)) \\ &= 1 - F_1(\alpha_1^{-1}(x)). \end{aligned}$$

Thus for any $u_1, \ldots, u_n$,

$$\begin{aligned}
C_{\alpha_1(X_1),\alpha_2(X_2),\ldots,\alpha_n(X_n)}(u_1,\ldots,u_n) &= C_{\alpha_1(X_1),\alpha_2(X_2),\ldots,\alpha_n(X_n)}(G_1(x_1),\ldots,G_n(x_n)) \\
&= P(\alpha_1(X_1) \le x_1, \ldots, \alpha_n(X_n) \le x_n) \\
&= P(X_1 > \alpha_1^{-1}(x_1), X_2 \le \alpha_2^{-1}1(x_2), \ldots, X_n \le \alpha_n^{-1}(x_n)) \\
&= P(X_2 \le \alpha_2^{-1}(x_2), \ldots, X_n \le \alpha_n^{-1}(x_n)) \\
&\quad -P(X_1 \le \alpha_1^{-1}(x_1), \ldots, X_n \le \alpha_n^{-1}(x_n)) \\
&= C_{\alpha_2(X_2),\ldots\alpha_n(X_n)}(G_2(x_2),\ldots,G_n(x_n)) \\
&\quad -C_{X_1,\ldots\alpha_n(X_n)}(F_1(\alpha_1^{-1}(x_1)), G_2(x_2),\ldots,G_n(x_n)) \\
&= C_{\alpha_2(X_2),\ldots\alpha_n(X_n)}(G_2(x_2),\ldots,G_n(x_n)) \\
&\quad -C_{X_1,\ldots\alpha_n(X_n)}(1-G_1(x_1)), G_2(x_2),\ldots,G_n(x_n)) \\
&= C_{\alpha_2(X_2),\ldots\alpha_n(X_n)}(u_2,\ldots,u_n) \\
&\quad -C_{X_1,\ldots\alpha_n(X_n)}(1-u_1,u_2,\ldots,u_n).
\end{aligned}$$

□

Example 1. In the bivariate case, it follows that:

(i) If α is strictly decreasing and β is strictly increasing, then

$$C_{\alpha(X)\beta(Y)}(u,v) = u - C_{XY}(u, 1-v).$$

(ii) If α is strictly increasing and β is strictly decreasing, then

$$C_{\alpha(X)\beta(Y)}(u,v) = v - C_{XY}(1-u, v).$$

(iii) If α and β are both strictly decreasing, then

$$C_{\alpha(X)\beta(Y)}(u,v) = u + v - 1 + C_{XY}(1-u, 1-v).$$

These copulas are so-called rotated copula. More details will be given in the next section.

Definition 2.15 (Empirical copula)**.** Let $(X_{11},\ldots,X_{1p}),\ldots,(X_{n1},\ldots,X_{np})$ be random samples. Then the *empirical copula function* is defined by

$$C_n(u_1,\ldots,u_p) := \frac{1}{n}\sum_{i=1}^{n}\prod_{j=1}^{p}\mathbf{1}\{F_{j,n}(X_{ij}) \le u_j\}, \tag{2.4}$$

where $F_{j,n}(x_j) := \dfrac{1}{n+1}\sum_{i=1}^{n}\mathbf{1}\{X_{ij} \le x_j\}, \quad 1 \le j \le p$. In the bivariate case,

$$C_n(u,v) := \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}\{F(X) \le u, G(Y) \le v\}. \tag{2.5}$$

2.3 Bivariate copula

Bivariate copulas, i.e. two-dimensional copulas, are the fundamental building blocks in this thesis. According to their characters, copulas can be separated into different classes. In this paper, two important classes will be introduced and used as candidates for the model selection. They are the Elliptical Copula Class and the Archimedean Copula Class. We first give the definition of the density function of a bivariate copula.

Definition 2.16. Let C be a two-dimensional two-times partial differential copula, then the function $c : [0,1]^2 \to [0,1]$ with

$$c(u,v) = \frac{\partial^2 C(u,v)}{\partial u \partial v}$$

is called the *copula density* of the copula C

Elliptical copulas

The most common and important copulas in Finance are the Gaussian (Normal) copula and the t-copula. Both belong to the Elliptical Copula Class. A detailed introduction can be found in Embrechts et al. [2001] and Fusai and Roncoroni [2008].

Definition 2.17 (Multivariate Gaussian copula)**.** Let Σ be a symmetric, positive definite matrix with unit diagonal entries. The multivariate Gaussian copula is defined as

$$C(u_1, \ldots, u_n) = \Phi_\Sigma(\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_n)),$$

where $\Phi(\cdot)$ denotes the univariate standard normal distribution function and $\Phi_\Sigma(\cdot)$ denotes the standardized multivariate normal distribution with correlation matrix Σ. The corresponding density function in a bivariate case is given in Table 2.2.

Definition 2.18 (T copula)**.** For a symmetric and positive definite matrix Σ with unit diagonal entries, let $T_v(\cdot;\Sigma)$ denote the standardized multivariate Student's t distribution with correlation matrix Σ and $\nu \geq 1$ degrees of freedom. The multivariate T copula is defined as

$$C(u_1, \ldots, u_n; \Sigma, \nu) = T_\nu(T_\nu^{-1}(u_1), \ldots, T_\nu^{-1}(u_n); \Sigma),$$

where $T_\nu^{-1}(\cdot)$ is the inverse of the univariate Student's t distribution function with ν degrees of freedom. The corresponding density function in a bivariate case is given in Table 2.2.

Archimedean Copulas

Archimedean copulas constitute an important class of copula due to their analytical tractability and ability to reproduce different dependency structures. For instance, according to the bivariate probability integral transform (BIPIT) established by Genest and Rivest [1993], the distribution function $K(t)$ can be rewritten in an explicit expression.

Table 2.2: Density functions of bivariate elliptical copulas and their parameter spaces

Copula	Density function	Parameter space
Normal	$c(u_1,u_2)=\frac{1}{\sqrt{1-\theta^2}}\exp\left\{\frac{\theta^2(x_1^2+x_2^2)-2\theta x_1x_2}{2(1-\theta^2)}\right\}$ where $x_1=\Phi^{-1}(u_1), x_2=\Phi^{-1}(u_2)$ and $\Phi^{-1}(\cdot)$ is the inverse of the standard univariate normal distribution function	$\theta\in(-\infty,\infty)$
t	$c_{\theta,\upsilon}(u_1,u_2)=$ $\frac{\Gamma(\frac{\upsilon+2}{2})/\Gamma(\frac{\upsilon}{2})}{\upsilon\pi \mathtt{d}t(x_1,\upsilon)\mathtt{d}t(x_2,\upsilon)\sqrt{1-\theta^2}}\times\left\{1+\frac{x_1^2+x_2^2-2\theta x_1x_2}{\upsilon(1-\theta^2)}\right\}^{-\frac{\upsilon+1}{2}}$ where $x_1=t_\upsilon^{-1}(u_1), x_2=t_\upsilon^{-1}(u_2)$. $\mathtt{d}t(\cdot,\dot{\upsilon})$ and $t_\upsilon^{-1}(\dot{\cdot})$ are the probability density and the quantize function, respectively, for the standard univariate Student t-distribution with υ degrees of freedom, expectation 0 and variance $\frac{\upsilon}{\upsilon-2}$.	$\theta\in(-\infty,\infty)$

Table 2.3: Bivariate Archimedean copula functions and their generators

Copula	Bivariate copula function	Generator $\phi(t)$
Clayton	$C_\theta(u,v)=\left(u^{-\theta}+v^{-\theta}-1\right)^{-\frac{1}{\theta}}$	$\frac{t^{-\theta}-1}{\theta}$
Gumbel	$C_\theta(u,v)=\exp\left[-\left((-\ln u)^{-\theta}+-(\ln v)^{-\theta}\right)^{\frac{1}{\theta}}\right]$	$\lvert\ln(t)\rvert^\theta$
Frank	$c_\theta(u_1,u_2)=-\theta g_1\frac{1+g_{u_1+u_2}}{(g_{u_1}g_{u_2}+g_1)^2}$	$\ln(\frac{1-e^{-\theta}}{1-e^{-\theta t}})$
BB1	$C_{\theta,\upsilon}(u,v)=\left\{1+\left[(u^{-\theta}-1)^\delta+(v^{-\theta}-1)^\delta\right]^{\frac{1}{\delta}}\right\}^{-\frac{1}{\theta}}$	$(t^{-\theta}-1)^\delta$
BB7	$C_{\theta,\delta}(u,v)=$ $1-\left(1-\left[(1-(1-u)^\theta)^{-\delta}+(1-(1-v)^\theta)^{-\delta}-1\right]^{-\frac{1}{\delta}}\right)^{\frac{1}{\theta}}$	$[1-(1-t)^\theta]^{-\delta}-1$

Table 2.4: Density functions of bivariate Archimedean copula and their parameter spaces

Copula	Density function	Parameter space
Clayton	$c_\theta(u,v) = (1+\theta)(uv)^{-1-\theta} \cdot \left(u^{-\theta} + v^{-\theta} - 1\right)^{-1/\theta-2}$ where perfect dependence:$\theta \to \infty$, Independence:$\theta \to 0$.	$\theta \in (0, \infty)$
Gumbel	$c_\theta(u,v) = C_\theta(u,v)(uv)^{-1} \times \left\{(-\ln v)^\theta + (-\ln v)^\theta\right\}^{-2+\frac{2}{\theta}} \times (\ln u \ln v)^{\theta-1} \times \left\{1 + (\theta-1)((-\ln u)^\theta + (-\ln v)^\theta)^{-\frac{1}{\theta}}\right\}$ where perfect dependence:$\theta \to \infty$, independence:$\theta = 1$.	$\theta \in [1, \infty)$
Frank	$c_\theta(u_1,u_2) = -\theta g_1 \dfrac{1 + g_{u_1+u_2}}{(g_{u_1}g_{u_2} + g_1)^2}$ where $g_z := g(z) = e^{-\theta z} - 1$.	$\theta \in (-\infty, \infty)$, $\theta \neq 0$
BB1	$c_{\theta,\delta}(u,v) =$ $\left\{1 + \left[(u^{-\theta}-1)^\delta + (v^{-\theta}-1)^\delta\right]^{\frac{1}{\delta}}\right\}^{\frac{1}{\theta}-2} \times \left[(u^{-\theta}-1)^\delta + (v^{-\theta}-1)^\delta\right]^{\frac{2}{\delta}-2} \times \left\{\theta\delta + 1 + \theta(\delta-1)\left[(u^{-\theta}-1)^\delta + (v^{-\theta}-1)^\delta\right]^{-\frac{1}{\delta}}\right\} \times (u^{-\theta}-1)^{\delta-1} u^{-\theta-1} (v^{-\theta}-1)^{-\delta-1} v^{-\theta-1}$	$\theta \in (0, \infty)$, $\delta \in [1, \infty)$
BB7	$c_{\theta,\delta}(u,v) = (-\dfrac{1}{\theta})(\dfrac{1}{\delta} - 1) \cdot h^{\frac{1}{\theta}-2} \cdot dvh \cdot duh - \dfrac{1}{\theta} \cdot h^{\frac{1}{\theta}-1} \cdot duvh$ where · $h = 1 - \left(\left(1-(1-u)^\theta\right)^{-\delta} - \left(1-(1-v)^\theta\right)^{-\delta} - 1\right)^{\frac{1}{\delta}}$ · $duh = -\theta\left((1-(1-u)^\theta)^{-\delta} - (1-(1-v)^\theta)^{-\delta} - 1\right)^{\frac{1}{\delta}-1}\left(1-(1-u)^\theta\right)^{-\delta-1}(1-u)^{\theta-1}$ · $dvh = -\theta\left((1-(1-u)^\theta)^{-\delta} - (1-(1-v)^\theta)^{-\delta} - 1\right)^{\frac{1}{\delta}-1}\left(1-(1-uv)\right)^{-\delta-1}(1-v)^{\theta-1}$ · $duvh = \frac{1}{\delta}(-\frac{1}{\delta}-1)\left((1-(1-u)^\theta)^{-\delta} - (1-(1-v)^\theta)^{-\delta} - 1\right)^{\frac{1}{\delta}-2} \cdot duS \cdot dvS$ · $duS = -\theta\delta\left((1-(1-u)^\theta)^{-\delta-1}(1-u)^{\theta-1}\right.$	$\theta \in [1, \infty)$, $\delta \in (0, \infty)$

An Archimedean copula *generator* is a convex and strictly decreasing function $\phi : [0,1] \to \mathbb{R}^+$ where the *pseudo-inverse of* ϕ is defined as a function $\phi^{[-1]} : [0,\infty] \to [0,1]$, such that

$$\phi^{[-1]}(t) := \begin{cases} \phi^{[-1]}(t) & \text{if } 0 \le t \le \phi(0) \\ 0 & \text{if } \phi(0) \le t < \infty \end{cases}.$$

If $\phi(0) = \infty$, the $\phi^{[-1]} = \phi^{-1}$, and the generator said to be *strict*. In this case, copula function has an explicit expression.

Definition 2.19 (bivariate Archimedean copula)**.** Let ϕ be an Archimedean copula generator. The function

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

is an Archimedean copula.

Remark 2.20 (Properties of Archimedean copulas).

(i) C is symmetric.

(ii) C is associative, i.e. $C(C(u_1, u_2), u_3) = C(u_1, C(u_2, u_3))$.

(iii) For any $c \ge 0$, $c\phi$ is also a generator of C.

There are a number of copulas which belong to the class of Archimedean copulas. In this paper, the Clayton, Gumbel, Frank, BB1 and BB7 copulas will be investigated. Table 2.3 lists their bivariate copula functions and generators. Table 2.4 lists the density functions and the parameter spaces, respectively.

Rotated copulas

Theorem 2.21. *Let X and Y be two continuous random variables with distributions F and G. Let C be the copula of (X, Y). Define $\bar{F}(X) := 1 - F(X)$ and $\bar{G}(Y) = 1 - G(Y)$, then it follows*

· *$\bar{F}(X)$ and $\bar{G}(Y)$ has copula*

$$C^{--}(\bar{F}(X), \bar{G}(Y)) = F(X) + G(Y) - 1 + C(1 - F(X), 1 - G(Y)).$$

· *$\bar{F}(X)$ and $G(Y)$ has copula*

$$C^{-+}(\bar{F}(X), G(Y)) = G(Y) - C(1 - F(X), G(Y)).$$

· *$F(X)$ and $\bar{G}(Y)$ has copula*

$$C^{+-}(F(X), \bar{G}(Y)) = F(X) - C(F(X), 1 - G(Y)).$$

Proof. Use Theorem 2.14 with $\alpha_1(X) = 1 - X \sim \bar{F}(X)$ and $\alpha_2(Y) = 1 - Y \sim \bar{G}(Y)$ in bivariate case. □

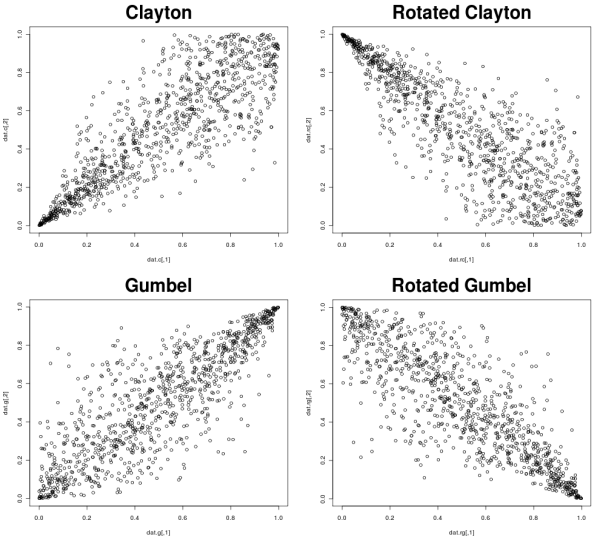


Figure 2.1: Plots for Example 2: Clayton and 90^o rotated Clayton data (top row), Gumbel and 90^o rotated Gumbel data (bottom row)

Definition 2.22. The transformation from Theorem 2.21 is called *rotation*. In the bivariate case the copula $C^{--}$is a 180 degree rotated copula, $C^{+-}$and C^{-+} are 90 degree rotated copulas in different directions.

Example 2.

- The 90 degree clockwise rotated Clayton copula:

$$C^{RC}(u,v) := u - C^{C}(u, 1-v),$$

  where $C^{C}$is a Clayton copula.

- The 90 grad clockwise rotated Gumbel copula:

$$C^{RG}(u,v) := u - C^{G}(u, 1-v),$$

  where $C^{G}$is a Gumbel copula.

From Example 2 one can see that rotated Archimedean copulas are not Archimedean copulas. One can also conclude that if (U,V) has a rotated copula, then $(U, 1-V)$ has the original copula, i.e.

$$(U,V) \sim C^R \;\Leftrightarrow\; (U, 1-V) \sim C. \tag{2.6}$$

Eq. (2.6) indicates an equivalent relationship between the rotated copula and the original one.

It is easy to prove that their density functions have the following relationship

Theorem 2.23. *Let c^R denote the density function of the rotated copula C^R and c denote the density function of the original copula C. Then*

$$c^R(u,v) = c(u, 1-v).$$

2.4 Pair-copula decomposition and vines

Using two-dimensional copulas one might construct general multivariate distributions by specifying the dependence and conditional dependence of selected pairs of random variables and all marginal distribution functions. In this section we introduce the pair-copula decomposition of a general multivariate distribution and illustrate this with some examples. This concept was first provided by [Bedford and Cooke, 2002]. We will briefly present this concept in this section. For more details we refer to [Aas et al., 2009].

Consider a random vector $\mathbf{X} =(X_1, \ldots, X_n)$ with a joint distribution H and margins $F_1(x_1) \ldots, F_n(x_n)$. Sklar's theorem ([Sklar, 1959]) states that H can be rewritten as

$$H(x_1, \ldots x_n) = C\big(F_1(x_1), \ldots, F_n(x_n)\big).$$

Using the chain rule the joint density f can be rewritten as

$$f(x_1, \ldots, x_n) = c_{1,\ldots,n}\big(F_1(x_1), \ldots, F_n(x_n)\big) \cdot f_1(x_1) \cdot \ldots \cdot f_n(x_n), \tag{2.7}$$

where $c_{1,\ldots n}(\cdot)$ is the n-variate copula density. In the bivariate case, Eq. (2.7) simplifies to

$$f(x_1, x_2) = c_{12}\big(F_1(x_1), F_2(x_2)\big) \cdot f_1(x_1) \cdot f_2(x_2), \tag{2.8}$$

where $c_{12}(\cdot, \cdot)$ is the appropriate *pair copula density* for the pair of transformed variables $U = F_1(x_1)$ and $V = F_2(x_2)$. For a conditional density in the two-dimensional case it easily follows that

$$f(x_1|x_2) = \frac{f(x_1, x_2)}{f(x_2)} \overset{(2.8)}{=} c_{12}(F_1(x_1), F_2(x_2)) \cdot f_1(x_1). \tag{2.9}$$

For three random variables we have that

$$\begin{aligned} f(x_1|x_2, x_3) &= c_{13|2}(F(x_1|x_2), F(x_3|x_2)) \cdot f(x_1|x_2) \\ &\overset{(2.9)}{=} c_{13|2}(F(x_1|x_2), F(x_3|x_2)) \cdot c_{12}(F_1(x_1), F_2(x_2)) \cdot f_1(x_1) \end{aligned}$$

for some appropriate pair copulas $c_{13|2}$ and $c_{12|3}$ applied to corresponding transformed variables.

On the other hand, the joint density f can be factorized as

$$f(x_1, \ldots, x_n) = f_n(x_n) \cdot f(x_{n-1}|x_n) \cdot f(x_{n-2}|x_{n-1}, x_n) \cdot \ldots \cdot f(x_1|x_2, \ldots x_n), \tag{2.10}$$

then using the decomposition (2.7) and (2.10) we have the pair-copula decomposition in the three-dimensional case:

$$\begin{aligned} f(x_1, x_2, x_3) &= f(x_3) f(x_2|x_3) f(x_1|x_2, x_3) \\ &= c_{12|3}\big(F(x_1|x_3), F(x_2|x_3)\big) c_{13}\big(F_1(x_1), F_3(x_3)\big) \\ &\quad \cdot c_{23}\big(F_2(x_2), F_3(x_3)\big) f_1(x_1) f_2(x_2) f_3(x_3). \end{aligned}$$

The general formula for the conditional density in d-dim is given as follows

$$f(x|\mathbf{v}) = c_{x,v_j|\mathbf{v}_{-j}}\big(F(x|\mathbf{v}_{-j}), F(v_j|\mathbf{v}_{-j})\big) \cdot f(x|\mathbf{v}_{-j})$$

with $\mathbf{v} = (v_1, \ldots, v_d)$ and $\mathbf{v}_{-j} = (v_1, \ldots, v_{j-1}, v_{j+1}, \ldots, v_d)$.

In conclusion, under appropriate regularity conditions a multivariate density can be expressed as a product of pair-copulas acting on several different conditional probability distributions. It is also clear that the construction is iterative by nature, and that given a specific factorization, there are still many different re-parametrizations.

Joe [1997] shows that for every v_j in the vector $\mathbf{v}$ the general formula of $F(x|\mathbf{v})$ is given as

$$F(x|\mathbf{v}) = \frac{\partial C_{x,v_j|\mathbf{v}_{-j}}(F(x|\mathbf{v}_{-j}), F(v|\mathbf{v}_{-j}))}{\partial F(v_j|\mathbf{v}_j)},$$

where $C_{ij|\mathbf{k}}$ is a bivariate copula distribution function. In the following we will use the function $h(x, v, \theta)$ to represent this conditional distribution function with x and v uniformly distributed, i.e. $f(x) = f(v) = 1$, $F(x) = x$ and $F(x) = x$. That is

$$h(x, v, \Theta) := F(x|v) = \frac{\partial C_{x,v}(x, v, \Theta)}{\partial v}, \tag{2.11}$$

where the second parameter of $h(\cdot)$ always corresponds to the conditioning variable and Θ denoted the set of parameters for the copula of the joint distribution function of x and v. Let $h^{-1}(u, v, \Theta)$ be the inverse of the h-function with respect to the first variable u, or equivalently the inverse of the conditional distribution function. An example for the h-function is given as follows.

Example 3. Let $c(u_1, u_2)$ denote the Gaussian copula density function with parameter θ. The h-function is given by

$$h(u_1, u_2, \theta) = \Phi\Big(\frac{\Phi^{-1}(u_1) - \theta\Phi^{-1}(u_2)}{\sqrt{1-\theta^2}}\Big)$$

and the inverse of the h-function is given by

$$h^{-1}(u_1, u_2, \theta) = \Phi\Big(\Phi^{-1}(u_1)\sqrt{1-\theta^2} + \theta\Phi^{-1}(u_2)\Big).$$

D-vine

Bedford and Cooke [2002], and Kurowicka and Cooke [2006] introduced the so-called *regular vines* as a graphical method to represent the multitude of the possible pair-copula decompositions. For example, there are 24 possible pair-copula decompositions for a four-dimensional distribution. The vines help to organize them. The class of regular vines consists of many different types of vines and embraces a large number of possible pair-copula decompositions. Up to four dimension, only two special classes of vines are involved, the *canonical vine (C-vine)* and the *D-vine*. In this thesis we will concentrate on the D-vine. More details about C-vine can be found in Aas et al. [2009].

As we known, there are many possible pair-copula constructions for a high dimensional distribution. Different decompositions lead to different models. Every model decomposes the density in a different way and each model gives a nested set of trees.

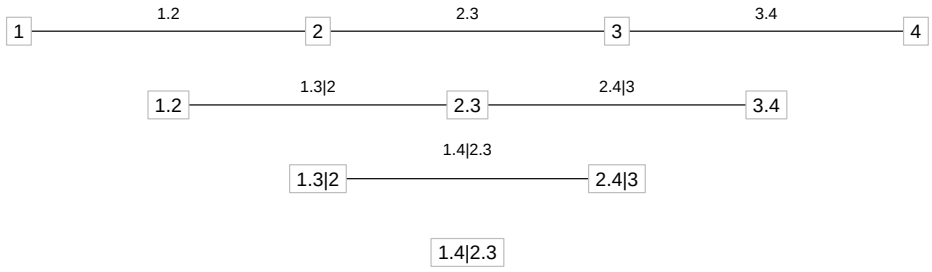


Figure 2.2: A D-vine with 4 variables, 4 trees and 6 edges.

Figure 2.2 shows a specific model for a four-dimensional D-vine. It consists of three trees $T_j,\ j = 1, 2, 3$. Tree T_j has $5 - j$ nodes and $4 - j$ edges. Each edge corresponds to a pair-copula density and the edge label corresponds to the subscript of the pair-copula density. For example edge $13|2$ corresponds to the pair copula $c_{13|2}(\cdot)$. The entire decomposition is defined by the $n(n-1)/2$ edges and the marginal densities of each variable. In D-vines every node has a maximum of two edges, i.e. neighbors. The nodes in Tree T_j are only necessary for determining the labels of the edges in the next tree T_{j+1}. As in Figure 2.2, two edges in T_j, which becomes node in T_{j+1}, are joined by an edge in T_{j+1} only if these edges in T_j share a common node.

Bedford and Cooke [2002] provided the density of an n-dimensional distribution in terms of a regular vine. We specialize to the D-vine here. The density $f(x_1, \ldots, x_n)$ corresponding to a D-vine can be written as

$$\begin{aligned} f(x_1, \ldots, x_n) &= \prod_{k=1}^{n} f(x_k) \prod_{j=1}^{n-1} \prod_{i=1}^{n-j} c_{i,i+j|i+1,\ldots i+j+i} \\ &\quad \big(F(x_i|x_{i+1}, \ldots, x_{i+j+1}), F(x_{i+j}|x_{i+1}, \ldots x_{i+j+1})\big), \end{aligned}$$

where index j identifies the trees, and i runs over the edges in each tree. In the following examples we want to provide insight into the general expression of the density function for D-vine structures in lower dimensional cases.

Example 4 (Three variables)**.** The general density for D-vine structures in a three-dimensional case is

$$\begin{aligned} f(x_1, x_2, x_3) &= f(x_1) \cdot f(x_2) \cdot f(x_3) \\ &\quad \cdot c_{12}\big(F_1(x_1), F_2(x_2)\big) \cdot c_{23}\big(F_2(x_2), F_3(x_3)\big) \cdot c_{13|2}\big(F(x_1|x_2), F(x_3|x_2)\big). \end{aligned}$$

Six possible permutations of (x_1,x_2,x_3) exist for this representation of the density, but only three of them leads to different decompositions.

Example 5 (Four variables)**.** The four-dimensional D-vine structure is given by

$$\begin{aligned} f(x_1, x_2, x_3x_4) &= f(x_1) \cdot f(x_2) \cdot f(x_3) \cdot f(x_4) \\ &\quad \cdot c_{12}\big(F_1(x_1), F_2(x_2)\big) \cdot c_{23}\big(F_2(x_2), F_3(x_3)\big) \cdot c_{34}\big(F_3(x_3), F_4(x_4)\big) \\ &\quad \cdot c_{13|2}\big(F(x_1|x_2), F(x_3|x_2)\big) \cdot c_{24|3}\big(F(x_2|x_3), F(x_4|x_3)\big) \\ &\quad \cdot c_{14|23}\big(F(x_1|x_2, x_3), F(x_4|x_2, x_3)\big). \end{aligned} \tag{2.12}$$

There are 12 different D-vine decomposition and each of them is unique.

For the general case with n variables, there are $n!$ possible ways of ordering the variables in tree T_1. Since we have undirected edges for all pairs (i, j) and arbitrary conditioning sets for D-vines, we can reverse the order in tree T_1 for a D-vine without changing the corresponding vine. Therefore we only have $n!/2$ different trees on the first level. Given such a tree T_1, the trees $T_2, T_3, \ldots, T_{n-1}$ are completely determined. This implies that the number of distinct D-vines on nodes n is given by $n!/2$.

2.5 Statistical hypothesis testing

In this section, we will review the basic definitions and theories for statistical hypothesis tests. In statistics, one is often interested in testing whether a certain hypothesis is true, for instance, a hypothesis of the parameter or the distribution function. These hypotheses are tested using the sampled data, i.e. a test is a method of making statistical decisions based on observations. Thus, it is called as statistical test.

Definition 2.24 (Statistical test problem, statistical test)**.** A *statistical test problem* consists of the null hypothesis H_0 and the alternative hypothesis H_1. H_0 and H_1 are disjoint and mutually exclusive. They retain the predication about the population distribution functions or the parameters of the object.
A test is a *two-sided* test if

$$H_0 : " = " \quad \text{versus} \quad H_1 : " \neq ".$$

The test is *one-sided* if

$$H_0 : " \leq " \quad \text{versus} \quad H_1 : " > " \qquad \text{or} \qquad H_0 : " \geq " \quad \text{versus} \quad H_1 : " < ".$$

If H_0 or H_1 consists of one point in the parameter space, the null hypothesis or the alternative is *simple.* If H_0 or H_1 contains more than one point, it is *composite.*
A statistical test with a suitable size provides a formal decision rule, which decides whether H_0 or H_1 is true.

In a statistical test, a *test statistic* or *statistic* can be virtually any measurable function of samples. It can be generally defined as follows:

Definition 2.25. A *test statistic* for the parameter θ is a function

$$T = g(X_{1,} \ldots, X_n),$$

where $X_1, \ldots, X_n$ are the random variables. Let $x_1, \ldots, x_n$ be the realizations, then $t = g(x_1, \ldots x_n)$ is the *estimated value* of T.

Definition 2.26 (Type I error, type II error)**.** In a statistical test for the test problem H_0 versus H_1 , we can make a

- *Type I error,* if we reject H_0 but H_0 is true.

· *Type II error*, if we accept H_0 but it is false.

Definition 2.27 (Significance test, significance level)**.** A statistical test is called a test with a significance level α , or a *significance test, when*

$$P(\textit{Type I error}) \leq \alpha,$$

where $0 \leq \alpha \leq 1$. Typical values of α are 0.1, 0.05, 0.01.

The result of a test is *statistical significant* if we reject the null hypothesis. Otherwise, the result is not statistical significant.

The rejection region is used in hypothesis testing. Let T be a test statistic. Possible values of T can be divided into two regions, the acceptance region and the rejection region. If the observed value of T comes out to be in the acceptance region, the null hypothesis (the one being tested) is accepted, or at any rate not rejected. If $T = t$ falls in the rejection region, the null hypothesis is rejected. The terms acceptance region and rejection region are subsets of the sample space.

Definition 2.28. Set up a hypothesis $H_0 : \Theta \in \Omega_H$ with the corresponding alternative $H_1 : \Theta \in \Omega_A$, then $\Omega_H \cup \Omega_A = \Omega$ and $\Omega_H \cap \Omega_A = \emptyset$ construct a partition of the parameter space. Define

$$C = \{\text{Rejection of } H_0\} = \{\mathbf{x} = (x_1, \ldots, x_n) : \mathbf{x} \in R\}$$

and

$$C^c = \{\text{Acceptance of } H_0\} = \{\mathbf{x} = (x_1, \ldots, x_n) : \mathbf{x} \notin R\},$$

where R is a subset of the sample space $\mathcal{X}$. Then R would be called as the *rejection region*, i.e. if $\mathbf{x} \in R$ is observed then reject H_0, and we would say that we do not reject H_0 if $\mathbf{x} \notin R$.

Tests are compared based on their power function. The *power function* of a test with rejection region R is defined as $\beta(\theta) = P_\theta(\mathbf{x} \in R)$. The *size* of a test is defined as $\sup_{\theta \in \Omega_H} \beta(\theta)$.

P-value is associated with the test statistic in a statistical test. It can be defined as follows

Definition 2.29. *P-value* is the probability of observing a test statistic that is as extreme or more extreme than currently observed assuming that the null hypothesis is true. It can be expressed as

$$p = \inf_{\alpha} \{\mathbf{x} \in R_\alpha\} = P(T \geq t_{obs} | H_0),$$

where t_{obs} denotes the observed value of T and R_α is a size α rejection region satisfying $P_\theta(R) \leq \alpha$. Expressed alternatively, the p-value is the smallest α value at which H_0 can be rejected, when the sample $\mathbf{x}$ is observed. If the p-value is equal or less than a given significance level α, then H_0 would be rejected.

The statistical testing process

(i) State the test problem with relevant null and alternative hypotheses to be tested.

(ii) Formulate the statistical assumptions; for example, assumptions about the statistical independence or about the form of the distributions of the observations.

(iii) Decide which test is appropriate, and use the relevant test statistic T.

(iv) Derive the distribution of the test statistic under the null hypothesis from the assumptions and determine which significance level α is appropriate.

(v) Derive the possible values of T and then determine the rejection region corresponding to the chosen significance level α.

(vi) Compute from the observations the observed value t_{obs} of the test statistic T.

(vii) Reject the null hypothesis H_0 if the observed value t_{obs} is in the critical region, and accept or "fail to reject" the hypothesis otherwise.

2.6 Determination of the p-value using bootstrapping

Bootstrap methods are computer based methods to assess the accuracy of statistical estimates such as standard errors, confidence intervals and P-values. The key idea is to resample from the original data, i.e. to create replicate data sets. More introductions about bootstrap can be found in Chernick [2008], Davison and Hinkley [1997] and Efron et al. [1993].

From last section we know that the P-value is defined as

$$p = P(T \geq t \mid H_0).$$

The bootstrap procedure for P-values can generally be designed as the following algorithm:

Algorithm 2.1 Bootstrapping P-value

(i) Estimate parameter $\hat{\theta}$ of the null model using the observations $\mathbf{y}$.

(ii) Assume $F_{\hat{\theta}}$ is the model to be tested. Draw B independent random samples $\mathbf{y}_1^*, \ldots, \mathbf{y}_B^*$ with replacement of the observations $\mathbf{y}$ from null model $F_{\hat{\theta}}$.

(iii) Estimate parameters $\hat{\theta}_1^*, \ldots, \hat{\theta}_B^*$ from samples $\mathbf{y}_1^*, \ldots, \mathbf{y}_B^*$.

(iv) Evaluate the test statistic value using $\hat{\theta}_j^*$ for bootstrap sample $\mathbf{y}_j^*$ for $j = 1, \ldots, B$ respectively and denote this value by t_j^*.

(v) Evaluate the observed value of test statistic from original observations $\mathbf{y}$ and denote it by t.

(vi) Approximate the P-value by

$$p_{boot} = \frac{\#\{j; t_j^* \geq t\}}{B}$$

for $j \in \{1, \ldots, B\}$.

Chapter 3

Kendall's process and the λ function

3.1 Revisiting Kendall's τ

Since copulas were introduced in the last chapter, we want to connect Kendall's τ to copulas. Recall that Remarks 2.7 (vi),

$$\tau = 4 \int_{[0,1]^2} F(x,y) \mathrm{d}F(x,y) - 1$$

gives a relationship between Kendall's τ and the joint distribution of bivariate random variables. A similar expression can be derived for a copula C, see Genest and Rivest [1993, 2001]. It can be expressed as an integral of C. Specifically,

Theorem 3.1. *Let $(X,Y)^t$ be a vector of continuous random variables with copula C. Then Kendall's τ for $(X,Y)^t$ is given by*

$$\tau(X,Y) = 4 \int_{[0,1]^2} C(u,v) dC(u,v) - 1. \tag{3.1}$$

Note that the integral in (3.1) is the expected value of the random variable $C(U,V)$, where $U,V \sim U(0,1)$ with joint distribution function C, i.e. $\tau(X,Y) = 4\mathbb{E}(C(U,V)) - 1$.

Proof. From the definition of Kendall's τ we know

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

where $(\tilde{X},\tilde{Y})$ is the independent copy of (X,Y), so that $H(x,y) = H(\tilde{x},\tilde{y}) = C(F(x),G(y))$. Since the random variables are all continuous,

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

Hence $\tau(X,Y) = 2P((X-\tilde{X})(Y-\tilde{X}) > 0) - 1$. But

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

and these probabilities can be evaluated by integrating over the distribution of one of the vectors $(X,Y)^t$ or $(\tilde{X},\tilde{Y})^t$. Hence

$$\begin{aligned} P(X>\tilde{X},Y>\tilde{Y}) &= P(\tilde{X}<X,\tilde{Y}<Y) \\ &= \int_{\mathbb{R}^2} P(\tilde{X}<x,\tilde{Y}<y)\mathrm{d}C(F(x),G(y)) \\ &= \int_{\mathbb{R}^2} C(F(x),G(y))\mathrm{d}C(F(x),G(y)). \end{aligned}$$

Applying the probability integral transform $u=F(x)$ and $v=G(y)$ yields

$$P(X>\tilde{X},Y>\tilde{Y}) = \int_{[0,1]^2} C(u,v)\mathrm{d}C(u,v).$$

Similarly,

$$\begin{aligned} P(X<\tilde{X},Y<\tilde{Y}) &= \int_{\mathbb{R}^2} P(\tilde{X}>x,\tilde{Y}>y)\mathrm{d}C(F(x),G(y)) \\ &= \int_{\mathbb{R}^2} \big(1-F(x)-G(y)+C(F(x),G(y))\big)\mathrm{d}C(F(x),G(y)) \\ &= \int_{[0.1]^2} (1-u-v+C(u,v))\mathrm{d}C(u,v). \end{aligned}$$

However, since C is the joint distribution function of a vector $(U,V)^t$ with $U(0,1)$ margins, $E(U)=E(V)=1/2$, and hence

$$P(X<\tilde{X},Y<\tilde{Y}) = 1-\frac{1}{2}-\frac{1}{2}+\int_{[0,1]^2} C(u,v)\mathrm{d}C(u,v).$$

Thus,

$$P((X-\tilde{X})(Y-\tilde{Y})>0) = 2\int_{[0.1]^2} C(u,v)\mathrm{d}C(u,v)$$

and the conclusion follows. □

If C stands for Gaussian or T copula, using the Theorem 3.1 one can obtain the following relationship between τ and the copula parameter θ (Lindskog et al. [2003])

$$\theta = \frac{\pi}{2}\sin(\tau).$$

Due to the complexity of the copula function in most cases, Eq. (3.1) can generally not be directly calculated. However, for Archimedean copulas, Kendall's τ can be expressed explicitly with the help of their generator.

Theorem 3.2. *Let X and Y be random variables with an Archimedean copula C generated by ϕ. Kendall's τ of X and Y is given by*

$$\tau = 1+4\int_0^1 \frac{\phi(t)}{\phi'(t)}dt.$$

Proof. See Embrechts et al. [2001], Genest and Rivest [1993] □

Since Kendall's τ is a function of the Archimedean copula generator ϕ, τ is uniquely identified by the generator. Table 3.1 collects the expressions of theoretical Kendall's τ for the copulas we use in this paper.

Table 3.1: Kendall's τ for different copula families

Copula	τ	Parameter spaces	Range of τ
Gaussian & T	$\frac{2}{\pi}\arcsin(\theta)$	$\theta \in (-\infty, \infty)$	$[-1, 1]$
Clayton	$\frac{\theta}{\theta+2}$	$\theta \in (0, \infty)$	$[0, 1]$
Gumbel	$\frac{\theta-1}{\theta}$	$\theta \in [1, \infty)$	$[0, 1]$
Frank	$1 - \frac{4}{\theta} + \frac{4D_1(\theta)}{\theta}$	$\theta \in (-\infty, \infty) \setminus \{0\}$	$[-1, 1]$
BB1	$1 - \frac{2}{\delta(\theta+2)}$	$\theta \in (0, \infty),\ \delta \in [1, \infty)$	$[0, 1]$
BB7	$1 - \frac{2}{\delta(2-\theta)} + \frac{4}{\theta^2\delta}B(\frac{2-2\theta}{\theta}+1, \delta+2)$	$\theta \in [1, \infty),\ \delta \in (0, \infty)$	$[0, 1]$

where $D_1(\theta) = \theta^{-1}\int_0^\theta \frac{x}{e^x-1}\mathrm{d}x$ stands for the first Debye function and $B(x, y) = \int_0^1 t^{x-1}(1-t)^{y-1}\mathrm{d}t = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}$ stands for the Beta function.

3.2 Kendall's process and λ-function

We first want to introduce the definition of the probability integral transform. If X is a continuous random variable with distribution function F, then the *probability integral transform* (PIT) of X is defined as $U := F(X)$. U is a uniformly distributed random variable on the unit interval $[0, 1]$, i.e. $P(F(X) \leq t) = t,\ \forall t \in [0, 1]$. In this paper we focus on the two-dimensional case.

Definition 3.3 (BIPIT)**.** Let X and Y be two random variables with marginal distributions F and G. X and Y have the joint continuous distribution function H. Then the random variable

$$H := H(X, Y)$$

is called the *bivariate probability integral transform* (BIPIT). H is a one-dimensional random variable.

In this section we will investigate the distribution function $K(\cdot)$ of the random variable $H = H(X, Y)$. Genest and Rivest [1993] firstly introduced the K and λ-function in a proposition and called them the "decomposition of Kendall's τ". This connection is given in Remark 3.6.

