Advances in Dependence Modeling: Multivariate Quantiles, Copula Level Curve Lengths, and Non-Simplified Vine Copulas

Zur Erlangung des akademischen Grades eines Doktors der Wirtschaftswissenschaften (Dr. rer. pol.) bei der Fakultät für Wirtschaftswissenschaften des Karlsruher Instituts für Technologie (KIT)

genehmigte Dissertation

von
M.Sc. Maximilian Daniel Coblenz

Tag der mündlichen Prüfung: 07. Dezember 2018

Referent: Prof. Dr. Oliver Grothe

Korreferent: Prof. Dr. Rainer Dyckerhoff

Karlsruhe, Dezember 2018

Danksagung

Nach mehrjähriger Tätigkeit in der Finanzindustrie erhielt ich von Herrn Professor Dr. Oliver Grothe die Möglichkeit, durch eine Promotion meinen Lebensweg zu verändern. Für diese Chance bin ich ihm sehr dankbar. Darüber hinaus möchte ich mich bei ihm für die hervorragende Betreuung, den anregenden Gedankenaustausch, seine Hilfsbereitschaft und die stetige Unterstützung während der Promotion herzlich bedanken.

Ich möchte mich zudem bei meinem Koautor und Korreferent Herrn Professor Dr. Rainer Dyckerhoff von der Universität zu Köln für den Austausch und die hervorragende Zusammenarbeit bedanken. Zudem danke ich der Prüferin Frau Professorin Dr. Melanie Schienle und dem Prüfungsvorsitzenden Herrn Prof. Dr. Oliver Stein.

Ich danke außerdem meinen weiteren Koautoren Herrn Professor Dr. Wolfgang Trutschnig und Manuela Schreyer für die hervorragende Zusammenarbeit. Mein Dank gilt auch den Gutachtern und Editoren der Journals sowie den Teilnehmern von diversen Konferenzen und Seminaren für die interessanten Anregungen.

Im Zuge des Networking Grant des Karlsruhe House of Young Scientists wurde es mir ermöglicht, für einige Tage Frau Professorin Dr. Ingrid Hobæk-Haff an der Universität Oslo in Norwegen zu besuchen. Weiterhin ermöglichte mir das Karlsruhe House of Young Scientists über den Research Travel Grant, einen dreimonatigen Forschungsaufenthalt bei Herrn Professor Dr. Marius Hofert an der University of Waterloo in Kanada durchzuführen. Ich möchte mich daher herzlich beim Karlsruhe House of Young Scientists und meinen Gastgebern für diese einmaligen Möglichkeiten bedanken.

Diese Arbeit entstand während meiner Zeit als wissenschaftlicher Mitarbeiter am Lehrstuhl Analytics und Statistik des Karlsruher Instituts für Technologie. Ich danke meinen Zimmerkolleginnen Laura Hersing und Anika Kaplan für die tolle Zusammenarbeit, Unterstützung und die offenen Ohren. Mein Dank gilt auch der guten Seele des Lehrstuhls, Brigitte Laufert, die mir immer weiterhelfen und mich unterstützen konnte. Darüber hinaus bedanke ich mich bei den HiWis Huidi, Leon, Niklas, Sebastian und Tim für ihre Hilfe.

Ich möchte mich außerdem bei all den Mitarbeitern des KIT bedanken, die mir den Arbeitsalltag in der Universität versüßt haben.

Zuletzt bedanke ich mich bei meinen Eltern Claudia und Harald für ihre Unterstützung während der Promotion und für ihren Rückhalt in schwierigen Lebensphasen. Ohne sie hätte ich es nicht so reibungslos geschafft. Die Arbeit ist ihnen gewidmet.

Kurzfassung

Copulas ermöglichen es, die Abhängigkeitsstruktur einer multivariaten Zufallsvariable von deren univariaten Randverteilungen zu trennen. Daher stellen sie ein mächtiges Werkzeug dar, die Abhängigkeit von multivariaten Daten zu untersuchen, zu modellieren und zu interpretieren. Diese Arbeit leistet einen Beitrag zur Abhängigkeitsmodellierung, indem mehrere Aspekte von Copulas behandelt werden.

Im Speziellen werden multivariate Quantile basierend auf Copulas analysiert, weil diese von steigender Bedeutung in Bereichen wie Hydrologie und Risikomanagement sind. Die nichtparametrische Schätzung von multivariaten Quantilen wird behandelt und mit Hilfe eines Smoothed Bootstraps erweitert. Darüber hinaus wird die Schätzunsicherheit mit Hilfe von Konfidenzbändern, welche eine ganzheitliche Sicht darauf ermöglichen, bewertet.

Daran anschließend werden die Längen von Copula-Höhenlinien diskutiert. Diese liefern eine neue Sicht auf Abhängigkeiten. Mehrere Eigenschaften der Längen von Copula-Höhenlinien werden bewiesen und ein neues Zusammenhangsmaß wird eingeführt. Dieses Maß wird weiterhin auf seine Eigenschaft als Konkordanzmaß überprüft. Außerdem werden einige theoretische Resultate für spezielle Copulafamilien hergeleitet.

Schließlich werden Vine-Copulas behandelt. Diese bieten eine Möglichkeit, eine mehrdimensionale Abhängigkeitsstruktur mit Hilfe von bivariaten (bedingten) Copulas zu modellieren. Es wird eine neue Methode entwickelt, die sogenannte Simplifying Assumption zu relaxieren. Diese ermöglicht es, große Datenmengen mit Hilfe von Vine-Copulas noch besser zu modellieren. Mehrere Aspekte des eingeführten Ansatzes, wie Simulation und Schätzung, werden untersucht. Die neue Technik verspricht eine hohe Anwendbarkeit in einem breiten Spektrum an Gebieten, die sich mit Datenanalyse beschäftigen.

Die in dieser Arbeit adressierten Themen werden theoretisch und mit Hilfe von Simulationsstudien analysiert. Darüber hinaus werden sie mit zahlreichen Beispielen und Anwendungen auf Datensätzen veranschaulicht, um weitere Einsichten über ihre Anwendbarkeit und ihre Möglichkeiten zu bekommen.

Abstract

Copulas allow to decompose the dependence structure of a multivariate random variable from its univariate marginal distributions. Thus, they provide a powerful tool to investigate, to model, and to interpret the dependence in multidimensional data. This thesis contributes to dependence modeling by examining several aspects of copulas.

In particular, multivariate quantiles based on copulas are studied because these are of increasing importance in areas such as hydrology and risk management. Nonparametric estimation of multivariate quantiles is treated and extended with a smoothed bootstrap. Furthermore, the estimation uncertainty is assessed via confidence regions, which provide a holistic view on this.

Subsequently, the lengths of copula level curves are discussed. These provide a new way of looking at dependence. Several properties of copula level curve lengths are proved and a new measure of association is introduced. This measure is further examined for its properties as a concordance measure. Also, many theoretical results for specific copula families are derived.

Finally, vine copulas are treated. These provide a way to model a multidimensional dependence structure via bivariate (conditional) copulas. A new method to relax the so-called simplifying assumption is developed. This method allows to model big data with vine copulas even better. Several aspects of the introduced approach, such as simulation and estimation, are examined. The new technique promises high applicability in a wide range of areas concerned with data analysis.

The topics addressed in this thesis are studied theoretically and with the help of simulation studies. Moreover, they are illustrated with numerous examples and applications on data sets in order to provide more insight into their applicability and capabilities.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The advent of high-dimensional data in many applications and areas of research poses new challenges and calls for statistical methods that can tackle these. In this thesis, we contribute to modeling dependence in multivariate data with the help of copulas. Moreover, we advance the theory of copulas. A copula is a specific type of multivariate distribution function which allows to separate the univariate marginal distributions from the mutual dependence of the random variables involved. It can be seen as one of the most general forms of dependence modeling.

Moreover, copulas offer great flexibility in handling and modeling dependence structures of multivariate data. Therefore, over the last years, they have become an important data analysis tool in various research areas, such as hydrology (Salvadori and De Michele, 2004; Salvadori, 2004), coastal engineering (Salvadori et al., 2015, 2016), medicine (Dalla Valle et al., 2017), risk management (Cousin and Di Bernardino, 2013; Di Bernardino et al., 2013), and finance (Patton, 2006; Chollete et al., 2009). This thesis contributes to the copula literature by developing new methods and extending existing techniques. Ultimately, the aim is to provide easily applicable data analysis tools based on copulas, which can be used in a wide range of applications.

In the following, the structure and contributions of the thesis are outlined, which are composed of the abstracts of the cited works. The next chapter gives an introduction to copulas and discusses some related topics, such as concordance measures, tail dependence, and Kendall's distribution function. Furthermore, the most important concepts for empirical estimation are outlined as well as the Hausdorff distance. In passing, we introduce some general notation which is used throughout this work. Some additional notation is provided in the individual chapters to keep them mostly self-contained.

In Chapter 3, we focus on the estimation of multivariate quantiles based on copulas. We provide a nonparametric estimation procedure for a notion of multivariate quantiles, which has been introduced by Salvadori et al. (2013). These quantiles are based on particular level sets of copulas and admit the usual probabilistic interpretation that a p-quantile comprises a probability mass p. We also explore the usefulness of a smoothed bootstrap in the estimation process. Our simulation results show that the nonparametric estimation procedure yields excellent results and that the smoothed bootstrap can be beneficially applied. The main purpose is to provide an easily applicable method for practitioners and applied researchers in domains such as hydrology and coastal engineering. This chapter is based on: Coblenz, M., Dyckerhoff, R., and Grothe, O. (2018), Nonparametric estimation of multivariate quantiles, *Environmetrics*, 29: e2488 (Coblenz et al., 2018b).

Subsequently, Chapter 4 deals with reliable methods to evaluate the precision of the estimated multivariate quantile sets. Therefore, we focus on two recently developed approaches to estimate confidence regions for level sets and extend them to provide confidence regions for multivariate quantiles based on copulas. These are basically level sets at a level which has to be estimated, as well. In a simulation study, we check coverage probabilities of the employed approaches. In particular, we focus on small sample sizes. One approach shows reasonable coverage probabilities, the second one obtains mixed results. Not only the bounded copula domain but also the additional estimation of the quantile level pose some problems. A small sample application gives further insight into the employed techniques. This chapter is based on: Coblenz, M., Dyckerhoff, R., and Grothe, O. (2018), Confidence Regions for Multivariate Quantiles, *Water*, 10: 996 (Coblenz et al., 2018a).

Chapter 5 stays in the domain of copula level sets, however, deals with different aspects of these. Motivated by the well-known fact that the surface of copulas is closely related to common dependence measures such as Spearman's ρ_S, we investigate level curves of bivariate copulas and study their lengths. To this end, we establish the length profile $L_C(t)$ which maps each level $t \in [0, 1]$ to the length of the respective level curve. Some basic properties of the length profile, such as continuity and differentiability with respect to t, are examined. Based on the length profile, a measure ℓ_C is defined, which can be interpreted as the average level curve length. ℓ_C is a measure of association, it is, however, not a concordance measure in general. Additionally, some further properties, such as closed-form formulas of ℓ_C for completely dependent copulas, are demonstrated. This chapter is based on: Coblenz, M., Grothe, O., Schreyer, M., and Trutschnig, W. (2018), On the length of copula level curves, *Journal of Multivariate Analysis*, 167: 347–365 (Coblenz et al., 2018c).

In Chapter 6 we leave copula level sets aside and turn towards so-called vine copulas, which offer a way to construct a d-dimensional ($d > 2$) copula from bivariate building blocks. A new method to relax the simplifying assumption in vine copulas is presented. It is particularly suitable to explore vine structures for non-simplified parts and promises high applicability in real world examples. Moreover, the method yields a graphical representation of non-simplified vines which can be used to interpret results and to gain further insights into the non-simplified parts of the vine copula. The considered approach divides the conditioning spaces of the vine structure into disjoint sets. Each part of the resulting tessellations is governed by a copula. In contrast to existing approaches which assume a constant copula family in order to model the conditional copula through its parameter, this allows to model a conditional copula composed of multiple copula families. Several methods to estimate such a tessellation are discussed. The estimation procedure is investigated in a simulation study and compared to simplified vines. We find that in non-simplified settings a vine copula with tessellation of conditioning spaces obtains better estimation results than simplified vines. In simplified settings, the introduced method performs on a par with simplified vines. This chapter is based on the unpublished work: Coblenz, M. (2018), Non-Simplified Vine Copulas via Tessellation of Conditioning Spaces, *Working Paper* (Coblenz, 2018).

Chapter 7 concludes this thesis. All proofs are deferred to the Appendix to keep the main body of the thesis more concise and crisp. Moreover, the Appendix contains some additional topics belonging to individual chapters.

Chapter 2

Preliminaries on Copulas and Related Concepts

In this chapter, we introduce copulas and some related concepts, which we frequently use throughout. In particular, we discuss empirical estimation of copulas. Also, some dependence concepts, such as concordance measures and tail dependence, are reviewed. Finally, we introduce some further notions, such as Kendall's distribution function and the Hausdorff distance, which come up repeatedly. For a very thorough treatment of copulas, we refer the reader to the excellent books by Nelsen (2006), Joe (2015), and Durante and Sempi (2016). A practical perspective on copula modeling give Genest and Favre (2007). Parts of the following sections are close to the material presented in Coblenz et al. (2018a,b,c).

2.1 Copulas

Throughout, we use capital letters for random variables and small letters for their realizations. If not stated otherwise, we assume all random variables to be continuous. Whenever densities are needed, we assume that these exist. The terms probability density function and cumulative distribution function are abbreviated as PDF and CDF, respectively.

Let $\mathbf{X}$ denote a d-variate random variable, $d \geq 2$. Furthermore, let $F(\mathbf{x}) = \mathbb{P}(\mathbf{X} \leq \mathbf{x})$ denote the distribution function of $\mathbf{X}$, where $\leq$ is interpreted componentwise, and let $F_i(x) = \mathbb{P}(X_i \leq x)$ denote the univariate marginal distributions. We assume the marginal distribution functions F_i to be continuous. According to Sklar's theorem (Sklar, 1959), there exists a unique copula C such that

$$F(x_1, \dots, x_d) = C(F_1(x_1), \dots, F_d(x_d)). \tag{2.1}$$

Conversely, any function F, which can be represented by a copula C of univariate marginal distributions, is a distribution function. Using the transformation $(U_1,\dots,U_d) = (F_1(X_1),\dots,F_d(X_d))$, the copula function C is a multivariate distribution function of uniform univariate random variables $U_1,\dots,U_d$ and, thus, its domain is the unit hyper-cube $[0,1]^d$. Equation (2.1) shows that we can separate the marginal distributions and the overall dependence structure of $\mathbf{X}$ and that the dependence structure is solely captured by the copula. We denote the family of all d-dimensional copulas by $\mathcal{C}^d$, where we drop the exponent for bivariate copulas.

A classical definition of copulas can also be given in terms of three properties. A function C is a copula (i) if it is *grounded*, i.e., $C(u_1,\dots,u_d) = 0$ if at least one $u_i = 0$, (ii) if it has *uniform univariate marginal distributions*, i.e., $C(u_1,\dots,u_d) = u_i$ if all components of $(u_1,\dots,u_d)$ are 1 except u_i, and (iii) if it is *n-increasing*, which means that the copula volume V_C of each box $B = [u_1,v_1] \times \cdots \times [u_d,v_d]$ on $[0,1]^d$ is non-negative. That is (Nelsen, 2006; Durante and Sempi, 2016),

$$V_C = \sum_{\mathbf{z} \in \times_{i=1}^{d}\{u_i,v_i\}} (-1)^{N(\mathbf{z})} C(\mathbf{z}) \geq 0, \tag{2.2}$$

where $\times_{i=1}^{d}\{u_i,v_i\}$ is the set of vertices of B and $N(\mathbf{z}) = \#\{k : z_k = x_k\}$. For $d = 2$ this boils down to

$$V_C = C(u_1,u_2) - C(u_1,v_2) - C(v_1,u_2) + C(v_1,v_2), \tag{2.3}$$

for all $0 \leq u_1 \leq v_1 \leq 1$ and $0 \leq u_2 \leq v_2 \leq 1$ (Nelsen, 2006). Also, from these properties it can be seen that any distribution function on $[0,1]^d$ with uniform univariate marginal distributions is a copula.

In the following, we introduce some important copulas and copula families. Stochastic independence is captured by the so-called independence copula Π, which is the product of the arguments, i.e.,

$$\Pi(u_1,\dots,u_d) = u_1 \cdots u_d. \tag{2.4}$$

Perfect negative and perfect positive dependence are expressed by the lower and upper Fréchet-Hoeffding bound W and M, respectively. In two dimensions both are copulas themselves, whereas in higher dimensions only M is a copula. They are given by

$$W(u_1,\dots,u_d) = \max\left\{1 - d + \sum_{i=1}^{d} u_i, 0\right\} \tag{2.5}$$

and

$$M(u_1, \ldots, u_d) = \min\{u_1, \ldots, u_d\}. \tag{2.6}$$

It can be shown that for any copula C it holds that (Nelsen, 2006)

$$W(u_1, \ldots, u_d) \leq C(u_1, \ldots, u_d) \leq M(u_1, \ldots, u_d). \tag{2.7}$$

From $(u_1, \ldots, u_d) = (F_1(x_1), \ldots, F_d(x_d))$ and Equation (2.1), we can construct copulas by inserting the marginal quantile functions F_i^{-1}, $i = 1, \ldots, d$, into F (Nelsen, 2006), i.e.,

$$C(u_1, \ldots, u_d) = F(F_1^{-1}(u_1), \ldots, F_d^{-1}(u_d)). \tag{2.8}$$

By this the Gauss copula and t-copula can be obtained from the multivariate normal distribution and the multivariate Student t-distribution. Let Φ_d be the d-dimensional standard normal CDF and $\mathbf{R}$ be a $d \times d$ correlation matrix. Furthermore, let Φ^{-1} be the inverse of the univariate standard normal CDF. Then, the Gauss copula is given by (Joe, 2015)

$$C(u_1, \ldots, u_d) = \Phi_d(\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_d); \mathbf{R}). \tag{2.9}$$

Analogously, let $t_{d,\nu}$ be the d-dimensional Student t CDF and t_ν^{-1} be the inverse of the univariate Student t CDF, where ν is a degrees of freedom parameter. The t-copula is defined by (Joe, 2015)

$$C(u_1, \ldots, u_d) = t_{d,\nu}(t_\nu^{-1}(u_1), \ldots, t_\nu^{-1}(u_d); \mathbf{R}, \nu). \tag{2.10}$$

An important family of copulas are the so-called Archimedean copulas. They are particularly popular because they are theoretically appealing and easy to handle. Archimedean copulas follow the law

$$C(u_1, \ldots, u_d) = \psi(\psi^{-1}(u_1) + \cdots + \psi^{-1}(u_d)), \tag{2.11}$$

where ψ is the so-called generator, or generator function, and ψ^{-1} is its (pseudo)-inverse. A generator ψ is a non-increasing, continuous function mapping $[0, \infty]$ to $[0, 1]$. It satisfies $\psi(0) = 1$ and $\psi(\infty) = 0$. Additionally, it is strictly decreasing on $[0, \inf\{t : \psi(t) = 0\}]$. If $\inf\{t : \psi(t) = 0\} = \infty$, the generator is called strict (Hofert, 2008).

Three Archimedean copulas, which frequently come up in this thesis, are the Clayton, Gumbel, and Frank copulas. Let θ be the parameter of the respective copula. The generator of the Clayton copula is $\psi(u) = (1 + u)^{-\frac{1}{\theta}}$, the generator of the Gumbel copula is $\psi(u) = \exp(-u^{-\frac{1}{\theta}})$, and the generator of the Frank copula is $\psi(u) = -(\ln(e^{-u}(e^{-\theta} - 1) + 1))/\theta$. There are many more properties and types of Archimedean copulas, which are not discussed here. Comprehensive overviews can be found, e.g., in Nelsen (2006), Hofert (2008), and Joe (2015). The next section deals with empirical estimation.

2.2 Empirical Estimation

In this section, we deal with empirical estimation of copulas, where we focus on nonparametric methods. Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be an i.i.d. sample of a random vector $\mathbf{X}$ and let x_{ij} be the ith component of the vector $\mathbf{X}_j$. The copula C of $\mathbf{X}$ may be estimated based on the so called pseudo-observations $\hat{\mathbf{U}}_j = (\hat{u}_{1j}, \ldots, \hat{u}_{dj})$, $j = 1, \ldots, n$. These can be obtained either by estimation of the marginal distributions $\hat{F}_i(x)$, i.e.,

$$\hat{\mathbf{U}}_j = \left(\hat{F}_1(x_{1j}), \ldots, \hat{F}_d(x_{dj})\right), \tag{2.12}$$

or by rank transformation of the data, i.e.,

$$\hat{\mathbf{U}}_j = \frac{1}{n+1}(\text{vector of componentwise ranks of } \mathbf{X}_j \text{ in } \mathbf{X}_1, \ldots, \mathbf{X}_n). \tag{2.13}$$

Note that estimation of the marginal distributions is prone to model misspecification (Genest and Favre, 2007). Hence, a rank transformation is often preferable. Furthermore, we want to point out that, based on the pseudo-observations, the copula can be estimated parametrically using maximum likelihood (Genest and Favre, 2007).

Using the pseudo-observations $\hat{\mathbf{U}}_j$, $j = 1, \ldots, n$, the copula can be estimated nonparametrically in different fashions. The estimator $\hat{C}$ is called empirical copula and obtained by the empirical distribution of the pseudo-observations

$$\hat{C}(\mathbf{u}) = \frac{1}{n} \sum_{j=1}^{n} \mathbf{1}_{\{\hat{\mathbf{U}}_j \leq \mathbf{u}\}}. \tag{2.14}$$

A second estimator $\hat{C}_h$ that we use later on is based on kernel density estimation. It is obtained by

$$\hat{C}_h(\mathbf{u}) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{K}_h(\Phi^{-1}(\mathbf{u}) - \Phi^{-1}(\hat{\mathbf{U}}_i)), \tag{2.15}$$

where $\mathbf{K}_h(\mathbf{x}) = \mathbf{K}(\mathbf{x}/h)$ is the scaled version of a suitable multivariate kernel $\mathbf{K}$ and Φ^{-1} is the inverse univariate standard normal CDF applied componentwise. Using a multiplicative kernel, this estimator is investigated in Omelka et al. (2009). The transformation Φ^{-1} circumvents potential boundary issues that can arise in the copula domain $[0,1]^d$. It is also recommended in Joe (2015). Apart from choosing a kernel, the estimator also requires a bandwidth parameter h. In this thesis, we choose $\mathbf{K}_h$ to be a multiplicative multivariate Gaussian kernel and $h = \left(\frac{4}{d+2}n\right)^{\frac{1}{d+4}}$, which is Silverman's rule of thumb (Silverman, 1986). As will become clear in Chapter 4, these choices are particularly easy to work with. In the next section, we deal with a univariate distribution function, which is important for multivariate quantiles (cf. Chapters 3 and 4).

2.3 Kendall's Distribution Function

A concept we frequently employ when dealing with multivariate quantiles (cf. Chapters 3 and 4) is Kendall's distribution function $K_C : [0,1] \longmapsto [0,1]$ (Barbe et al., 1996; Genest and Rivest, 2001; Nelsen et al., 2003) given by

$$K_C(p) = \mathbb{P}(C(U_1, \dots, U_d) \leq p), \tag{2.16}$$

where $p \in [0,1]$ is a probability level and the $U_1, \dots, U_d$ are distributed according to Copula C. $K_C(p)$ gives the probability that for a realization $u_1, \dots, u_d$ of $U_1, \dots, U_d$ it holds that $C(u_1, \dots, u_d) \leq p$.

According to Barbe et al. (1996), the Kendall distribution function can be estimated nonparametrically from a sample of size n by

$$\hat{K}_C(p) = \frac{1}{n} \sum_{j=1}^{n} \mathbf{1}_{\{V_{j,n} \leq p\}}, \tag{2.17}$$

where $V_{j,n} = \frac{\#\{k \neq j | \mathbf{X}_k \leq \mathbf{X}_j\}}{n-1}$. The inverse of $K_C(p)$ is estimated using

$$\hat{K}_C^{-1}(p) = \inf\{t | \hat{K}_C(t) \geq p\} \tag{2.18}$$

for $0 < p < 1$ and $\hat{K}_C^{-1}(0) = 0$ and $\hat{K}_C^{-1}(1) = 1$. One can show that plim $\hat{\mathrm{K}}_{\mathrm{C}}(\mathrm{p}) = \mathrm{K}_{\mathrm{C}}(\mathrm{p})$ for $p \in [0,1]$ (Barbe et al., 1996) and that plim $\hat{\mathrm{K}}_{\mathrm{C}}^{-1}(\mathrm{p}) = \mathrm{K}_{\mathrm{C}}^{-1}(\mathrm{p})$ for $p < 1$ (Serfling, 1980), where plim denotes convergence in probability. Also, the empirical Kendall distribution is strongly consistent for its population counterpart (Ghoudi and Rémillard, 1998). Some important notions of dependence are reviewed in the next section.

2.4 Dependence Concepts

Since copulas contain the complete dependence structure of a multivariate random variable, they are linked to many different dependence concepts and measures. Copulas are deeply connected to ranks (Nelsen, 2006) and, thus, to two important rank correlation measures: Kendall's τ and Spearman's ρ_S. These are also measures of concordance and indicate the overall strength of dependence. For a proper definition of concordance measures we refer the reader to Section 5.5.2.

In the following, let $C \in \mathcal{C}$ be a bivariate copula. Then, Kendall's τ can be computed as (Nelsen, 2006)

$$\begin{aligned} \tau &= 4E[C(U_1, U_2)] - 1 && (2.19)\\ &= 4\int_0^1 \int_0^1 C(u_1, u_2) dC(u_1, u_2) - 1. && (2.20) \end{aligned}$$

Furthermore, Spearman's ρ_S can be calculated as (Nelsen, 2006)

$$\begin{aligned} \rho_S &= 12\int_0^1 \int_0^1 (C(u_1, u_2) - u_1u_2) du_1 du_2 && (2.21)\\ &= 12\int_0^1 \int_0^1 u_1u_2 dC(u_1, u_2) - 3 && (2.22)\\ &= 12\int_0^1 \int_0^1 C(u_1, u_2) du_1 du_2 - 3. && (2.23) \end{aligned}$$

From the first line we can see that Spearman's ρ_S can be expressed as the difference of the volumes below the graphs of copula C and the independence copula. This interpretation is further examined in Chapter 5, where we look at the length of copula level curves.

Another important concept, which we encounter later on, is lower and upper tail dependence. Lower tail dependence measures the strength of dependence of two random variables close to the point (0,0) in the unit square, whereas upper tail dependence measures the strength of dependence close to the point (1,1) in the unit square. Lower tail dependence λ_L and upper tail dependence λ_U can be calculated as (Nelsen, 2006)

$$\lambda_L = \lim_{t \to 0^+} \frac{C(t,t)}{t} \tag{2.24}$$

and

$$\lambda_U = 2 - \lim_{t \to 1^-} \frac{1 - C(t,t)}{1 - t}. \tag{2.25}$$

We want to point out that there are d-dimensional ($d > 2$) extensions for the concepts introduced above, see, e.g., Schmid and Schmidt (2007), Joe et al. (2010), and Joe (2015). However, these are not straightforward to establish. Furthermore, there is a plethora of other dependence concepts, such as positive quadrant dependence or stochastically increasing positive dependence (Joe, 2015). However, these topics are outside the scope of this thesis and, thus, are not discussed further here. In the last section, we introduce the Hausdorff distance, which we repeatedly use when dealing with level sets of copulas.

2.5 Hausdorff Distance

In order to measure the distance between copula level sets on the one hand and for consistency results related to level sets on the other hand, we need the Hausdorff distance between two non-empty sets $A, B \subset \mathbb{R}^d$, which we will denote by $\delta_H(A, B)$ (Rockafellar and Wets, 1998). This is especially important when we deal with multivariate quantiles, see Chapter 3 and Chapter 4. Let $\delta(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|$ be the Euclidean distance between two points $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$. We define the distance between a point $\mathbf{x} \in \mathbb{R}^d$ and a set $A \subset \mathbb{R}^d$ as $\delta(\mathbf{x}, A) = \inf_{\mathbf{y} \in A} \delta(\mathbf{x}, \mathbf{y})$, where we can use min instead of inf if A is closed. The Hausdorff distance δ_H can then be defined as follows.

Definition 1 (Hausdorff Distance). *For non-empty subsets $A, B \subset \mathbb{R}^d$ the* Hausdorff distance $\delta_H(A, B)$ *is defined by*

$$\delta_H(A, B) = \max\{\sup_{\boldsymbol{x} \in A} \delta(\boldsymbol{x}, B), \sup_{\boldsymbol{x} \in B} \delta(\boldsymbol{x}, A)\}. \tag{2.26}$$

In general, the Hausdorff distance may be infinite. However, since we consider only subsets of the compact set $[0, 1]^d$ in this thesis, the Hausdorff distance is always finite. The Hausdorff distance is *not* a metric, but only a pseudometric, i.e., $\delta_H(A, B) = 0$ does in general not imply that $A = B$. In fact $\delta_H(A, B) = 0$ if and only if $clA = clB$, where cl denotes the closure of a set.

It is possible to define convergence of sets in the Hausdorff distance, or for short *Hausdorff convergence*, also see, e.g., Dyckerhoff (2017) for more details on this topic.

Definition 2 (Hausdorff Convergence). *Let $(A_n)_{n \in \mathbb{N}}$ be a sequence of non-empty subsets of $\mathbb{R}^d$. The sequence $(A_n)_{n \in \mathbb{N}}$ is said to be* Hausdorff convergent *to a non-empty set A, if $\lim_{n \to \infty} \delta_H(A_n, A) = 0$.*

However, it should be noted that, since δ_H is only a pseudometric, the Hausdorff limit is not unique. In fact, if a sequence (A_n) of sets is convergent to a limit A, then every set B for which $clA = clB$ is also a limit of the sequence.

In the next chapters, we put the concepts introduced here to use. We begin by estimating multivariate quantiles based on copulas.

Chapter 3

Nonparametric Estimation of Multivariate Quantiles

This chapter is based on Coblenz et al. (2018b). Note that the figures in this chapter are under the copyright of John Wiley & Sons, Ltd. A permission to reuse them in this work is granted by the copyright holder.

3.1 Introduction

It is important to assess and quantify risk in complex environments. A statistical approach to do so is using quantiles. They provide an easy way to measure extreme events and their corresponding probabilities. Up to now, quantiles are largely used in a univariate setting. However, the increase in complexity and the availability of larger data sets has motivated researchers to transfer the concept into dimensions higher than one and to propose different definitions of multivariate quantiles, see, e.g., Serfling (2002), Di Bernardino et al. (2013), and Salvadori et al. (2013). Multivariate quantiles are especially useful when it is not possible or reasonable to combine the random variables involved into one single variable of interest, i.e., when no univariate analysis on the combined variables is possible. This is particularly the case in application domains such as hydrology, where, e.g., flood peak and flood volume may not be meaningfully combined into one variable of interest.

Historically, hydrology and coastal engineering are areas where first examples of multivariate quantile approaches gained attention in applications and could provide a more realistic picture than univariate approaches, see Yue and Rasmussen (2002), Salvadori (2004), and Salvadori and De Michele (2004) for early references. Additionally, international guidelines in these fields led researchers to consider multivariate approaches more closely (Salvadori et al., 2016). For example, Chebana and Ouarda (2011) use multivariate quantiles in a bivariate setting of frequency analysis of floods. They model dependencies of flood volume and flood peak via a copula approach and analyze the resulting combinations for a given risk level. Requena et al. (2013) use multivariate quantiles based on copulas in hydrologic dam design. In Salvadori et al. (2015) copula-based multivariate quantiles are used to measure the probability of structural failure in coastal and offshore engineering, measured by return period and design quantile. Recently, multivariate quantiles are also present in applications of financial risk management. There, they lead to multivariate extensions of risk measures such as value at risk and expected shortfall. References in this field using a copula approach are, e.g., Cousin and Di Bernardino (2013) and Di Bernardino and Prieur (2014).

In general, the definition of multivariate quantiles is not a trivial task, since $\mathbb{R}^d$, $d > 1$, has no total ordering. For this reason, various attempts to define multivariate quantiles are present in the literature; see, e.g., Serfling (2002) for an overview. The notion of multivariate quantiles we use in this thesis is linked to copulas. A multivariate quantile set for given level p is defined as the level set of a copula whose probability mass sums up to p (cf. Definition 4). In our opinion, this is the most natural way to define multivariate quantiles. Equivalent definitions are also used by Salvadori et al. (2013) and Salvadori et al. (2014). It differs from approaches used by, e.g., Di Bernardino and Prieur (2014), where the p-th quantile is defined to be the p-level set of the copula (and not the level set comprising probability mass p). The quantiles of the latter case do not allow for the interpretation that a proportion of $p \cdot 100\%$ of the data lies below the quantile. The definition used in this chapter preserves such an interpretation, but is empirically more involved, since not only level curves have to be estimated, but also the probability mass below the curves. Therefore, one needs estimates of the inverse of the Kendall distribution function (cf. Section 2.3).

The contributions of this chapter are as follows: We give theoretical validity to the considered quantile definition by showing that a copula function is uniquely determined by the multivariate quantile sets. Furthermore, we introduce a nonparametric estimation procedure and establish consistency of the estimators. These results are new in the literature on the notion of multivariate quantiles we use here. To further improve the method's accuracy (in particular in small sample sizes), we suggest a smoothed bootstrap procedure. The procedure extends the original sample with additional points and improves the accuracy of the estimated quantiles. A simulation study investigates the finite sample performance of the estimator for smoothed and original data in detail. Finally, the main purpose of the chapter is to endow practitioners and researchers with an easy to understand, flexible and easy to implement, yet powerful, estimation procedure for multivariate quantiles. We therefore illustrate our method on a flood peak and flood volume data set comprising 33 observations, where we greatly benefit from the smoothed bootstrap approach.

The chapter is organized as follows: In the next section, our definition of multivariate quantiles is motivated and discussed. Section 3.3 introduces the nonparametric estimation procedure, establishes the convergence results, and shows how a smoothed bootstrap can be beneficially incorporated in the estimation procedure. A simulation study is conducted in Section 3.4 in order to address questions of accuracy and variation of the approach in finite samples. The chapter concludes with an application example on hydrology data in Section 3.5 and some final remarks in the last section. All proofs can be found in the accompanying paper Coblenz et al. (2018b).

3.2 Multivariate Quantile Sets

In this section, we discuss the notion of multivariate quantiles we will use throughout the chapter. We also give some accompanying theorems ensuring uniqueness of the quantile sets with respect to the considered dependence structure. This provides theoretical validity that the considered sets may be interpreted as quantiles.

Since $\mathbb{R}^d$, $d > 1$, has no total ordering, there are several notions of multivariate quantile sets. Serfling (2002) reviews different approaches. Other works on multivariate quantiles include Tibiletti (1993) and Chaudhuri (1996). Extensive work on the topic is also done by Cousin and Di Bernardino (2013), Di Bernardino et al. (2013) and Di Bernardino and Rullière (2013), who use multivariate quantiles in financial risk management.

To motivate our definition of multivariate quantiles (Definition 4 below), let us first reconsider a univariate quantile

$$x_p = F_X^{-1}(p) \tag{3.1}$$

for the simplest case of a continuous random variable X with strictly increasing distribution function F_X. Univariate quantiles exhibit two features. First, the position of x_p is uniquely determined by the level p of the distribution function F_X. Second, we have $\mathbb{P}(X \in [-\infty, x_p]) = p$, i.e., the level set of the univariate distribution F_X at level p comprises a probability mass of p, which is the probabilistic interpretation one would expect of a quantile. Note, that the second feature is more crucial than the first when interpreting quantiles.

As a direct extension of the first feature in the univariate case, we can focus on the level sets at level p defined by a multivariate distribution function F, or equivalently copula C, of a random vector.

Definition 3 (Level Sets)**.** *Let $\boldsymbol{X}$ denote a d-variate random variable, $d \geq 2$, with distribution function F and corresponding copula C. Then,*

$$S'_p(F) = \{\boldsymbol{x} \in \mathbb{R}^d : F(\boldsymbol{x}) \leq p\} \tag{3.2}$$

and

$$S'_p(C) = \{\boldsymbol{u} \in [0,1]^d : C(\boldsymbol{u}) \leq p\}, \tag{3.3}$$

for $p \in [0,1]$, are called level sets of distribution function F and copula C, respectively.

The sets $S'_p(C)$ are considered as multivariate quantiles in a series of papers, cf. Di Bernardino et al. (2013) and Di Bernardino and Rullière (2013).

We use $S'_p(C)$ and $S'_p(F)$ as the starting points on the way to our notion of multivariate quantiles and provide some theoretical justification. The next theorem shows that a copula C and multivariate distribution F are uniquely determined by their respective level sets $S'_p(C)$ and $S'_p(F)$.

Theorem 1. *Let F be a distribution with continuous marginal distribution functions and C its corresponding copula. Then,*

(i) the level sets $(S'_p(C))_{p\in[0,1]}$ uniquely characterize the copula C;

(ii) the level sets $(S'_p(F))_{p\in[0,1]}$ uniquely characterize the distribution function F.

The proofs can be found in Coblenz et al. (2018b), Appendix A. We also refer the reader to Trutschnig (2012) for some further results on copula level sets.

To motivate our notion of multivariate quantiles further, focus on $S'_p(C)$ and note that the level sets lack the usual probabilistic interpretation expected from quantiles:

$$\mathbb{P}(\mathbf{U} \in S'_{p'}(C)) \neq p'. \tag{3.4}$$

Hence, the probability for a random vector to fall into the p'-th level set is not p' and the second feature of the univariate case above is not resembled. To facilitate such an interpretation, we relabel the sets by their enclosed probability mass. To this end, recall that Kendall's distribution function $K_C(p)$ yields the probability that $C(U_1, U_2, \ldots, U_d)$ stays below or is equal to level p, see Section 2.3. Thus, in general for $d \geq 2$ we have

$$\mathbb{P}(\mathbf{U} \in S'_p(C)) = \mathbb{P}(C(U_1, \ldots, U_d) \leq p) = K_C(p) \geq p. \tag{3.5}$$

In the following, consider the class $\mathcal{C}_K^d$ of copulas for which Kendall's distribution function K_C is strictly increasing and continuous, and with $\mathcal{F}_K$ the class of distributions for which the copula is in $\mathcal{C}_K^d$. For distributions in $\mathcal{F}_K$ (or copulas in $\mathcal{C}_K^d$) the inverse $K_C^{-1} : [0,1] \longmapsto [0,1]$ exists and is also strictly increasing and continuous. We can now define the notion of multivariate quantile sets, which we use throughout the rest of the chapter (see also the motivations in Salvadori et al. (2013) and Salvadori et al. (2014)):

Definition 4 (Multivariate Quantile)**.** *Let $q_p = K_C^{-1}(p)$. For a copula $C \in \mathcal{C}_K^d$ and $p \in [0,1]$ a multivariate quantile set is defined as*

$$S_p(C) := S'_{q_p}(C) = \{\boldsymbol{u} \in [0,1]^d : C(\boldsymbol{u}) \leq q_p\}. \tag{3.6}$$

Analogously, we can define the quantile sets in terms of the distribution function $F(\boldsymbol{x})$ and denote them by $S_p(F)$,

$$S_p(F) := S'_{q_p}(F) = \{\boldsymbol{x} \in [0,1]^d : F(\boldsymbol{x}) \leq q_p\}. \tag{3.7}$$

For continuous distribution functions, the bijection between $S_p(C)$ and $S_p(F)$ is defined by the marginal distribution functions F_j, $j = 1, \ldots, d$. Namely, if $(X_1, X_2, \ldots, X_d) \in S_p(F)$ then $(F_1(X_1), F_2(X_2), \ldots, F_d(X_d)) \in S_p(C)$ and vice versa.

Just like in the case of the level sets $S'_p(C)$ and $S'_p(F)$, a copula C and multivariate distribution F are uniquely determined by their multivariate quantile sets $S_p(C)$ and $S_p(F)$, respectively. This is shown in the next theorem, which is proved in Coblenz et al. (2018b), Appendix A.

Theorem 2. *Let C be a copula in class $\mathcal{C}_K^d$ and F be a distribution in class $\mathcal{F}_K$. Then,*

(i) the quantile sets $(S_p(C))_{p\in[0,1]}$ uniquely characterize the copula C;

(ii) the quantile sets $(S_p(F))_{p\in[0,1]}$ uniquely characterize the distribution F.

Some remarks on and properties of the defined multivariate quantile sets:

1. We have $(S_p(C)) \subset (S_{p'}(C))$ for $p < p'$ and $\mathbb{P}\left(\bigcup_{p\in[0,1]} S_p(C)\right) = 1$ as well as $\mathbb{P}\left(\bigcap_{p\in[0,1]} S_p(C)\right) = 0$. Furthermore, $\mathbb{P}(S_p(C)) = p$. Similar properties hold for $(S_p(F))_{p\in[0,1]}$.

2. In the bivariate case the "direction" of the quantile sets is from "bottom left" to "upper right" in the same sense as univariate quantiles measure the probability mass on their left hand side. In general for $d \geq 2$, the quantile sets are oriented along the diagonal of the d-dimensional unit hypercube from $\mathbf{0}^d$ to $\mathbf{1}^d$. If, e.g., in a financial context, $\mathbf{X} = (X_1, \ldots, X_d)$ denote market returns of assets, then $S_p(F)$ might be called a multivariate extension of value at risk at level p and $S_p(C)$ is the corresponding copula version. See also the work of Cousin and Di Bernardino (2013) on this topic.

3. Similarly, one can define the quantile sets above based on the survival function $\bar{F}(\mathbf{x}) = P(X_1 > x_1, \ldots, X_d > x_d)$ and the corresponding survival copula $\bar{C}$. The orientation of these is again along the diagonal of the d-dimensional hypercube, however from $\mathbf{1}^d$ to $\mathbf{0}^d$ in this case. The details of such a definition are completely analogous to the ones above. This definition coincides with case 4 in Salvadori et al. (2016). Note that when using rank transformations the survival function approach and the approach discussed above can be made equivalent if data values are multiplied by -1.

4. One might also be interested in the quantile sets

$$\mathbb{P}\left(C(U_1, \ldots, U_d) > K_C^{-1}(p)\right) = 1 - \mathbb{P}(S_p(C)) = 1 - p, \tag{3.8}$$

   which coincides with case 3 in Salvadori et al. (2016) and Definition 5 in Chapter 4. Treatment of these quantile sets is completely analogous to the one for $S_p(C)$ by noting that both are complementary sets in the unit hypercube.

Since for continuous distributions $(S_p(F_X))_{p\in[0,1]}$ and $(S_p(C_F))_{p\in[0,1]}$ are connected via the marginal distributions, we can choose which one to use in an application. However, in contrast to a general distribution function F, the domain of every copula C is the unit hypercube and thus bounded. This makes it easier and more tractable to work with multivariate quantiles based on copulas, which is particularly relevant when it comes to graphical presentation of results. Thus, in the following we will focus on the quantile sets $S_p(C)$ related to a copula C.

3.3 Nonparametric Estimation Procedure

In this section we treat nonparametric estimation of the quantile sets $S_p(C)$ as defined above. We establish convergence results and discuss computation of the boundaries of the quantile sets in practice. Additionally, to improve smoothness and accuracy of the estimated level set, we explore how to incorporate a smoothed bootstrap into the procedure.

3.3.1 Consistency of Proposed Estimators

Let $\mathbf{X}$ be a d-variate random vector and let $\mathbf{X}_1, \mathbf{X}_2, \cdots \overset{iid}{\sim} \mathbf{X}$, all defined on a common probability space $(\Omega, \mathcal{A}, P)$. Consider an estimator $\hat{C}_{\mathbf{X}_1,\ldots,\mathbf{X}_n}$ (or short $\hat{C}_n$) of the copula C of $\mathbf{X}$. Note that in this section we do not assume this estimator to be the empirical copula in Equation (2.14).

In a first step, we are interested in the convergence of the level sets

$$\hat{S}'_{n,p} := \hat{S}'_p(\hat{C}_n) = \{\mathbf{u} \in \mathbb{R}^d | \hat{C}_n(\mathbf{u}) \leq p\}, \tag{3.9}$$

to their population counterpart $S'_p := S'_p(C)$. For convergence of the level sets, we consider the notion of convergence in the Hausdorff distance. The definition as well as some important properties of the Hausdorff distance are given in Section 2.5. The following theorem states that the empirical level sets $\hat{S}'_{n,p}$ are strongly uniform consistent estimators for S'_p.

Theorem 3. *Assume that $\hat{C}_n$ is strongly uniform consistent, i.e.,*

$$\|\hat{C}_n - C\|_\infty = \sup_{\boldsymbol{u}\in[0,1]^d} |\hat{C}_n(\boldsymbol{u}) - C(\boldsymbol{u})| \xrightarrow{a.s.} 0\,. \tag{3.10}$$

Assume further that the copula C is strictly increasing, i.e, $x_i < y_i$, $i = 1, \ldots, n$, implies $C(x_1, \ldots, x_n) < C(y_1, \ldots, y_n)$.

Then, almost surely,

$$\lim_{n\to\infty} \sup_{p\in[0,1]} \delta_H(S'_p, \hat{S}'_{n,p}) = 0\,, \tag{3.11}$$

i.e., the level sets converge uniformly to their population counterparts.

The proof can be found in Coblenz et al. (2018b), Appendix A.

We now consider the multivariate quantile sets

$$S_p(C) = S'_{q_p}(C), \tag{3.12}$$

where $q_p = K_C^{-1}(p)$ and K_C denotes the Kendall distribution function. Furthermore, let C be in the class $\mathcal{C}^d_K$ of copulas for which Kendall's distribution function K_C is strictly increasing and continuous. Assume that for estimating K_C we use a strongly consistent estimator $\hat{K}_n$. We denote by $\hat{K}_n^{-1}$

the usual left-continuous quasi-inverse of $\hat{K}_n$. The multivariate quantile sets $S_p := S_p(C)$ are estimated by

$$\hat{S}_{n,p} := \hat{S}_p(\hat{C}_n) = \{\mathbf{u} \in \mathbb{R}^d | \hat{C}_n(\mathbf{u}) \leq \hat{K}_n^{-1}(p)\}. \tag{3.13}$$

We pose the same question as for the level sets $\hat{S}'_{n,p}$ and S'_p, i.e., do the multivariate quantile sets $\hat{S}_{n,p}$ converge almost surely to their population counterpart S_p? Note, that the situation is more complicated than for the level sets $\hat{S}'_{n,p}$, since now also the level $q_p = K_C^{-1}(p)$ has to be estimated. In general, almost sure convergence of the level sets $\hat{S}'_{n,p}$ to S'_p together with almost sure convergence of the inverse of the empirical Kendall distribution function $\hat{q}_{p,n} = \hat{K}_n^{-1}(p)$ to $q_p = K^{-1}(p)$ is not sufficient to guarantee almost sure convergence of $\hat{S}'_{n,\hat{q}_{p,n}}$ to S'_{q_p}. To guarantee this type of convergence, we need the level sets $\hat{S}'_{n,p}$ to converge not only pointwise to S'_p but also to *converge continuously* to S'_p. An equivalent condition to *continuous convergence* is the notion of uniform convergence on compact sets, or short *compact convergence*. However, note that uniform convergence of the level sets (and thus also compact and continuous convergence) has already been shown in Theorem 3. For a proper definition and the connection between these notions of convergence, we refer the reader to Coblenz et al. (2018b), Appendix A.

We now state the main theorem of this section.

Theorem 4. *Assume that $\hat{C}_n$ is strongly uniform consistent, i.e.,*

$$\|\hat{C}_n - C\|_\infty = \sup_{\boldsymbol{u} \in [0,1]^d} |\hat{C}_n(\boldsymbol{u}) - C(\boldsymbol{u})| \xrightarrow{a.s.} 0\,. \tag{3.14}$$

and that $\hat{K}_n$ is a strongly consistent estimator for the Kendall distribution function K_C.

Assume further that the copula C is strictly increasing, i.e, $x_i < y_i$, $i = 1, \ldots, n$, implies $C(x_1, \ldots, x_n) < C(y_1, \ldots, y_n)$ and that the Kendall distribution function is continuous and strictly monotone.

Then, almost surely,

$$\lim_{n\to\infty} \delta_H(S_p, \hat{S}_{n,p}) = 0\,, \tag{3.15}$$

i.e., the multivariate quantile sets $\hat{S}_{n,p}$ converge almost surely to their population counterparts S_p. In other words $\hat{S}_{n,p}$ is a strongly consistent estimator for S_p.

The proof can be found in Coblenz et al. (2018b), Appendix A.

Some remarks on the assumptions of Theorem 4 and on related results:

1. It has been shown in Deheuvels (1979, 1980) that the empirical copula $\hat{C}$ converges with probability one uniformly to its population counterpart C. Thus, the empirical copula as defined in Equation (2.14) is a suitable estimator that fulfills the assumptions of the theorem.

2. The empirical Kendall distribution as defined in Equation (2.17) is strongly consistent for its population counterpart, see Ghoudi and Rémillard (1998).

3. There are some papers related to level set estimation. However, these either treat different function classes, e.g., densities, (Baillo et al., 2001), only treat fixed levels of the set (Cuevas et al., 2006), or assume continuous distribution estimators (Di Bernardino et al., 2013). Hence, these are not applicable to our specific case here.

4. Note, that these results also guarantee consistency for an upper-level set estimation, i.e., case 3 in Salvadori et al. (2016) and Definition 5 in Chapter 4, since it is the complementary set. Additionally, consistency for the survival approach, i.e., case 4 in Salvadori et al. (2016), is also covered.

3.3.2 Shape of the Estimated Quantile Sets

The basic equation for nonparametric estimation of $S_p(C)$ is

$$\hat{S}_p(\hat{C}) = \{\mathbf{u} \in \mathbb{R}^d | \hat{C}(\mathbf{u}) \leq \hat{K}_C^{-1}(p)\}, \tag{3.16}$$

where $\hat{C}(\mathbf{u})$ is the empirical copula and $\hat{K}_C^{-1}(p)$ is an estimator of the inverse Kendall's distribution function as defined in Equations (2.14) and (2.18). According to Theorem 4, this estimator is consistent. An estimated set $\hat{S}_p(\hat{C})$ is bounded by a polytope, which consists partly of the outer planes of the $[0,1]^d$-hypercube and partly of $(d-1)$-dimensional orthogonal hyperplanes. Figure 3.1 left panel shows examples of $\hat{S}_p(\hat{C})$ in the 2-dimensional case from a sample with size $n = 100$ of a Clayton copula ($\theta = 2$). Black curves indicate the corresponding theoretical boundaries.