Proposition 3.4. *Let U and V be uniform random variables whose dependency function $C(u, v)$ is of the form $\phi^{-1}(\phi(u) + \phi(v))$ for some convex decreasing function ϕ defined on $(0, 1]$ with the property that $\phi(1) = 0$. Set $Z := \phi(U)/(\phi(U) + \phi(V))$, $W := C(X, Y)$, and $\lambda(w) := \phi(w)/\phi'(w)$ for $0 \leq w \leq 1$. Then*

(i) Z is uniformly distributed on $[0, 1]$.

(ii) W is distributed as $K(w) = w - \lambda(w)$ on $[0, 1]$.

(iii) Z and W are independent random variables.

If (U, V) has copula C, then $K(t)$ and $\lambda(t)$ are formally defined as

Definition 3.5. Let X and Y be two continuous random variables with joint distribution function H and marginal distribution functions F and G. Let $U = F(X)$ and $V = G(Y)$. If (U, V) has copula C, then the *Kendall distribution function,* or *copula distribution function* $K(t)$ for $0 \leq t \leq 1$ is defined as

$$K_\theta(t) := P(H(X, Y) \leq t) = P(C(F(X), G(Y))) = P(C_\theta(U, V) \leq t) \tag{3.2}$$

and the λ-function is defined as

$$\lambda_\theta(t) := t - K_\theta(t).$$

Remark 3.6. Let C be a copula and K be its Kendall distribution function.

(i) $t \leq K_\theta(t), \quad \forall t \in [0, 1]$.

(ii) $K_\theta(0^-) = 0$.

(iii) $\tau(X, Y) = 3 - 4 \int_0^1 K_\theta(t)\mathrm{d}t$.

(iv) Let F be a right-continuous distribution function such that $F(0^-) = 0$ and $F(t) \geq t \; \forall t \in [0, 1]$. Then there exists a copula C such that $K_\theta(t) = F(t) \quad \forall t \in [0, 1]$.

All proofs of Remarks 3.6 can be found in Nelsen et al. [2003] and Nelsen et al. [2001]. Genest and Rivest [1993] shows that K can be estimated non-numerically by the empirical distribution function of pseudo-observations, which are defined as follows:

Definition 3.7. Let U and V be defined as in Definition 3.5, then U and V are two $U(0, 1)$ distributed random variables. If $\mathbf{u} = (u_1, \ldots, u_n)$ and $\mathbf{v} = (v_i, \ldots, v_n)$ are the observations of U and V, then the *pseudo-observations* are $w_i := C_n(u_i, v_i)$ for $i = 1, \ldots, n$ and the *empirical version* of $K(t)$ is simply the empirical distribution of the empirical copula function. It is defined as

$$K_n(t) := \frac{1}{n} \sum_{j=1}^{n} \mathbf{1}\{w_j \leq t\} \tag{3.3}$$

using the sample w_j, where C_n is the empirical copula function defined in Eq. (2.5).

K and λ-functions for Archimedean copulas

We already know from Section 2.3 that the generator of an Archimedean copula has all the properties given in Proposition 3.4. Therefore one can conclude that the BIPIT form for an Archimedean copula is given as:

$$K(t) = t - \lambda(t), \qquad \text{and} \quad \lambda(t) = \frac{\phi(t)}{\phi'(t)}, \quad \forall t \in [0, 1],$$

where $\phi'(t) = \partial\phi(t)/\partial t$. Since the generator of an Archimedean copula family is explicit, the copula distribution function K and the λ-function have explicit expression. Table 3.2 list a few Archimedean copulas with the expressions of their generators and $\lambda(t)$.

Table 3.2: The Generator and $\lambda(t)$ for Archimedean copula families

Copula	$\phi(t)$	$\lambda(t)$
Clayton	$\frac{t^{-\theta}-1}{\theta}$	$-\frac{t(1-t^{\theta})}{\theta}$
Gumbel	$\lvert\ln(t)\rvert^{\theta}$	$\frac{t\ln(t)}{\theta}$
Frank	$\ln(\frac{1-e^{-\theta}}{1-e^{-\theta t}})$	$-\frac{1-e^{-\theta t}}{\theta e^{-\theta t}}\ln(\frac{1-e^{-\theta}}{1-e^{-\theta t}})$
BB1	$(t^{-\theta}-1)^{\delta}$	$\frac{1}{\theta\delta}\frac{t^{-\theta}-1}{t^{-1-\theta}}$
BB7	$[1-(1-t)^{\theta}]^{-\delta}-1$	$-\frac{(1-(1-t)^{\theta})^{-\delta}-1}{\theta\delta(1-t)^{\theta-1}(1-(1-t)^{\theta})^{-\delta-1}}$

K and λ-functions for elliptical copulas

If we look at the Gaussian and the Student t copulas, their copula functions do not have the form

$$C(u,v)=\phi^{-1}(\phi(u)+\phi(v))$$

as defined in Proposition 3.4. Therefore no explicit formula of K or the λ-function can be found as for Archimedean copulas. However, once the parameter θ of the copula is known or estimated, one can use a nonparametric method based on the theoretical cdf of the copulas $C_\theta(U,V)$ to simulate $K_\theta(t)$ and $\lambda_\theta(t)$. If random sample pairs (u_i,v_i) are drawn from C_θ for $i=1,\ldots,N$, then the corresponding BIPIT form for the copula function is given by

$$K_\theta^N(t):=\frac{1}{N}\sum_{i=1}^{N}\mathbf{1}\{C_\theta(u_i,v_i)\le t\}. \tag{3.4}$$

Since $K_\theta^N \overset{weak}{\to} K_\theta$ for a large sample size of (U,V) (shown by Genest and Rivest [1993]), i.e. $N\to\infty$, the $K_\theta^N(t)$ is a consistent approximation for K_θ and $\lambda_\theta^N(t):=t-K_\theta^N(t)$ for $\lambda_\theta(t)$.

Note that Eq. (3.4) is a general nonparametric simulation procedure. It is applied not only for Elliptical copula families but also for other copula families. For instance, we can use Eq. (3.4) to derive $K_\theta^N(t)$ for Archimedean copulas alternatively.

K and λ-functions for rotated copulas

We know from the "Rotated copulas" subsection that the rotated copula function can be expressed with the original copula function. For example, the 90 degree clockwise rotated copula can be rewritten as

$$C^R(u,v)=u-C(u,1-v).$$

Obviously, rotated Archimedean copulas do not belong to the Archimedean Copula Class. Therefore, we do not have analytical expressions of $K(t)$ and $\lambda(t)$ for rotated Archimedean

or Elliptical copulas. This leads to a complexity of plotting $K(t)$ and $\lambda(t)$. However, Eq. (2.6) indicates an equivalent relationship between a rotated copula and the original copula,

$$(U, V) \sim C^R \Leftrightarrow (U, 1 - V) \sim C.$$

If C^R is a rotated Archimedean copula, then C is the corresponding original Archimedean copula. Thus, we want to compute $K(t)$ and $\lambda(t)$ of $(u_i, 1 - v_i)$ instead of $K(t)$ and $\lambda(t)$ of (u_i, v_i). Since the $K(t)$ and $\lambda(t)$ of $(u_i, 1 - v_i)$ have analytical BIPIT form, Computing $K(t)$ and $\lambda(t)$ of $(u_i, 1 - v_i)$ would be easier and faster than directly computing $K(t)$ and $\lambda(t)$ of (u_i, v_i) in this case. Unfortunately Elliptical copulas do not have this advantage. There is no different between using rotated data $(u_i, 1-v_i)$ and using original data (u_i, v_i), since their $K(t)$ and $\lambda(t)$ cannot be expressed explicitly.

If we do not want to rotate data, another way is to use the rotated copula model C^R as we introduced before. In this case the expressions of $K(t)$ and $\lambda(t)$ are never explicit. Therefore we use the BIPIT form to derive $K(t)$ and $\lambda(t)$ for rotated copulas. Using Eq. (3.4) we have

$$K_\theta^N(t) := \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\{C_\theta^R(u_i, v_i) \leq t\}, \tag{3.5}$$

where copula C_θ^R belongs to Rotated Copula Class.

Kendall's process

Genest and Rivest [1993] suggest that the empirical process

$$\mathbb{K}_n(t) := \sqrt{n}(K_n(t) - K_{\hat{\theta}_n}(t))$$

is referred to as *Kendall's process,* where $\hat{\theta}_n$ is the estimate of the underlying copula parameter and $K_{\hat{\theta}_n}(t)$ is the theoretical Kendall function using $\hat{\theta}_n$.

Since K_n is the nonparametric estimator of the theoretical K_θ, it is natural to examine the degree of the closeness by comparing K_n and $K_{\hat{\theta}_n}$ for different underlying copulas. A graphical examination for this is to compare the plotted K_n to the plotted $K_{\hat{\theta}_n}$ on one graph. In this way, one can observe differences between K_n and $K_{\hat{\theta}_n}$ for each copula model. This method will be introduced and examined in Chapter 4. If we want to measure this difference more precisely, then Kendall's process would be a good choice, since Kendall's process measures exactly the distance between these two functions with a scale $\sqrt{n}$. A Goodness of Fit (GOF) test based on this process will be given in Chapter 5.

Chapter 4

Graphical methods to determine copula family in two dimensions

4.1 Scatter plot

The traditional way to get a first impression of the dependency structure of random pair (U, V) is to look at the scatter plot of pairs $(u_1, v_1), \ldots, (u_n, v_n)$ from the data set. As a graphical method for model selection, one can compare the scatter plot of the pairs $(u_i, v_i) = (\frac{r_i}{n+1}, \frac{s_i}{n+1})$ with sample data generated from bivariate $C_{\hat{\theta}}$. (r_i, s_i) stands for the ranks of (u_i, v_i) among $u_1, \ldots, u_n$ and $v_1, \ldots, v_n$. Such a representation is given in Example 6.

For sampling bivariate data from a copula C, one can use a simple Algorithm 4.1. See Genest and Favre [2007].

Example 6. Test data is generated from the Clayton copula with $\theta = 5$ ($\tau = 0.71$) and $n = 500$. In the panels of Figure 4.1, 5 sample data are generated from 5 different copula families with same τ. These 5 sample data are compared with the test data, respectively.

Figure 4.1 provides the scatter plots of all 5 possible copulas. Since the Clayton sample data covers the test data most efficiently in plot, the Clayton copula is the best fit. This conclusion is logical, since the data set is simulated from Clayton copula.

Algorithm 4.1 Sampling bivariate copula data

---

1 Generate U from a Uniform distribution on $(0, 1)$.

2 Given $U = u$, generate V from the conditional distribution

$$Q_u(v) = P\big(V \leq v | U = u\big) = \frac{\partial}{\partial u} C(u, v)$$

by setting $V = Q_u^{-1}(U^*)$, where U^* is another $U(0, 1)$ distributed random variable. If an explicit expression for $Q_u^{-1}(U^*)$ dose not exist, the value $v = Q_u^{-1}(u^*)$ can be determined by trial and error, or using the bisection method.

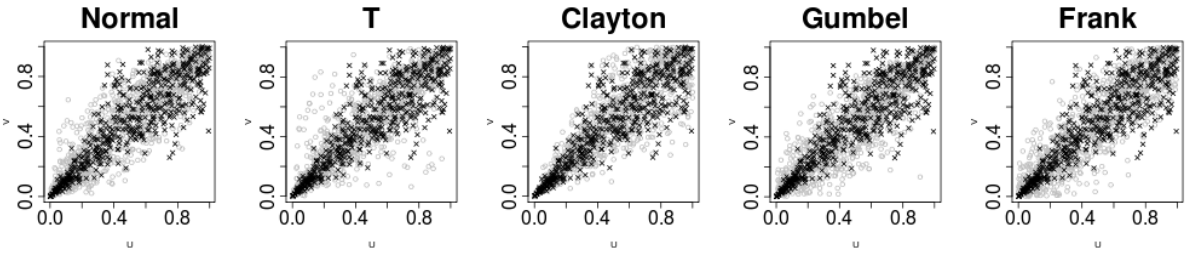


Figure 4.1: Scatter plot of 500 (u, v) generated from Clayton ($\theta = 5$, black "x") vs. 1000 data simulated from 5 different copulas with $\tau = 0.71$ (gray points). For the T copula the degree of freedom (df) has been fixed with 3.

Example 6 shows that the model selection with scatter plots can be successful, if the sample size is large. However, this result can be optimistic since our test data are generated directly from a copula. If the dependency structure of the data does not follows any investigated copula model very well, it could be harder to discover an appropriate copula model for the data at hand.

4.2 Contour Plots

A contour plot is a graphical technique for representing a three-dimensional surface by plotting constant z slices (contours) on a two-dimensional format as

$$\left\{k = f(x, y),\ k, x, y \in \mathbb{R}\right\} .$$

That is, given a value for k_i, lines are drawn for connecting the (x_i, y_i) coordinates where that k_i value occurs. The contour plot is an alternative to a 3-D surface plot. Since the contour plot is drawn from the function f, one can distinguish contour plots between different functions f. Let $f(x, y) = c(x, y)$, where $c(\cdot, \cdot)$ is the density function of copula $C \in \mathcal{C}$ and $\mathcal{C}$ is the set of all models we want to test. Then $k = c(u, v)$.

Consider a copula $C(u, v)$ for $u, v \in (0, 1)$. In this section, two kinds of contour plots will be studied, a contour plot on the original scale and a contour plot on the normal scale. The contour plot of

$$\{k = c(u, v),\ u, v \in (0, 1),\ k \in \mathbb{R}\}$$

is called a *contour plot on the original scale.* If we transform U and V to standard normal distributed random variables, i.e. $Z_1 := \Phi^{-1}(U)$ and $Z_2 := \Phi^{-1}(V)$, then the contour plot of

$$\{k = c(z_1, z_2),\ z_1 = \Phi^{-1}(u), z_2 = \Phi^{-1}(v),\ u, v \in (0, 1)\}$$

is called a *contour plot on the normal scale.*

Contour plots on the original scale

Original scale means random variables U and V have uniform distributions. Thus, the contour plots are plots of bivariate copula density functions. These functions have already

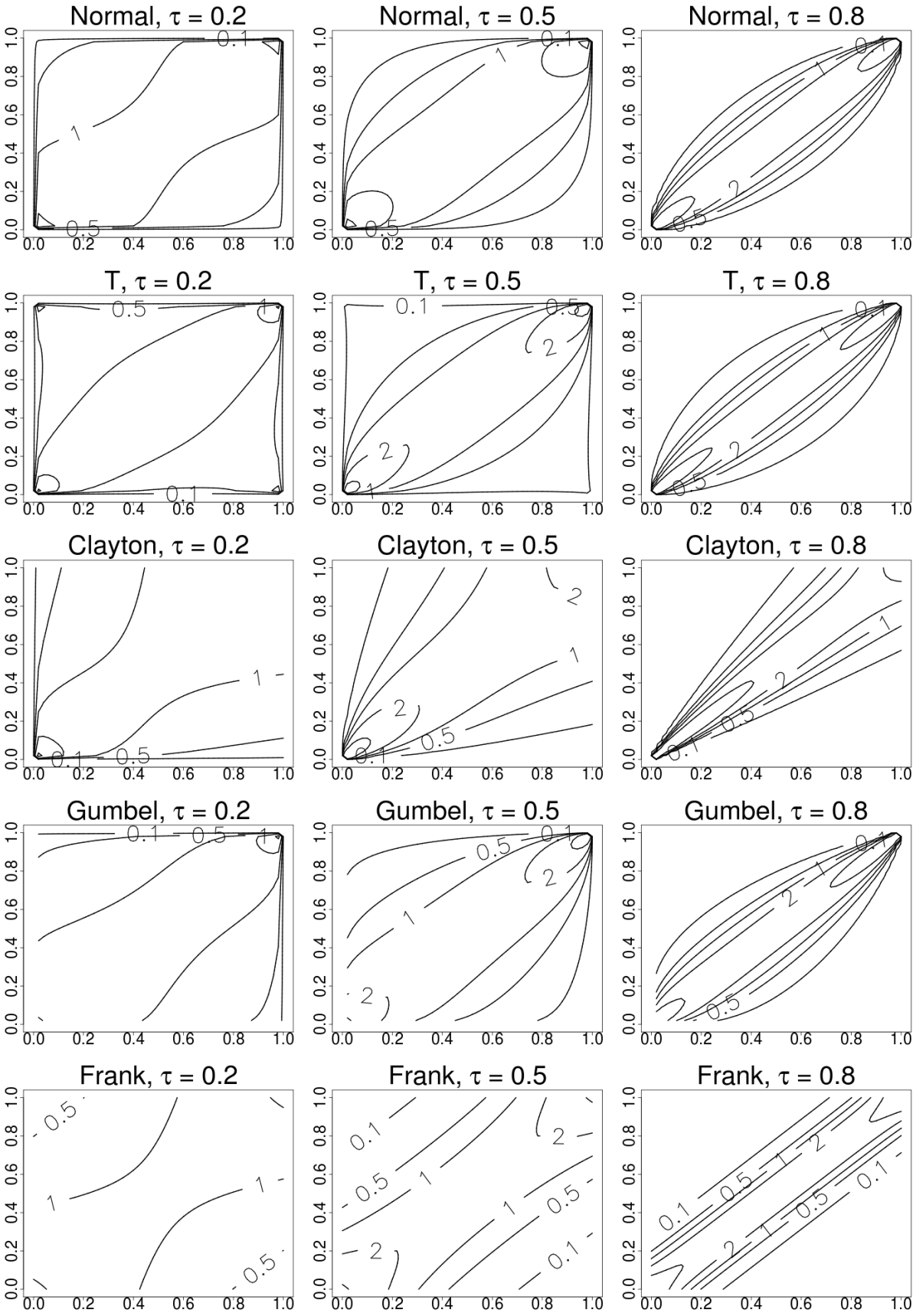


Figure 4.2: Theoretical contour plot with original scale of theoretical copula density function of Normal (1st row), T (2nd row) , Clayton (3rd row), Gumbel (4th row) and Frank (5th row) copulas. Values of Kendall's τ have been fixed with $\tau = 0.2$ (left column), 0.5 and $0,8$ (right column). In each contour plot, 5 levels $(0.1, 0.5, 1, 2, 5)$ were intended to be plotted, but some of them do not exist on the evaluated plotting range. For example, in the left panels with $\tau = 0.2$.

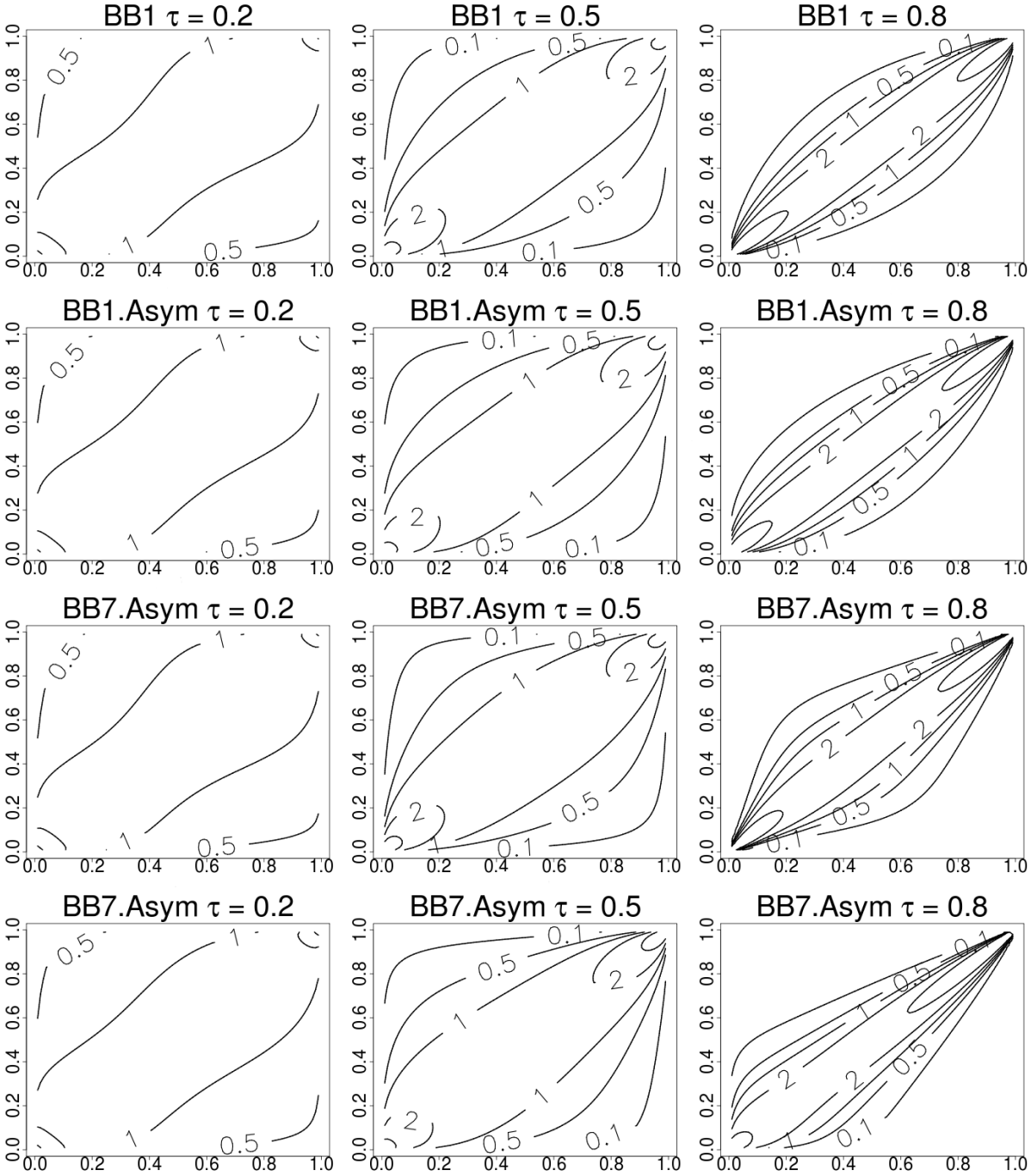


Figure 4.3: Theoretical contour plot with original scale of theoretical copula density function of BB1 with symmetrical tail dependence (1st row,$\Lambda_U = \Lambda_L$), BB1 with asymmetrical tail dependence (2nd row, $\Lambda_U = 2\Lambda_L$) , BB7 with symmetrical tail dependence (3rd row) and BB7 with asymmetrical tail dependence (4th row) copulas. Values of Kendall's τ have been fixed with $\tau = 0.2$ (left column), 0.5 and $0,8$ (right column). In each plot 5 contour levels $(0.1, 0.5, 1, 2, 5)$ were intended to be plotted, but some of them do not exist on the evaluated plotting range. For example, in the left panels with $\tau = 0.2$.

been listed in Table 2.2 and Table 2.4 for the copulas studied in this thesis. Figure 4.2 and Figure 4.3 displays the contour plots of these density functions using different Kendall's $\tau = 0.2, 0.5, 0.8$ with 5 levels $(0.1, 0.5, 1, 2, 5)$. Note that BB1 and BB7 copulas are investigated with two different parameter choices, respectively. For BB1 and BB7 copulas symmetrical tail dependence is defined by $\Lambda_U = \Lambda_L$ and asymmetrical tail dependence is defined by $\Lambda_U = 2\Lambda_L$, where Λ_u denotes the coefficient of upper tail dependence and Λ_L denotes the coefficient of lower tail dependence.

Contour plots on the normal scale

To construct a contour plot on the normal scale, we transform the copula data to data with standard normal margins. Contour plots based on this transformed data are more convenient to interpret, since the transformed data often have a unimodal density. This is not true for copula data. Departure from elliptical contour lines indicate departure from a normal copula. Thus, it is easier to distinguish between copula models with copula density $c_\theta(\cdot, \cdot)$.

More precisely, let U and V be $U(0, 1)$ distributed, then $Z_1 := \Phi^{-1}(U)$ and $Z_2 := \Phi^{-1}(V)$ are $\mathcal{N}(0, 1)$ distributed. The joint density of (z_1, z_2) is given by

$$f_\theta(z_1, z_2) = c_\theta(\Phi(u), \Phi(v)) \cdot \phi(u) \cdot \phi(v),$$

where $\phi(\cdot)$ is the standard normal density function. We call the distribution of (Z_1, Z_2) a meta copula distribution with standard normal margins. Table 4.1 gives the joint pdf of (z_1, z_2) based on several copula choices.

Figure 4.4 displays the theoretical contour plots of Normal, Student t, Clayton and Gumbel copulas with normal margins having fixed Kendall's $\tau = 0.2, 0.5, 0.8$. For T copula the degree of freedom has been fixed to 3. Figure 4.5 displays the theoretical contour plots of BB1 and BB7 with symmetrical and asymmetrical tail dependence. Symmetrical dependence means $\Lambda_U = \Lambda_L$ and asymmetrical dependence means $\Lambda_U = 2\Lambda_L$, where Λ_u stands for the coefficient of upper tail dependence and Λ_L stands for the coefficient of lower tail dependence, respectively. Both Figures consist of contour plots with 5 levels $(0.02, 0.04, 0.1, 0.2, 0.5)$.

In order to derive the empirical version of contour plots of any data, one needs to estimate the density function. The kernel smoothing method can be used here, where a normal kernel is chosen for the kernel smoothing.

Example 7. Test data is simulated from a Clayton copula with sample size $n = 1000$ and parameter $\theta = 2$, i.e. $\tau = 0.2$, and transformed to normal margins

Figure 4.6 is a contour plot of the estimated density function of the test data using Kernel smoothing. Comparing Figure 4.6 to the middle panel of row 3 of Figure 4.4 we can see that the contour plots are similar, showing that we can estimate the theoretical contour plots using data.

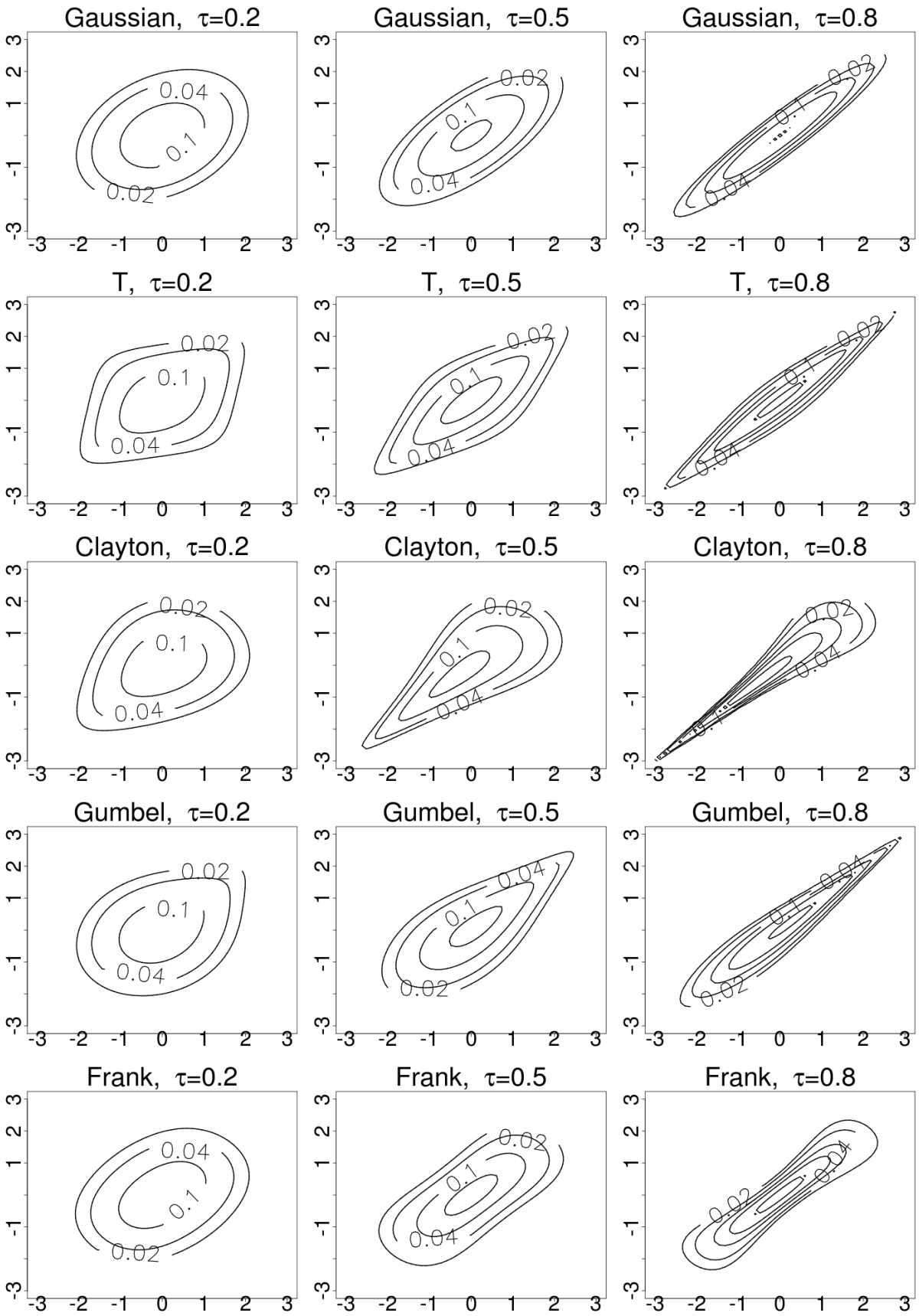


Figure 4.4: Theoretical contour plots of 5 different copulas with normal margins having fixed Kendall's τ: Normal (top row), T (2nd row), Clayton (3rd row), Gumbel (4th row), Frank (bottom row). Left column corresponds for $\tau = 0.2$. 2nd column corresponds for $\tau = 0.5$. right column corresponds for $\tau = 0.8$. In each plot, 5 contour levels $(0.1, 0.5, 1, 2, 5)$ were intended to be plotted, but some of them do not exist on the evaluated plotting range. For example, in the left panels with $\tau = 0.2$.

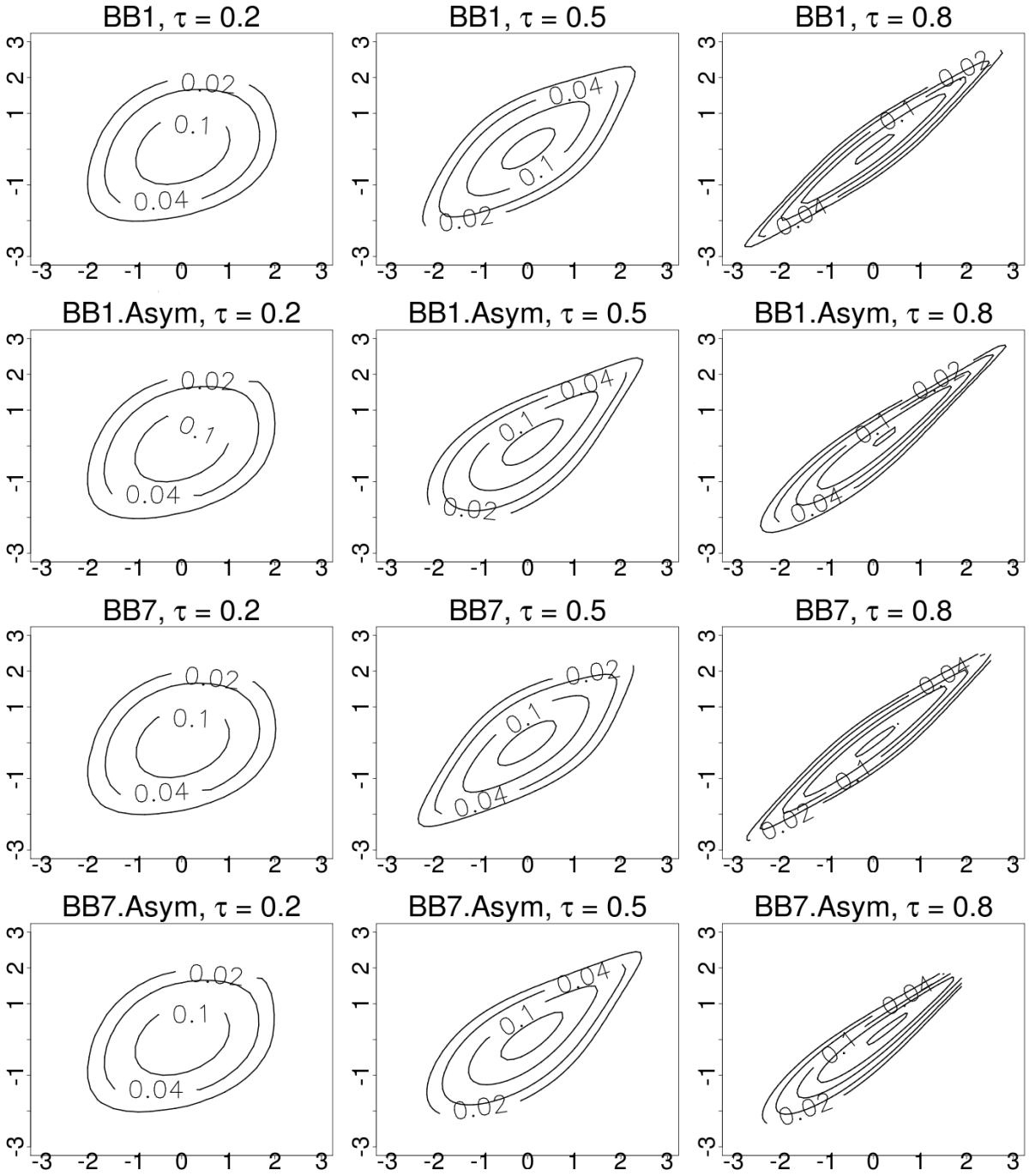


Figure 4.5: Theoretical contour plots of BB1 and BB7 copulas with normal margins having fixed Kendall's τ: BB1 with symmetrical tail dependence (top row, $\Lambda_U = \Lambda_L$),BB1 with symmetrical tail dependence (2nd row, $\Lambda_U = 2\Lambda_L$), BB7 with symmetrical tail dependence (3rd row, $\Lambda_U = \Lambda_L$) and BB7 with symmetrical tail dependence (bottom row, $\Lambda_U = 2\Lambda_L$). Left column corresponds for $\tau = 0.2$. 2nd column corresponds for $\tau = 0.5$. Right column corresponds for $\tau = 0.8$. In each plot, 5 contour levels $(0.1, 0.5, 1, 2, 5)$ were intended to be plotted, but some of them do not exist on the evaluated plotting range. For example, in the left panels with $\tau = 0.2$.

Table 4.1: Joint density of meta copula distributions with standard normal margins

Copula	Density function
Normal	$f_{(z_1,z_2)}(z_1,z_2) = \frac{2\pi}{\sqrt{1-\theta^2}} \exp\left\{\frac{z_1^2+z_2^2-2\theta z_1 z_2}{2(1-\theta^2)}\right\}$
T	$f_{(z_1,z_2)}(z_1,z_2) = \frac{2\pi\Gamma(\frac{\nu+2}{2})/\Gamma(\frac{\nu}{2})}{\nu\pi \mathrm{d}t(w_1,\nu)\mathrm{d}t(w_2,\nu)\sqrt{1-\theta^2}}$ $\times\left\{1+\frac{w_1^2+w_2^2-2\theta w_1 w_2}{\nu(1-\theta^2)}\right\}^{-\frac{\nu+1}{1}} \times \exp\left(\frac{z_1^2+z_2^2}{2}\right)$
Clayton	$f_{(z_1,z_2)}(z_1,z_2) = 2\pi(1+$ $\theta)(\Phi(z_1)\Phi(z_2))^{-1-\theta}\times\left(\Phi(z_1)^{-\theta}+\Phi(z_2)^{-\theta}-1\right)^{-\frac{1}{\theta}-2} \times \exp\left(\frac{z_1^2+z_2^2}{2}\right)$
Gumbel	$f_{(z_1,z_2)}(z_1,z_2) =$ $2\pi C_{12}(u_1,u_2)(u_1u_2)^{-1} \times \left[(-\ln u_1)^\theta+(-\ln u_2)^\theta\right]^{-2+\frac{2}{\theta}}\times(\ln u_1 \ln u_2)^{\theta-1} \times$ $\left\{1+(\theta-1)\left[(-\ln u_1)^\theta+(-\ln u_2)^\theta\right]^{-\frac{1}{\theta}}\right\} \times \exp\left(\frac{z_1^2+z_2^2}{2}\right)$ where $C_{12}(u_1,u_2) = \exp\left\{-\left[(-\ln u_1)^\theta+(-\ln u_2)^\theta\right]^{-\frac{1}{\theta}}\right\}$ and $(u_1,u_2) = (\Phi(z_1),\Phi(z_2))$.
Frank	$f_{(z_1,z_2)}(z_1,z_2) = 2\pi\theta g_1 \frac{1+g_{\Phi z_1+\Phi z_2}}{(g_{\Phi z_1}g_{\Phi z_2}+1)^2} \times \exp\left(\frac{z_1^2+z_2^2}{2}\right)$ where $g_z = e^{-\theta z-1}$.

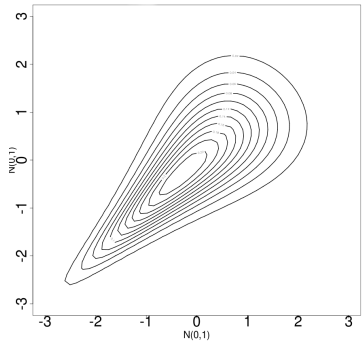


Figure 4.6: Contour plot of the estimated density function using kernel smoothing with a normal kernel

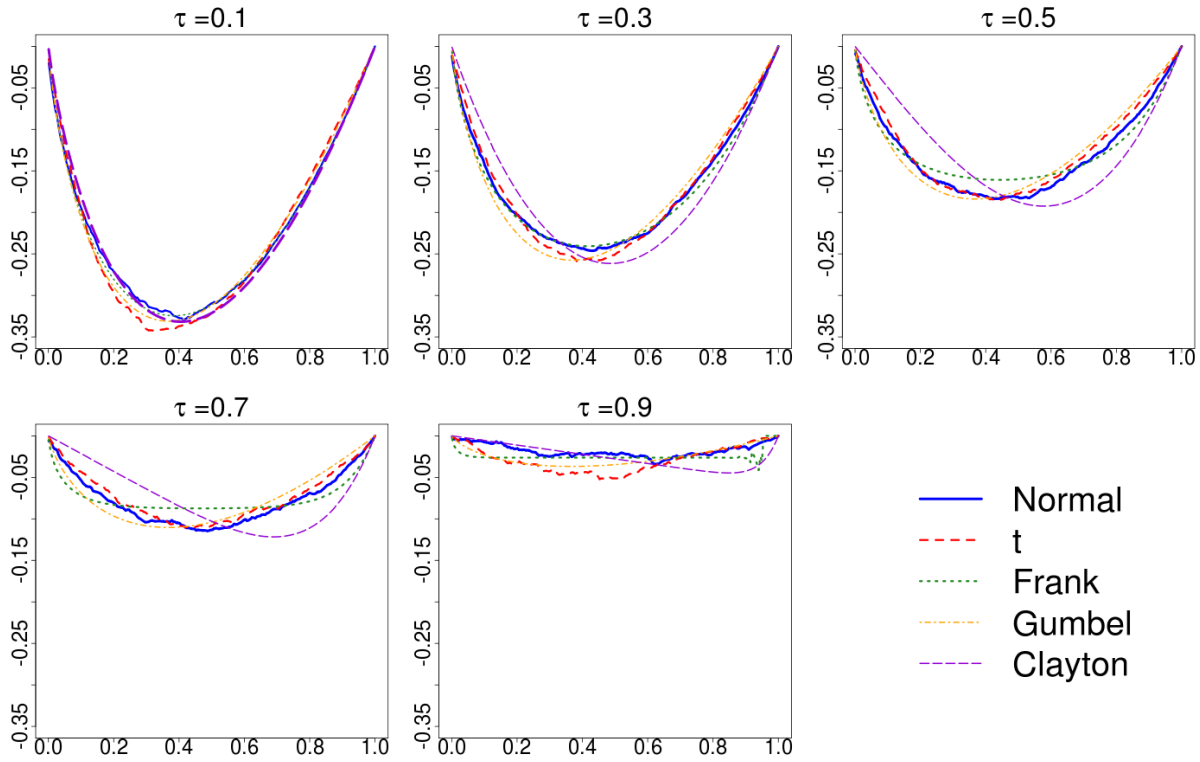


Figure 4.7: Theoretical $\lambda(t)$-plots for Normal (thick solid line), T with $df = 3$(thick dashed line), Frank (dotted line), Gumbel (dot-dashed line) and Clayton (long-dashed line) with $\tau = 0.1$(Top-left), 0.3(Top-middle), 0.5 (Top-right)m, 0.7 (Bottom-left) and 0.9 (Bottom-right). ($\lambda(t)$ was simulated for Normal and T based on a N=5000 sample.)