In practice, it can be cumbersome to construct the boundaries of an estimated set $\hat{S}_p(\hat{C})$ for graphical representation. This is especially true for dimensions higher than 2. A construction scheme for $d = 2$ is: Let $p' = \hat{K}_C^{-1}(p)$. One can start at point $(\lfloor np' \rfloor, 1)$, where $\lfloor \cdot \rfloor$ denotes the floor function and n is the sample size. Let u_i, $i = 1, \ldots, n$, be the transformed observations in the unit square. The first point of intersection (i.e., corner of the first step) of the orthogonal lines can be found by checking for the

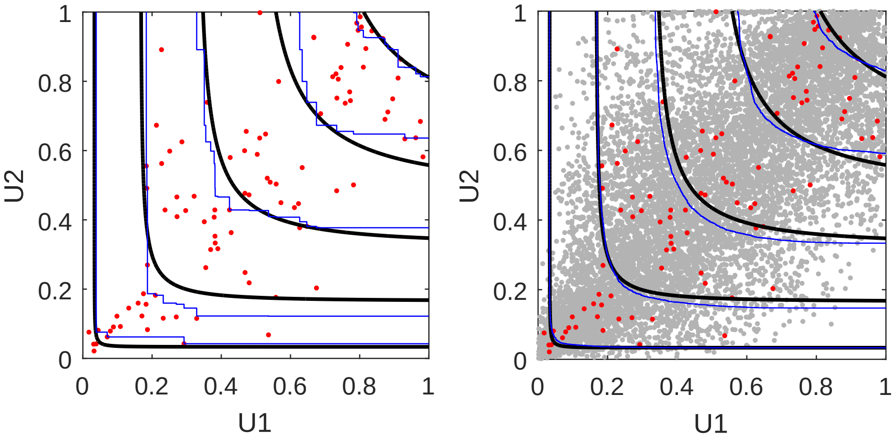

Figure 3.1: Left Panel: Empirical north-eastern boundaries of the quantile sets for a Clayton copula with parameter $\theta = 2$ and sample size $n = 100$. Red dots denote sample observations. The blue boundaries refer to values $p = 0.05, 0.25, 0.5, 0.75, 0.95$ (from left to right). The corresponding theoretical boundaries are in black. Right Panel: Empirical north-eastern boundaries of the quantile sets for a Clayton copula when using a smoothed bootstrap. Parameters are as in the left panel. Gray points refer to the smoothed bootstrap observations (Coblenz et al., 2018b).

largest second component of the u_i's in the rectangle $[0, \lfloor np' \rfloor] \times [0, 1]$. Let this component be denoted by y. Then, the corner has the coordinates $(\lfloor np' \rfloor, y)$. The second corner can be found by checking for the smallest first component of the u_i's in the rectangle $[\lfloor np' \rfloor, 1] \times [0, y]$. Let this component be denoted by x. Then, the second corner has the coordinates (x, y). This procedure can be iterated in an alternating manner, until the last corner with first component equal to 1 is reached.

3.3.3 Smoothed Bootstrap Algorithm

In particular for smaller sample sizes, the shape of the level sets of the empirical copula as shown in the left panel of Figure 3.1 may be coarse and jagged. We propose a smoothed bootstrap procedure to artificially increase the sample size, which helps to smooth the shape. It also improves the accuracy of the estimated quantile sets, as will be shown in the simulation study.

A smoothed bootstrap is a two step procedure, which is equivalent to a combination of bootstrapping (Efron, 1979) and kernel density estimation (Li and Racine, 2007). First, the sample is bootstrapped. Second, sample points are generated from the bootstrapped sample by drawing from a kernel smoothed density. Further references are, e.g., Silverman and Young (1987), Hall et al. (1989), de Angelis and Young (1992), and Ho and Lee (2005). In the following, we provide a detailed description of how to use a smoothed bootstrap in the multivariate quantile estimation process.

For the procedure it is necessary to decide on both a kernel function $\mathbf{K}$ and a bandwidth parameter h. Both choices are somewhat arbitrary and affect the results. However, in our case a Gaussian kernel is suggested, because it fits well with the assigned standard normal margins of the transformed data on $\mathbb{R}^d$ as will become clear later. The Gaussian kernel is the multivariate normal distribution with mean 0 and covariance $\mathbf{\Sigma}$. The bandwidth parameter h controls the covariance matrix $\mathbf{\Sigma}$ used in the Gaussian kernel. This is done by setting in the Gaussian kernel the variance to h^2 and the correlation to $\hat{\rho}h^2$, where $\hat{\rho}$ is the empirical correlation of the original (but transformed) sample. Note, that $\hat{\rho}$ will exist because we operate on a transformed sample with standard normal margins as described below.

In kernel density estimation, the bandwidth parameter h for smoothing can be chosen by different approaches. For an overview see, e.g., Li and Racine (2007). Since we cannot compute the optimal bandwidth parameter h for our estimation problem theoretically, we ran some thorough simulation experiments testing several rules of thumb. These included a rule of thumb for CDF and univariate quantile estimation suggested by Azzalini (1981), the well-known Silverman's rule of thumb (Silverman, 1986), and a recently proposed data driven rule of thumb by Botev et al. (2010). The simulations showed that Silverman's rule of thumb performed most satisfyingly. For the 2-dimensional case, Silverman's rule of thumb is obtained by setting $h = n^{-1/6}$.

When using kernel methods in the $[0, 1]^d$ copula domain, one usually faces boundary problems, since the kernels centered around observations may overlap the boundaries. To circumvent such problems, in the first step we assign standard normal distributed margins to the $i = 1 \ldots n$ original (pseudo-)

observations, i.e.,

$$u_{1i}, u_{2i}, \ldots, u_{di} \rightarrow \Phi^{-1}(u_{1i}), \Phi^{-1}(u_{2i}), \ldots, \Phi^{-1}(u_{di}). \tag{3.17}$$

This procedure is also suggested in Joe (2015).

We operate now on the transformed data on $\mathbb{R}^d$ and bootstrap the sample $n_{sb} >> n$ times, i.e., we draw n_{sb} points with replacement from the n original transformed points. Denote a bootstrapped point as $\mathbf{x_i^b}$, $i = 1 \ldots n_{sb}$. The smoothed bootstrap is completed by adding a point $\mathbf{y_i}$, $i = 1, \cdots, n_{sb}$, drawn from the Gaussian kernel to each $\mathbf{x_i^b}$, i.e.,

$$\mathbf{x_i^{sb}} = \mathbf{x_i^b} + \mathbf{y_i}. \tag{3.18}$$

Note, that this amounts to drawing a realization of a multivariate normal distribution $N(\mathbf{x_i^b}, \mathbf{\Sigma})$ for each bootstrapped point $\mathbf{x_i^b}$. Because of the Gaussian margins, the use of Gaussian kernels within this procedure is optimal for fitting the margins (Silverman, 1986). Furthermore, a Gaussian kernel in this setup allows to transform back the points from the smoothed bootstrap to the unit hypercube by

$$\Phi\left(\frac{x_{1i}^{sb}}{\sqrt{1+h^2}}\right), \Phi\left(\frac{x_{2i}^{sb}}{\sqrt{1+h^2}}\right), \ldots, \Phi\left(\frac{x_{di}^{sb}}{\sqrt{1+h^2}}\right) \rightarrow u_{1i}^{sb}, u_{2i}^{sb}, \ldots, u_{di}^{sb}, \tag{3.19}$$

where $\Phi(\cdot)$ is the univariate standard normal distribution. Thus, by using a Gaussian kernel transformation by ranks can be avoided.

In general, the usefulness of the smoothed bootstrap is a controversial issue (de Angelis and Young, 1992). To further investigate the effect of the smoothed bootstrap on the underlying dependence structure of a sample, we employ a nonparametric test developed in Remillard and Scaillet (2009). The null hypothesis of the test states that the empirical copulas of two samples are the same. We proceed by first sampling from a copula. Second, the smoothed bootstrap procedure described above is used to generate a smoothed bootstrap sample. On this pair of samples, the test is conducted. We carry this out for Clayton, Gauss, and Gumbel copulas with different parameter settings ($\theta_{Clayton} = 0.5, 2, 5$, $\rho = -0.9, -0.5, 0.5, 0.9$ and $\theta_{Gumbel} = 2, 3, 5$) ten times each. In order to keep the runtime low, the original sample comprises 50 points and the smoothed bootstrap sample 500 points. Approximate p-values are obtained by 100 iterations of the wild bootstrap procedure suggested in Remillard and Scaillet (2009). Figure 3.2 shows box plots of the test results. In summary, the test never rejected the null hypothesis even at the 10% significance level, thus yielding evidence that the smoothed bootstrap as specified above does not alter the inherent dependence structure of a given sample severely – at least not for the analyzed copula types.

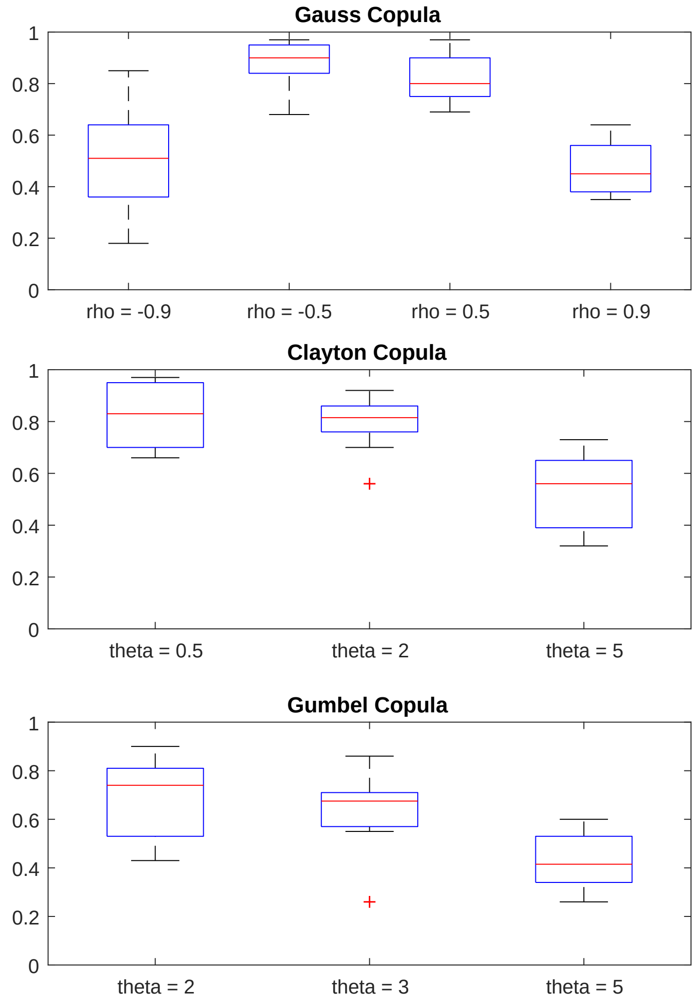

Figure 3.2: Box plots of the results when testing for equal empirical copulas of sample and smooth bootstrapped sample. Each of the original samples comprises 50 data points. The smoothed bootstrap samples comprise 500 data points each. The test is repeated ten times per scenario. In summary, the test was never rejected even for a significance level of 10%, thus providing evidence that a smoothed bootstrap does not dilute the dependence structure of the original data (Coblenz et al., 2018b).

The smoothed bootstrap sample extends the original sample in the copula domain, thus yielding a sample of size $n + n_{sb}$. In this respect it is different from jittering strategies that deal with tied ranks (Pappadà et al., 2017). The smoothed bootstrap multivariate quantile sets $\hat{S}_p^{sb}(\hat{C})$ can be estimated based on the extended sample. Figure 3.1 right panel shows an example of a smoothed bootstrap sample of $n_{sb} = 10,000$ points in gray, as well as the estimated quantile sets from the extended sample $\hat{S}_p^{sb}(\hat{C})$ in blue, for $p = 0.05, 0.25, 0.5, 0.75, 0.95$. The corresponding theoretical boundaries are shown in black. The other parameters are as in the left panel. Clearly, the smoothed bootstrap helps to approximate the true shape of the quantile sets.

In the next section, we show with a simulation study that the smoothed bootstrap can provide both additional accuracy and lower variance to the estimation of multivariate quantiles.

3.4 Simulation Study

It is clear from Figure 3.1 that an estimated quantile set from a smoothed bootstrap sample is less coarse and jagged than the original one and therefore favored by the practitioner. In this section we want to further quantify the following questions: How accurate are the estimation procedures of the multivariate quantiles for different dependence structures and levels p for small to medium sample sizes in terms of bias and variance? What is the benefit of using the smoothed bootstrap in the estimation process? In order to answer these questions, we conduct a simulation study.

We run Monte Carlo (MC) simulations for typical 2-dimensional copulas: Gauss copula, Clayton copula and Gumbel copula. In particular, the Gumbel copula is often used in hydrological applications, see, e.g., Chebana and Ouarda (2011). For the Gauss copula, we set parameter $\rho = -0.9, -0.5, 0.5, 0.9$ in order to test high and moderate negative and positive correlation in the data. For the Clayton copula, we set parameter $\theta_{Clayton} = 0.5, 2, 5$, and for the Gumbel copula, we set parameter $\theta_{Gumbel} = 2, 3, 5$ in order to test for different strengths of dependence. We estimate the $p = 0.05, 0.25, 0.5, 0.75, 0.95$ quantiles. The simulation will thus show a broad picture of the procedure's performance over the whole unit square. We also test different sample sizes starting with a low $n = 50$, further using 100 and 200 observations up to $n = 500$.

For each parameter combination from above, we run $1,000$ MC repetitions and employ the nonparametric estimation procedure described in Section 3.3 in each repetition. For each estimated quantile set, the enclosed theoretical copula volume (i.e., the enclosed theoretical probability mass, cf. Equation

(2.3)) is calculated. Next, we apply a smoothed bootstrap in order to extend each MC sample by additional points drawn from the kernel smoothed density using a Gaussian kernel with Silverman's rule of thumb (Silverman, 1986). Note, that in practice a lot of points should be added (i.e., $100,000$ to $1,000,000$; also cf. Section 3.5). However, in order to keep the simulation runtime low, we choose to extend each MC sample by $n_{sb} = 10,000$ points.

Tables 3.1, 3.2 and 3.3 show the simulation results. In order to assess the usefulness of the smoothed bootstrap, we report results for both the unsmoothed estimation and the smoothed estimation, separately. The tables show the mean, mean squared error (times 100), minimum and maximum of the enclosed theoretical copula volume over the $1,000$ MC repetitions. The following can be observed:

1. As expected, the quantile set estimation gets more accurate the larger the sample size. Overall, for the Gauss, Clayton and Gumbel copulas the unsmoothed estimation procedure provides satisfying results for sample size $n = 500$ (and thus for $n > 500$).

2. For lower samples sizes and $p = 0.05$ and $p = 0.25$, the unsmoothed estimation procedure is highly biased. Moreover, in the Gauss copula case for negative correlation structure it vastly overestimates the true value. This is due to the fact, that in this scenario very few observations are in the lower left area of the unit square. This makes estimation of low quantile sets tremendously harder. In contrast to that, the smoothed estimation procedure greatly reduces the bias in these cases.

3. Bias reduction can also be observed in the Clayton copula and Gumbel copula cases, where only positive dependence structures are tested. In particular in low sample sizes of $n = 50$ and $n = 100$, the smoothed estimation procedure provides less biased results compared to the unsmoothed version.

4. In addition to the bias reductions mentioned above, the smoothed bootstrap lowers estimation variance. The mean squared errors over the MC iterations are lower in most cases for the smoothed procedure compared to the unsmoothed version. Also, the minimum and maximum of the smoothed bootstrap estimates are closer to each other in the smoothed version. Hence, the estimation results are clustered more tightly around the true value. This also indicates that by applying a smoothed bootstrap dramatic deviations from the true value are less severe.

The next section illustrates the nonparametric estimation procedure in a real world application on flood peak and flood volume data.

p	ρ_{Gauss}	n	P	MSE	$P_{\min}$	$P_{\max}$	P^{sb}	MSE^{sb}	$P^{sb}_{\min}$	$P^{sb}_{\max}$
0.05	-0.9	50	0.26	4.82	0.11	0.47	0.07	0.08	0.02	0.18
		100	0.17	1.53	0.07	0.32	0.06	0.04	0.02	0.13
		200	0.10	0.33	0.04	0.20	0.06	0.02	0.03	0.10
		500	0.07	0.07	0.03	0.13	0.05	0.01	0.03	0.09
	-0.5	50	0.14	1.02	0.02	0.34	0.06	0.08	0.01	0.17
		100	0.09	0.23	0.02	0.22	0.06	0.04	0.02	0.13
		200	0.07	0.08	0.02	0.17	0.05	0.02	0.02	0.11
		500	0.06	0.02	0.03	0.10	0.05	0.01	0.03	0.08
	0.5	50	0.08	0.30	0.01	0.29	0.06	0.06	0.01	0.15
		100	0.06	0.08	0.01	0.16	0.06	0.03	0.02	0.13
		200	0.06	0.03	0.01	0.12	0.05	0.02	0.02	0.10
		500	0.05	0.01	0.03	0.10	0.05	0.01	0.03	0.08
	0.9	50	0.07	0.21	0.01	0.25	0.06	0.05	0.01	0.14
		100	0.06	0.07	0.01	0.17	0.05	0.03	0.02	0.12
		200	0.05	0.03	0.02	0.12	0.05	0.02	0.02	0.10
		500	0.05	0.01	0.02	0.09	0.05	0.01	0.03	0.08
0.25	-0.9	50	0.33	1.46	0.12	0.59	0.26	0.22	0.13	0.42
		100	0.30	0.46	0.16	0.48	0.26	0.11	0.16	0.37
		200	0.27	0.18	0.17	0.40	0.25	0.06	0.18	0.33
		500	0.26	0.06	0.20	0.34	0.25	0.03	0.20	0.32
	-0.5	50	0.29	0.65	0.12	0.53	0.26	0.21	0.14	0.40
		100	0.27	0.26	0.12	0.43	0.26	0.13	0.16	0.37
		200	0.26	0.12	0.16	0.38	0.25	0.06	0.19	0.35
		500	0.25	0.04	0.20	0.32	0.25	0.03	0.21	0.31
	0.5	50	0.26	0.39	0.09	0.45	0.26	0.22	0.12	0.42
		100	0.26	0.22	0.13	0.40	0.26	0.13	0.15	0.38
		200	0.25	0.10	0.14	0.36	0.25	0.06	0.17	0.33
		500	0.25	0.04	0.20	0.33	0.25	0.03	0.20	0.32
	0.9	50	0.27	0.44	0.10	0.47	0.26	0.20	0.13	0.40
		100	0.25	0.19	0.11	0.38	0.25	0.12	0.15	0.37
		200	0.25	0.09	0.17	0.34	0.25	0.06	0.19	0.32
		500	0.25	0.04	0.19	0.32	0.25	0.03	0.20	0.31
0.5	-0.9	50	0.55	0.73	0.31	0.77	0.51	0.29	0.36	0.69
		100	0.52	0.30	0.34	0.67	0.50	0.15	0.37	0.63
		200	0.51	0.14	0.41	0.61	0.50	0.08	0.41	0.59
		500	0.51	0.06	0.43	0.58	0.50	0.04	0.44	0.56
	-0.5	50	0.50	0.45	0.29	0.73	0.50	0.28	0.33	0.67
		100	0.51	0.27	0.32	0.68	0.50	0.17	0.38	0.62
		200	0.50	0.12	0.39	0.60	0.50	0.08	0.41	0.58
		500	0.50	0.05	0.44	0.57	0.50	0.04	0.44	0.56
	0.5	50	0.50	0.51	0.31	0.71	0.50	0.28	0.33	0.67
		100	0.50	0.26	0.36	0.63	0.50	0.16	0.37	0.61
		200	0.50	0.13	0.38	0.60	0.50	0.09	0.40	0.59
		500	0.50	0.05	0.44	0.58	0.50	0.04	0.45	0.56
	0.9	50	0.50	0.49	0.30	0.74	0.50	0.29	0.33	0.68
		100	0.50	0.25	0.32	0.64	0.50	0.15	0.35	0.62
		200	0.50	0.13	0.36	0.61	0.50	0.09	0.40	0.58
		500	0.50	0.05	0.43	0.58	0.50	0.04	0.43	0.56
0.75	-0.9	50	0.76	0.30	0.57	0.92	0.75	0.20	0.58	0.87
		100	0.76	0.19	0.57	0.87	0.75	0.11	0.62	0.86
		200	0.75	0.10	0.65	0.85	0.75	0.06	0.67	0.84
		500	0.75	0.04	0.68	0.81	0.75	0.03	0.69	0.80
	-0.5	50	0.76	0.37	0.49	0.91	0.75	0.22	0.56	0.87
		100	0.75	0.19	0.61	0.88	0.75	0.11	0.61	0.85
		200	0.75	0.10	0.62	0.84	0.75	0.06	0.64	0.82
		500	0.75	0.04	0.69	0.81	0.75	0.03	0.70	0.80
	0.5	50	0.75	0.39	0.53	0.92	0.75	0.23	0.59	0.89
		100	0.75	0.18	0.60	0.87	0.75	0.11	0.64	0.86
		200	0.75	0.09	0.64	0.83	0.75	0.06	0.67	0.82
		500	0.75	0.04	0.69	0.82	0.75	0.03	0.70	0.80
	0.9	50	0.75	0.38	0.51	0.93	0.75	0.21	0.59	0.90
		100	0.74	0.20	0.60	0.86	0.75	0.12	0.63	0.85
		200	0.75	0.10	0.65	0.85	0.75	0.06	0.67	0.83
		500	0.75	0.04	0.69	0.81	0.75	0.03	0.70	0.80
0.95	-0.9	50	0.95	0.10	0.78	1.00	0.94	0.06	0.84	0.99
		100	0.94	0.05	0.84	0.99	0.95	0.03	0.88	0.99
		200	0.95	0.02	0.89	0.98	0.95	0.01	0.89	0.98
		500	0.95	0.01	0.91	0.97	0.95	0.01	0.92	0.97
	-0.5	50	0.94	0.11	0.80	1.00	0.95	0.06	0.83	0.99
		100	0.94	0.06	0.84	0.98	0.95	0.03	0.88	0.98
		200	0.95	0.03	0.89	0.99	0.95	0.01	0.90	0.98
		500	0.95	0.01	0.91	0.97	0.95	0.01	0.92	0.97
	0.5	50	0.94	0.11	0.80	1.00	0.95	0.06	0.84	0.99
		100	0.94	0.06	0.84	0.99	0.95	0.03	0.87	0.99
		200	0.95	0.03	0.87	0.98	0.95	0.01	0.89	0.98
		500	0.95	0.01	0.92	0.98	0.95	0.01	0.92	0.97
	0.9	50	0.95	0.09	0.78	1.00	0.95	0.05	0.86	0.99
		100	0.94	0.06	0.84	0.99	0.95	0.03	0.89	0.98
		200	0.95	0.03	0.90	0.99	0.95	0.01	0.91	0.98
		500	0.95	0.01	0.91	0.98	0.95	0.01	0.92	0.98

Table 3.1: Simulation results for a 2-dimensional Gauss copula. Index *sb* indicates smoothed bootstrap results.

p	θ_{Clayton}	n	P	MSE	$P_{\min}$	$P_{\max}$	P^{sb}	MSE^{sb}	$P^{sb}_{\min}$	$P^{sb}_{\max}$
0.05	0.5	50	0.08	0.31	0.00	0.27	0.06	0.06	0.01	0.15
		100	0.06	0.10	0.01	0.21	0.06	0.03	0.01	0.14
		200	0.06	0.03	0.02	0.13	0.05	0.02	0.02	0.11
		500	0.05	0.01	0.02	0.09	0.05	0.01	0.03	0.08
	2	50	0.07	0.19	0.01	0.25	0.05	0.05	0.01	0.15
		100	0.06	0.06	0.01	0.15	0.05	0.02	0.02	0.12
		200	0.05	0.03	0.01	0.11	0.05	0.01	0.02	0.09
		500	0.05	0.01	0.02	0.08	0.05	0.01	0.03	0.08
	5	50	0.07	0.15	0.01	0.26	0.05	0.05	0.01	0.15
		100	0.05	0.06	0.01	0.18	0.05	0.02	0.02	0.12
		200	0.05	0.03	0.01	0.14	0.05	0.01	0.02	0.12
		500	0.05	0.01	0.02	0.09	0.05	0.01	0.02	0.08
0.25	0.5	50	0.26	0.46	0.09	0.49	0.25	0.21	0.11	0.43
		100	0.26	0.20	0.14	0.41	0.25	0.12	0.15	0.36
		200	0.25	0.09	0.17	0.34	0.25	0.06	0.19	0.33
		500	0.25	0.04	0.18	0.32	0.25	0.03	0.18	0.30
	2	50	0.26	0.47	0.10	0.45	0.24	0.21	0.13	0.40
		100	0.26	0.21	0.13	0.40	0.25	0.11	0.14	0.37
		200	0.25	0.10	0.17	0.36	0.25	0.06	0.17	0.31
		500	0.25	0.04	0.20	0.32	0.25	0.03	0.20	0.30
	5	50	0.27	0.44	0.10	0.46	0.25	0.19	0.12	0.40
		100	0.25	0.20	0.13	0.42	0.25	0.11	0.16	0.39
		200	0.25	0.10	0.16	0.34	0.25	0.06	0.18	0.33
		500	0.25	0.04	0.19	0.31	0.25	0.03	0.19	0.29
0.5	0.5	50	0.51	0.52	0.27	0.72	0.50	0.29	0.30	0.66
		100	0.50	0.25	0.33	0.67	0.50	0.15	0.37	0.62
		200	0.50	0.13	0.39	0.61	0.50	0.08	0.41	0.59
		500	0.50	0.05	0.43	0.56	0.50	0.04	0.43	0.55
	2	50	0.50	0.52	0.24	0.75	0.49	0.30	0.29	0.67
		100	0.50	0.26	0.33	0.67	0.49	0.16	0.38	0.61
		200	0.50	0.14	0.37	0.63	0.50	0.09	0.40	0.61
		500	0.50	0.05	0.43	0.57	0.50	0.04	0.44	0.56
	5	50	0.51	0.48	0.29	0.74	0.50	0.26	0.34	0.66
		100	0.50	0.26	0.33	0.65	0.50	0.15	0.34	0.62
		200	0.50	0.12	0.39	0.63	0.49	0.08	0.41	0.59
		500	0.50	0.05	0.42	0.57	0.50	0.04	0.43	0.56
0.75	0.5	50	0.75	0.39	0.54	0.90	0.75	0.22	0.55	0.87
		100	0.75	0.21	0.60	0.88	0.75	0.12	0.64	0.86
		200	0.75	0.10	0.65	0.84	0.75	0.06	0.68	0.83
		500	0.75	0.04	0.67	0.80	0.75	0.03	0.69	0.80
	2	50	0.76	0.36	0.53	0.92	0.75	0.20	0.57	0.88
		100	0.75	0.18	0.56	0.87	0.75	0.12	0.62	0.85
		200	0.75	0.10	0.64	0.84	0.75	0.07	0.66	0.83
		500	0.75	0.04	0.69	0.81	0.75	0.03	0.70	0.80
	5	50	0.75	0.42	0.50	0.95	0.75	0.23	0.56	0.89
		100	0.74	0.21	0.60	0.86	0.75	0.12	0.62	0.85
		200	0.75	0.10	0.65	0.84	0.75	0.06	0.66	0.82
		500	0.75	0.04	0.68	0.82	0.75	0.03	0.69	0.80
0.95	0.5	50	0.95	0.10	0.81	1.00	0.95	0.05	0.84	0.99
		100	0.94	0.06	0.83	0.99	0.95	0.02	0.87	0.98
		200	0.95	0.03	0.88	0.98	0.95	0.01	0.90	0.98
		500	0.95	0.01	0.91	0.98	0.95	0.01	0.92	0.97
	2	50	0.95	0.09	0.81	1.00	0.95	0.04	0.88	0.99
		100	0.94	0.06	0.83	0.99	0.95	0.02	0.89	0.99
		200	0.95	0.03	0.87	0.98	0.95	0.01	0.91	0.98
		500	0.95	0.01	0.90	0.97	0.95	0.01	0.92	0.97
	5	50	0.95	0.10	0.82	1.00	0.96	0.04	0.88	0.99
		100	0.94	0.06	0.85	0.99	0.95	0.02	0.89	0.99
		200	0.95	0.03	0.88	0.99	0.95	0.01	0.91	0.98
		500	0.95	0.01	0.91	0.97	0.95	0.01	0.92	0.97

Table 3.2: Simulation results for a 2-dimensional Clayton copula. Index *sb* indicates smoothed bootstrap results.

p	θ_{Gumbel}	n	P	MSE	$P_{\min}$	$P_{\max}$	P^{sb}	MSE^{sb}	$P^{sb}_{\min}$	$P^{sb}_{\max}$
0.05	2	50	0.08	0.25	0.00	0.29	0.06	0.06	0.01	0.15
		100	0.06	0.08	0.01	0.17	0.06	0.03	0.02	0.12
		200	0.06	0.03	0.02	0.13	0.05	0.01	0.03	0.11
		500	0.05	0.01	0.03	0.10	0.05	0.01	0.03	0.09
	3	50	0.08	0.22	0.00	0.25	0.06	0.06	0.01	0.16
		100	0.06	0.07	0.01	0.16	0.05	0.03	0.02	0.11
		200	0.05	0.03	0.02	0.11	0.05	0.01	0.02	0.09
		500	0.05	0.01	0.03	0.09	0.05	0.01	0.03	0.08
	5	50	0.07	0.19	0.00	0.26	0.06	0.06	0.01	0.17
		100	0.06	0.06	0.01	0.14	0.05	0.03	0.01	0.12
		200	0.05	0.02	0.02	0.11	0.05	0.01	0.02	0.10
		500	0.05	0.01	0.03	0.09	0.05	0.01	0.03	0.08
0.25	2	50	0.26	0.41	0.10	0.53	0.26	0.22	0.15	0.45
		100	0.25	0.19	0.12	0.42	0.26	0.12	0.15	0.38
		200	0.25	0.10	0.17	0.36	0.25	0.07	0.18	0.35
		500	0.25	0.04	0.19	0.31	0.25	0.03	0.21	0.30
	3	50	0.27	0.45	0.09	0.50	0.26	0.22	0.12	0.42
		100	0.25	0.19	0.14	0.41	0.25	0.11	0.17	0.36
		200	0.25	0.09	0.15	0.35	0.25	0.06	0.17	0.35
		500	0.25	0.04	0.18	0.31	0.25	0.03	0.19	0.30
	5	50	0.26	0.44	0.09	0.51	0.25	0.22	0.12	0.41
		100	0.25	0.20	0.14	0.42	0.25	0.12	0.16	0.38
		200	0.25	0.09	0.15	0.36	0.25	0.06	0.18	0.34
		500	0.25	0.04	0.19	0.31	0.25	0.03	0.20	0.30
0.5	2	50	0.50	0.49	0.25	0.73	0.50	0.30	0.30	0.66
		100	0.50	0.23	0.34	0.64	0.50	0.15	0.37	0.61
		200	0.50	0.11	0.40	0.63	0.50	0.08	0.43	0.60
		500	0.50	0.05	0.44	0.57	0.50	0.04	0.44	0.56
	3	50	0.50	0.49	0.25	0.69	0.50	0.30	0.31	0.65
		100	0.50	0.23	0.33	0.64	0.50	0.15	0.36	0.61
		200	0.50	0.12	0.39	0.62	0.50	0.08	0.41	0.59
		500	0.50	0.04	0.42	0.56	0.50	0.03	0.45	0.55
	5	50	0.50	0.51	0.25	0.69	0.50	0.29	0.31	0.64
		100	0.50	0.24	0.34	0.67	0.50	0.15	0.36	0.62
		200	0.50	0.12	0.37	0.61	0.50	0.08	0.39	0.59
		500	0.50	0.05	0.44	0.57	0.50	0.04	0.44	0.56
0.75	2	50	0.75	0.36	0.55	0.90	0.75	0.22	0.59	0.87
		100	0.75	0.20	0.60	0.88	0.75	0.13	0.63	0.85
		200	0.75	0.10	0.65	0.83	0.75	0.06	0.66	0.81
		500	0.75	0.04	0.68	0.81	0.75	0.03	0.69	0.80
	3	50	0.75	0.36	0.54	0.90	0.75	0.22	0.59	0.87
		100	0.74	0.19	0.60	0.85	0.75	0.11	0.61	0.85
		200	0.75	0.10	0.63	0.83	0.75	0.07	0.65	0.83
		500	0.75	0.04	0.69	0.81	0.75	0.03	0.69	0.80
	5	50	0.75	0.41	0.53	0.91	0.74	0.23	0.59	0.87
		100	0.74	0.20	0.56	0.88	0.75	0.12	0.61	0.85
		200	0.75	0.10	0.65	0.84	0.75	0.07	0.66	0.84
		500	0.75	0.04	0.68	0.81	0.75	0.03	0.68	0.79
0.95	2	50	0.95	0.10	0.82	1.00	0.94	0.06	0.84	0.99
		100	0.94	0.06	0.84	0.99	0.94	0.03	0.88	0.99
		200	0.95	0.03	0.90	0.99	0.95	0.02	0.90	0.99
		500	0.95	0.01	0.91	0.98	0.95	0.01	0.91	0.97
	3	50	0.95	0.11	0.76	1.00	0.94	0.06	0.83	1.00
		100	0.94	0.06	0.84	0.99	0.95	0.03	0.88	0.98
		200	0.95	0.03	0.89	0.98	0.95	0.02	0.90	0.97
		500	0.95	0.01	0.92	0.98	0.95	0.01	0.92	0.97
	5	50	0.94	0.10	0.78	1.00	0.94	0.06	0.84	0.99
		100	0.94	0.06	0.83	0.99	0.95	0.03	0.86	0.98
		200	0.95	0.03	0.89	0.98	0.95	0.02	0.91	0.98
		500	0.95	0.01	0.91	0.97	0.95	0.01	0.92	0.97

Table 3.3: Simulation results for a 2-dimensional Gumbel copula. Index *sb* indicates smoothed bootstrap results.

3.5 Application

We apply the nonparametric estimation procedure to flood event data. The aim of this section is, first, to exemplify the overall applicability of the proposed methods. Second, we discuss some effects when dealing with small samples and how the smoothed bootstrap can overcome these. Third, we demonstrate some further benefits of the smoothed bootstrap approach considering the quantile set structure.

The data is taken from Yue et al. (1999). It is also used in Chebana and Ouarda (2011), who compute multivariate quantiles for the data with a different approach. The data set comprises flood peak and flood volume of the Ashuapmushuan basin in Quebec, Canada, for the years 1963-1995. This results in 33 observations. Figure 3.3 upper left panel shows a scatter plot of the data points. Since the notion of multivariate quantiles used here measures probabilities of small values – like univariate quantiles do – we multiplied each data value by -1. Thus, large values for flood peak and flood volume correspond to quantiles at small levels. This is equivalent to a survival approach and ensures the (graphical) orientation of the multivariate quantiles as applied here (also cf. case 4 in Salvadori et al. (2016)).

The data is transformed to the unit square using ranks. This is shown in Figure 3.3 lower left panel. Clearly, the data exhibits positive dependence. The quantile set $S_{0.1}(C)$ is estimated according to the procedure outlined in Section 3.3.2. The result is shown in Figure 3.4 as a blue line. Red dots represent the transformed observations. Note that the user does not have to decide on a specific copula in the procedure, which is generally the case for nonparametric approaches.

We then use the smoothed bootstrap method to improve the quantile estimation results. Therefore, we transform the copula pseudo-observations via the inverse standard normal distribution. Then, we add 1 million smoothed bootstrap points. Smoothing is done with Silverman's rule of thumb (Silverman, 1986) and a Gaussian kernel as described in Section 3.3.3. The extended sample is illustrated in Figure 3.3 upper right panel. Blue dots indicate the original sample, whereas gray dots are the smoothed bootstrap points. Note, that the margins are standard normal by construction. Then the extended sample is transformed back to the unit square. This is depicted in Figure 3.3 lower right panel. On the extended sample $S_{0.1}(C)$ is estimated again.

Figure 3.4 shows a comparison of the quantile set from the original sample in blue and of the quantile set from the smoothed bootstrap in black. The quantile set of the smoothed bootstrap sample is smoother than that of the original sample. Also, the black line sags more towards the lower left corner of the unit square. Based on the simulation study, we can conclude that the

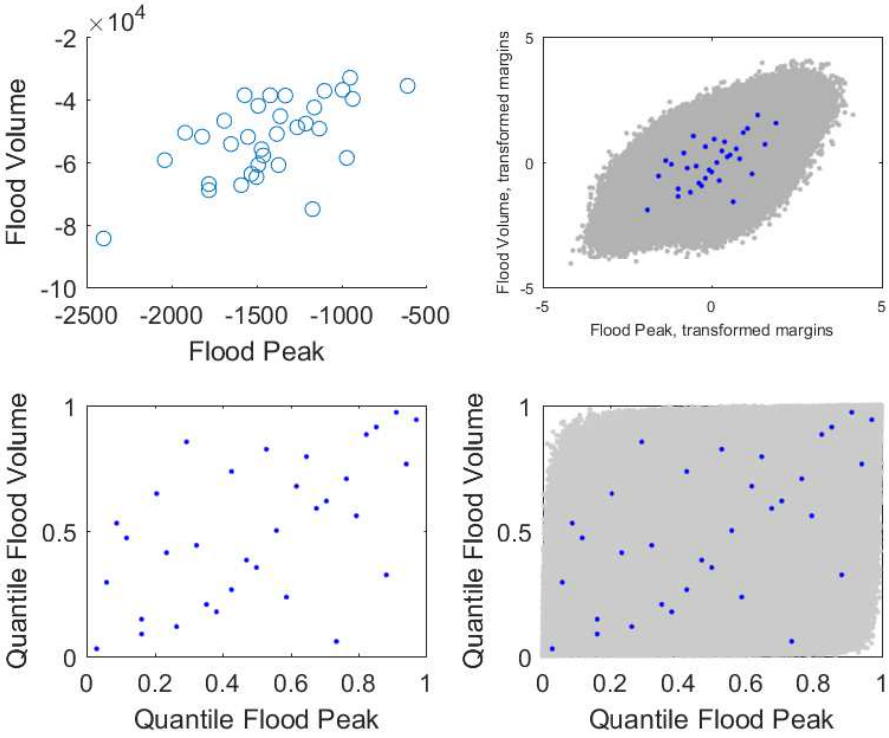

Figure 3.3: Upper Left: The original flood peak and volume data. Lower Left: Data rank transformed to $[0,1]^2$ domain. One can see the positive dependence. Upper Right: The sample after the smoothed bootstrap shown in the domain with standard normal margins. Blue dots indicate the original sample, gray dots indicate the added smoothed bootstrap sample. Lower Right: Smoothed bootstrap sample rank transformed to $[0,1]^2$ domain. Blue dots again indicate the original sample, gray dots indicate the added smoothed bootstrap sample (Coblenz et al., 2018b).

smoothed bootstrap estimator gives a more realistic picture considering the shapes of the smoothed and unsmoothed boundaries. Additionally, due to the small sample size, there is an upward bias in estimating $K_C^{-1}(p)$ from the original sample, which is not apparent in the smoothed bootstrap estimator. These effects are illustrated in the tables in Section 3.4. The simulation study indicated that in small samples for $p = 0.05$ up to $p = 0.25$ the unsmoothed procedure overestimates the true values, whereas the smoothed bootstrap mitigates this.

To give some further interpretation, each point can be thought of corre-

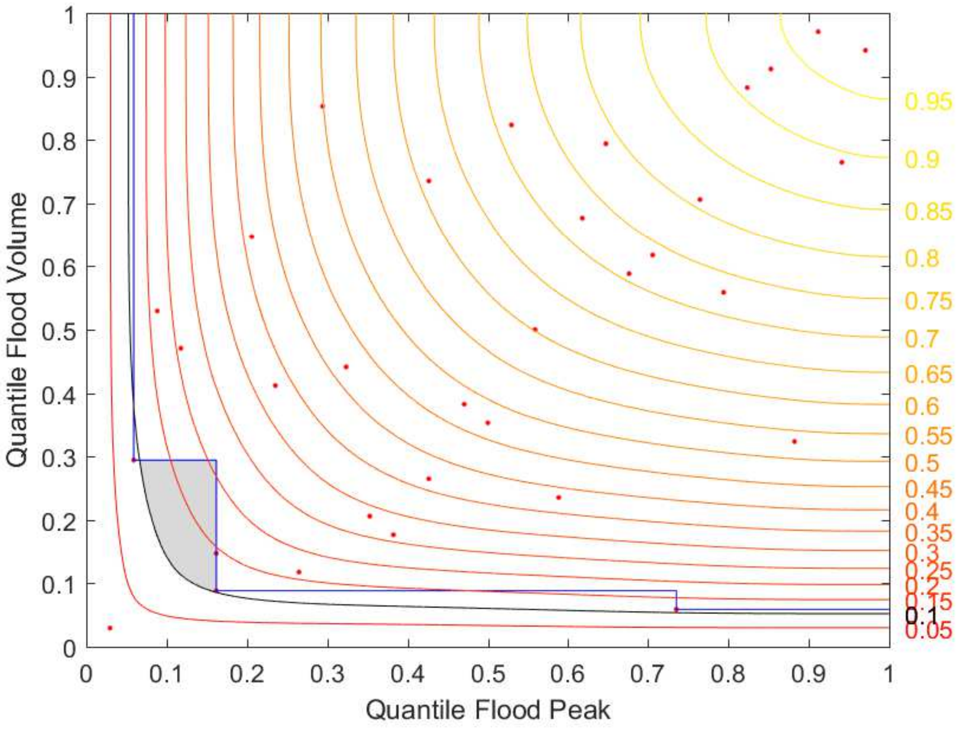

Figure 3.4: Results of the estimation procedure shown in the copula domain. Red dots are the original data points. The blue and black lines depict the unsmoothed quantile and smoothed quantile set for $p = 0.1$, respectively. The red-colored lines depict quantile sets using the smoothed bootstrap for $p = 0.05, 0.1, \ldots, 0.95$ (from left to right). The smoothed bootstrap procedure delivers a broader picture of the underlying dependence structure. Also, as can be seen for the points in the shaded area, the unsmoothed procedure can lead to underestimation of risk (Coblenz et al., 2018b).

sponding to a specific hydrograph. For points on the quantile boundary, the hydrographs are equivalently extreme, e.g., in the sense of the Kendall Return Period as defined in Salvadori et al. (2011), Definition 4. The points below the boundary correspond to hydrographs that represent more extreme events. Once the quantile set is estimated, a new point – potentially stemming from a hydrograph – is easily classified as within the set by transforming it to the unit square and checking its position. Conversely, any points from within the quantile can be transformed back to the volume-peak domain. Thus, multivariate quantiles can provide important input, e.g., for dam design simulation (Requena et al., 2013) and structural failure approaches (Salvadori et al., 2015).

Figure 3.4 also shows the estimated quantile sets of the smoothed bootstrap procedure for $p = 0.05, 0.1, \ldots, 0.95$ as red-colored lines. The points in the shaded area are in the estimated 10% quantile set when using the unsmoothed procedure. However, as can be seen in the figure some of them are in the 15% and 20% quantile sets of the smoothed bootstrap procedure (and some of them are in quantile sets of even higher level). These points would yield a higher Kendall Return Period (Salvadori et al., 2011) than expected from the unsmoothed procedure and are thus more common events than anticipated. This small sample bias is reduced by the smoothed bootstrap procedure, which delivers a more accurate picture of the quantile set structure.

In addition to that, the smoothed bootstrap procedure is capable of giving a more fine-grained picture of quantile sets at very low levels. These are particularly important for identifying extreme events. Figure 3.5 shows a comparison of unsmoothed and smoothed bootstrap procedure for the estimated quantile sets at level $p = 0.005, 0.01, \ldots, 0.04$. Lines in black correspond to the smoothed bootstrap procedure whereas lines in blue correspond to the unsmoothed procedure. A first thing to note is that the unsmoothed procedure is unable to produce different estimates for $p = 0.005, \ldots, 0.03$ and $p = 0.035, 0.04$, respectively. Also, for quantile sets at levels lower than $p = 0.005$ the estimated quantile set boundary remains at the left blue line. Thus, the unsmoothed procedure is incapable of producing reliable results for low quantiles in small samples. In contrast to that, the smoothed bootstrap procedure still gives a well nuanced picture.

3.6 Concluding Remarks

In this chapter, we introduce a nonparametric estimation procedure for multivariate quantile sets. The quantile sets are specific level sets of a copula. The interpretation links nicely to the univariate case, since the probability mass p enclosed by the quantile set corresponds to the p-quantile. Also, the presented quantile sets characterize the copula itself, which provides theoretical validity of the approach.

The proposed estimation procedure comprises estimating the copula, Kendall's distribution function and its inverse, and the quantile sets. We show consistency of the estimator. Furthermore, we design a smoothed bootstrap to increase accuracy of the estimator. In order to shed light on the procedure's performance in finite samples, we conduct a simulation study. We find

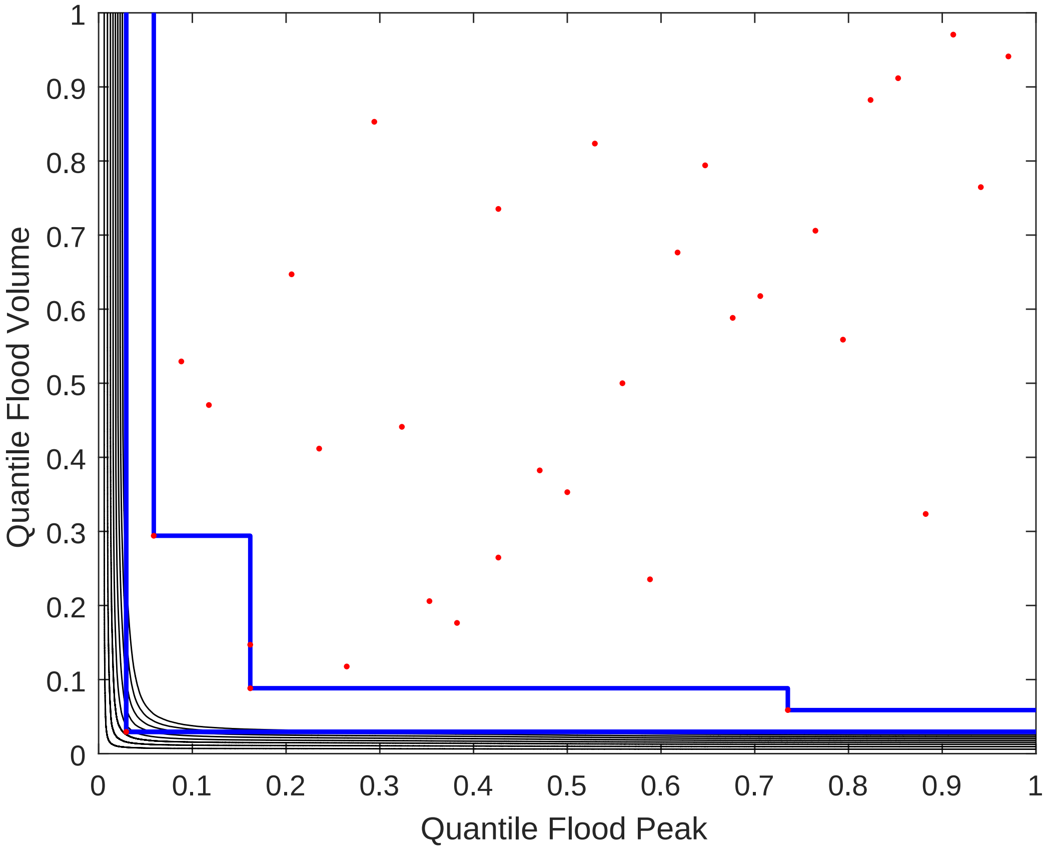

Figure 3.5: Results of the estimation procedure shown in the copula domain. Red dots are the original data points. The blue lines depict the quantile sets for $p = 0.005, 0.01, \ldots, 0.04$ of the unsmoothed procedure. According to the unsmoothed procedure the quantile sets for $p = 0.005, \ldots, 0.03$ and $p = 0.035, 0.04$ are the same, respectively. The black lines depict the same quantile sets using the smoothed bootstrap. Clearly, the smoothed bootstrap procedure shows a more fine-grained picture for quantiles at very low levels (Coblenz et al., 2018b).

that the smoothed bootstrap provides improvements in accuracy, in particular in small sample sizes. Furthermore, the smoothed bootstrap reduces mean squared error. The smoothed bootstrap might also be an interesting topic for future research in hydrology when dealing with ties in the data, where randomization strategies are needed, cf. Pappadà et al. (2017).

The type of multivariate quantile sets employed here has potential use in applications of hydrology, coastal engineering, and finance. We illustrate this by providing an application example in hydrology using 33 sample points.

The data is from the Ashuapmushuan basin in Quebec, Canada, for the years 1963-1995. We estimate a multivariate quantile for flood peak and volume. The example shows that the procedure is flexible and easy to implement, yet produces reliable results. This is especially the case when sample sizes are small. Extending particularly small samples by smoothed bootstrap points can provide additional insight to the analysis. We hope that practitioners and researchers recognize the potential of the nonparametric estimation procedure presented here and will utilize it in their domain of interest.

In the next chapter, we further deal with multivariate quantiles. However this time, the focus is on assessing the uncertainty of the estimation via confidence regions and not on the estimation procedure itself.

Chapter 4

Confidence Regions for Multivariate Quantiles

This chapter is based on Coblenz et al. (2018a).

4.1 Introduction

The track record of multivariate quantiles in hydrology is long and started with the papers by Yue and Rasmussen (2002), Salvadori and De Michele (2004), and Salvadori (2004). Quickly, a growing literature on this topic with an application focus arose, see, e.g., Chebana and Ouarda (2011), Salvadori et al. (2014), and Salvadori et al. (2015). A thorough overview of the current state of the art can be found in Salvadori et al. (2016). The notion of multivariate quantile we use in this chapter is based on copulas. It has the nice feature that a $100\% \cdot p$ multivariate quantile separates the copula domain into two sets, one comprising p, the other comprising $1 - p$ of the total probability mass. Some theoretical aspects can be found in, e.g., Salvadori and De Michele (2007), Salvadori et al. (2013), and the previous chapter based on Coblenz et al. (2018b).

Not only the estimation of multivariate quantiles as outlined in the previous chapter is important, but also an assessment of the estimation uncertainty. Confidence regions can be an essential tool for doing this. In contrast to pointwise confidence bands, confidence regions provide a holistic precision analysis of multivariate quantiles. For example, Serinaldi (2013, 2016) constructs confidence regions for multivariate quantiles based on highest density regions (Hyndman, 1996). Yet, in principle any approach for constructing confidence regions of level sets is applicable since the multivariate quantiles considered are specific level sets.

We attempt to fill this research gap and contribute to the existing literature on multivariate quantiles in several ways. First, we extend two recently developed approaches for construction of level set confidence regions by Mammen and Polonik (2013) and Chen et al. (2017) to the estimation problem at hand. Note that the multivariate quantiles considered here are level sets at specific levels of the copula. However, in contrast to the cited works, where the levels are known and fixed in advance, the level of the multivariate quantile has to be estimated. Second, we check the coverage probabilities of the extended methods by a simulation study in order to investigate their reliability. Finally, we apply the methods on a small sample of flood data to gain further insights.

The chapter is structured as follows: The next section introduces the notion of multivariate quantiles used here. The confidence region approaches by Mammen and Polonik (2013) and Chen et al. (2017) are discussed in Section 4.3. Moreover, they are extended to multivariate quantile estimation. In Section 4.4, a simulation study is conducted in order to explore the strengths and weaknesses of the considered methods. The chapter is concluded by an application on a small sample of flood data and a discussion of some further aspects.

4.2 Notational Preliminaries

This section introduces the notion of multivariate quantiles we use throughout this chapter. In the following, we make use of the concepts on empirical estimation, Kendall's distribution function, and the Hausdorff distance as outlined in Chapter 2. Additionally, some specific notation and preliminaries are covered.

Subsequently, we define the notion of multivariate quantiles we use in this chapter. It has been used previously in, e.g., Salvadori et al. (2013, 2016). It is slightly different as in the previous chapter for reasons of (notational) convenience when used in the confidence region methods later on. We want to point out that the methods developed in this chapter can be easily transferred to the multivariate quantiles as defined in Chapter 3. In the following, let $\mathcal{C}_K^d$ denote the class of copulas for which the Kendall distribution function K_C is strictly increasing and continuous.