4.3 K-plots

K-plots consist of comparing the empirical distribution K_n given in (3.3) with the fitted theoretical distribution $K_{\hat{\theta}}$ defined in (3.2). Here we substitute $\hat{\theta}$ for θ. The procedure is to plot K_n and $K_{\hat{\theta}}$ on one graph and see how well they agree. Since $\lambda_\theta(t) = t - K_\theta(t)$, $\lambda_n(t)$ and $\lambda_{\hat{\theta}}(t)$ can also be compared. After a study of plotting λ and K functions, the plots of λ-function more clearly display the distance between lines drawn form functions than the Kendall distribution function K. This means that more efficient results can be obtained by plotting the $\lambda-$function. Therefore, in the following study and examples only $\lambda(t)$ will be plotted.

Figure 4.7 displays the behavior of theoretical $\lambda(t)$ for 5 different copulas: Normal, T, Clayton, Gumbel and Frank. Kendall's τ is set to be $\tau = 0.1, 0.3, 0.5, 0.7, 0.9$. Note that for Normal and T copulas $\lambda(t)$ is simulated with Eq. (3.4) with $N = 5000$. The second parameter of T copula, degrees of freedom, has been fixed to $df = 3$.

Figure 4.8 displays the behavior of theoretical $\lambda(t)$ for BB1 and BB7 copulas with symmetrical tail dependence ($\lambda_U = \lambda_L$) and asymmetrical tail dependence ($\lambda_U = 2\lambda_L$).

The copula parameters corresponding to $\tau = 0.1, 0.3, 0.5, 0.7, 0.9$ for the different copula families are listed in Table 4.2.

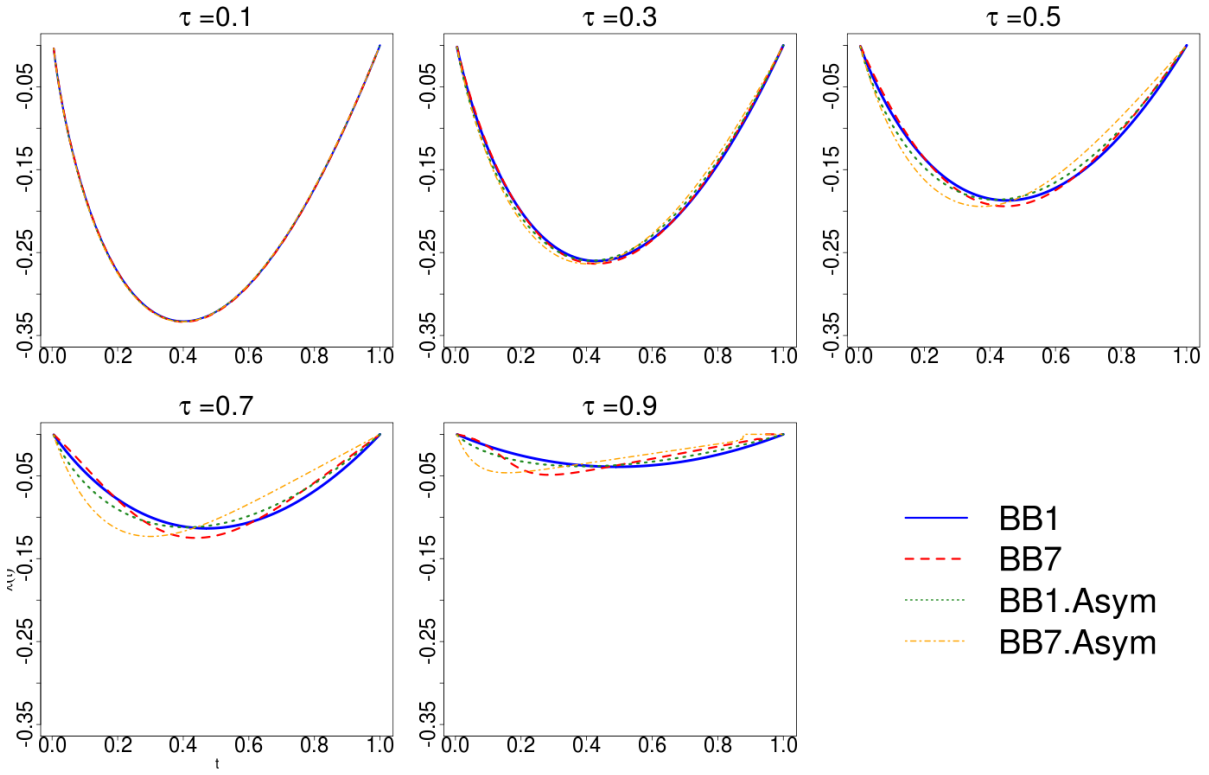


Figure 4.8: Theoretical $\lambda(t)$-plots for symmetric (i.e. $\lambda_U = \lambda_L$) BB1 (thick solid line), symmetric BB7 (thick dashed line), Frank (dotted line), asymmetric (i.e. $\lambda_U = 2\lambda_L$) BB1 (thin-dot-dashed line) and asymmetric BB7 (thin long-dashed line) with $\tau = 0.1$(Top-left), 0.3(Top-middle), 0.5 (Top-right)m, 0.7 (Bottom-left) and 0.9 (Bottom-right).

Table 4.2: Copula parameter values for different copula families having fixed Kendall's τ (for T copula df=3)

Parameter	Copula	$\tau = 0.1$	$\tau = 0.3$	$\tau = 0.5$	$\tau = 0.7$	$\tau = 0.9$
θ	Normal and T	0.16	0.45	0.71	0.89	0.99
θ	Clayton	0.22	0.86	2	4.67	18
θ	Gumbel	1.11	1.43	2	3.33	10
θ	Frank	0.91	2.92	5.73	11.39	38.09
θ	Sym. BB1 ($\lambda_u = \lambda_L$)	0.18	0.38	0.54	0.71	0.89
δ	Sym. BB1	1.02	1.20	1.56	2.42	6.59
θ	Sym. BB7	1.01	1.27	1.84	3.42	13.60
δ	Sym. BB7	0.19	0.54	1.14	2.73	12.91
θ	Asym. BB1 ($\lambda_u = 2\lambda_L$)	0.16	0.27	0.29	0.24	0.1
δ	Asym. BB1	1.02	1.25	1.73	2.94	9.08
θ	Asym. BB7	1.02	1.39	2.28	4.72	17.24
δ	Asym. BB7	0.18	0.4	0.61	0.8	0.94

Table 4.3: Estimated parameters in **Example 8** and **Example 9**

Example 8						Example 9					
Clayton: $\theta = 3,\ \tau = 0.6$						T: $\theta = 0.5, df = 3, \tau = 0.6$					
Copula	θ	δ	df	Λ_U	Λ_L	Copula	θ	δ	df	Λ_U	Λ_L
Normal	0.80	-	-	0	0	Normal	0.51	-	-	0	0
T	0.80	-	4.68	0.02	0.02	T	0.51	-	2.73	0.13	0.13
Clayton	3.0	-	-	-	0.79	Clayton	1.05	-	-	-	0.52
Gumbel	2.51	-	-	0.68	-	Gumbel	1.52	-	-	0.42	-
Frank	7.9	-	-	-	-	Frank	3.43	-	-	-	-
BB1	2.85	1.04	-	0.05	0.79	BB1	0.27	1.36	-	0.34	0.15
BB7	1.06	3.01	-	0.08	0.79	BB7	1.47	0.54	-	0.40	0.27

Examples

The following examples give graphical tests using the λ-function. The empirical distribution λ_n is compared with the theoretical one λ_θ of different copula families. The plot of λ_n and the plot of λ_θ should be close to each other, when the data are sufficiently abundant and the model is good.

The left graph in Figure 4.9 shows that the test in **Example 8** works well: Clayton fit the test data well, as well as BB1 and BB7. These 3 models can hardly be distinguished in the plot. For **Example 9**, The right graph in Figure 4.9 suggests T copula. The best model from the Archimedean Copula Class might be BB7. Table 4.3 lists the estimated parameters in both examples.

Example 8. Test data is simulated from a Clayton copula with parameter 3 and sample size 1000.

Example 9. Test data is simulated from a T copula with parameter $\theta = 0.5,\ df = 3$ and sample size 1000.

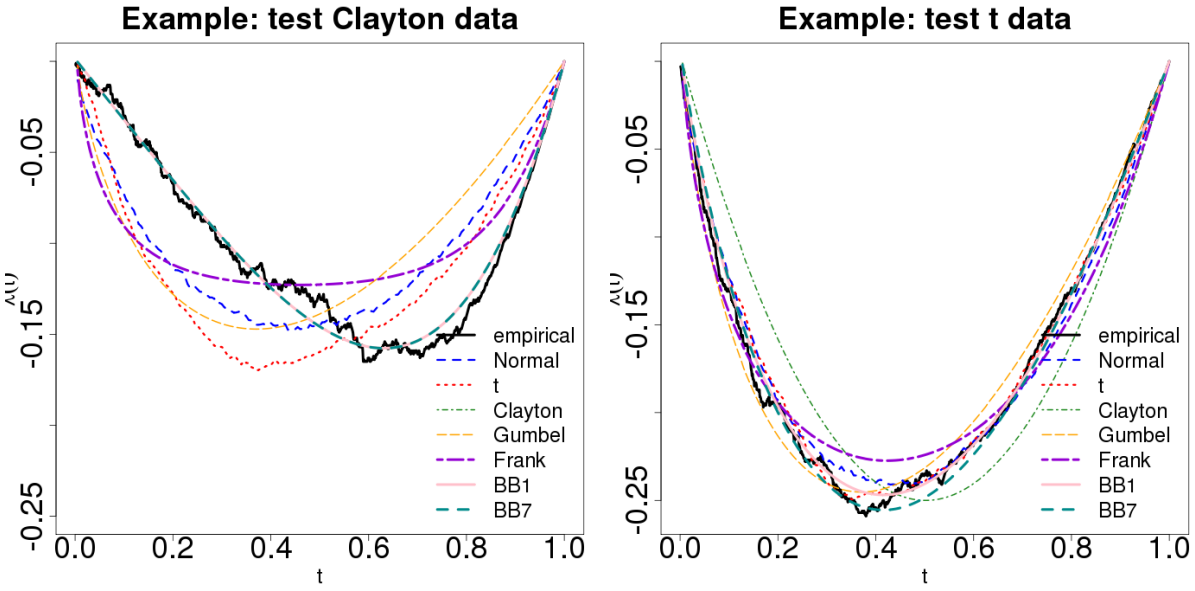


Figure 4.9: Examples of $\lambda_n(t)$ vs. $\lambda_{\hat{\theta}}(t)$: Example 8 (left) use simulated Clayton data (n=1000) with $\theta = 3$, $\tau = 0.6$. Example 9 (right) use simulated T data (n=1000) with $\theta = 0.5(\tau = 0.615)$ and $df = 3$

Chapter 5

Goodness of fit tests based on the copula and Kendall's process

This chapter describes two rank-based procedures that have been recently proposed for testing the goodness-of-fit (GOF) of any class of d-variate copulas. These tests are based on the empirical copula and the Kendall's process. Both are well established by Genest et al. [2006], Genest and Rémillard [2008] and Genest et al. [2009]. Many other possible Goodness-of-Fit tests for copulas have been proposed in literature, see Berg [2009] for details.

5.1 Copula goodness of fit

One fundamental problem for copulas is to determining whether a given family of copulas appropriately models the dependency structure of an observed data set. Hence, GOF tests are needed. Let $\mathbf{X}$ be a continuous d-variate random vector with distribution function F, margins $F_1, ..., F_d$, and unique underlying copula C. We want to test the null hypothesis

$$H_0 : C \in \mathcal{C} = \{C_\theta : \theta \in \mathcal{O}\}, \tag{5.1}$$

i.e., $C = C_{\theta_0}$ for some $\theta_0 \in \mathcal{O}$.

The GOF test for copulas is basically a special case of the general problem: testing multivariate distributed model, but is complicated due to the unspecified marginal distributions. The use of empirical margins introduces many nuisance parameters. This complicates the deduction of the asymptotic distribution properties for the tests. Thus p-values are found by bootstrap.

Generally, GOF tests for multivariate distribution functions based on an empirical process are bootstrap based, where the empirical process is a measure of the distance between the empirical distribution F_n and the theoretical distribution F_θ with a scale $\sqrt{n}$, i.e.

$$\mathbb{G} := \sqrt{n}(F_n - F_\theta).$$

Thus these tests are special cases of Algorithm 2.1. The bootstrap procedure for determining the p-value of a goodness of fit based on an empirical process $\mathbb{G}$ can be described as follows:

Given independent copies $\mathbf{X_1}, ..., \mathbf{X}_n$ of a random vector $\mathbf{X}$ with cumulative distribution function $F : \mathbb{R}^d \to \mathbb{R}$, suppose it is desired to test

$$H_0 : F \in \mathcal{F} = \{F_\theta : \theta \in \mathcal{O}\},$$

the hypothesis that F comes from a parametric family of distributions whose members are indexed by a parameter θ belonging to an open set $\mathcal{O} \subset \mathbb{R}^p$. To achieve this goal, a natural way to proceed consists of measuring the difference between the empirical distribution function F_n, defined for all $\mathbf{x} \in \mathbb{R}^d$ by

$$F_n(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(\mathbf{X}_i \leq \mathbf{x})$$

and a parametric estimate $F_{\hat{\theta}}$ of F derived under H_0 from estimate $\hat{\theta} = T_n(\mathbf{X}_1, ..., \mathbf{X}_n)$ of the true parameter value θ_0. Then the Goodness-of-Fit procedure is based on a continuous $S_n = f(\mathbb{G})$ of the fitted empirical process

$$\mathbb{G}_n := \sqrt{n}(F_n - F_{\hat{\theta}}),$$

where $f(\cdot)$ is a known continuous function. For the GOF tests introduced in next sections, f is considered as a specified integral of the squared empirical process.

To test the hypothesis or derive the P-value, Stute et al. [1993] suggest a bootstrap procedure as in Algorithm 5.1.

In the following, We will introduce two GOF tests which are based on two different specified empirical processes.

5.2 Two tests based on empirical copula process

Let $\mathbf{v}_1 = (v_{11}, \ldots, v_{n1}), \ldots, \mathbf{v}_d = (v_{1d}, \ldots, v_{nd})$ be $U(0,1)$ distributed random samples with copula C. Suppose it is desired to test the null hypothesis (5.1). Naturally, we want to compare the "distance" between the empirical copula C_n defined in (2.4), i.e.

$$C_n(\mathbf{u}) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{v_{i1} \leq u_1, \ldots, v_{id} \leq u_d\}, \quad \mathbf{u} = (u_1, \ldots, u_d) \in [0,1]^d, \tag{5.2}$$

and the parametric estimate $C_{\hat{\theta}}$, where $\hat{\theta}$ is an estimate of the unknown parameter θ. Based on this concept, Genest and Rémillard [2008] established the *empirical copula process*

$$\mathbb{C}_n := \sqrt{n}(C_n - C_{\hat{\theta}})$$

which measures the distance $C_n - C_{\hat{\theta}}$ with a scale $\sqrt{n}$.

Genest and Rémillard [2008] considered rank-based versions of the familiar Carmer von Mises and Kolmogorov Smirnov statistics in combination with $\mathbb{C}_n$:

$$S_n^{\mathbb{C}_n} := \int_{[0,1]^d} \mathbb{C}_n^2(\mathbf{u}) \mathrm{d}C_n(\mathbf{u}) \quad \text{and} \quad T_n^{\mathbb{C}_n} = \sup_{\mathbf{u} \in [0,1]^d} |\mathbb{C}_n(\mathbf{u})|. \tag{5.3}$$

Algorithm 5.1 Bootstrap procedure for goodness-of-fit based on empirical process

(i) Compute the empirical F_n with given $\mathbf{X}_1, \ldots, \mathbf{X}_n$.

(ii) Compute $\hat{\theta} = T_n(\mathbf{X}_1, ..., \mathbf{X}_n)$ and the parametric estimate $F_{\hat{\theta}}$.

(iii) Compute $S_n = f(\mathbb{G}_n)$.

For $k = 1, \ldots, B$,

(i) Generate n independent observations $\mathbf{X}^*_{1,k}, \ldots, \mathbf{X}^*_{n,k}$ from distribution $F_{\hat{\theta}}$.

(ii) Compute $\hat{\theta}^*_k = T_n(\mathbf{X}^*_{1,k}, ..., \mathbf{X}^*_{n,k})$ and the estimate $F_{\hat{\theta}^*_k}$.

(iii) Compute for each $\mathbf{x} \in \mathbb{R}^d$:

$$F^*_{n,k} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(\mathbf{X}^*_{i,k} \leq \mathbf{x}).$$

(iv) Compute $S^*_{n,k}$, where

$$S^*_{n,k} = f(\mathbb{G}^*_{n,k}) \text{ and } \mathbb{G}^*_{n,k} := \sqrt{n}(F^*_{n,k} - F_{\hat{\theta}^*_k}).$$

End for
Compute P-value:

$$p_{boot} = \frac{1}{B} \sum_{k=1}^{B} \mathbf{1}\{S^*_{n,k} \geq S_n\}.$$

We call GOF tests based on these statistics *tests based on the empirical copula process.*

The distributions of the test statistics $S_n^{\mathbb{C}_n}$and $T_n^{\mathbb{C}_n}$ depend in particular on the copula under H_0 and the unknown parameter θ, therefore a distribution of the test statistic cannot be tabulated and p-values can only be obtained via Monte Carlo methods. Large values of these statistics lead to a rejection of H_0. Genest et al. [2006] proved the convergence of the process $\mathbb{C}_n$ and showed that tests based on $S_n^{\mathbb{C}_n}$ and $T_n^{\mathbb{C}_n}$ are consistent: if $C \notin \mathcal{C}$, then H_0 would be rejected with probability 1 as $n \to \infty$. The specific bootstrap procedure for the tests based on empirical copula process is described as in Algorithm 5.2.

The GOF test with respect to the statistic $S_n^{\mathbb{C}_n}$ has been implemented as a R-function *gofCopula()* in the R-package "copula". Only five copula families, Normal, T, Clayton, Gumbel and Frank, are applied in this package.

5.3 Two tests based on Kendall's transform

These tests are explored by Genest and Rivest [1993], and Wang and Wells [2000]. Let $\mathbf{X} = (X_1, \ldots, X_d)$ be a continuous d-variate random vector with distribution function F, margins $F_1, \ldots, F_d$ and unique underlying copula C. Let $U_i := F_i(X_i)$ for $i = 1, \ldots, d$, then the joint distribution of $\mathbf{U} = (U_1, \ldots, U_d)$ is C. Suppose we are interested in Hypothesis (5.1), i.e.

$$H_0 : C \in \mathcal{C} = \{C_\theta : \theta \in \mathcal{O}\}.$$

Now under H_0, the vector $\mathbf{U}$ is distributed as C_θ for some $\theta \in \mathcal{O}$. Let K_θ denote the Kendall distribution function of C_θ, and K_n denote the corresponding empirical Kendall distribution function which is an estimator of K_θ. Hence, $C_\theta(\mathbf{U})$ has distribution K_θ. Through the Kendall process

$$\mathbb{K}_n(t) = \sqrt{n}(K_n(t) - K_{\theta_n}(t))$$

one can test

$$H_0^{''} : K \in \mathcal{K}_0 = \{K_\theta : \theta \in \mathcal{O}\}.$$

Because $H_0 \subset H_0^{''}$, the non-rejection of $H_0^{''}$ does not entail the acceptance of H_0. Consequently, tests based on the empirical process $\mathbb{K}_n(t)$ are not generally consistent. However, in the case of bivariate Archimedean copulas $H_0^{''}$ and H_0 are equivalent. Recall the definition of K for Archimedean Copula Class, we see that $K(t)$ is uniquely defined with the generator $\phi(t)$, i.e.

$$K(t) = \frac{\phi(t)}{\phi'(t)}.$$

The foregoing expression yields the inversion

$$\phi(t) = \exp\Big[\int_{t_o}^{t} \frac{1}{v - K(v)} dv\Big],$$

where $0 < t_0 < 1$ is an arbitrary constant. Thus the Archimedean generator ϕ is completely determined by K. In the meantime, the definition of Archimedean Copula Class

$$C_\theta(u, v) = \phi_\theta^{-1}(\phi_\theta(u) + \phi_\theta(v))$$

Algorithm 5.2 Parametric bootstrap for $S_n^{\mathbb{C}_n}$ and $T_n^{\mathbb{C}_n}$

M is our n-dimensional data set. $\mathcal{C}$ is the set of models to be tested.

(i) Compute the observed empirical copula C_n with data set M.

(ii) Select a model $C_\theta \in \mathcal{C}$ ($H_0 : C = C_\theta$ for some $\theta \in \mathcal{O}$).

(iii) Estimate the parameter θ of C_θ and denote the estimate as $\hat{\theta}$.

(iv) Compute the observed value of $S_n^{\mathbb{C}_n}$or $T_n^{\mathbb{C}_n}$in (5.3) with C_n and $C_{\hat{\theta}}$.

(v) **For** k=1:B,

(a) Generate a random sample $M_k = (\mathbf{U}_{1,k}^*, \ldots, \mathbf{U}_{n,k}^*)$ from distribution C_{θ_n}.

(b) Compute the empirical copula using samples

$$C_{n,k}^*(\mathbf{u}) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{\mathbf{U}_{i,k}^* \leq \mathbf{u}\}$$

at each $\mathbf{u} \in [0,1]^d$.

(c) Use the sample M_k and the same method as before to estimate the parameter $\hat{\theta}_k^*$of C_θ.

(d) Compute the value of $S_{n,k}^{\mathbb{C}_n *}$ or $T_{n,k}^{\mathbb{C}_n *}$ with $C_{n,k}^*$ and $C_{\hat{\theta}_k^*}$.

End for

(vi) Approximate the P-value for the testing problem (5.1)

$$p_{boot} = \frac{1}{B}\sum_{k=1}^{B} \mathbf{1}\{S_{n,k}^{\mathbb{C}_n *} \geq S_n^{\mathbb{C}_n}\}$$

using $S_n^{\mathbb{C}_n}$, or

$$p_{boot} = \frac{1}{B}\sum_{k=1}^{B} \mathbf{1}\{T_{n,k}^{\mathbb{C}_n *} \geq T_n^{\mathbb{C}_n}\}$$

using $T_n^{\mathbb{C}_n}$.

shows that Archimedean copulas are uniquely determined by their generator ϕ. Therefore

$$C = C_\theta \Leftrightarrow K = K_\theta.$$

More discussion about this limitation can be found in Wang and Wells [2000] andGenest et al. [2006]. The specific test statistics for this GOF test are given by

$$S_n = \int_0^1 |\mathbb{K}_n(t)|^2 \mathtt{d}K_{\theta_n}(t) \quad \texttt{and} \quad T_n = \sup_{0 \leqslant t \leqslant 1} |\mathbb{K}_n(t)|. \tag{5.4}$$

Since $\mathbb{K}_n$ is an empirical process, these tests are special cases of Algorithm 5.1.Genest et al. [2006] show that, the expressions of S_n and T_n can be computed as follows:

$$S_n = \frac{n}{3} + n \sum_{j=1}^{n-1} K_n^2(\frac{j}{n}) \left\{ K_{\theta_n}(\frac{j+1}{n}) - K_{\theta_n}(\frac{j}{n}) \right\} - n \sum_{j=1}^{n-1} K_n(\frac{j}{n}) \left\{ K_{\theta_n}^2(\frac{j+1}{n}) - K_{\theta_n}^2(\frac{j}{n}) \right\} \tag{5.5}$$

and

$$T_n = \sqrt{n} \max_{i=0,1; 0 \leqslant j \leqslant n-1} \left| -K_n(\frac{j}{n}) - K_{\theta_n}(\frac{j+i}{n}) \right|. \tag{5.6}$$

Proof. Eq. (5.5) and (5.6) can be straightforward calculated using discretization

$$\begin{aligned} T_n &= \sup_{0 \leq t \leq 1} |\mathbb{K}_n(t)| \\ &= \sqrt{n} \sup_{0 \leq t \leq 1} |K_n(t) - K_{\theta_\mathbf{n}}(t)| \\ &\overset{\text{discretization}}{=} \sqrt{n} \max_{i=0,1;\ 0 \leq j \leq n-1} \left| K_n(\frac{j}{n}) - K_{\theta_n}(\frac{j+i}{n}) \right|, \end{aligned}$$

$$\begin{aligned} S_n &= \int_0^1 |\mathbb{K}_n(t)|^2 \mathtt{d}K_{\theta_n}(t) \\ &= n \int_0^1 \left[K_n^2(t) - 2K_n(t)K_{\theta_n}(t) + K_{\theta_n}^2 \right] \mathtt{d}K_{\theta_n}(t) \\ &= n \int_0^1 K_n^2(t) \mathtt{d}K_{\theta_n}(t) - n \int_0^1 2K_n(t)K_{\theta_n}(t) \mathtt{d}K_{\theta_n}(t) + n \int_0^1 K_{\theta_n}^2 \mathtt{d}K_{\theta_n}(t) \\ &= n \int_0^1 K_n^2(t) \mathtt{d}K_{\theta_n}(t) - n \int_0^1 K_n(t) \mathtt{d}K_{\theta_n}^2(t) + \frac{n}{3} \\ &\overset{\text{discretization}}{=} \frac{n}{3} + n \sum_{j=1}^{n-1} K_n^2(\frac{j}{n}) \left\{ K_{\theta_n}(\frac{j+1}{n}) - K_{\theta_n}(\frac{j}{n}) \right\} \\ &\quad -n \sum_{j=1}^{n-1} K_n(\frac{j}{n}) \left\{ K_{\theta_n}^2(\frac{j+1}{n}) - K_{\theta_n}^2(\frac{j}{n}) \right\}. \end{aligned}$$

□

Algorithm 5.3 Bootstrap procedure of Goodness-of-fit based on Kendall's transform in the case for S_n

M is our n-dimensional data set and $\mathcal{C}$ is the set of models we want to test.

(i) Compute the empirical K_n (Eq. (3.3)) with data set M.

(ii) Select a model $C_\theta \in \mathcal{C}$ ($H_0 : C = C_\theta$ for some $\theta \in \mathcal{O}$).

(iii) Use M to estimate the copula parameter θ of C_θ. Denote the estimated parameter as $\hat{\theta}$ and the selected model as $C_{\hat{\theta}}$.

(iv) Compute the theoretical $K_{\hat{\theta}}$ of the selected model $C_{\hat{\theta}}$.

(v) Use K_n and $K_{\hat{\theta}}$ to compute the observed value of the test statistic S_n in Eq. (5.5).

(vi) **For** k=1:B,

(a) Generate a random sample M_k from $C_{\hat{\theta}}$. M_k has the same size as M.

(b) Use sample M_k to estimate θ of C_θ and denote the estimate as $\hat{\theta}^*_k$.

(c) Compute the empirical $K^*_{n,k}$ from sample M.

(d) Compute the theoretical $K^*_{\hat{\theta}_k}$ of $C_{\hat{\theta}^*_k}$.

(e) Determine the value of the test statistic $S^*_{n,k}$ with $K^*_{n,k}$ and $K^*_{\theta_{n,k}}$.

(vii) If $S^*_{1:B} \leqslant \ldots \leqslant S^*_{B:B}$ denote the ordered values of the test statistics $S^*_{n,1}, \ldots, S^*_{n,B}$, an estimate of the critical value of the test at level α based on S_n is given by

$$S^*_{\lfloor (1-\alpha)N \rfloor : N} \quad \texttt{and} \quad \frac{1}{B}\#\{j : S^*_j \geqslant S_n\}$$

yields an estimate of the P-value associated with the observed value S_n of the statistic.

The use of these statistics has the advantage that simple formulas are available for S_n and T_n in terms of the ranks of the observations, which is not the case for the empirical copula process. Formal testing procedures based on these statistics consist of rejecting $H_0 : C \in \mathcal{C}$ when the observed value of S_n or T_n is greater than the $100(1-\alpha)$th percentile of its distribution under the null hypothesis. As the asymptotic distributions of S_n and T_n both depend on the unknown copula C_θ and on θ, approximate P- values for these statistics must be found again via simulation. The bootstrap methodology required to compute associated P-values proceeds as in Algorithm 5.3. This algorithm only shows the case for S_n. For T_n is analogous.

5.4 Simulation studies

Since we introduced the tests based on Kendall's transform, In this section a large number of repeated simulation studies are conducted to assess the performance of these tests for various classes of copula models under the null hypothesis and under the alternative. We are interested in two characteristics of the tests: their ability to maintain their nominal level, arbitrarily fixed at $\alpha = 5\%$ throughout the study, and their power under a variety of alternatives.

Scenario design

For this investigation four scenarios are designed. In each scenario, one bivariate copula family is specified by $\tau = 0.5$ and by a tail dependence parameter if any, and is used under the null hypothesis. Seven copula models, Normal (N), T, Clayton (C), Gumbel (G), Frank (F), BB1 and BB7 copulas, are considered under the alternative in every test throughout the simulation studies.

Each scenario begins with selecting a bivariate copula model C_θ and generating a random data set M with size n from C_θ. The GOF tests based on Kendall's transform are conducted with this data set for the null hypothesis

$$H_0 : C^* \in \mathcal{C}' := \{C(\cdot,\cdot|\theta), \theta \in \Theta\}, \tag{5.7}$$

where $C(\cdot,\cdot|\theta)$ denote the specified bivariate parametric copula family and is selected from $\{C^N, C^T, C^C, C^G, C^F, C^{BB1}, C^{BB7}\}$. For each of these seven different $C^* \in \mathcal{C}'$, two P-values $p^{S_n(C^*)}$ and $p^{T_n(C^*)}$ are estimated by using Algorithm (5.3). At last, for every scenario with the choice of C_θ we replicate this concept R times and compare the estimated p-values with a pre-selected significant level $\alpha = 0.05$. The entire process is described as in Algorithm (5.4).

The four scenarios are summarized in Table 5.1. Since we want to see the influence by using different data size additionally, each scenario is separated into two sub scenarios depending on the data size.

Since we fixed $\tau = 0.5$ for all scenarios and additionally $\Lambda_U = 2\Lambda_L$ for Sc2 , the corresponding copula parameters and their tail dependence in every scenario can be directly determined using formulas in Table 3.1 and Table 2.1. All of these values are listed in Table 5.2.

Algorithm 5.4 Scenario design for tests based on Kendall's transform

Select a bivariate copula C_θ and fix the parameter θ with a selected Kendall's τ and a second parameter condition if any. Let R be the number of replication.
For $k = 1 : R$

(i) Generate a sample data set from C_θ and denoted as M_k.

(ii) For every $C^* \in \mathcal{C}' := \{C^N, C^T, C^C, C^G, C^F, C^{BB1}, C^{BB7}\}$:

(a) determine the parameter estimate $\hat{\theta}$ of C^* and denote the corresponding model as $C^*_{\hat{\theta}}$.

(b) Run Algorithm 5.3 with M_k to estimate $p_k^{S_n(C^*_{\hat{\theta}})}$ and $p_k^{T_n(C^*_{\hat{\theta}})}$.

(c) Compare $p_k^{S_n(C^*_{\hat{\theta}})}$ and $p_k^{T_n(C^*_{\hat{\theta}})}$ to a significant level α, respectively.

Table 5.1: Selected scenarios

Scenario	Sub scenario	Data comes from	Parameter choice	Data size n
Sc1	**Sc1-1**	bivariate Normal	$\tau = 0.5$	200
	Sc1-2	copula family		800
Sc2	**Sc2-1**	bivariate BB1	$\tau = 0.5$	200
	Sc2-2	copula family	$\Lambda_U = 2\Lambda_L$	800
Sc3	**Sc3-1**	bivariate Clayton	$\tau = 0.5$	200
	Sc3-2	copula family		800
Sc4	**Sc4-1**	bivariate Gumbel	$\tau = 0.5$	200
	Sc4-2	copula family		800

Table 5.2: Parameters for the true copula models in four scenarios: Sc1 - Sc4.

Models	θ	δ	Λ_U	Λ_L
N (Sc1)	0.70	-	0	0
BB1 (Sc2)	0.29	1.73	0.51	0.25
C (Sc3)	2	-	0	0.707
G (Sc4)	2	-	0.586	0

Table 5.3: The average values of 100 estimated parameters using MLE and based on 7 copula families for each scenarios

Model		N	T		C	G	F	BB1		BB7	
True	Sc	θ	θ	$\bar{\nu}$	θ	θ	θ	θ	δ	θ	δ
N	Sc1-1	0.700	0.698	18.723	1.983	1.991	5.649	0.428	1.570	1.700	0.908
	Sc1-2	0.700	0.696	20.000	1.952	1.976	5.575	0.417	1.561	1.692	0.889
BB1	Sc2-1	0.700	0.706	9.043	1.986	1.993	5.644	0.290	1.750	1.985	0.823
	Sc2-2	0.699	0.704	7.574	1.950	1.975	5.574	0.288	1.729	1.956	0.808
C	Sc3-1	0.708	0.697	7.522	2.043	2.021	5.780	1.924	1.037	1.047	2.011
	Sc3-2	0.705	0.693	6.053	1.991	1.995	5.663	1.938	1.016	1.022	1.978
G	Sc4-1	0.702	0.697	9.142	1.998	1.999	5.672	0.066	1.930	2.223	0.505
	Sc4-2	0.705	0.702	8.251	1.994	1.997	5.665	0.023	1.974	2.275	0.457

For processing these scenarios, we set the number of replications to 100 (i.e. 100 random samples are generated in each sub scenario) and the number of bootstraps in each test to 1000, i.e. $R = 100$ and $B = 1000$. For simulating the theoretical Kendall function $K_{\hat{\theta}}$ of Normal and T copulas in the second bootstrap level we use $N = 5000$. For the copula parameter estimation we use the Maximum Likelihood Estimation (MLE) with the R-function *CopulaEstimator2D()* in the R-package *CDvineMLE()* provided by Dr. Almeida and Mr. Schepsmeier.

In each sub scenario the parameters of seven copula families are estimated by using 100 random samples, respectively. Three of these families have two parameters. Thus, we obtain $100 \cdot (1 \cdot 4 + 2 \cdot 3) = 1000$ parameter estimates for each sub scenario. Table 5.3 reports the average values of the 100 estimated parameters of each alternative copula model for every sub scenario, i.e.

$$\bar{\mu} := \frac{1}{R} \sum_{k=1}^{R} \hat{\mu}_k \text{ for } \mu \in \{\theta, \nu, \delta\} \text{ and } R = 100.$$

Table 5.3 shows that the estimated parameters of one-parameter copula families are nearly unchanged between different copula families. This is not surprising, since we fixed $\tau = 0.5$ for all scenarios.

For the 100 parameter estimates of every model for different scenarios, their histograms are plotted in Figure 5.1 and their boxplots are given in Figure 5.2 and Figure 5.3. Both kinds of plots report that the variation of the estimated values is generally symmetric and decreasing when data size is larger (from $n = 200$ to $n = 800$), respectively. Since a histogram counts the number of observations that fall into each of the disjoint categories and represents the density of the underlying data, we can receive a visual impression of the distribution with the help of the histogram. Especially, Figure 5.1 shows the estimates are centralized to their corresponding mean in the most cases. Therefore, the normal distribution is considered for each case.

Table 5.4 reports the average values of empirical Kendall's τ and Figure 5.4 lists the histograms and boxplots of empirical Kendall's τ for different scenarios.

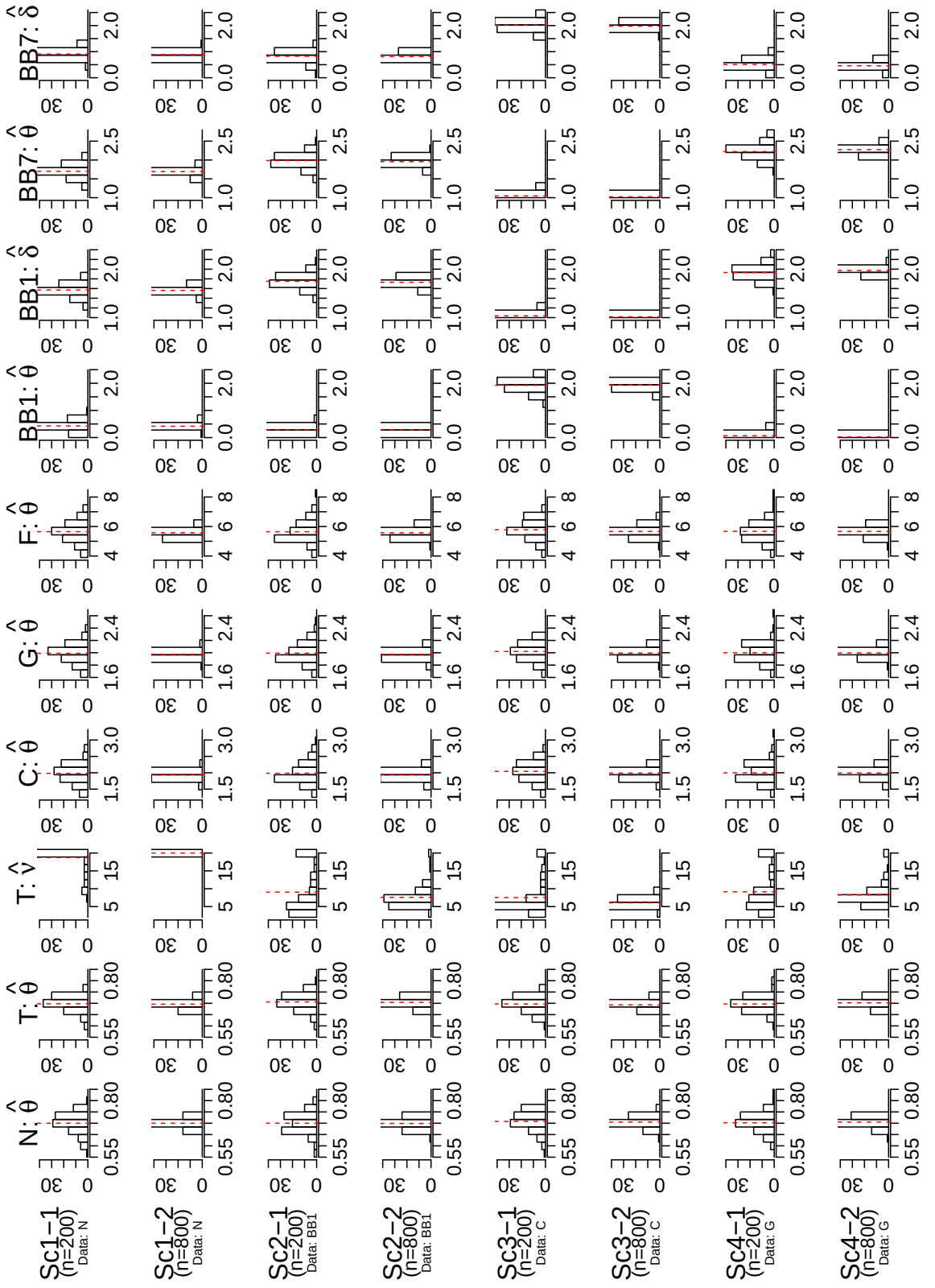


Figure 5.1: Histograms of the estimated parameters of the 7 alternative copula models in all 8 sub scenarios. The red vertical dashed lines show the locations of $E(\hat{\theta})$, $E(\hat{\nu})$ or $E(\hat{\delta})$ in the corresponding plots.

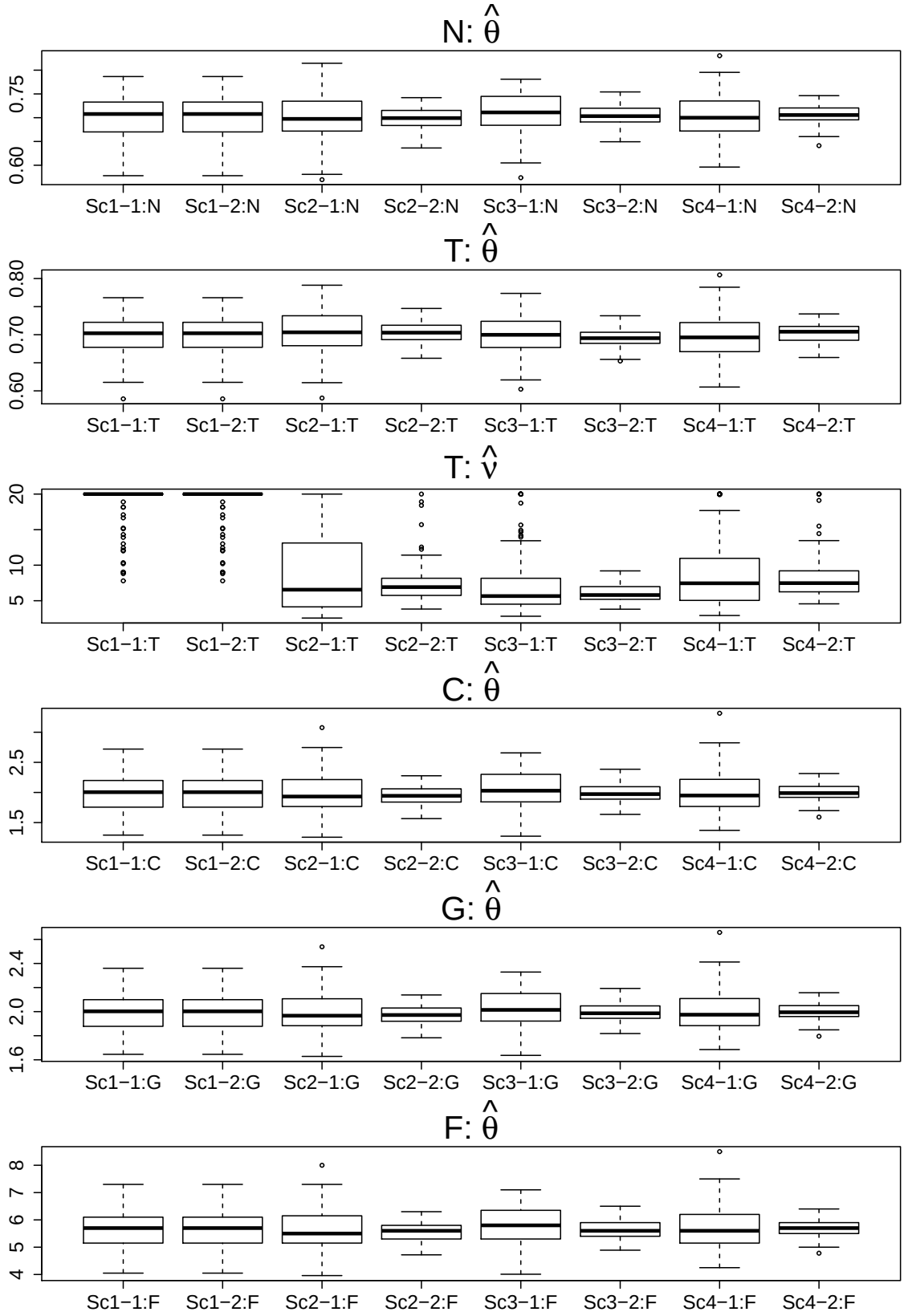


Figure 5.2: Cross comparison of the boxplots of the estimated parameters from the same copula models in different sub scenarios. For the models: Normal, T, Clayton, Gumbel and Frank copulas. (Part 1)

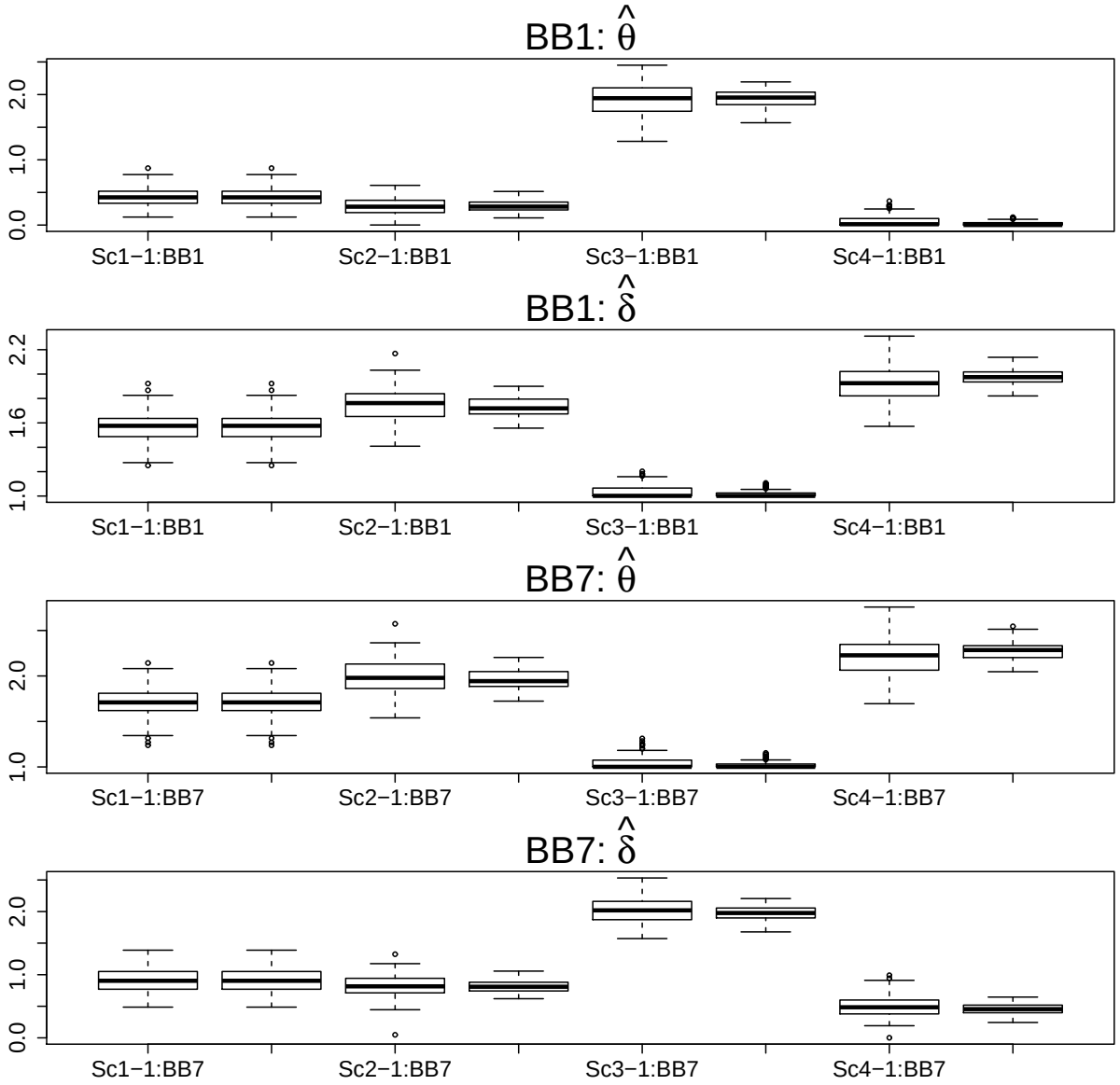


Figure 5.3: (Continued) Cross comparison of the boxplots of the estimated parameters from the same copula models in different sub scenarios. For the models: BB1 and BB7 copulas. (Part 2)

Table 5.4: The average values of empirical Kendall's τ for each scenario

Sc	Sc1-1	Sc1-2	Sc2-1	Sc2-2	Sc3-1	Sc3-2	Sc4-1	Sc4-2
$\bar{\tau}$	0.495	0.493	0.495	0.493	0.502	0.498	0.497	0.499

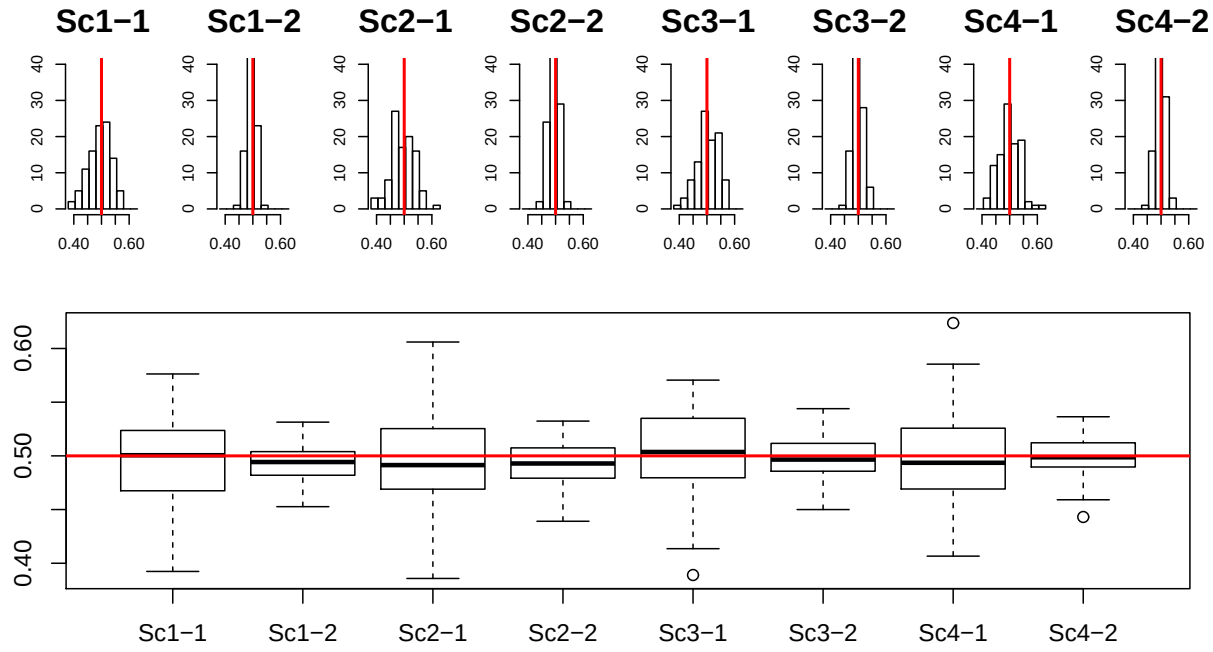


Figure 5.4: Histograms and boxplots of empirical Kendall's τ in every sub scenario. Red lines corresponds to $\tau = 0.5$.

Results

Once again, eight sub scenarios are conducted depending on different data size. In the 100 replications of each sub scenario, seven models need to be tested and testing each model delivers two different p-values as results, i.e. from S_n and T_n. Thus, we receive $7 \cdot 2 \cdot 100 = 1400$ p-values for each sub scenario in total.

Figures 5.5, 5.6, 5.7 and 5.8 display the histograms of the p-values estimated from 100 replications for each model in all four scenarios, respectively. Remember that each scenario is separated into two sub scenarios with size 200 and 800. Seven candidate models and two different p-values for each test lead to $2 \cdot 7 \cdot 2 = 28$ Histograms in each figure.

The histograms show that p-values of the true model in each scenario are more or less uniformly distributed. If the alternative models exist which are very close to the true model, their corresponding p-values are uniformly distributed as well. The question about whether two models are close to each other will be discussed later.

Additionally, Figures 5.5 - 5.8 also report that Statistic S_n derives a higher p-value than T_n in most cases. For example, one can compare the histogram for the BB1 copula in case Sc1-1_Sn to the histogram for the BB1 copula in case Sc1-1_Tn in Figures 5.5. With this conclusion, S_n will be considered as a more efficient result than T_n and we prefer to compute S_n via parametric bootstrapping described in Algorithm (5.3) to estimate p-values. Studies in Genest et al. [2009] also confirmed the same conclusions.

After p-values are estimated, the most common way to assess the significance of GOF tests is to compare their p-values to a significance level, respectively. We selected $\alpha = 0.05$ here. Small p-value leads to a rejection of the tested model. Equivalently, the p-value should be high, if the tested copula family is our true copula family. Thus, we can count

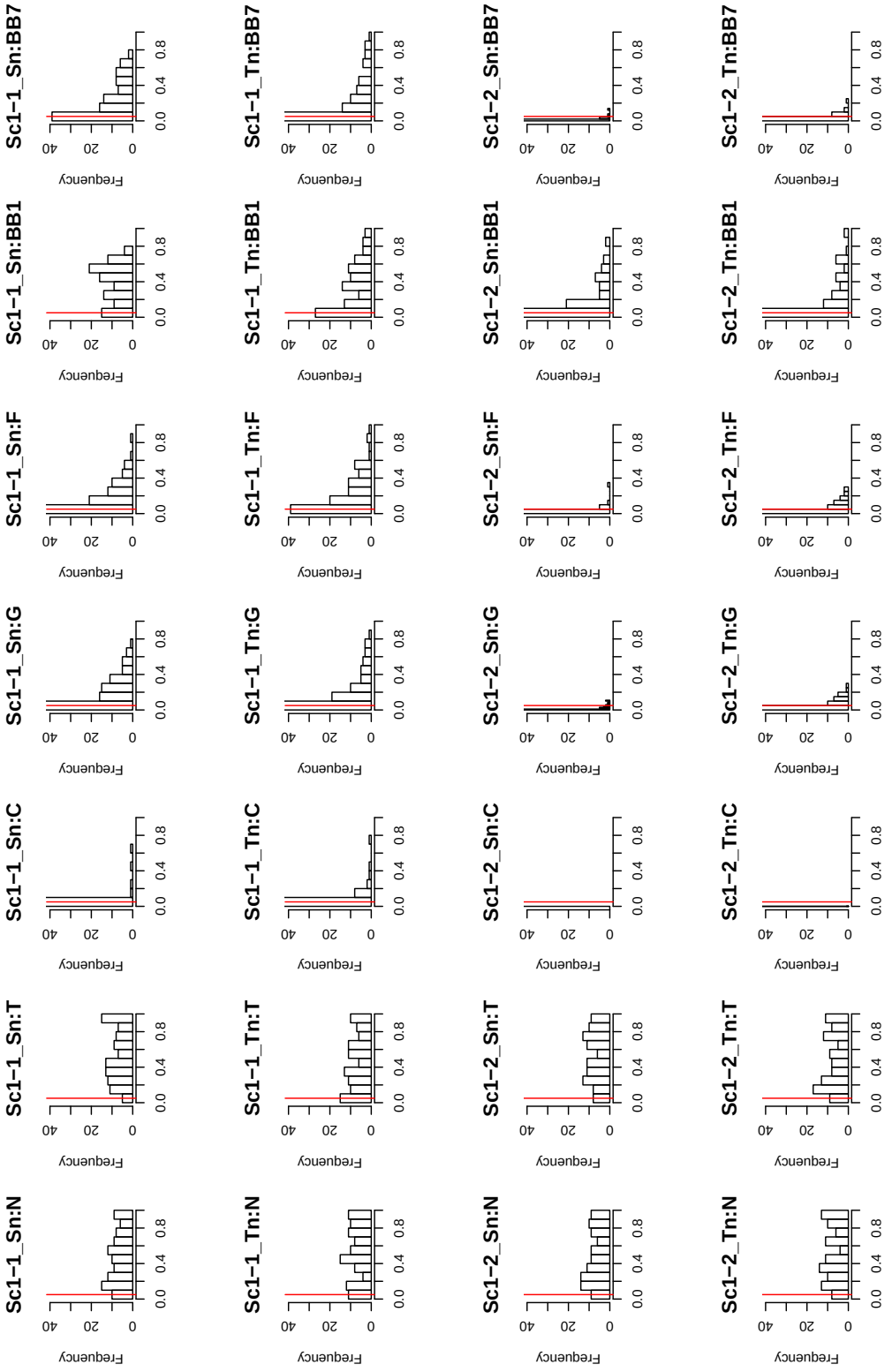


Figure 5.5: Scenario 1: Data set comes from Normal copula with $\tau = 0.5$. Sc1-1 denotes the sub scenario corresponding to a data set with $n = 200$ and SC1-2 denotes the sub scenario corresponding to a data set with $n = 800$. Each Histogram of p-values is plotted with a vertical line $\alpha = 0.05$. 7 plots in each row stand for 7 test models, respectively. They are N, T, C ,G, F, BB1 and BB7 from left to right. 1st row lists the plots of p-value of S_n for these 7 test models in Sc1-1, 2nd row for p-value of T_n in Sc1-1, 3rd row for p-value of S_n in Sc1-2 and 4th row for p-value of T_n in Sc1-2.

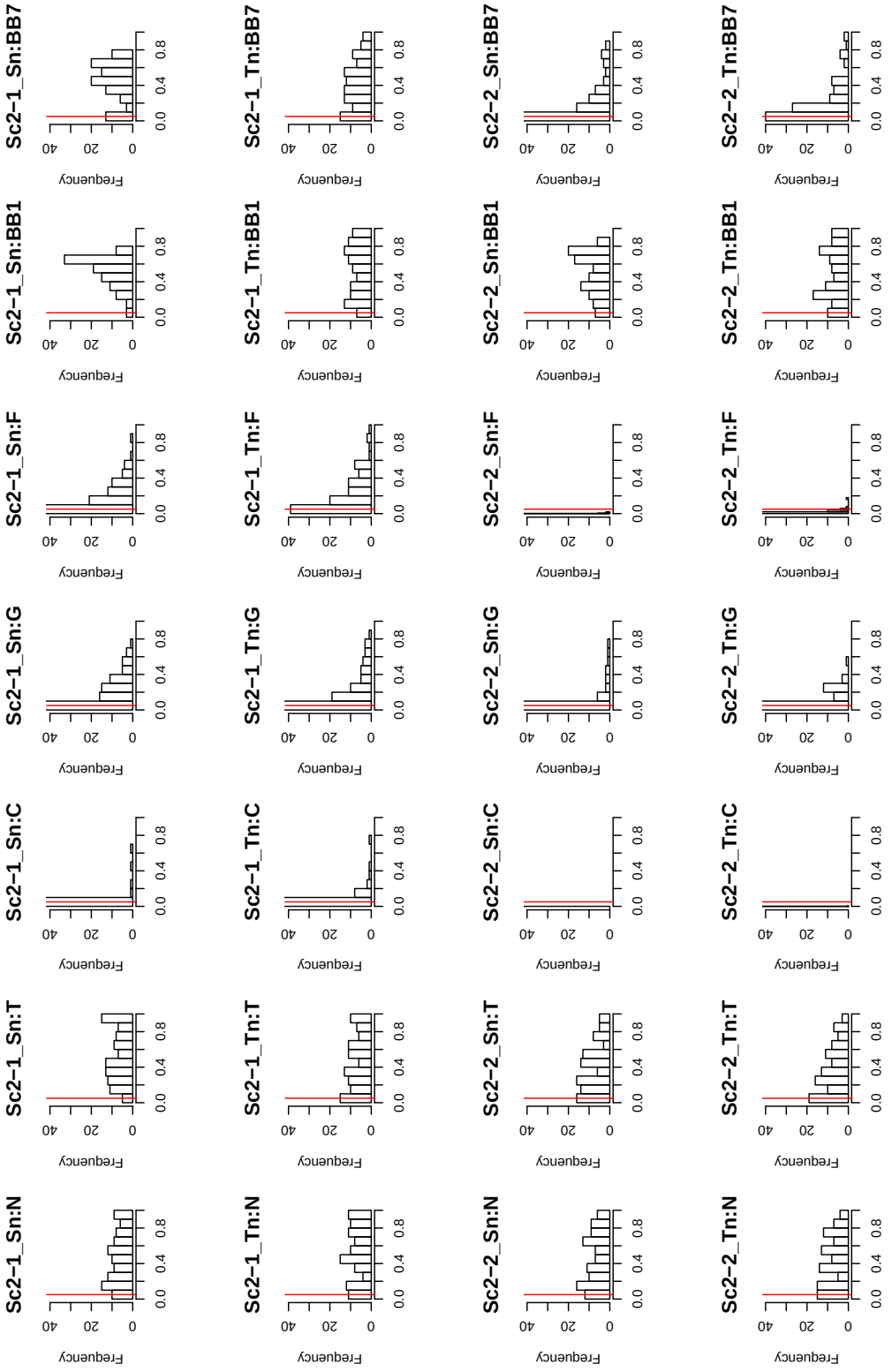


Figure 5.6: Scenario 2: Data set comes from BB1 copula with $\tau = 0.5$ and $\Lambda_U = 2\Lambda_L$. Sc2-1 denotes the sub scenario corresponding to a data set with $n = 200$ and SC2-2 denotes the sub scenario corresponding to a data set with $n = 800$. Each Histogram of p-values is plotted with a vertical line $\alpha = 0.05$. 7 plots in each row stand for 7 test models, respectively. They are N, T, C ,G, F, BB1 and BB7 from left to right. 1st row lists the plots of p-value of S_n for these 7 test models in Sc2-1, 2nd row for p-value of T_n in Sc2-1, 3rd row for p-value of S_n in Sc2-2 and 4th row for p-value of T_n in Sc2-2.

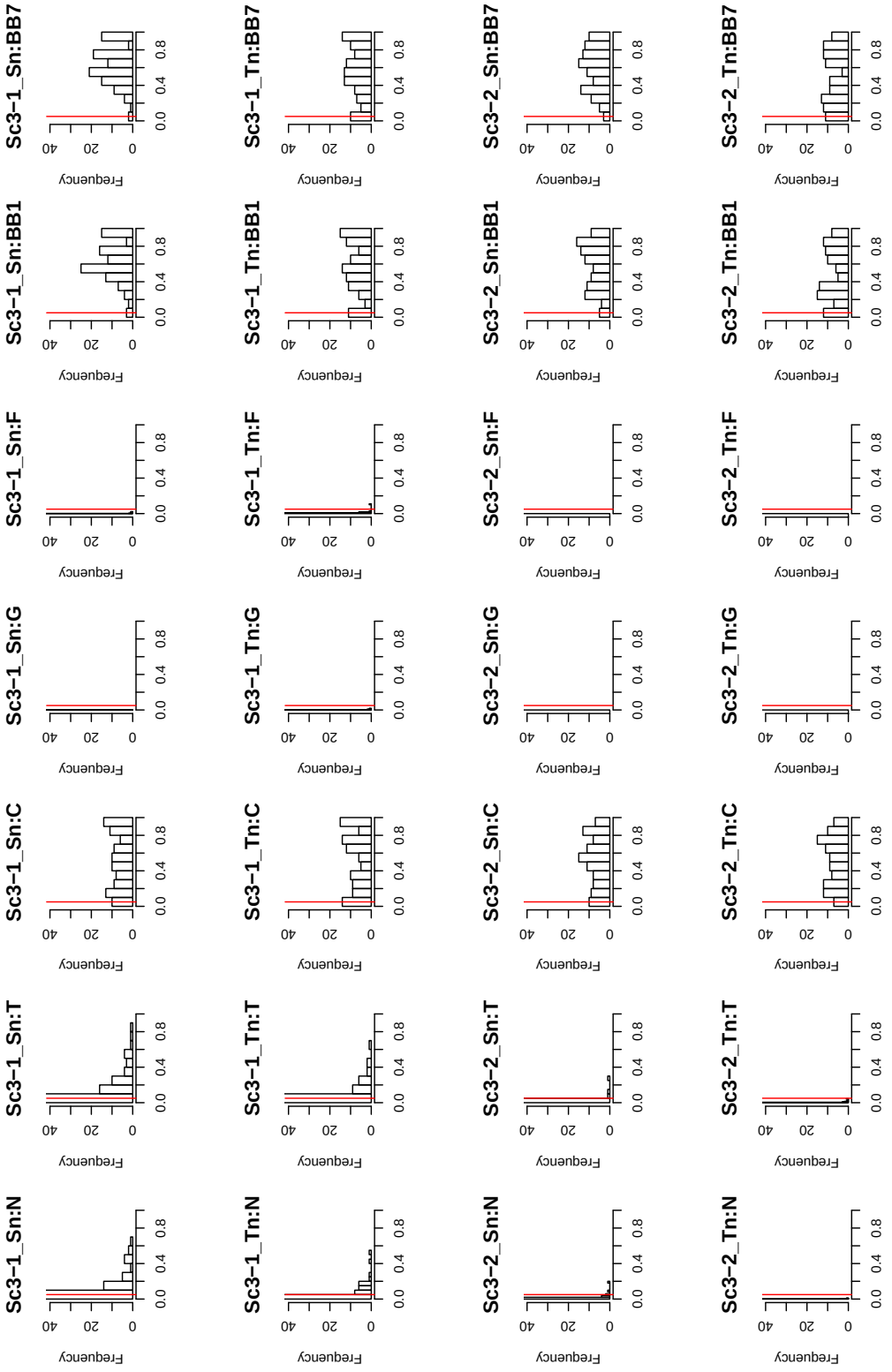


Figure 5.7: Scenario 3: Data set comes from Clayton copula with $\tau = 0.5$. Sc3-1 denotes the sub scenario corresponding to a data set with $n = 200$ and SC3-2 denotes the sub scenario corresponding to a data set with $n = 800$. Each Histogram of p-values is plotted with a vertical line $\alpha = 0.05$. 7 plots in each row stand for 7 test models, respectively. They are N, T, C ,G, F, BB1 and BB7 from left to right. 1st row lists the plots of p-value of S_n for 7 test models in Sc3-1, 2nd row for p-value of T_n in Sc3-1, 3rd row for p-value of S_n in Sc3-2 and 4th row for p-value of T_n in Sc3-2.

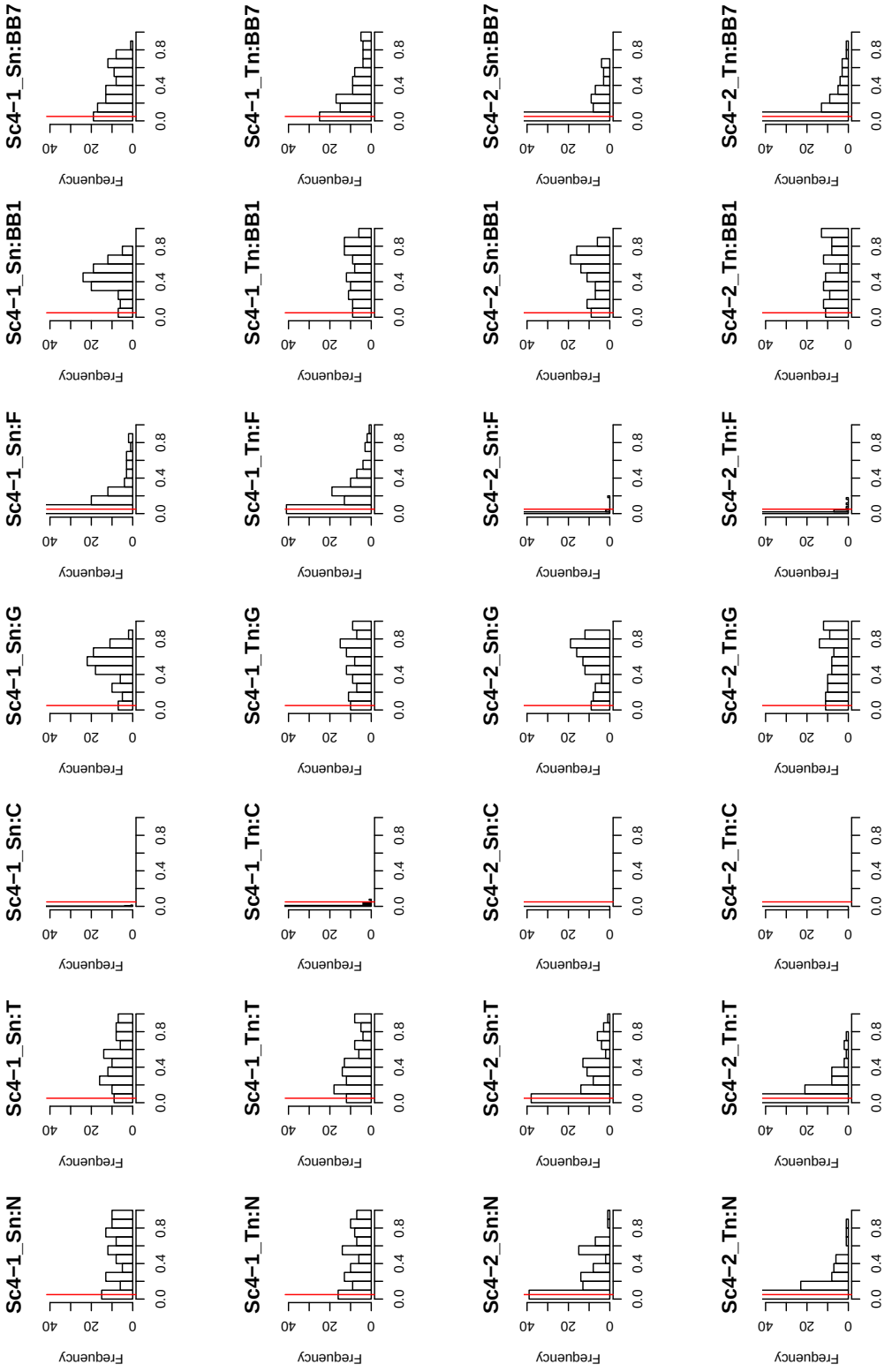


Figure 5.8: Scenario 4: Data set comes from Gumbel copula with $\tau = 0.5$. Sc4-1 denotes the sub scenario corresponding to a data set with $n = 200$ and SC4-2 denotes the sub scenario corresponding to a data set with $n = 800$. Each Histogram of p-values is plotted with a vertical line $\alpha = 0.05$. 7 plots in each row stand for 7 test models, respectively. They are N, T, C ,G, F, BB1 and BB7 from left to right. 1st row lists the plots of p-value of S_n for 7 test models in Sc4-1, 2nd row for p-value of T_n in Sc4-1, 3rd row for p-value of S_n in Sc4-2 and 4th row for p-value of T_n in Sc4-2.

Table 5.5: Number of times that the p-values < 0.05

Scenario	$m(C_\theta^*)$	C_θ^* (B=1000, R=100)						
($\tau = 0.5$)		N	T	C	G	F	BB1	BB7
Sc1-1: (N, n=200)	S_n	**6**	2	90	32	30	12	33
$\Lambda_U = \Lambda_L = 0$	T_n	**7**	8	75	37	22	20	36
Sc1-2: (N, n=800)	S_n	**4**	4	100	98	92	37	93
$\Lambda_U = \Lambda_L = 0$	T_n	**4**	7	100	89	75	41	76
Sc2-1: (BB1, n=200)	S_n	6	2	90	30	33	**3**	8
$\Lambda_U = 2\Lambda_L = 0.508$	T_n	7	8	75	32	36	**3**	9
Sc2-2: (BB1,n=800)	S_n	7	7	100	79	100	**5**	39
$\Lambda_U = 2\Lambda_L = 0.508$	T_n	9	11	100	63	96	**6**	31
Sc3-1: (C, n=200)	S_n	63	45	**4**	100	100	2	1
$\Lambda_L = 0.707$	T_n	74	66	**5**	100	99	5	5
Sc3-2: (C, n=800)	S_n	97	97	**7**	100	100	1	1
$\Lambda_L = 0.707$	T_n	100	100	**6**	100	100	6	5
Sc4-1: (G, n=200)	S_n	6	6	100	**3**	36	1	14
$\Lambda_U = 0.586$	T_n	7	8	99	**5**	25	2	14
Sc4-2: (G, n=800)	S_n	19	27	100	**3**	99	7	45
$\Lambda_U = 0.586$	T_n	35	39	100	**7**	97	10	48

how many times p-values are smaller than α from 100 replications of each test, respectively, i.e. $m(C_\theta^*) \in \{m_{S_n(C_\theta^*)}, m_{T_n(C_\theta^*)}\}$ with

$$m_{S_n(C_\theta^*)} := \#\{j : \ p_k^{S_n(C_\theta^*)} < 0.05, \ j = 1, \ldots, 100\} \tag{5.8}$$

and

$$m_{T_n(C_\theta^*)} := \#\{j : \ p_k^{T_n(C_\theta^*)} < 0.05, \ j = 1, \ldots, 100\}. \tag{5.9}$$

If the selected alternative model fit the data well, we expect a small value of $m_{S_n(C_\theta^*)}$ and $m_{T_n(C_\theta^*)}$. Table 5.5 lists the values of $m_{S_n(C_\theta^*)}$ and $m_{T_n(C_\theta^*)}$ in each sub scenario.

If we denote this true model in each scenario as C_θ^{TRUE}, then $m_{S_n(C_\theta^{TRUE})}$ and $m_{T_n(C_\theta^{TRUE})}$ should be small. In the mean time, we are expecting large values of $m_{S_n(C_\theta^*)}$ and $m_{T_n(C_\theta^*)}$ for $C^* \in \mathcal{C}' \backslash \{C^{TRUE}\}$. In Table 5.5, the values of $m_{S_n(C_\theta^{TRUE})}$ and $m_{T_n(C_\theta^{TRUE})}$ for each scenario have been displayed in bold text. For instance, the true model should be the Normal copula in Sc1. Therefore the values of $m_{S_n(C_\theta^N)}$ and $m_{T_n(C_\theta^N)}$ (first column under Scenario Sc1) in Table 5.5 are bold and they should be small.

As can be seen from this table, the number of times that p-value$<$ 0.05 is relative small when there is no misspecification. However, in several cases the true model is not the only model with few small p-values, especially when data size is small (n=200).

When we look at the results for Sc1 in Table 5.5 where Normal copula is our true model, we see that $m_{Sn(C_\theta^N)}$ and $m_{Tn(C_\theta^N)}$ have the results as we expected. However, $m_{Sn(C_\theta^T)}$ and $m_{Tn(C_\theta^T)}$ for T copula are also very low. Therefore T copula seems to be a appropriate model in this case too. The same problem can be found in all other scenarios. Especially in Scenario Sc3, Clayton copula should be the best model, but BB1 and BB7 copulas

delivered an even better results of $m_{Sn(C_\theta^*)}$ and $m_{T_n(C_\theta^*)}$ than the Clayton copula (see 9th to 12th rows in Table 5.5).

These problems can be caused by the closeness between two copula families under certain parameter choices. For example, the data set is generated from a Clayton copula with $\tau = 0.5$, $\Lambda_U = 0$ and $\Lambda_L = 0.707$ in Sc3. We can fit the same parameter values for BB1 or BB7 copula as well, so that BB1 or BB7 copulas have the same tail dependencies and Kendall's τ as the Clayton copula. Thus, we have a BB1 or BB7 copula close to the Clayton model in Sc3.

Discussion about the closeness between copula families

A common way to discover closeness between copula families is to compare their contour plots. Since we want to investigate the closeness between copulas which we used in the scenarios, we consider the copula families defined in $\mathcal{C}'$. These copulas should be fixed with same or similar parameter choices. Thus, the following copulas have been selected from $\mathcal{C}'$:

N: Normal copula with $\tau = 0.5$. For Normal copulas it always holds that $\Lambda_U = \Lambda_L$.

T3: T copula with $\tau = 0.5$ and the degree of freedom $\nu = 3$, so that $\Lambda_U = \Lambda_L = 0.048$.

T5: T copula with $\tau = 0.5$ and the degree of freedom $\nu = 5$, so that $\Lambda_U = \Lambda_L = 0.021$.

C: Clayton copula with $\tau = 0.5$, so that $\Lambda_U = 0$ and $\Lambda_L = 0.707$.

G: Gumbel copula with $\tau = 0.5$, so that $\Lambda_U = 0.586$ and $\Lambda_L = 0$.

F: Frank copula with $\tau = 0.5$ (F).

BB1(S): BB1 copula with $\tau = 0.5$ and symmetrical tail dependence $\Lambda_U = \Lambda_L = 0.446$.

BB1(A): BB1 copula with $\tau = 0.5$ and asymmetrical tail dependence $\Lambda_U = 2\Lambda_L = 0.508$.

BB1(C): BB1 copula with $\tau = 0.5$ and $\Lambda_U = \Lambda_U^C = 0$,$\Lambda_L = \Lambda_L^C = 0.707$.

BB1(G): BB1 copula with $\tau = 0.5$ and $\Lambda_L = \Lambda_L^G = 0$,$\Lambda_U = \Lambda_U^G = 0.586$.

BB7(S): BB7 copula with $\tau = 0.5$ and symmetrical tail dependence $\Lambda_U = \Lambda_L = 0.544$.

BB7(A): BB7 copula with $\tau = 0.5$ and asymmetrical tail dependence $\Lambda_U = 2\Lambda_L = 0.646$.

BB7(C): BB7 copula with $\tau = 0.5$ and $\Lambda_U = \Lambda_U^C = 0$,$\Lambda_L = \Lambda_L^C = 0.707$.

BB7(G): BB7 copula with $\tau = 0.5$ and $\Lambda_L = \Lambda_L^G = 0$,$\Lambda_U = \Lambda_U^G = 0.586$.

14 copulas are considered with parameters that come from same source, i.e. $\tau = 0.5$. Note that if we set $\tau = 0.5$ and $\Lambda_U = \Lambda_L$ for BB1 and BB7 copulas, then Λ_U and Λ_L cannot be zero. Therefore we will compare the Normal copula to BB1 and BB7 copulas with $\Lambda_U = \Lambda_L \neq 0$. Figure 5.9 displays the contour plots of these 14 copulas, respectively.

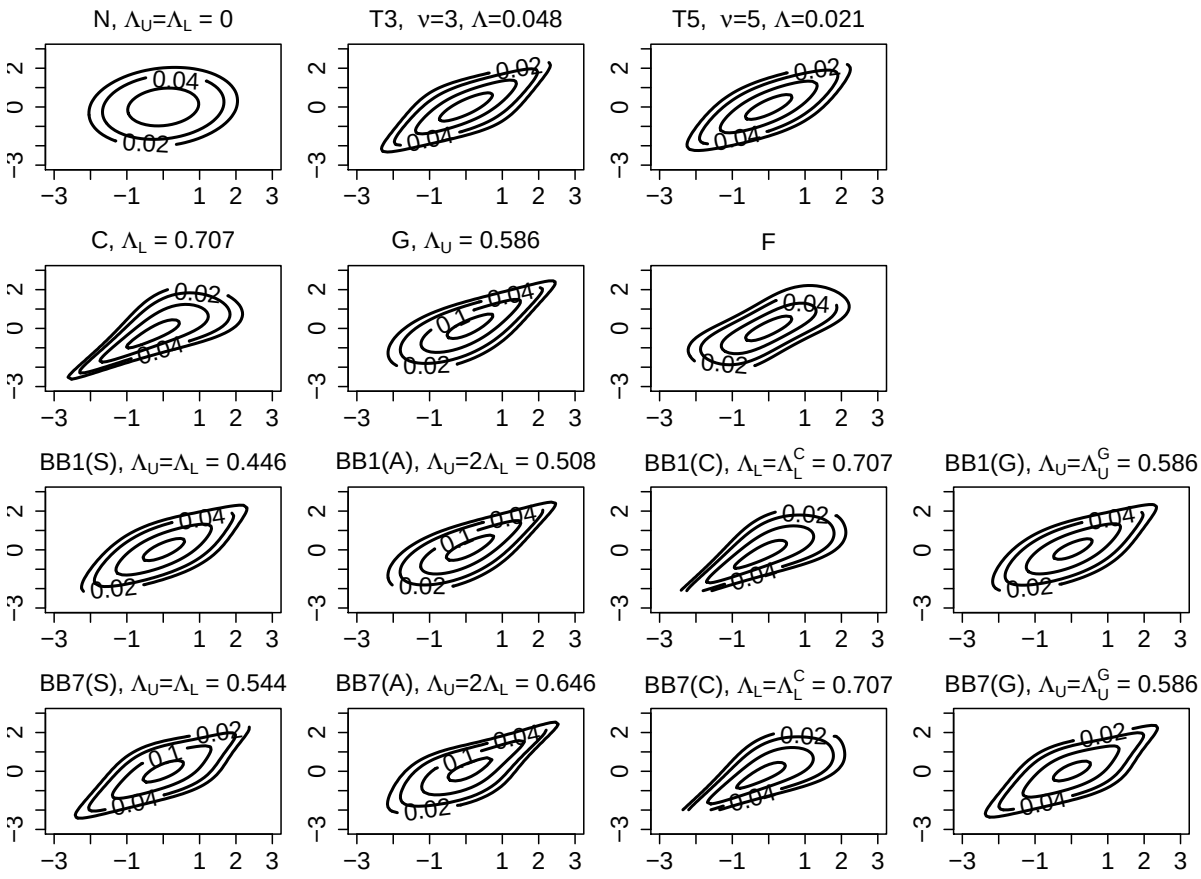


Figure 5.9: Theoretical contour plots between 7 candidates models under same parameter choices, i.e. $\tau = 0.5$. Five contour levels (0.02, 0.04, 0.1, 0.2 and 0.5) were intended to be plotted. Some of them do not exist on the plotting range.

As we can see in Figure 5.9, several plots are close to each other. For instance, the contour plot of Clayton copula (C) and plots of BB1 and BB7 with $\Lambda_L = \Lambda_L^C = 0.707$ (BB1(C) and BB7(C)) are almost identical. The similarity between these certain plots leads to a closeness of copula families. This explains why we have close results in Table 5.5 and in the histograms of estimated p-values.

Alternatively, one can use the *Kullback-Leibler information criterion* (KLIC) (Kullback and Leibler [1951]) to investigate the closeness between copula families. KLIC measures the distance between the true but unknown distribution and a specified, approximating model with estimates $\hat{\theta}$ of the true values θ (note that these are not the parameters of the true distribution). Vuong [1989] defines the KLIC as

$$KLIC := \int h_0(x) \log \left[\frac{h_0(x)}{f(x|\hat{\theta})} \right] dx = E_0[\log h_0(x)] - E_0\big[\log f(x|\hat{\theta})\big], \tag{5.10}$$

where X is a random variable following the true distribution with density $h_0(\cdot)$ and E_0 denotes the expectation with respect to this true distribution which is approximated by $f(\cdot|\hat{\theta})$.

Since our investigation is based on bivariate copulas and the copula densities are given, we can use the KLIC (5.10) to measure the closeness between two bivariate copula families. The KLIC between the true copula C_0 and an alternative copula C_1 can be rewritten as

$$KLIC(C_0, C_1) := \int_{[0,1]^2} c_0(u,v) \log\Big[\frac{c_0(u,v)}{c_1(u,v)}\Big] \mathrm{d}u\mathrm{d}v,$$

where c_0 and c_1 denote the copula densities corresponding to copulas C_0 and C_1, respectively. Obviously, the KLIC function is asymmetric.