Definition 5 (Salvadori et al. (2013)). *For a copula $C \in \mathcal{C}_K^d$ and $p \in [0,1]$ a multivariate quantile is defined as*

$$S_p(C) := \{\boldsymbol{u} \in [0,1]^d : C(\boldsymbol{u}) \geq K_C^{-1}(p)\}. \tag{4.1}$$

We can now write $\mathbb{P}(S_p(C)) = \mathbb{P}(C(\mathbf{u}) \geq K_C^{-1}(p)) = 1 - \mathbb{P}(C(\mathbf{u}) \leq K_C^{-1}(p)) = 1 - K_C(K_C^{-1}(p)) = 1 - p$. Hence, the boundary of the $p \cdot 100\%$ multivariate quantile partitions the copula domain into a set comprising probability mass p and a set comprising probability mass $1 - p$, which is a nice feature of this particular definition. Furthermore, the shape of the boundary is determined by the shape of the level curve of the copula. The level curve reflects the distribution of the probability mass and the strength of dependence between the involved variables (see, e.g., Chapter 5) which transfers to the quantile definition here. For further motivation and theoretical considerations of this approach, see Salvadori et al. (2016) or Coblenz et al. (2018b). Note that because $\mathbb{R}^d$, $d > 1$, has no total ordering, there are many other notions of multivariate quantiles, see, e.g., Tibiletti (1993), Chaudhuri (1996), Serfling (2002), Chebana and Ouarda (2011), and Di Bernardino et al. (2013). However, we do not consider these further here.

S_p can be estimated either by

$$\hat{S}_p(\hat{C}) = \{\mathbf{u} \in \mathbb{R}^d | \hat{C}(\mathbf{u}) \geq \hat{K}_C^{-1}(p)\}, \tag{4.2}$$

or by

$$\hat{S}_p(\hat{C}_h) = \{\mathbf{u} \in \mathbb{R}^d | \hat{C}_h(\mathbf{u}) \geq \hat{K}_C^{-1}(p)\}, \tag{4.3}$$

where $\hat{K}_C^{-1}$ is as defined in Equation (2.18), $\hat{C}$ is the empirical copula (2.14), and $\hat{C}_h$ is the kernel estimated copula (2.15). The estimator $\hat{S}_p(\hat{C})$ is consistent, cf. Theorem 4. An algorithm to construct the estimator on a given bivariate copula sample can be found in Section 3.3.2.

We want to point out that the estimators (4.2) and (4.3) can be used for cases 3 and 4 in Salvadori et al. (2016). Also, the estimators cover the multivariate quantiles used in in the previous chapter. Furthermore, note that we use a nonparametric approach for multivariate quantile estimation here. Parametric and semiparametric estimators can be found, e.g., in Salvadori and De Michele (2007) and Salvadori et al. (2013). In the next section we introduce two approaches to construct confidence regions for level set estimation and extend them to multivariate quantiles.

4.3 Confidence Regions for Multivariate Quantiles

In this section, we introduce the approaches by Mammen and Polonik (2013) and Chen et al. (2017). These construct confidence regions for estimated level sets. The Mammen and Polonik (2013) method is applicable to level sets of any functions. On the other hand, the method by Chen et al. (2017) is developed for densities. Both approaches assume a fixed level for which the level set is estimated. This is different from what we need in the context of multivariate quantiles. Thus, we make necessary extensions to the approaches in order to make them applicable to estimated multivariate quantiles. We introduce each method in turn and extend it. Moreover, we address some computational aspects.

We want to point out that there are other methods to construct confidence regions for multivariate quantiles. For example, Serinaldi (2013, 2016) follows a quite different approach based on highest density regions (Hyndman, 1996). This method constructs confidence regions which are centered at the distribution of points on a level set. In contrast to that, the approaches by Mammen and Polonik (2013) and Chen et al. (2017) yield confidence regions that bound the multivariate quantile. Thus, the techniques are principally incomparable to one another. Therefore, we do not consider approaches based on highest density regions further.

4.3.1 Approach extended from Mammen and Polonik (2013)

The approach by Mammen and Polonik (2013) is based on the supremum distance between a function and its estimate on a specific set of points. It can be used to construct confidence regions for level sets of the form $L = \{\mathbf{x} \in \mathbb{R}^d : f(\mathbf{x}) \geq 0\}$ of an arbitrary function $f : \mathbb{R}^d \to \mathbb{R}$. Note that by using the function $h(\mathbf{x}) = f(\mathbf{x}) - \lambda$, instead, one can construct confidence regions for level sets of f at any level λ. In the following, let $L^- = \{\mathbf{x} \in \mathbb{R}^d : f(\mathbf{x}) > 0\}$ and let n denote the sample size. The approach seeks to find sets $\hat{L}_\ell$ and $\hat{L}_u$ such that

$$\mathbb{P}(\hat{L}_\ell \subset L^- \text{ and } L \subset \hat{L}_u) \xrightarrow{n\to\infty} 1 - \alpha, \tag{4.4}$$

where $1 - \alpha$ is the confidence level. The sets $\hat{L}_\ell$ and $\hat{L}_u$ are estimated by

$$\hat{L}_\ell = \{\mathbf{x} \in \mathbb{R}^d : \hat{f}(\mathbf{x}) > \hat{b}_n\} \text{ and } \hat{L}_u = \{\mathbf{x} \in \mathbb{R}^d : \hat{f}(\mathbf{x}) \geq -\hat{b}_n\}, \tag{4.5}$$

where $\hat{f}$ is an estimator of f and $\hat{b}_n$ is an estimator of the $1 - \alpha$ quantile of $Z = \sup_{\mathbf{x}\in\mathbb{R}^d:|f(\mathbf{x})|\leq\beta} |\hat{f}(\mathbf{x}) - f(\mathbf{x})|$. Since the distribution of Z is unknown, Mammen and Polonik (2013) suggest using a bootstrap.

The approach above is not directly applicable to the multivariate quantiles S_p since it assumes the level λ to be fixed. In contrast to that, estimation of S_p requires estimation of the level $K_C^{-1}(p)$. Thus, we have to extend the method. Let $\mathbf{U}_1, \ldots, \mathbf{U}_n$ be a d-dimensional copula sample. Then, the approach by Mammen and Polonik (2013) is extended and applied as in Algorithm 1.

Algorithm 1 Extension of Mammen and Polonik (2013).

```
1: Choose the level $p$ and the confidence level $1-\alpha$.
2: Estimate $K_C^{-1}(p)$ and $S_p$ on $\mathbf{U}_1, \ldots, \mathbf{U}_n$ according to Equation (2.18) and Equation (4.2), respectively.
3: Determine $\Delta_n = \{\mathbf{u} \in \{\mathbf{U}_1, \ldots, \mathbf{U}_n\} : -\beta_n \leq \hat{C}(\mathbf{u}) - \hat{K}_C^{-1} \leq \beta_n\}$, where $\beta_n = n^{-1/2}$ and $\hat{C}$ is the empirical copula in Equation (2.14).
4: Draw $n_{bs}$ bootstrap samples $\mathbf{U}_1^*, \ldots, \mathbf{U}_n^*$. Repeat Step 2 on each of these.
5: Let $\hat{C}_i$ and $\hat{K}_{C,i}^{-1}$ be the empirical copula and estimated inverse Kendall function of the $i$th bootstrap sample. Determine $Z_i = \max_{\mathbf{u} \in \Delta_n} |\hat{C}_i(\mathbf{u}) - \hat{K}_{C,i}^{-1} - \hat{C}(\mathbf{u}) + \hat{K}_C^{-1}|$ for each $i = 1, \ldots, n_{bs}$.
6: Estimate $b_n$ as the empirical $1-\alpha$-quantile of $Z = (Z_1, \ldots, Z_{n_{bs}})$.
7: The confidence region of $\hat{S}_p(\hat{C})$ is determined by the two sets $\hat{S}_\ell = \{\mathbf{v} \in [0,1]^d : \hat{C}(\mathbf{v}) > \hat{K}_C^{-1}(p) - \hat{b}_n\}$ and $\hat{S}_u = \{\mathbf{v} \in [0,1]^d : \hat{C}(\mathbf{v}) \geq \hat{K}_C^{-1}(p) + \hat{b}_n\}$.
```

Note that by incorporating the estimation of $K_C^{-1}(p)$ into Step 5, we account for the estimation uncertainty of $\hat{S}_p(\hat{C})$ and $\hat{K}_C^{-1}(p)$ simultaneously. Thus, we propose to use $h(x) = C(x) - K_C^{-1}(p)$ and we can write

$$\sup_{\mathbf{x} \in \mathbb{R}^d} \left| \hat{h}(\mathbf{x}) - h(\mathbf{x}) \right| = \sup_{\mathbf{x} \in \mathbb{R}^d} \left| \left(\hat{C}(\mathbf{x}) - \hat{K}_C^{-1}(p) \right) - \left(C(\mathbf{x}) - K_C^{-1}(p) \right) \right| \tag{4.6}$$

$$= \sup_{\mathbf{x} \in \mathbb{R}^d} \left| \left(\hat{C}(\mathbf{x}) - C(\mathbf{x}) \right) - \left(\hat{K}_C^{-1}(p) - K_C^{-1}(p) \right) \right| \tag{4.7}$$

$$\leq \sup_{\mathbf{x} \in \mathbb{R}^d} \left| \hat{C}(\mathbf{x}) - C(\mathbf{x}) \right| + \left| \hat{K}_C^{-1}(p) - K_C^{-1}(p) \right|. \tag{4.8}$$

Recall that both the empirical copula and the Kendall distribution function are strongly consistent (Deheuvels, 1979, 1980; Ghoudi and Rémillard, 1998) and that $K_C^{-1}(p)$ is strongly consistent when $K_C(p)$ is continuous and strictly monotone (cf. Section 2.3). Hence, the term above converges to 0 for $n \to \infty$ and we expect the approach to be a valid extension of Mammen and Polonik (2013). Furthermore, the approach is easy to implement and is computationally very efficient.

Figure 4.1 shows an exemplary application of the approach on a bivariate Clayton copula sample (left panel) and a bivariate Gumbel copula sample (right panel) of size 100 each, where we have bootstrapped $1,000$ times. The blue line depicts the boundary of the estimated multivariate quantile $\hat{S}_p(\hat{C})$, whereas the gray line depicts the theoretical boundary of S_p with $p = 0.9$.

The orange and green lines are the boundaries of the sets $\hat{S}_\ell$ and $\hat{S}_u$ for $\alpha = 0.1$ and $\alpha = 0.05$, respectively.

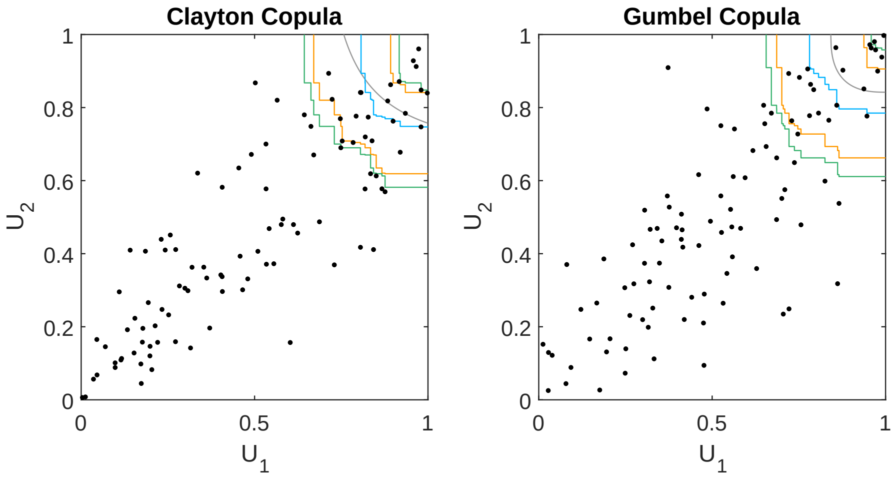

Figure 4.1: Example of the extended Mammen and Polonik (2013) approach. The blue line shows the estimated boundary of the multivariate quantile, the gray line shows the theoretical multivariate quantile boundary for $p = 0.9$. The orange and green lines depict the confidence regions for $\alpha = 0.1$ and $\alpha = 0.05$, respectively. Left Panel: Clayton copula sample with $\theta = 3$ (i.e., Kendall's $\tau = 0.6$) and $n = 100$. Right Panel: Gumbel copula sample with $\theta = 2.5$ (i.e., Kendall's $\tau = 0.6$) and $n = 100$ (Coblenz et al., 2018a).

4.3.2 Approach extended from Chen et al. (2017)

The approach by Chen et al. (2017) is based on the Hausdorff distance δ_H (see Section 2.5) between an estimated level set and its theoretical counterpart. Note that in the following we present Method 1 in Chen et al. (2017). The second method in Chen et al. (2017) is very similar to the approach in the previous section. Additionally, Chen et al. (2017) state that the approach by Mammen and Polonik (2013) should yield better results compared to their second method.

Chen et al. (2017) focus on confidence regions for density level sets of the form $L = \{\mathbf{x} \in \mathbb{R}^d : f_h(\mathbf{x}) = \lambda\}$, where f_h is the convolution of a density f and a kernel $\mathbf{K}$. Given a sample, L can be estimated with a kernel density estimator $\hat{f}_h$ of f_h as $\hat{L} = \{\mathbf{x} \in \mathbb{R}^d : \hat{f}_h(\mathbf{x}) = \lambda\}$. Let W be the Hausdorff distance between L and $\hat{L}$, i.e., $W = \delta_H(L, \hat{L})$. The confidence region of $\hat{L}$ is then

$$\hat{R} = \bigcup_{\mathbf{x} \in \hat{L}} \{\mathbf{y} : ||\mathbf{x} - \mathbf{y}|| \leq w_n\}, \tag{4.9}$$

where w_n is the $1 - \alpha$ quantile of W. This amounts to drawing a sphere of radius w_n around each point in $\hat{L}$. It can be shown that

$$\mathbb{P}(L \subset \hat{R}) \geq 1 - \alpha, \tag{4.10}$$

where $1 - \alpha$ is the confidence level. Since the distribution of W is unknown, bootstrapping is suggested by Chen et al. (2017).

Similar to the approach in the previous section, the method by Chen et al. (2017) is not directly applicable to multivariate quantiles. However, not only the estimation of K_C^{-1} has to be considered, but also that copulas are distribution functions and not densities. Additionally, the method of Chen et al. (2017) assumes an unbounded domain which is not the case in a copula context. Again, let $\mathbf{U}_1, \dots, \mathbf{U}_n$ be a d-dimensional copula sample. We extend the approach in Algorithm 2.

Algorithm 2 Extension of Chen et al. (2017).

```
1: Choose the level $p$ and the confidence level $1 - \alpha$.
2: Estimate $K_C^{-1}(p)$ on $\mathbf{U}_1, \dots, \mathbf{U}_n$ according to Equation (2.18).
3: Estimate $S_p$ based on the kernel density estimate $\hat{C}_h$ (cf. Equation (2.15))
   on $\mathbf{U}_1, \dots, \mathbf{U}_n$ using Silverman's rule of thumb (Silverman, 1986) and a
   Gaussian kernel according to Equation (4.3).
4: Draw $n_{bs}$ bootstrap samples $\mathbf{U}_1^*, \dots, \mathbf{U}_n^*$. Repeat Step 2 and Step 3 on
   each of these.
5: Determine the Hausdorff distance $\delta_H$ between $\hat{S}_p(\hat{C}_h)$ of the original sample
   and each bootstrapped $\hat{S}_p^i(\hat{C}_h^i), i = 1, \dots, n_{bs}$, where $\hat{C}_h^i$ is the kernel
   density estimated copula on bootstrap sample $i$.
6: Estimate $w_n$ as the empirical $1 - \alpha$-quantile of $\hat{W} =
   (\delta_H(\hat{S}_p, \hat{S}_p^1), \dots, \delta_H(\hat{S}_p, \hat{S}_p^{n_{bs}}))$.
7: The confidence region is $\bigcup_{\mathbf{x} \in \hat{S}_p(\hat{C}_h)} B(\mathbf{x}, \hat{w}_n)$, where $B(\mathbf{x}, \hat{w}_n) = \{\mathbf{y} : ||\mathbf{x} -
   \mathbf{y}|| \leq \hat{w}_n\}$.
```

This method is computationally more demanding than the approach by Mammen and Polonik (2013). Note that issues caused by the bounded copula domain are circumvented by using the Probit transformation in the kernel based estimator $\hat{C}_h$ (cf. Equation (2.15)). Thus, standard kernel density estimation can be used, which is readily available in pertinent statistical software. The result of Step 3 in Algorithm 2 is a set of finitely many points

$\mathbf{x} \in \mathbb{R}^d$ which make up the boundary of the multivariate quantile $\hat{S}_p(\hat{C}_h)$ in the space $\mathbb{R}^d$. Recall from the previous chapter that by using a Gaussian kernel and Silverman's rule of thumb (Silverman, 1986) for the bandwidth h, a point $\mathbf{x}$ on the boundary can be transformed back to the copula domain $[0, 1]^d$ by

$$\Phi\left(\frac{\mathbf{x}}{\sqrt{1+h^2}}\right) \to \mathbf{u}, \tag{4.11}$$

where Φ is the univariate standard normal CDF applied componentwise. This allows us not only to compute the Hausdorff distance in Step 5 on the bounded copula domain but also to construct subsequently the confidence regions in $[0, 1]^d$. Note that we interpolate the points on the multivariate quantile boundary of the kernel density estimation linearly, which introduces a small numerical imprecision to the Hausdorff distance calculation. By incorporating the estimation of $K_C^{-1}(p)$ in Step 4, we account for its estimation uncertainty simultaneously.

Figure 4.2 shows exemplary confidence region estimation results on a bivariate Clayton copula sample (left panel) and bivariate Gumbel copula sample (right panel) of size 100 each. We have used $1,000$ bootstrap samples for each plot. The color coding is as in Figure 4.1 above: The blue line depicts the boundary of the estimated multivariate quantile $\hat{S}_p(\hat{C}_h)$, whereas the gray line depicts the theoretical boundary of S_p for $p = 0.9$. The orange and green lines are the boundaries of the confidence region $\bigcup_{\mathbf{x}\in\hat{S}_p(\hat{C}_h)} B(\mathbf{x}, \hat{w}_n)$ for $\alpha = 0.1$ and $\alpha = 0.05$, respectively.

Note that we have extended the approaches by Mammen and Polonik (2013) and Chen et al. (2017) in several aspects to make them applicable for the estimation of multivariate quantiles. It is not quite clear, whether they retain their statistical properties and how they behave on small sample sizes. In particular, it is interesting to investigate the proposed confidence level $1 - \alpha$ via coverage probabilities. We do this with a simulation study in the next section.

4.4 Simulation Study

We investigate in a simulation study whether the extended approaches introduced in Sections 4.3.1 and 4.3.2 hold their proposed confidence level $1 - \alpha$ via coverage probabilities. In particular, we focus on small sample sizes as they are found in hydrology applications. For both approaches we consider the same simulation settings. We simulate samples of sizes $n = 100, 200$ from Gauss, Clayton, and Gumbel copulas, where we restrict ourselves to the bivariate case. The Gauss copula has a parameter ρ corresponding to a

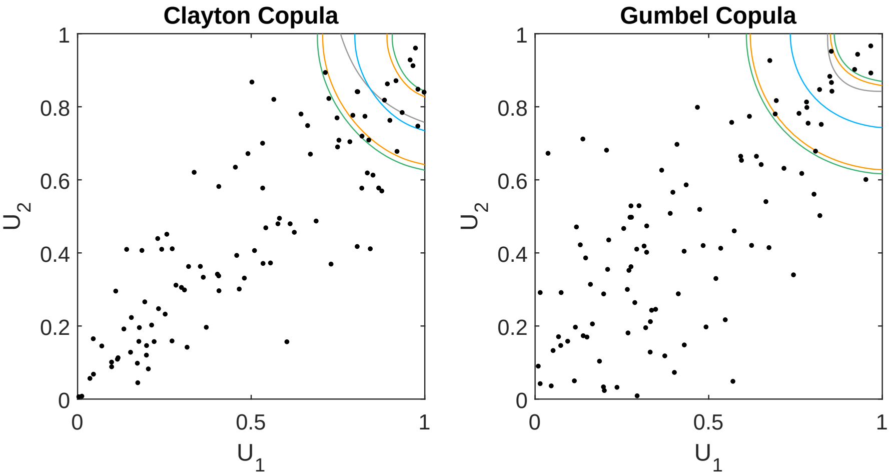

Figure 4.2: Example of the extended Chen et al. (2017) approach. The blue line shows the estimated boundary of the multivariate quantile based on a kernel density estimated copula, the gray line shows the theoretical multivariate quantile boundary for $p = 0.9$. The orange and green lines depict the boundaries of the confidence region for $\alpha = 0.1$ and $\alpha = 0.05$, respectively. Left Panel: Clayton copula sample with $\theta = 3$ (i.e., Kendall's $\tau = 0.6$) and $n = 100$. Right Panel: Gumbel copula sample with $\theta = 2.5$ (i.e., Kendall's $\tau = 0.6$) and $n = 100$ (Coblenz et al., 2018a).

Kendall's τ of $-0.8, -0.5, 0, 0.5, 0.8$, whereas for the Clayton and the Gumbel copula settings the parameters correspond to a Kendall's τ of $0.3, 0.5, 0.8$. Note that in the Gauss case a Kendall's τ of 0 corresponds to independence.

For each setting, we estimate confidence regions for the $p = 0.1, 0.5, 0.9$ multivariate quantile to get a better picture of the performance on the whole copula domain. Confidence regions are estimated at the 90% and 95% confidence levels. For this, we use $1,000$ bootstraps for the Mammen and Polonik (2013) approach and 200 bootstraps for the Chen et al. (2017) approach, due to the high computation times of the latter. Each simulation setting is repeated $1,000$ times to obtain reliable results. The coverage probability is calculated by checking whether the theoretical multivariate quantile boundary lies within the estimated confidence region in each simulation run. For example, Figure 4.1 and Figure 4.2 show cases, where the theoretical multivariate quantile is covered by the confidence region.

The coverage probabilities for the extended Mammen and Polonik (2013) approach can be found in Table 4.1. The first sanity check which can be made is that the 95% confidence region exhibits higher coverage probabilities than the 90% confidence region, which is the case throughout. Most of the settings for the 10% and 50% multivariate quantiles show more conservative coverage probabilities than the respective confidence level would suggest. In contrast to that, particularly the negative dependence settings for $p = 0.9$ exhibit too low coverage probabilities. Too high and too low coverage probabilities could be due to the estimation uncertainty of K^{-1} and the bounded copula domain. The results over the different sample sizes are very similar. We conclude from this that the estimator works quite well for small sample sizes. Overall, the results for the Mammen and Polonik (2013) approach are reasonable.

		$1-\alpha = 90\%$						$1-\alpha = 95\%$					
		$n = 100$			$n = 200$			$n = 100$			$n = 200$		
Copula	τ \ p	0.1	0.5	0.9	0.1	0.5	0.9	0.1	0.5	0.9	0.1	0.5	0.9
Gauss	-0.8	100	97.7	82.6	100	97.8	80.7	100	99.2	91.9	100	98.8	89.1
	-0.5	100	93.3	82.0	100	93.1	84.3	100	96.3	90.8	100	97.2	92.3
	0	99.4	89.6	85.2	99.4	89.8	86.0	99.8	94.4	93.1	100	95.1	93.0
	0.5	97.1	92.0	91.3	98.1	89.6	89.6	98.7	96.1	96.1	99.5	94.9	94.7
	0.8	96.5	92.7	94.5	96.7	90.6	93.0	98.0	96.6	97.2	98.8	95.5	97.4
Clayton	0.3	98.5	91.6	89.6	98.1	91.9	86.4	99.5	96.0	95.2	99.1	95.8	93.8
	0.5	97.1	91.1	90.2	97.4	88.9	88.7	98.7	94.9	95.6	98.9	94.2	95.5
	0.8	97.2	93.7	92.9	97.2	93.6	92.7	98.2	96.5	96.7	98.4	97.2	97.2
Gumbel	0.3	98.5	89.4	88.8	98.3	89.9	87.8	99.7	93.7	95.3	99.7	95.1	94.6
	0.5	98.1	91.6	91.7	98.2	89.7	90.3	99.4	96.0	96.6	99.1	94.3	95.4
	0.8	96.8	93.1	94.9	96.5	89.9	94.1	98.7	97.1	97.9	98.4	95.2	97.2

Table 4.1: Simulation results for the extended Mammen and Polonik (2013) approach. The overall coverage probabilities are reasonable.

The results of the extended Chen et al. (2017) approach can be found in Table 4.2. Most of the coverage probabilities are too low. In particular, confidence regions for high dependence seem to be problematic. In contrast to that, the results are reasonable for low to medium strong dependence, i.e., $\tau \in [-0.5, 0.5]$. This could be due to several effects. First, the bounded copula domain could be an issue. Second, the original approach by Chen et al. (2017) was developed for densities and not for copulas which are distribution functions. Also, the estimation of K_C^{-1} is present in the approach. However, we do not think that the latter plays an important role since the results for the Mammen and Polonik (2013) approach are good where estimation of K_C^{-1} is necessary, too. Finally, we calculate the coverage probabilities by checking whether the level curve at level $K_C^{-1}(p)$ of the underlying copula C is within the boundaries of the constructed confidence set since we are actually interested in the level curves of the copula C. In contrast to that, the approach of Chen et al. (2017) aims to estimate confidence regions for the level curves of a convolution of the copula C and the kernel $\mathbf{K}_h$, whereby a certain smoothness and limit behavior of the results is ensured. This could lead to the biased coverage probabilities in our case.

		$1-\alpha=90\%$						$1-\alpha=95\%$					
		$n=100$			$n=200$			$n=100$			$n=200$		
Copula	τ \ p	0.1	0.5	0.9	0.1	0.5	0.9	0.1	0.5	0.9	0.1	0.5	0.9
Gauss	-0.8	0.0	0.5	93.0	0.0	0.0	76.8	0.2	2.7	96.8	0.0	0.0	93.2
	-0.5	96.1	87.0	93.9	89.2	69.0	93.0	98.1	94.8	96.9	95.4	81.9	97.0
	0	86.1	91.5	94.7	89.2	91.9	93.1	91.5	95.9	97.8	94.3	96.2	97.1
	0.5	79.5	83.1	90.0	78.5	80.2	89.6	87.9	90.0	95.9	86.7	86.6	95.6
	0.8	71.3	63.0	79.7	62.8	51.7	70.9	80.1	74.6	89.0	74.3	65.5	81.6
Clayton	0.3	77.2	87.8	93.4	75.5	84.4	93.3	84.4	93.0	97.4	85.6	91.7	96.7
	0.5	69.8	77.5	94.0	65.3	73.4	91.9	79.5	85.0	97.3	76.3	82.9	96.3
	0.8	58.1	51.7	87.6	43.8	37.0	87.8	71.0	65.4	94.0	57.7	50.5	92.9
Gumbel	0.3	84.7	89.3	91.4	88.4	88.1	90.2	91.0	94.5	96.7	93.5	93.8	95.2
	0.5	83.8	82.8	88.3	82.8	80.6	86.4	90.1	90.2	94.7	90.5	89.4	92.9
	0.8	74.1	67.6	71.5	67.5	53.8	66.3	84.0	77.4	82.2	77.9	67.3	78.3

Table 4.2: Simulation results for extended Chen et al. (2017) approach. The overall results are mixed.

In conclusion, the simulation study shows a reasonable performance of the extended Mammen and Polonik (2013) method. On the other hand, results for the extended Chen et al. (2017) method are mixed. They are, however, reasonably precise for low to medium strong dependence. In summary, we advise practitioners to use the Mammen and Polonik (2013) approach for construction of multivariate quantile confidence regions. In the next section, we apply the introduced methods on a small hydrology related data set to gain further insights.

4.5 Application

We apply the two confidence region approaches on a small data set with a hydrology context. The data can be found in Yue et al. (1999) and it is also used in Section 3.5. Recall that it comprises 33 yearly maximum values of flood peak and flood volume of the Ashuapmushuan basin in Quebec, Canada. The observations were collected in the period 1963-1995. In a first step, we rank-transform the data to obtain the pseudo-observations in the copula domain $[0,1]^2$. Figure 4.3 shows a scatter plot of the original data and the rank-transformed data. The data exhibit positive dependence with a Kendall's τ of approximately 0.41.

In a second step, we estimate the 90% (i.e., $p = 0.9$) multivariate quantile with the two estimators $\hat{S}_p(\hat{C})$ and $\hat{S}_p(\hat{C}_h)$. The estimation results are shown in Figure 4.4. As can be seen, the two estimated boundaries nicely overlap. For comparison purposes we additionally estimate a parametric copula model. A Clayton copula with parameter $\theta \approx 1.4$ fits the data best

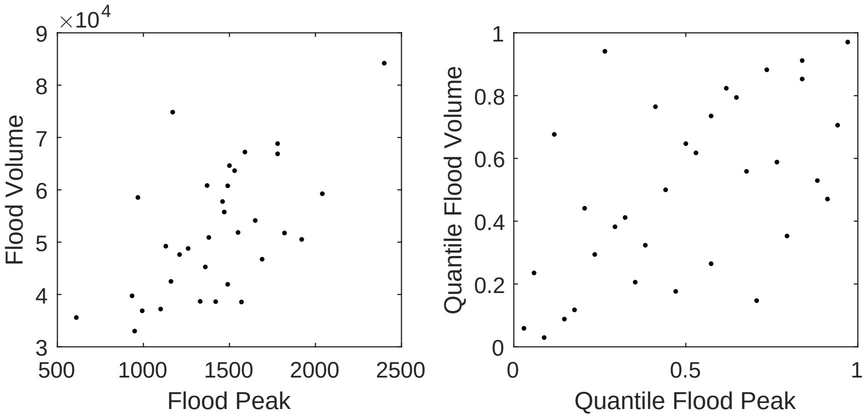

Figure 4.3: Left Panel: 33 flood peak and flood volume observations from the Ashuapmushuan basin in Quebec, Canada. Right Panel: The same data, but rank-transformed to the copula domain $[0, 1]^2$ (Coblenz et al., 2018a).

among Gumbel, Frank, Gauss, and t-copulas. The estimated boundary is shown in Figure 4.4 as a red line and is close to the nonparametric estimates.

In a third step, we apply the extended method of Mammen and Polonik (2013) as introduced in Section 4.3.1 to the data. The result of this can be seen in Figure 4.4. The orange and green step curves depict the confidence region boundaries for confidence levels 90% and 95%, respectively. Recall that the boundary of the 90% multivariate quantile partitions the copula domain into a set comprising 10% of the probability mass, which lies to the upper right of the boundary, and a set comprising 90% of the probability mass, which lies to the lower left of the boundary. Counting the points within the confidence region boundaries, we obtain between 33% and 3% of the points for the 90% confidence region and between 36% and 3% of the points for the 95% confidence region. Thus, the confidence regions seem wide which has to be related to the small sample size, though.

Next, we also apply the extended method of Chen et al. (2017) as introduced in Section 4.3.2. Figure 4.4 shows the results. The orange and green smooth lines depict the confidence region boundaries for confidence levels 90% and 95%, respectively. With the same calculations as above, both the 90% confidence region and the 95% confidence region enclose between 21% and 3% of the points. Thus, the confidence regions are tighter than those of the Mammen and Polonik (2013) method. This can also be seen in Figure 4.4. Clearly, the approach of Chen et al. (2017) gives a tighter confidence region on the lower end, whereas the two approaches give similar results on the upper end. This has to be put in light of the simulation study which

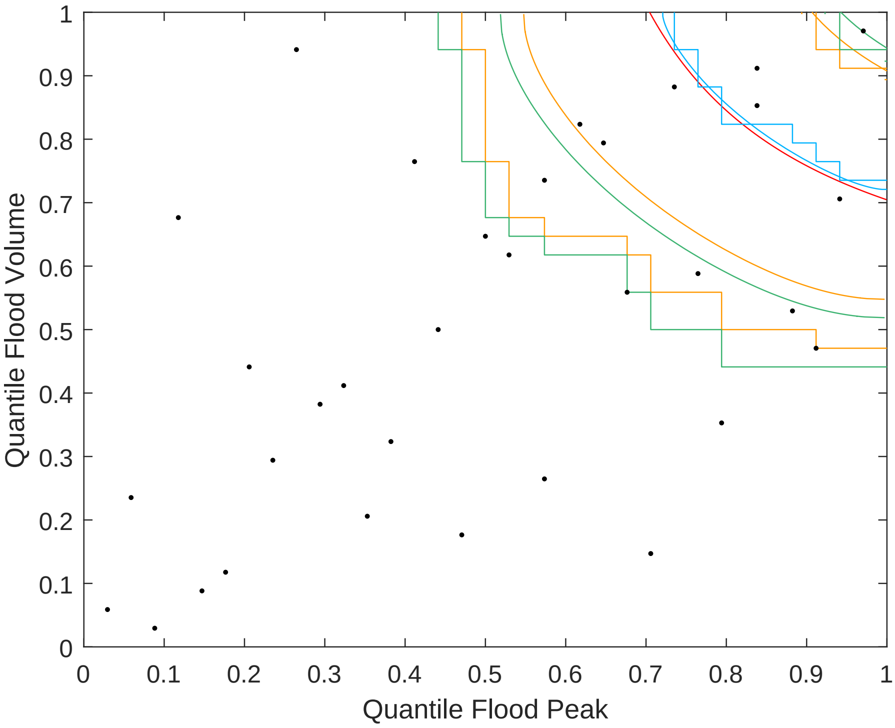

Figure 4.4: Combined estimation results of the multivariate quantile with both confidence region methods. Boundaries of the estimated multivariate quantiles $\hat{S}_p(\hat{C})$ and $\hat{S}_p(\hat{C}_h)$ are shown as a blue step curve and a blue smooth curve, respectively. The red curve refers to the boundary of the multivariate quantile of a Clayton copula that is parametrically estimated on the data. Confidence regions of $\hat{S}_p(\hat{C})$ for confidence levels 90% and 95% are depicted as orange and green step curves, whereas the confidence regions of $\hat{S}_p(\hat{C}_h)$ are shown as orange and green lines for the respective confidence levels (Coblenz et al., 2018a).

shows too liberal coverage probabilities for the method of Chen et al. (2017) in the considered case $p = 0.9$ and moderate positive dependence.

Furthermore, we analyze the secondary return period as defined in Salvadori et al. (2011). The estimated secondary return period given by the multivariate quantile is $\frac{1}{1-\hat{K}_C(P)} = 10$ years. For the Mammen and Polonik (2013) approach the confidence regions suggest a secondary return period between 3 and 33 years and between 2.75 and 33 years for the 90% and 95% confidence levels, respectively. The confidence regions of the Chen et al.

(2017) approach suggest a secondary return period between 4.7 and 33 years and between 4.1 and 33 years for the 90% and 95% confidence levels, respectively. Thus, the confidence regions can also be used to assess the precision of the implied secondary return period of the multivariate quantile.

Finally, we want to stress again the advantages of having confidence regions for multivariate quantiles in a hydrology context. Not only give confidence regions a statistical insight into the estimation uncertainty present, e.g., Figure 4.4 shows that these are very wide and more data would be needed for a reliable estimate of the multivariate quantile, but also they are helpful to the design of infrastructures. Since the true multivariate quantile boundary lies within the confidence region boundaries at the specified confidence level, the points within the confidence region should be considered when planning, e.g., new dams. In particular, a point from within the region between the lower boundary of the confidence region and the multivariate quantile boundary could actually be a point with (true) secondary return period of 10 years and thus would be rarer than the estimated multivariate quantile suggests. Conversely, a point from within the region between the upper boundary of the confidence regions and the multivariate quantile boundary could have a lower (true) secondary return period and thus would occur more often than might be expected from considering the estimated multivariate quantile boundary only.

4.6 Concluding Remarks

We extend the two approaches by Mammen and Polonik (2013) and Chen et al. (2017) for construction of confidence regions for level sets to make them applicable in a multivariate quantile context. This involves incorporating the estimation of the quantile level via the inverse Kendall distribution function K_C^{-1} and also adjusting for the bounded copula domain.

The simulation study shows reasonable coverage probabilities for the extended Mammen and Polonik (2013) method. Some of the coverage probabilities are too conservative. However, in particular for negative dependence and high quantile levels the approach yields too low coverage probabilities. On the other hand, the extended Chen et al. (2017) method shows mixed results. Overall, the coverage probabilities are too liberal. However, they show a reasonable precision for low to medium strong dependence. An application on a small hydrology-related data set illustrated some further aspects of the approaches.

On a final note, we want to point out that we tried to keep the extension of the methods as simple as possible. The approaches could be extended in several further ways. For example, a smoothed bootstrap along the lines of Section 3.3.3 can be incorporated into the analysis. We leave this for future research, though. We hope that practitioners in hydrology and other fields find the considered approaches helpful and easy to apply to their problems at hand.

In the next chapter, we investigate level sets, and in particular level curves, of bivariate copulas more closely. It turns out, that the level curve lengths of copulas contain information with respect to the dependence structure and we study this relationship.

Chapter 5

On the Length of Copula Level Curves

This chapter is based on Coblenz et al. (2018c).

5.1 Introduction

The surface of copulas, and in particular the volume determined by the surface, is informative with respect to dependence. It is an integral part of well-known concordance measures, such as Spearman's ρ_S (cf. Equation (2.21)). It therefore seems natural to study some further aspects of copula surfaces, such as the length of the contour lines of copula surfaces.

The importance of level curves and level sets of copulas has already been recognized in several areas such as risk management, hydrology, and coastal engineering. Cousin and Di Bernardino (2013) and Di Bernardino and Rullière (2013) use level curve approaches for financial risk management. Level sets and level curves in the form of multivariate quantiles are also relevant in risk analysis in hydrology and coastal engineering, see, e.g., Salvadori et al. (2013, 2016) and the previous two chapters.

We establish a function $L_C(t)$ which we call the length profile of the copula C. The length profile maps each level $t \in [0, 1]$ to the length of the respective (lower) t level curve of the copula. Some basic properties of the length profile, such as continuity, differentiability, and monotonicity with respect to t, are investigated. Based on the length profile, we define the so-called length measure ℓ_C, which is the average length of the level curves. Although ℓ_C is a measure of association, it does not fulfill all the properties of concordance measures.

Over the course of our investigation, we obtain some interesting theoretical and mathematically relevant results on the general class of bivariate copulas. Complementing the latter, we establish results for some specific copula classes, such as Archimedean copulas and shuffles of M. Additionally, a nice closed-form formula for ℓ_C for mutually completely dependent copulas is derived. The investigation is accompanied by numerous examples that show some important features of the length profile and the length measure.

The chapter is structured as follows: The next section introduces some specific notation and comprises some further preliminaries on copulas. Section 5.3 deals with the so-called radius-vector function and its connection to copula level curves. The length profile is established in Section 5.4. After giving a definition of the length profile, we investigate some of its properties more closely, which are then used when introducing the length measure in Section 5.5. Subsequently, we check whether the length measure is a concordance measure. Some concluding remarks complete the chapter. The proofs are either deferred to the Appendix or can be found in the accompanying paper Coblenz et al. (2018c).

5.2 Further Notation and Preliminaries

This section introduces some specific notation used in this chapter and comprises some additional preliminaries on copulas. Let $\mathcal{K}([0,1]^2) = \mathcal{K}$ denote the family of all non-empty compact subsets of $[0,1]^2$. For the (topological) interior of a set A we write $\text{int}(A)$. To simplify notation, given $E, F \in \mathcal{K}$ we write $d(E,F) = \max_{a\in E} \min_{b\in F} \|b-a\|$ as well as $d(a,F) = \min_{b\in F} \|b-a\|$. This implies $\delta_H(E,F) = \max\{d(E,F), d(F,E)\}$, where δ_H is the Hausdorff distance (cf. Section 2.5). The one-dimensional Hausdorff measure (Federer, 1996) is denoted by H_1, the d-dimensional Lebesgue measure is denoted by λ_d, where we drop the index in the case $d = 1$.

Throughout this chapter, we deal with 2-dimensional copulas exclusively. We denote the family of all two-dimensional copulas by $\mathcal{C}$. Furthermore, $\Pi \in \mathcal{C}$ denotes the independence copula, and $W \in \mathcal{C}$ and $M \in \mathcal{C}$ denote the lower and upper Fréchet-Hoeffding bound, respectively, also cf. Section 2.1. The upper Fréchet-Hoeffding bound will also be referred to as minimum copula. A copula C is called exchangeable if $C(u,v) = C(v,u)$ for any $u, v \in [0,1]$. Otherwise, the copula is called non-exchangeable. From the definition of a copula it can be shown that the non-exchangeable counterpart $C^\top(u,v) := C(v,u)$ is also a copula. $C^\top$ is commonly known as the transpose of C. An important family of exchangeable copulas is the class of Archimedean copulas, see Section 2.1.

We also make use of the notion of a shuffle of M, or briefly shuffle. Every straight shuffle $C \in \mathcal{C}$ of the minimum copula can be expressed in terms of a permutation $\boldsymbol{\pi} \in \sigma_n$, where σ_n denotes all bijections on $\{1, \ldots, n\}$ for some $n \in \mathbb{N}$, and a vector $\boldsymbol{s} = (s_i)_{i=0}^n$, where $0 = s_0 < \ldots < s_n = 1$. If $\boldsymbol{s} = (i/n)_{i=0}^n$, the corresponding copula C_n is given by $C_n(u, v) = \lambda([0, u] \cap h^{-1}([0, v]))$, where $h(u) := u - i/n + (\pi_{i+1} - 1)/n$ for $u \in [i/n, (i + 1)/n]$, $i = 0, ..., n - 1$. Thus, the support of a straight shuffle is a permutation of the support of M vertically cut into strips. We remark that the term shuffle is more general and comprises structures that are not straight, i.e., where the strips of the minimum copula can be flipped around their vertical symmetry axis. In case all of the strips are flipped, the shuffle is also called flipped shuffle or shuffle of W. A more detailed discussion of shuffles is given in Nelsen (2006), Chapter 3.2.3, and Trutschnig and Fernández Sánchez (2013).

Shuffles of M belong to the class of mutually completely dependent copulas, which we denote by $\mathcal{C}_{cd}$. Recall that a copula C is called mutually completely dependent if there exists a λ-preserving bijection $h : [0, 1] \to [0, 1]$, such that C concentrates all mass on the graph of h, i.e., if we have $\mu_C(\text{graph}(h)) = 1$, where μ_C is the doubly stochastic measure corresponding to the copula C. The class of all λ-preserving bijections on $[0, 1]$ is denoted by $\mathcal{T}$. For equivalent definitions of (mutual) complete dependence we refer the reader to Trutschnig (2012). Furthermore, note that $\mathcal{C}_{cd}$ is dense in $(\mathcal{C}, d_\infty)$ and that empirical checkmin copulas (M-interpolations of the empirical copula, see Mikusiński and Taylor (2010)) are mutually completely dependent.

For every copula $C \in \mathcal{C}$ there exists a Markov kernel (i.e., a regular conditional distribution) $K_C^* : [0, 1] \times \mathcal{B}([0, 1]) \to [0, 1]$ fulfilling

$$\int_{[0,1]} K_C^*(x, G_x)\, d\lambda(x) = \mu_C(G), \tag{5.1}$$

for every $G \in \mathcal{B}([0, 1]^2)$, where $G_x := \{y \in [0, 1] : (x, y) \in G\}$ denotes the x-section of $G \in \mathcal{B}([0, 1]^2)$ for every $x \in [0, 1]$. In particular,

$$\int_{[0,1]} K_C^*(x, F)\, d\lambda(x) = \lambda(F) \tag{5.2}$$

for every $F \in \mathcal{B}([0, 1])$ (see Kallenberg (1997) for more details). We will refer to K_C^* simply as Markov kernel of C.

5.3 Radius-Vector Functions and Copula Level Curves

The t level set of a copula $C \in \mathcal{C}$ is the collection of points $(u, v) \in [0, 1]^2$ for which $C(u, v) = t$ holds. Note that, in general, the t level set of a copula can contain a plateau, i.e., $\mathrm{int}(C^{-1}(\{t\})) \neq \emptyset$, and thus, there are cases where it is not a curve. As a consequence, since we deal with the length of copula level curves, we need a more general notion of level curve. For every $C \in \mathcal{C}$ and $t \in [0, 1]$ the lower t level set is given by

$$[C]_t = \big\{(u, v) \in [0, 1]^2 : C(u, v) \geq t\big\}. \tag{5.3}$$

The *radius-vector function* R_C of C (Trutschnig, 2012; Molchanov, 2005) is defined as

$$R_C(t, \varphi) := R_t^C(\varphi) := \max\{s \geq 0 : (1, 1) + s(\cos(\varphi), \sin(\varphi)) \in [C]_t\}, \tag{5.4}$$

for every $t \in [0, 1]$ and $\varphi \in I := [\pi, 3\pi/2]$. The radius-vector function R_C allows to parametrize the (lower) t level curve $\Gamma_C(t)$ of C via

$$\Gamma_C(t) = \big\{\gamma_t^C(\varphi) : \varphi \in I\big\} \in \mathcal{K}, \tag{5.5}$$

where $\gamma_t^C(\varphi) := (u_t(\varphi), v_t(\varphi)) = (1, 1) + R_C(t, \varphi)(\cos(\varphi), \sin(\varphi))$. Hence, $\Gamma_C(t)$ is the boundary of $[C]_t$ in $(0, 1)^2$.

Some basic properties of R_C are collected in the following lemma from Trutschnig (2012).

Lemma 5 (Trutschnig (2012))**.** *For every $C \in \mathcal{C}$ the radius-vector function R_C has the following properties:*

(a) For every fixed $t \in [0, 1]$ the function $\varphi \mapsto R_C(t, \varphi)$ is continuous.

(b) For every fixed $\varphi \in [\pi, 3\pi/2]$ the function $t \mapsto R_C(t, \varphi)$ is left-continuous and strictly decreasing.

(c) The function $\varphi \mapsto u_t(\varphi)$ is monotonically non-decreasing and continuous. The function $\varphi \mapsto v_t(\varphi)$ is monotonically non-increasing and continuous.

(d) For every compact subinterval $J \subset (\pi, 3\pi/2)$ there exists a constant L such that all functions $\varphi \mapsto R_C(t, \varphi)$ are Lipschitz continuous with common Lipschitz constant L on J.

(e) If $(t_n)_{n\in\mathbb{N}}$ is monotonically decreasing with limit $t \in [0, 1)$ and $R_C(t_n, \varphi)$ converges to $R_C(t, \varphi)$ for every φ then $\lim_{n\to\infty} \delta_H([C]_{t_n}, [C]_t) = 0$ holds.

The following results show that uniform/pointwise convergence of copulas with $\mathrm{int}(C^{-1}(\{t\})) = \emptyset$ implies uniform convergence of the radius-vector function R_t^C. Moreover, along the same lines of Trutschnig (2012) and Fernández Sánchez and Trutschnig (2015), we prove that this also implies Hausdorff convergence of t level curves for all but at most countably many t in $[0, 1]$. Note that $\mathrm{int}(C^{-1}(\{t\})) = \emptyset$ is not equivalent to C being strictly increasing. It is easy to show that a strictly increasing copula indeed contains no plateaus in its level sets. However, the minimum copula is a counter-example for the reverse statement. Thus, by $\mathrm{int}(C^{-1}(\{t\})) = \emptyset$ a weaker restriction is imposed on the copula.

We begin with some useful properties of the mapping $\Gamma_C : [0, 1] \to \mathcal{K}$. The proofs of the following lemma and the two theorems thereafter can be found in Coblenz et al. (2018c).

Lemma 6. $\Gamma_C : [0, 1] \to \mathcal{K}$ *has the following properties:*

(1.) Γ_C *is left continuous on* $(0, 1]$.

(2.) Γ_C *has a discontinuity in* $t \in (0, 1)$ *if and only if* $\mathrm{int}(C^{-1}(\{t\})) \neq \emptyset$.

The following theorem shows that uniform/pointwise convergence of copulas implies convergence of their level curves.

Theorem 7. *Suppose that* $(C_n)_{n\in\mathbb{N}}$ *is a sequence of copulas converging pointwise to* $C \in \mathcal{C}$ *and that* $t \in (0, 1)$ *fulfills* $\mathrm{int}(C^{-1}(\{t\})) = \emptyset$. *Then,*

$$\lim_{n\to\infty} \delta_H(\Gamma_{C_n}(t), \Gamma_C(t)) = 0 \tag{5.6}$$

holds.

Theorem 8. *Suppose that* $(C_n)_{n\in\mathbb{N}}$ *is a sequence of copulas converging pointwise to* $C \in \mathcal{C}$. *Then for every* $t \in [0, 1]$ *with* $\mathrm{int}(C^{-1}(\{t\})) = \emptyset$ *the corresponding radius-vector functions converge uniformly, i.e.,*

$$\lim_{n\to\infty} \|R_t^{C_n} - R_t^C\|_\infty = 0. \tag{5.7}$$

In the next section, we investigate the length of copula level curves more closely and introduce the length profile.

5.4 The Length Profile

Based on $\Gamma_C(t)$ and the one-dimensional Hausdorff measure H_1, the length of a copula level curve can be calculated as $H_1(\Gamma_C(t))$. For specific copulas, the length of a copula level curve can be stated in terms of the usual formula for the length of a curve.

Theorem 9. *Let $C \in \mathcal{C}$ be continuously differentiable on $[0,1]^2$ and assume $\partial C(u,v)/\partial v > 0$ for all $(u,v) \in (0,1]^2$. Further, let $t \in (0,1)$. Then, the length of the level curve of C at level t is*

$$H_1(\Gamma_C(t)) = \int_t^1 \sqrt{1 + [v_t'(u)]^2} du = \int_t^1 \sqrt{1 + \left[\frac{\frac{\partial C}{\partial u}(u, v_t(u))}{\frac{\partial C}{\partial v}(u, v_t(u))} \right]^2} du, \quad (5.8)$$

where $v_t : [t,1] \mapsto [t,1]$ is a unique, continuously differentiable function fulfilling $C(u, v_t(u)) = t$, for any $u \in [t,1]$.

The proof of this theorem can be found in Appendix A.

Note that if only $\partial C(u,v)/\partial u > 0$, then, by analogue reasoning, there exists a function $u_t(v)$ instead for which Theorem 9 holds true. The condition $\partial C(u,v)/\partial v > 0$ (apart from being necessary to obtain a well-defined length of a level curve in Theorem 9) means that the conditional distribution of U given V contains 0 in its support.

Corollary 10. *Let $C(u,v) = \psi(\psi^{-1}(u) + \psi^{-1}(v))$ be a bivariate Archimedean copula with continuously differentiable generator ψ on $[0,1]$ and let $t \in (0,1)$. Then,*

$$H_1(\Gamma_C(t)) = \int_t^1 \sqrt{1 + \left[\frac{(\psi^{-1})'(u)}{(\psi^{-1})'(\psi(\psi^{-1}(t) - \psi^{-1}(u)))} \right]^2} du. \quad (5.9)$$

The result follows from straightforward calculations. Appendix B contains specific formulas for some well-known copulas, such as Gauss, t-, and Clayton copula.

We can now define the *length profile* of a copula.

Definition 6 (Length Profile)**.** *Let C be a copula. The function $L_C : [0,1] \to [0,\infty)$, defined by $L_C(t) := H_1(\Gamma_C(t))$, is called length profile of C.*

Remark 1.

1. *For the lower and upper Fréchet-Hoeffding bound simple geometric considerations yield $L_W(t) = \sqrt{2} \cdot (1-t)$ and $L_M(t) = 2 \cdot (1-t)$.*

2. *For the independence copula the length profile is given by*

$$L_\Pi(t) = \int_t^1 \sqrt{1 + t^2/u^4} du. \quad (5.10)$$

Just like the so-called K-plots (Genest and Rivest, 1993; Genest and Boies, 2003), the length profile might prove to be a useful tool for graphical inspection of dependence structures. Figure 5.1 shows length profiles of Gauss, Clayton, and Gumbel copulas for different values of Spearman's ρ_S. The black lines in each panel represent the length profiles of the upper Fréchet-Hoeffding bound (top), the independence copula (middle), and the lower Fréchet-Hoeffding bound (bottom). Clearly, different shapes of the length profiles are observable for different copulas. For example, the length profiles of the Clayton copula seem to go more directly into the upper left corner than the length profiles of Gauss and Gumbel copulas. A connection of the length profile (or more specifically of the derivative of the length profile) to lower tail dependence (see Section 2.4) might manifest itself there. Also, all the length profiles seem to be monotonically decreasing and convex, a property that does, however, not hold in general, as we are going to show subsequently.

In what follows we try to prove the results in a general setting, i.e., for all bivariate copulas $C \in \mathcal{C}$ – if some properties do not hold in the general case, we provide counterexamples. The first result concerns bounds for the length profile. The proof can be found in Appendix A.

Theorem 11. *Let $C \in \mathcal{C}$. Then, for any $t \in (0,1]$*

$$\sqrt{2}(1-t) \leq L_C(t) \leq 2(1-t). \tag{5.11}$$

In particular, note that from Theorem 11 it immediately follows that $0 \leq L_C(t) \leq 2$.

In the following, we study measurability, continuity, and differentiability of the length profile. Although the length profile is not continuous and not monotonically increasing in general (cf. Example 1), it is lower semi-continuous $\lambda-$almost everywhere. The proof can be found in Coblenz et al. (2018c).

Theorem 12. *Suppose that $\operatorname{int}(C^{-1}(\{t_0\})) = \emptyset$. Then, the length profile L_C is lower semi-continuous at the point $t_0 \in [0,1]$.*

From this we get the following result, implying that the length profile is integrable. The proof can be found in Coblenz et al. (2018c).

Theorem 13. *The length profile L_C is a Borel measurable function.*

The next two theorems show that for specific copulas the length profile is continuous and differentiable. Their proofs can be found in Appendix A. Example 1, however, shows that in general the length profile is neither continuous nor differentiable.

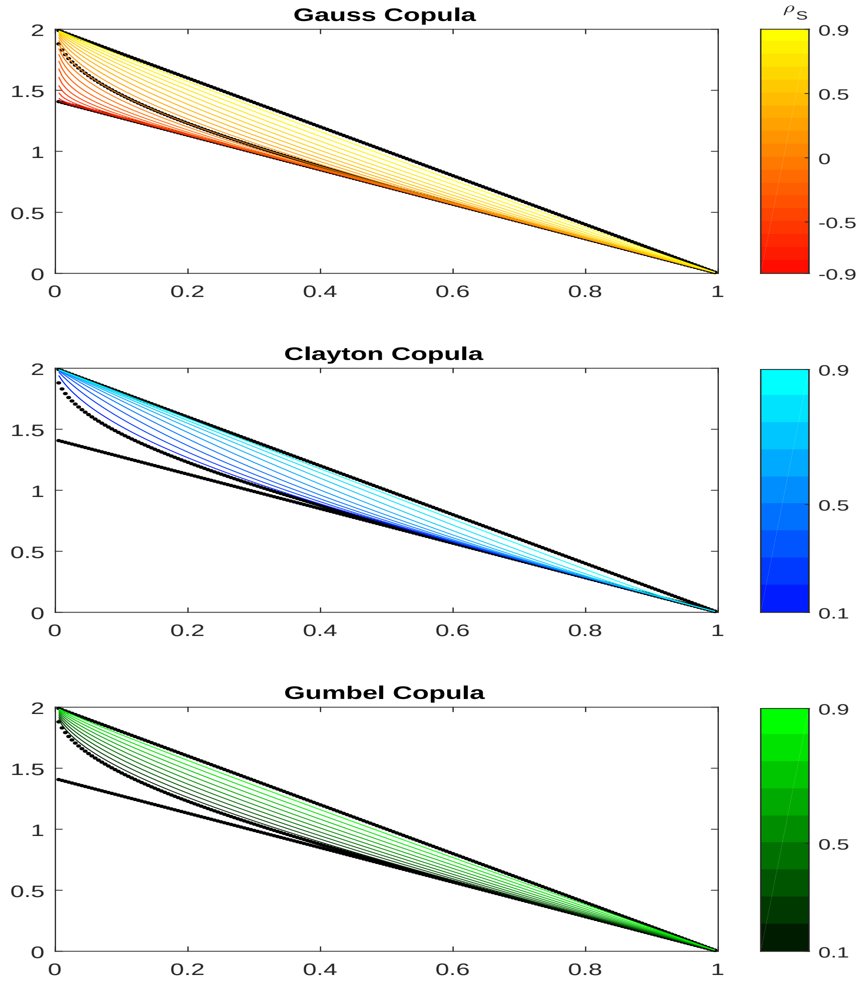

Figure 5.1: Length profiles of Gauss (top), Clayton (middle), and Gumbel (bottom) copulas for different values of Spearman's ρ_S: $\rho_S \in \{-0.9, \dots, 0.9\}$ (Gauss) and $\rho_S \in \{0.1, \dots, 0.9\}$ (Clayton and Gumbel). The black lines in each panel represent from top to bottom the length profiles of the upper Fréchet-Hoeffding bound, the independence copula, and the lower Fréchet-Hoeffding bound (Coblenz et al., 2018c).

Theorem 14. *Let $C \in \mathcal{C}$ be a continuously differentiable copula. Furthermore, assume that $\partial C(u, v)/\partial v > 0$, for $(u, v) \in (0, 1]^2$. Then L_C is a continuous function on $(0, 1]$.*

Theorem 15. *Let $C \in \mathcal{C}$ be a twice continuously differentiable copula with $\partial C(u, v)\partial v > 0$, for $(u, v) \in (0, 1]^2$. Also, let v be the unique, twice differentiable function, implicitly defined by $C(u, v(t, u)) = t$. Then L_C is differentiable on $(0, 1)$ and*

$$\frac{d}{dt} L_C(t) = \int_t^1 \frac{\frac{\partial}{\partial u} v(t, u) \frac{\partial^2}{\partial t \partial u} v(t, u)}{\sqrt{1 + \left[\frac{\partial}{\partial u} v(t, u)\right]^2}} du - \sqrt{1 + \left[\frac{\partial}{\partial u} v(t, t)\right]^2}, \tag{5.12}$$

where

$$\frac{\partial}{\partial u} v(t, u) = -\frac{\frac{\partial C(u, v(t, u))}{\partial u}}{\frac{\partial C(u, v(t, u))}{\partial v}}, \frac{\partial}{\partial u} v(t, t) = -\frac{1}{\frac{\partial C(t, 1)}{\partial v}},$$

$$\text{and } \frac{\partial^2}{\partial t \partial u} v(t, u) = -\frac{\frac{\partial^2 C(u, v(t, u))}{\partial t \partial u}}{\frac{\partial C(u, v(t, u))}{\partial t \partial v}}.$$

The following example shows that in general the length profile is neither continuous nor differentiable. Note that the example also shows that the length profile is not convex.

Example 1. Let $C \in \mathcal{C}$ be the shuffle of W as depicted in Figure 5.2, left panel. Clearly, the length profile of copula C (Figure 5.2, right panel) is not continuous. ■

Assuming convex level curves and the conditions of Theorem 9 it can be shown that the length profile is ordered.

Theorem 16. *Let C_1, C_2 be copulas with convex level curves that satisfy the assumptions of Theorem 9. If $C_1 \leq C_2$ on $[0, 1]^2$, then $L_{C_1} \leq L_{C_2}$ on $[0, 1]$.*

The proof can be found in Appendix A. Note that Theorem 16 applies to Archimedean copulas with continuously differentiable and strict generator. As before, the result does not hold in general: In the following example, two copulas are given that violate ordering of the length profile. The idea behind the example is to find two copulas, where one has step-shaped level sets and one has level sets that are almost straight lines, and alter them accordingly until the desired properties arise.

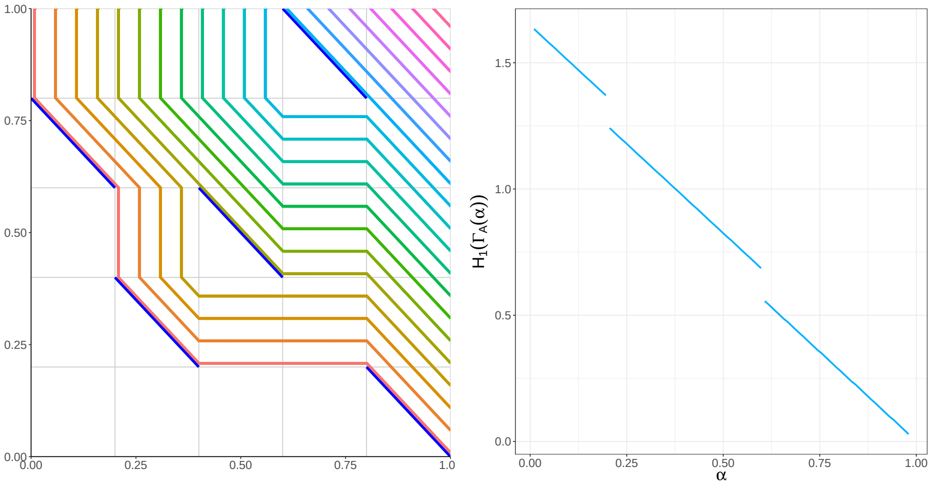

Figure 5.2: Left panel: Support (blue lines) and level curves of the copula. Right panel: Length profile of the copula (Coblenz et al., 2018c).

Example 2. Let C_1 and C_2 be the two copulas depicted in Figure 5.3 and Figure 5.4, left and middle panels, where the functional form of C_1 is given in Appendix C and C_2 is a shuffle of M. It holds that $C_1 \geq C_2$ on $[0, 1]^2$. Figure 5.4, right panel, shows the length profile of the two copulas. Clearly, $L_{C_1}(\alpha) \leq L_{C_2}(\alpha)$ for some $\alpha \in [0, 1]$. Thus, the length profile is not ordered in general. ■

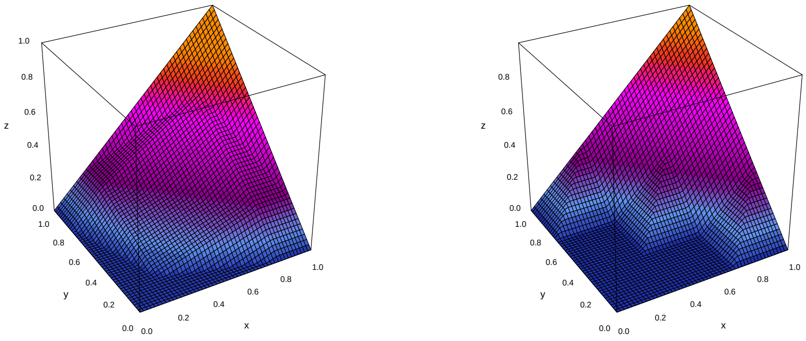

Figure 5.3: Left panel: Graph of copula C_1. Right panel: Graph of copula C_2 (Coblenz et al., 2018c).

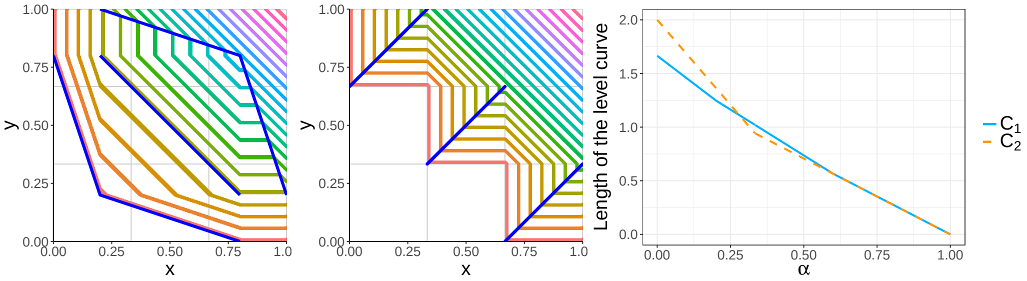

Figure 5.4: Level curves of copula C_1 (left panel) and copula C_2 (middle panel). Blue lines depict the supports of the two copulas. The right panel depicts the length profiles of C_1 and C_2 in blue and orange, respectively (Coblenz et al., 2018c).

The following theorem, which is proved in Coblenz et al. (2018c), implies that monotonicity holds for a dense subclass:

Theorem 17. *For every straight shuffle $C_n \in \mathcal{C}$ of the minimum copula with permutation $\boldsymbol{\pi} \in \sigma_n$, $n \in \mathbb{N}$, the length profile $L_{C_n}(t)$ is monotonically decreasing in t.*

The next example shows that, in general, the length profile is not monotonically decreasing.

Example 3. Consider the copula C_3, which is depicted in Figure 5.5 and Figure 5.6, left panel. The functional form of the copula is given in Appendix C.

Now consider the level curves for $t_1 = C_3(0.3, 0.5) = 11269/50000 = 0.22538$ and $t_2 = C_3(0.5, 0.5) = 5759/25000 = 0.23036$. The level curves for t_1 and t_2 are shown in Figure 5.6. The length of the level curves is $H_1(\Gamma_{C_3}(t_1)) = 1.29412$ and $H_1(\Gamma_{C_3}(t_2)) = 1.35251$, respectively. Hence, generally the length profile is not monotonically decreasing. ■

We close this section with a theorem which shows that the length profile does not characterize a copula in general. The proof can be found in Appendix A.

Theorem 18. *Let $C \in \mathcal{C}$ be a non-exchangeable copula and $C^\top$ be its transpose. Then, $L_C(t) = L_{C^\top}(t)$ for any $t \in [0, 1]$.*

Note that, in general, a copula is characterized by its level curves, cf. Theorem 1. However, Theorem 18 reveals that by aggregation of a level curve via its length some information about the underlying copula is lost. Based on the length profile, we introduce a measure of association which we will refer to as *length measure*.

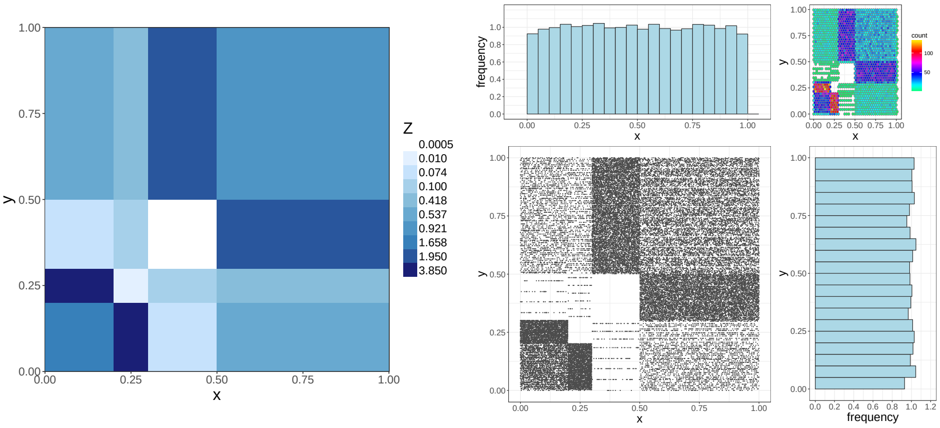

Figure 5.5: Density (left panel) and sample (right panel) of copula C_3 (Coblenz et al., 2018c).

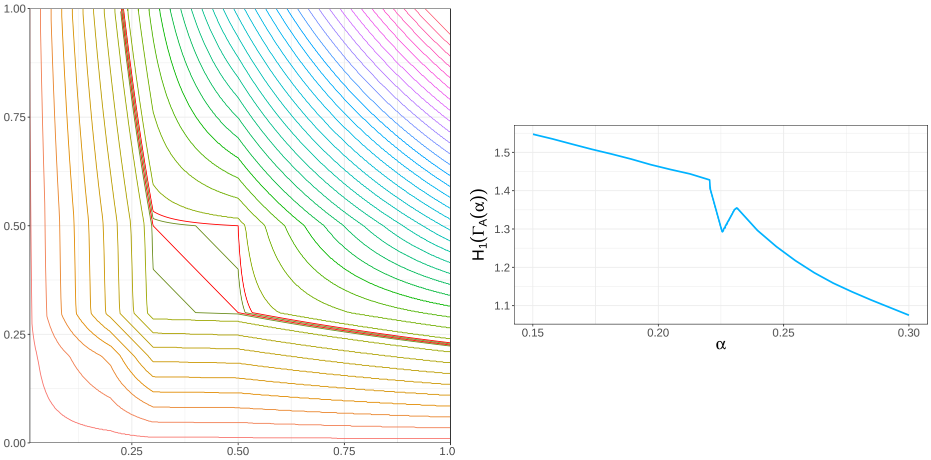

Figure 5.6: Left panel: Level curves of copula C_3. Red lines are the level curves at level t_1 and t_2. Right panel: The length profile of copula C_3 (Coblenz et al., 2018c).

5.5 The Length Measure ℓ_C

We first define the length measure and subsequently check whether the length measure is a concordance measure. Furthermore, we provide some insights with respect to mutually completely dependent copulas and empirical estimation.

5.5.1 Definition of ℓ_C

We begin with a definition of the length measure.

Definition 7 (Length Measure)**.** *For $C \in \mathcal{C}$*

$$\ell_C := \int_0^1 L_C(t)dt \tag{5.13}$$

is called length measure of the copula C.

Recall that the surface of copulas (and, to this end, the volume below the copula graph) comprises information on dependence and is hence a crucial part of well-known concordance measures. For example, Spearman's ρ_S is the difference of the volumes below the graphs of a given copula C and the independence copula (Nelsen, 2006)

$$\rho_S = 12 \int_0^1 \int_0^1 (C(u,v) - uv)dudv. \tag{5.14}$$

Furthermore, using Fubini's Theorem, Spearman's ρ_S can be represented as

$$\rho_S \propto \int_0^1 \lambda_2([C]_t)dt. \tag{5.15}$$

Since level curves form a part of the level sets $[C]_t$ and are a specific representation of the surface, the nice and intuitive geometric interpretation of ℓ_C justifies its use: It is the average of the level curve lengths of the copula C.

The following corollary, which is proved in Appendix A, provides bounds of the length measure.

Corollary 19. *For each $C \in \mathcal{C}$ it holds that*

$$\frac{\sqrt{2}}{2} \leq \ell_C \leq 1. \tag{5.16}$$

In particular, $\ell_W = \sqrt{2}/2$, $\ell_M = 1$ and $\ell_\Pi \approx 0.7652$.

Table 5.1 shows some values of ℓ_C for well-known copulas at different Kendall's τ. Since there are no closed-form solutions to the integral in (5.13), a numerical procedure was used to calculate the values. The results clearly show that ℓ_C cannot be transformed into Kendall's τ since the values between different copulas would have to match. Note that we have used a very high precision for the computations and seemingly equal values are indeed different in truncated decimal places.

Copula			$\ell_W = 0.707$			$\ell_\Pi = 0.765$			$\ell_M = 1$	
	τ	-0.9	-0.8	-0.7	-0.6	-0.5	-0.4	-0.3	-0.2	-0.1
Gauss	ℓ	0.707	0.708	0.709	0.712	0.716	0.722	0.730	0.740	0.751
Frank	ℓ	0.707	0.708	0.710	0.712	0.716	0.722	0.729	0.739	0.751
	τ	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Gauss	ℓ	0.781	0.800	0.820	0.842	0.866	0.891	0.918	0.945	0.972
Clayton	ℓ	0.783	0.803	0.824	0.847	0.870	0.895	0.920	0.946	0.973
Gumbel	ℓ	0.781	0.800	0.820	0.843	0.867	0.892	0.918	0.945	0.972
Frank	ℓ	0.782	0.800	0.821	0.843	0.867	0.892	0.918	0.945	0.973

Table 5.1: Values of ℓ_C for Gauss, Clayton, Gumbel, and Frank copula for different Kendall's τ.

Remark 2. *Note that there is a nonlinear transformation $f : \mathbb{R} \to \mathbb{R}$,*

$$f(x) = ax^b + c, \tag{5.17}$$

with $a \approx -0.1326$, $b \approx -8.015$ and $c \approx 1.133$, such that $f(\ell_W) = -1$, $f(\ell_\Pi) = 0$ and $f(\ell_M) = 1$.

Subsequently, we show that the length measure is a measure of association. In addition to that, we check the properties of a concordance measure.

5.5.2 Is ℓ_C a Measure of Association and a Concordance Measure?

An important tool for assessing dependence in the context of copulas are concordance measures, such as Kendall's τ and Spearman's ρ_S, see Section 2.4. In the following, we investigate whether the length measure is a concordance measure. We begin with a definition of concordance measures.

Definition 8 (Scarsini (1984)). *A numeric measure of association κ ($= \kappa_{X,Y} = \kappa_C$) between two continuous random variables X, Y with copula C is a measure of concordance if the following properties are satisfied:*

1. *κ is defined for every pair X, Y of continuous random variables;*
2. *$-1 \leq \kappa_{X,Y} \leq 1$, $\kappa_{X,X} = \kappa_M = 1$ and $\kappa_{X,-X} = \kappa_W = -1$;*
3. *$\kappa_{X,Y} = \kappa_{Y,X}$;*
4. *if X and Y are independent, then $\kappa_{X,Y} = \kappa_\Pi = 0$;*
5. *$\kappa_{-X,Y} = \kappa_{X,-Y} = -\kappa_{X,Y}$;*

6. if C_1 and C_2 are two copulas such that $C_1 \leq C_2$ on $[0,1]^2$ then $\kappa_{C_1} \leq \kappa_{C_2}$;

7. if $(X_n, Y_n)_{n\in\mathbb{N}}$ is a sequence of continuous random variables with copulas C_n and if $(C_n)_{n\in\mathbb{N}}$ converges pointwise to C, then $\lim_{n\to\infty} \kappa_{C_n} = \kappa_C$.

Note that Property 1 is satisfied since we did not restrict ourselves to certain copulas. Also note that Properties 2 and 4 are satisfied by applying the transformation in Remark 2. Additionally, the length measure is a function of the copula, and thus, invariant under strictly increasing transformations of the marginals. According to some sources, these properties already qualify the length measure to be a measure of association (Lebrun and Dutfoy, 2009). The following theorem shows that Property 3 is satisfied as well. The proof can be found in Appendix A.

Theorem 20. *Let X, Y be continuous random variables, where (X, Y) has copula $C_{X,Y}$ and (Y, X) has copula $C_{Y,X}$. Then,*

$$\ell_{C_{X,Y}} = \ell_{C_{Y,X}}. \tag{5.18}$$

The next example shows that Property 5 is generally not fulfilled.

Example 4. In this example, we show that there exists no bijection $T : [1/\sqrt{2}, 1] \to [-1, 1]$ such that $T(\ell_C)$ fulfills Property 5 of a concordance measure.

In a first step, notice that such a bijection would have to fulfill $T(\ell_W) = T(1/\sqrt{2}) = -1$, $T(\ell_\Pi) = 0$, and $T(\ell_M) = T(1) = 1$. Furthermore, recall the following lemma which follows immediately from the properties of a copula.

Lemma 21. *If X and Y are two continuous random variables such that (X, Y) has copula $C_{(X,Y)}$, then*

(1.) $(-X, Y)$ has copula $C_{(-X,Y)}(u, v) := v - C_{(X,Y)}(1 - u, v)$ and

(2.) $(X, -Y)$ has copula $C_{(X,-Y)}(u, v) := u - C_{(X,Y)}(u, 1 - v)$.

Let X and Y denote two continuous random variables with copula C. According to Lemma 21, the copula of $(X, -Y)$ is given by $C^*(u, v) := u - C(u, 1 - v)$. If $T(\ell_C)$ is a concordance measure, $T(\ell_C) = -T(\ell_{C^*})$ has to hold. Suppose that (X, Y) is a continuous random vector with copula C and that C is the flipped shuffle of M as depicted in Figure 5.7, left panel.

The length profile of C is given by

$$L_C(t) = \begin{cases} 2 & \text{if } t = 0 \\ 1 + \sqrt{2}(\frac{1}{2} - t), & \text{if } t \in (0, \frac{1}{2}] \\ \sqrt{2}(1 - t), & \text{if } t > \frac{1}{2}. \end{cases} \tag{5.19}$$

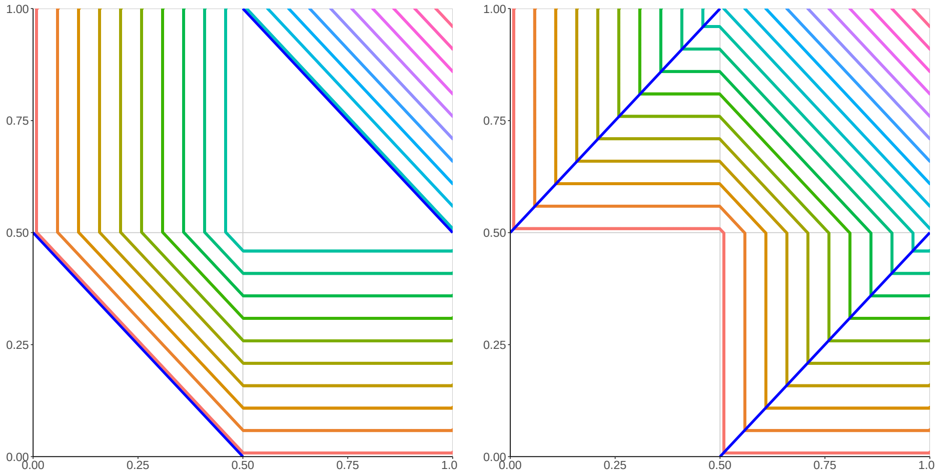

Figure 5.7: Left panel: Support of copula C (blue lines) and some level curves. Right panel: Support of copula C^* (blue lines) and some level curves (Coblenz et al., 2018c).

Thus, the length measure calculates to

$$\ell_C = \int_0^1 L_C(t)\,dt = \frac{1}{4}\left(2+\sqrt{2}\right). \tag{5.20}$$

On the other hand, for the copula $C^*(u,v) = u - C(u, 1-v)$ (Figure 5.7, right panel) the length profile is given by

$$L_{C^*}(t) = \begin{cases} 2 & \text{if } t = 0 \\ 4(\frac{1}{2}-t)+\sqrt{2}t, & \text{if } t \in (0, \frac{1}{2}] \\ \sqrt{2}(1-t), & \text{if } t > \frac{1}{2}. \end{cases} \tag{5.21}$$

This implies

$$\ell_{C^*} = \frac{1}{4}\left(2+\sqrt{2}\right) = \ell_C, \tag{5.22}$$

and thus, $T(\ell_C) = T(\ell_{C^*})$. Since property 5 of a concordance measure requires $T(\ell_C) = -T(\ell_{C^*})$, we get $T(\ell_C) = T(\ell_{C^*}) = 0$. Recalling that $\ell_\Pi \neq \ell_C$ and that $T(\ell_\Pi) = 0$, this contradicts bijectivity of T. ■

Concordance ordering (Property 6 above) for copulas with convex level curves is established in the following theorem, which is a direct consequence of Theorem 16.

Theorem 22. *Let C_1, C_2 be copulas with convex level curves that fulfill the assumptions of Theorem 9. If $C_1 \leq C_2$ on $[0,1]^2$, then*

$$\ell_{C_1} \leq \ell_{C_2}. \tag{5.23}$$

Note that, e.g., Archimedean copulas with continuously differentiable and strict generator fulfill the assumptions of Theorem 22. However, concordance ordering of the length measure is not fulfilled in general. To verify this assertion, reconsider Example 2 which showed that the length profiles for two given copulas $C_1 \geq C_2$ are not ordered. The length measures of these two copulas are given by $\ell_{C_1} = (8 + 17\sqrt{2} + 2\sqrt{10})/50 \approx 0.767$ and $\ell_{C_2} = (1 + \sqrt{2})/3 \approx 0.805$ and, hence, are not ordered either.

Property 7 is fulfilled for copulas with convex level curves. This is shown in the following result.

Theorem 23. *Let $(C_n)_{n \in \mathbb{N}}$ be a sequence of copulas with convex level curves that converges to $C \in \mathcal{C}$. Furthermore, let C and C_n fulfill the assumptions of Theorem 5, for each $n \in \mathbb{N}$. Then,*

$$\lim_{n \to \infty} \ell_{C_n} = \ell_C. \tag{5.24}$$

The proof can be found in Appendix A. The previous theorem is, e.g., fulfilled for a sequence of Archimedean copulas that converges to an Archimedean copula, where all generators involved are continuously differentiable and strict. One can show that Property 7 does not hold in general (cf. Example 7). Some further concepts which are introduced in the next section are needed for this result. Overall, the length measure is not a concordance measure.

Note that one could also consider an integral of the form $\int_0^1 L_C(t) - L_\Pi(t) dt$ and define this as the length measure. However, except for some rescaling, this measure has very similar properties to ℓ_C and also suffers the same consequences. In particular, Example 4 can be directly extended, which means that Property 5 of a concordance measure is not fulfilled. In addition to that, note that Property 7 of a concordance measure also cannot be ensured since the integral can be separated which then involves computation of ℓ_C. In order to construct a dependence measure, one could also consider the integrals $\int_0^1 |L_C(t) - L_\Pi(t)| dt$ or $\int_0^1 (L_C(t) - L_\Pi(t))^2 \, dt$ instead. However, Corollary 19 shows that $\ell_\Pi - \ell_W < \ell_M - \ell_\Pi$, which would violate the properties of a dependence measure stating that maximum dependence should be assigned to situations where one variable is a (not necessarily monotonic) function of the other variable, see Schweizer and Wolff (1981). The next section introduces some further concepts related to the length measure.

5.5.3 Further Considerations for the Class $\mathcal{C}_{cd}$

In this section we establish some nice features of the length profile for the class of mutually completely dependent copulas. Since every copula is Lipschitz continuous, the co-area formula (Federer, 1996) holds and we can express ℓ_C as follows for every $C \in \mathcal{C}$:

$$\ell_C = \int_{(0,1)^2} \|\nabla C(u,v)\|_2 \, d\lambda_2(u,v). \tag{5.25}$$

In fact, the classical co-area formula yields

$$\int_{(0,1)^2} \|\nabla C(u,v)\|_2 \, d\lambda_2(u,v) = \int_{[0,1]} \int_{C^{-1}(\{t\})} dH_1(u,v) d\lambda(t). \tag{5.26}$$

Example 5. We calculate ℓ_C for the copula C considered in Example 4, i.e., C is the flipped shuffle of M as in Figure 5.7, left panel. Then, C is symmetric and the Markov kernel of C is given by $K_C^*(u,[0,v]) = \mathbf{1}_{[0,v]}(h(u))$, whereby $h : [0,1] \to [0,1]$ is the λ-preserving transformation given by

$$h(u) = (\tfrac{1}{2} - u)\mathbf{1}_{[0,\frac{1}{2}]}(u) + (\tfrac{3}{2} - u)\mathbf{1}_{(\frac{1}{2},1]}(u). \tag{5.27}$$

Using symmetry of C and the form of the kernel, it follows immediately that $\|\nabla C\|_2$ (i) is equal to $\sqrt{2}$ on a set (consisting of two triangles) of λ_2-measure $1/4$, (ii) equal to 1 on a set (consisting of two squares) of λ_2-measure $1/2$, and (iii) 0 elsewhere. Using equation (5.25), we get $\ell_C = \sqrt{2}/4 + 1/2$. ■

Example 5 can be generalized to the class of mutually completely dependent copulas $\mathcal{C}_{cd}$. Given a λ-preserving bijection h, we let C_h denote the corresponding mutually completely dependent copula in the sequel (Li et al., 1998; Trutschnig, 2011). It is straightforward to verify that $C_h(v,u) = C_{h^{-1}}(u,v)$ holds for every $(u,v) \in [0,1]^2$ and $h \in \mathcal{T}$. Now define the following sets

$$\Omega_{\sqrt{2}} = \big\{(u,v) \in [0,1]^2 : h(u) \le v, h^{-1}(v) \le u\big\} \tag{5.28}$$

$$\begin{aligned}\Omega_1 = \big\{(u,v) \in [0,1]^2 : h(u) \le v \text{ and } h^{-1}(v) > u \\ \text{OR } h(u) > v \text{ and } h^{-1}(v) \le u\big\}\end{aligned} \tag{5.29}$$

$$\Omega_0 = \big\{(u,v) \in [0,1]^2 : h(u) > v, h^{-1}(v) > u\big\} = (\Omega_{\sqrt{2}} \cup \Omega_1)^c. \tag{5.30}$$

These sets comprise areas, where $\|\nabla C(u,v)\|_2$ is equal to $\sqrt{2}$, 1 or 0 (which is the reason for choosing the indices in this way). Since the Markov kernel of every $C \in \mathcal{C}_{cd}$ only attains values in $\{0,1\}$, we get the following result, which is proved in Coblenz et al. (2018c).

Theorem 24. *For every $C = C_h \in \mathcal{C}_{cd}$ the following formula holds:*

$$\ell_C = \sqrt{2}\,\lambda_2(\Omega_{\sqrt{2}}) + \lambda_2(\Omega_1) = (\sqrt{2}-1)\,\lambda_2(\Omega_{\sqrt{2}}) + 1 - \lambda_2(\Omega_0) \tag{5.31}$$

Given $h \in \mathcal{T}$, let the transformation $T_h : [0,1]^2 \to [0,1]^2$ be defined by

$$T_h(u,v) = \big(h^{-1}(v), h(u)\big), \tag{5.32}$$

for all $u, v \in [0,1]$. Then, T_h is λ_2-preserving and bijective and maps Ω_0 to $\Omega_{\sqrt{2}}$ since

$$\begin{aligned} T^{-1}(\Omega_{\sqrt{2}}) &= \big\{(u,v) \in [0,1]^2 : (h^{-1}(v), h(u)) \in \Omega_{\sqrt{2}}\big\} && (5.33)\\ &= \big\{(u,v) \in [0,1]^2 : h(h^{-1}(v)) \leq h(u) \\ &\qquad\qquad \text{and } h^{-1}(h(u)) \leq h^{-1}(v)\big\} && (5.34)\\ &= \big\{(u,v) \in [0,1]^2 : v \leq h(u) \text{ and } u \leq h^{-1}(v)\big\}. && (5.35) \end{aligned}$$

Since the latter set has the same λ_2-measure as Ω_0, we have proved the following even handier formula for the length measure:

Corollary 25. *For every $C = C_h \in \mathcal{C}_{cd}$ the following formula holds:*

$$\ell_C = 1 - (2-\sqrt{2})\lambda_2(\Omega_0) \tag{5.36}$$

Example 6. Consider the two shuffles $C_{h_1}, C_{h_2} \in \mathcal{C}_{cd}$ depicted in Figure 5.8. For C_{h_1} (left panel) we have $\lambda_2(\Omega_0) = 4/25$ from which, using Equation (5.36), we get $\ell_{C_{h_1}} = 1 - 4(2-\sqrt{2})/25 = (17+\sqrt{2})/25$. For C_{h_2} (right panel) we have $\lambda_2(\Omega_0) = 18/100$ from which, using equation (5.36) again, we get $\ell_{C_{h_2}} = 1 - 18(2-\sqrt{2})/100 = (64+\sqrt{2})/100$. ■

The following example shows that, in general, the length measure does not converge uniformly, implying that Property 7 of a concordance measure (cf. Section 5.5.2) does not hold in general.

Example 7. Fix $m \in \mathbb{N}$, set $n = m^2$ and let π denote the permutation of $\{1, \ldots, n\}$ given by $\pi(m(j-1)+k) = m(k-1)+j$ for $k, j = 1, \ldots, m$, and $\boldsymbol{s} = (i/n)_{i=0}^{n}$. Letting C_n denote the straight shuffle corresponding to $\boldsymbol{s}$ and π, it is well-known (Nelsen (2006), Theorem 3.2.2.) that the sequence $(C_n)_{n\in\mathbb{N}}$ converges uniformly to the independence copula Π. Figure 5.9 depicts the shuffle C_n for some values of $n^2 = m$.

It is easy to verify that the length profile of C_n not only converges to 2 as $t \to 0$, but also converges to 0 as $t \to 1$. The same is true for the length profile of Π. Some basic calculations show that for $C_n = C_{m^2}$ we get $\lambda_2(\Omega_0) = ((m-1)/m)^2/4$. Thus, the length measure is $\ell_{C_n} = \ell_{C_{m^2}} = (\sqrt{2}-2)\,((m-1)/m)^2/4 + 1$, which converges to $(\sqrt{2}-2)/4 + 1 \approx 0.8535$ as $n \to \infty$. In contrast to that, the length measure of Π is $\ell_\Pi \approx 0.7652$. Thus, Property 7 of a concordance measure is not fulfilled. ■

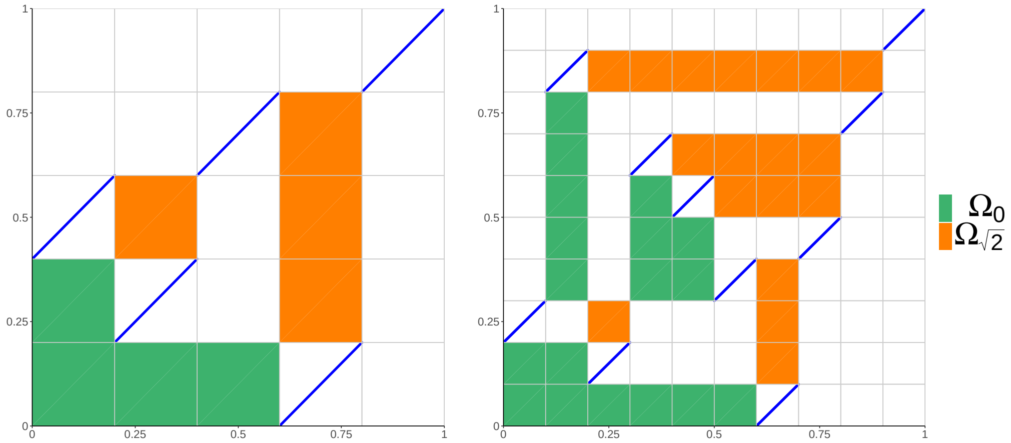

Figure 5.8: Supports of two shuffles $C_{h_1}, C_{h_2} \in \mathcal{C}_{cd}$ in blue and the corresponding sets $\Omega_{\sqrt{2}}$ and Ω_0 (orange and green) (Coblenz et al., 2018c).

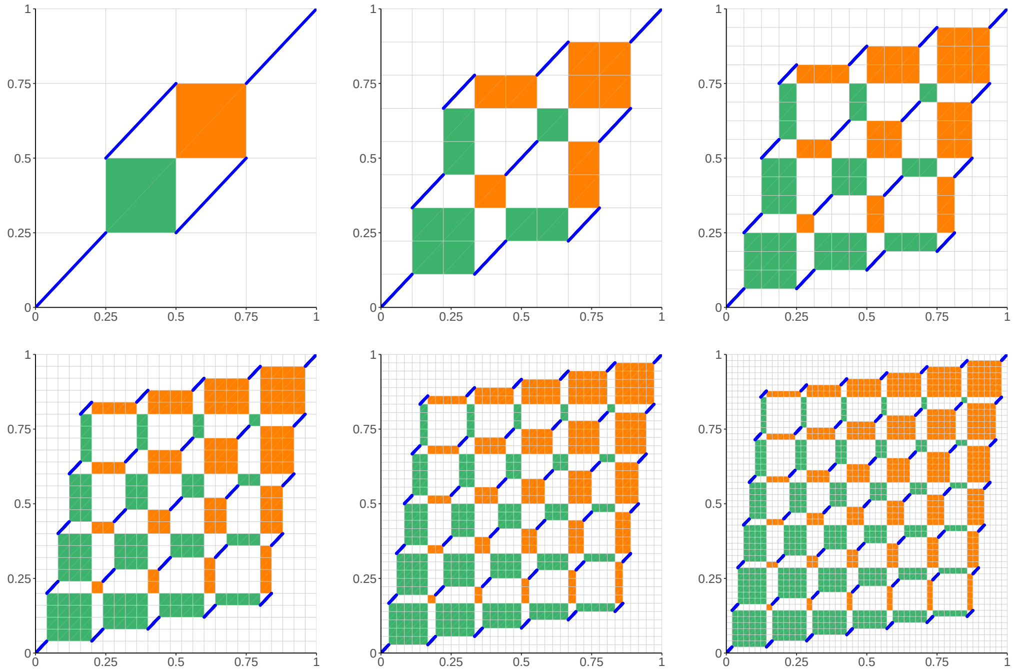

Figure 5.9: Ω_0 (green) and $\Omega_{\sqrt{2}}$ (orange) of $M_n = M_{m^2}$ for $m = 2, 3, 4, 5, 6$ (Coblenz et al., 2018c).

Before providing some concluding remarks in the last section, we close this section with a remark on empirical estimation of the length measure ℓ_C.

Remark 3. *Using Theorem 24 and Corollary 25, an empirical estimator for ℓ_C based on the set Ω_0 seems viable. Given n rank-transformed pseudo-observations, each single observation can be thought of as falling into a box of an $n \times n$ grid. Assuming no ties, each observation is in a grid box with unique row and column. The boxes containing an observation are then transformed to a box containing the strip of a straight shuffle (cf. Figure 5.8), i.e., equivalent to using empirical checkmin copulas. The set Ω_0 can be estimated by counting the boxes belonging to it. The length measure is then estimated by Formula* (5.36)*. Note that the previous results (especially Example 7) suggest that this estimator does not work.*

We conduct a small Monte Carlo simulation of $1,000$ *runs in order to assess the estimator. Table 5.2 shows results for the independence, Clayton and Gauss copulas for different sample sizes n. Clearly, the estimator is biased upwards from the true values, which are shown as well. Interestingly, for the independence copula the estimator yields the value of the shuffle C_n in Example 7 ($\ell_{C_n} \approx$ 0.8535), which showed that the length measure does not converge in general. Although not reported here, this upward bias can be seen for t-, Gumbel, and Frank copulas as well as for further parameters of the Gauss and Clayton copulas.*

Copula	n	mean($\hat{\ell}_C$)	std($\hat{\ell}_C$)	Copula	n	mean($\hat{\ell}_C$)	std($\hat{\ell}_C$)
Π	100	0.855	0.010	Clayton, $\theta = 2$	100	0.927	0.008
$\ell_\Pi = 0.765$	1000	0.854	0.003	$\ell_C = 0.870$	1000	0.927	0.002
Gauss, $\rho = -0.5$	100	0.807	0.009	Gauss, $\rho = 0.5$	100	0.903	0.008
$\ell_C = 0.727$	1000	0.805	0.003	$\ell_C = 0.827$	1000	0.903	0.003

Table 5.2: Estimation results for independence, Clayton, and Gauss copulas.

5.6 Concluding Remarks

We investigate the level curve lengths of bivariate copulas more closely. To this end, we introduce the length profile $L_C(t)$ and establish the length measure ℓ_C as the average length of the level curves of copula C. Some properties of the length profile and the length measure are examined. In particular, the length measure is shown to be a measure of association, however, it is not a concordance measure.

Some work on the length of copula level curves remains to be done. First, the empirical estimation of the length measure is an open question. A first attempt using results for mutually completely dependent copulas did not work out (cf. Section 5.5.3). A further idea would be to estimate the level curves of the copula in a first step and take the average of the lengths of the estimated curves in a second step. Second, the length profile and the length measure can be extended to dimensions $d > 2$ by considering the volume of level surfaces. Hence, a multivariate measure of association is possible.

Finally, it would be interesting to investigate whether there are links between the length of level curves and concordance measures or other notions of dependence. Since Spearman's ρ_S is calculated via the copula surface, it might be worth trying to relate it to the length of level curves, which is a component of the surface. Particularly, we are investigating whether

$$\int_0^1 L_C(t)\omega_C(t)dt \propto \rho_S, \tag{5.37}$$

where $\omega_C(t)$ is a (complicated) suitable weight which depends on the copula C. Another route we are still analyzing is whether the length profile is linked to lower tail dependence λ_L, cf. Section 2.4. In particular, we conjecture that for some subclasses of copulas

$$\left.\frac{dL_C}{dt}\right|_{t=0} = -\frac{2}{\lambda_L}, \tag{5.38}$$

i.e., the first derivative of the length profile $L_C(t)$ at $t = 0$ for suitable copulas C is inversely proportional to the lower tail dependence λ_L.

In the next chapter, we leave the universe of copula level sets and level curves and turn towards modeling d-dimensional copulas, $d > 2$. This is done with so-called vine copulas, which allow to express a d-dimensional copula as a composition of bivariate copulas.

Chapter 6

Non-Simplified Vine Copulas via Tessellation of Conditioning Spaces

This chapter is based on Coblenz (2018).

6.1 Introduction

In recent years, conditional copula models and, in particular, vine copulas have become an important tool for modeling multivariate dependence structures, see, e.g., Chollete et al. (2009), Hobæk Haff et al. (2016), Dalla Valle et al. (2017), Hincks et al. (2018), and Callau Poduje and Haberlandt (2018). In contrast to other multivariate distributions and copula models, vine copulas offer a flexible and easy way of modeling high-dimensional data. Moreover, they can be represented graphically as a sequence of trees, which makes them appealing for interpretation of dependence structures.

The vine copula literature started with the seminal works by Joe (1996), Cooke (1997), and Bedford and Cooke (2001, 2002). A comprehensive overview of past and recent developments can be found on the vine copula homepage maintained by the Czado group at the Technical University of Munich (`http://www.vine-copula.org`). Vine copulas, sometimes also called pair-copula constructions or vines for short, can be motivated by the following two observations: On the one hand, there is only a limited availability of flexible copulas in dimensions three or higher, and on the other hand, there is a plethora of bivariate copulas available. Vine copulas move between the poles of these statements by constructing a d-variate ($d > 2$) distribution from bivariate building blocks.

In order to make theoretical and practical treatment of vine copulas convenient, the so-called simplifying assumption is imposed on the vine structure. The simplifying assumption states that conditioning variables enter only through the arguments of conditional copulas but do not exert an effect on the functional form of the copulas. This assumption is controversially discussed in the vine copula literature (Hobæk Haff et al., 2010; Acar et al., 2012; Stöber et al., 2013). In this chapter, we develop a new method to relax the simplifying assumption. It relies on introducing a tessellation on the conditioning spaces of the conditional copulas. In contrast to existing methods, our model thereby allows the copula family to change within a conditional copula and, thus, offers a different approach to modeling dependence structures.

We contribute to the vine copula literature in several aspects. The main contribution is the development of the non-simplified vine model that works with tessellations of conditioning spaces. The introduced non-simplified vine method is particularly suitable to investigate data sets for non-simplified vine structures and can be seen as a data analysis tool complementary to simplified vines. Thereby, we expect the simplifying assumption to be valid in a lot of real-world examples. If a non-simplified vine is needed, only few of the involved conditional copulas will be affected, and of these especially the ones in lower levels of the vine since these affect the dependence structure the most (Joe et al., 2010; Joe, 2011a,b). The developed method can be used to relax the simplifying assumption for these conditional copulas and, thus, the overall model becomes more appropriate.

Furthermore, we introduce a graphical representation of the developed model to easily interpret estimated non-simplified vines. This is particularly helpful for data exploration. Also, we develop simulation algorithms that are based on an extension of existing algorithms for simplified vines. For estimation purposes, we use well-established techniques from decision trees. These make the estimation computationally feasible and, thus, provide an alternative to simplified vines. Finally, in passing we provide a comprehensive literature overview of recent simplified and non-simplified vine techniques.

The chapter is organized as follows: Section 6.1.1 introduces vine copulas and Section 6.1.2 discusses the simplifying assumption. In Section 6.2 we introduce the new non-simplified vine structure which is based on tessellations of the conditioning variable spaces. Additionally, we develop a graphical representation of the structure. We deal with theoretical and computational aspects of simulation and estimation in Sections 6.3 and 6.4. In an extensive simulation study, we first investigate the estimation procedure for the introduced non-simplified vine model in Section 6.5. Furthermore, in the second part of the simulation study the model is compared to simplified vines in various simplified and non-simplified settings. Section 6.6 contains an application example and Section 6.7 concludes the chapter.

6.1.1 Vine Copulas

This section gives an introduction to vine copulas. These are most easily represented in terms of densities and in the following we assume that these all exist. The corresponding distribution functions and copulas can be obtained by the usual integration. For the sake of simplicity, we assume random variables to be continuous throughout. However, note that vine copulas can generally be extended to discrete random variables (Joe, 2015; Zilko and Kurowicka, 2016).

Let $\mathbf{X} = (X_1, \dots, X_d)$ be a vector of d continuous random variables with joint density $f(x_1, \dots, x_d)$ and univariate marginal densities $f_1(x_1), \dots, f_d(x_d)$. Furthermore, let $F_1(x_1), \dots, F_d(x_d)$ and $c(F_1(x_1), \dots, F_d(x_d))$ denote the corresponding univariate marginal distribution functions and copula density, respectively. Using the pseudo-observations $u_i = F_i(x_i)$, $i = 1, \dots, d$, the copula density can also be written as $c(F_1(x_1), \dots, F_d(x_d)) = c(u_1, \dots, u_d)$. Subsequently, for given indices i, j, $i_1, \dots, i_k$, we write for short $f_i := f_i(x_i)$, $u_{i|i_1,\dots,i_k} := F_{i|i_1,\dots,i_k}(x_i|x_{i_1}, \dots, x_{i_k})$, $c_{ij} := c_{ij}(u_i, u_j)$, and $c_{ij|i_1,\dots,i_k} := c_{ij|i_1,\dots,i_k}(u_{i|i_1,\dots,i_k}, u_{j|i_1,\dots,i_k}; u_{i_1}, \dots, u_{i_k})$.

For simplicity, we consider the case $d = 3$ first and decompose $f(x_1, x_2, x_3)$ as follows

$$f(x_1, x_2, x_3) = f_3 \cdot f_{2|3} \cdot f_{1|23} = f_1 \cdot f_2 \cdot f_3 \cdot c_{13} \cdot c_{23} \cdot c_{12|3}, \tag{6.1}$$

where we have used $f_{2|3} = \frac{f_{23}}{f_3} = c_{23} \cdot f_2$ and $f_{1|23} = \frac{f_{123}}{f_{23}} = \frac{f_{12|3}}{f_{2|3}} = c_{12|3} \cdot f_{1|3} = c_{12|3} \cdot c_{13} \cdot f_1$. Moreover, note that

$$c_{123} = c_{13} \cdot c_{23} \cdot c_{12|3}. \tag{6.2}$$

Equation (6.2) is an example of a 3-dimensional vine copula. As can be seen, a vine copula is a multivariate distribution function that is built of bivariate (conditional) copulas. The concept can be easily extended to dimensions $d > 3$ by decomposing the multivariate densities in an iterative manner as above (Aas et al., 2009). Note that a thorough theoretical treatment of conditional copulas can be found in Patton (2006).