In order to have a symmetric version of the KLIC, we take the mean of two KLIC functions, i.e

$$KLIC_{sym}(C_0, C_1) := \frac{1}{2}\big(KLIC(C_0, C_1) + KLIC(C_1, C_0)\big).$$

If C_0 and C_1 are close to each other, one can conclude directly from the definition of KLIC that $KLIC(C_0,C_1)$ and $KLIC_{sym}(C_0, C_1)$ are close to zero.

With help of the R-function *KLICfunction()* written by Mr. Brechmann, we computed the values of symmetrized KLIC function for all 14 selected copulas in Table 5.6.

Figure 5.10 reports the relationships and distances between every two KLIC values listed in Table 5.6 graphically. Depending on 14 different given true models 14 plots (horizontal lines) are drawn in each panel of Figure 5.10.

With the results from Table 5.6, we can explain the problems in Table 5.5 efficiently.

- For Sc1: We look at the first 4 rows in Table 5.5 at first. The results for Normal and T copulas are very close. Therefore we expect that Normal and T copula families are close to each other under certain parameter choices. Reading from Table 5.6 we see that $KLIC_{sym}(C^N, C^{T3}) = 0.051$ and $KLIC_{sym}(C^N, C^{T5}) = 0.021$. This means that Normal and T copulas are getting close to each other when the degree of freedom is increasing. Since the data set in Sc1 is generated from a Normal copula with $\tau = 0.5$ and the T copula has two parameters, we estimate the degree of freedom of T copula by using maximum likelihood estimation with the data set. It yields $\hat{\nu} = 20$ and $KLIC(C^N, C^{T20}) = 0$, where C^{T20} denoted the T copula with $\tau = 0.5$ and $\nu = 20$. Therefore Normal and T copulas are identical with $\tau = 0.5$ and $\nu = 20$.

- For Sc2: Table 5.6 reports that

  $$KLIC_{sym}(C^{BB1(Asym)}, C^N) = 0.016$$

  and

  $$KLIC_{sym}(C^{BB1(Asym)}, C^{T5}) = 0.006.$$

  This explains why the values in the 1st and 2nd columns for Sc2 in Table 5.5 are also very low.

- For Sc3: Identifiability exist between Clayton and BB1 and BB7 copulas. Table 5.6 shows that $KLIC_{sym}(C^C, C^{BB1(C)}) = 0$ and $KLIC_{sym}(C^C, C^{BB7(C)}) = 0$. This indicates that they are almost identical when BB1 and BB7 copulas are fixed with $\tau = 0.5$ and $\Lambda_L = \Lambda_L^C = 0.707$.

Table 5.6: Values of symmetrized KLIC between 14 copula families

Sc	**Sc1**			**Sc3**	**Sc4**		**Sc2**							
Family	N	T3	T5	C	G	F	BB1(A)	BB1(S)	BB1(C)	BB1(G)	BB7(A)	BB7(N)	BB7(C)	BB7(G)
N	0	0.051	0.021	0.130	0.038	0.034	0.016	0.015	0.13	0.038	0.048	0.032	0.109	0.032
T3	0.051	0	0.006	0.121	0.04	0.075	0.016	0.014	0.121	0.04	0.034	0.017	0.012	0.017
T5	0.021	0.006	0	0.113	0.029	0.048	0.006	0.004	0.112	0.029	0.029	0.121	0.112	0.013
C	0.130	0.121	0.113	0	0.248	0.160	0.148	0.097	0	0.248	0.234	0.106	0	0.146
G	0.038	0.040	0.029	0.248	0	0.068	0.018	0.032	0.247	0	0.008	0.045	0.247	0.027
F	0.034	0.075	0.048	0.160	0.068	0	0.056	0.060	0.160	0.068	0.105	0.096	0.159	0.095
BB1(A)	0.016	0.016	0.006	0.148	0.018	0.056	0	0.005	0.148	0.014	0.012	0.011	0.147	0.006
BB1(S)	0.015	0.014	0.004	0.097	0.032	0.060	0.005	0	0.097	0.036	0.030	0.004	0.097	0.008
BB1(C)	0.130	0.121	0.113	0	0.247	0.160	0.148	0.097	0	0.247	0.234	0.106	0	0.146
BB1(G)	0.038	0.040	0.029	0.248	0	0.068	0.014	0.036	0.247	0	0.008	0.045	0.246	0.026
BB7(A)	0.048	0.034	0.029	0.234	0.008	0.105	0.012	0.030	0.234	0.008	0	0.028	0.233	0.012
BB7(N)	0.032	0.017	0.012	0.106	0.045	0.096	0.011	0.004	0.106	0.045	0.028	0	0.106	0.034
BB7(C)	0.109	0.120	0.112	0	0.247	0.159	0.147	0.097	0	0.246	0.233	0.106	0	0.145
BB7(G)	0.032	0.017	0.013	0.146	0.027	0.095	0.006	0.008	0.146	0.026	0.012	0.035	0.145	0

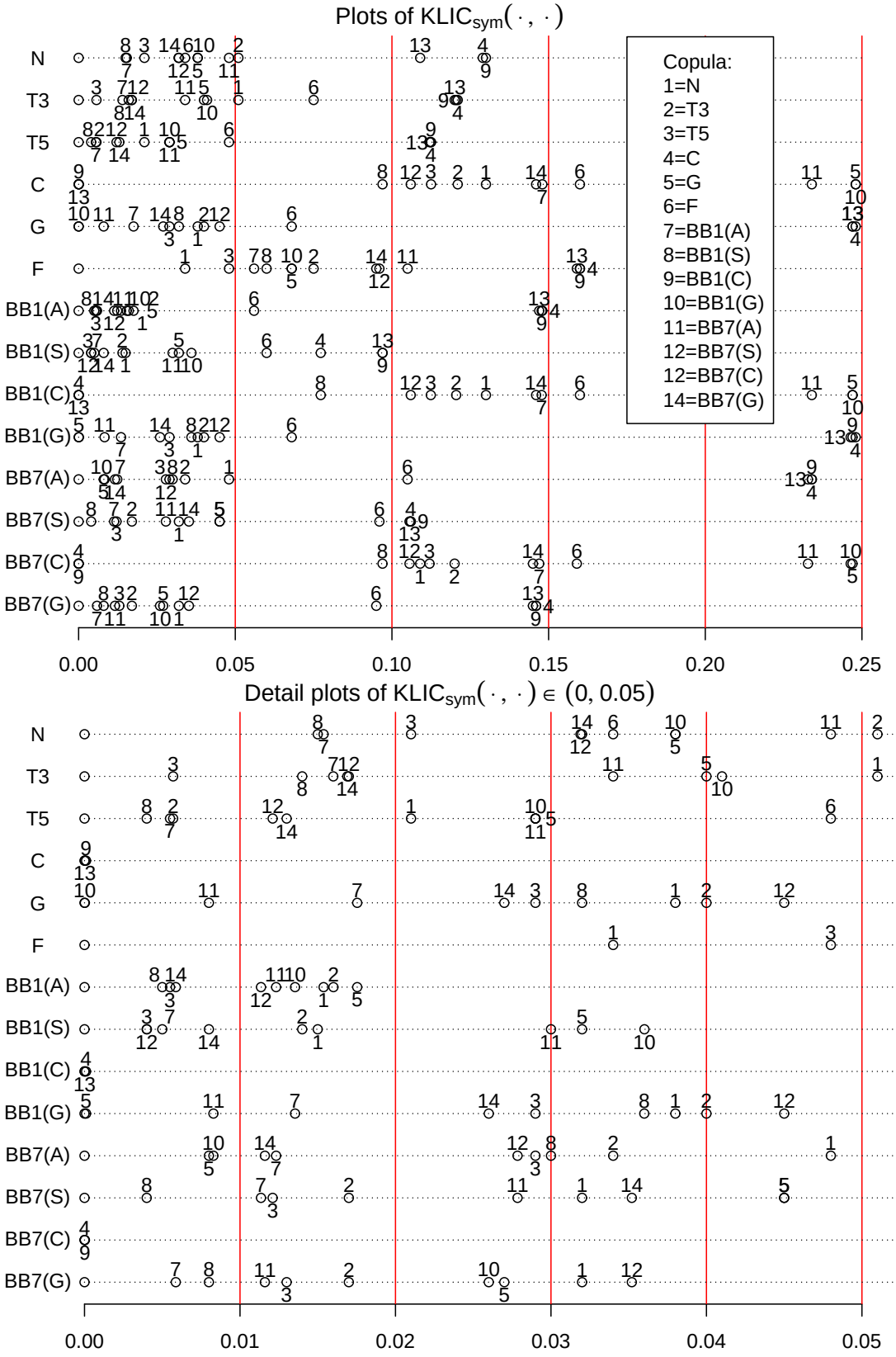


Figure 5.10: Graphical comparison of the $KLIC_{sym}$-values in Table 5.6 : Plots of all $KLIC_{sym}$-values in Table 5.6 corresponding to 14 different models (top) and the detail plots of the top graph for $KLIC_{sym}(\cdot,\cdot) \leq 0.05$ (bottom).

Table 5.7: Average $KLIC_{sym}$ values based on estimated parameters between true model and model used for estimation for each scenario..

	$E[KLIC_{sym}(C^{TRUE}, C^*)]$							
	C^{TRUE}	C^{Approx}						
Sc	Model	N	T	C	G	F	BB1	BB7
Sc1-1	N (n=200)	**0.000**	0.002	0.128	0.038	0.034	0.018	0.040
Sc1-2	N (n=800)	**0.000**	0.002	0.126	0.037	0.033	0.016	0.036
Sc2-1	BB1 (n=200)	0.013	0.011	0.155	0.019	0.058	**0.000**	0.013
Sc2-2	BB1 (n=800)	0.011	0.012	0.150	0.018	0.056	**0.000**	0.013
Sc3-1	C (n=200)	0.133	0.117	**0.000**	0.252	0.164	0.002	0.002
Sc3-2	C (n=800)	0.129	0.112	**0.000**	0.246	0.159	0.000	0.001
Sc4-1	G (n=200)	0.038	0.029	0.246	**0.000**	0.068	0.003	0.015
Sc4-2	G (n=800)	0.038	0.028	0.247	**0.000**	0.068	0.001	0.010

· For Sc4: we have the conclusion that a Gumbel copula and a BB1 copula with $\Lambda_L = \Lambda_L^G = 0.586$ are very close, since $KLIC_{sym}(C^{BB1(G)}, C^G) = 0$.

Since the KLIC function is implemented on the copula parameter level, we want to compute the $KLIC_{sym}$ between the true and alternative models in each scenarios using the parameter estimated from 100 replications. Respectively, for each model pairs we obtain 100 estimated $KLIC_{sym}$ values from 100 replications. Table 5.7 lists the average values of the 100 estimated $KLIC_{sym}$ values between every true copula and copula used for estimation for every sub scenario, i.e.

$$E[KLIC_{sym}(C^{TRUE}, C^*)] := \overline{KLIC_{sym}(C^{TRUE}, C^{Approx})} = \frac{1}{100}\sum_{k=1}^{100} KLIC_{sym}(C^{TRUE}_{\hat{\theta}_k^{TRUE}}, C^{Approx}_{\hat{\theta}_k^{Approx}}).$$

Figure 5.11 displays the boxplots of the estimated $KLIC_{sym}$ values between true and alternative models for each scenario. From Table 5.7 and Figure 5.11 we can see that larger sample size leads to lower $KLIC_{sym}$.

Now we have computed the number of times that $p < 0.05$ and the symmetrical KLICs for each case, i.e $m(C^*)$ and $E[KLIC_{sym}(C^{TRUE}, C^*)]$ with $C^* \in \mathcal{C}'$ (Eq.(5.7)) and $m(C^*) \in \{m_{S_n(C^*)}, m_{T_n(C^*)}\}$ (Eq. (5.8) and (5.9)). We want to compare them by plotting on the same graph. We did this in Figure 5.12. A higher $KLIC_{sym}$ leads to a lower p-value estimated for a copula model C^* and this small p-value leads to a departure of C^* from the true model C^{TRUE} or equivalently a higher value of $KLIC_{sym}(C^{TRUE}, C^*)$, therefore a monotone relationship should be observed between $m(C^*)$ and $E[KLIC_{sym}(C^{TRUE}, C^*)]$ in Figure 5.12.

As we expected, Figure 5.12 showed the monotone increasing relationships between $m(C^*)$ and $E[KLIC_{sym}(C^{TRUE}, C^*)]$ for all 7 candidate models in the most (sub) scenarios. For scenario Sc4-2 with the Gumbel copula as true model, we did not observe the monotone increasing relationships between BB7, Normal and T copulas in the bottom panel of Figure 5.12. In this case $KLIC_{sym}(C^G, C^*)$ is larger than $KLIC_{sym}(C^G, C^{BB7})$ but $m(C^*)$ is smaller than $m(C^{BB7})$ for $C^* \in (C^N, C^T)$. Since our true model is Gumbel copula, for Normal and T copulas we expect large values of $m(C^*)$. This contravention

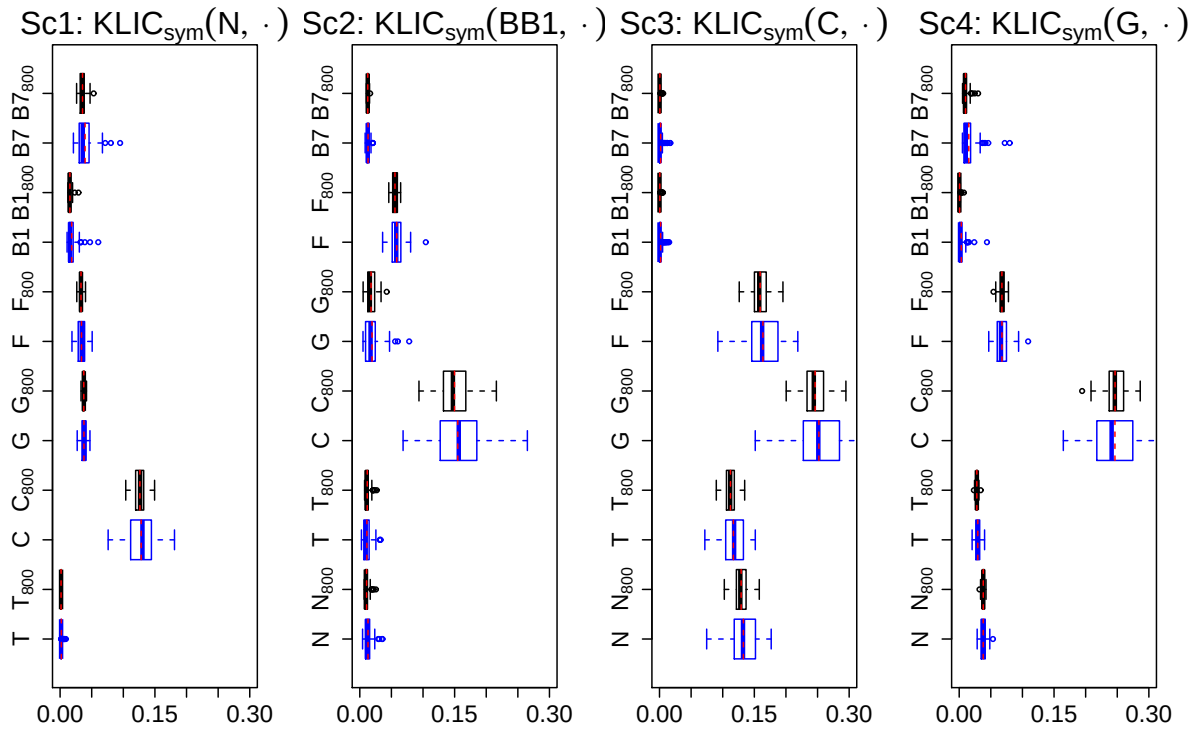


Figure 5.11: Boxplots of the estimated $KLIC_{sym}$ values between true and alternative models for each scenarios. Labels in y-axis without subscript corresponds to data size $n = 200$ and with subscript corresponds to $n = 800$. B1 and B7 stand for BB1 and BB7.

can be caused by the non-parametric simulation of the theoretical Kendall functions K_θ of Normal and T copulas on the second bootstrap level of the GOT tests. One this level the Kendall functions of Normal and T copula families are simulated using the empirical procedure. Hence, the estimates $K_{\hat{\theta}}$ as well as the test statistics and the corresponding p-values of Gumbel, Normal and T copulas can be close to each other. These conclusions lead to the small and close results of their KLIC functions, which explain the non-monotone problem in the bottom panel of Figure 5.12.

To prove this conclusion we conduct the tests of Normal and T copula families for scenarios Sc4 again and use the bootstrapping samples with size $10,000$ to simulate their theoretical Kendall function at this time, i.e.

$$K_\theta^N(t) := \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\{C_\theta(u_i, v_i) \leq t\}$$

with $N = 10,000$ and C_θ being Normal or T copula. Note that $N = 5000$ was used in the previous simulation studies. We list the number of times that $p < 0.05$ for this study in the last two columns of Table 5.8. The histogram of the resulted p-values are displayed in Figure 5.13. Results in Table 5.8 shows that the number of times that $p < 0.05$ generally get higher with $N = 10,000$. However, Large N only has a significant effect for large data size.

For the same reason we plotted Figure 5.12, we compared the number of times that $p < 0.05$ whit their KLIC values in Figure 5.14 for these new simulations. As introduced

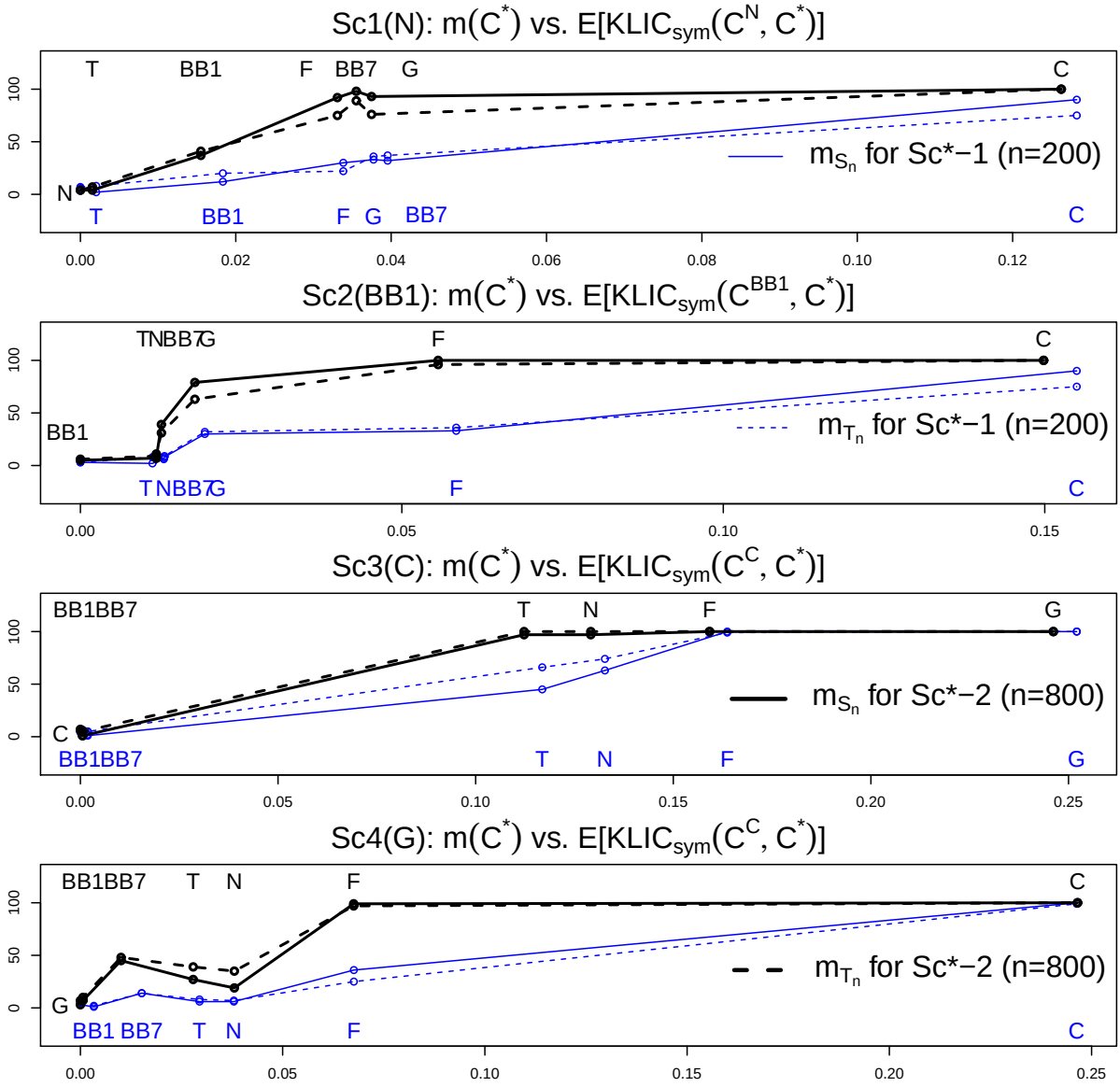


Figure 5.12: $m(C^*)$ vs. $E[KLIC_{sym}(C^{TRUE}, C^*)]$ for $C^* \in \mathcal{C}'$ (Eq.(5.7)) and $m(C^*) \in \{m_{S_n(C^*)}, m_{T_n(C^*)}\}$ (Eq. (5.8) and (5.9))

Table 5.8: Comparison between number of times that $p < 0.05$ for the tests of Normal and T copulas for Sc4 using $N = 5000$ and $N = 10,000$. Columns 1-5 list the results for $N = 5000$ which are captured from Table 5.5.

Sc4 ($\tau = 0.5$)	Using statistic						N=5000		N=10,000	
		C	G	F	BB1	BB7	N	T	N	T
n=200	S_n	100	3	36	1	14	6	6	**7**	**7**
	T_n	99	5	25	2	14	7	8	**8**	**9**
n=800	S_n	100	3	99	7	45	19	27	**39**	**43**
	T_n	100	7	97	10	48	35	39	**54**	**57**

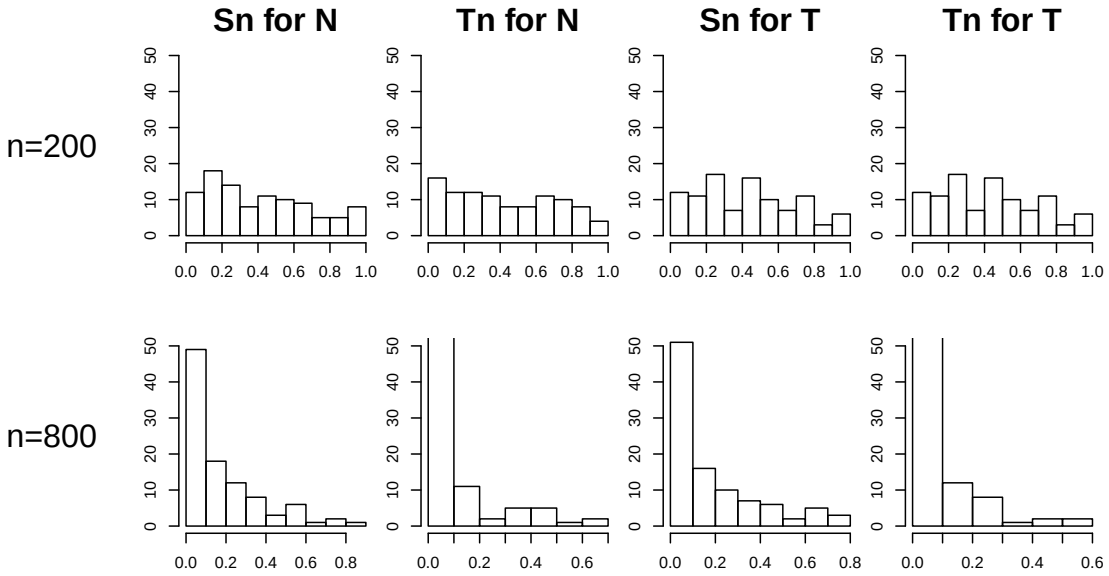


Figure 5.13: Histogram of the number of times that $p < 0.05$ for the tests of Normal and T copulas for Sc 4 using $N = 10,000$

before that the KLIC values only depended on the estimated copula parameters and we only used the empirical procedure for the theoretical Kendall functions of Normal and T copula families, therefore the number of times that $p < 0.05$ for other copula models and the KLIC values for all copula models including Normal and T copulas have not been changed. Figure 5.12 shows clearly an improvement comparing to the bottom panel of Figure 5.12, which prove that an adequate large N will lead us to observe a monotone relationship.

In this subsection we illustrated and explained the identifiability problem between copula families. Considering their conclusions, the results we get from the simulation studies are surprising but can be explained. The tests based on Kendall's transform seem to be acceptable. Genest et al. [2006], Genest et al. [2009] and Berg [2009] established similar simulation studies and proposed the same conclusion. Comparing their numerical studies to ours, there are some advantages and disadvantages:

(i) A large number of replications (10,000 random samples) are conducted in Genest et al. [2006], Genest et al. [2009] and Berg [2009]. Thus, the results from their studies are more precise. However, these simulations required more time and computer power.

(ii) In our studies we tested BB1 and BB7 copulas. Both models are two-parameter copulas and have flexible upper and lower tail dependencies. Thereby, they are able to cover a large band of possibilities of data modeling. Unfortunately, these two families are not considered in other simulation studies. Their investigation are mostly limited to T and other one-parameter copulas.

(iii) The identifiability problem between copula families is illustrated and discussed. Although other studies suffered from the same problem, similar investigation in this

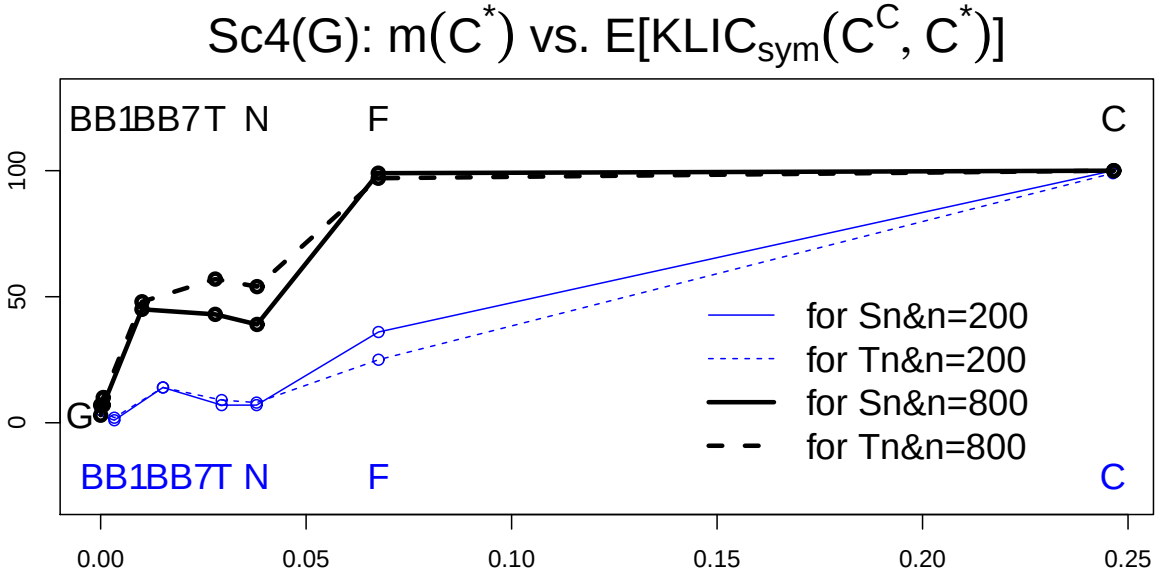


Figure 5.14: $m(C)$ vs. $E[KLIC_{sym}(C^G, C^*)]$ for Sc4 with Gumbel copula as rue model and using $N = 10,000$ for Normal and T copulas. Note that $C^* \in \mathcal{C}'$ (Eq.(5.7)) and $m(C^*) \in \{m_{S_n(C^*)}, m_{T_n(C^*)}\}$ (Eq. (5.8) (5.9)).

direction were not reported.

Effect of bootstrap size

In this simulation study, we want to briefly discuss the impact of the number of bootstraps (B) on p-values. Remember that we used 1000 Bootstraps (B=1000) in the previous simulations.

Obviously, the larger B is, the better results we will have. When is B large enough in practice? To investigate this discussion, we want to compare estimated P-values with different values of B.

Table 5.9 reports the simulated P-values for testing a Clayton copula model. Data set is generated from Clayton copula with size $n = 1000$ and $\tau = 0.5$. P-values are estimated via parametric bootstrapping in Algorithm 5.3 for two runs of B=100, two runs of B=1000, two runs of B=10,000 and one run of B=100,000. For example, the first two columns in Table 5.9 show the p-values derived from two runs of B=100. In these tests, we can observe the discrepancy between the p-values. This illustrates the importance of taking larger B to insure reliable conclusions. As can be seen in the last three columns in Table 5.9, taking B=10,000 instead of B=100,000 does not changed the estimated p-values. Thereby, $B = 10,000$ seems perfectly adequate in practice, although one could certainly get by less if limited in time or computing power. Given that the sample comes from a Clayton copula, it is unsurprising that all tests here lead to a non-rejection of Clayton model.

Genest and Favre [2007] verified the same conclusions by testing a different sample with a smaller size. One can refer to Genest and Favre [2007] as an alternative.

Altogether, with the illustrations and discussions of both kind simulation studies we

Table 5.9: P-values estimated by Parametric Bootstrap using different size of Bootstrap (B) for testing the Goodness-of-Fit based on Kendall's transform of the Clayton copula model. Data set is generated from Clayton copula with size $n = 1000$ and $\tau = 0.5$.

	P-values based on a run of						
Using Statistic	B= 100	B=100	B=1000	B=1000	B=10,000	B=10,000	B=100,000
S_n	0.17	0.25	0.167	0.152	0.148	0.149	0.149
T_n	0.26	0.19	0.229	0.218	0.231	0.230	0.231

can observe increasing power of the tests with data size and number of bootstraps, and conclude that the results from the Scenario Studies are reasonable and reliable. In practice, the given data set usually has a large size. Thus, tests based on such data set will be even more accurate.

Chapter 6

Applications to the Norwegian return data

In previous chapters we introduced graphical methods and Goodness-of-Fit tests, and tested them with simple examples or scenarios. Now we want to apply these methods altogether on a same data set, so that we can have a complete exploratory data analysis (EDA) . Since we focus on the pair-copula decomposition of D-vines in this thesis, in this chapter we will carry out the D-vine decomposition step by step.

6.1 Learning data set

The data set is converted from Norwegian data with an appropriate time series model. Norwegian data consists of four time series of daily data:

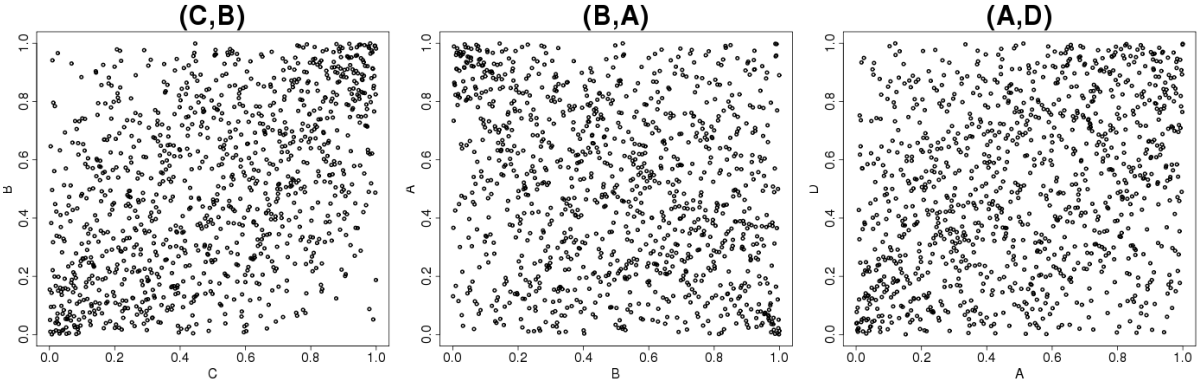
- The Norwegian stock index
- The MSCI world stock index
- The Norwegian bond index
- The SSBWG hedged bond index

Each index contains $N = 1094$ daily return values for the period from 04.10.2000 to 08.07.2003. Since this data represents time series data, it is unreasonable to assume an i.i.d marginal structure. Aas et al. [2009] showed that an AR(1)-Garch(1,1) fits the marginal time series structure well, i.e. they fitted the following model for log-return x_i of each index for every $i = 1, \ldots, N$,

$$\begin{aligned} x_{i,t} &= c_i + \alpha_i x_{i,t-1} + \sigma_{i,t} z_{i,t}, \\ \sigma^2_{i,t} &= a_{i,0} + a_i \epsilon^2_{i,t-1} + b_i \sigma^2_{i,t-1}, \end{aligned}$$

where $\epsilon_{i,t-1} = \sigma_{i,t} z_{i,t}$ and $z_{i,t}$ is i.i.d. with $E[z_{i,t}]$=0 and $Var[z_{i,t}]$=1. Since we are interested in the dependency structure among the 4 return series, the analysis is performed on the

Figure 6.1: Pairs-plots of A, B, C and D



standardized residual $z_{i,t}$. For each index, we adapt the standardized residual $z_{i,t}$ to $(0,1)$-uniform variable $u_{i,t}$ using the empirical distribution functions. For all $i = 1, \ldots, N$,

$$u_{i,t} := \frac{1}{N} \sum_{j=1}^{N} \mathbf{1}\{z_{j,t} \leq z_{i,t}\}.$$

For further notations, we denote the variable with observations $u_{i,t}$ converted from the four indices as follows:

A for the variable from **Norwegian stock index**.

B for the variable from **MSCI world stock index**.

C for the variable from **Norwegian bond index**.

D for the variable from **SSBWG hedged bond index**.

Thus, we have 4 new variables and each of them has 1094 observations and approximated $U(0,1)$ distribution. They build our data set for the further modeling and the data set has size 4×1094.

The same data set is used in Aas et al. [2009]. More discussion about the conversion of the Norwegian data can be found in Aas et al. [2009].

6.2 Assessment of dependence

Before a copula model for the pair (U, V) is sought, visual tools are used to check for the presence of dependence. The scatter plots of all possible pairs between A, B, C and D are shown in Figure 6.1. Scatter plots of pairs (C, B) and (A, D) point to the existence of positive relationships between two variables, respectively. (B, A) seems to have a negative relationship.

Table 6.1: Estimated τ_n and P-value corresponding to testing for independence for pairs of variables from the transformed Norwegian return data

τ_n	B	A	D	P-value	B	A	D
C	0.313	-0.128	0.110		0	0	0
B		-0.158	-0.041			0	0.044
A			0.196				0

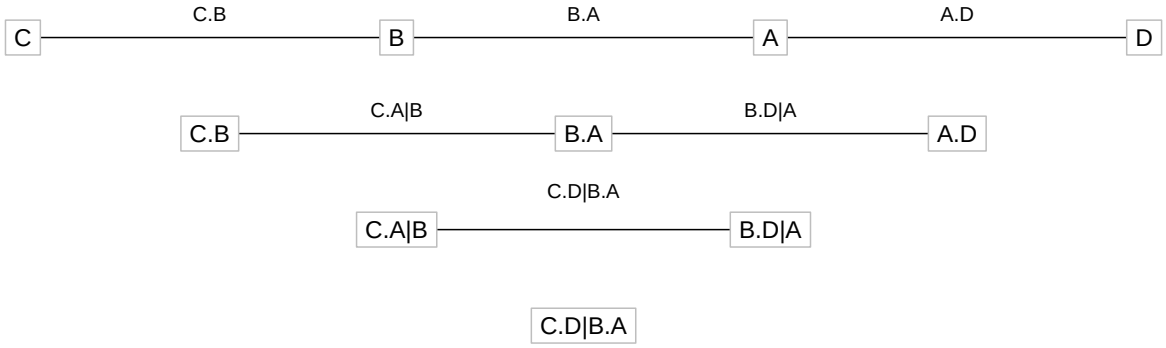


Figure 6.2: Selected D-vine structure of Norwegian data

Testing independence based on Kendall's τ

To quantify the degree of dependence in each pair, $\hat{\tau}_n$ is estimated by using Eq. (2.1) and the P-values of the independence test based on $\hat{\tau}_n$ are determined by using Eq. (2.2). These values are listed in Table 6.1. The results indicate the existence of dependence for all 6 pairs at significance level $\alpha = 0.05$.

Table 6.1 shows B seems to be the central factor of our data set, since (B, C) and (B, A) have the strongest dependencies. These relationships can be explained economically. Note that A and D are converted from the Norwegian stock and bond indices, and B is converted from the MSCI world stock index data. Since Stock and bond, as the two most important financial assets, are heavily related to each other and the economy of Norway is connected to the economy of the world, it is understandable that the Norwegian stock and bond indices depend highly on the world stock index.

6.3 Selecting an appropriate D-vine decomposition

To select a D-vine structure for our data, we need to find the first tree as the first step. Since D-vine structure are uniquely determined by the first tree, we can determine the D-vine structure for our data right after the first step.

Consider the transformed Norwegian return data, 6.1 reports that the strongest dependence between two variables exists in pair (C, B), since a larger value of $|\hat{\tau}_n|$ indicates stronger dependence. If we order the 6 pairs depending on $|\hat{\tau}_n|$, it would be

$$|\hat{\tau}_n(C,B)| \geq |\hat{\tau}_n(B,A)| \geq |\hat{\tau}_n(C,A)| \geq |\hat{\tau}_n(A,D)| \geq |\hat{\tau}_n(C,D)| \geq |\hat{\tau}_n(D,B)|,$$

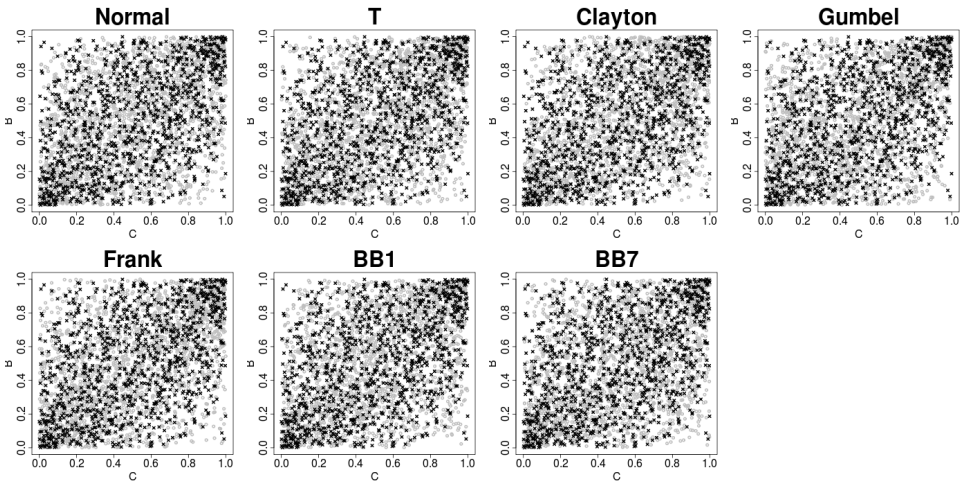


Figure 6.3: Scatter plot for the pair (C, B) (black points) vs. data (gray points) simulated form Normal (top-left), T (Top-middle), Clayton (top-right), Gumbel (bottom-left) and Frank (bottom-right) with same $\tau = 0.313$.

then the first tree in a D-vine structure is uniquely determined as C-B-A-D. After Tree 1 is determined, the entire D-vine structure is clear. We draw this structure in Figure 6.2. As shown in Figure 6.2, the D-vine structure has 3 trees and 6 edges. We denote the 3 trees from top to bottom as T_1, T_2 and T_3 . 6 edges mean that we have 6 bivariate copula models to find. Since T_1 has 3 edges, we start the model selection with finding 3 copula models in T_1.

6.4 Model selection

For modeling the dependence of each pair displayed in Figure 6.2, the same copula families as before are considered. They are Normal, T, Clayton, Gumbel, Frank, BB1 and BB7 copula families.

Copula selection on the first level

Before we compute the h-functions (Eq. (2.11)) used in trees T_2 and T_3, we need to select an adequate copula family for each pair in T_1, i.e. (C, B), (B, A) and (A, D). Since there are have 7 candidates to test for each pair, in order to reduce the complexity of the GOF tests we first try to filter out a few models by using the graphical methods described in Chapter 4 at first. Therefore we start the model selection with these graphical methods.

Scatter plot

As can be seen in Chapter 4, we can use the scatter plot not only to get a first impression of data but also to compare the data with a random sample drawn from selected theoretical copula families. Figure 6.3 shows the scatter plots of the pair (C, B) (black points) and the random sample from 7 different copulas (gray points) respectively. Unfortunately, these

plots do not give us enough information to eliminate any copula family, since we cannot decide which copula sample covers the data better . One reason for this could be that the dependence between variables C and B is not strong enough. Naturally, scatter plots of the other two pairs (B, A) and (A, D) will be even less useful, since results in Table 6.1 shows that these 2 pairs have weaker dependencies than (C, B).