Note that the decomposition in (6.1) is not unique. By permuting indices, the following two decompositions are equivalent alternatives

$$\begin{aligned} f(x_1, x_2, x_3) &= f_1 \cdot f_2 \cdot f_3 \cdot c_{12} \cdot c_{23} \cdot c_{13|2} & (6.3)\\ &= f_1 \cdot f_2 \cdot f_3 \cdot c_{12} \cdot c_{13} \cdot c_{23|1}. & (6.4) \end{aligned}$$

For $d = 3$ there are 3 distinct vine copulas (the ones shown above), for $d = 4$ there are 24 distinct vine copulas and for $d = 5$ there are already 480 distinct vine copulas. For an arbitrary dimension d, there are $2^{(d-2)(d-3)/2} \cdot \frac{d!}{2}$ distinct vine copula decompositions (Morales-Nápoles, 2011; Cooke et al., 2015; Joe, 2015). Appendix D shows the vine arrays (cf. Section 6.3) of all 4-dimensional vine copulas.

Vine copulas have a nice graphical representation as a multi-layered tree structure, which is called vine and is thus the namesake of vine copulas. It was developed in Cooke (1997) and Bedford and Cooke (2001, 2002).

Definition 9 (Regular Vine)**.** *A regular vine* (or R-vine) $\mathcal{V}$ *is a sequence of trees* $T_1, \ldots, T_{d-1}$ *with nodes* N_i *and edges* E_i, $i = 1, \ldots, d-1$, *that has the following properties:*

1. T_1 *is a tree with nodes* $N_1 = \{1, \ldots, d\}$ *and edges* E_1. *For* $i = 2, \ldots, d-1$, T_i *is a tree with nodes* $N_i = E_{i-1}$ *and edges* E_i.
2. *For nodes joined by an edge in* T_i, *the corresponding edges in* T_{i-1} *must share a common node.*

Note that Property 2 is sometimes called *proximity condition* (Dißmann et al., 2013; Joe, 2015). Figure 6.1 shows the graphical representation of an exemplary 5-dimensional R-vine with the following density decomposition

$$f(x_1, \ldots, x_5) = c_{12} \cdot c_{23} \cdot c_{34} \cdot c_{35} \cdot c_{13|2} \cdot c_{24|3} \cdot c_{25|3} \cdot c_{14|23} \cdot c_{45|23} \cdot c_{15|234} \cdot \prod_{k=1}^{5} f_k. \quad (6.5)$$

In the following, we need a notation for an edge e in tree T_i in order to obtain a density formula for R-vines. It depends on the two nodes m and n which are joined by e in T_i (and which are themselves edges in T_{i-1}). We define the sets

$$\begin{aligned} l_e &= \min\{j : j \in (V(m) \cup V(n)) \backslash D_e\} && (6.6)\\ r_e &= \max\{j : j \in (V(m) \cup V(n)) \backslash D_e\} && (6.7)\\ D_e &= V(m) \cap V(n), && (6.8) \end{aligned}$$

where $V(m) = \{l_m, r_m, D_m\}$, $V(n) = \{l_n, r_n, D_n\}$, and D_e is the conditioning set in edge e, which is the empty set in T_1. Now, an edge e in tree T_i can be denoted by $e = l_e r_e | D_e$, where $l_e < r_e$. Hence, the density of a d-dimensional R-vine is (Czado, 2010)

$$f(x_1, \ldots, x_d) = \prod_{k=1}^{d} f_k \cdot \prod_{i=1}^{d-1} \prod_{e \in E_i} c_{l_e, r_e | D_e}. \quad (6.9)$$

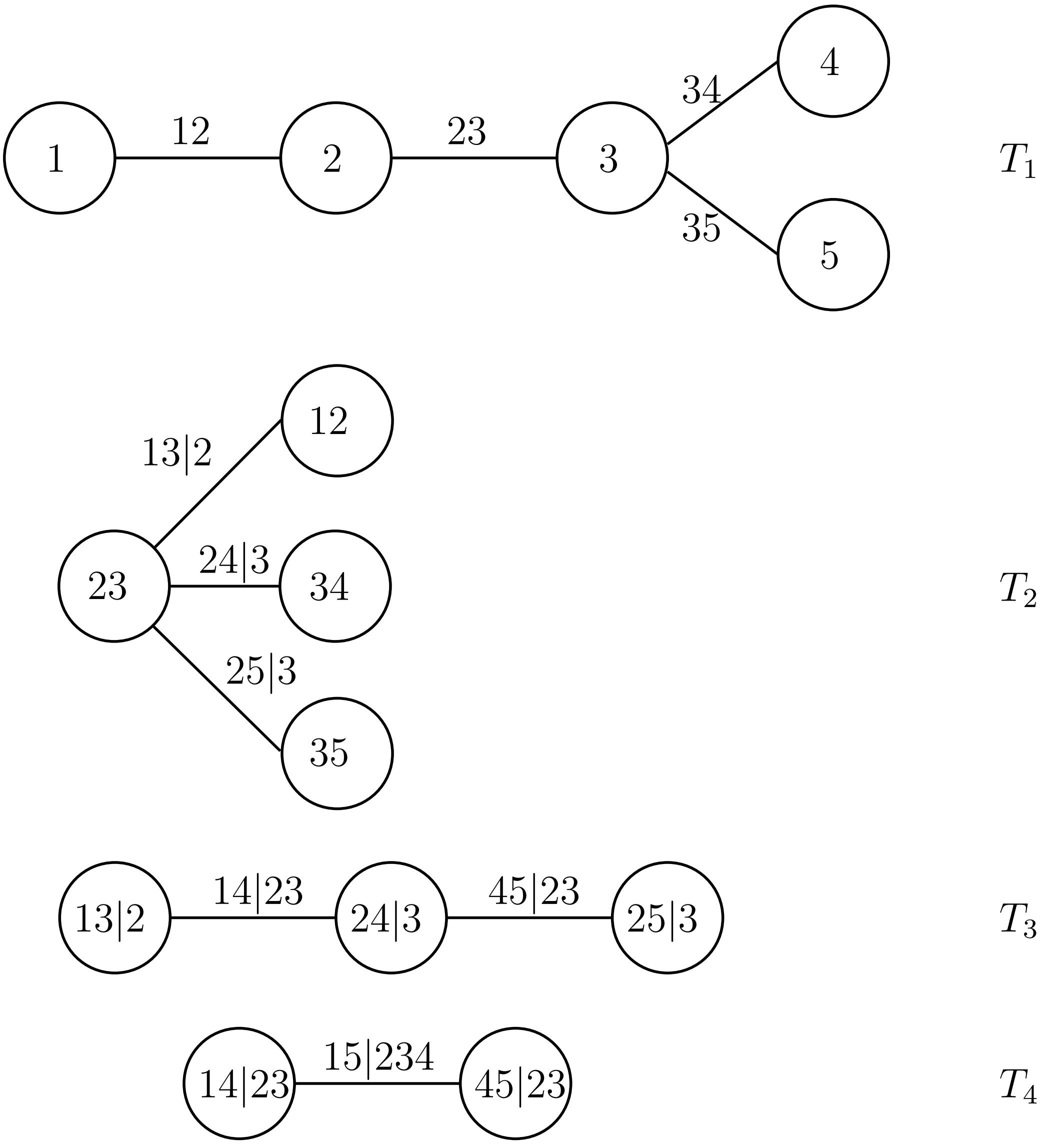

Figure 6.1: A 5-dimensional R-vine.

There are two special cases of R-vines: *Drawable* vines (D-vines) and *canonical* vines (C-vines). D-vines are regular vines where each node in tree T_i, $i = 1, \dots, d-1$, has a maximum of 2 edges attached to it. Figure 6.2 shows an exemplary 5-dimensional D-vine with density decomposition

$$f(x_1, \dots, x_5) = c_{12} \cdot c_{23} \cdot c_{34} \cdot c_{45} \cdot c_{13|2} \cdot c_{24|3} \cdot c_{35|4} \cdot c_{14|23} \cdot c_{25|34} \cdot c_{15|234} \cdot \prod_{k=1}^{5} f_k. \quad (6.10)$$

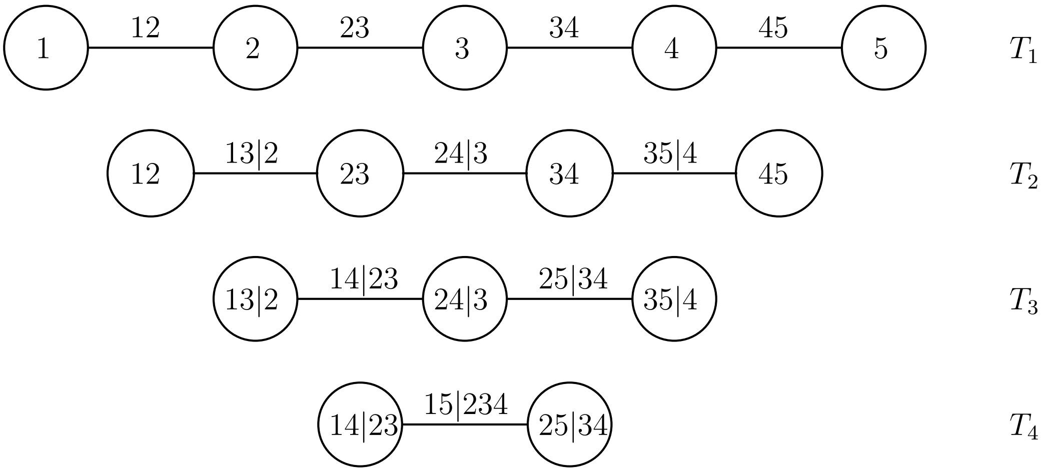

Figure 6.2: A 5-dimensional D-vine.

On the other hand, C-vines are regular vines where each tree T_i has a unique node with $d-i$ edges. Figure 6.3 shows an exemplary 5-dimensional C-vine with density decomposition

$$f(x_1,\ldots,x_5)=c_{12}\cdot c_{13}\cdot c_{14}\cdot c_{15}\cdot c_{23|1}\cdot c_{24|1}\cdot c_{25|1}\cdot c_{34|12}\cdot c_{35|12}\cdot c_{45|123}\cdot\prod_{k=1}^{5} f_k. \quad (6.11)$$

There are R-vines which are neither D-vines nor C-vines. The R-vine in Figure 6.1 is an example. Proper R-vines, i.e., R-vines which are neither D-vines nor C-vines, exist only for dimensions five and higher. For $d=3$ all possible vines are D-vines, for $d=4$ there are 24 distinct vines of which 12 are D-vines and 12 are C-vines (cf. Appendix D).

Both D-vines and C-vines allow a general density decomposition which is notationally lighter compared to R-vines. A D-vine density can be expressed as (Czado, 2010)

$$f(x_1,\ldots,x_d)=\prod_{k=1}^{d} f_k\cdot\prod_{j=1}^{d-1}\prod_{i=1}^{d-j} c_{i,i+j|i+1,\ldots,i+j-1}, \quad (6.12)$$

whereas a C-vine density is obtained by (Czado, 2010)

$$f(x_1,\ldots,x_d)=\prod_{k=1}^{d} f_k\cdot\prod_{j=1}^{d-1}\prod_{i=1}^{d-j} c_{j,j+i|1,\ldots,j-1}. \quad (6.13)$$

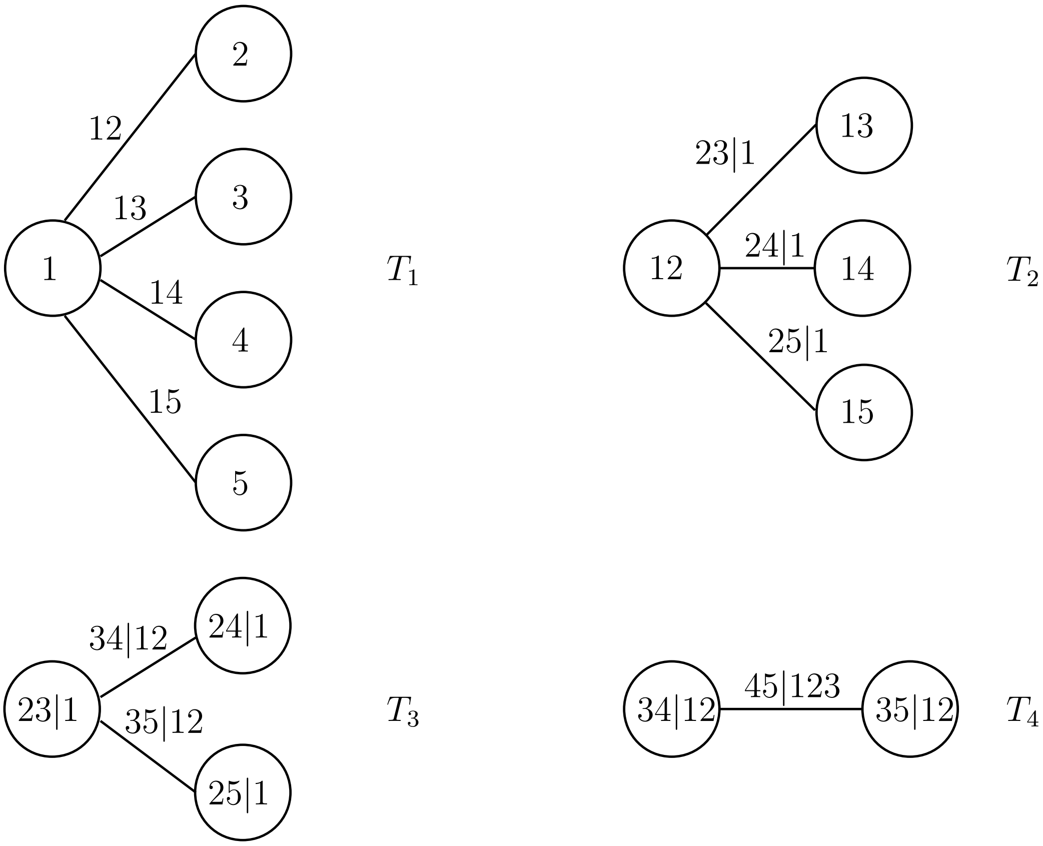

Figure 6.3: A 5-dimensional C-vine.

An active research area is the investigation of the strength, type, and structure of dependence a vine copula can have. There are propositions for tail dependence, which state that upper and lower tail dependence are most strongly influenced by the copulas in the first tree T_1 (Joe et al., 2010; Joe, 2011b). Also, reflection symmetry is determined by the conditional copulas in a vine structure (Joe, 2011b). Finally, there is some work comparing the strength of dependence between D-, C-, and R-vines (Joe, 2011a).

There is many more to vine copulas, such as vine equivalent classes, which is outside the scope of this thesis. The interested reader can find a detailed treatise of further concepts connected to vine copulas in the excellent book by Joe (2015). The next section deals with the so-called simplifying assumption which is an important concept in the vine copula universe.

6.1.2 The Simplifying Assumption and its Relaxations

This section introduces the simplifying assumption and reviews methods to relax it. We start again with the R-vine copula in Figure 6.1 which has the copula density decomposition (6.5). Let us now have a closer look, e.g., at the conditional copula density $c_{13|2}$ within the decomposition which is fully written-out

$$c_{13|2} = c(u_{1|2}, u_{3|2}; u_2), \tag{6.14}$$

where $u_{1|2} = F_{1|2}(x_1|x_2)$ and $u_{3|2} = F_{3|2}(x_3|x_2)$. Thus, not only the arguments of $c_{13|2}$ depend on the conditioning variable but also the copula itself. This observation applies to all conditional copulas that are part of a vine decomposition. Since this makes theoretical and practical handling of vine copulas, e.g., in simulation and estimation, very complicated, it is often assumed that the dependence of conditioning variables only enters through the arguments of the vine copula model but not also through the conditional copula functions. This assumption is commonly called the *simplifying assumption*.

The simplifying assumption is not merely a theoretical convenience. There are multivariate distributions for which any R-vine decomposition is simplified. Examples are the multivariate Gaussian and Student-t distributions as well as the Clayton copula (Hobæk Haff et al., 2010; Stöber et al., 2013). Other special cases for which this is true can be found in Joe (2015). Moreover, for some multivariate distributions there is a specific R-vine decomposition which is simplified, but not all R-vine decompositions are simplified. A list of these cases can be found in Joe (2015), too.

The simplifying assumption is a heavily discussed topic, see, e.g., Hobæk Haff et al. (2010), Acar et al. (2012), and Stöber et al. (2013). It is mostly conceived as a convenient tool both for theoretical analysis and for applications (Killiches et al., 2017). However, since the simplifying assumption is not always fulfilled in real data, there is a growing literature exploring different strategies to relax it. These strategies can be roughly categorized into two groups: (1) approaches that model the conditional copulas in a vine directly and (2) approaches that model the parameter(s) of the conditional copulas involved as a function of the conditioning variable(s). Into the first category fall Schellhase and Spanhel (2018) who model a non-simplified vine copula by using penalized splines for the conditional copulas. Further approaches for this category are still lacking. In contrast to that, the second category offers a broader spectrum of methods. They all rely on the assumption that the parametric form of the conditional copula remains the same for all values of the conditioning variables.

Acar et al. (2012) model the conditional copula parameter as a locally linear function of the conditioning variable. Let $c_{ij|k}$ be a conditional copula with one conditioning variable X_k and $\eta_{ij|k}$ a twice continuously differentiable function on the support of X_k. Then, the copula parameter $\theta_{ij|k}$ can be modeled as

$$\theta_{ij|k}(x_k^*) = g(\eta_{ij|k}(x_k^*)) \approx g(\eta_{ij|k}(x_k) + \eta'_{ij|k}(x_k)(x_k^* - x_k)), \tag{6.15}$$

where x_k^* is a point in the neighborhood of x_k and g is a link function that maps from $\mathbb{R}$ into the parameter space, thus ensuring that only viable parameter values are obtained (e.g., for a Gaussian copula, the parameter has to stay between -1 and 1). The copula parameter is then estimated via a kernel weighted maximum likelihood approach. So far, this approach can only be used for one conditioning variable and thus is infeasible for vine copulas of dimension four or higher.

Other strategies in the second category rely on Proposition 1 in Hobæk Haff et al. (2010) which states that if the Kendall's τ of a conditional copula in a given vine structure depends on the conditioning variable(s), the vine is non-simplified. Note that the reverse statement does not hold in general, i.e., the Kendall's τ of the conditional copula in question can be independent of the conditioning variable(s) and the vine decomposition can still be non-simplified. Hence, these techniques try to model Kendall's τ as a function of the conditioning variable(s) and subsequently convert it into the parameter of the copula family which is used to model the conditional copula. For a lot of copula families, the second step can be easily executed since closed-form formulas between the copula parameter and Kendall's τ are available. Let g be a link function that maps from $\mathbb{R}^d$ to $[-1, 1]$. Then, the Kendall's τ $\tau_{ij|i_1,\dots,i_k}$ of $c_{ij|i_1,\dots,i_k}$ can be modeled as

$$\tau_{ij|i_1,\dots,i_k}(x_{i_1}, \dots, x_{i_k}) = g(x_{i_1}, \dots, x_{i_k}). \tag{6.16}$$

Lopez-Paz et al. (2013) set $g(x_{i_1}, \dots, x_{i_k}) = 2\Phi(f(x_{i_1}, \dots, x_{i_k})) - 1$, where Φ is the Gaussian CDF and f is some nonlinear function. In Vatter and Nagler (2018), g is viewed as a generalized additive model with linear and nonlinear components. An advantage of this is that standard estimation procedures for generalized additive models can be used. Also, this approach allows to incorporate exogenous variables into the estimation process.

We want to point out that there is a growing literature on modeling conditional copulas outside of a vine copula context. These comprise time-dependent conditional copulas (Fermanian and Wegkamp, 2012), single-index copulas (Fermanian and Lopez, 2018), conditional copulas with generalized additive models (Vatter and Chavez-Demoulin, 2015) and Gaussian processes (Levi and Craiu, 2018), as well as modeling of conditional copulas with other non- and semiparametric methods (Acar et al., 2011, 2013; Abegaz et al., 2012; Gijbels et al., 2011, 2012; Veraverbeke et al., 2011). These are not considered further here. However, it would be interesting to adapt these to vine copulas and test their modeling capabilities.

In summary, approaches from the second category above try to relax the simplifying assumption by letting the parameter of the conditional copula depend on the conditioning variables. They are limited in the sense that a conditional copula in a vine must stem from one chosen copula family. It cannot be composed of different copula families. The next section develops a technique that allows the copula family to change within a conditional copula.

6.2 Tessellation of Conditioning Spaces

In this section, we develop a new strategy to relax the simplifying assumption. It is different from the approaches discussed in the previous section since it does not assume a constant parametric form of the conditional copulas involved. Also, it is efficient in terms of computing times (cf. Section 6.6). Additionally, a nice graphical representation of the non-simplified vine structure is introduced. The developed approach can be used in data analysis to investigate vine structures for non-simplified parts. It lends some ideas from Derumigny and Fermanian (2017) and Kurz and Spanhel (2018) since the conditioning space is partitioned into boxes. However, our general intention and the partitioning technique used here are quite different from the existing literature. We strongly focus on estimation and rely on decision tree techniques to simplify the estimation task.

The basic idea of the new method is to partition the conditioning space into disjoint sets and estimate a conditional copula on each part of the resulting tessellation separately. Let the conditional copula in the vine structure we want to model be $c_{ij|i_1,\dots,i_k} = c_{ij|i_1,\dots,i_k}(u_{i|i_1,\dots,i_k}, u_{j|i_1,\dots,i_k}; u_{i_1}, \dots, u_{i_k})$. The space of the pseudo-observations $U_{i_1}, \dots, U_{i_k}$ of the conditioning variables $X_{i_1}, \dots, X_{i_k}$ is $S_k := [0,1]^k$ since there are k variables we condition on. We now seek a partition of S_k into m subsets B_l, $l = 1, \dots, m$.

Definition 10 (Tessellation). *A collection of subsets B_l, $l = 1, \dots, m$, of a specific partition of S_k, such that*

$$\bigcup_{l=1}^{m} B_l = S_k \tag{6.17}$$

and

$$B_{l_1} \cap B_{l_2} = \emptyset, \text{ for all } l_1 \neq l_2, \tag{6.18}$$

is called tessellation. *It is denoted by* $\mathbb{B} := \{B_1, \ldots, B_m\}$.

In general, a tessellation $\mathbb{B}$ can comprise any mutually non-overlapping partition of S_k. However, to simplify the estimation of $c_{ij|i_1,\ldots,i_k}$, we restrict ourselves to partitions that can be represented by a decision tree. That is, the tessellation is comprised of partitions with linear, axis-aligned boundaries. It can be represented by a decision tree that splits a conditioning variable's space into discrete (ordered) intervals in its nodes. Each interval is then considered a new node on the next decision tree level that splits the variable space of the next variable. This is repeated until all conditioning variable spaces are split and S_k is fully partitioned. Note that the decision tree can generate multiple intervals at each node and, thus, we do not restrict ourselves to a binary decision tree. Usually, a variable can be split repeatedly at different tree levels. However, in order to make the estimation task easier later on, we exclude this possibility, i.e., once a variable is split in a given tree level it cannot be split again in subsequent tree levels. As an example, consider Figure 6.4 which shows a tessellation $\mathbb{B} = \{B_1, B_2, B_3, B_4\}$ for a 2-dimensional conditioning variable space of u_2 and u_3 in the left panel. The tessellation can be represented by a decision tree as shown in the right panel.

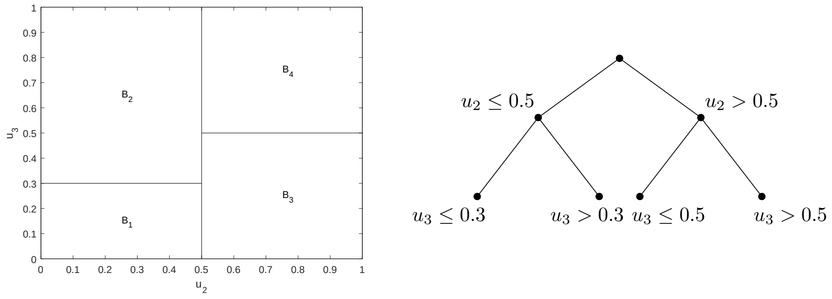

Figure 6.4: Left panel: Tessellation $\mathbb{B} = \{B_1, B_2, B_3, B_4\}$ of the conditioning space $S_2 = [0,1]^2$ spanned by u_2 and u_3. The partitions have linear, axis-aligned boundaries. Right panel: The corresponding representation via a decision tree.

We assume that on every $B_l \in \mathbb{B}$ a copula density $c_{ij}^{B_l}$ is specified which does not depend on the conditioning variable further. That is, the copula density $c_{ij|i_1,\ldots,i_k}$ can be represented as

$$c_{ij|i_1,\ldots,i_k} = \begin{cases} c_{ij}^{B_1}, & (u_{i_1},\ldots,u_{i_k}) \in B_1 \\ \vdots & \\ c_{ij}^{B_m}, & (u_{i_1},\ldots,u_{i_k}) \in B_m \end{cases}, \tag{6.19}$$

where $c_{ij}^{B_l} = c(u_{i|i_1,\ldots,i_k}, u_{j|i_1,\ldots,i_k}; (u_{i_1},\ldots,u_{i_k}) \in B_l)$. Thus, on each $B_l \in \mathbb{B}$ the copula stays constant. The assumption that the copula densities $c_{ij}^{B_l}$, $l = 1,\ldots,m$, do not depend on the conditioning variables can be further relaxed, e.g., by making the parameter dependent on the conditioning variable(s) as in the approaches discussed in the previous section. We leave this for future research and do not pursue this idea further here.

Note that this approach contains the simplified vine structure as a special case, i.e., when all copulas on a given tessellation are the same or the tessellation contains S_k as the only element. If the tessellation is estimated from the data (cf. Sections 6.4.2–6.4.4), the simplified and non-simplified vines are nested. This is in line with a non-simplified vine model, where the parameters of the conditional copulas are functions of the conditional variables, because the parameters can be restricted to a constant value and, thus, a simplified vine is obtained (Killiches et al., 2016). In contrast to that, the models are not nested in a statistical sense when the tessellation is predefined (cf. Section 6.4.1) since then they represent different distributions which cannot be transformed into each other by a parameter restriction in the richer model, i.e., the non-simplified vine with tessellated conditioning spaces (Greene, 2012).

The decision tree tessellation we employ not only has advantages for estimation purposes (cf. Section 6.4) but also yields a nice graphical representation of the respective vine structure. Figure 6.5 shows an exemplary 4-dimensional D-vine, where on levels T_2 and T_3 the simplifying assumption is violated. The conditioning space tessellations are incorporated in their decision tree representation at the respective edges of the vine. Additionally, the corresponding Kendall's τ values are shown. This graphical representation has two main advantages. First, one can see immediately which copulas violate the simplifying assumption. Second, via the decision tree representation one can get a first impression how strongly the simplifying assumption is violated and how fine-grained the tessellation actually is. Hence, it can be a powerful tool for exploratory data analysis.

In the next two sections we deal with simulation and estimation of the proposed vine copula structure. Moreover, we discuss concepts necessary for computational treatment.

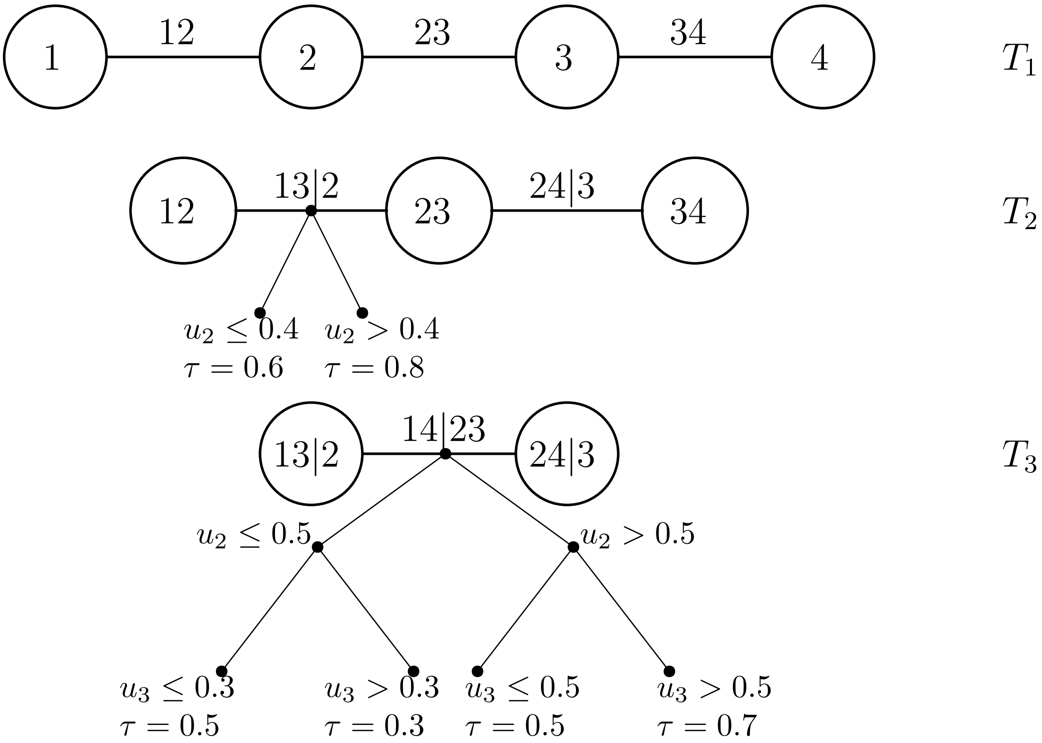

Figure 6.5: Exemplary graphical representation of a non-simplified D-vine. The simplifying assumption is violated on levels T_2 and T_3. The tessellations of the conditioning spaces are incorporated via the decision tree representation at the edges where the simplifying assumption is violated. Also, Kendall's τ is shown for each partition.

6.3 Simulation

This section deals with generating a sample from a non-simplified vine copula with tessellation of conditioning spaces. We extend well-known simulation algorithms for simplified vine copulas for that purpose. In the following, we introduce some further concepts which will come in handy when treating computational aspects of both simulation and estimation later on.

For simulation, we need the conditional distribution of U_i given U_j. The first concept we introduce is the so-called *h-function*, which delivers just this. The h-function is a (univariate) conditional distribution based on a bivariate copula (Aas et al., 2009) and is an integral part of simulation and estimation of vines. Note that in general the h-function is not the uniform distribution.

Let $F_{ij}(x_i, x_j)$ be a bivariate continuously differentiable distribution function of two continuous random variables X_i and X_j. Then, the conditional distribution $F_{i|j}(x_i|x_j)$ of X_i given X_j can be obtained by (Joe, 2015)

$$F_{i|j}(x_i|x_j) = \lim_{\epsilon \to 0^+} \frac{\mathbb{P}(X_i \leq x_i, x_j \leq X_j \leq x_j + \epsilon)}{\mathbb{P}(x_j \leq X_j \leq x_j + \epsilon)} = \frac{\frac{\partial F_{ij}}{\partial x_j}}{f_j(x_j)}, \tag{6.20}$$

where $f_j(x_j)$ is the density of X_j. Note that this also holds for bivariate distributions that are conditioned on further random variables $X_{i_1}, \dots, X_{i_k}$ and even holds for d-variate distributions, $d > 2$. Now observe that a bivariate copula C_{ij} is a distribution function of uniformly distributed random variables $U_i = F_i(X_i)$ and $U_j = F_j(X_j)$, i.e., $f_i(u_i) = f_j(u_j) = 1$. Then, we can express the h-function h_{ij} of copula C_{ij} as (Aas et al., 2009; Joe, 2015)

$$h_{ij}(u_i, u_j) = C_{i|j} = \frac{\partial C_{ij}(u_i, u_j)}{\partial u_j}. \tag{6.21}$$

The inverse h-function, h^{-1}, also plays an important role as will become clear later.

A second concept we employ is the so-called vine array $\mathbf{A}$ of a vine copula. A vine array represents a d-dimensional vine via a $d \times d$ upper-triangular matrix $\mathbf{A}$ with elements (a_{ij}), $i, j = 1, \dots, d$ (Joe, 2015). It simplifies computational treatment of vines tremendously. The first row and the diagonal of the vine array represent the sequence in which the variables are joined via the copulas in the first tree T_1 of a vine copula. The remaining entries of the upper-triangular rows are determined by the sequence the variables are joined via copulas in the subsequent trees T_2 to T_{d-1}.

This is best explained by an example. Consider the R-vine in Figure 6.1 again. The corresponding vine array $\mathbf{A}_R$ is given by

$$\mathbf{A}_R = \begin{pmatrix} 1 & 1 & 2 & 3 & 3 \\ & 2 & 1 & 2 & 2 \\ & & 3 & 1 & 4 \\ & & & 4 & 1 \\ & & & & 5 \end{pmatrix}.$$

In the first tree, we start with variable 1 and join variable 2 to it. Then, we join 3 to 2, 4 to 3, and finally, 5 to 3. Thus, the first row of $\mathbf{A}_R$ is $(1, 1, 2, 3, 3)$ and the diagonal elements are $(1, 2, 3, 4, 5)$ because the copulas c_{12}, c_{23}, c_{34}, and c_{35} are present. This means that a copula c_{ij} on an edge in the first tree T_1 of a d-dimensional vine can be indexed by the vine array elements via $c_{a_{1k}a_{kk}}$, $k = 2, \dots, d$.

Advancing to the second tree T_2 of the R-vine in Figure 6.1, we see that variables 1 and 3, 2 and 4, and 2 and 5 are joined by the copulas $c_{13|2}, c_{24|3}$, and $c_{25|3}$, respectively. Thus, the remaining entries in the second row of $\mathbf{A}_R$ are $(1, 2, 2)$. Note that the elements in the first row now contain the respective conditioning variables. This is always the case. For a given tree T_k and column $m \geq k+1$ of a vine array $\mathbf{A}$, the rows $1, \ldots, k-1$ of $\mathbf{A}$ contain the conditioning variable set.

In the third tree T_3, variables 1 and 4 as well as 4 and 5 are joined by the copulas $c_{14|23}$ and $c_{45|23}$ yielding the remaining entries $(1, 4)$ in the third row of $\mathbf{A}_R$. Note that here the elements in the first and second row contain the conditioning variables. Finally, variables 1 and 5 are joined in the fourth tree T_4 by copula $c_{15|234}$ which is reflected by the entry $a_{45} = 1$. Again, rows one to three in the last column contain the conditioning variables.

Since every variable has to be included in the first tree of a vine copula, the diagonal a_{kk} of a vine array contains every variable exactly once. It is always possible – by potentially renumbering the variables – to construct a vine array $\mathbf{A}$ that fulfills $a_{kk} = k$ for its diagonal entries (Joe, 2015). This is particularly important for computational treatment of vine copulas. We call a vine array that suffices $a_{kk} = k$ on its diagonal entries, an *ordered* vine array.

Definition 11 (Ordered Vine Array)**.** *A vine array* $\mathbf{A}$ *of a d-dimensional vine copula which fulfills* $(a_{kk}) = k, k = 1 \ldots, d$, *on its diagonal elements, is called* ordered *vine array.*

Sometimes, a vine array which additionally fulfills $a_{k-1,k} = k-1$ is called vine array in natural order (Joe, 2015).

Furthermore, note that in each column k of an ordered vine array, each of the variables involved appears exactly once because of the construction rules of a vine copula (cf. Definition 9). Apart from that, the entries in a general R-vine do not necessarily follow a pattern. This is different for vine arrays of D-vines and C-vines. The (ordered) vine arrays $\mathbf{A}_D$ and $\mathbf{A}_C$ of the D-vine in Figure 6.2 and the C-vine in Figure 6.3, respectively, are as follows

$$\mathbf{A}_D = \begin{pmatrix} 1 & 1 & 2 & 3 & 4 \\ & 2 & 1 & 2 & 3 \\ & & 3 & 1 & 2 \\ & & & 4 & 1 \\ & & & & 5 \end{pmatrix}, \mathbf{A}_C = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ & 2 & 2 & 2 & 2 \\ & & 3 & 3 & 3 \\ & & & 4 & 4 \\ & & & & 5 \end{pmatrix}.$$

As can be seen, both vine arrays obey a specific pattern. For a general ordered D-vine array the off-diagonal elements fulfill $(a_{k,k+1}) = 1, (a_{k,k+2}) = 2, \ldots,$ $k = 1, \ldots, d$, whereas for a general ordered C-vine array the rows fulfill $(a_{1k}) = 1, (a_{2k}) = 2, \ldots, k = 1, \ldots, d$. For further examples of vine arrays, we refer the reader to Appendix D which lists the vine arrays of all 4-dimensional vine copulas.

We are now equipped to treat the simulation of non-simplified vine copulas with tessellation of conditioning spaces. We start with a general simulation algorithm for multivariate distributions which we subsequently apply to simplified vine copulas. Afterwards, we introduce simulation algorithms for non-simplified D-, C-, and R-vines which are directly extended from algorithms for simplified vines. Finally, we show an exemplary sample of a non-simplified vine with tessellated conditioning spaces.

A simulation algorithm for general multivariate distributions can be obtained from the so-called Rosenblatt transform (Rosenblatt, 1952). Let $F(x_1, \ldots, x_d)$ be the d-dimensional distribution of the random variable $\mathbf{X} = (X_1, \ldots, X_d)$ from which we want to sample. Also, let the univariate marginal distributions be denoted by $F_1(x_1), \ldots, F_d(x_d)$ and the conditional distributions be denoted by $F_{2|1}(x_2|x_1), F_{3|12}(x_3|x_1, x_2), \ldots, F_{d|1,\ldots,d-1}(x_d|x_1, \ldots, x_{d-1})$. Furthermore, let $F^{-1}_{2|1}(x_1|x_2), F^{-1}_{3|12}(x_3|x_1, x_2), \ldots, F^{-1}_{d|1,\ldots,d-1}(x_d|x_1, \ldots, x_{d-1})$ be the inverses of the conditional distributions. Let $W_1, \ldots, W_d$ be i.i.d. uniformly distributed and let the random vector $(Y_1, \ldots, Y_d)$ be given by

$$Y_1 = F_1^{-1}(W_1) \tag{6.22}$$

$$Y_2 = F_{2|1}^{-1}(W_2|Y_1) \tag{6.23}$$

$$Y_3 = F_{3|12}^{-1}(W_3|Y_1, Y_2) \tag{6.24}$$

$$\vdots \tag{6.25}$$

$$Y_d = F_{d|1,\ldots,d-1}^{-1}(W_d|Y_1, \ldots, Y_{d-1}). \tag{6.26}$$

It can be shown that $(Y_1, \ldots, Y_d)$ is distributed according to F (Rosenblatt, 1952; Joe, 2015).

Since the presented algorithm uses conditional distributions, it can be easily applied to copulas and simplified vine copulas by using h-functions and their inverses. To demonstrate this, we execute the algorithm for a 3-dimensional D-vine with the following vine array

$$\mathbf{A} = \begin{pmatrix} 1 & 1 & 2 \\ & 2 & 1 \\ & & 3 \end{pmatrix}.$$

This exemplary D-vine structure comprises the copulas $c_{12}, c_{23}, c_{13|2}$. Recall that a copula is a distribution of uniformly distributed random variables. Hence, in order to generate a random vector (U_1, U_2, U_3) that is distributed according to the D-vine, we draw i.i.d. uniformly distributed W_1, W_2, W_3 and compute

$$U_1 = W_1 \tag{6.27}$$
$$U_2 = C^{-1}_{2|1}(W_2|U_1) = h^{-1}_{12}(W_2, U_1) \tag{6.28}$$
$$U_3 = C^{-1}_{3|12}(W_3|U_1, U_2) = h^{-1}_{23}(h^{-1}_{13|2}(W_3, h_{12}(U_1, U_2)), U_2), \tag{6.29}$$

where the indexing of the h-functions and their inverses corresponds to the indices of the involved copulas. Note that in the third line we have to account for the fact that U_1 and U_3 depend on each other via the copula $C_{13|2}$ and, thus, we cannot just use $C^{-1}_{3|2}$ to obtain U_3 as we have used $C^{-1}_{2|1}$ in the second line to obtain U_2. However, the term $h^{-1}_{13|2}(W_3, h_{12}(U_1, U_2))$, which yields a value for $C_{3|2}(U_3, U_2)$, takes care of this. Thus, we can simulate from an arbitrary vine copula by stacking h-functions and their inverses as appropriate. Computationally, the correct stacking order for R-vines can be obtained from the vine array. For D- and C-vines the stacking order can be directly inferred from their imposed structure. Algorithms 8 to 10 in Appendix E show simulation algorithms for simplified D-, C-, and R-vines.

In the following, we extend the simulation algorithms for simplified vine copulas to non-simplified vine copulas with tessellated conditioning spaces as introduced in Section 6.2. For the algorithms, we need an auxiliary $(d-1) \times (d-1)$ array **fam** that contains in each row k the indices of the copulas in tree T_k as they appear in the respective ordered vine array. For example, the arrays $\mathbf{fam}^{\mathrm{R}}$, $\mathbf{fam}^{\mathrm{D}}$, and $\mathbf{fam}^{\mathrm{C}}$ of the vine copulas in Figures 6.1, 6.2, and 6.3, respectively, are

$$\mathbf{fam}^{\mathrm{R}} = \begin{pmatrix} 12 & 23 & 34 & 35 \\ 13|2 & 24|3 & 25|3 & 0 \\ 14|23 & 45|23 & 0 & 0 \\ 15|234 & 0 & 0 & 0 \end{pmatrix},$$

$$\mathbf{fam}^{\mathrm{D}} = \begin{pmatrix} 12 & 23 & 34 & 45 \\ 13|2 & 24|3 & 35|4 & 0 \\ 14|23 & 25|34 & 0 & 0 \\ 15|234 & 0 & 0 & 0 \end{pmatrix},$$

and

$$\mathbf{fam}^{\mathrm{C}} = \begin{pmatrix} 12 & 13 & 14 & 15 \\ 23|1 & 24|1 & 25|1 & 0 \\ 34|12 & 35|12 & 0 & 0 \\ 45|123 & 0 & 0 & 0 \end{pmatrix}.$$

In addition to that, let $\mathbf{B}$ be an auxiliary $(d-1) \times (d-1)$ array along the lines of the array **fam** that contains a data structure to save the corresponding tessellation of the **fam** array. Algorithms 3, 4, and 5 show the extended versions of Algorithms 8, 9, and 10 in Appendix E. The blue writing marks the differences. As can be seen, an extension is straightforward by checking how to stack the h-functions and their inverses h^{-1} according to the tessellations in the respective tree levels that are contained in $\mathbf{B}$. This is always possible because in order to simulate the next variable U_k we already know the variables $U_1, \ldots, U_{k-1}$. Thus, the respective tessellations are known and we know how to transform the variables via the h- and h^{-1}-functions.

Algorithm 3 Simulation of d-dimensional non-simplified D-vine

```
Initialize two auxiliary $d \times d$ arrays $(a_{ij})$ and $(b_{ij})$, an auxiliary $(d-1) \times (d-1)$ array **fam** that contains in each row $k$ the indices of the copulas in tree $T_k$ as they appear in an ordered D-vine array, and an auxiliary array $\mathbf{B}$ that contains a data structure for the corresponding tessellation in **fam**.
Generate $d$ independent uniformly distributed random variables $w_1, \ldots, w_d$.
Set $u_1 = w_1$, $a_{11} = w_1$, $b_{11} = w_1$.
for $i = 2, \ldots, d$
    $a_{i1} = w_i$
    for $j = 2, \ldots, i$
        $a_{ij} = h^{-1}_{\mathbf{fam}(\mathrm{i-j+1,j-1}),\mathbf{B}(i-j+1,j-1)}(a_{i,j-1}, b_{i+1,j-1})$
    end for
    $u_i = a_{ii}$
    $b_{ii} = a_{ii}$
    for $j = i-1, \ldots, 1$
        $b_{ij} = h_{\mathbf{fam}(\mathrm{i-j,j}),\mathbf{B}(i+j,j)}(b_{i-1,j}, a_{i,j+1})$
    end for
    Select the tessellations according to $u_1, \ldots, u_i$ from $\mathbf{B}$.
end for
Return $(u_1, \ldots, u_d)$.
```

Figure 6.6 shows the scatter plot matrix of a sample from the vine structure depicted in Figure 6.5. Copulas c_{12}, c_{23}, c_{34} are Clayton copulas with parameters $\theta = 2, 3, 4$, respectively. Copula $c_{24|3}$ is a Gumbel copula with parameter $\theta = 4$. Copula $c_{13|2}$ is given by

$$c_{13|2} = \begin{cases} c_{13}^{\leq 0.4}, & u_2 \leq 0.4 \\ c_{13}^{>0.4}, & u_2 > 0.4 \end{cases}, \tag{6.30}$$

where $c_{13}^{\leq 0.4}$ and $c_{13}^{>0.4}$ are Gauss copulas with parameter $\rho = -0.7, 0.7$, re-

Algorithm 4 Simulation of d-dimensional non-simplified C-vine

```
Initialize an auxiliary $(d-1)\times(d-1)$ array **fam** that contains in each row $k$ the indices of the copulas in tree $T_k$ as they appear in an ordered C-vine array and initialize an auxiliary array $\mathbf{B}$ that contains a data structure to select the corresponding tessellation in **fam**.
Generate $d$ independent uniformly distributed random variables $w_1, \dots, w_d$.
Set $u_1 = w_1$, $u_2 = h^{-1}_{\mathbf{fam}(1,1)}(w_2, w_1)$.
Select the tessellations according to $u_1, u_2$ from $\mathbf{B}$.
for $i = 3, \dots, d$
    $t = w_i$
    for $j = i-1, \dots, 1$
        $t = h^{-1}_{\mathbf{fam}(\mathrm{j,i-j}),\mathbf{B}(j,i-j)}(t, w_j)$
    end for
    $u_i = t$
    Select the tessellations according to $u_1, \dots, u_i$ from $\mathbf{B}$.
end for
Return $(u_1, \dots, u_d)$.
```

spectively. Copula $c_{14|23}$ is given by

$$c_{14|23} = \begin{cases} c_{14}^{B_1}, & u_2 \leq 0.5, u_3 \leq 0.3 \\ c_{14}^{B_2}, & u_2 \leq 0.5, u_3 > 0.3 \\ c_{14}^{B_3}, & u_2 > 0.5, u_3 \leq 0.5 \\ c_{14}^{B_4}, & u_2 > 0.5, u_3 > 0.5 \end{cases}, \tag{6.31}$$

where the tessellation is as in Figure 6.4 and $c_{14}^{B_1}, c_{14}^{B_4}$ are Gauss copulas with parameter $\rho = -0.5, 0.5$, respectively, and $c_{14}^{B_2}, c_{14}^{B_3}$ are Frank copulas with parameter $\theta = 5, 10$, respectively.

In the next section, we develop an algorithm to estimate the introduced non-simplified vine structure from a sample.

6.4 Estimation

In this section, we deal with the estimation of non-simplified vine copulas with tessellated conditioning spaces. We first assume the vine structure (i.e., the vine array $\mathbf{A}$) to be known. Similar to the simulation algorithm, the estimation framework can be adapted from simplified vine copula estimation.

For simplified vine copulas with known vine structure, there are parametric (Joe, 2005; Hobæk Haff, 2013), semiparametric (Genest et al., 1995;

Algorithm 5 Simulation of d-dimensional non-simplified R-vine

```
Input ordered vine array $\mathbf{A} = (a_{kj})$
Initialize three auxiliary $d \times d$ arrays $(q_{ij})$ , $(v_{ij})$, and $(z_{ij})$ as well as an auxiliary $(d-1) \times (d-1)$ array **fam** that contains in each row $k$ the indices of the copulas in tree $T_k$ as they appear in the ordered R-vine array $\mathbf{A}$ and an auxiliary array $\mathbf{B}$ that contains a data structure to select the corresponding tessellation in **fam**.
Compute upper-triangular matrix $\mathbf{M} = (m_{kj})$, where $m_{kj} = \max\{a_{1j}, \dots, a_{kj}\}$, for $k = 1, \dots, j-1$, $j = 2, \dots, d$.
Generate $d$ independent uniformly distributed random variables $w_1, \dots, w_d$.
Set $u_1 = w_1$, $u_2 = h^{-1}_{\mathbf{fam}(1,1)}(w_2, w_1)$, $q_{22} = w_2$, $v_{12} = h_{\mathbf{fam}(1,1)}(u_1, u_2)$.
Select the tessellations according to $u_1, u_2$ from $\mathbf{B}$.
for $j = 3, \dots, d$
    $q_{jj} = w_j$
    for $l = j-1, \dots, 2$
        if $a_{lj} = m_{lj}$ then $s = q_{l,a_{lj}}$, else $s = v_{l-1,m_{lj}}$
        $z_{lj} = s$
        $q_{lj} = h^{-1}_{\mathbf{fam}(l,j-1),\mathbf{B}(l,j-l)}(q_{l+1,j}, s)$
    end for
    $q_{1j} = h^{-1}_{\mathbf{fam}(1,j-1)}(q_{2j}, u_{a_{1j}})$
    $u_j = q_{1j}$
    Select the tessellations according to $u_1, \dots, u_j$ from $\mathbf{B}$.
    $v_{1j} = h_{\mathbf{fam}(1,j-1)}(u_{a_{1j}}, u_j)$
    for $l = 2, \dots, j-1$
        $v_{lj} = h_{\mathbf{fam}(l,j-1),\mathbf{B}(l,j-l)}(z_{lj}, q_{lj})$
    end for
end for
Return $(u_1, \dots, u_d)$.
```

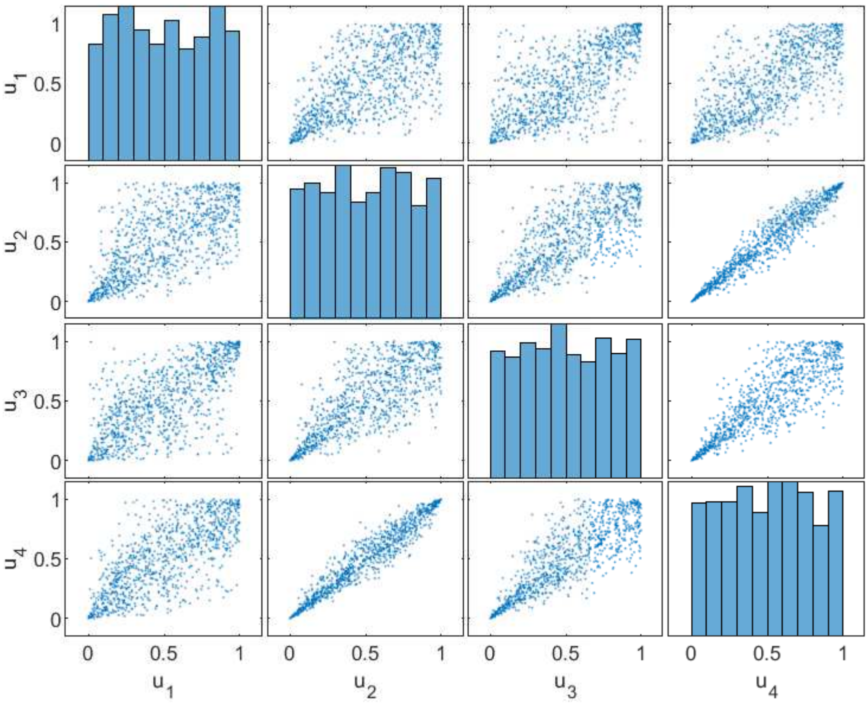

Figure 6.6: Scatter plot matrix of a sample from the vine structure depicted in Figure 6.5.

Hobæk Haff, 2013), and nonparametric (Hobæk Haff and Segers, 2015; Nagler and Czado, 2016) estimation approaches. We focus on the so-called stepwise semiparametric (SSP) method as introduced in Aas et al. (2009) and formally treated in Hobæk Haff (2013). The estimator is consistent (Hobæk Haff, 2013) and has good finite sample performance compared to other simplified vine copula estimators (Hobæk Haff, 2012).

The SSP is semiparametric because it estimates the copulas parametrically and the univariate marginal distributions via rank transformation of the original sample. Thus, it circumvents model specification errors from the univariate marginal distributions (Genest et al., 1995; Hobæk Haff, 2013). It is stepwise because it assumes that the log-likelihood of the vine copula can be maximized by a stepwise maximization of the log-likelihoods of the single copulas in a vine structure. Let θ_i be the parameter vector of the copulas in tree T_i, $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be a d-dimensional sample of size n, and $\mathbf{U}_1, \ldots, \mathbf{U}_n$ be the corresponding pseudo-observations. Furthermore, let the notation be as in Equation (6.9) for an R-vine density. Then, the log-likelihood function we maximize can be written as

$$\ell(\theta_1, \ldots, \theta_{d-1}; \mathbf{U}_1, \ldots, \mathbf{U}_n) = \sum_{i=1}^{n} \sum_{j=1}^{d-1} \sum_{e \in E_j} \log(c_{l_e, r_e | D_e}). \tag{6.32}$$

The SSP estimator now makes use of the fact that in order to estimate the parameters θ_i on level i of the vine, only the estimated parameter vectors $\hat{\theta}_1, \ldots, \hat{\theta}_{i-1}$ of the previous levels $1, \ldots, i-1$ are necessary. Estimation of the copula family can be incorporated by using, e.g., the Akaike Information Criterion (AIC) (Akaike, 1973). The SSP estimation process for simplified vine copulas is summarized in Algorithm 6.

Algorithm 6 Stepwise Semi-Parametric Estimation of d-dimensional Simplified Vine Copula.

```
1: Input sample $\mathbf{X}_1, \ldots, \mathbf{X}_n$ and vine structure $\mathbf{A}$.
2: Rank transform the sample $\mathbf{X}_1, \ldots, \mathbf{X}_n$ to pseudo-observations $\hat{\mathbf{U}}_1, \ldots, \hat{\mathbf{U}}_n$.
3: for $j = 1, \ldots, d-1$
4:     Estimate $\theta_j$ by maximizing $\ell(\hat{\theta}_1, \ldots, \hat{\theta}_{j-1}, \theta_j; \hat{\mathbf{U}}_1, \ldots, \hat{\mathbf{U}}_n)$. Choosing the copula family can be incorporated by using the AIC (Akaike, 1973).
5:     Use $\hat{\theta}_j$ and the $h$-functions of the copulas in tree $T_j$ to obtain the pseudo-observations of the next level $j+1$.
6: end for
7: Return $\hat{\theta}_1, \ldots, \hat{\theta}_i$.
```

Estimation of a non-simplified vine with tessellated conditioning spaces affects Algorithm 6 in Steps 4 and 5. Not only the tessellation $\mathbb{B}$ has to be estimated in order to maximize the log-likelihood function but also the tessellation influences which h-function is used to obtain the pseudo-observations for the next vine level. Two main issues have to be addressed for estimating the tessellation $\mathbb{B}$ of a given conditional copula. First, we have to deal with finding the points at which the conditioning space is partitioned, i.e., the points at which a dependence change is present. And second, we have to find the sequence in which the variables are partitioned since this potentially influences the resulting tessellation. In the following, we discuss some general concepts to address these issues.

Finding the points at which the conditioning space is partitioned boils down to finding a distributional change. This can be achieved, e.g., by statistical tests, such as a permutation test for Kendall's τ (Janssen and Pauls, 2003; Omelka and Pauly, 2012; Chung and Romano, 2013; DiCiccio and Romano, 2017) or the test by Remillard and Scaillet (2009), which compares two empirical copulas. Furthermore, techniques from change-point detection (Hawkins, 1987; Chen and Gupta, 1997; Guégan and Zhang, 2010) can be utilized. An important concept we will frequently employ is equal-frequency binning, which is a well-established strategy for dealing with continuous variables in decision trees (Dougherty et al., 1995). It allows us to artificially partition the conditioning space into bins and compare the copulas of the bins to each other.

Although not investigated in detail in this thesis, we want to comment on consistency in the following. The problem of finding a tessellation is similar to the problem of finding the vine structure (cf. Section 6.4.6). In both cases a structure is determined in a first step and the copulas within the imposed structure are estimated via likelihood maximization in a second step. Additionally, with the exception of the change-point detection approach (cf. Section 6.4.4), the proposed methods below do not rely on maximum likelihood to estimate the tessellation. Just as in the case of estimating a vine structure, a proper measure of distance between the theoretical tessellation and its estimate is difficult to obtain. In summary, it is difficult to define what would actually be meant by consistency of the tessellation of a conditioning space. Consistency is further discussed in Section 6.4.4 when the method relying on change-point detection is introduced.

Based on the general concepts above, in the next sections we develop specific techniques to estimate the tessellation. For this, we consider, without loss of generality, the estimation of the bivariate conditional copula $c_{12|3,\ldots,d}$ on a sample of size n. Thus, the conditioning space for which we want to estimate the tessellation is $S_{d-2} = [0,1]^{d-2}$. Note that using the usual notation, $c_{12|3,\ldots,d}$ is the copula of the random variables $U_{1|3,\ldots,d}$ and $U_{2|3,\ldots,d}$. We deal with finding the tessellation given a known sequence of variables first and subsequently extend this to finding the variable sequence simultaneously. For notational simplicity, let the given variable sequence be $U_3, \ldots, U_d$, i.e., analogous to Figure 6.4 we want to estimate a tessellation $\hat{\mathbb{B}}$ of S_{d-2} that can be represented as a decision tree. In the following, we develop four approaches to estimate the tessellation.

6.4.1 Tessellation Estimation with Predefined Equal-Frequency Binning

The first approach is very simple and estimates a tessellation by a predefined equal-frequency binning (Dougherty et al., 1995) of the conditioning space.

We denote this estimator by EF for the rest of the thesis. In particular, we propose to use $k_1 = \min\left\{10, \left\lfloor \frac{n}{100} \right\rfloor\right\}$ equal-frequency bins, $b_1^{(3)}, \dots, b_{k_1}^{(3)}$, for the first variable U_3. That is, at most 10 bins are created. The same is repeated for the next variable U_4, such that on each of the bins $b_1^{(3)}, \dots, b_{k_1}^{(3)}$, $k_2 = \min\left\{10, \left\lfloor \frac{n_{b_1^{(3)}}}{100} \right\rfloor\right\}$ bins are created, where $n_{b_1^{(3)}}$ is the number of points with $U_3 \in b_1^{(3)}$. This is repeated for all conditioning variables until either all variables are binned or no more bins can be created because the number of points is below 100 per bin. Hence, the minimum number of points in one bin is 100 and we need at least 200 points in order to create two bins, which ensures a sufficient sample size for the parametric copula estimation. On the resulting tessellation a copula is estimated on each partition.

Note that this approach is very simplistic and leads to a tessellation with equally many points in each bin. However, as will become clear shortly, it has an advantage over the other presented methods because it can deal computationally very efficiently with large, high-dimensional data sets due to its predefined nature.

6.4.2 Tessellation Estimation with Permutation Tests

The second estimation approach uses a permutation test (Janssen and Pauls, 2003; Omelka and Pauly, 2012; Chung and Romano, 2013; DiCiccio and Romano, 2017) on Kendall's τ between the variables $U_{1|3,\dots,d}$ and $U_{2|3,\dots,d}$ and is denoted by PT for the rest of the thesis. We start with the first conditioning variable U_3 and divide its space into k bins, $b_1^{(3)}, \dots, b_k^{(3)}$. The bins are placed using equal-frequency binning. The number of bins is determined by $k = \left\lfloor \frac{n}{100} \right\rfloor$.

Now, we try to reduce the number of bins by testing for a difference of Kendall's τ in each pair of adjacent bins $b_i^{(3)}$ and $b_{i+1}^{(3)}$. This is done using permutation tests, i.e., in total $k-1$ tests. In case the null hypothesis of equal Kendall's τ cannot be rejected, we merge the two adjacent bins $b_i^{(3)}$ and $b_{i+1}^{(3)}$ into one bin $b_{i,i+1}^{(3)}$. When two or more consecutive tests do not reject the null hypothesis we merge all the involved bins $b_i^{(3)}, \dots, b_{i+l}^{(3)}$ into a new bin $b_{i,\dots,i+l}^{(3)}$. This is a simplification. One could also test again as soon as two bins are merged into a new bin, e.g., conduct a permutation test on $b_{i,i+1}^{(3)}$ and $b_{i+2}^{(3)}$. However, note that in such a step-wise aggregation setting the sequence of tested intervals is important and changing it could result in different bins. The remaining $l \leq k$ bins are renumbered according to $b_{(1)}^{(3)}, \dots, b_{(l)}^{(3)}$.

This procedure is then repeated for the next conditioning variable U_4.

Note that now the (remaining) bins $b^{(3)}_{(1)}, \ldots, b^{(3)}_{(l)}$, $l \leq k$, from the first variable have to be considered in turn, i.e., the described procedure is repeated for U_4 on each relabeled bin $b^{(3)}_{(i)}$ of the first variable U_3. By generating new bins for each conditioning variable in turn, the data set is cut into smaller and smaller pieces with fewer and fewer points. Note that we decided to have at least 100 points in each bin. Thus, at some point we cannot partition the conditioning space any further and the resulting bins make up the estimated tessellation $\hat{\mathbb{B}}$. This might occur before all the conditioning variables are partitioned and the procedure stops prematurely at this point. Thus, not all the conditioning variables have to be partitioned. As soon as the procedure stops, a copula is estimated on each partition of the resulting tessellation $\hat{\mathbb{B}}$.

In order to account for the multiple testing problem in this procedure, we use the Holm-Bonferroni method (Holm, 1979) to control the significance level α. It is more powerful than the usual Bonferroni procedure, which uses $\frac{\alpha}{k-1}$ to adjust for multiple testing (Aickin and Gensler, 1996; Bender and Lange, 2001). Moreover, in contrast to the methods of Šidák (1967) or Hochberg (1988), it neither imposes an independence assumption nor certain kinds of dependence structures. The method works by first ordering the p-values obtained from the $k-1$ permutation tests from lowest to highest. Let these p-values and the corresponding null hypotheses be denoted by $p_{(1)}, \ldots, p_{(k-1)}$ and $H_{(1)}, \ldots, H_{(k-1)}$, respectively. Then, determine the minimal index r for which $p_{(r)} > \frac{\alpha}{k-r}$ and reject all hypotheses $H_{(1)}, \ldots, H_{(r-1)}$. For $r = 1$ no null hypothesis is rejected and if no such r exists all null hypotheses are rejected. The correction is used for the first variable and on each bin for subsequent variables anew. This is sensible since, first, the total number of final partitions is unknown in advance and, second, each partition represents a new testing problem.

Finally, note that in the PT estimator Kendall's τ is tested. Thus, we expect that a dependence change, which involves different copula families exhibiting the same Kendall's τ, is not detected by this method.

6.4.3 Tessellation Estimation with the Remillard and Scaillet (2009) Test

The third estimator works exactly as the PT approach but uses the test by Remillard and Scaillet (2009) instead of a permutation test. We denote the estimator by RS for the rest of the thesis. The Remillard and Scaillet (2009) test compares the empirical copulas of two samples for equality. Hence, the test is potentially able to detect cases where different copula families exhibit the same Kendall's τ. Note that the Remillard and Scaillet (2009) test is computationally extremely burdensome. Thus, it is imperative to compare two samples which have at most 200 points each. This is automatically ensured by choosing the number of bins according to the rule outlined in the PT approach.

6.4.4 Tessellation Estimation with Change-Point Detection

The fourth estimation approach uses techniques from change-point detection (Hawkins, 1987; Chen and Gupta, 1997; Guégan and Zhang, 2010) and is denoted by CP for the rest of the thesis. Change-point detection is a wide field. There is some work on change-point detection for distribution functions and particularly copulas (Holmes et al., 2013; Bücher et al., 2014). However, these methods are computationally very burdensome due to the use of a multiplier bootstrap to calculate p-values (Bücher and Kojadinovic, 2016a,b). In addition to that, these methods only detect one change-point at a time. Therefore, we adapt the method by Chen and Gupta (1997) which combines the so-called binary splitting – a consistent method introduced by Vostrikova (1981) – and the Bayesian Information Criterion (BIC) (Schwarz, 1978).

We begin with the conditioning variable U_3 and estimate a copula $\hat{C}_{12|3,\dots,d}$ of $U_{1|3,\dots,d}$ and $U_{2|3,\dots,d}$ on the whole data set. Then, we compute its BIC denoted by $BIC(n)$, which serves as a reference point. Let $u_{3,(1)},\dots,u_{3,(n)}$ be the ordered sequence of U_3 values in the data set. For $t = 100,\dots,n-100$, split the data set into two parts fulfilling $U_3 \in \{u_{3,(1)},\dots,u_{3,(t)}\}$ and $U_3 \in \{u_{3,(t+1)},\dots,u_{3,(n)}\}$, estimate a copula of $U_{1|3,\dots,d}$ and $U_{2|3,\dots,d}$ on both, and compute the total BIC of the two estimates. In total, $n-199$ copula tuples are estimated with respective total BIC denoted as $BIC(100),\dots,BIC(n-100)$. Then, determine

$$\hat{t} = \operatorname*{arg\,min}_{100 \le t \le n-100} BIC(t). \tag{6.33}$$

If $BIC(n) < BIC(\hat{t})$ there is no change point in the data set and we do not split. Otherwise, there is a change point at $\hat{t}$ and we split the data set into two sets fulfilling $U_3 \in \{u_{3,(1)},\dots,u_{3,(\hat{t})}\}$ and $U_3 \in \{u_{3,(\hat{t}+1)},\dots,u_{3,(n)}\}$. On these two sets the new reference copulas are estimated and the procedure is repeated analogously on each. This is done until either no new change point is detected or the number of points to detect a change point is below 200. This generates an ensemble of bins $b_1^{(3)},\dots,b_l^{(3)}$ on which in the next step the procedure is repeated for variable U_4. Again the estimation of the tessellation is stopped as soon as all conditioning variables are treated or because there are too few points, i.e., less than 200 per bin.

In contrast to the previous two approaches which aggregate intervals and are thus bottom-up, binary splitting is a top-down method. Its disadvantage is that we have to try each point in turn which is computationally burdensome. However, since we have to estimate the copulas on each partition in order to detect change points, we do not have to re-estimate the copula on each part of the resulting estimated tessellation $\hat{\mathbb{B}}$. In the CP estimator both the tessellation and the copulas are estimated at the same time, whereas this is done sequentially in the previous methods. This is a trait of CP which can be used for a consistency result.

Since binary splitting is consistent (Vostrikova, 1981), a consistency result for the CP estimator seems feasible. However, there are several obstacles which have to be considered. First, it is important to ensure the overall consistency of the stepwise semiparametric estimator. Second, the interplay of estimating the tessellation and estimating the copula on each partition has to be taken into account. Third, if there is more than one conditioning variable, choosing the variable sequence influences the estimator and thus consistency. As soon as the variable sequence is chosen, it is fixed and a wrong variable order will (potentially) yield a wrong tessellation. Hence, we would need a variable sequence estimator that somehow chooses the correct sequence of variables directly or, alternatively, with a rate that increases for increasing sample size. As will become clear in the next section, it is difficult to choose a variable sequence and, thus, even more difficult to find a criterion which fits into a consistency framework. However, we conjecture that the CP estimator is consistent for 3-dimensional vine copulas, where the variable sequence does not play a role.

6.4.5 Estimation of the Variable Sequence

The remaining issue is how to choose the variable sequence. It is clear that the variable sequence may have a profound effect on the resulting tessellation because we split each variable only once and, therefore, different sequences can lead to different estimated tessellations. We suggest a heuristic for selecting a variable sequence which is motivated by heuristics that use Kendall's τ to estimate the structure of a vine copula, e.g., cf. Dißmann et al. (2013). First, all the conditioning variables are equal-frequency binned as in the PT estimator above. On each bin, Kendall's τ between $U_{1|3,\ldots,d}$ and $U_{2|3,\ldots,d}$ is calculated. The variable with the highest absolute difference of Kendall's τ between adjacent bins is first in the sequence. The second variable in the sequence is the one with the second highest absolute difference of Kendall's τ between adjacent bins, and so on. Thus, when using the approaches above, we maximize the possibility of creating lot of bins for the first variable, leading to fewer points in the subsequent pass for the next variable. This maximizes the chance to stop the procedures prematurely, i.e., before all variables have to be considered.

6.4.6 Estimation of the Vine Structure

To conclude this section, we deal with the case of an unknown vine structure. Since this is not the main focus of this chapter, we give a prospect of this topic only. Vine structure selection is an active research topic. Since the number of vines increases exponentially with the dimension d (cf. Section 6.1.1), trying out all possible vine structures is computationally infeasible. Instead, there is a number of heuristics which can be employed. Kurowicka (2011) suggests a top-down approach that tries to infer the vine structure from partial correlations. Also, there are some Bayesian approaches that treat the vine structure as a latent variable (Czado et al., 2013). One of the most prominent heuristics is the bottom-up algorithm by Dißmann et al. (2013) which we extend in the following. Recently, the algorithm was extended to create a vine that is as least non-simplified as possible (Kraus and Czado, 2017) by using the test statistic for non-simplified vines developed by Kurz and Spanhel (2018).

The algorithm by Dißmann et al. (2013) tries to exploit the fact that the (conditional) copulas in lower trees influence the dependence structure of the vine the most (Joe, 2011a,b). Thus, the algorithm maximizes the Kendall's τs of the (conditional) copulas in each tree. This is done computing a maximal spanning tree via Prim's algorithm (Prim, 1957; Dijkstra, 1959) and extracting the entries for the vine array from this. The estimation of the (conditional) copulas and the preparation of the pseudo-observations for the next tree level is done according to Algorithm 6. Algorithm 7 shows an extended version of the procedure by Dißmann et al. (2013), which accounts for estimation of a tessellation of conditioning spaces.

In the next section, we conduct a simulation study in order to investigate the different tessellation estimation strategies. Furthermore, we compare the non-simplified vine estimation to simplified vines.

6.5 Simulation Study

We split the simulation study into two parts. The first part comprises an analysis of the tessellation estimation techniques as outlined in Section 6.4. In particular, we compare these in order to determine whether there is a best strategy for the estimation task at hand. The second part explores how

Algorithm 7 Extension of the Algorithm by Dißmann et al. (2013) for Non-Simplified Vines.

```
Input sample $\mathbf{X}_1, \ldots, \mathbf{X}_n$.
Rank transform the sample $\mathbf{X}_1, \ldots, \mathbf{X}_n$ to pseudo-observations $\hat{\mathbf{U}}_1, \ldots, \hat{\mathbf{U}}_n$.
Calculate the Kendall's $\tau$s $\hat{\tau}_{i,j}$ for each variable pair $\{i, j\}$.
Compute the spanning tree which maximizes the sum of absolute Kendall's $\tau$s via Prim's algorithm (Prim, 1957; Dijkstra, 1959). Let $e = \{i, j\}$ be the edges of the spanning tree. Thus,
    $$\max_{i,j} \sum_e |\hat{\tau}_e|.$$
For each edge $e = \{i, j\}$ in the spanning tree estimate a copula $\hat{c}_{ij}$ and transform the observations via the corresponding $h$-functions for the next level.
for $k = 2, \ldots, d - 1$
    Calculate the Kendall's $\tau$s $\hat{\tau}_{i,j|D_e}$ for each conditional variable pair that can be part of the tree $T_k$. This is determined by the proximity condition (cf. Definition 9, Property 2).
    Compute the spanning tree which maximizes the sum of absolute Kendall's $\tau$s of these edges, i.e.,
        $$\max_{i,j|D_e} \sum_e |\hat{\tau}_e|.$$
    For each edge $e = \{i, j|D_e\}$ in the spanning tree, estimate a non-simplified conditional copula $\hat{c}_{ij|D_e}$ with tessellated conditioning spaces and transform the observations via the corresponding $h$-functions for the next level.
end for
Return estimated vine array $\hat{\mathbf{A}}$.
```

the non-simplified vine with tessellation of conditioning spaces compares to simplified vines in different estimation settings. This part of the simulation study shows how the developed non-simplified vine can complement simplified vine estimation. Furthermore, it gives insight into the question on when it is important to use a non-simplified vine model.

6.5.1 Tessellation Strategy

In this part of the simulation study, we analyze which of the tessellation estimation approaches based on a permutation test (PT), based on the Remillard and Scaillet (2009) test (RS), and based on change-point detection (CP) works best. Note that we leave out the equal-frequency (EF) estimator since it operates on a predefined tessellation and is, thus, incomparable to the other methods which adapt to a specific sample. We consider six scenarios in total. Five of these utilize a 3-dimensional D-vine and one uses a 4-dimensional D-vine. Each scenario comprises several settings which vary in terms of the tessellation of the conditional copulas. For each setting in each scenario a sample of size $n = 1,000$ is generated and the estimators PT, RS, and CP are employed. The set of copulas out of which the best fitting copula is chosen comprises the t-, Gauss, Clayton, Gumbel, and Frank copulas. The number of Monte Carlo runs is 200. In the following, Scenarios 1 to 6 are described in detail. Table 6.1 gives a concise overview.

In Scenarios 1 to 5 we simulate from the 3-dimensional D-vine

$$c_{123} = c_{12} c_{23} c_{13|2}, \tag{6.34}$$

where the conditional copula $c_{12|3}$ contains a scenario-specific tessellation $\mathbb{B}$ of its conditioning space $S_1 = [0, 1]$. The copulas c_{12} and c_{23} are fixed throughout the scenarios to a Gauss and Gumbel copula with parameters that correspond to a Kendall's τ of 0.5, respectively. For the estimation task, these are assumed to be known such that we do not deal with model misspecification from the first tree T_1. Thus, the correct pseudo-observations for copula $c_{12|3}$ can be calculated in each sample and we can compare the tessellation strategies in a controlled environment. Considering that the used sample size of $n = 1,000$ is high, we expect our results to be robust in a realistic estimation setting, where the true copulas in the lower trees are unknown.

Scenarios 1–4 are structured in a way that the dependence structure in adjacent partitions is increasingly different in subsequent scenarios. In contrast to that, Scenario 5 covers the case, where the dependence structures only differ through the copula families. In Scenarios 1 and 2, each part of the

	Setting	Tessellation	Copulas	Kendall's τ
Scenario 1	1	$[0,0.5] - (0.5,1]$	G – G	$0.4 - 0.8$
	2	$[0,0.5] - (0.5,1]$	C – C	$0.4 - 0.8$
	3	$[0,0.4] - (0.4,0.8] - (0.8,1]$	G – G – G	$0.4 - 0.5 - 0.8$
	4	$[0,0.4] - (0.4,0.8] - (0.8,1]$	C – C – C	$0.4 - 0.5 - 0.8$
	5	$[0,0.2] - (0.2,0.4] - (0.4,0.8] - (0.8,1]$	G – G – G – G	$0.2 - 0.4 - 0.6 - 0.8$
	6	$[0,0.2] - (0.2,0.4] - (0.4,0.8] - (0.8,1]$	C – C – C – C	$0.2 - 0.4 - 0.6 - 0.8$
Scenario 2	1	$[0,0.5] - (0.5,1]$	C – Gu	$0.4 - 0.8$
	2	$[0,0.5] - (0.5,1]$	t – G	$0.4 - 0.8$
	3	$[0,0.4] - (0.4,0.8] - (0.8,1]$	C – Gu – C	$0.4 - 0.5 - 0.8$
	4	$[0,0.4] - (0.4,0.8] - (0.8,1]$	t – G – t	$0.4 - 0.5 - 0.8$
	5	$[0,0.2] - (0.2,0.4] - (0.4,0.8] - (0.8,1]$	C – Gu – C – Gu	$0.2 - 0.4 - 0.6 - 0.8$
	6	$[0,0.2] - (0.2,0.4] - (0.4,0.8] - (0.8,1]$	t – G – t – G	$0.2 - 0.4 - 0.6 - 0.8$
Scenario 3	1	$[0,0.5] - (0.5,1]$	G – G	$(-0.6) - 0.6$
	2	$[0,0.5] - (0.5,1]$	t – t	$(-0.6) - 0.6$
	3	$[0,0.4] - (0.4,0.8] - (0.8,1]$	G – G – G	$(-0.5) - 0.5 - (-0.5)$
	4	$[0,0.4] - (0.4,0.8] - (0.8,1]$	t – t – t	$(-0.5) - 0.5 - (-0.5)$
	5	$[0,0.2] - (0.2,0.4] - (0.4,0.8] - (0.8,1]$	G – G – G – G	$0.5 - (-0.3) - (-0.7) - 0.5$
	6	$[0,0.2] - (0.2,0.4] - (0.4,0.8] - (0.8,1]$	t – t – t – t	$0.5 - (-0.3) - (-0.7) - 0.5$
Scenario 4	1	$[0,0.5] - (0.5,1]$	t – Gu	$(-0.6) - 0.6$
	2	$[0,0.5] - (0.5,1]$	t – G	$(-0.6) - 0.6$
	3	$[0,0.4] - (0.4,0.8] - (0.8,1]$	G – Gu – t	$(-0.5) - 0.5 - (-0.5)$
	4	$[0,0.4] - (0.4,0.8] - (0.8,1]$	t – G – t	$(-0.5) - 0.5 - (-0.5)$
	5	$[0,0.2] - (0.2,0.4] - (0.4,0.8] - (0.8,1]$	C – t – G – Gu	$0.5 - (-0.3) - (-0.7) - 0.5$
	6	$[0,0.2] - (0.2,0.4] - (0.4,0.8] - (0.8,1]$	t – G – t – G	$0.5 - (-0.3) - (-0.7) - 0.5$
Scenario 5	1	$[0,0.5] - (0.5,1]$	C – Gu	$0.7 - 0.7$
	2	$[0,0.5] - (0.5,1]$	t – G	$0.7 - 0.7$
	3	$[0,0.5] - (0.5,1]$	t – G	$(-0.7) - (-0.7)$
	4	$[0,0.5] - (0.5,1]$	F – G	$(-0.7) - (-0.7)$
	5	$[0,0.4] - (0.4,0.8] - (0.8,1]$	C – Gu – G	$0.7 - 0.7 - 0.7$
	6	$[0,0.4] - (0.4,0.8] - (0.8,1]$	G – C – t	$0.7 - 0.7 - 0.7$
	7	$[0,0.4] - (0.4,0.8] - (0.8,1]$	t – F – G	$(-0.7) - (-0.7) - (-0.7)$
	8	$[0,0.4] - (0.4,0.8] - (0.8,1]$	t – G – F	$(-0.7) - (-0.7) - (-0.7)$
Scenario 6	1	$[0,0.5] \times [0,0.4] - (0.5,1] \times [0,0.6] - (0.5,1] \times (0.6,1] - [0,0.5] \times (0.4,1]$	C – Gu – G – G	$0.3 - 0.7 - 0.3 - 0.7$
	2	$[0,0.5] \times [0,0.4] - (0.5,1] \times [0,0.6] - (0.5,1] \times (0.6,1] - [0,0.5] \times (0.4,1]$	C – C – C – C	$0.7 - 0.2 - 0.5 - 0.2$
	3	$[0,0.5] \times [0,0.4] - (0.5,1] \times [0,0.6] - (0.5,1] \times (0.6,1] - [0,0.5] \times (0.4,1]$	t – C – G – C	$(-0.5) - 0.7 - (-0.5) - 0.7$
	4	$[0,0.5] \times [0,0.4] - (0.5,1] \times [0,0.6] - (0.5,1] \times (0.6,1] - [0,0.5] \times (0.4,1]$	t – t – t – t	$(-0.5) - 0.5 - (-0.5) - 0.5$
	5	$[0,0.5] \times [0,0.4] - (0.5,1] \times [0,0.6] - (0.5,1] \times (0.6,1] - [0,0.5] \times (0.4,1]$	C – G – F – Gu	$0.7 - 0.7 - 0.7 - 0.7$
	6	$[0,0.5] \times [0,0.4] - (0.5,1] \times [0,0.6] - (0.5,1] \times (0.6,1] - [0,0.5] \times (0.4,1]$	t – G – t – G	$(-0.7) - (-0.7) - (-0.7) - (-0.7)$

Table 6.1: Overview of simulation scenarios for the tessellation estimation. Each partition of a tessellation is separated by a hyphen. The third column shows the copulas in the same order as the tessellation. Also, the same order applies for the Kendall's τ in the last column. The following abbreviations for copulas are used: G – Gauss, C – Clayton, Gu – Gumbel, F – Frank. For the t-copula, the degrees of freedom are set to $\nu = 3$ throughout, i.e., only the second parameter ρ controls the strength of dependence.

tessellation is governed by a copula with positive Kendall's τ, where adjacent partitions exhibit a different Kendall's τ. The settings differ in the number of partitions (two, three, or four) and the copula families employed. In Scenario 1, each setting uses either a Gauss or a Clayton copula throughout, i.e., between partitions the copula family stays the same but the Kendall's τ varies. In contrast to that, in Scenario 2 also the copula family varies between adjacent partitions. Table 6.1 gives a detailed overview of the settings within Scenarios 1 and 2.

Scenarios 3 and 4 are similar to the previous two scenarios, however, involve also negative values for Kendall's τ. Scenario 3 features settings, where the partitions are governed by the same copula family (Gauss or t) but exhibit a different Kendall's τ in adjacent partitions. Scenario 4 comprises settings where also the copula families in adjacent partitions of the tessellation are different. Finally, Scenario 5 comprises settings where the copula families between adjacent partitions differ but Kendall's τ is constant over the whole tessellation. Table 6.1 gives a detailed overview of the settings within Scenarios 3, 4, and 5.

In Scenario 6, we simulate from the 4-dimensional D-vine

$$c_{1234} = c_{12}c_{23}c_{34}c_{13|2}c_{24|3}c_{14|23}, \tag{6.35}$$

where the conditional copula $c_{14|23}$ contains a tessellation of its conditioning space $S_2 = [0,1]^2$. The tessellation is fixed for all settings, the copula families and Kendall's τ vary, though. Copula c_{12} is a Gauss copula with parameter corresponding to a Kendall's τ of 0.5, copula c_{23} is a Clayton copula with parameter corresponding to a Kendall's τ of 0.4, and copula c_{34} is a Gumbel copula with parameter corresponding to a Kendall's τ of 0.7. The conditional copulas $c_{13|2}$ and $c_{24|3}$ are simplified and are both Gauss copulas with parameters translating to a Kendall's τ of 0.5 and -0.5, respectively. We assume the copulas in the first two trees T_1 and T_2 to be known, such that no model misspecification occurs for these trees. Again, due to the used sample size of $n = 1,000$, we expect the results to be valid for real estimation tasks. In addition to the tessellation strategy, the variable sequence is estimated in Scenario 6, as well. The settings within Scenario 6 differ in terms of Kendall's τ and the copula families in the partitions. Table 6.1 gives a detailed overview.

Figures 6.7 – 6.11 show the results for the 3-dimensional Scenarios 1 – 5. The bar charts show how often a change of dependence was detected at which level of the conditional variable U_2. The PT, RS, and CP estimators are depicted in orange, blue, and green, respectively. The black vertical lines represent the true tessellation in the respective setting. For PT and RS a significance level of $\alpha = 0.1$ is chosen. In the following, we interpret the results for each Scenario in turn.

Scenario 1 covers a change in Kendall's τ, where the copula family stays the same for all partitions of the tessellation. The three tessellation estimation approaches do equally well in the first two settings, where only two partitions are present and the difference in Kendall's τ is relatively high at 0.4. However, the performance of PT and RS dwindles as the difference in the Kendall's τ gets smaller and the number of partitions increases in Settings 3 – 6. The estimator CP is clearly superior to the latter two in these settings. PT and RS underestimate the number of partitions in these cases, whereas CP tends to overestimate the number of partitions in all settings. This overarching effect is further discussed below.

In Scenario 2 not only Kendall's τ changes but also the copula family. The estimation results are very similar to those of Scenario 1. In the first two settings the three approaches do equally well, whereas CP is superior in the last four settings. It is not surprising that PT yields similar results as in Scenario 1 since it can only detect changes in Kendall's τ by construction of the test. However, the RS estimator compares the empirical copulas and, thus, should perform better when two samples from different copulas are compared. In fact, the test tends to estimate the true change points in the tessellation slightly better in Scenario 2 compared to Scenario 1. Nonetheless, the RS estimator's overall estimation performance is disappointing.

Scenario 3 comprises a change from positive to negative Kendall's τ values, where the copula family stays the same for the whole tessellation. Across all settings the three approaches PT, RS, and CP perform equally well. This is no surprise since the absolute changes in Kendall's τ are quite big. The same outcome can be observed in Scenario 4, where additionally the copula family changes between partitions. Overall, the three approaches seem to perform equally well when the dependence strength (as measured by Kendall's τ) exhibits large changes between partitions.

In contrast to the previous scenarios, Scenario 5 comprises settings, where Kendall's τ is constant over the whole tessellation and only the copula family changes between the partitions. The estimation approach PT barely detects a dependence change at all. This is to be expected because by construction the permutation test only checks for a difference in Kendall's τ. Also, the RS estimator yields poor performance. This is surprising since the test should be able to detect different dependence structures since it works on the empirical copulas directly (Remillard and Scaillet, 2009). Furthermore, PT and RS do not hold their significance level, which is an additional hint that they are misspecified in such settings. The CP approach on the contrary obtains good results and is vastly superior to the other two estimators.

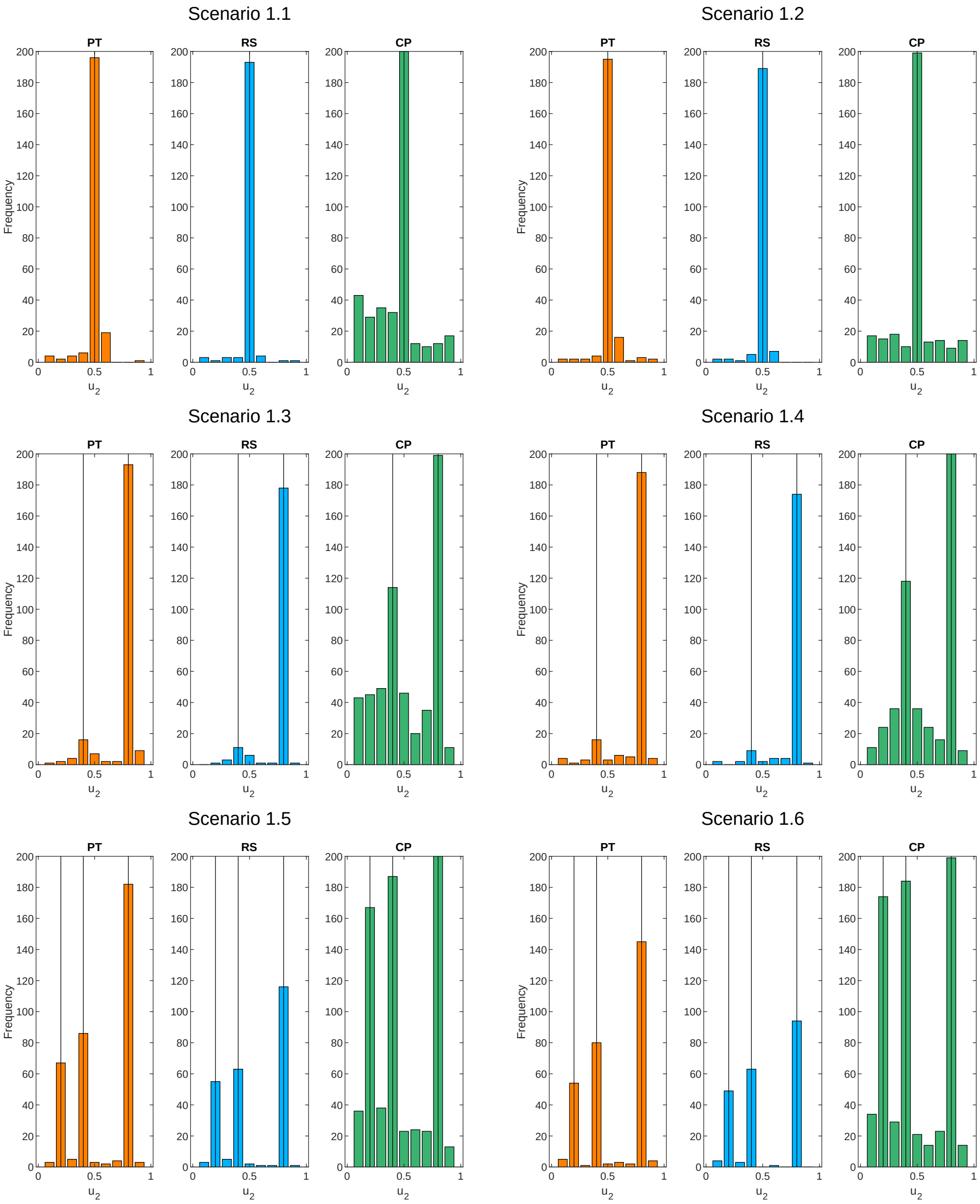

Figure 6.7: Tessellation estimation results for Scenario 1. The estimators PT, RS, and CP are depicted as orange, blue, and green bars, respectively. Black vertical lines indicate the true tessellation. Particularly in Settings 3 – 6, CP performs superior compared to PT and RS.

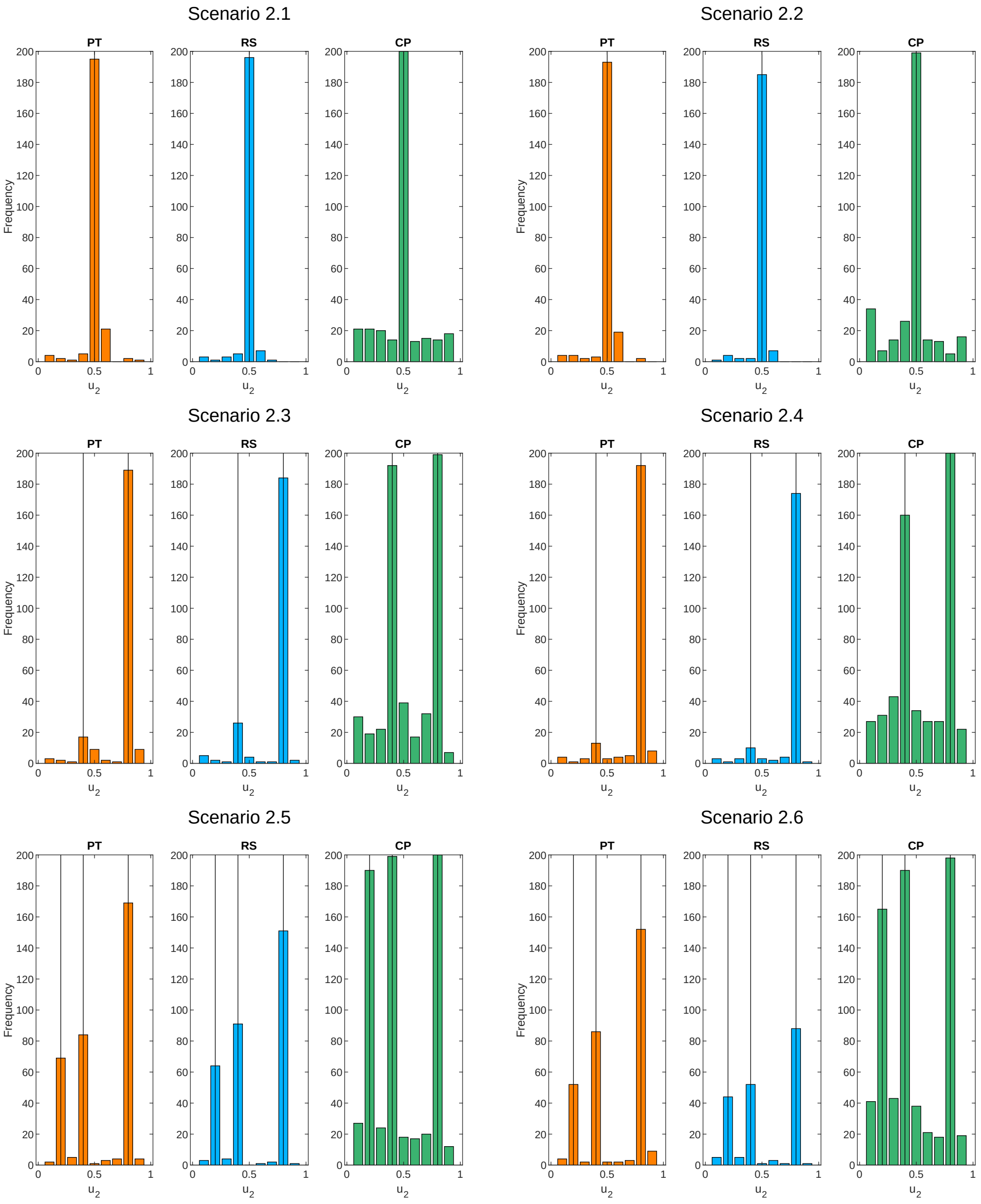

Figure 6.8: Tessellation estimation results for Scenario 2. The estimation approaches PT, RS, and CP are depicted as orange, blue, and green bars, respectively. Black vertical lines indicate the true tessellation. Just like in Scenario 1, CP performs much better than PT and RS in Settings 3 – 6.

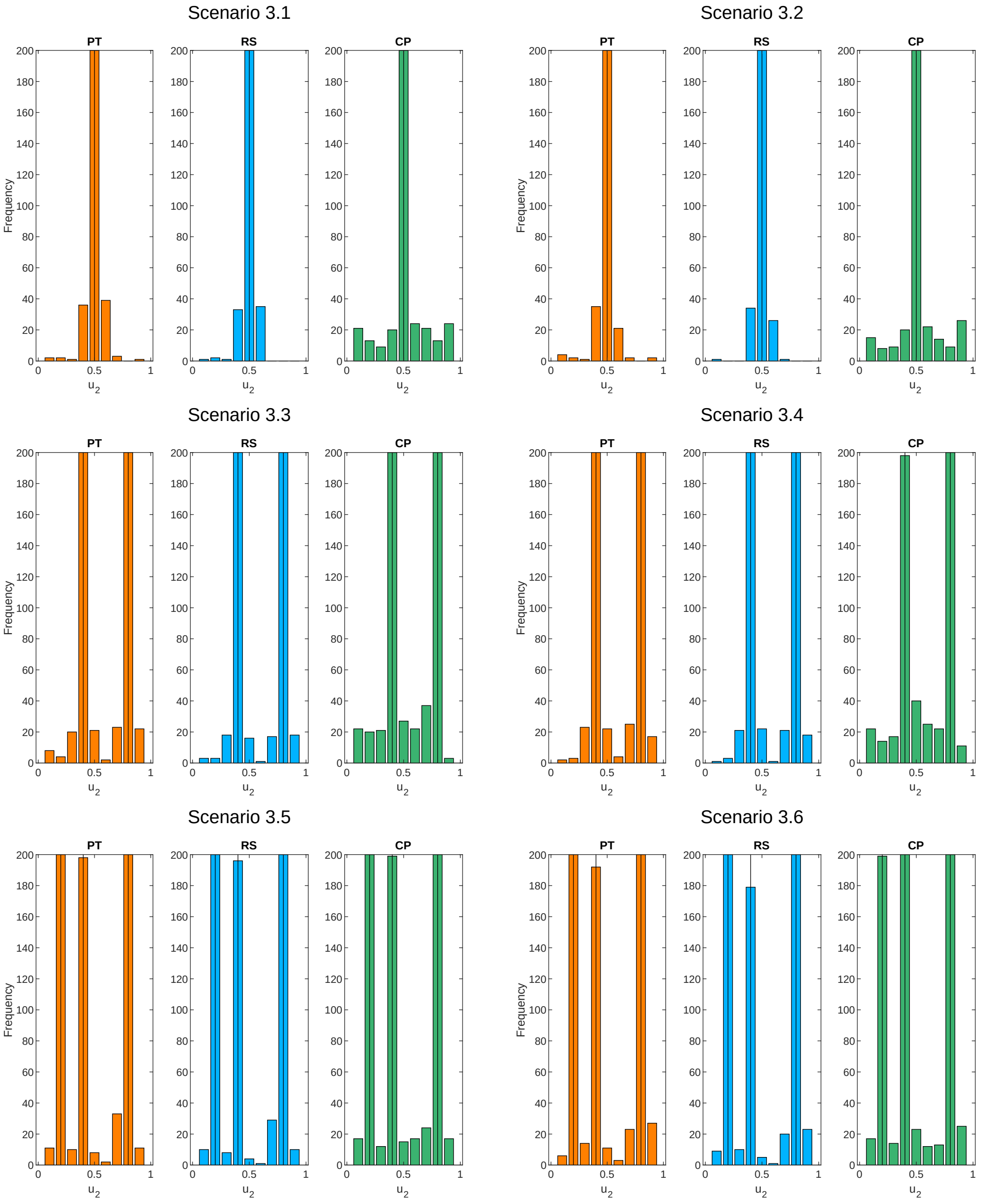

Figure 6.9: Tessellation estimation results for Scenario 3. The estimators PT, RS, and CP are depicted as orange, blue, and green bars, respectively. Black vertical lines indicate the true tessellation. The three approaches are on a par across all settings.

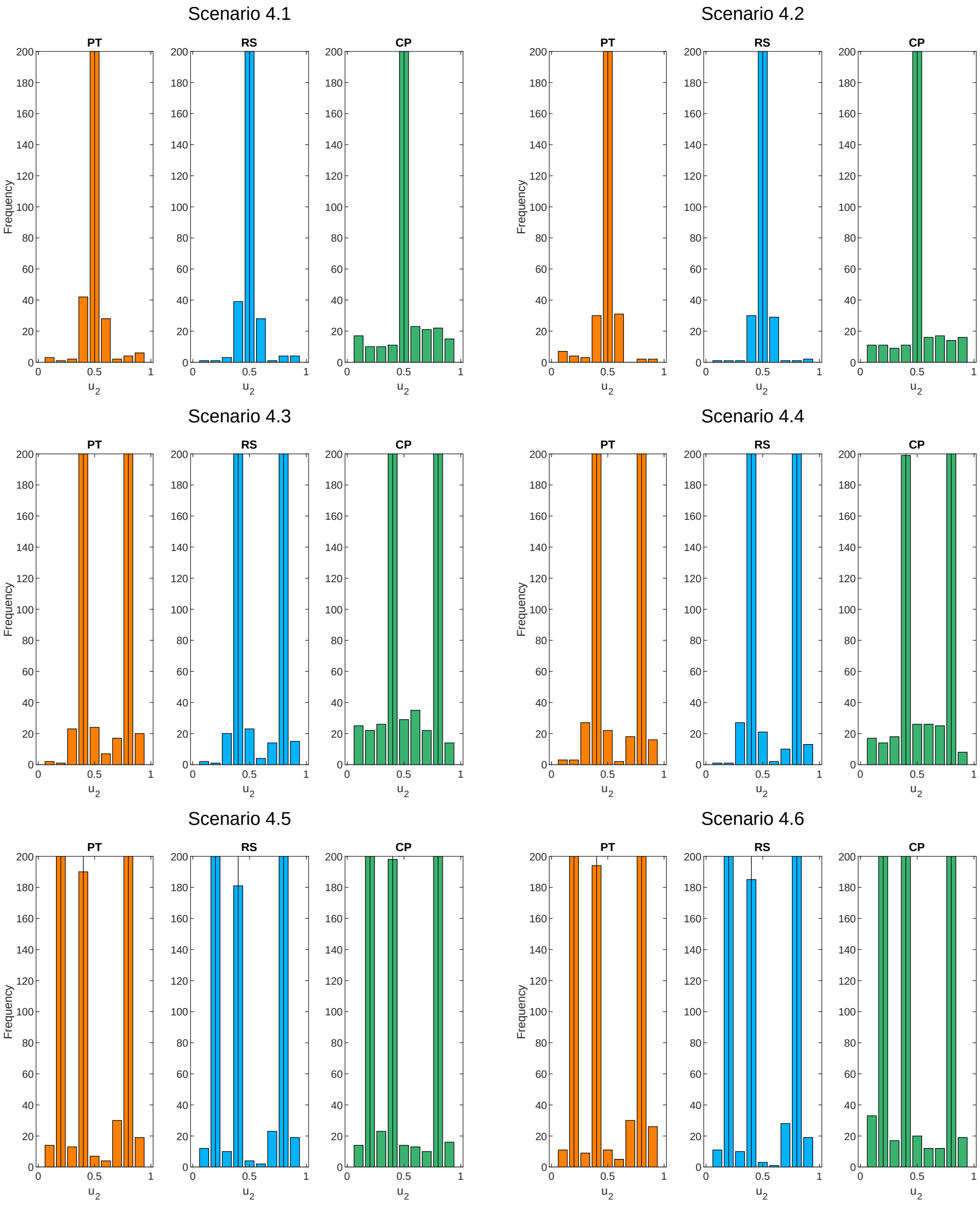

Figure 6.10: Tessellation estimation results for Scenario 4. The approaches PT, RS, and CP are depicted as orange, blue, and green bars, respectively. Black vertical lines indicate the true tessellation. Just like in Scenario 3, the three approaches are on a par across all settings.

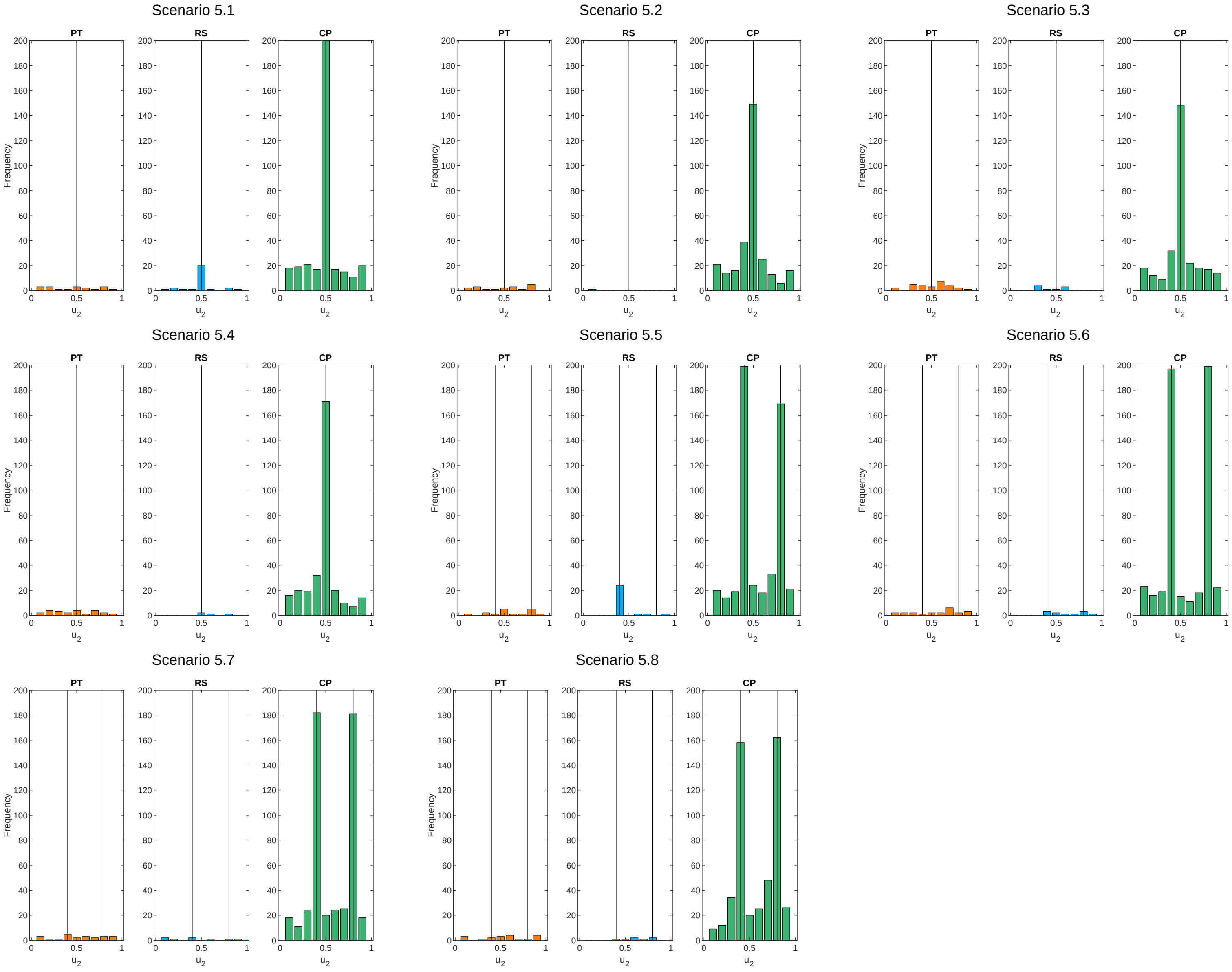

Figure 6.11: Tessellation estimation results for Scenario 5. The estimators PT, RS, and CP are depicted as orange, blue, and green bars, respectively. Black vertical lines indicate the true tessellation. CP clearly performs superior to PT and RS. The latter two hardly detect a dependence change on the conditioning space at all.