Contour plots on the normal scale

Table 6.2 lists the contour plots on the normal scale for all 3 pairs (C, B), (A, D) and $(B, \mathbf{1} - A)$, where $\mathbf{1}$ is a vector with length k and entries 1 and $(B, \mathbf{1} - A)$ stands for the 90 degree clockwise rotation of the pair (B, A). The reason for testing $(B, \mathbf{1} - A)$ instead of (B, A) is that with $\tau_n(B, A) < 0$ Clayton, Gumbel, BB1 and BB7 copulas cannot be tested, since they are not defined for negative Kendall's τ. By using this data rotation we have that

$$\tau_n(B, \mathbf{1} - A) = -\tau_n(B, A) > 0.$$

Hence, they can be tested for $(B, \mathbf{1} - A)$. In the "Rotated copulas" section we showed the equivalent relationship

$$(U, V) \text{ has copula C} \Leftrightarrow (U, \mathbf{1} - V) \text{ has rotated copula } C^{+-}$$

and $c^{+-}(u, v) = c(u, 1 - v)$. This means if any rotated copula C^{+-} fits the data pair $(B, \mathbf{1} - A)$, then copula C fits the pair (B, A). Thus, testing $(B, \mathbf{1} - A)$ is equivalent to testing (B, A). Alternatively, one can also draw the contour plot of rotated copula density functions of (B, A). This procedure is more complex due to the copula function rotation, since a rotated Archimedean copula does not belong to the Archimedean Copula Class.

The first row in Table 6.2 displays the empirical contour plots. The 2nd to 8th rows list the contour plots for 7 different theoretical copulas. Parameters of these 7 copulas are estimated from 3 pairs, respectively.

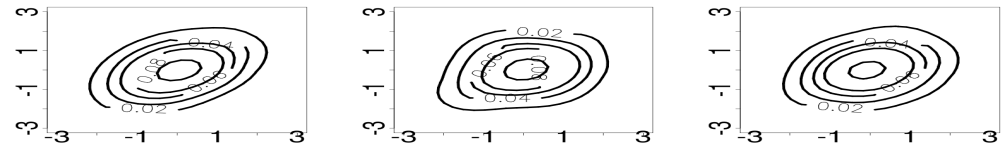
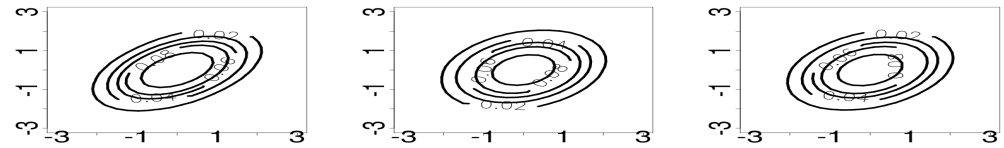
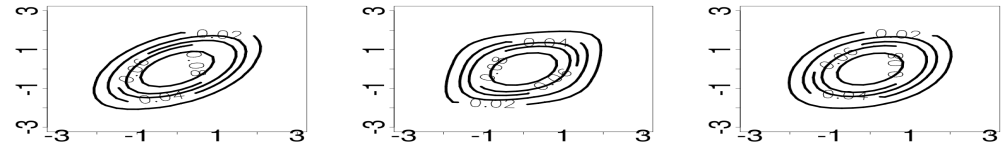
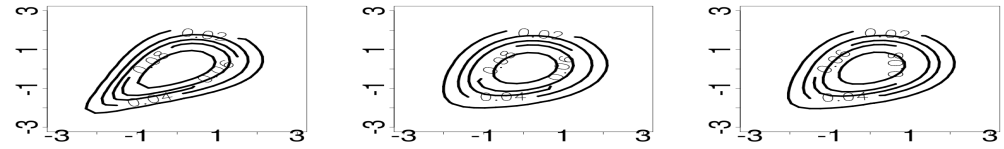
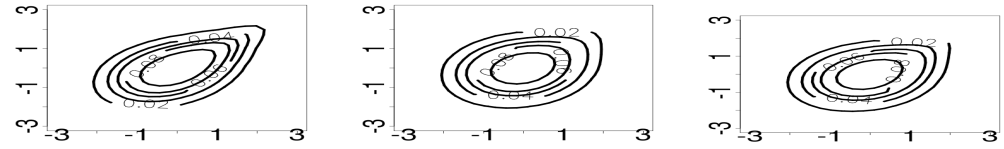
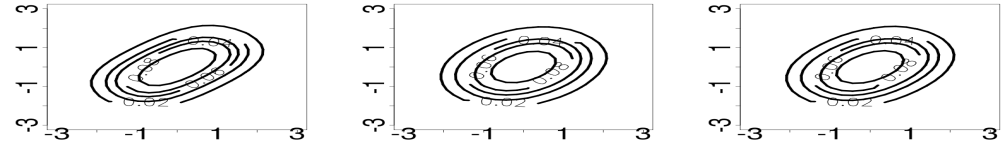
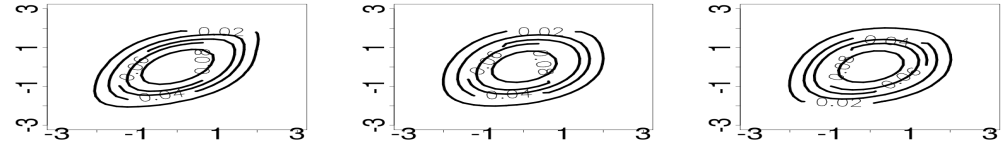
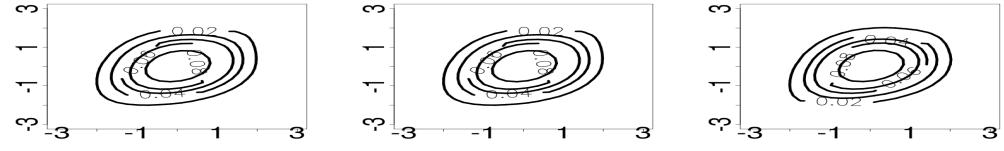
Contour plots in Table 6.2 suggest a rejection of the Gumbel model for all 3 pairs, since we did not detect a clear upper tail dependence in every empirical contour plot. Clayton copula also seems to be a bad choice for Pairs (C, B) and (A, D), since their empirical contour plots did not show a clear lower tail dependence as well. The empirical contour plot of $(B, \mathbf{1} - A)$ has a small tendency to V-notch in the lower left corner. Therefore, besides Normal and T copulas, Clayton copula may be also a good choice for $(B, \mathbf{1} - A)$.

λ-plots

Figure 6.4 shows the plots of empirical $\lambda_n(t)$ and 7 corresponding theoretical $\lambda(t)$ for pairs (C, B), (A, D) and $(B, \mathbf{1} - A)$, respectively. The reason to use $(B, \mathbf{1} - A)$ instead of (B, A) in λ-plots is the same as in contour plots. For λ-plots, the λ-function of rotated Archimedean copulas must be simulated, since a rotated Archimedean copula does not have explicit BIPIT form. This leads to less accuracy of the test, which is the case for (B, A). Note that for the simulation we use sample size N=5000.

Figure 6.4 shows Figure 6.4 in detail (the valley of each plot in Figure 6.4) so that we can distinguish between lines more efficiently.

Table 6.2: Empirical Contour plots (1st row) for all 3 pairs (C, B) (left), $(B, 1-A)$ and (A, D) (right) and theoretical Contour plots of different copula families. Contour levels (0.02, 0.04, 0.06, 0.08, 0.12) are plotted in each plot.

Copula	(C,B), $\tau_n = 0.313$	(B,**1**-A), $\tau_n = -0.158$	(A,D), $\tau_n = 0.196$E
Emp.			
Normal			
T			
Clayton			
Gumbel			
Frank			
BB1			
BB7			

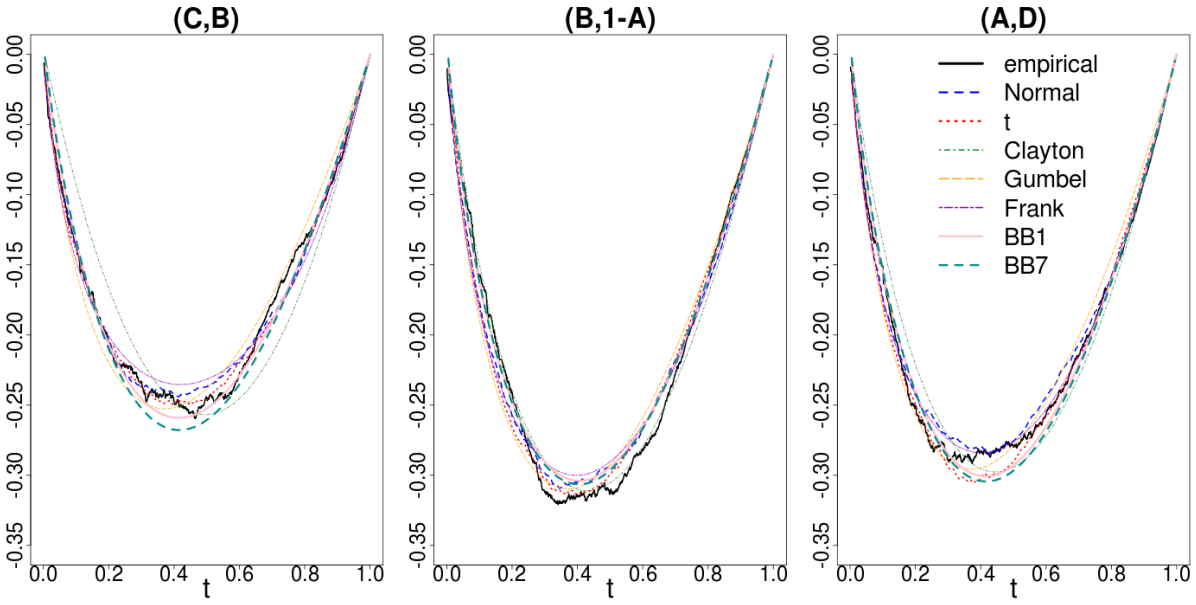


Figure 6.4: Plots of empirical $\lambda_n(t)$ and theoretical $\lambda(t)$ for all 3 pairs (C,B) (left, $\tau_n = 0.313$), (B,**1** -A) (middle, $\tau_n = -0.158$) and (A,D) (right, $\tau_n = 0.196$).

The plots in Fig 6.4 suggest that T and Normal copula might be the best models for all 3 pairs. The left plot for (C,B) indicate that BB1, BB7 , Gumbel and Clayton copulas do not fit well. The right plot for (A, D) rejects the Clayton, BB1 and BB7 copulas. From the middle plots for $(B, \mathbf{1} - A)$**,** we can hardly tell which model is better. The hypothesis we made out of the contour plots seems not to be confirmed. Therefore, we do not want to filter out any model before the GOF tests. We need to test all 7 models in the GOF tests.

GOF tests

Since we are not able to eliminate any candidates using the graphical methods, we have to run the GOF tests for all 7 models. For the same reasons mentioned in the graphical tests, all 7 models will be tested for the rotated data $(B, \mathbf{1} - A)$. From the "Copula" section, we know that Normal, T and Frank copula families are defined for $\tau \in (-1, 1)$. Therefore, we will test these 3 models for data pair (B, A) as well. Table 6.3 lists the p-values of both kinds of GOF tests developed in Chapter 5, the tests based on Kendall's transform and the tests based on empirical copula process. We denote the p-values of tests base on Kendall's transform as $p^{S_n(\mathbb{K}_n)}$ and $p^{T_n(\mathbb{K}_n)}$ corresponding to the two different statistics S_n and T_n. As mentioned before, we use the R-function *gofCopula()* in the R-package *copula* to compute the p-value of tests based on empirical process and denote the p-value as $p^{S_n(\mathbb{C}_n)}$. Note that *gofCopula()* is implemented only for the test statistic S_n and only for 5 models. Thus, we use $p^{S_n(\mathbb{C}_n)}$ here as a reference.

Table 6.3 reports that T copula has the highest p-values for pairs (C, B) and (A, D) in the tests based on Kendall's transform. The values of $p^{S_n(\mathbb{C}_n)}$ also confirm this. Thereby we select T copula with no doubt. However, the decision for (B, A) is difficult to make, since

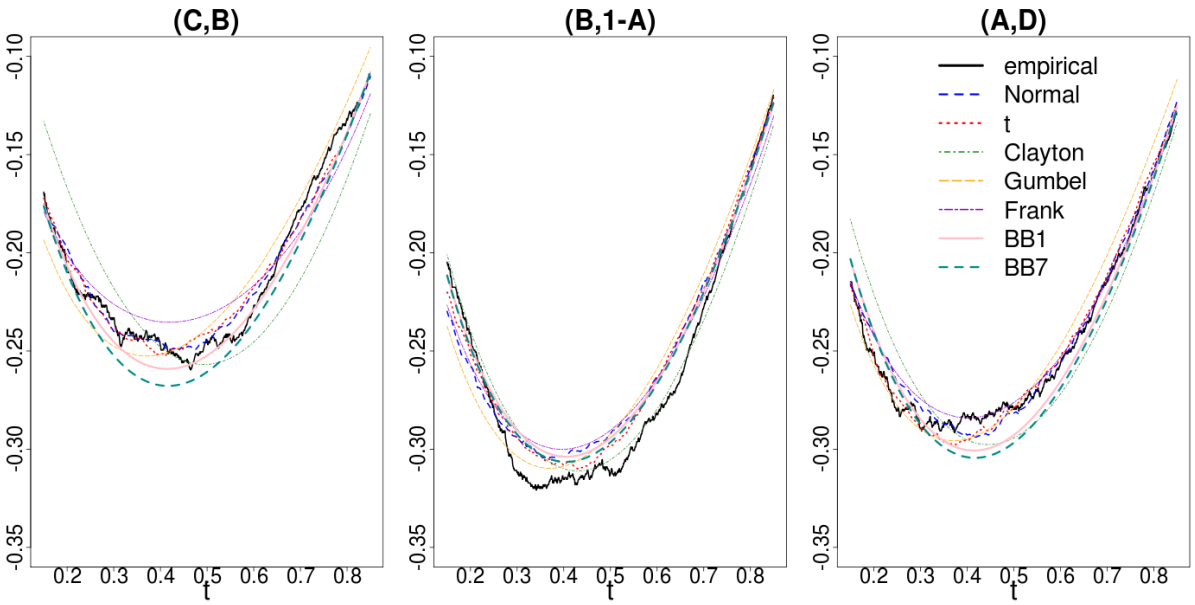


Figure 6.5: Details plot of Figure 6.4.

Table 6.3: P-values of the tests based on Kendall's transform (denoted as $p^{S_n(\mathbb{K}_n)}$ and $p^{T_n(\mathbb{K}_n)}$) and the tests based on empirical process (denoted as $p^{S_n(\mathbb{C}_n)}$) for data pair in Tree T_1

Pair	P-values	N	T	C	G	F	BB1	BB7
(C, B)	$p^{S_n(\mathbb{K}_n)}$	0.731	**0.958**	0	0	0.016	0.393	0.045
	$p^{T_n(\mathbb{K}_n)}$	0.627	**0.856**	0	0.022	0.076	0.470	0.068
	$p^{S_n(\mathbb{C}_n)}$	0.40	**0.427**	0	0.002	0.001	-	-
(B, A)	$p^{S_n(\mathbb{K}_n)}$	0.653	**0.847**	-	-	0.475	-	-
	$p^{T_n(\mathbb{K}_n)}$	0.726	**0.816**	-	-	0.601	-	-
	$p^{S_n(\mathbb{C}_n)}$	0.001	**0.082**	-	-	0	-	-
$(B, \mathbf{1} - A)$	$p^{S_n(\mathbb{K}_n)}$	0.034	**0.434**	**0.388**	0	0	0.075	0.145
	$p^{T_n(\mathbb{K}_n)}$	0.024	**0.326**	**0.317**	0	0	0.105	0.179
	$p^{S_n(\mathbb{C}_n)}$	0.001	0.009	**0.030**	0	0	-	-
(A, D)	$p^{S_n(\mathbb{K}_n)}$	0.361	**0.575**	0.001	0.004	0.019	0.296	0.157
	$p^{T_n(\mathbb{K}_n)}$	0.372	**0.392**	0.006	0.002	0.026	0.350	0.167
	$p^{S_n(\mathbb{C}_n)}$	0.492	**0.574**	0	0	0.220	-	-

we detected two appropriate models for $(B, \mathbf{1} - A)$. We look at the results for $(B, \mathbf{1} - A)$. Although we have that

$$p^{S_n(\mathbb{K}_n)}(T) = 0.434 > p^{S_n(\mathbb{K}_n)}(C) = 0.388 \text{ and } p^{T_n(\mathbb{K}_n)}(T) = 0.326 > p^{T_n(\mathbb{K}_n)}(C) = 0.317, \tag{6.1}$$

these values are very close to each other. Since we used 1000 bootstraps in these tests and the scenario study in the last chapter showed that the p-value can be varied in a restricted small interval by replication, the relationships in (6.1) do not always hold. With $\tau_n(B, \mathbf{1} - A) = 0.158$ the KLIC function shows

$$KLIC_{sym}(T, C) = 0.024.$$

Therefore it is hard to say which one between the Clayton and the T copulas is better. Since we are using financial data here, and the experience we get from several tests tell us that the T copula is usually a good fit, we select T copula here.

Although we selected the T copula as our model in this step, we are still interested in the case that the rotated Clayton copula is chosen. This alternative path will be briefly investigated in the "Final model" section.

Copula selection for the next levels

Now we need to find appropriate models for the pairs in Trees T_2 and T_3. The same routine will be done as in the last section. Remember that for the nodes C, B, A and D in Tree T_1 we used their observations as the underlying data directly. From now on we need to compute h-functions defined in Eq. (2.11) first and use the resulted data as our underlying data in each node. For instance, we want to determine the copula which corresponds to the first edge CA|B in Tree 2, i.e. $C_{CA|B}(F(C|B), F(A|B)))$. Thus, $F(C|B)$ and $F(A|B)$ are needed. They can be computed by the h-functions. For example,

$$F(A|B) = h(A, B, \hat{\Theta}) = \frac{\partial C_{A,B}(A, B, \hat{\Theta})}{\partial B}$$

where B and A are known and $\hat{\Theta}$ and $C_{A,B}(A, B, \hat{\Theta}) = C_{B,A}(B, A, \hat{\Theta})$ were determined in the last level. Thereby $F(A|B)$ can be computed, as well as $F(C|B)$. For sampling data from the h-functions, we used the C-function *Hfunc()* from the R-package *CDVineMLE* which is provided by Mr. Schepsmeier and Dr. Almeida.

In Tree T_2, we have the copulas $C_{CA|B}(F(C|B), F(A|B))$ and $C_{BD|A}(F(B|A), F(D|A))$ to determine. We start with modeling $C_{CA|B}$.

Using the same procedure as in T_1, we compute the empirical $\tau_n(F(C|B), F(A|B))$ and test the independence first. One finds

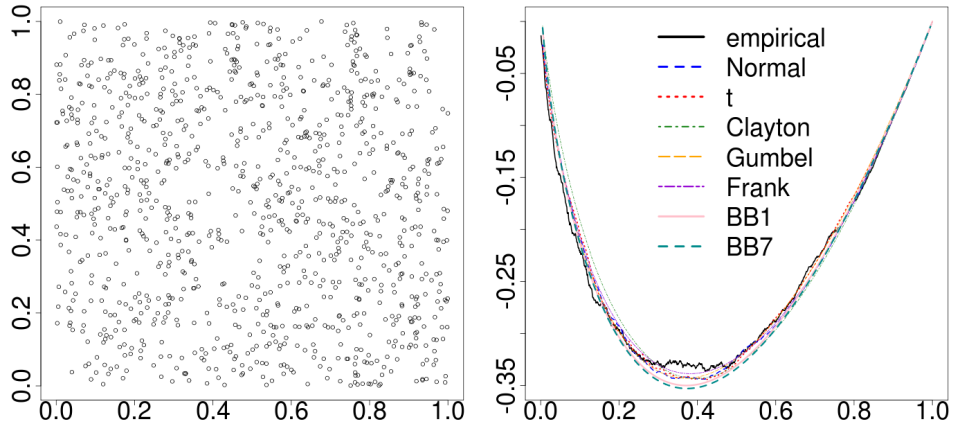
$$\tau_n(F(C|B), F(A|B)) = -0.068,$$

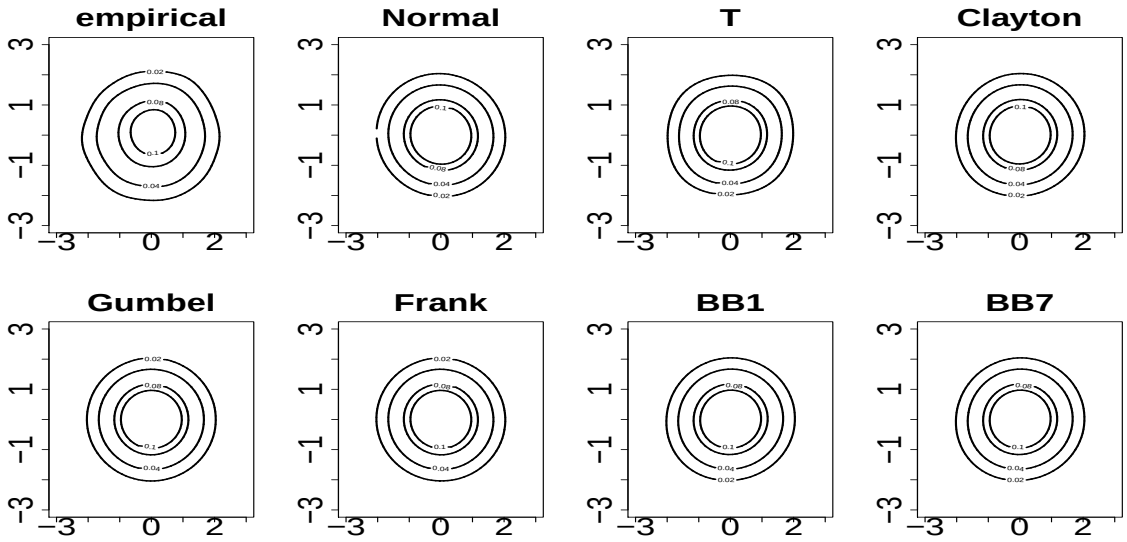
and the p-values of the corresponding independence test is given by

$$p = 0.001.$$

Table 6.4: Scatter plot of $(F(C|B), F(A|B))$; λ-plots and contour plots (plotted contour levels: 0.02, 0.04, 0.08, 0.1, 0.2) of rotated data $(F(C|B), \mathbf{1} - F(A|B))$; GOF tests of the copula $C_{CA|B}$.

$C_{CA|B}(F(C|B), F(A|B)),\ \ \tau_n = -0.068$





Pair	P-value	N	T	C	G	F	BB1	BB7
$(F(C\|B), \mathbf{1} - F(A\|B))$	$p^{S_n(\mathbb{K}_n)}$	**0.76**	0.566	0.005	0.484	0.441	0.128	0.055
	$p^{T_n(\mathbb{K}_n)}$	**0.826**	0.648	0.008	0.573	0.505	0.141	0.062
	$p^{S_n(\mathbb{C}_n)}$	0.407	0.248	0.009	**0.680**	0.449	-	-
$(F(C\|B), F(A\|B))$	$p^{S_n(\mathbb{K}_n)}$	0.653	**0.817**	-	-	0.271	-	-
	$p^{T_n(\mathbb{K}_n)}$	0.739	**0.958**	-	-	0.135	-	-
	$p^{S_n(\mathbb{C}_n)}$	0.388	0.295	-	-	**0.438**	-	-

Possible copulas: Normal, T, G, F, BB1

Selected copula: T

Therefore they are weakly negative dependent. For the same reasons described in the test of (B, A), we use the rotated $(F(C|B), 1 - F(A|B))$ to plot the λ-function and contour plots. All these plots and the results from the GOF tests are summarized in Table 6.4.

According to the λ-plots, we may reject the BB1 and BB7 models. However, we cannot distinguish between the contour plots. Since τ_n is near zero, we expect a weak dependence structure between variables and the dependencies in tails cannot be observed in contour plots. Therefore, it is unsurprising to have this conclusion. Tests based on Kendall's transform suggest Normal and T copulas as our best model. The estimated degree of freedom of T copula is $\hat{\nu} = 20$. Thus, T and Normal copula are almost identical in this case. We select T copula in this step, since T copula appears to have relatively higher p-values in Table 6.4.

Now we look at Edge BD|A, i.e. $C_{BD|A}(F(B|A), F(D|A))$. The empirical τ_n and the p-value of the corresponding independence test are given by

$$\tau_n(F(B|A), F(D|A)) = 0.011 \quad \text{and} \quad p = 0.583.$$

This test reports that $F(B|A)$ and $F(D|A)$ are independent. Thus, we have

$$C_{BD|A}(F(B|A), F(D|A)) = 1$$

and the corresponding h-function for the next level can be simplified as

$$F(D|BA) = \frac{\partial C_{BD|A}(F(D|A), F(B|A))}{\partial F(B|A)} = F(D|A).$$

It means that Node BD|A in Tree T_3 inherit the information of Node AD directly.

Obviously, an independent copula is selected in this step. If we have to select a model from the 7 candidate, Normal copula would be a good choice, since independence leads to a equivalent choice between copulas and Normal copula is the best known and implemented model.

We now have all models selected for Tree 2. It remains to find the model for the last edge CD|BA. One finds

$$\tau_n(F(C|BA), F(D|BA)) = -0.09$$

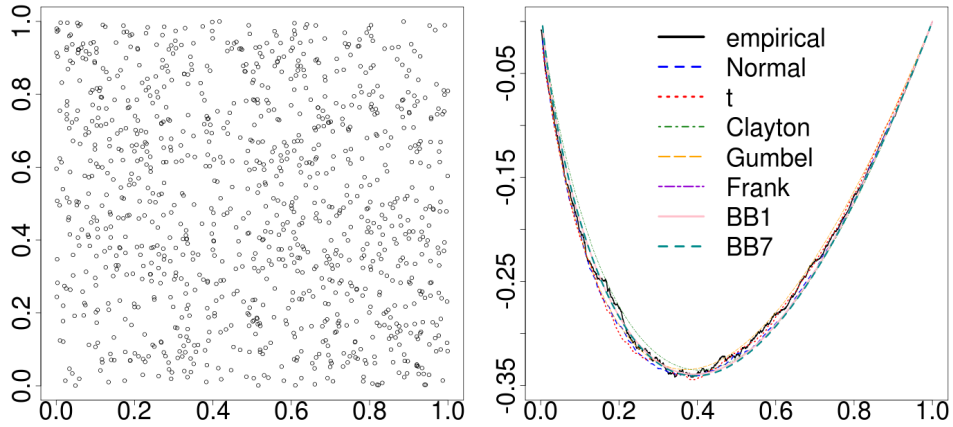
and p-value of the corresponding independence test is 0. Thereby, $F(C|BA)$ and $F(D|BA)$ are negative dependent. Graphical and GOF tests need to be carried out for $(F(C|BA), 1 - F(D|BA))$. Table 6.5 lists the Scatter plot, contour plots, plots of λ-function and the p-values of GOF tests.

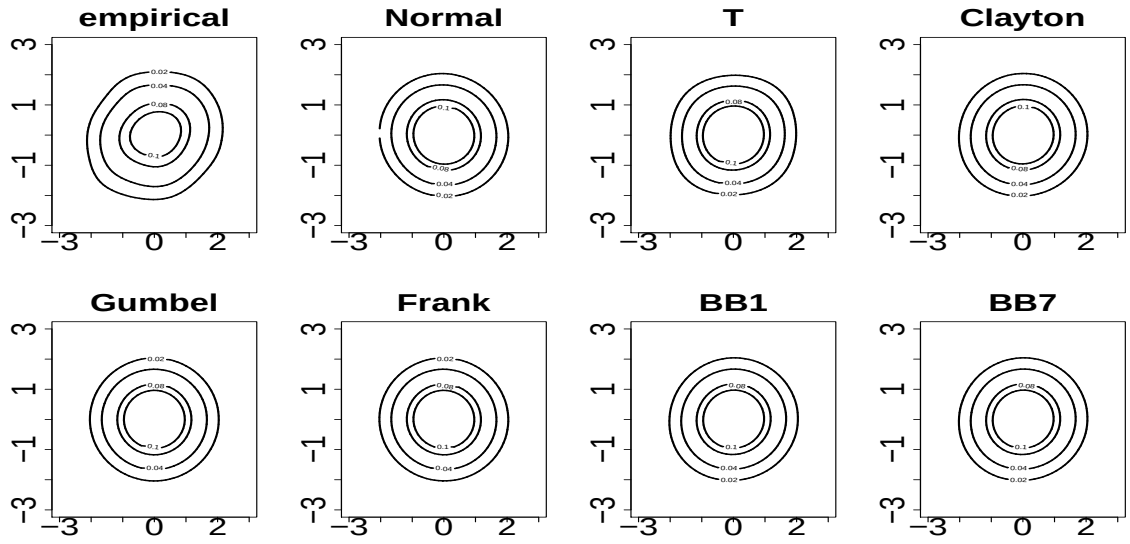
Unfortunately, contour plots or the λ-plots between different copulas are so close to each other, they cannot help us to reject any of the 7 copulas. The scatter plot does not show any dependence structure. This demonstrates the similarity between contour plots and λ-plots. The same conclusion can also be found in GOF tests. Results show that every copula can be considered as an acceptable model for $C_{CD|BA}$, since the p-values are greater than 0.05 overall. We suggest the T , BB1 and BB7 copulas, since they have the highest p-values. With $\tau_n(F(C|BA), \mathbf{1} - F(D|BA)) = 0.09$ we have the values of KLIC functions as follows,

$$KLIC_{sym}(T, BB1) = KLIC_{sym}(T, BB7) = 0.02 \text{ and } KLIC_{sym}(BB1, BB7) = 0.$$

Table 6.5: Scatter plot of $(F(C|BA), F(D|BA))$; λ-plots and contour plots (plotted contour levels: 0.02, 0.04, 0.08, 0.1, 0.2) of rotated data $(F(C|BA), \mathbf{1} - F(D|BA))$; GOF tests of the copula .

$C_{CD|BA}(F(C|BA), F(D|BA)), \ \tau_n = -0.09$



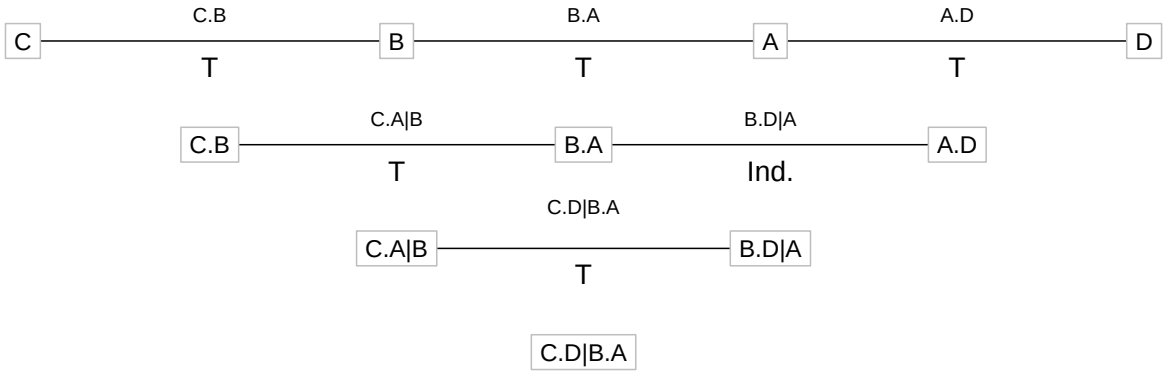


Pair	P-value	N	T	C	G	F	BB1	BB7
$(F(C\|BA), \mathbf{1} - F(D\|BA))$	$p^{S_n(\mathbb{K}_n)}$	0.488	**0.825**	0.44	0.091	0.231	0.839	0.775
	$p^{T_n(\mathbb{K}_n)}$	0.472	**0.752**	0.389	0.098	0.428	0.772	0.759
	$p^{S_n(\mathbb{C}_n)}$	0.883	**0.878**	0.211	0.361	0.651	-	-
$(F(C\|BA), F(D\|BA))$	$p^{S_n(\mathbb{K}_n)}$	0.518	**0.896**	-	-	0.717	-	-
	$p^{T_n(\mathbb{K}_n)}$	0.511	**0.666**	-	-	0.593	-	-
	$p^{S_n(\mathbb{C}_n)}$	0.882	**0.895**	-	-	0.644	-	-

Possible copulas: Normal, T, G, F, BB1

Selected copula: T

Figure 6.6: Final model, where Ind. stands for independent copula.



Thus, the choice between T, BB1 and BB7 copulas are equivalent. We choose T copula here.

Eventually, we completed the D-vine decomposition and determined at least one appropriate model for the pair-copula in each step. Therefore the multivariate (or four-dimensional) density function for our data set can be determined using Eq. (2.12) with our selected copula models and given margins.

6.5 Final model

Using the D-vine structure shown in Figure 6.2 and the selected models in each step, the final model can be expressed graphically as in Figure 6.6.

As we mentioned at the end of subsection 6.4, we found two appropriate models for C_{BA}. The choice between T and the rotated Clayton copulas was difficult. Eventually we selected the T copula. What will change, if we select the rotated Clayton C^{RC}? First, we need to recompute $F(A|B)$ and $F(B|A)$ with C^{RC}:

$$F(A|B) = \frac{\partial C^{RC}(A,B)}{\partial B} = \frac{\partial B - \partial C^{C}(\mathbf{1}-A,B)}{\partial B} = 1 - \frac{\partial C^{C}(B,\mathbf{1}-A)}{\partial B}$$

and

$$F(B|A) = \frac{\partial C^{RC}(B,A)}{\partial A} = \frac{\partial B - \partial C^{C}(B,\mathbf{1}-A)}{\partial A} = \frac{\partial C^{C}(B,\mathbf{1}-A)}{\partial \mathbf{1}-A}.$$

Then we need to find new models for $C_{CA|B}$ and $C_{BD|A}$, since $F(A|B)$ and $F(B|A)$ are changed in this case. We start with computing τ_n and testing independence as usual. For $C_{BD|A}$ one finds

$$\tau_n(F(B|A), F(D|A)) = 0.006 \text{ and } p = 0.754,$$

where $F(D|A)$ remains unchanged. Therefore $F(B|A)$ and $F(D|A)$ are still independent and our selected model is still the independent copula in this step.

For $C_{CA|B}$ we have

$$\tau_n(F(C|B), F(A|B)) = -0.068 \text{ and } p = 0.001.$$

Table 6.6: P-values of the tests based on Kendall's transform for $C_{CA|B}$

Pair	P-value	N	T	C	G	F	BB1	BB7
$(F(C\|B), \mathbf{1} - F(A\|B))$	$p^{S_n(\mathbb{K}_n)}$	0.239	0.172	0.007	0.508	0.528	0.085	0.037
	$p^{T_n(\mathbb{K}_n)}$	0.408	0.072	0.024	0.444	0.399	0.067	0.033
$(F(C\|B), F(A\|B))$	$p^{S_n(\mathbb{K}_n)}$	**0.579**	0.31	-	-	0.154	-	-
	$p^{T_n(\mathbb{K}_n)}$	**0.51**	0.274	-	-	0.08	-	-

Table 6.7: P-values of the tests based on Kendall's transform for $C_{CD|BA}$

Pair	P-value	N	T	C	G	F	BB1	BB7
$(F(C\|BA), \mathbf{1} - F(D\|BA))$	$p^{S_n(\mathbb{K}_n)}$	0.664	0.81	0.47	0.106	0.256	**0.885**	**0.852**
	$p^{T_n(\mathbb{K}_n)}$	0.672	0.652	0.586	0.257	0.409	**0.974**	**0.956**
$(F(C\|BA), F(D\|BA))$	$p^{S_n(\mathbb{K}_n)}$	0.764	**0.888**	-	-	0.689	-	-
	$p^{T_n(\mathbb{K}_n)}$	0.65	**0.963**	-	-	0.502	-	-

Review the results in Subsection 6.4, we find that τ_n and the p-values remain approximately unchanged, whether T or rotated Clayton copula is selected for C_{BA}. According to the results from the GOF tests (Table 6.6), we select Normal copula for $C_{CA|B}$ here. The last step is to find new model for $C_{CD|BA}$. With

$$\tau_n(F(C|BA), F(D|BA)) = -0.089 \text{ and } p = 0$$

we find negative dependence between variables in this step. P-values of GOF tests in Table 6.7 suggest T, BB1 and BB7 copulas. We still select T copula for $C_{CD|BA}$.

Consequently, we receive an alternative final model if we select rotated Clayton copula for C_{BA}. This alternative model is displayed in Figure 6.7.

Table 6.8 lists the estimated parameters of each candidate copula model in the tests for the final model. Table 6.9 lists the estimated parameters for the alternative final model. Since we selected a different copula family for (B, A) in the alternative final model than in the first final model, the environments are only changed for the trees T_2 and T_3. Consequently, in Table 6.9 we only list the values of estimated parameters for the 3 pairs in T_2 and T_3.

Figure 6.7: Alternative model, where Ind. stands for independent copula.

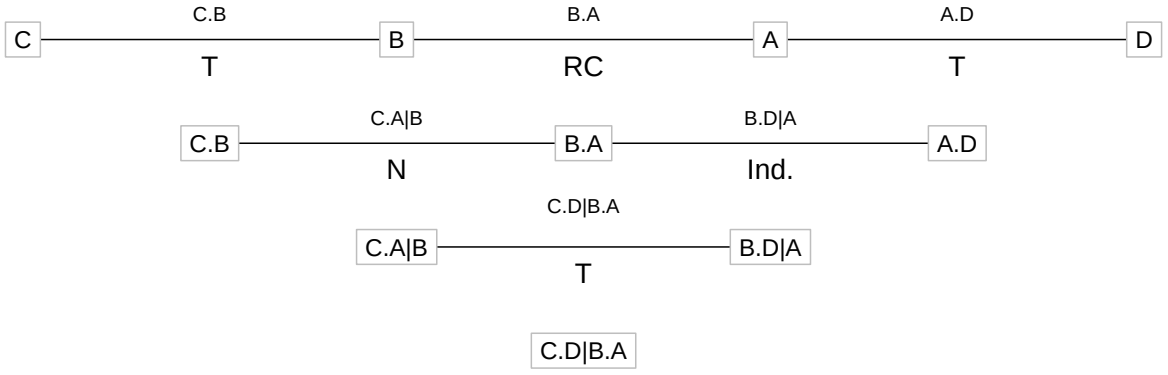


Table 6.8: Estimated parameters of 7 models in each test for the final model

Tree	Pair	Parameter	Model						
			N	T	C	G	F	BB1	BB1
T_1	(C, B)	θ	0.473	0.473	0.913	1.456	3.07	0.267	1.305
	Edge: C-B	δ or ν	-	16.545	-	-	-	1.255	0.455
		Λ_U	0	0	-	0.391	-	0.263	0.299
		Λ_L	0	0	0.468	-	-	0.127	0.219
	$(B, \mathbf{1} - A)$	θ	0.246	0.246	0.376	1.188	1.45	0.207	1.128
	Edge: B-A	δ or ν	-	4.23	-	-	-	1.099	0.273
		Λ_U	0	0.053	-	0.208	-	0.121	0.151
		Λ_L	0	0.053	0.158	-	-	0.048	0.079
	(A, D)	θ	0.303	0.3	0.488	1.244	1.82	0.268	1.088
	Edge: A-D	δ or ν	-	14.432	-	-	-	1.081	0.334
		Λ_U	0	0	-	0.254	-	0.102	0.109
		Λ_L	0	0	0.242	-	-	0.092	0.125
T_2	$(F(C\|B), \mathbf{1} - F(A\|B))$	θ	0.106	0.106	0.145	1.073	0.61	0.01	1.046
	Edge: CB-BA	δ or ν	-	20	-	-	-	1.046	0.036
		Λ_U	0	0	-	0.092	-	0.06	0.06
		Λ_L	0	0	0.00	-	-	0	0
	$(F(B\|A, F(D\|A))$	θ	0.017	0.017	0.022	1.011	0.09	0.037	1.018
	Edge: BA-AD	δ or ν	-	17.803	-	-	-	1.012	0.041
		Λ_U	0	0	-	0.015	-	0.016	0.024
		Λ_L	0	0	0	-	-	0	0
T_3	$(F(C\|BA), F(D\|BA))$	θ	0.141	0.141	0.199	1.099	0.81	0.126	1.018
	Edge: CA|B-BD|A	δ or ν	-	20	-	-	-	1.021	0.142
		Λ_U	0	0	-	0.121	-	0.028	0.024
		Λ_L	0	0	0.03	-	-	0.004	0.008

Table 6.9: Estimated parameters of 7 models in each test for the alternative final model. The values are only listed for pairs which are changed from the first final model.

Tree	Pair	Parameter	Model						
			N	T	C	G	F	BB1	BB1
T_2	$(F(C\|B), \mathbf{1} - F(A\|B))$ Edge: CB-BA	θ	0.107	0.107	0.147	1.073	0.61	0.009	1.042
		δ or ν	-	20	-	-	-	1.044	0.034
		Λ_U	0	0	-	0.093	-	0.058	0.056
		Λ_L	0	0	0.09	-	-	0	0
	$(F(B\|A, F(D\|A))$ Edge: BA-AD	θ	0.01	0.01	0.013	1.006	0.05	0.041	1.008
		δ or ν	-	20	-	-	-	1.0004	0.041
		Λ_U	0	0	-	0.009	-	0.0005	0.011
		Λ_L	0	0	0	-	-	0	0
T_3	$(F(C\|BA), F(D\|BA))$ Edge: CA|B-BD|A	θ	0.14	0.14	0.196	1.098	0.8	0.131	1.017
		δ or ν	-	20	-	-	-	1.02	0.145
		Λ_U	0	0	-	0.12	-	0.026	0.023
		Λ_L	0	0	0.029	-	-	0.005	0.008

Chapter 7

Applications to Asian exchange rates

7.1 Learning data set

Our data set consists of 5 time series of daily exchange rates of different Asian countries from 27. November 2006 until 25. November 2010. Each time series contains 1022 observations. The basic currency is US-Dollar, e.g. 1 US-Dollar = 100 Japanese Yen. For the further notation we named the exchange rates notated with US-Dollar as follows: **JPY** (Japanese Yen), **KRW** (South-Korean Won), **INR** (Indian Rupee), **SGD** (Singapore Dollar) and **THB** (Thai Baht). A graphical illustration of the exchange rates and their log-returns can be seen in Figure 7.1.

For the following applications, the observations of each variable must be independent over time. Thus, we want to find an appropriate model for the serial log-returns of each exchange rate. We test three different models with Normal and T innovations:

- GARCH(1,1) model (Bollerslev [1986]).
- AR(1)-GARCH(1,1) model, i.e. AR(1) process (Brockwell and Davis [2002] and Brockwell and Davis [2009]) with GARCH(1,1) noise.
- ARMA(1,1)-GARCH(1,1) model, i.e. ARMA(1,1) process (Brockwell and Davis [2002] and Brockwell and Davis [2009]) with GARCH(1,1) noise.

After the residuals of the log-returns are filtered by these models, we use the Ljung-Box-Test (Ljung and Box [1978]) to test the null hypothesis that there is no autocorrelation left in the standardized (R) and squared standardized residuals (R^2). The p-values of these tests are listed in Table 7.1.

According to the results reported in Table 7.1, ARMA(1,1)-GARCH(1,1) is preferred in general. Thus for each series $i \in$ (INR, THB, SGD, KRW, JPY) we use the ARMA(1,1)-GARCH(1,1) to model the log-return x_i and filter the standardized residual z_i. The ARMA(1,1)-GARCH(1,1) is given as

$$\begin{aligned} x_{i,t} &= \mu_i + \Phi_i x_{i,t-1} + \sigma_{i,t} z_{i.t} + \Theta_i \sigma_{i.t-1} z_{i.t-1} \\ \sigma^2_{i,t} &= \omega + \alpha_i \epsilon^2_{i,t-1} + \beta_i \sigma^2_{i,t-1} \end{aligned}$$

where $E[z_{i,t}] = 0$ and $\mathrm{Var}(z_{i,t}) = 0$ and $\epsilon_{i,t-1} = \sigma_{i,t} z_{i,t}$. Since we are mainly interested in estimating the dependence structure of the risk factor, the standardized residual vectors

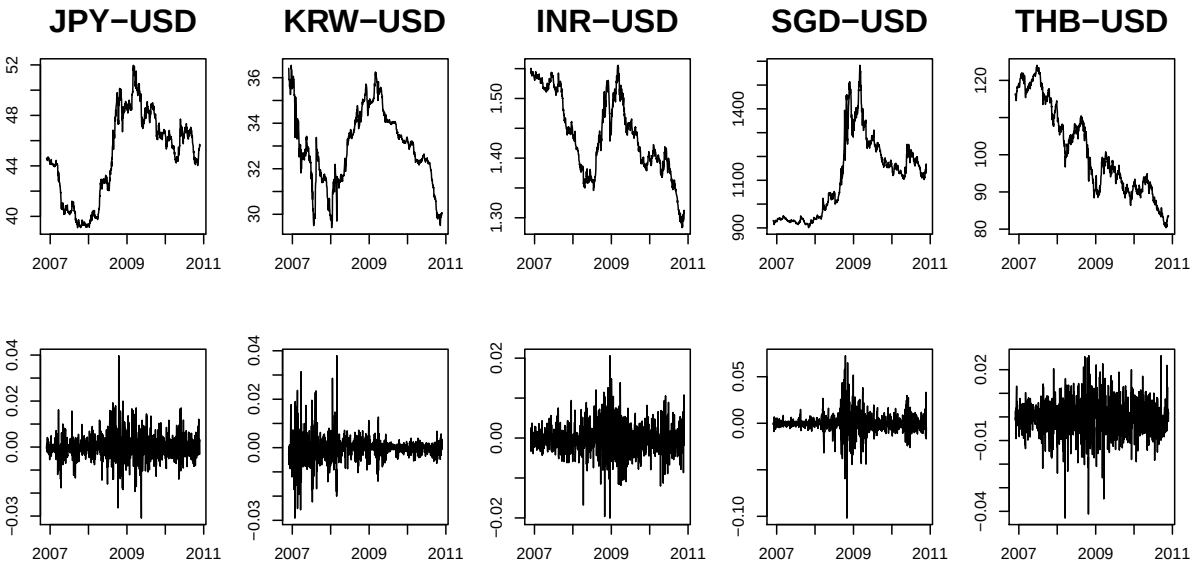


Figure 7.1: Asian exchange rates from 25. November 2006 to 25. November 2010

Table 7.1: p-values of the Ljung-Box-test of Asian log-returns using different models for the standardized residuals R and the squared standardized residuals R^2

	GARCH(1,1)				**AR(1)-GARCH(1,1)**				**ARMA(1,1)-GARCH(1,1)**			
	N		t		N		t		N		t	
	R	R^2	R	R^2	R	R^2	R	R^2	R	R^2	R	R^2
INR	0.07	0.92	0.07	0.91	0.09	0.98	0.07	0.95	**0.09**	**0.98**	0.05	0.95
THB	0	0.43	0.01	0.06	0.05	0.42	0.03	0.33	**0.74**	**0.61**	0.19	0.45
SGD	**0.46**	0.66	0.45	0.6	0.46	0.66	0.39	0.58	0.45	**0.68**	0.4	0.56
KRW	0.67	0.74	0.64	0.7	0.75	0.73	0.46	0.71	**0.85**	**0.74**	0.45	0.71
JPY	0.81	0.67	0.8	0.66	0.89	0.65	0.88	0.62	**0.9**	**0.65**	0.89	0.63

Table 7.2: Five selected D-vine structures for Tree 1 and the sum of their absolute empirical Kendall's τ

D-vine Structure	Sum of $\hat{\tau}_n$
INR-THB-SGD-KRW-JPY	**0.943**
INR-KRW-SGD-THB-JPY	0.868
THB-SGD-KRW-INR-JPY	0.887
THB-INR-SGD-KRW-JPY	0.853
THB-SGD-INR-KRW-JPY	0.86

are transformed to uniform variables using the empirical distribution functions before further modeling.

7.2 Copula selection

In the upper panels of Figure 7.2 we drawn the pairs plot of the transformed residuals. In the lower panels of Figure 7.2 we showed their empirical Kendall's τ and the p-values of their independence tests based on the empirical Kendall's τ. Positive dependencies are detected between KRW, INR, SGD and THB. For the pairs contained JPY their dependencies seem to be negative and/or very weak.

Based on the values of empirical Kendall's τ and the results of their independent tests displayed in Figure 7.2, we consider several D-vine structures for this level and select the one which maximize the sum of the corresponding absolute values of $\hat{\tau}_n$ (Table 7.2). According to Table 7.2 we determine the order of the data pairs on the first level of the D-vine structure as follows:

$$\text{INR} \overset{0.147}{-} \text{THB} \overset{0.287}{-} \text{SGD} \overset{0.0.361}{-} \text{KRW} \overset{-0.148}{-} \text{JPY}$$

where the values of their empirical Kendall's τ are listed associatively. The corresponding whole D-vine structure are mapped in Figure 7.3.

Since the whole D-vine structure is known, further modeling turns to the copula selection for each data pair displayed in the D-vine map (Figure 7.3). we conduct the selection by using the same methods as for the Norwegian data in the last section. Once again we use the EDA tools such as λ-plot, Contourplot and GOF tests based on Kendall's transform to find the appropriate pair-copula models and summarize their plots and results for each pair within tables, respectively. If we observe negative dependencies, i.e. $\hat{\tau}_n < 0$, for a data pair $(\mathbf{u}, \mathbf{v})$, we use the rotated pair $(\mathbf{u}, 1 - \mathbf{v})$ for copula selection.

Table 7.3-7.6 list the EDA results for all pairs on Tree 1. During the copula selection on Tree 1 we received several acceptable copula models for the same data pair. the choice between these models was difficult to make. For instance, the selection between T, BB1 and BB7 copulas for pair (SGD,KRW) was basically equivalent. we selected BB1 copula since the p-values for this model are slightly higher than the others. As for the pairs (INR, THB) and (THB, SGD) their Contourplots and λ-plots showed that the Frank copulas were the best fits, which were also proved by the GOF tests.

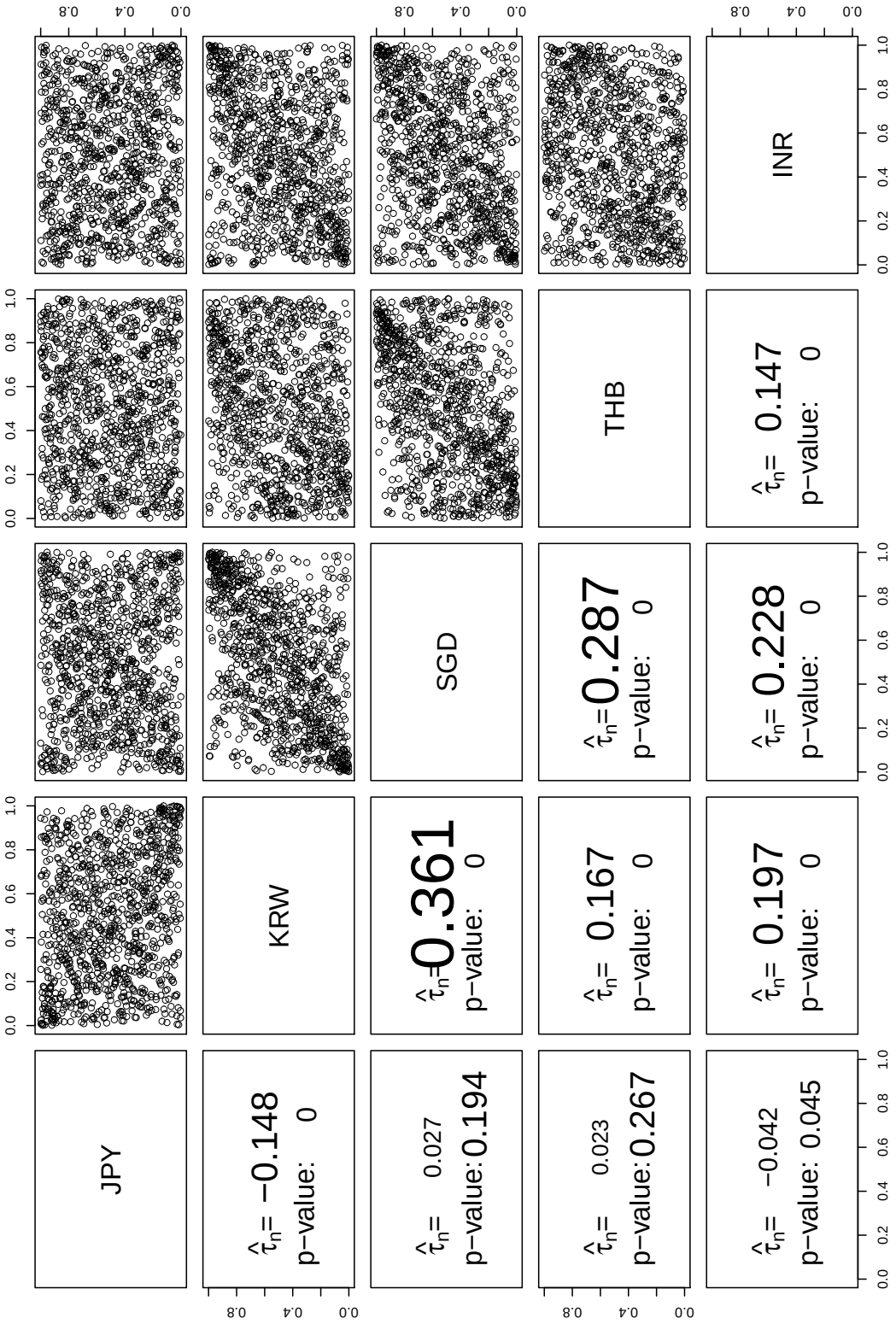
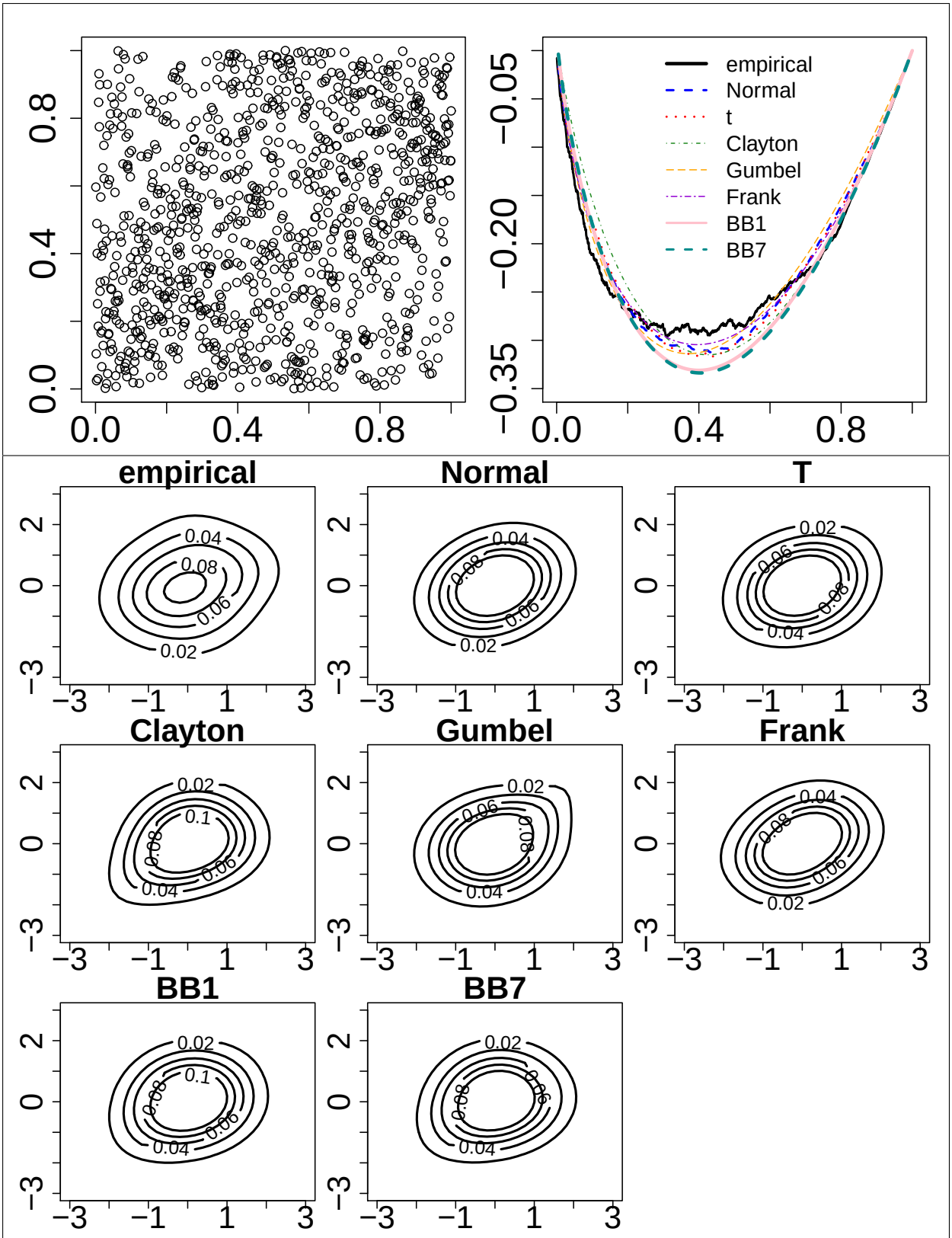


Figure 7.2: Upper panels: Pairs plot of the copula data converted from Asian exchange rates.
Lower panels: The according empirical Kendall's τ and the p-values of their independent tests.

Table 7.3: Pairs plot, λ-plot, Contourplot, and tests based on Kendall's transform for the data pair (INR, THB) on Tree 1

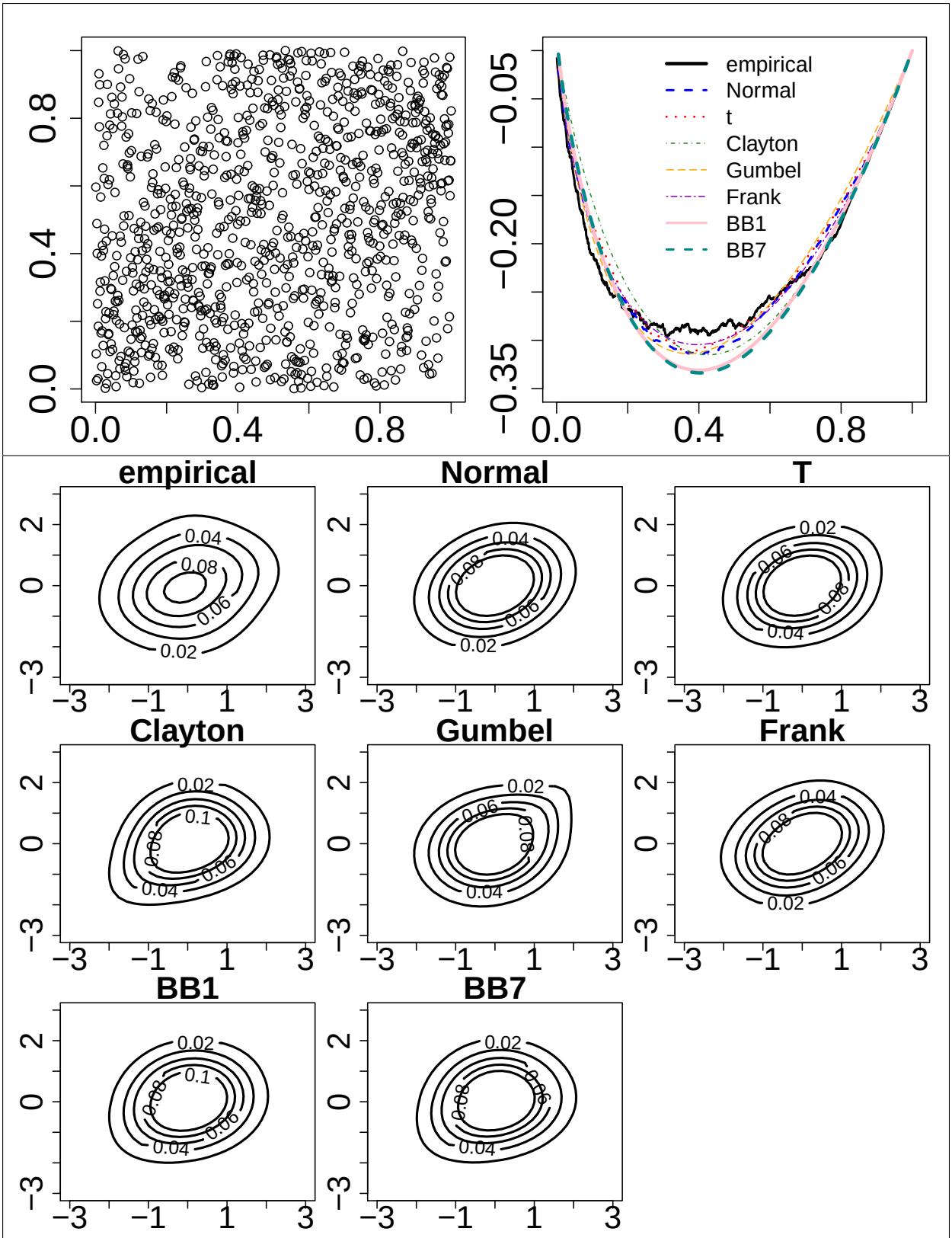
Copula $C_{INR,THB}$ ($\hat{\tau}_n = 0.147$)



P-value	N	T	C	G	F	BB1	BB7	Possible copula(s):
p^{S_n}	**0.074**	**0.070**	0	0.016	**0.078**	0	0	N, T, F
p^{T_n}	**0.105**	**0.091**	0	0.021	**0.162**	0	0	Selected copula: **F**

Table 7.4: Pairs plot, λ-plot, Contourplot, and tests based on Kendall's transform for the data pair (THB, SGD) on Tree 1

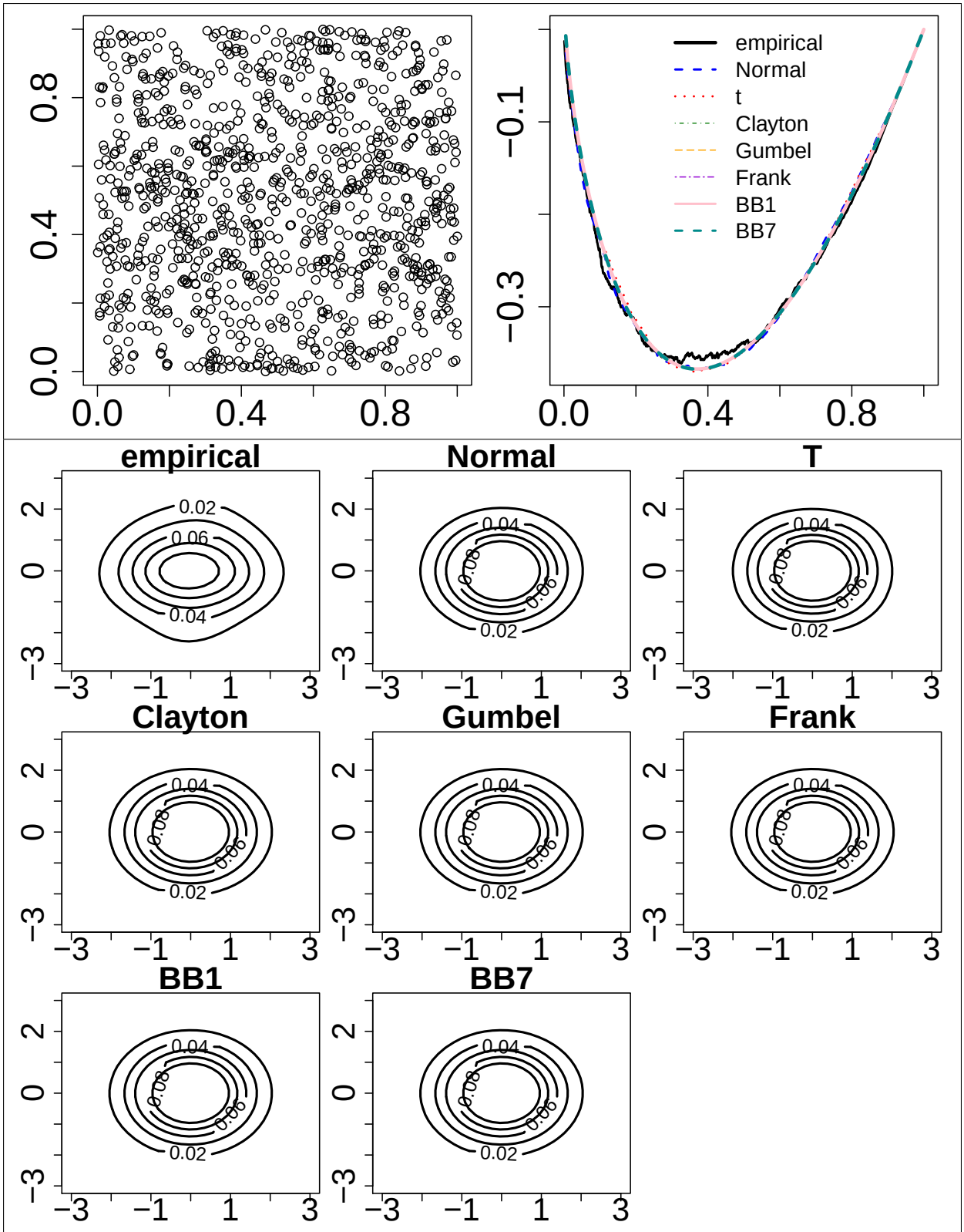
Copula $C_{THB,SGD}$ ($\hat{\tau}_n = 0.287$)



P-value	N	T	C	G	F	BB1	BB7	Possible copula(s):
p^{S_n}	**0.204**	0.060	0	0.011	**0.995**	0	0	N, , F
p^{T_n}	**0.259**	0.035	0	0.067	**0.991**	0	0	Selected copula: **F**

Table 7.5: Pairs plot, λ-plot, Contourplot, and tests based on Kendall's transform for the data pair (SGD, KRW) on Tree 1

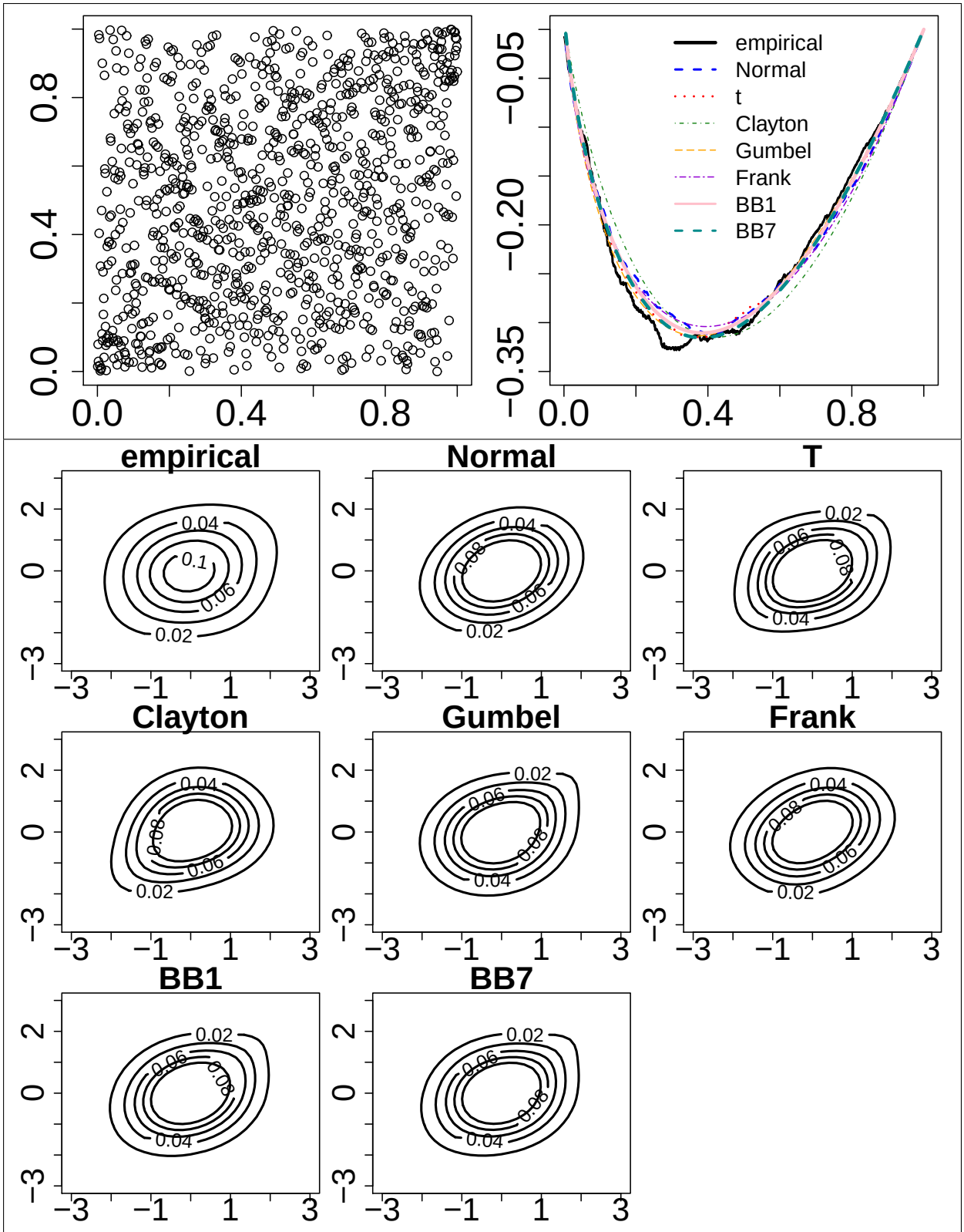
Copula $C_{SGD,KRW}$ $(\hat{\tau}_n = 0.361)$



P-value	N	T	C	G	F	BB1	BB7	Possible copula(s):
p^{S_n}	**0.304**	**0.432**	0	0.007	0	**0.447**	**0.412**	N, T, BB1,BB7
p^{T_n}	**0.297**	**0.415**	0	0	0	**0.416**	**0.403**	Selected copula: **BB1**

Table 7.6: Pairs plot, λ-plot, Contourplot, and tests based on Kendall's transform for the data pair (KRW, 1-JPY) on Tree 1

Copula $C_{KRW,JPY}$ $(\hat{\tau}_n = -0.148)$



	P-value	N	T	C	G	F	BB1	BB7	Possible copula(s):
p^{S_n}		**0.118**	**0.475**	0	0.007	0	**0.065**	**0.087**	N, T, BB1, BB7
p^{T_n}		**0.276**	**0.306**	0	0.001	0	**0.087**	**0.174**	Selected copula: **T**

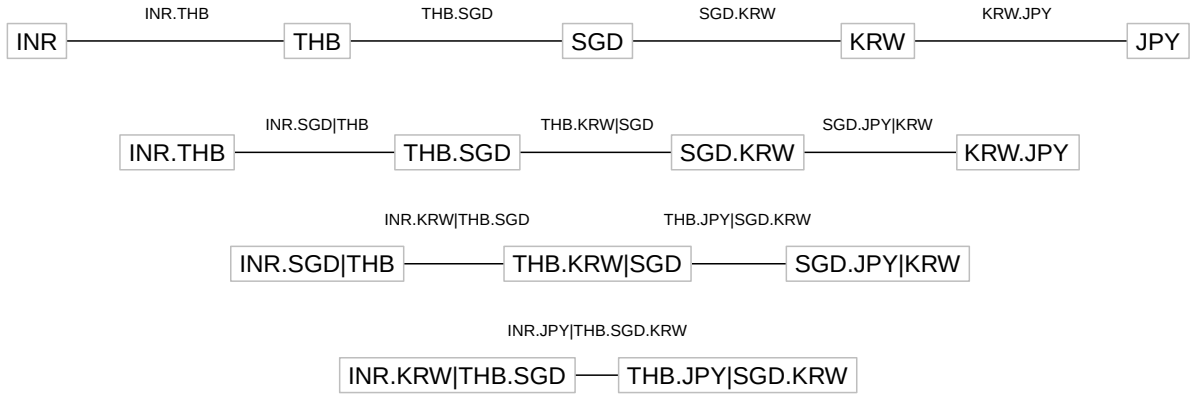


Figure 7.3: D-vine structure for the Asian exchange rate data

The Summarized EDA results for pairs on Tree 2 are listed in Table 7.7-7.9. Since the main dependence structures have been captured on the first tree, on the second level we observe weak dependencies in every pairs plot and thereby obtain low values of empirical Kendall's τ. Based on this fact choosing appropriate copula models on this level could be ambiguous, for instance the model section for (THB,SGD, SGD.KRW) and (SGD.KRW, KRW.JPY). The reason we choose Normal and Gumbel copulas for these two pairs is that they have the highest p^{S_n}. As we concluded in the simulation studies p-values based on statistic S_n is more efficient than the p-values based on T_n, therefore p^{T_n} is not the primary criteria here. For pair (INR.SGD, THB.SGD) we choose Frank copula since it has the highest p-value overall.

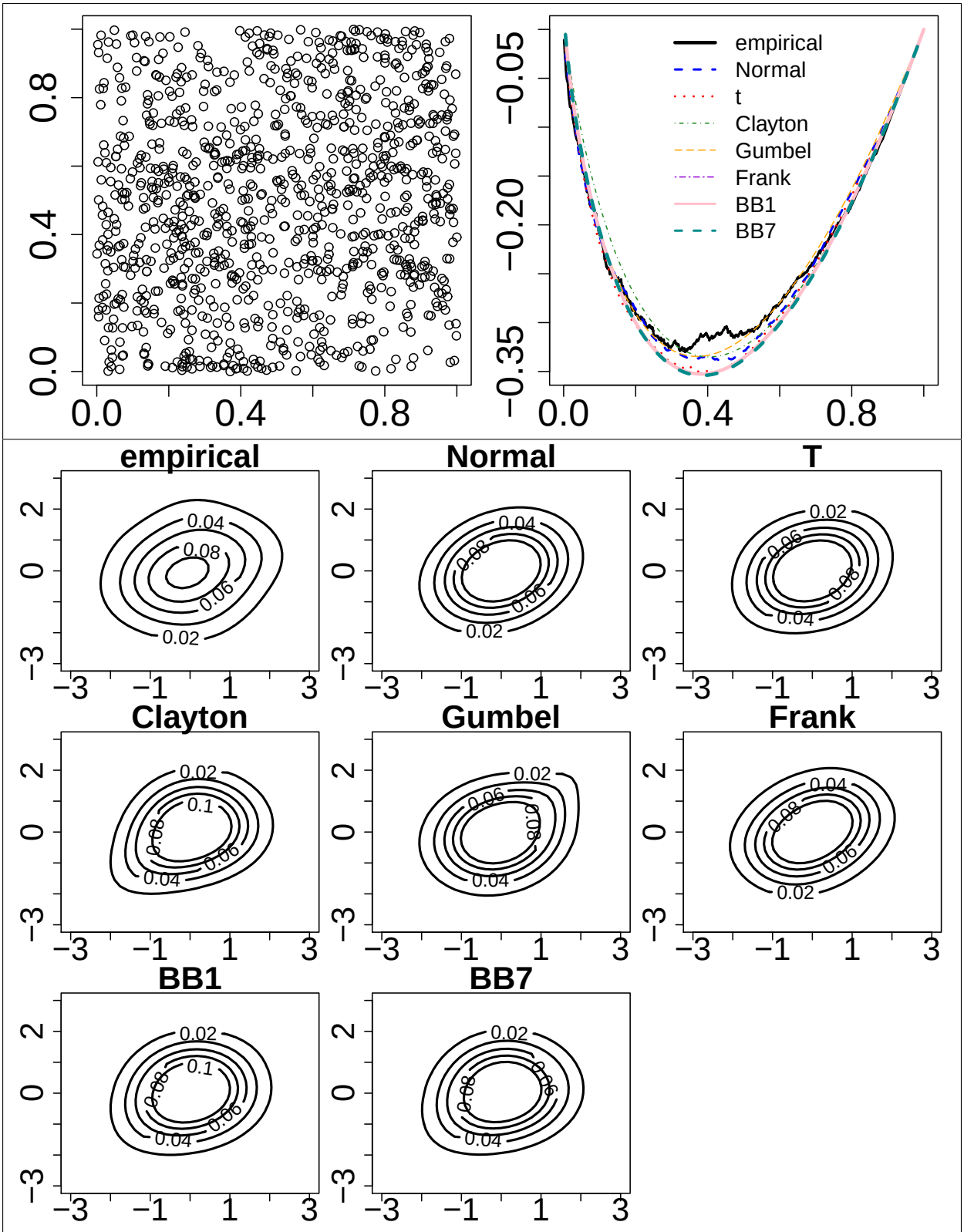
Table 7.10 and 7.11 report the results for the copula selection on Tree 3. One can see that the empirical Kendall's τ are almost zero, i.e. -0.002 and 0.001. Thus, the data are likely independent and we can hardly distinguish their λ-plots and the Contourplots. In this case, all models are accepted. According to the P-values listed in both tables, GOF tests suggest BB1 copula for the first pair of Tree 3 and Normal copula for the second since they have the highest p-values overall. However, our choices are not exclusive. One can certainly use any other model.

Table 7.12 shows the results for the last copula pair on Tree 4. Pairs plot shows that data is independent, which is also proved by the empirical Kendall's τ ($\hat{\tau}_n = 0$) . Thus, we do not expect that any differences between the Contour plots or λ-plots. GOF tests show that all models fit the data very well. We select the Normal copula here.

After a copula model is selected for Edge INR.KRW|THB.SGD - THB.JPY|SGD.KRW, we accomplished the D-vine decomposition. With the D-vine structure showed in Figure 7.3 and the selected copula models for each pair we are able to archive the final model and approach the appropriate multivariate distribution for our data set.