Table 6.2 shows the average number of detected dependence changes for each setting and estimator. One detected dependence change corresponds to two partitions in the tessellation, two detected dependence changes correspond to three partitions in the tessellation and so on. Overall, the estimation approaches PT and RS tend to underestimate the number of partitions in Scenarios 1, 2, and 5. In contrast to that, CP on average overestimates the number of partitions in all scenarios. Overestimation of the number of partitions is, however, not a severe issue. Estimating a change of dependence where the true model does not have one results in fitting two copulas on data generated from the same model. Hence, for increasing sample size, we expect these estimated copulas to be very close in dependence structure.

Another look at this can be taken from the bias-variance tradeoff perspective. Estimating too many dependence changes increases the variance of the estimator but does not increase bias. In contrast to that, a missed dependence change biases the estimate and decreases variance. The CP estimator is capable of detecting the correct number and appropriate position of the true partitions more often than PT and RS. Thus, it exhibits a lower bias but potentially a higher variance.

Figures 6.12 – 6.17 show the results for Scenario 6. The top rows of the figures show the estimated tessellation of each MC iteration stacked. The true tessellation is represented by the blue lines at the bottom and top of the stacked estimated tessellations. In the bottom row, the figures show histograms of detected changes on the conditioning space. The true tessellation is depicted by the white lines at the bottom of the histograms. Good performance is indicated by detection of changes close to the points $(0, 0.4), (0.5, 0), (0.5, 0.4), (0.5, 0.6), (0.5, 1)$, and $(1, 0.6)$. In the following, the performance of the three approaches is discussed.

The PT estimator shows a lot of variability in the estimated tessellations in Settings 1 – 4. However, its performance suffers in the last two settings, where only few dependence changes are detected. Whereas in the first four settings Kendall's τ and the copula family differs between partitions, in the last two settings only the copula family changes and Kendall's τ is constant. Thus, these findings are in line with the 3-dimensional case, where the performance of PT drops in settings with constant Kendall's τ. Overall, the estimated tessellations do not seem to reflect the true underlying tessellation as can be seen by inspecting the histograms. Most of the time, the conditioning space is partitioned along the variable U_3 first, even though in the true model the variable U_2 has to be partitioned first. This effect is also discussed below.

The RS estimator shows a similar performance like PT. In Settings 1 – 4 the conditioning space is cut into horizontal strips most of the time. Thus,

Scenario 1	1	2	3	4	5	6	7	8
PT	1.16	1.14	1.18	1.15	1.78	1.48	–	–
RS	1.05	1.03	1.01	0.99	1.24	1.07	–	–
CP	1.95	1.54	2.81	2.37	3.55	3.46	–	–
Scenario 2	1	2	3	4	5	6	7	8
PT	1.16	1.14	1.17	1.17	1.71	1.56	–	–
RS	1.08	1.01	1.13	1.01	1.58	1.0	–	–
CP	1.68	1.64	2.78	2.85	3.53	3.66	–	–
Scenario 3	1	2	3	4	5	6	7	8
PT	1.42	1.34	2.50	2.48	3.37	3.38	–	–
RS	1.36	1.31	2.38	2.44	3.33	3.26	–	–
CP	1.72	1.61	2.76	2.74	3.50	3.51	–	–
Scenario 4	1	2	3	4	5	6	7	8
PT	1.44	1.40	2.47	2.46	3.39	3.43	–	–
RS	1.41	1.33	2.40	2.38	3.26	3.29	–	–
CP	1.64	1.52	2.86	2.66	3.44	3.56	–	–
Scenario 5	1	2	3	4	5	6	7	8
PT	0.09	0.09	0.14	0.12	0.09	0.11	0.12	0.10
RS	0.15	0.01	0.05	0.02	0.14	0.06	0.04	0.04
CP	1.69	1.49	1.45	1.54	2.58	2.60	2.51	2.47

Table 6.2: Average number of detected dependence changes for Scenarios 1 – 5. CP overestimates the number of changes in Scenarios 1 – 4, whereas PT and RS do so in Scenarios 3 and 4 only. In Scenario 5, solely CP keeps a proper level of detected changes, whereas PT and RS severely underestimate the number of changes.

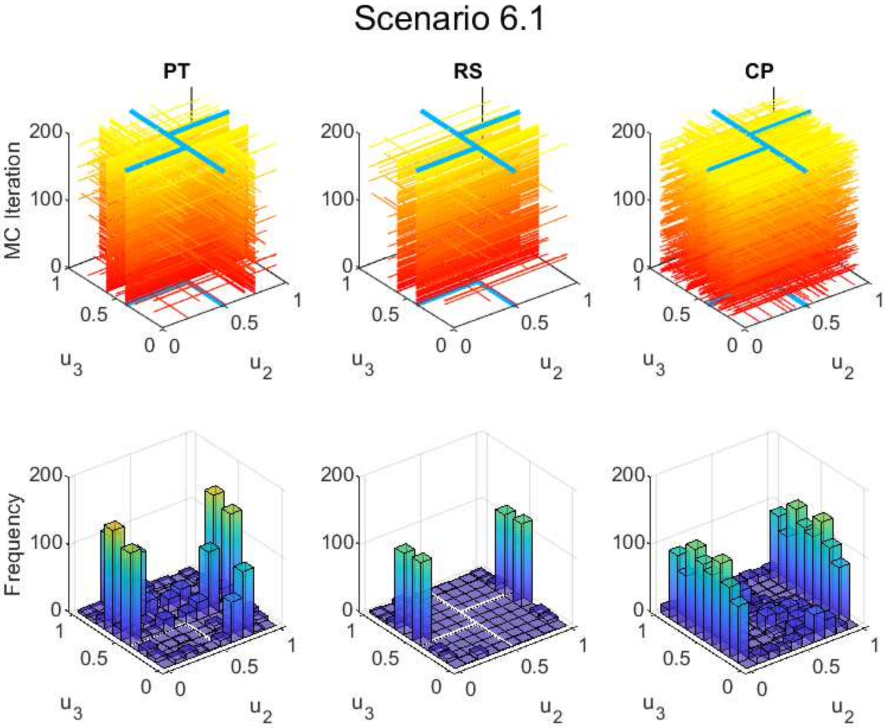

Figure 6.12: Tessellation estimation results for Scenario 6, Setting 1. The estimators PT, RS, and CP are shown from left to right in each row. The top row shows the estimated tessellation of each MC iteration stacked. The blue lines at the bottom and top depict the true tessellation. The histograms in the second row show where a change in dependence is detected. The white lines on the bottom of the histograms depict the true tessellation. PT and CP show some variability in the estimated tessellation, whereas RS seems to cut the conditioning space horizontally most of the time.

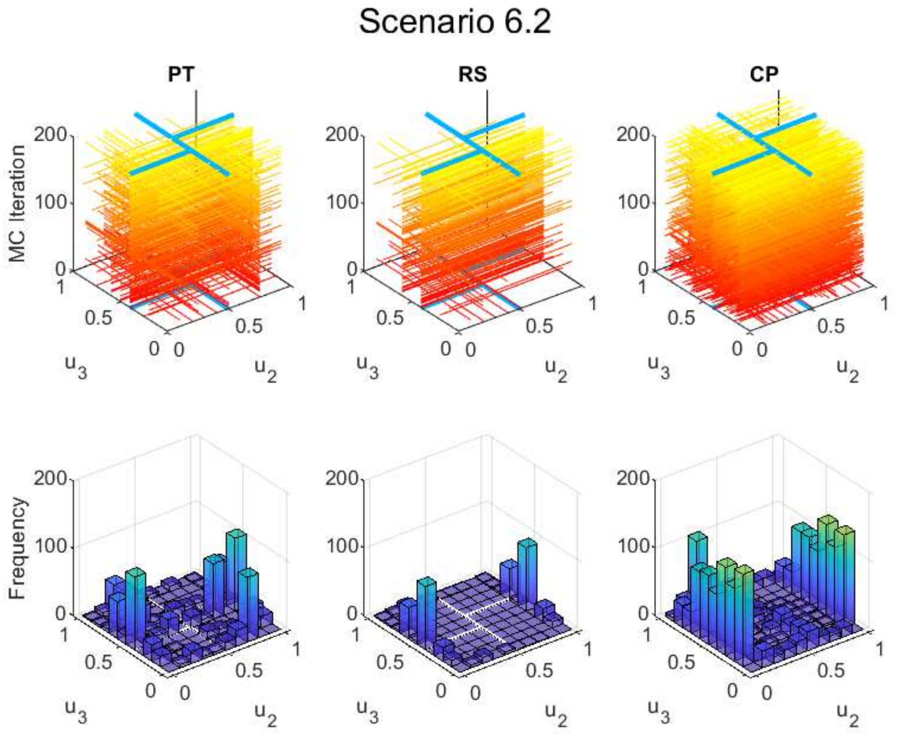

Figure 6.13: Tessellation estimation results for Scenario 6, Setting 2. The estimation approaches PT, RS, and CP are shown from left to right in each row. The top row shows the estimated tessellation of each MC iteration stacked. The blue lines at the bottom and top depict the true tessellation. The histograms in the second row show where a change in dependence is detected. The white lines on the bottom of the histograms depict the true tessellation. Again, PT and CP show some variability in the estimated tessellation, whereas RS seems to cut the conditioning space in horizontal strips most of the time. Also, RS estimates fewer tessellations than the other approaches.

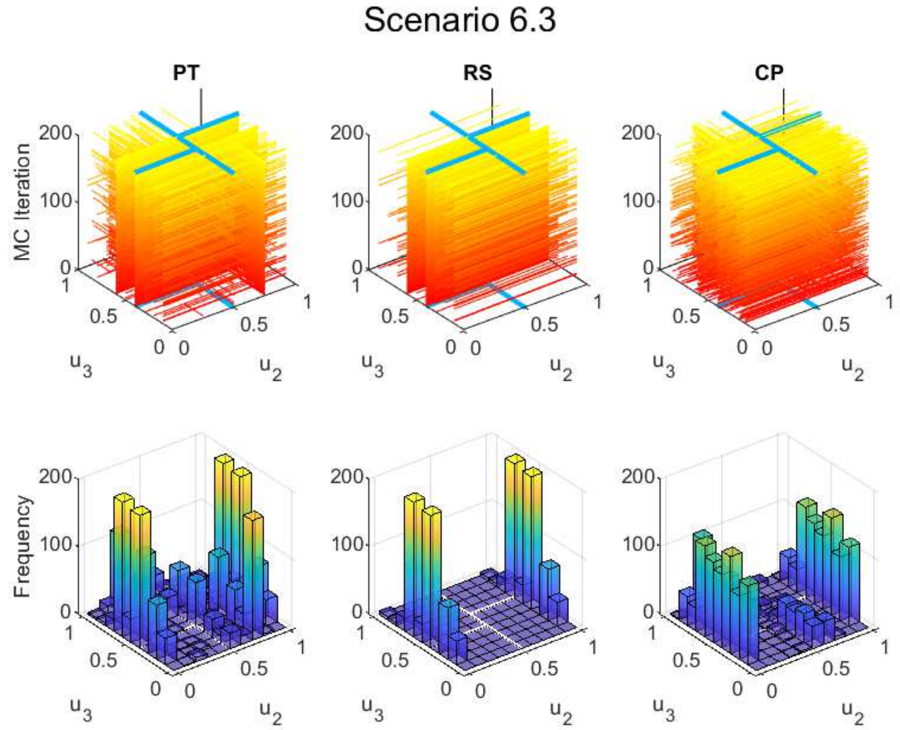

Figure 6.14: Tessellation estimation results for Scenario 6, Setting 3. The estimators PT, RS, and CP are shown from left to right in each row. The top row shows the estimated tessellation of each MC iteration stacked. The blue lines at the bottom and top depict the true tessellation. The histograms in the second row show where a change in dependence is detected. The white lines on the bottom of the histograms depict the true tessellation. As in the first two settings, PT and CP show some variability in the estimated tessellation, whereas RS seems to cut the conditioning space horizontally most of the time.

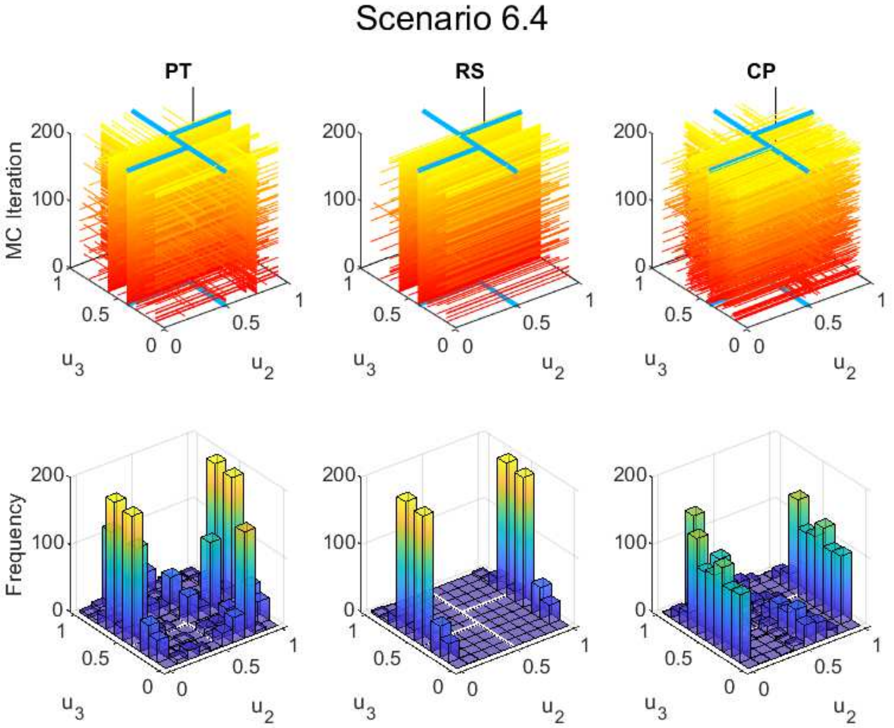

Figure 6.15: Tessellation estimation results for Scenario 6, Setting 4. PT, RS, and CP are shown from left to right in each row. The top row shows the estimated tessellation of each MC iteration stacked. The blue lines at the bottom and top depict the true tessellation. The histograms in the second row show where a change in dependence is detected. The white lines on the bottom of the histograms depict the true tessellation. PT and CP show some variability in the estimated tessellation, whereas RS seems to cut the conditioning space horizontally most of the time.

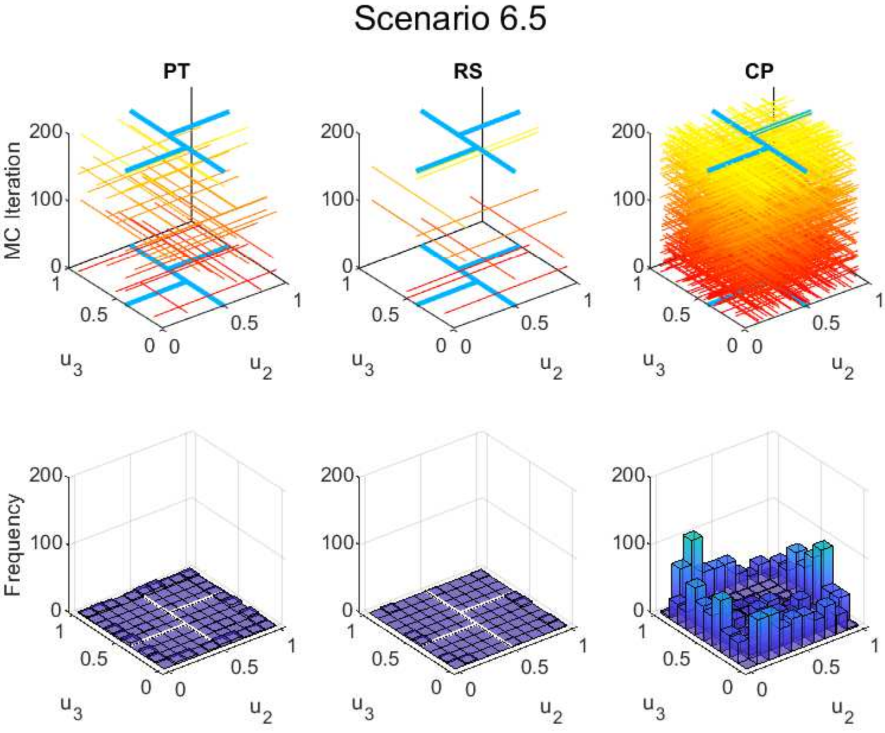

Figure 6.16: Tessellation estimation results for Scenario 6, Setting 5. The estimators PT, RS, and CP are shown from left to right in each row. The top row shows the estimated tessellation of each MC iteration stacked. The blue lines at the bottom and top depict the true tessellation. The histograms in the second row show where a change in dependence is detected. The white lines on the bottom of the histograms depict the true tessellation. PT and RS barely detect any tessellation at all, whereas CP maintains a high number of tessellations.

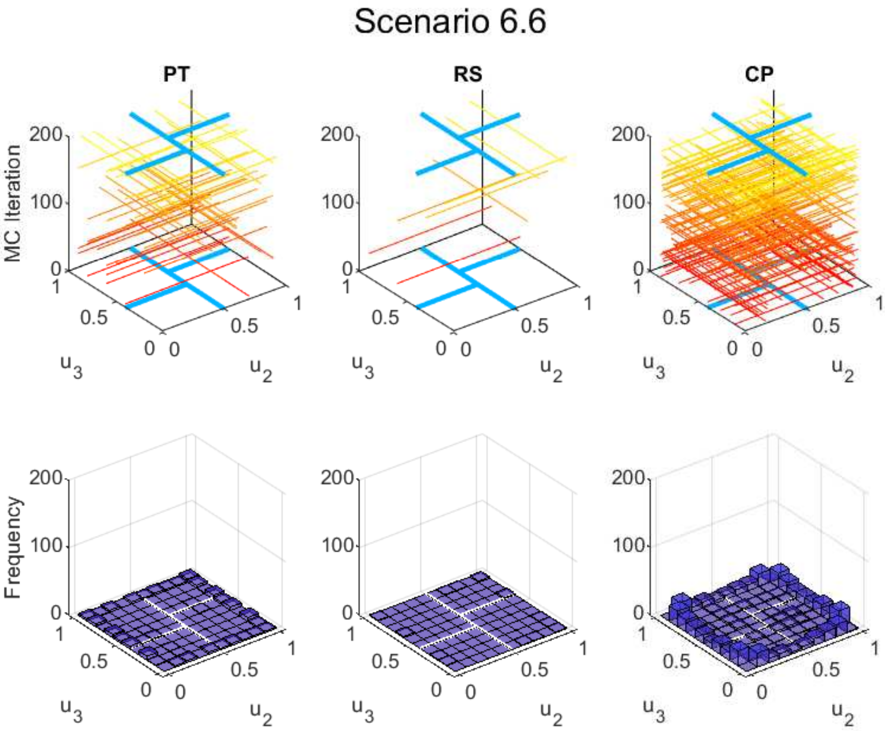

Figure 6.17: Tessellation estimation results for Scenario 6, Setting 6. The estimation approaches PT, RS, and CP are shown from left to right in each row. The top row shows the estimated tessellation of each MC iteration stacked. The blue lines at the bottom and top depict the true tessellation. The histograms in the second row show where a change in dependence is detected. The white lines on the bottom of the histograms depict the true tessellation. As in the previous setting, PT and RS barely detect any tessellation. Also, CP struggles in this setting.

the true tessellation is not reflected well. Also, the approach performs poorly in the last two settings, which is again in line with the findings from Scenario 5. Hence, PT and RS struggle with cases where Kendall's τ stays constant but the dependence structure changes via the involved copula families only.

The CP estimator shows high variability in the estimated tessellations across Settings 1 – 5. However, the true tessellation is rarely recovered. As can be seen from the histograms, compared to the other two estimators, CP detects a dependence change more often inside the unit square. Moreover, in contrast to the other two approaches, CP is also capable of estimating a tessellation in the last two settings, which is in line with the findings from Scenario 5. Merely Setting 6 appears to be very difficult. This could be due to the employed Gauss and t-copulas, which are very tough to distinguish.

Table 6.3 shows the proportion of U_2 being selected as the first variable in the tessellation estimation of Scenario 6. Note that the variable sequence is the same for each estimator since we assume the first two tree levels to be known and, thus, the pseudo-observations for tree T_3 are the same regardless of the tessellation estimation strategy. It can be seen that the introduced heuristic rarely (correctly) chooses U_2 as the first variable in Settings 1 – 4. In the last two settings, Kendall's τ is constant on the whole tessellation. Since the heuristic only compares differences in Kendall's τ a value of 50% is expected, which is also reflected in the empirical proportions. The heuristic determines the variable sequence by the direction, i.e., the variable, with the most variability in Kendall's τ. A future direction for research could be to find another criterion for selecting the variable sequence.

Setting	1	2	3	4	5	6
Fraction	0.085	0.14	0.0	0.005	0.50	0.47

Table 6.3: Proportion of U_2 being the first variable for which the conditioning space is split in Scenario 6.

Overall, the results of Scenario 6 are mixed compared to the 3-dimensional settings. In particular, the estimators struggle to obtain the true tessellation. This can be due to several overlapping effects, which are the heuristic to determine the variable sequence in which the conditioning space is partitioned and the ability of the estimators to detect a change in dependence. Nonetheless, CP seems to perform best across all settings. In particular, it is able to estimate a tessellation a fair amount of time in the Settings 5 and 6. Therefore, among the introduced estimators, CP is the most appropriate.

In conclusion, CP performs equally good or superior to PT and RS in the 3-dimensional Scenarios 1 – 5. Moreover, it shows a higher estimation performance in the 4-dimensional Scenario 6. Estimation is especially difficult when the copula family changes but Kendall's τ is constant and CP also works in these difficult settings. Thus, we recommend using CP for tessellation estimation. The next section compares simplified vines to non-simplified vines with tessellated conditioning spaces in different estimation settings.

6.5.2 Comparison to Simplified Vines

In this part of the simulation study, we explore the estimation behavior and performance of non-simplified vines with tessellated conditioning spaces and compare them to simplified vines. First, we want to give some motivation and intuition for this. The four panels of Figure 6.18 show data generated from four different models in green: a simplified vine, a non-simplified vine with a tessellation of the conditioning space, and two non-simplified vines, where the copula parameter is a function of the conditioning variable U_2. On each data set a simplified vine, a non-simplified vine with estimator EF (equal-frequency), and a non-simplified vine with estimator CP (change-point detection) are estimated and new points generated from the fitted models. The black points correspond to the simplified vine and the blue and orange points correspond to the non-simplified vine EF and CP, respectively. As can be seen, the simplified vine model is represented equally well by all three fitted models. In contrast to that, the non-simplified models are better represented by the estimated non-simplified vines. A slight exception is the non-simplified model where the copula parameter is a linear function of the conditioning variable. Here, the simplified vine fits well, too. These effects are more closely inspected in this part of the simulation study.

The simulation study consists of two scenarios which comprise several settings. For each setting a sample of size $n = 1,000$ is generated and three models are estimated on the samples – a simplified vine, a non-simplified vine using EF, and a non-simplified vine using CP. Each experiment is repeated 200 times. Table 6.4 gives an overview of the scenarios. Also, these are briefly discussed in the following.

Scenario 1 is 3-dimensional and uses the D-vine in Equation (6.34). The copulas c_{12} and c_{23} are Gumbel and Clayton copulas, respectively, with parameters corresponding to a Kendall's τ of 0.6 each. The conditional copula $c_{13|2}$ varies over the different settings. In two settings the conditioning space is tessellated. Two further settings employ a simplified vine. In the remaining two settings the parameter of the conditional copula is a function of the conditional variable U_2, whereby the copula family is fixed. Table 6.4 gives a detailed overview of the settings within Scenario 1.

Scenario 2 uses the 4-dimensional D-vine in Equation (6.35). The copulas

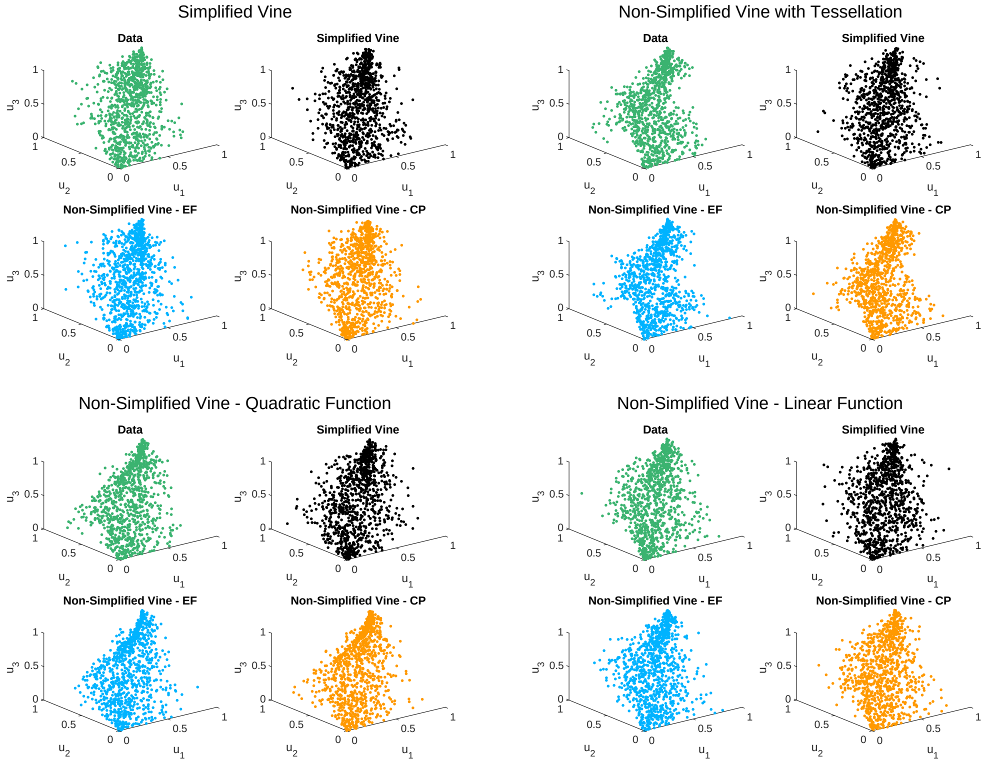

Figure 6.18: Each panel shows 3-dimensional data simulated from a vine structure in green. On these a simplified vine, non-simplified vine EF, and non-simplified vine CP are estimated. Then, new points are generated from the estimated models. The points from the simplified vine are shown in black, the points from the non-simplified vine EF are shown in blue, and the points from the non-simplified vine CP are shown in orange. Upper left panel: The true model is a simplified vine. All three estimated models seem to work. Upper right panel: The true model is a non-simplified vine with tessellated conditioning spaces. Clearly, the simplified vine does not resemble the true model. Lower left and lower right panels: The true model is a non-simplified vine, where the parameter of the conditional copula is a quadratic (lower left panel) or linear (lower right panel) function of the conditional variable U_2. Again, the simplified vine does not resemble the true model well, whereas the two non-simplified vines with tessellated conditioning spaces seem to reflect an appropriate dependence structure.

	Setting	Conditional Copula	Tessellation	Copulas	Kendall's τ
Scenario 1	1	$c_{13\|2}$	$[0, 0.5]$ – $(0.5, 1]$	G – G	(-0.6) – 0.6
	2	$c_{13\|2}$	$[0, 0.5]$ – $(0.5, 1]$	C – Gu	0.3 – 0.8
	3	$c_{13\|2}$	–	independence	–
	4	$c_{13\|2}$	–	F	0.7
	5	$c_{13\|2}$	–	C	u_2^2
	6	$c_{13\|2}$	–	G	$-0.5 + u_2$
Scenario 2	1	$c_{13\|2}$	$[0, 0.5]$ – $(0.5, 1]$	C – Gu	0.4 – 0.8
	1	$c_{14\|23}$	$[0, 0.5] \times [0, 0.5]$ – $(0.5, 1] \times [0, 0.5]$ – $(0.5, 1] \times (0.5, 1]$ – $[0, 0.5] \times (0.5, 1]$	G – G – G – G	0.6 – (-0.6) – 0.6 – (-0.6)
	2	$c_{13\|2}$	$[0, 0.7]$ – $(0.7, 1]$	t – G	0.4 – 0.8
	2	$c_{14\|23}$	$[0, 0.3] \times [0, 0.5]$ – $(0.3, 1] \times [0, 0.6]$ – $(0.3, 1] \times (0.6, 1]$ – $[0, 0.3] \times (0.5, 1]$	C – Gu – F – G	0.7 – 0.4 – (-0.4) – 0.3
	3	$c_{13\|2}$	–	independence	–
	3	$c_{14\|23}$	–	independence	–
	4	$c_{13\|2}$	–	F	0.7
	4	$c_{14\|23}$	–	F	(-0.5)
	5	$c_{13\|2}$	–	F	u_2
	5	$c_{14\|23}$	–	G	$0.5(u_2 + u_3)$
	6	$c_{13\|2}$	–	F	$0.9 \sin(2\pi u_2)$
	6	$c_{14\|23}$	–	G	$2u_2u_3 - 1$

Table 6.4: Overview of simulation scenarios for the second part of the simulation study. Each partition of a tessellation is separated by a hyphen. The fourth column shows the copulas in the same order as the tessellation. Also, the same order applies for the Kendall's τ in the last column. The following abbreviations for copulas are used: G – Gauss, C – Clayton, Gu – Gumbel, F – Frank. For the t-copula, the degrees of freedom are set to $\nu = 3$ throughout, i.e., only the second parameter ρ controls the strength of dependence.

c_{12}, c_{23}, and c_{34} are Gumbel, Clayton, and Gauss copulas with parameters translating to a Kendall's τ of 0.6, 0.6, and -0.4, respectively. The copula $c_{24|3}$ is a Clayton copula with a parameter corresponding to a Kendall's τ of 0.4. The copulas $c_{13|2}$ and $c_{14|23}$ vary over the settings. As before, two settings employ a tessellation of the conditioning spaces, two settings are simplified vines, and in two settings the copula families are fixed, however, their parameters are functions of the conditioning variables. Table 6.4 gives a detailed overview of Scenario 2.

In each setting a simplified vine and a non-simplified vine with tessellation of conditioning spaces is estimated. We assume the vine structure to be known throughout. All other components of the vine, i.e., the copulas and the tessellation(s) have to be estimated. For the overall estimation, we use the SSP estimator (cf. Algorithm 6). For the tessellation estimation we employ both the estimator EF and the estimator CP since the latter performed best in the first part of the simulation study. The set of copulas, out of which the best fitting copula is chosen in each step of the SSP estimation, comprises the t-, Gauss, Clayton, Gumbel, and Frank copulas. For each fitted model the AIC (Akaike, 1973) and BIC (Schwarz, 1978) are computed. Additionally, a Vuong model comparison test for non-nested models (Vuong, 1989) is conducted on each pair of fitted models. For the comparison of the simplified vine and the non-simplified CP vine, also a likelihood ratio test (Greene, 2012) is used because these two models are nested.

Table 6.5 and Figures 6.19 and 6.20 show the results of the simulation study. In Settings 3 and 4 of Scenarios 1 and 2 the true model is simplified. When using CP as the tessellation estimator, it is possible to actually estimate a simplified vine. In fact, this happens here. Of the 200 repetitions, a simplified vine is estimated 112 times in Setting 3 and 139 times in Setting 4 of Scenario 1, as well as 27 times in Setting 3 and 36 times in Setting 4 of Scenario 2. For the remaining settings a simplified vine is never estimated. This already shows a reasonable performance of the non-simplified vine model when using the CP estimator in 3 dimensions.

Table 6.5 shows the average AIC and BIC values of the fitted models with standard deviations in brackets. In Settings 1 and 2 of both scenarios, the non-simplified vines exhibit a superior fit compared to the simplified model as measured by AIC and BIC. Thus, when the true model is a non-simplified vine with tessellated conditioning spaces, AIC and BIC suggest that estimating a non-simplified vine using EF or CP is more appropriate than estimating a simplified vine, which was to be expected.

In Settings 3 and 4 of both scenarios the true model is a simplified vine. Nonetheless, the non-simplified vine using CP is on a par with the simplified vine in terms of AIC and BIC. In fact, the average values are very close in

		AIC			BIC	
Scenario 1	Simplified	EF	CP	Simplified	EF	CP
1	-2785 (96)	-3420 (96)	-3463 (92)	-2766 (96)	-3357 (96)	-3431 (93)
2	-3231 (113)	-3715 (109)	-3746 (108)	-3211 (113)	-3354 (110)	-3713 (108)
3	-2399 (96)	-2395 (96)	-2402 (96)	-2385 (96)	-2336 (96)	-2380 (97)
4	-3832 (108)	-3823 (109)	-3833 (109)	-3817 (108)	-3764 (109)	-3814 (108)
5	-2663 (100)	-3181 (116)	-3185 (115)	-2646 (100)	-3122 (116)	-3125 (114)
6	-2414 (96)	-2596 (96)	-2607 (96)	-2395 (96)	-2537 (96)	-2551 (96)
Scenario 2	Simplified	EF	CP	Simplified	EF	CP
1	-4005 (110)	-4984 (115)	-4794 (240)	-3967 (110)	-4783 (116)	-4673 (242)
2	-3567 (108)	-4412 (121)	-4394 (129)	-3528 (108)	-4246 (121)	-4292 (131)
3	-3369 (107)	-3345 (108)	-3367 (108)	-3340 (107)	-3182 (107)	-3314 (110)
4	-4385 (101)	-4342 (102)	-4378 (102)	-4356 (101)	-4180 (102)	-4326 (105)
5	-3977 (106)	-5205 (147)	-5103 (171)	-3936 (106)	-5017 (146)	-4922 (170)
6	-3711 (117)	-5591 (144)	-5468 (136)	-3672 (117)	-5418 (143)	-5293 (136)

Table 6.5: Average AIC and BIC values for each setting and estimated model. Standard deviations are given in brackets next to each value. A lower value indicates a better model fit.

some cases. This is also due to the observation above that in these settings the CP estimator actually fits a simplified vine in quite a high number of cases. In contrast to that, EF performs worse than the simplified vine, particularly in terms of BIC. Here, the substantial amount of parameters used in the EF approach exacts its toll.

In Setting 5 of the two scenarios, the copula parameters of the true models are a nonlinear function of the conditioning variables. In this case, the two non-simplified vine models perform vastly superior compared to the simplified model. Thus, a non-simplified vine model with tessellated conditioning spaces is a better approximation to the true non-simplified model than a simplified vine. A similar picture shows Setting 6 of the two scenarios, where the copula parameters are linear functions of the conditioning variables. Where the simplified vine model compares well to the two non-simplified models in the 3-dimensional case, it is strikingly worse in the 4-dimensional case.

Comparing the two non-simplified vine estimators, they yield quite similar AIC and BIC values overall. Since the tessellation estimation in CP counts towards the number of parameters in the model (number of parameters from tessellation = number of partitions -1), the high number of estimated copulas due to the predefined partitioning in EF is compensated.

The results from the information criteria are corroborated by the Vuong model comparison test for non-nested models (Vuong, 1989). Figure 6.19 shows box plots of the pairwise comparisons for all settings. The test statistic follows a standard normal distribution (Vuong, 1989) and the 1% and 99% quantiles of the standard normal distribution are shown as black horizontal lines in the panels.

In the simplified vs non-simplified EF comparison, a significant positive test statistic indicates better performance of the simplified vine model, a significant negative test statistic indicates better performance the non-simplified vine model. The simplified vine is preferred for all truly simplified models, whereas the non-simplified vine is preferred for all truly non-simplified models. This is in line with the results from the AIC and BIC values. The same interpretation of the test statistic sign holds for the simplified vs non-simplified CP comparison. Again, if the true model is non-simplified the test indicates that the non-simplified estimator is more appropriate. However, the test shows that if the true model is simplified, the two competitors perform equally well. This is due to the number of simplified vines the estimation yields for CP. Finally, the two non-simplified vine models are compared. Here, a significant positive test statistic indicates a better fit of CP. Clearly, the CP estimator dominates EF in all simplified settings.

Since the simplified vine and the non-simplified CP vine are nested (cf. Section 6.2), the Vuong model comparison test is not appropriate. Thus, for this pair we also conduct a likelihood ratio test (Greene, 2012), which accounts for nested models. Figure 6.20 shows box plots of the p-values of the test. A p-value of 1% is depicted as a black line and a significant test indicates that the unrestricted model, i.e., the non-simplified CP vine, is superior compared to the simplified vine. Clearly, the non-simplified CP vine performs better in the non-simplified settings, whereas the two models are equally good in the simplified settings. This is in line with the results above.

We also report the variable sequence chosen in Scenario 2. Table 6.6 shows the proportion of U_2 being selected first. Note that the variable sequence of EF and CP can differ here since all vine levels are estimated assuming unknown copulas and, thus, different pseudo-observations are obtained for the estimation in T_3. However, the proportions of the two models are very close.

Setting	1	2	3	4	5	6
EF	0.51	0.01	0.47	0.465	0.885	0.48
CP	0.54	0.015	0.46	0.495	0.80	0.475

Table 6.6: Proportion of U_2 being the first variable for which the conditioning space is split in Scenario 2. Because the first and second tree levels are estimated as well, the variable sequence heuristic can yield different results for EF and CP.

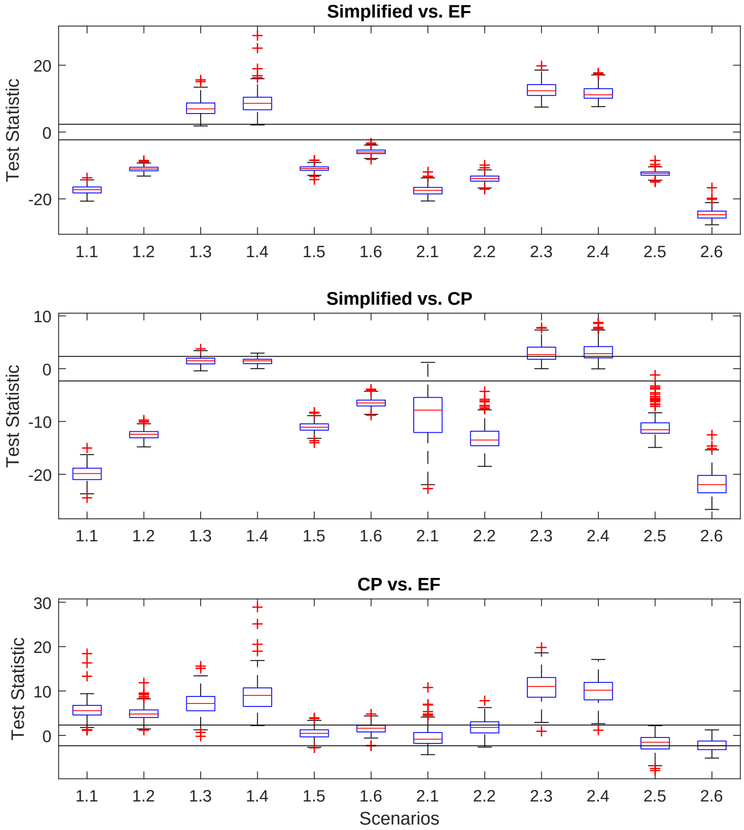

Figure 6.19: Box plots of Vuong model comparison test statistics. The test statistic follows a standard normal distribution (Vuong, 1989). Black horizontal lines depict the 1% and 99% quantiles of the standard normal distribution. Upper panel: Comparison of simplified vine and non-simplified EF vine. Positive significant test statistics indicate that a simplified vine model is superior, whereas negative significant test statistics indicate that a non-simplified EF vine model is superior. Middle panel: Comparison of simplified vine and non-simplified CP vine. Positive significant test statistics indicate that a simplified vine model is superior, whereas negative significant test statistics indicate that a non-simplified CP vine model is superior. Note that these models are nested. Lower panel: Comparison of non-simplified CP vine and non-simplified EF vine. Positive significant test statistics indicate that a non-simplified CP vine model is superior, whereas negative significant test statistics indicate that a non-simplified EF vine model is superior.

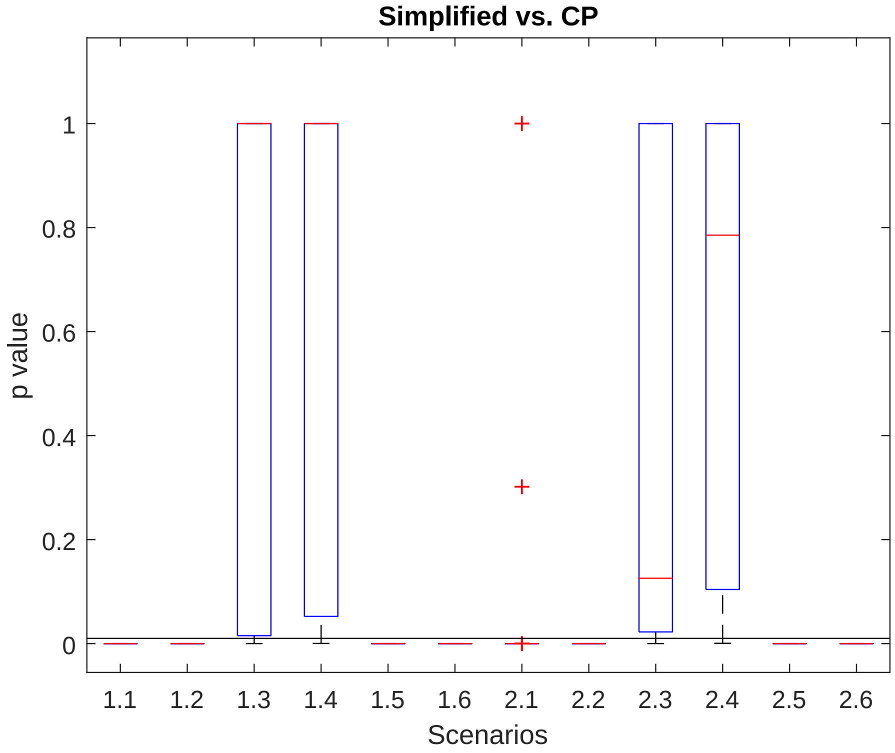

Figure 6.20: Box plots of likelihood ratio test p-values. The black horizontal line depicts a p-value of 1%. The nested simplified vine and non-simplified CP vine are compared. A significant test indicates that the unrestricted model, i.e., the non-simplified CP vine, is superior.

In Setting 1, the variable sequence does not play a role because the conditioning space is evenly split. Consequently, the proportion is close to 50% for both approaches. Setting 2 shows again that the heuristic can yield the wrong variable order since here variable U_2 should be first which is not well reflected in the empirical proportion. In Settings 3 and 4, where a simplified vine is the true model and the variable sequence does not play a role, the proportion is close to 50%. Interestingly, in the last two settings, where the copula parameter is a function of the conditioning variable, the variable sequence is close to 50% in only one setting. Hence, the variable sequence seems to matter in the estimation of non-simplified vine models that do not exhibit a tessellation of conditioning spaces.

This part of the simulation study reveals some interesting findings. The non-simplified vine using CP suggests itself as a good way for an exploratory data analysis. On the one hand, it can support the use of a simplified vine, when it yields a simplified vine in the estimation process. On the other hand, it can be used as a good approximation to a non-simplified vine even if in the true model the conditioning space is not tessellated. In comparison to EF, the CP estimator obtains superior estimation results in some settings. However, the EF estimator should not be discarded since it offers a relatively easy and, in terms of computational time, cheap alternative to CP for very high dimensions and large data sets. In the next section, we shed some further light on the findings of the simulation study via an application of the developed estimators to a real-world data set.

6.6 Application

We apply the developed estimation method for non-simplified vines to the well-known hydro-geochemical stream and sediment reconnaissance data from Cook and Johnson (1986). It is also used by Acar et al. (2012) in their analysis of simplified vines and they find that the simplifying assumption is not valid on the data set. The data set consists of 655 data points of logarithms of measured concentrations from several elements in Montrose, Colorado. We use the elements cobalt (Co), scandium (Sc), and titanium (Ti). Figure 6.21 shows a scatter plot matrix of the original data. The rank-transformed data can be seen in Figure 6.24, top row. Clearly, all variables exhibit positive dependence.

We estimate a simplified vine, a non-simplified vine using EF, a non-simplified vine using CP, and a non-simplified vine according to the method by Acar et al. (2012), which uses a kernel density estimation to obtain a function of the conditional copula parameter. The vine structure is estimated

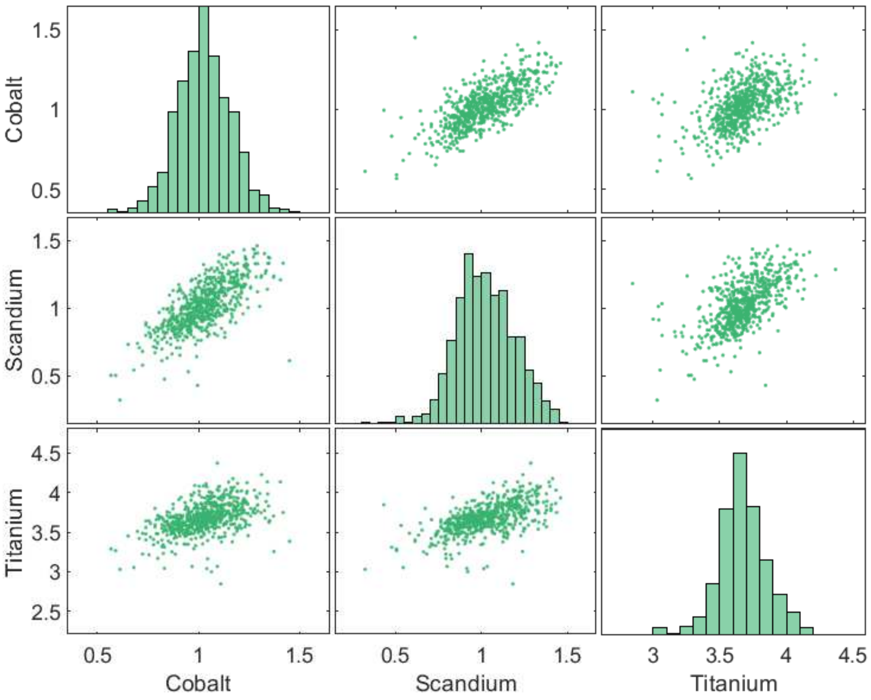

Figure 6.21: Scatter plot of the hydro-geochemical stream and sediment reconnaissance data, original scales.

according to the algorithm of Dißmann et al. (2013). Since we deal with 3-dimensional data, the order of the variables in the first tree has to be chosen only. Co and Sc exhibit the highest Kendall's τ of 0.54. Sc and Ti exhibit the second highest Kendall's τ of 0.44. Thus, the first tree level is Co-Sc-Ti and the conditional copula of interest is $c_{CoTi|Sc}$. The copulas c_{CoSc} and c_{ScTi} are estimated as t-copulas with parameters $\rho_{CoSc} = 0.74$, $\nu_{CoSc} = 8.02$ and $\rho_{ScTi} = 0.62$, $\nu_{ScTi} = 5.93$, respectively. Converting these parameter values back to Kendall's τ yields 0.53 and 0.43, which is very close to the empirical values.

Figure 6.22 shows a plot of the conditioning variable Sc rank-transformed against the conditional Kendall's τ $\tau_{CoTi|Sc}$. The black line depicts $\tau_{CoTi|Sc}$ of the simplified copula. The non-simplified EF and CP models correspond to the blue and orange step functions. These also implicitly show the estimated tessellation because each partition belongs to one step. For comparison purposes the green curve shows the estimated $\tau_{CoTi|Sc}$ of the method by Acar et al. (2012), where we have used a Gaussian kernel and Silverman's rule of thumb (Silverman, 1986). Note that the increasing values of the green curve close to the boundaries 0 and 1 are an artifact of the kernel density estimation in a bounded domain and do therefore not reflect the true behavior of the conditional Kendall's τ appropriately. Remarkably, the step functions follow the path of the green curve. This indicates that the developed estimators EF and CP are capable of approximating other non-simplified vine models. Furthermore, this analysis shows again that the simplifying assumption indeed seems to be violated on this data set.

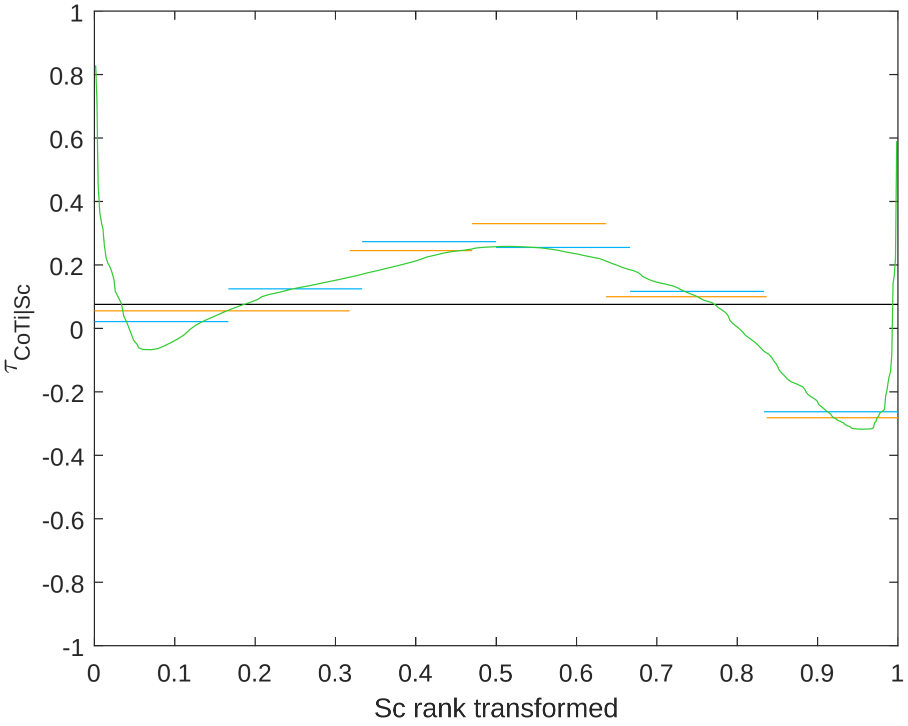

Figure 6.22: Plot of Sc rank transformed and $\tau_{CoTi|Sc}$ according to different models. The black line depicts $\tau_{CoTi|Sc}$ of a simplified vine. The blue and orange steps depict $\tau_{CoTi|Sc}$ of the non-simplified EF vine and the non-simplified CP vine, respectively. For comparison purposes, the green curve shows $\tau_{CoTi|Sc}$ obtained by the method of Acar et al. (2012).