Table 7.7: Pairs plot, λ-plot, Contourplot, and tests based on Kendall's transform for the rotated data pair (INR.THB, 1-THB.SGD) on Tree 2

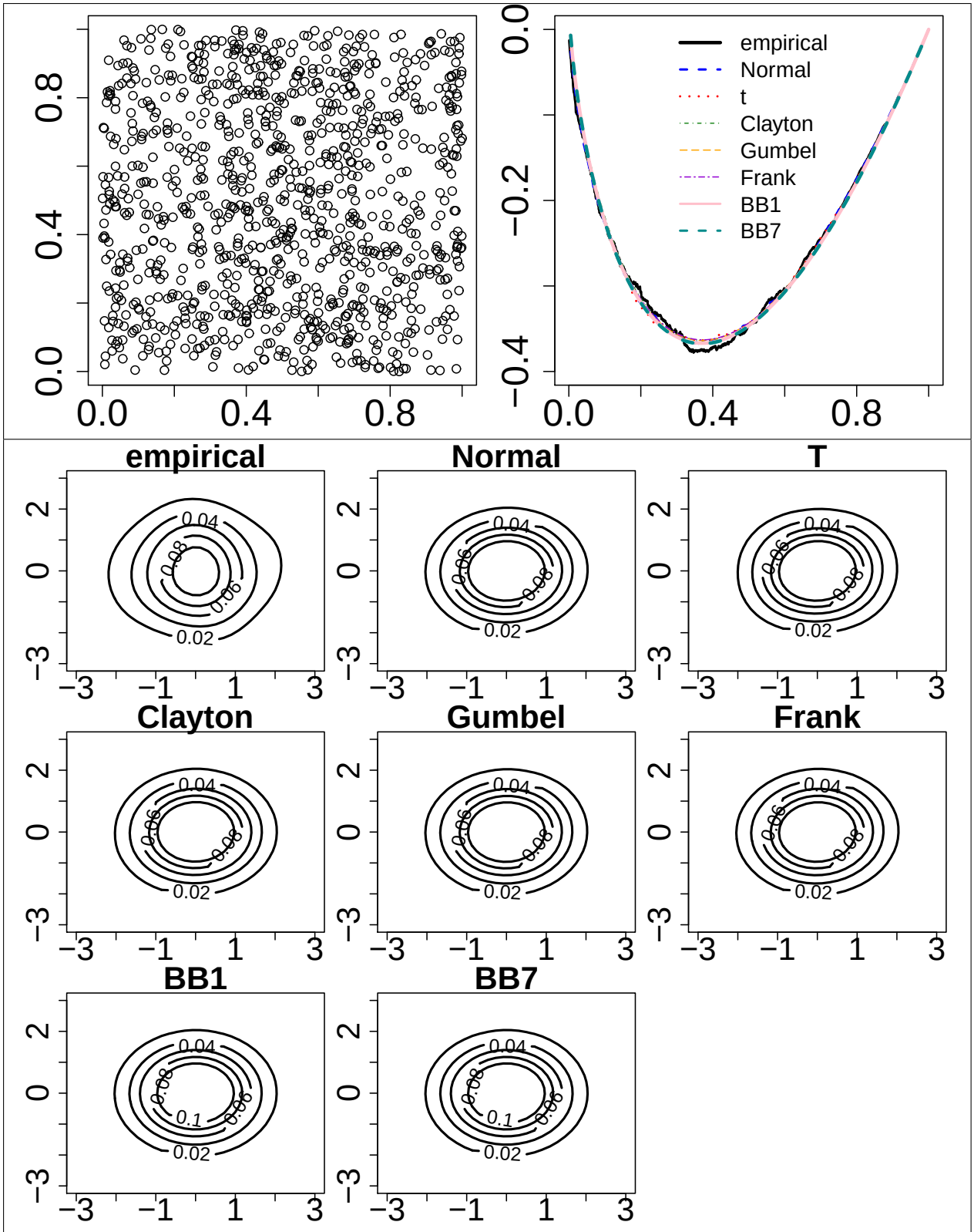
Copula $C_{INR,SGD|THB}$ ($\hat{\tau}_n = -0.09$)



P-value	N	T	C	G	F	BB1	BB7	Possible copula(s):
p^{S_n}	**0.542**	**0.33**	0.008	**0.304**	**0.628**	0	0.001	N, T, G,F
p^{T_n}	**0.295**	**0.147**	0.022	**0.21**	**0.331**	0	0.001	Selected copula: **F**

Table 7.8: Pairs plot, λ-plot, Contourplot, and tests based on Kendall's transform for the rotated data pair (THB.SGD, 1-SGD.KRW) on Tree 2

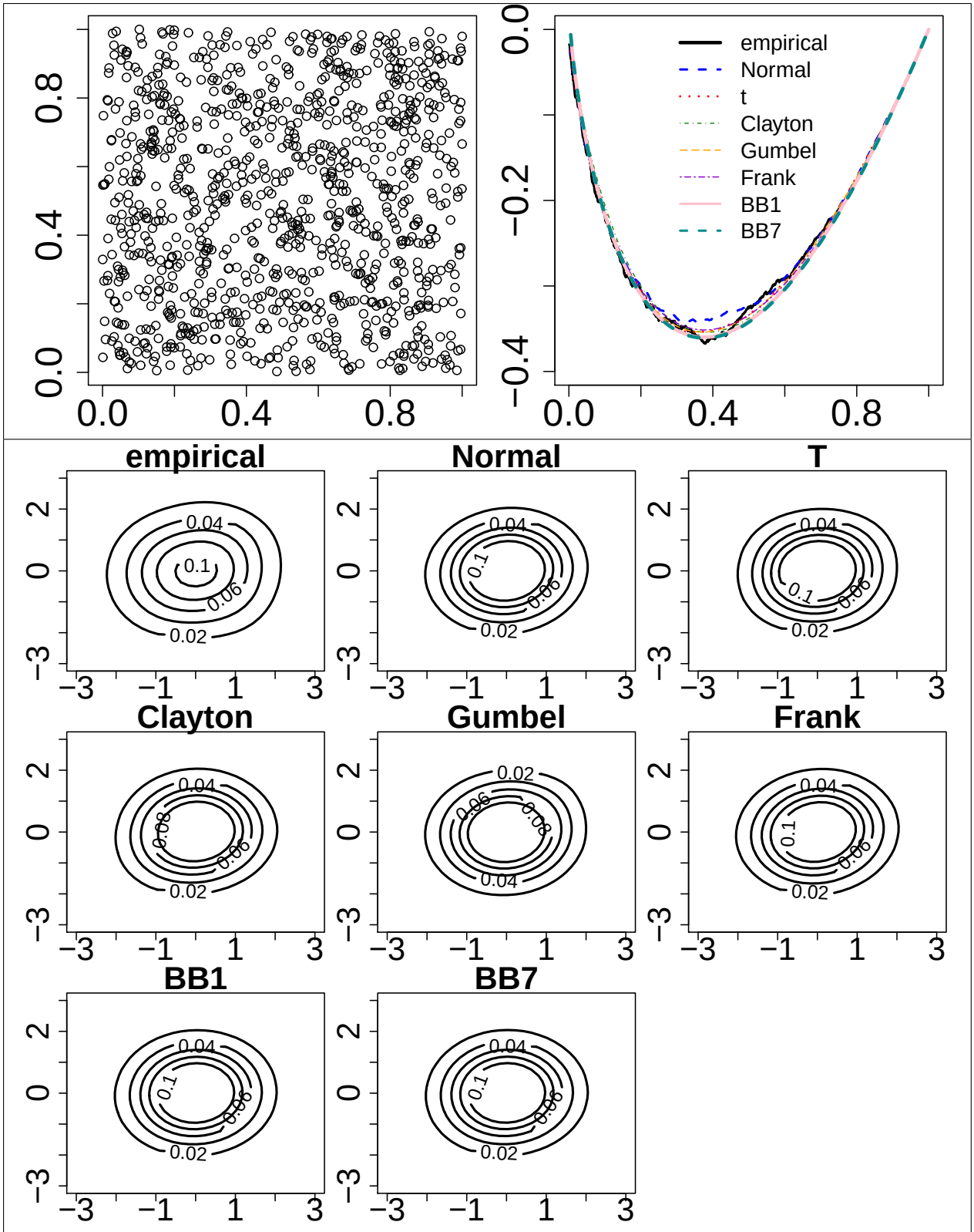
Copula $C_{THB.KRW|SGD}$ ($\hat{\tau}_n = -0.011$)



P-value	N	T	C	G	F	BB1	BB7	Possible copula(s):
p^{S_n}	**0.803**	**0.753**	**0.782**	**0.693**	**0.652**	**0.575**	**0.519**	all 7 models
p^{T_n}	**0.64**	**0.561**	**0.609**	**0.761**	**0.760**	**0.566**	**0.524**	Selected copula: **N**

Table 7.9: Pairs plot, λ-plot, Contourplot, and tests based on Kendall's transform for the data pair (SGD.KRW, KRW.JPY) on Tree 2

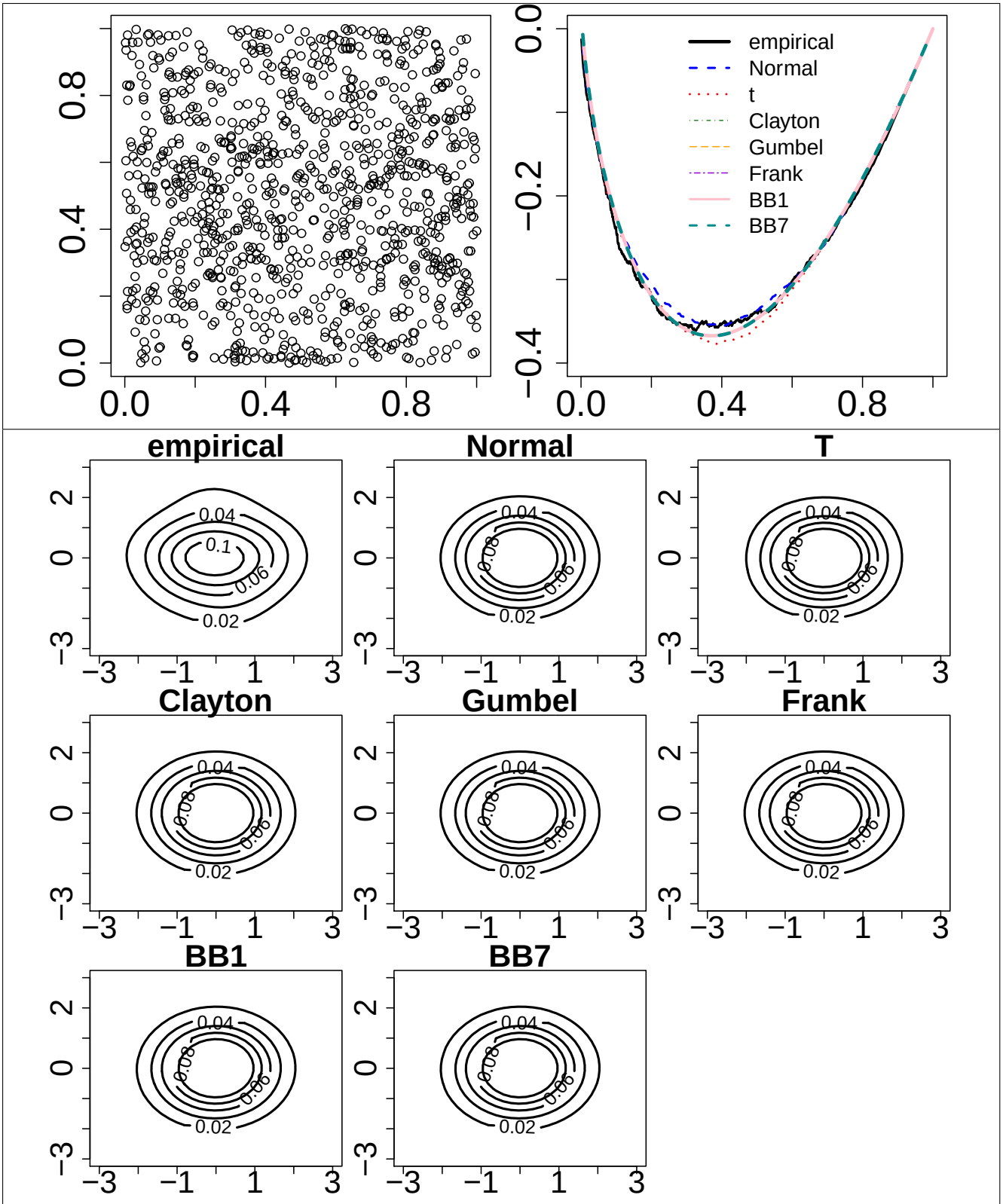
Copula $C_{SGD,JPY|KRW}$ $(\hat{\tau}_n = 0.038)$



P-value	N	T	C	G	F	BB1	BB7	Possible copula(s):
p^{S_n}	**0.697**	**0.705**	**0.508**	**0.937**	**0.891**	**0.9**	**0.64**	All models
p^{T_n}	**0.680**	**0.859**	**0.687**	**0.891**	**0.9**	**0.586**	**0.511**	Selected copula: **G**

Table 7.10: Pairs plot, λ-plot, Contourplot, and tests based on Kendall's transform for the rotated data pair (INR.SGD|THB, 1-THB.KRW|SGD) on Tree 3

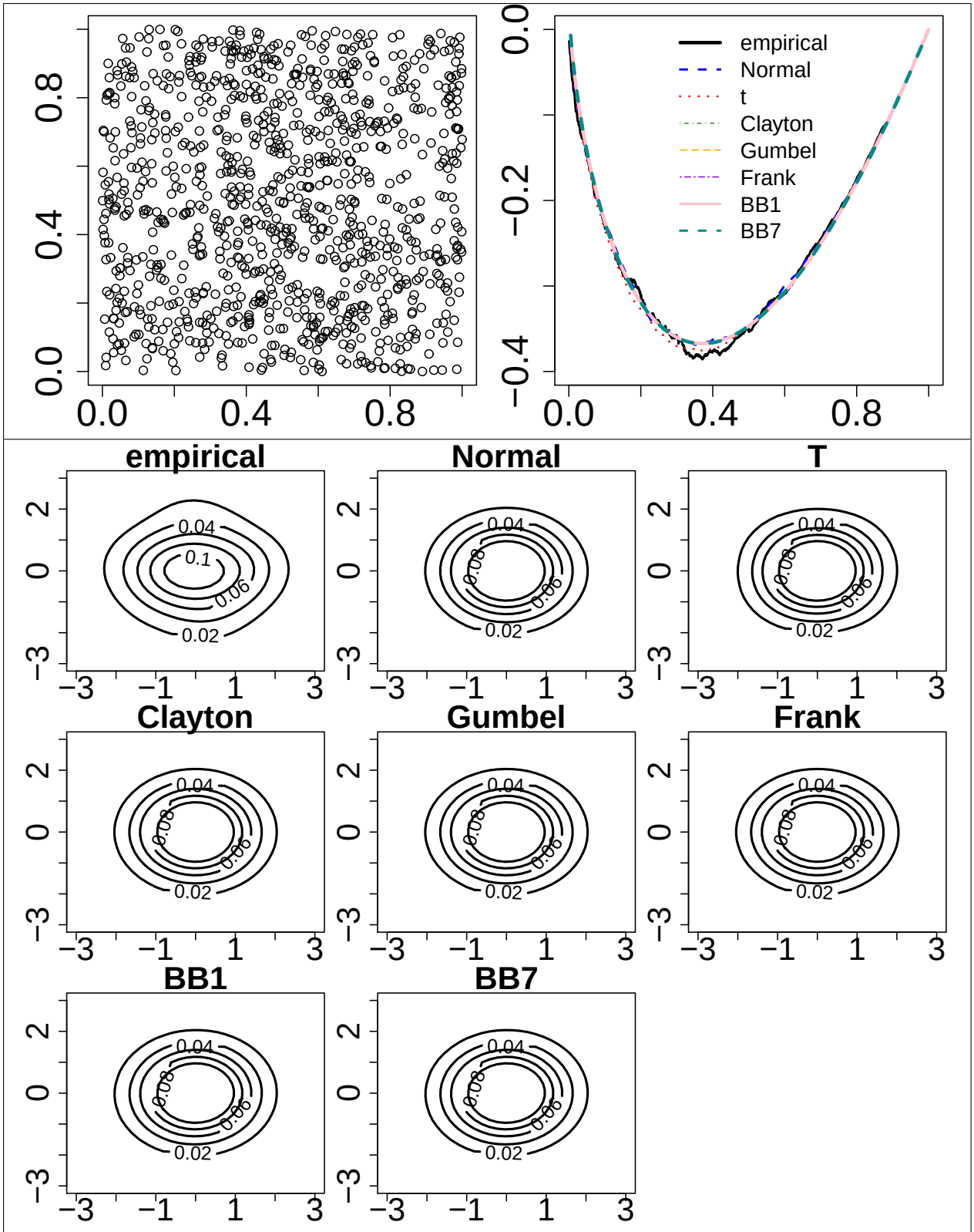
Copula $C_{INR.KRW|THB.SGD}$ ($\hat{\tau}_n = -0.002$)



P-value	N	T	C	G	F	BB1	BB7	Possible copula(s):
p^{S_n}	**0.639**	**0.884**	**0.456**	**0.476**	**0.780**	**0.904**	**0.902**	All models
p^{T_n}	**0.457**	**0.611**	**0.543**	**0.653**	**0.664**	**0.924**	**0.920**	Selected copula: **RBB1**

Table 7.11: Pairs plot, λ-plot, Contourplot, and tests based on Kendall's transform for the rotated data pair (THB.KRW|SGD, 1-SGD.JPY|KRW) on Tree 3

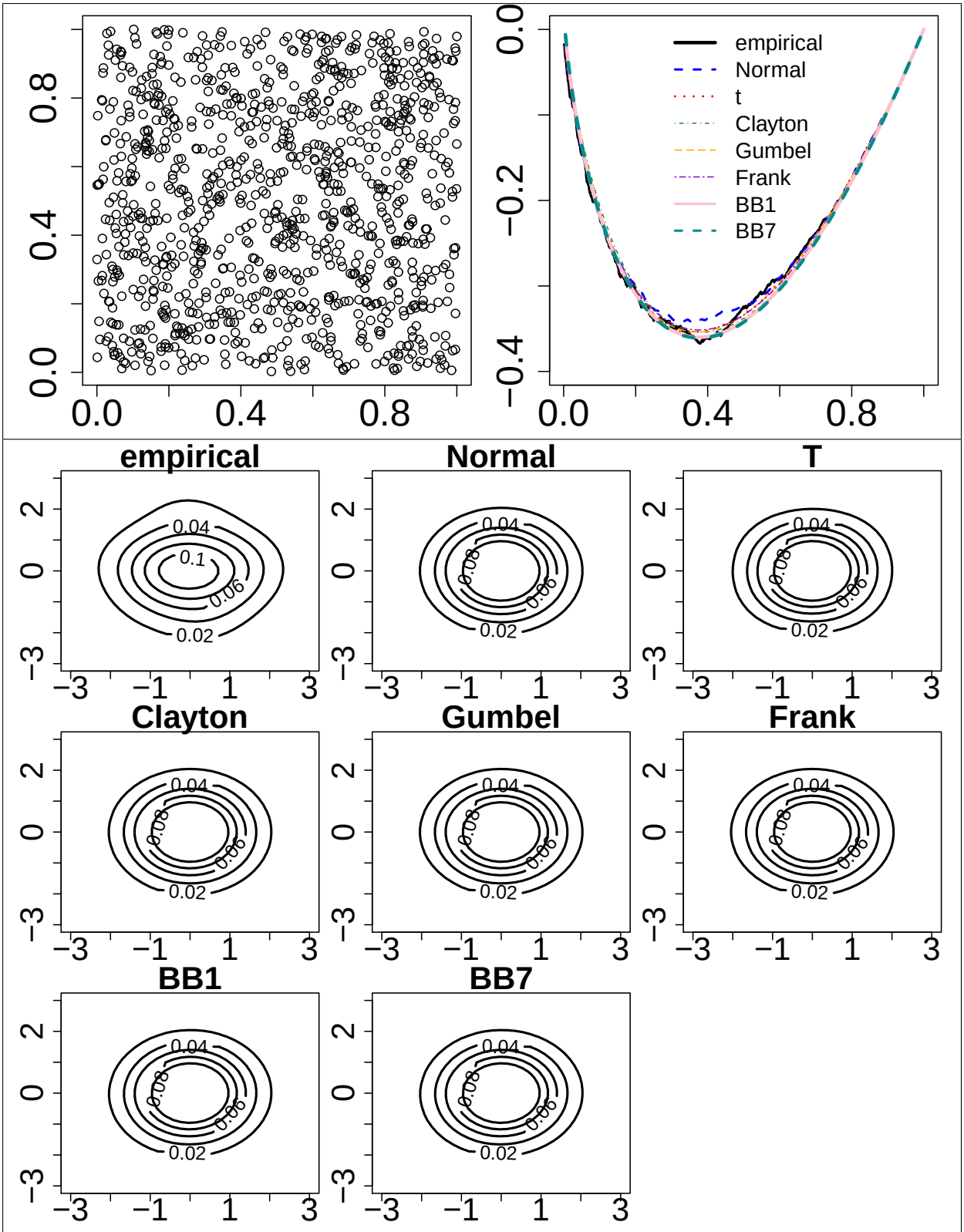
Copula $C_{THB.JPY|SGD.KRW}$ $(\hat{\tau}_n = -0.001)$



P-value	N	T	C	G	F	BB1	BB7	Possible copula(s):
p^{S_n}	**0.817**	**0.54**	**0.51**	**0.589**	**0.478**	**0.39**	**0.575**	All models
p^{T_n}	**0.902**	**0.368**	**0.434**	**0.602**	**0.39**	**0.544**	**0.552**	Selected copula: **N**

Table 7.12: Pairs plot, λ-plot, Contourplot, and tests based on Kendall's transform for the last data pair (INR.KRW|THB.SGD, THB.JPY|SGD.KRW) on Tree 4

Copula $C_{INR.JPY|THB.SGD.KRW}$ $(\hat{\tau}_n = 0)$



P-value	N	T	C	G	F	BB1	BB7	Possible copula(s):
p^{S_n}	**0.92**	**0.634**	**0.878**	**0.891**	**0.752**	**0.889**	**0.9035**	All models
p^{T_n}	**0.994**	**0.745**	**0.823**	**0.888**	**0.604**	**0.893**	**0.894**	Selected copula: **N**

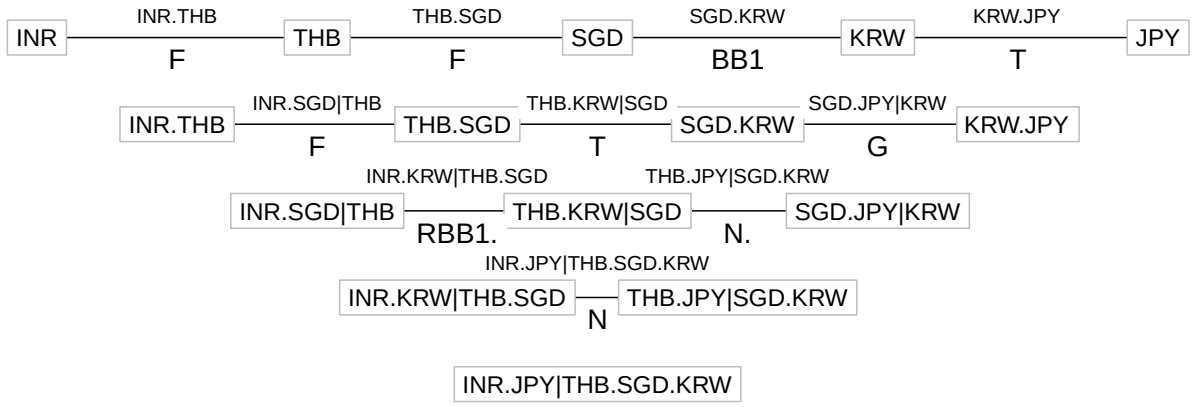


Figure 7.4: Model 1: Final model without independent copulas for Asian exchange rate data. The capital "R" in Figure stands for rotated copula.

7.3 Final model and discussion

Our final model is displayed in Figure 7.4, which displays the D-vine structure and the pair-copula models selected in the last section. This final model did not considered the independent tests based on empirical Kendall's τ and thereby the independent copulas. We call this final model without independent copulas as **Model 1**.

Note that during the model selection we used rotated data pairs if they have negative empirical Kendall's τ and selected the Frank, T and Normal copulas for these rotated data pair such as (INT.THB, 1-THB.SGD) and (THB.KRW|SGD, 1-SGD.JPY|KRW). In fact these 3 copula families are applied not only for positive but also for negative Kendall's τ. In such cases, we selected the original versions of these 3 copulas as our selected models, e.g. we selected Frank copula for Pair (INT.THB, THB.SGD) since GOF tests suggest Frank copula for (INT.THB, 1-THB.SGD). Unfortunately, The BB1 copula family dose not have this property. That is why we notated the rotated BB1 copula (RBB1) for Edge INR.SGD|THB - THB.KRW|SGD in Figure 7.4.

As can be seen in Model 1 different copula families were selected as the pair-copula models, which was other from the last application. In Section "Applications to Norwegian data" we only selected T copula as our pair-copula models basically. In general, T-copula is favored for modeling multivariate financial data. A number of studies have shown that T copula is generally a better fit of such data than other copulas. Since financial data usually exhibits symmetrical dependencies, T copula would be a appropriate choice. However, Asian financial market is basically build by developing countries and their economic situation is relatively unstable. Based on this fact and the current worldwide recession, our data set might not belong to the "standard" financial data. Therefore in the last section we observed asymmetrical or "unusual" dependencies between data pairs such as (INR, THB) and (THB, SGD), and selected Frank copula for them. Another interesting fact was that JPY depends weakly on other Asian exchange rates. Possible explanation of this is that the Japanese economy or its Export-import mainly depends on the large market such as Europe and North America.

Table 7.13: Estimated parameters in Model 1 using sequential estimation. The estimated parameters for the selected models are displayed in bold.

Tree	Pair	Par.	N	T	C	G	F	BB1	BB7
1	(INR, THB)	θ	0.229	0.229	0.335	1.172	**1.34**	0.17	1.002
		ν or δ	-	20	-	-	**-**	1.025	0.204
	(THB,SGD)	θ	0.436	0.436	0.806	1.403	**2.77**	0.253	1.13
		ν or δ	-	20	-	-	**-**	1.156	0.395
	(SGD,KRW)	θ	0.537	0.537	1.132	1.566	3.65	**0.35**	1.286
		ν or δ	-	4.243	-	-	-	**1.314**	0.591
	(KRW, 1-JPY)	θ	0.23	**0.23**	0.346	1.173	1.35	0.08	1.172
		ν or δ	-	**6.632**	-	-	-	1.14	0.152
2	(INR.THB, 1-THB.SGD)	θ	0.142	0.142	0.199	1.099	**0.81**	0.068	1.001
		ν or δ	-	20	-	-	**-**	1.009	0.775
	(THB.SGD, 1-SGD.KRW)	θ	0.018	**0.018**	0.023	1.011	0.1	0.001	1.001
		ν or δ	-	**20**	-	-	-	1.002	0.001
	(SGD.KRW, KRW.JPY)	θ	0.059	0.059	0.079	**1.039**	0.34	0.03	1.007
		ν or δ	-	20	-	**-**	-	1.002	0.037
3	(INR.SGD|THB,	θ	0.001	0.001	0.003	1.002	0.01	**0.001**	1.001
	1-THB.KRW|SGD)	ν or δ	-	20	-	-	-	**1.001**	0.001
	(THB.KRW|SGD,	θ	**0.001**	0.001	0.003	1.001	0.01	0.001	1.001
	1-SGD.JPY|KRW)	ν or δ	-	20	-	-	-	1.001	0.001
4	(INR.KRW|THB.SGD,	θ	**0.001**	0.001	0.003	1.001	0.01	0.013	1.001
	THB.JPY|SGD.KRW)	ν or δ	**-**	20	-	-	-	1.001	0.014

For reference and further analysis we list the estimated copula parameters for Model 1 in Table 7.13. Since the data pair on Tree 3 and 4 behaved independently and their empirical Kendall's τ are almost zero, the parameter estimates are near to their theoretical limits. The corresponding sequential Log-likelihood value for each selected pair-copula and the full Log-likelihood value are listed in Table 7.14

On the last two trees of Model 1, data pairs behaved independently. Therefore the choices between copula models for these pairs were equivalent. Besides the 7 tested copula families, one can also consider the independent copula, since small Kendall's τ suggests the independence. The advantage of using the independent copula is that one does not have to fit a parameter and the computation is easier and faster. Therefore we list the p-values of the independence tests based on empirical Kendall's τ in Table 7.15.

According to the results reported in Table 7.15 we can use the independent copula for the pairs (THB.SGD, 1-SGD.KRW). In this case the underlying data calculated by corresponding h-function at Node THB.KRW|SGD will be changed. Hence, all pairs in Trees 3 and 4 will be updated and their copula models might be selected differently. We illustrate an alternative final model by considering the independent copulas for these pairs and call this alternative model as **Model 2**. We summarize the scatter plots, the empirical Kendall's τ and the p-values of the independent tests for the 3 updated pairs in Table 7.16.

The scatter plots and the p-values listed in Table 7.16 suggest the independent cop-

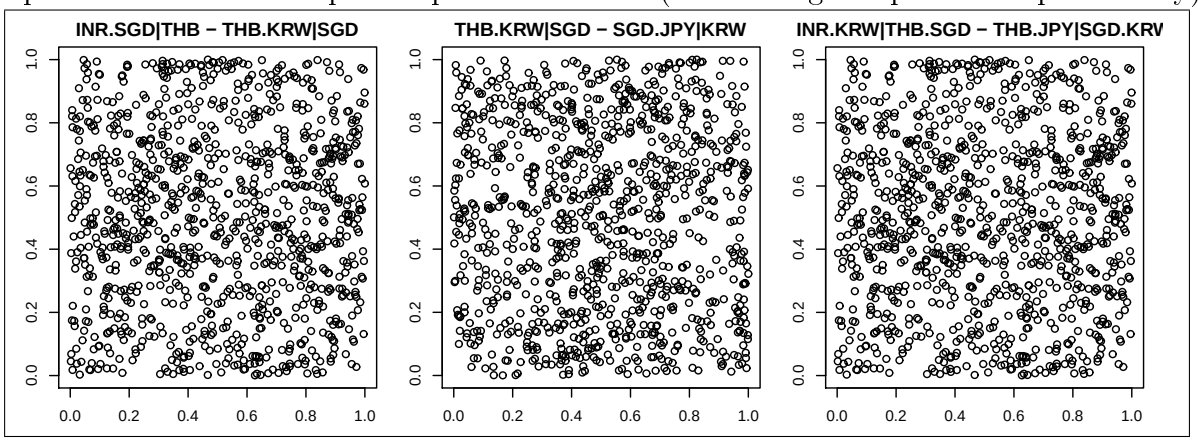
Table 7.14: The sequential Log-likelihood values for each selected pair-copula in Model 1 and the full Log-likelihood value

Tree	Pair	Selected copula	Log-likelihood
1	(INR, THB)	F	24.751
	(THB,SGD)	F	95.684
	(SGD,KRW)	BB1	172.953
	(KRW, JPY)	T	44.031
2	(INR.THB, THB.SGD)	F	19.717
	(THB.SGD, SGD.KRW)	T	-0.617
	(SGD.KRW, KRW.JPY)	G	13.359
3	(INR.SGD|THB, 1-THB.KRW|SGD)	BB1	-0.132
	(THB.KRW|SGD, SGD.JPY|KRW)	N	0.013
4	(INR.KRW|THB.SGD, THB.JPY|SGD.KRW)	N	0
Sum of Log-Likelihood values:			369.042
Full Log-likelihood values for Model 1:			371.222

Table 7.15: P-values of the independence tests based on empirical Kendall's τ for each pair in Model 1

Tree	Pair	$\hat{\tau}_n$	p-value
1	INR - THB	0.147	0
	THB - SGD	0.287	0
	SGD - KRW	0.361	0
	KRW - JPY	-0.148	0
2	INR.THB - THB.SGD	-0.09	0
	THB.SGD - SGD.KRW	-0.011	**0.594**
	SGD.KRW - KRW.JPY	0.038	**0.07**
3	INR.SGD|THB - THB.KRW|SGD	-0.002	**0.943**
	THB.KRW|SGD - SGD.JPY|KRW	-0.001	**0.975**
4	INR.KRW|THB.SGD - THB.JPY|SGD.KRW	0	**0.991**

Table 7.16: Scatter plots, empirical Kendall's τ and the p-values of the corresponding independent tests for the updated pairs in Model 2 (considering independent copula family)



Updated pair	$\hat{\tau}_n$	p-value of the ind. test	Selected copula
INR.SGD|THB - THB.KRW|SGD	-0.004	**0.859**	Ind.
THB.KRW|SGD - SGD.JPY|KRW	-0.013	**0.553**	Ind.
INR.KRW|THB.SGD - THB.JPY|SGD.KRW	-0.004	**0.859**	Ind.

ulas as same as in Table 7.15. Therefore we use independent copulas for the 2 Pairs (INR.SGD|THB, THB.KRW|SGD) and (THB.KRW|SGD, SGD.JPY|KRW in Tree 3. After updating the underlying data for the last pair, the results in Table 7.16 also suggest the independent copula. All together we use the independent copulas for all these updated pairs as well as for Pair (THB.SGD, SGD.KRW). Thus, Model 2 can be described in Figure 7.5.

Since we use the independent copulas for the updated pairs and the the rest pairs in the D-vine structure and along with their estimated parameters in Model 1 are unchanged, no further parameter estimations are needed and we can simply set the corresponding sequential Log-likelihood values for these updated pairs to 0. Table 7.17 list the sequential and full Log-likelihood values for Model 2.

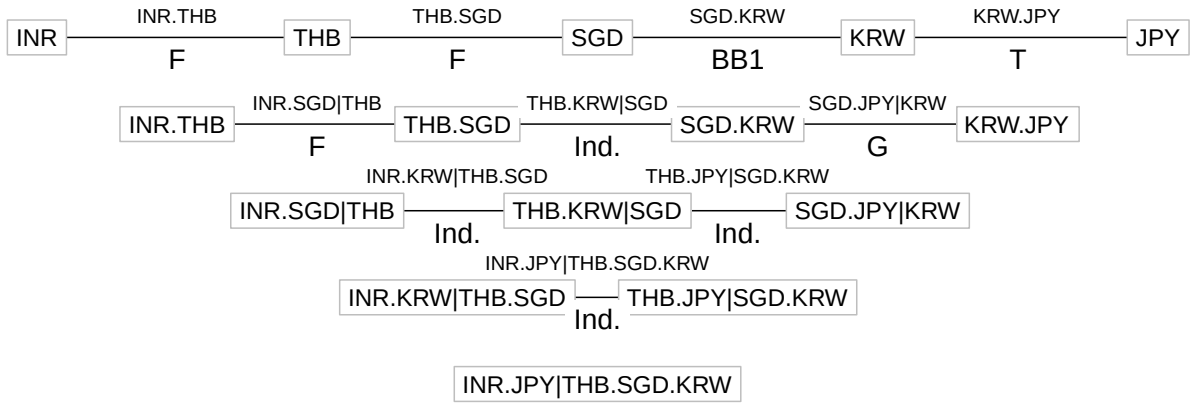


Figure 7.5: Model 2: Final model with independent copulas for Asian exchange rate data. The capital "R" in Figure stands for rotated copula.

Table 7.17: The sequential Log-likelihood values for each selected pair-copula in Model 2 and the full Log-likelihood value. * denote the updated pairs comparing to Model 1.

Tree	Pair	Selected copula	Log-likelihood	
1	(INR, THB)	F	24.751	
	(THB,SGD)	F	95.684	
	(SGD,KRW)	BB1	172.953	
	(KRW, JPY)	T	44.031	
2	(INR.THB, THB.SGD)	F	19.717	
	(THB.SGD, SGD.KRW)	Ind.	0	*
	(SGD.KRW, KRW.JPY)	G	13.359	
3	(INR.SGD|THB, 1-THB.KRW|SGD)	Ind.	0	*
	(THB.KRW|SGD, SGD.JPY|KRW)	Ind.	0	*
4	(INR.KRW|THB.SGD, THB.JPY|SGD.KRW)	Ind.	0	
Sum of Log-Likelihood values:			370.777	
Full Log-likelihood values for Model 1:			373.14	

Chapter 8

Summary and conclusions

This diploma thesis considered the measuring and modeling of complex dependencies with multivariate data and the D-vine pair-copula construction (PCC) introduced by Aas et al. [2009]. Our main object was to find appropriate exploratory data analysis (EDA) tools to determine the D-vine structure and the pair-copula families. For this purpose we first introduced some copula related measures like Kendall's τ, tail dependence and λ-function, and reviewed some copula families and their properties combining with these measures. By using these measures we developed and illustrated graphical method such as contour plots and λ-plots, as well as goodness-of-fit (GOF) tests based on bootstrap procedures to identify the pair-copulas.

Unfortunately, the theory surrounding goodness-of-fit tests for some copula family classes is still incomplete. As evidenced in Section 5.3, the consistency of the tests based on Kendall's process is limited for many copula families (See also Genest et al. [2006]). On the other hand, the number of bootstraps and data size played key roles for the variation of p-value estimates or the significance of the tests. Thus, the main emphasis here was testing the reliability of the goodness-of-fit approaches based on Kendall's process. For this investigation, we simulated this test procedure with different specified scenarios and replicated the simulations in a large number. From the analysis presented in Section 5.4, it appears that several copula families provide acceptable models in each scenario. However, this is not surprising. According to the research resulting from Kullback-Leibler information criterion (KLIC), these acceptable models are always close to each other. Therefore, it is unlikely that choosing between them would make any difference.

A second simulation study is conducted to investigate the impact of the number of bootstraps on p-values. From these two simulation studies we observed increasing power with sample size and dependence. Large data size and a large number of bootstraps are expensive for computation. Since the results from our simulation studies is satisfying, the number of bootstraps B=1000 seems to be acceptable.

In our applications the D-vine decompositions were conducted by using the developed EDA tools and two different financial data set . In almost every step of the D-vine decomposition we found several acceptable copula models according to the GOF tests, which confirmed the conclusions from the Scenarios study in Section 5.4. Consequently, different decompositions exist and lead to different final models. We illustrated an alternative final model by selecting different pair-copula models other than the first model in the

applications of Norwegian return data. Both final models fit the data set well.

Because of the closeness between copula families, we faced a situation that multiple copula families were accepted for a same data set in the scenarios study and the application. Therefore, further research is needed to find a new or second comparison method between the alternative acceptable pair-copula models and/or between alternative decompositions. One possible approach in this direction is to use the Vuong- and the Clarke-test (Vuong [1989], Clarke [2007] and Erhardt [2006]), and then examine the so-called "misspecifications". This test is already discussed and illustrated for C-vine in Schepsmeier [2010].

In the thesis we introduced a special class of copulas, rotated copulas, for modeling the data with negative dependence. However, for simplicity reasons we used the data rotation instead of the rotated copulas in the application and previous examples. Even though the rotated data worked well with the developed methods, the performance of the EDA tools using the rotated copula families can be considered for future research. Coincidentally, we tested seven copula families in this thesis. There are still a large number of copula families which also have good properties. For future investigations we could implement more families to cover a larger band of possibilities.

Many GOF tests which were not introduced in this thesis have been proposed and studied in other research. Thus, another classic problem which has existed for years is to find the best overall performing tests. Berg [2009], Genest et al. [2006] and Genest et al. [2009] did some work in this direction by doing a power comparison study between a dozen approaches. Unfortunately, no single approach can strictly dominate the others, irrespective of the circumstances.

Jiying Luo

Bibliography

Kjersti Aas, Claudia Czado, Arnoldo Frigessi, and Henrik Bakken. Pair-copula constructions of multiple dependence. *Insurance: Mathematics and Economics*, 44(2):182 – 198, 2009.

T. Bedford and R.M. Cooke. Probability density decomposition for conditionally dependent random variables modeled by vines. *Annals of Mathematics and Artificial Intelligence*, 32(1):245–268, 2001.

T. Bedford and R.M. Cooke. Vines: A new graphical model for dependent random variables. *Annals of Statistics*, 30(4):1031–1068, 2002.

D. Berg. Copula goodness-of-fit testing: an overview and power comparison. *The European Journal of Finance*, 15(7):675–701, 2009.

T. Bollerslev. Generalized autoregressive conditional heteroskedasticity. *Journal of econometrics*, 31(3):307–327, 1986. ISSN 0304-4076.

P.J. Brockwell and R.A. Davis. *Introduction to time series and forecasting*. Springer Verlag, 2002. ISBN 0387953515.

P.J. Brockwell and R.A. Davis. *Time series: theory and methods*. Springer Verlag, 2009. ISBN 1441903194.

M.R. Chernick. *Bootstrap methods: A guide for practitioners and researchers*. Wiley-Interscience, 2008.

K.A. Clarke. A simple distribution-free test for nonnested model selection. *Political Analysis*, 2007. ISSN 1047-1987.

A.C. Davison and D.V. Hinkley. *Bootstrap methods and their application*. Cambridge Univ Pr, 1997.

B. Efron, R. Tibshirani, and R.J. Tibshirani. *An introduction to the bootstrap*. Chapman & Hall/CRC, 1993.

P. Embrechts, F. Lindskog, and A. McNeil. Modelling Dependence with Copulas. 2001.

V. Erhardt. Verallgemeinerte Poisson und Nullen¡§uberschu?-Regressionsmodelle mit regressiertem Erwartungswert, Dispersions- und Nullen¡§uberschu?-Parameter und eine Anwendung zur Patentmodellierung. *Diplomarbeit*, 2006.

G. Fusai and A. Roncoroni. *Implementing models in quantitative finance: methods and cases*. Springer Verlag, 2008.

C. Genest and A.C. Favre. Everything you always wanted to know about copula modeling but were afraid to ask. *Journal of Hydrologic Engineering*, 12:347, 2007.

C. Genest and L.P. Rivest. On the multivariate probability integral transformation. *Statistics & Probability Letters*, 53:391–399, 2001.

C. Genest, B. Rémillard, and D. Beaudoin. Goodness-of-fit tests for copulas: A review and a power study. *Insurance: Mathematics and Economics*, 44(2):199–213, 2009.

Christian Genest and Bruno Rémillard. Validity of the parametric bootstrap for goodness-of-fit testing in semiparametric models. *Ann. Inst. Henri Poincaré Probab. Stat.*, 44(6): 1096–1127, 2008.

Christian Genest and Louis-Paul Rivest. Statistical inference procedures for bivariate Archimedean copulas. *J. Amer. Statist. Assoc.*, 88(423):1034–1043, 1993.

Christian Genest, Jean-François Quessy, and Bruno Rémillard. Goodness-of-fit procedures for copula models based on the probability integral transformation. *Scand. J. Statist.*, 33(2):337–366, 2006.

H. Joe. *Multivariate models and dependence concepts*. Chapman & Hall/CRC, 1997.

MG Kendall. A new measure of rank correlation. *Biometrika*, 30(1-2):81, 1938.

S. Kullback and R.A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, pages 79–86, 1951.

D. Kurowicka and R. Cooke. *Uncertainty analysis with high dimensional dependence modelling*. Wiley New York, 2006.

F. Lindskog, A. Mcneil, and U. Schmock. Kendall's tau for elliptical distributions. *Credit Risk–measurement, evaluation and management, Bol, Nakhaeizade et al., eds. Physica–Verlag, Heidelberg*, pages 149–156, 2003.

GM Ljung and GEP Box. On a measure of lack of fit in time series models. *Biometrika*, 65(2):297–303, 1978.

R.B. Nelsen. *An introduction to copulas*. Springer Us, 2006.

R.B. Nelsen, J.J. Quesada-Molina, J.A. Rodríguez-Lallena, and M. Úbeda-Flores. Distribution functions of copulas: a class of bivariate probability integral transforms. *Statistics & Probability Letters*, 54(3):277–282, 2001.

R.B. Nelsen, J.J. Quesada-Molina, J.A. Rodríguez-Lallena, and M. Úbeda-Flores. Kendall distribution functions* 1. *Statistics & Probability Letters*, 65(3):263–268, 2003.

Ulf Schepsmeier. Maximum likelihood estimation of C-vine pair-copula constructions based on bivariate copulas from different densities. *Diplomarbeit*, 2010.

A. Sklar. Fonctions de répartition à n dimensions et leurs marges. *Publ. Inst. Statist. Univ. Paris*, 8:229–231, 1959.

W. Stute, W.G. Manteiga, and M.P. Quindimil. Bootstrap based goodness-of-fit-tests. *Metrika*, 40(1):243–256, 1993.

Q.H. Vuong. Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica: Journal of the Econometric Society*, 57(2):307–333, 1989.

Weijing Wang and Martin T. Wells. Model selection and semiparametric inference for bivariate failure-time data. *J. Amer. Statist. Assoc.*, 95(449):62–76, 2000.