It is also interesting to look at the computing times for the estimators. A simplified vine was estimated in 0.7 seconds on a standard desktop computer. The non-simplified vine using EF was estimated in 1.4 seconds and the non-simplified vine using CP was estimated in approximately 5 minutes. This shows that EF can be a computationally efficient alternative to CP, particularly when the data set is very large. The estimation of the method by Acar et al. (2012) took several hours. Thus, for 3-dimensional data, the non-simplified vine model with tessellated conditioning spaces can be a time-efficient alternative.

Figure 6.23 shows the graphical representation of the non-simplified EF and CP vines. Some adjacent partitions of the tessellations have similar values of Kendall's τ. However, the copula families of two adjacent partitions always differ. The copula families are Tawn – Survival Gumbel – Gumbel – Plackett – Tawn – t for EF and Tawn – Gumbel – Plackett – Gumbel – t for CP. Note that we have used a bigger set of copulas for this estimation task. On this data set, both the EF and CP estimators seem to be appropriate alternatives.

In order to compare the different models further, we generate new points from the estimated simplified vine and the estimated non-simplified EF and CP models. The results are shown in Figure 6.24. The top row contains a scatter plot matrix and a 3-dimensional plot of the rank-transformed data. The remaining three rows contain the same plots for the simplified vine data, non-simplified vine EF data, and non-simplified vine CP data from top to bottom. From inspection of the scatter plot matrices, the three models seem to fit the data equally well. However, the 3-dimensional plots reveal some differences. In particular, the simplified model appears to be too uniform along the diagonal of the unit cube. In contrast to that, the non-simplified EF and CP models remedy this. Moreover, the non-simplified CP model seems to capture the behavior of the data towards the point $(1, 0, 0)$ best.

The application gives some further insights into the developed non-simplified vine with tessellated conditioning spaces. Interestingly, it is capable of approximating other non-simplified vine approaches. At the same time, having the flexibility to change the copula family in the conditional copula(s) seems to be a valuable trait of the proposed method. The next section gives some concluding remarks and points out directions for future research.

6.7 Concluding Remarks

A new approach to relax the simplifying assumption in vine copula models is developed. It partitions the conditioning spaces of the involved conditional copulas. This allows to define a different copula family on each part of the resulting tessellation, which sets the introduced method apart from existing techniques. Furthermore, the non-simplified vine model can be graphically

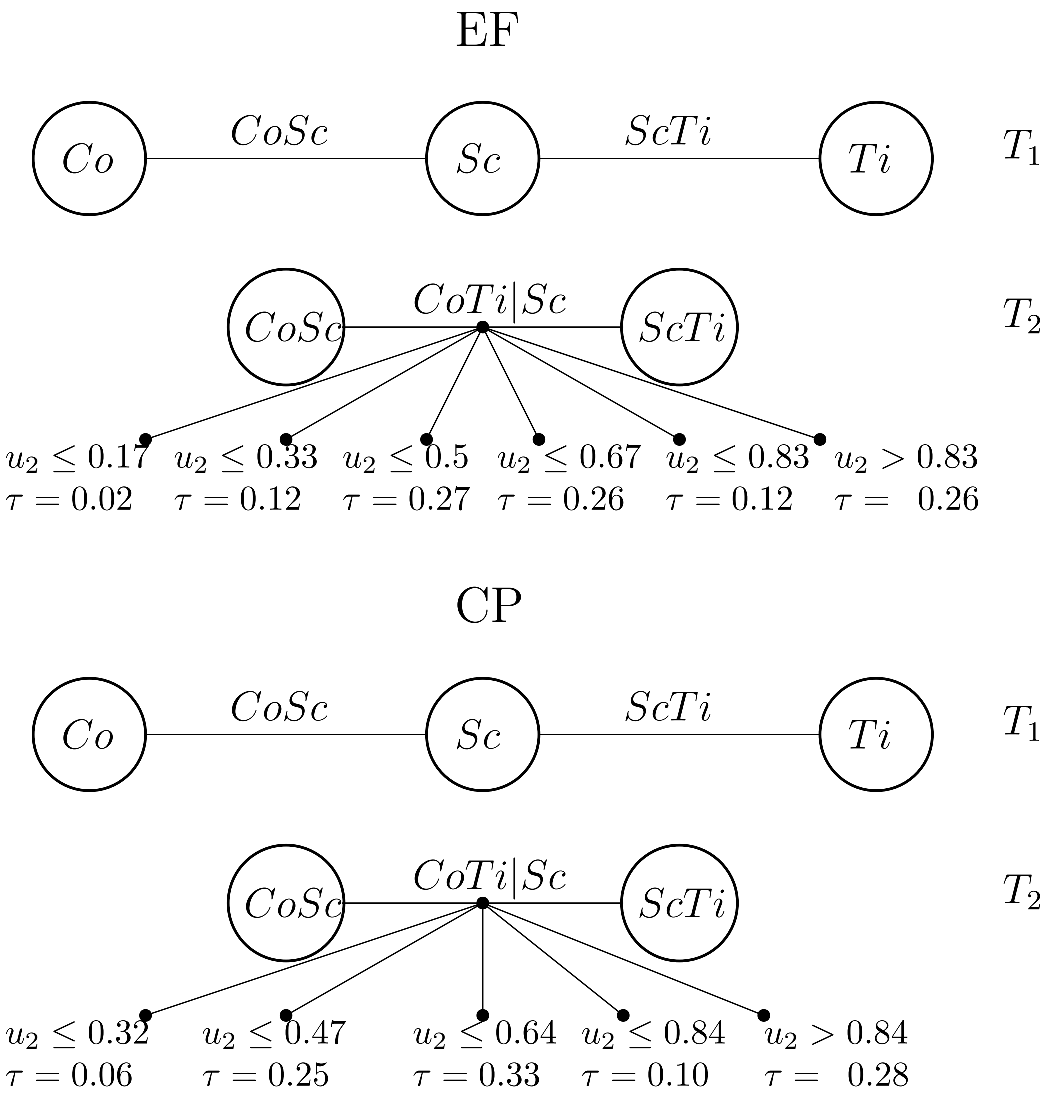

Figure 6.23: Estimated non-simplified vines on the hydro-geochemical stream and sediment reconnaissance data set.

displayed, which makes interpretation of the dependence structure easier.

Furthermore, we develop simulation and estimation algorithms. The simulation algorithms are extended from existing algorithms of simplified vines. For the estimation we employ the stepwise semiparametric estimator and combine it with techniques from decision trees in order to account for the tessellation estimation. We suggest four different tessellation estimation strategies, of which two use a statistical test (PT and RS), one uses methods from change-point detection (CP), and one uses equal-frequency binning (EF).

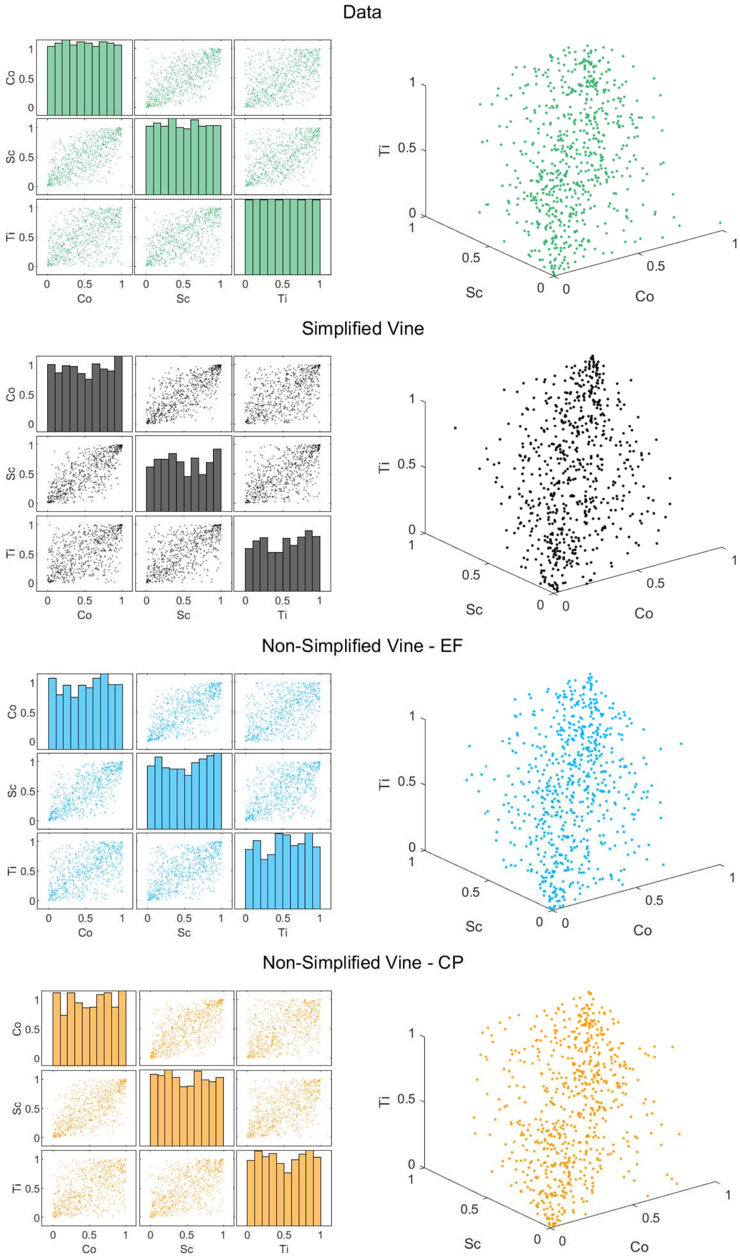

Figure 6.24: Scatter plots of original data rank-transformed (green), of points simulated from the fitted simplified vine (black), of points simulated from the fitted non-simplified EF vine (blue), and of points simulated from the non-simplified CP vine (orange).

A simulation study shows that among PT, RS, and CP, the CP approach performs best.

In a second simulation study, we compare the developed model to simplified vines in various simplified and non-simplified settings. If the true model is non-simplified, the developed approach performs better than a simplified vine. However, if the true model is simplified, the non-simplified and simplified vine structure are almost equally good. This indicates that the non-simplified vine with tessellation of conditioning spaces can be used both as a data exploration tool and a complementary model to simplified vines. An application on a well-known data set sheds some more light on the introduced techniques.

There are several directions for future research. First, it would be interesting to find a different heuristic to choose the variable sequence for the tessellation estimation. Second, we restricted ourselves to tessellations which can be represented by a decision tree because it makes their estimation more convenient. However, one can think of more complicated tessellations which are, e.g., not axis-aligned. A suitable estimation technique would have to be developed, though. Third, we constrained the parameter of the copula within a partition of a tessellation to be constant. This constraint can also be dropped and the approach thereby extended. Finally, it would be interesting to compare the introduced approach to other non-simplified vine models in a large-scale simulation study to gain further insights into the estimation capabilities of the models and their advantages and disadvantages.

The next chapter summarizes this thesis and provides some concluding remarks for further research directions.

Chapter 7

Conclusion

This thesis contributes to the use of copulas in multivariate data. Also, advancements with respect to theoretical aspects of copulas are made. In particular, we develop a nonparametric estimation procedure for multivariate quantiles based on copulas (see Chapter 3). The procedure is further enhanced by incorporating a smoothed bootstrap. Our simulation results and application example show the usefulness of the introduced techniques. Furthermore, we deal with methods to evaluate the precision of the multivariate quantiles (see Chapter 4). To this end, we extend recently introduced confidence region approaches and apply them to multivariate quantiles. This is done by incorporating the uncertainty of the estimation of the levels of the multivariate quantiles in these approaches. Moreover, we check their coverage probabilities in a simulation study.

We deal with theoretical aspects of bivariate copulas by studying the length of their level curves (see Chapter 5). We find that these are linked to the underlying dependence structure. For our investigation, we introduce the length profile, which maps each level $t \in [0, 1]$ to the corresponding length of the level curve at level t. Based on the length profile, we develop the length measure ℓ_C, which is the average length of the level curves of a copula C. We show several properties of both the length profile and the length measure. Additionally, we check whether the length measure is a concordance measure.

Finally, we introduce a new method to relax the simplifying assumption in vine copulas (see Chapter 6). The developed technique partitions the conditioning spaces of the vine structure into disjoint sets, where each part is governed by its own copula. The method yields an informative graphical representation of the estimated vine that can be used to interpret estimation results. We investigate the estimation procedure in a simulation study and find that it performs well in non-simplified settings and on a par with simplified vines in simplified settings.

Some work on the presented topics remains to be done. For example, it would be interesting to combine the smoothed bootstrap with the confidence region approaches for multivariate quantile estimation. Furthermore, there are open questions with respect to the connection of copula level curve lengths and some dependence concepts. For instance, it would be exciting to find a link to existing concordance measures, such as Spearman's ρ_S, or to lower tail dependence for some copula families. Finally, it would be interesting to extend the introduced non-simplified vine model to allow for even more flexible tessellations. Additionally, we are in need of a better criterion to choose (or estimate) the variable sequence when estimating the tessellations.

Appendix A

Proofs

A.1 Proof of Theorem 9

Proof. The claim follows from applying the implicit function theorem (Loomis and Sternberg, 1980) and using the standard formula for the length of a graph of a function. □

A.2 Proof of Theorem 11

Proof. For $t \in (0, 1]$ basic geometric considerations yield

$$L_W(t) = \sqrt{(1-t)^2 + (1-t)^2} = \sqrt{2} \cdot (1-t) \tag{A.1}$$

and

$$L_M(t) = 2 \cdot (1-t) \tag{A.2}$$

(also cf. Remark 1). The inequality follows by recalling that, (i), the lower t level set of a copula lies inside the triangle bounded by the level sets of the lower and upper Fréchet-Hoeffding bounds, (ii), the level sets of a copula cannot overlap, and (iii), Lemma 5 part c). □

A.3 Proof of Theorem 14

Proof. Let $\epsilon \in (0, 1/2)$, $I = [\epsilon, 1-\epsilon]$, $\varphi_1(t) := t$, and $\varphi_2(t) := 1$. Furthermore, using the implicit function theorem (Loomis and Sternberg, 1980), there exists a unique continuously differentiable function v for which $C\left(u, v(t, u)\right) = t$. Since $\partial v(t, u)/\partial u$ is continuous in t and u, so is

$$f(t, u) = \sqrt{1 + \left[\frac{\partial}{\partial u} v(t, u)\right]^2} \tag{A.3}$$

as a composition of continuous functions. By writing

$$L_C(t) = \int_{\varphi_1(t)}^{\varphi_2(t)} f(t,u)du \tag{A.4}$$

and considering that φ_1, φ_2 are continuous on $[0,1]$, the continuity of parameter-dependent integrals can be used (Wilson, 1958), and it follows that L_C is continuous on I. Using Theorem 11 and letting $t \to 1^-$, L_C is also continuous in $t = 1$ which completes the proof. □

A.4 Proof of Theorem 15

Proof. The claim follows directly by using the Leibniz integral rule.

Also notice that $v(t,t) = 1$ and

$$\frac{\partial C\,(t,1)}{\partial u} = \lim_{\epsilon \to 0^+} \frac{C(t+\epsilon,1) - C(t,1)}{(t+\epsilon) - t} = \lim_{\epsilon \to 0^+} \frac{(t+\epsilon) - t}{(t+\epsilon) - t} = 1.$$

□

A.5 Proof of Theorem 16

For the proof of Theorem 16, the following lemma is needed.

Lemma 26. *Let $a < b$ and f, g be continuously differentiable, convex functions on $[a,b]$, where $f(a) = g(a), f(b) = g(b)$ and $g(x) \leq f(x)$, for any $x \in [a,b]$. Then*

$$\int_a^b \sqrt{1 + (g'(t))^2}dt \geq \int_a^b \sqrt{1 + (f'(t))^2}dt. \tag{A.5}$$

Proof. Define $\varphi : \mathbb{R} \to \mathbb{R},\ \varphi(t) := (1+t^2)^{\frac{1}{2}}$, which is a twice continuously differentiable, convex function on $\mathbb{R}$. Using convexity of φ, for any $t \in [a,b]$,

$$\varphi(g'(t)) \geq \varphi(f'(t)) + (g'(t) - f'(t))\,\varphi'\,(f'(t))\,. \tag{A.6}$$

Integration yields

$$\int_a^b \varphi(g'(t))dt \geq \int_a^b \varphi(f'(t))dt + \int_a^b (g'(t) - f'(t))\,\varphi'\,(f'(t))\,dt. \tag{A.7}$$

It now suffices to show that $\int_a^b (g'(t) - f'(t)) \cdot \varphi'\,(f'(t))\,dt \geq 0$. This can be done by applying the second mean value theorem for definite integrals and using, (i), $f \geq g$, and (ii), $\varphi' \circ f'$ is nondecreasing on $[a,b]$ due to the convexity of φ and f. □

We can now prove Theorem 16 by contradiction.

Proof. Let $t \in (0, 1)$ be arbitrary but fixed. Let v_1, v_2 be the level curve functions at level t of C_1 and C_2, respectively. Both are continuous and differentiable, convex by assumption, and begin in $u = t$ and end in $u = 1$.

Now, suppose that $v_2(u) > v_1(u)$. Since $s \mapsto C_1(u, s)$ is strictly increasing in s and $C_1 \leq C_2$,

$$C_1\left(u, v_2(u)\right) > C_1\left(u, v_1(u)\right) = t = C_2\left(u, v_2(u)\right) \geq C_1\left(u, v_2(u)\right), \tag{A.8}$$

which is a contradiction. Thus, $v_2(u) \leq v_1(u)$ on $[t, 1]$. The claim now follows from Lemma 26. □

A.6 Proof of Theorem 18

Proof. The claim follows directly by observing that each level curve of C is the level curve of $C^\top$ reflected at the angle bisector of the unit cube. □

A.7 Proof of Corollary 19

Proof. The claim is a direct consequence of Theorem 11 and the monotonicity of integration.

Straightforward calculation yields $\ell_W = \sqrt{2}/2$ and $\ell_M = 1$. $\ell_\Pi \approx 0.7652$ can be obtained by numerical integration of

$$\ell_\Pi = \int_0^1 \left(\int_t^1 \sqrt{1 + \frac{t^2}{u^4}} du \right) dt. \tag{A.9}$$

□

A.8 Proof of Theorem 20

Proof. For symmetric copulas the claim follows directly. For asymmetric copulas the claim follows from Theorem 18. □

A.9 Proof of Theorem 23

Proof. In a first step, we show that the length profiles $L_{C_n}(t)$ of C_n converge to the length profile $L_C(t)$ of C for $t \in (0, 1)$. For $t \in \{0, 1\}$ this is trivially fulfilled. Using the notation in Theorem 5, let $v_{t,n}$ and v_t be the level curve of C_n and C, respectively. Since by assumption $v_{t,n}$, $n \in \mathbb{N}$, is a sequence of convex functions, the limit v_t is also convex. Moreover, the first derivatives of $v_{t,n}$ converge to the first derivative of v_t, i.e., $\lim_{n\to\infty} v'_{t,n} = v'_t$ (Rockafellar, 1970). Since

$$\begin{aligned}
\int_t^1 |v'_{t,n}(u)| du &= -\int_t^1 v'_{t,n}(u) du && \text{(A.10)}\\
&= v_{t,n}(t) - v_{t,n}(1) && \text{(A.11)}\\
&= 1 - t && \text{(A.12)}\\
&= v_t(t) - v_t(1) && \text{(A.13)}\\
&= \int_t^1 |v'_t(u)| du, && \text{(A.14)}
\end{aligned}$$

Scheffé's Lemma (Riesz, 1928; Scheffé, 1947) yields L^1 convergence. It follows

$$\begin{aligned}
|L(v_{t,n}) - L(v_t)| &= \left| \int_t^1 \sqrt{1 + \left[v'_{t,n}(u)\right]^2} du - \int_t^1 \sqrt{1 + \left[v'_t(u)\right]^2} du \right| && \text{(A.15)}\\
&\leq \int_t^1 \left| \sqrt{1 + \left[v'_{t,n}(u)\right]^2} - \sqrt{1 + \left[v'_t(u)\right]^2} \right| du && \text{(A.16)}\\
&\leq \int_t^1 \left| v'_{t,n}(u) - v'_t(u) \right| \longrightarrow 0 \quad \text{for } n \to \infty, && \text{(A.17)}
\end{aligned}$$

where we have used the mean value theorem and the fact that the absolute value of the derivative of $\sqrt{1+x^2}$ is bounded from above by 1.

We will now show that the length measure converges, as well. From Theorem 7 we have $L_C(t) \leq 2(1-t)$ on $[0,1]$. The dominated convergence theorem now completes the proof

$$\ell_{C_n} = \int_0^1 L_{C_n}(t) dt \to \int_0^1 L_C(t) dt = \ell_C. \qquad \text{(A.18)}$$

□

Appendix B

Level Curve Lengths for Specific Copulas

B.1 Gaussian Copula

Let Φ and Φ^{-1} be the univariate standard normal CDF and its inverse, respectively. According to Aas et al. (2009),

$$\frac{\partial C(u,v)}{\partial u} = \Phi\left(\frac{\Phi^{-1}(v) - \rho \cdot \Phi^{-1}(u)}{\sqrt{1-\rho^2}}\right) \quad \text{and} \tag{B.1}$$

$$\frac{\partial C(u,v)}{\partial v} = \Phi\left(\frac{\Phi^{-1}(u) - \rho \cdot \Phi^{-1}(v)}{\sqrt{1-\rho^2}}\right). \tag{B.2}$$

Thus,

$$L\left(C_t\right) = \int_t^1 \sqrt{1 + \left[\frac{\Phi\left(\frac{\Phi^{-1}(v_t(u)) - \rho \cdot \Phi^{-1}(u)}{\sqrt{1-\rho^2}}\right)}{\Phi\left(\frac{\Phi^{-1}(u) - \rho \cdot \Phi^{-1}(v_t(u))}{\sqrt{1-\rho^2}}\right)}\right]^2}\, du. \tag{B.3}$$

B.2 t-Copula

Let t_ν and t_ν^{-1} be the univariate t CDF and its inverse, respectively. According to Aas et al. (2009),

$$\frac{\partial C(u,v)}{\partial u} = t_{\nu+1}\left(\frac{t_\nu^{-1}(v) - \rho \cdot t_\nu^{-1}(u)}{\sqrt{\frac{(\nu+[t_\nu^{-1}(u)]^2)\cdot(1-\rho^2)}{\nu+1}}}\right) \quad \text{and} \tag{B.4}$$

$$\frac{\partial C(u,v)}{\partial v} = t_{\nu+1}\left(\frac{t_\nu^{-1}(u) - \rho \cdot t_\nu^{-1}(v)}{\sqrt{\frac{(\nu+[t_\nu^{-1}(v)]^2)\cdot(1-\rho^2)}{\nu+1}}}\right). \tag{B.5}$$

Thus,

$$L\left(C_t\right) = \int\limits_t^1 \sqrt{1 + \left[\frac{t_{\nu+1}\left(\frac{t_\nu^{-1}(v)-\rho\cdot t_\nu^{-1}(u)}{\sqrt{\frac{(\nu+[t_\nu^{-1}(u)]^2)\cdot(1-\rho^2)}{\nu+1}}}\right)}{t_{\nu+1}\left(\frac{t_\nu^{-1}(u)-\rho\cdot t_\nu^{-1}(v)}{\sqrt{\frac{(\nu+[t_\nu^{-1}(v)]^2)\cdot(1-\rho^2)}{\nu+1}}}\right)}\right]^2}\, du. \tag{B.6}$$

B.3 Clayton Copula

The generator of the Clayton copula is $\psi(u) = (1+u)^{-\frac{1}{\theta}}$. Using Corollary 10,

$$L\left(C_t\right) = \int\limits_t^1 \sqrt{1 + \left[\frac{1}{u^{\theta+1}\cdot\left(1+t^{-\theta}-u^{-\theta}\right)^{\frac{\theta+1}{\theta}}}\right]^2}\, du. \tag{B.7}$$

B.4 Gumbel Copula

The generator of the Gumbel copula is $\psi(u) = \exp(-u^{-\frac{1}{\theta}})$. Using Corollary 10,

$$L\left(C_t\right) = \int\limits_t^1 \sqrt{1 + \left[\frac{(-\ln(u))^{\theta-1}\cdot\exp\left(-\left[(-\ln(t))^\theta - (-\ln(u))^\theta\right]^{\frac{1}{\theta}}\right)}{u\cdot\left[(-\ln(t))^\theta - (-\ln(u))^\theta\right]^{\frac{\theta-1}{\theta}}}\right]^2}\, du. \tag{B.8}$$

B.5 Frank Copula

The generator of the Frank copula is $\psi(u) = -(\ln(e^{-u}(e^{-\theta}-1)+1))/\theta$. Using Corollary 10,

$$L\left(C_t\right) = \int\limits_t^1 \sqrt{1 + \left[\frac{\exp(-\theta u)\cdot(\exp(-\theta t)-1)\cdot(\exp(-\theta)-1)}{(\exp(-\theta u)-1)^2\cdot\left[1+\frac{\exp(-\theta t)-1}{\exp(-\theta u)-1}\cdot(\exp(-\theta)-1)\right]}\right]^2}\, du. \quad \text{(B.9)}$$

Appendix C

Supplementary Material for the Examples of Chapter 5

C.1 Example 2

Let $u_0 := \sqrt{2/5}/5$. The copula C_1 is given by

$$C_1(u,v) := \begin{cases} 0 & (v \leq \frac{4-5u}{15} \wedge u \geq \frac{1}{5}) \\ & \vee(u \leq \frac{1}{5} \wedge v \leq \frac{4-15u}{5}) \\ u & u \leq \frac{1}{5} \wedge v \geq \frac{4}{5} \\ v & v \leq \frac{1}{5} \wedge u \geq \frac{4}{5} \\ -1+u+v & (v \geq \frac{16-5u}{15} \wedge u \leq \frac{4}{5}) \\ & \vee(u \geq \frac{4}{5} \wedge v \geq \frac{16-15u}{5}) \\ u_0\sqrt{(v-\frac{4-5u}{15})^2+(u-\frac{4}{5}-3v)^2} & v \leq \frac{1}{5} \wedge v \geq \frac{4-5u}{15} \wedge u \leq \frac{4}{5} \\ u_0\sqrt{(u-\frac{4-5v}{15})^2+(v-\frac{4}{5}-3u)^2} & u \leq \frac{1}{5} \wedge u \geq \frac{4-5v}{15} \wedge v \leq \frac{4}{5} \\ u_0\sqrt{(u-\frac{1}{5})^2+(\frac{1-5u}{15})^2}+u_0\sqrt{(\frac{1-5v}{15})^2+(\frac{1}{5}-v)^2} & u \geq \frac{1}{5} \wedge v \geq \frac{1}{5} \wedge v \leq -u+1 \\ u_0\sqrt{(\frac{5u-4}{15})^2+(\frac{4}{5}-u)^2}-\frac{23}{25}+u+v & u \leq \frac{4}{5} \wedge v \leq \frac{4}{5} \wedge v \geq -u+1 \\ v-u_0\sqrt{(v-\frac{1}{5})^2+(\frac{5v-1}{15})^2} & u \geq \frac{4}{5} \wedge v \geq \frac{1}{5} \wedge v \leq \frac{16-15u}{5} \\ u-u_0\sqrt{(u-\frac{1}{5})^2+(\frac{5u-1}{15})^2} & v \geq \frac{4}{5} \wedge u \geq \frac{1}{5} \wedge u \leq \frac{16-15v}{5} \end{cases} .$$

C.2 Example 3

The copula C_3 is given by

$$C_3(u,v) := \begin{cases} \frac{829uv}{500} & 5u \leq 1 \wedge 5v \leq 1 \\ \frac{10u(10v+493)+4930v+41117}{200000} & \frac{3}{10} < u < \frac{1}{2} \wedge \frac{3}{10} < v < \frac{1}{2} \\ \frac{25u(5v+384)+9600v-3016}{12500} & \frac{1}{5} < u \leq \frac{3}{10} \wedge \frac{1}{5} < v \leq \frac{3}{10} \\ \frac{u(185v+1736)}{2500} & 5u \leq 1 \wedge \frac{3}{10} < v \leq \frac{1}{2} \\ \frac{(185u+1736)v}{2500} & \frac{3}{10} < u \leq \frac{1}{2} \wedge 5v \leq 1 \\ \frac{u(1343v+1157)}{2500} & 5u \leq 1 \wedge 2v > 1 \\ \frac{(1343u+1157)v}{2500} & 5v \leq 1 \wedge 2u > 1 \\ \frac{25u(100v+741)-130v-233}{25000} & \frac{1}{5} < u \leq \frac{3}{10} \wedge \frac{3}{10} < v \leq \frac{1}{2} \\ \frac{25(100u+741)v-130u-233}{25000} & \frac{3}{10} < u \leq \frac{1}{2} \wedge \frac{1}{5} < v \leq \frac{3}{10} \\ \frac{25u(209v+291)+298(v-1)}{12500} & \frac{1}{5} < u \leq \frac{3}{10} \wedge 2v > 1 \\ \frac{1}{500}(209u+291)v + \frac{149(u-1)}{6250} & \frac{1}{5} < v \leq \frac{3}{10} \wedge 2u > 1 \\ \frac{u(5759v+491)+491(v-1)}{6250} & 2u > 1 \wedge 2v > 1 \\ u\left(\frac{77v}{20} - \frac{274}{625}\right) & 5u \leq 1 \wedge \frac{1}{5} < v \leq \frac{3}{10} \\ \left(\frac{77u}{20} - \frac{274}{625}\right)v & \frac{1}{5} < u \leq \frac{3}{10} \wedge 5v \leq 1 \\ \frac{10(9751u-4751)v-21791(u-1)}{50000} & 2u \geq 1 \wedge \frac{3}{10} < v \leq \frac{1}{2} \\ \frac{10u(9751v-4751)-21791(v-1)}{50000} & v \geq \frac{1}{2} \wedge \frac{3}{10} \leq u \leq \frac{1}{2} \end{cases}.$$

Appendix D

4-dimensional Vine Copulas

This appendix lists the vine arrays (cf. Section 6.3) of all 4-dimensional vine copulas. In total there are 24 distinct 4-dimensional vine copulas, 12 of which are D-vines and 12 of which are C-vines.

D-vines

$$\mathbf{D}_1 = \begin{pmatrix} 1 & 1 & 2 & 3 \\ & 2 & 1 & 2 \\ & & 3 & 1 \\ & & & 4 \end{pmatrix}, \mathbf{D}_2 = \begin{pmatrix} 2 & 2 & 1 & 3 \\ & 1 & 2 & 1 \\ & & 3 & 2 \\ & & & 4 \end{pmatrix}, \mathbf{D}_3 = \begin{pmatrix} 2 & 2 & 3 & 1 \\ & 3 & 2 & 3 \\ & & 1 & 2 \\ & & & 4 \end{pmatrix}$$

$$\mathbf{D}_4 = \begin{pmatrix} 2 & 2 & 3 & 4 \\ & 3 & 2 & 3 \\ & & 4 & 2 \\ & & & 1 \end{pmatrix}, \mathbf{D}_5 = \begin{pmatrix} 1 & 1 & 3 & 2 \\ & 3 & 1 & 3 \\ & & 2 & 1 \\ & & & 4 \end{pmatrix}, \mathbf{D}_6 = \begin{pmatrix} 1 & 1 & 3 & 4 \\ & 3 & 1 & 3 \\ & & 4 & 1 \\ & & & 2 \end{pmatrix}$$

$$\mathbf{D}_7 = \begin{pmatrix} 1 & 1 & 2 & 4 \\ & 2 & 1 & 2 \\ & & 4 & 1 \\ & & & 3 \end{pmatrix}, \mathbf{D}_8 = \begin{pmatrix} 1 & 1 & 4 & 2 \\ & 4 & 1 & 4 \\ & & 2 & 1 \\ & & & 3 \end{pmatrix}, \mathbf{D}_9 = \begin{pmatrix} 2 & 2 & 4 & 1 \\ & 4 & 2 & 4 \\ & & 1 & 2 \\ & & & 3 \end{pmatrix}$$

$$\mathbf{D}_{10} = \begin{pmatrix} 3 & 3 & 2 & 1 \\ & 2 & 3 & 2 \\ & & 1 & 3 \\ & & & 4 \end{pmatrix}, \mathbf{D}_{11} = \begin{pmatrix} 3 & 3 & 1 & 2 \\ & 1 & 3 & 1 \\ & & 2 & 3 \\ & & & 4 \end{pmatrix}, \mathbf{D}_{12} = \begin{pmatrix} 2 & 2 & 1 & 4 \\ & 1 & 2 & 1 \\ & & 4 & 2 \\ & & & 3 \end{pmatrix}$$

C-vines

$$\mathbf{C}_1 = \begin{pmatrix} 1 & 1 & 1 & 1 \\ & 2 & 2 & 2 \\ & & 3 & 3 \\ & & & 4 \end{pmatrix}, \mathbf{C}_2 = \begin{pmatrix} 1 & 1 & 1 & 1 \\ & 3 & 3 & 3 \\ & & 2 & 2 \\ & & & 4 \end{pmatrix}, \mathbf{C}_3 = \begin{pmatrix} 1 & 1 & 1 & 1 \\ & 4 & 4 & 4 \\ & & 2 & 2 \\ & & & 3 \end{pmatrix}$$

$$\mathbf{C}_4 = \begin{pmatrix} 2 & 2 & 2 & 2 \\ & 1 & 1 & 1 \\ & & 3 & 3 \\ & & & 4 \end{pmatrix}, \mathbf{C}_5 = \begin{pmatrix} 2 & 2 & 2 & 2 \\ & 3 & 3 & 3 \\ & & 1 & 1 \\ & & & 4 \end{pmatrix}, \mathbf{C}_6 = \begin{pmatrix} 2 & 2 & 2 & 2 \\ & 4 & 4 & 4 \\ & & 1 & 1 \\ & & & 3 \end{pmatrix}$$

$$\mathbf{C}_7 = \begin{pmatrix} 3 & 3 & 3 & 3 \\ & 1 & 1 & 1 \\ & & 2 & 2 \\ & & & 4 \end{pmatrix}, \mathbf{C}_8 = \begin{pmatrix} 3 & 3 & 3 & 3 \\ & 2 & 2 & 2 \\ & & 1 & 1 \\ & & & 4 \end{pmatrix}, \mathbf{C}_9 = \begin{pmatrix} 3 & 3 & 3 & 3 \\ & 4 & 4 & 4 \\ & & 1 & 1 \\ & & & 2 \end{pmatrix}$$

$$\mathbf{C}_{10} = \begin{pmatrix} 4 & 4 & 4 & 4 \\ & 1 & 1 & 1 \\ & & 2 & 2 \\ & & & 3 \end{pmatrix}, \mathbf{C}_{11} = \begin{pmatrix} 4 & 4 & 4 & 4 \\ & 2 & 2 & 2 \\ & & 1 & 1 \\ & & & 3 \end{pmatrix}, \mathbf{C}_{12} = \begin{pmatrix} 4 & 4 & 4 & 4 \\ & 3 & 3 & 3 \\ & & 1 & 1 \\ & & & 2 \end{pmatrix}$$

Appendix E

Simulation of Simplified Vines

The presented algorithms for simulation from a d-dimensional simplified vine copula are adapted from Joe (2015). For the algorithms, we need an auxiliary $(d-1)\times(d-1)$ array **fam** that contains in each row k the indices of the copulas in tree T_k as they appear in the respective ordered vine array. Examples for such arrays can be found in Section 6.3.

Algorithm 8 Simulation of d-dimensional simplified D-vine

```
Initialize two auxiliary $d \times d$ arrays $(a_{ij})$ and $(b_{ij})$ and an auxiliary $(d-1) \times (d-1)$ array **fam** that contains in each row $k$ the indices of the copulas in tree $T_k$ as they appear in an ordered D-vine array.
Generate $d$ independent uniformly distributed random variables $w_1, \dots, w_d$.
Set $u_1 = w_1$, $a_{11} = w_1$, $b_{11} = w_1$.
for $i = 2, \dots, d$
    $a_{i1} = w_i$
    for $j = 2, \dots, i$
        $a_{ij} = h^{-1}_{\mathbf{fam}(\mathrm{i-j+1,j-1})}(a_{i,j-1}, b_{i+1,j-1})$
    end for
    $u_i = a_{ii}$
    $b_{ii} = a_{ii}$
    for $j = i-1, \dots, 1$
        $b_{ij} = h_{\mathbf{fam}(\mathrm{i-j,j})}(b_{i-1,j}, a_{i,j+1})$
    end for
end for
Return $(u_1, \dots, u_d)$.
```

Algorithm 9 Simulation of d-dimensional simplified C-vine

```
Initialize an auxiliary $(d-1) \times (d-1)$ array $\mathbf{fam}$ that contains in each row $k$ the indices of the copulas in tree $T_k$ as they appear in an ordered C-vine array.
Generate $d$ independent uniformly distributed random variables $w_1, \ldots, w_d$.
Set $u_1 = w_1$, $u_2 = h^{-1}_{\mathbf{fam}(1,1)}(w_2, w_1)$.
for $i = 3, \ldots, d$
    $t = w_i$
    for $j = i-1, \ldots, 1$
        $t = h^{-1}_{\mathbf{fam}(\mathrm{j},\mathrm{i}-\mathrm{j})}(t, w_j)$
    end for
    $u_i = t$
end for
Return $(u_1, \ldots, u_d)$.
```

Algorithm 10 Simulation of d-dimensional simplified R-vine

```
Input ordered vine array $\mathbf{A} = (a_{kj})$
Initialize three auxiliary $d \times d$ arrays $(q_{ij})$ , $(v_{ij})$, and $(z_{ij})$ as well as an
  auxiliary $(d-1) \times (d-1)$ array **fam** that contains in each row $k$ the
  indices of the copulas in tree $T_k$ as they appear in the ordered R-vine
  array $\mathbf{A}$.
Compute upper-triangular matrix $\mathbf{M} = (m_{kj})$, where $m_{kj} =$
  $\max\{a_{1j}, \dots, a_{kj}\}$, for $k = 1, \dots, j-1$, $j = 2, \dots, d$.
Generate $d$ independent uniformly distributed random variables
  $w_1, \dots, w_d$.
Set $u_1 = w_1$, $u_2 = h^{-1}_{\mathbf{fam}(1,1)}(w_2, w_1)$, $q_{22} = w_2$, $v_{12} = h_{\mathbf{fam}(1,1)}(u_1, u_2)$.
**for** $j = 3, \dots, d$
    $q_{jj} = w_j$
    **for** $l = j-1, \dots, 2$
        if $a_{lj} = m_{lj}$ then $s = q_{l,a_{lj}}$, else $s = v_{l-1,m_{lj}}$
        $z_{lj} = s$
        $q_{lj} = h^{-1}_{\mathbf{fam}(\mathrm{l,j-1})}(q_{l+1,j}, s)$
    **end for**
    $q_{1j} = h^{-1}_{\mathbf{fam}(1,\mathrm{j}-1)}(q_{2j}, u_{a_{1j}})$
    $u_j = q_{1j}$
    $v_{1j} = h_{\mathbf{fam}(1,\mathrm{j}-1)}(u_{a_{1j}}, u_j)$
    **for** $l = 2, \dots, j-1$
        $v_{lj} = h_{\mathbf{fam}(\mathrm{l,j-1})}(z_{lj}, q_{lj})$
    **end for**
**end for**
Return $(u_1, \dots, u_d)$.
```

Bibliography

Aas, K., Czado, C., Frigessi, A., and Bakken, H. (2009). Pair-copula constructions of multiple dependence. *Insurance: Mathematics and Economics*, 44(2):182–198.

Abegaz, F., Gijbels, I., and Veraverbeke, N. (2012). Semiparametric estimation of conditional copulas. *Journal of Multivariate Analysis*, 110:43–73.

Acar, E. F., Craiu, R. V., and Yao, F. (2011). Dependence Calibration in Conditional Copulas: A Nonparametric Approach. *Biometrics*, 67(2):445–453.

Acar, E. F., Craiu, R. V., and Yao, F. (2013). Statistical testing of covariate effects in conditional copula models. *Electronic Journal of Statistics*, 7:2822–2850.

Acar, E. F., Genest, C., and Nešlehová, J. (2012). Beyond simplified pair-copula constructions. *Journal of Multivariate Analysis*, 110:74–90.

Aickin, M. and Gensler, H. (1996). Adjusting for Multiple Testing When Reporting Research Results: the Bonferron Vs Holm Methods. *American Journal of Public Health*, 86(5):726–729.

Akaike, H. (1973). Information Theory and an Extension of the Maximum Likelihood Princicple. In Petrov, B. and Csaki, E., editors, *Proceedings of the 2nd International Symposium of Information Theory*, pages 267–281, Budapest. Akademiai Kiado.

Azzalini, A. (1981). A Note on the Estimation of a Distribution Function and Quantiles by Kernel Method. *Biometrika*, 68(1):326–328.

Baillo, A., Cuesta-Albertos, J. A., and Cuevas, A. (2001). Convergence rates in nonparametric estimation of level sets. *Statistics & Probability Letters*, 53:27–35.

Barbe, P., Genest, C., Ghoudi, K., and Rémillard (1996). On Kendall's Process. *Journal of Multivariate Analysis*, 58(0048):197–229.

Bedford, T. and Cooke, R. M. (2001). Probability density decomposition for conditionally dependent random variables modeled by vines. *Annals of Mathematics and Artificial Intelligence*, 32(1-4):245–268.

Bedford, T. and Cooke, R. M. (2002). Vines: A New Graphical Model for Dependent Random Variables. *The Annals of Statistics*, 30(4):1031–1068.

Bender, R. and Lange, S. (2001). Adjusting for multiple testing - When and how? *Journal of Clinical Epidemiology*, 54(4):343–349.

Botev, A. Z. I., Grotowski, J. F., and Kroese, D. P. (2010). Kernel Density Estimation via Diffusion. *The Annals of Statistics*, 38(5):2916–2957.

Bücher, A. and Kojadinovic, I. (2016a). A dependent multiplier bootstrap for the sequential empirical copula process under strong mixing. *Bernoulli*, 22(2):927–968.

Bücher, A. and Kojadinovic, I. (2016b). Dependent multiplier bootstraps for non-degenerate U-statistics under mixing conditions with applications. *Journal of Statistical Planning and Inference*, 170:83–105.

Bücher, A., Kojadinovic, I., Rohmer, T., and Segers, J. (2014). Detecting changes in cross-sectional dependence in multivariate time series. *Journal of Multivariate Analysis*, 132:111–128.

Callau Poduje, A. C. and Haberlandt, U. (2018). Spatio-Temporal Synthesis of Continuous Precipitation Series Using Vine Copulas. *Water*, 10:862.

Chaudhuri, P. (1996). On a Geometric Notion of Quantiles for Multivariate Data. *Journal of the American Statistical Association*, 91(434):862–872.

Chebana, F. and Ouarda, T. (2011). Multivariate quantiles in hydrological frequency analysis. *Environmetrics*, 22(1):63–78.

Chen, J. and Gupta, A. K. (1997). Testing and Locating Variance Changepoints with Application to Stock Prices. *Journal of the American Statistical Association*, 92(438):739–747.

Chen, Y. C., Genovese, C. R., and Wasserman, L. (2017). Density Level Sets: Asymptotics, Inference, and Visualization. *Journal of the American Statistical Association*, 112(520):1684–1696.

Chollete, L., Heinen, A., and Valdesogo, A. (2009). Modeling International Financial Returns with a Multivariate Regime-switching Copula. *Journal of Financial Econometrics*, 7(4):437–480.

Chung, E. and Romano, J. P. (2013). Exact and Asymptotically Robust Permutation Tests. *The Annals of Statistics*, 41(2):484–507.

Coblenz, M. (2018). Non-Simplified Vine Copulas via Tessellation of Conditioning Spaces. Working Paper.

Coblenz, M., Dyckerhoff, R., and Grothe, O. (2018a). Confidence Regions for Multivariate Quantiles. *Water*, 10:996. doi: 10.3390/w10080996.

Coblenz, M., Dyckerhoff, R., and Grothe, O. (2018b). Nonparametric estimation of multivariate quantiles. *Environmetrics*, 29:e2488. doi: 10.1002/ENV.2488.

Coblenz, M., Grothe, O., Schreyer, M., and Trutschnig, W. (2018c). On the length of copula level curves. *Journal of Multivariate Analysis*, 167:347–365. doi: 10.1016/j.jmva.2018.06.001.

Cook, R. D. and Johnson, M. E. (1986). Generalized Burr-Pareto-Logistic Distributions With Applications to a Uranium Exploration Data Set. *Technometrics*, 28(2):123–131.

Cooke, R. M. (1997). Markov And Entropy Properties Of Tree And Vine Dependent Variables. In *Proceedings of the ASA Section of Bayesian Statistical Science*.

Cooke, R. M., Kurowicka, D., and Wilson, K. (2015). Sampling, conditionalizing, counting, merging, searching regular vines. *Journal of Multivariate Analysis*, 138:4–18.

Cousin, A. and Di Bernardino, E. (2013). On multivariate extensions of Value-at-Risk. *Journal of Multivariate Analysis*, 119:32–46.

Cuevas, A., González-Manteiga, W., and Rodríguez-Casal, A. (2006). Plug-in estimation of general level sets. *Australian and New Zealand Journal of Statistics*, 48(1):7–19.

Czado, C. (2010). Pair-Copula Constructions of Multivariate Copulas. In Jaworski, P., Durante, F., Härdle, W. K., and Rychlik, T., editors, *Copula Theory and Its Applications*, pages 93–109, Berlin, Heidelberg. Springer Berlin Heidelberg.

Czado, C., Brechmann, E. C., and Gruber, L. (2013). Selection of Vine Copulas. In Jaworski, P., Durante, F., and Härdle, W. K., editors, *Copulae in Mathematical and Quantitative Finance*, pages 17–37, Berlin, Heidelberg. Springer Berlin Heidelberg.

Dalla Valle, L., Leisen, F., and Rossini, L. (2017). Bayesian non-parametric conditional copula estimation of twin data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 67(3):523–548.

de Angelis, D. and Young, G. A. (1992). Smoothing The Bootstrap. *International Statistical Review*, 60(1):45–56.

Deheuvels, P. (1979). La fonction de dépendance empirique et ses propriétés. un test non paramétrique d'indépendance. *Académie Royale de Belgique. Bulletin de la Classe des Sciences*, 65(6):274–292.

Deheuvels, P. (1980). Nonparametric test of independence. In Raoult, J.-P., editor, *Statistique non paramétrique asymptotique*, volume 821 of *Lecture Notes in Mathematics*, pages 95–107. Springer, Berlin.

Derumigny, A. and Fermanian, J.-D. (2017). About tests of the "simplifying" assumption for conditional copulas. *Dependence Modeling*, 5:154–197.

Di Bernardino, E., Laloë, T., Maume-Deschamps, V., and Prieur, C. (2013). Plug-in estimation of level sets in a non-compact setting with applications in multivariate risk theory. *ESAIM: Probability and Statistics*, 17:236–256.

Di Bernardino, E. and Prieur, C. (2014). Estimation of multivariate conditional-tail-expectation using Kendall's process. *Journal of Nonparametric Statistics*, 26(2):241–267.

Di Bernardino, E. and Rullière, D. (2013). Distortions of multivariate distribution functions and associated level curves: Applications in multivariate risk theory. *Insurance: Mathematics and Economics*, 53(1):190–205.

DiCiccio, C. J. and Romano, J. P. (2017). Robust Permutation Tests For Correlation And Regression Coefficients. *Journal of the American Statistical Association*, 112(519):1211–1220.

Dijkstra, E. W. (1959). A Note on Two Problems in Connexion with Graphs. *Numerische Mathematik*, 1:269–271.

Dißmann, J., Brechmann, E., Czado, C., and Kurowicka, D. (2013). Selecting and estimating regular vine copulae and application to financial returns. *Computational Statistics & Data Analysis*, 59:52–69.

Dougherty, J., Kohavi, R., and Sahami, M. (1995). Supervised and Unsupervised Discretization of Continuous Features. *Machine Learning Proceedings 1995*, pages 194–202.

Durante, F. and Sempi, C. (2016). *Principles of copula theory.* CRC/Chapman & Hall, Boca Raton, FL.

Dyckerhoff, R. (2017). Convergence of depths and depth-trimmed regions. Working Paper, arXiv:1611.08721v2.

Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, 7(1):1–26.

Federer, H. (1996). *Geometric measure theory.* Classics in mathematics. Springer, Berlin.

Fermanian, J.-D. and Lopez, O. (2018). Single-index copulas. *Journal of Multivariate Analysis*, 165:27–55.

Fermanian, J. D. and Wegkamp, M. H. (2012). Time-dependent copulas. *Journal of Multivariate Analysis*, 110:19–29.

Fernández Sánchez, J. and Trutschnig, W. (2015). Conditioning-based metrics on the space of multivariate copulas and their interrelation with uniform and levelwise convergence and Iterated Function Systems. *Journal of Theoretical Probability*, 28(4):1311–1336.

Genest, C. and Boies, J. C. (2003). Detecting dependence with Kendall plots. *The American Statistician*, 57(4):275–284.

Genest, C. and Favre, A.-C. (2007). Everything You Always Wanted to Know about Copula Modeling but Were Afraid to Ask. *Journal of Hydrologic Engineering*, 12(4):347–368.

Genest, C., Ghoudi, K., and Rivest, L.-P. (1995). A Semiparametric Estimation Procedure Of Dependence Parameters In Multivariate Families of Distributions. *Biometrika*, 82(3):543–552.

Genest, C. and Rivest, L.-P. (1993). Statistical Inference Procedures for Bivariate Archimedian Copulas. *Journal of the American Statistical Association*, 88(423):1034–1043.

Genest, C. and Rivest, L.-P. (2001). On the multivariate probability integral transform. *Statistics & Probability Letters*, 53(4):391–399.

Ghoudi, K. and Rémillard, B. (1998). Empirical processes based on pseudo-observations. In Szyskowicz, B., editor, *Asymptotic Methods in Probability and Statistics, A Volume in Honour of Miklós Csörgö*, pages 171–197. Elsevier, Amsterdam.

Gijbels, I., Omelka, M., and Veraverbeke, N. (2012). Multivariate and functional covariates and conditional copulas. *Electronic Journal of Statistics*, 6:1273–1306.

Gijbels, I., Veraverbeke, N., and Omelka, M. (2011). Conditional copulas, association measures and their applications. *Computational Statistics and Data Analysis*, 55(5):1919–1932.

Greene, W. H. (2012). *Econometric analysis*. Pearson series in economics. Pearson, Boston, Mass., 7. edition.

Guégan, D. and Zhang, J. (2010). Change analysis of a dynamic copula for measuring dependence in multivariate financial data. *Quantitative Finance*, 10(4):421–430.

Hall, P., Diciccio, T. J., and Romano, J. P. (1989). On Smoothing and the Bootstrap. *The Annals of Probability*, 17(2):692–704.

Hawkins, D. L. (1987). A test for a change point in a parametric model based on a maximal wald-type statistic. *Sankhya*, 49(Series A, Pt. 3):368–376.

Hincks, T., Aspinall, W., Cooke, R., and Gernon, T. (2018). Oklahoma's induced seismicity strongly linked to wastewater injection depth. *Science*, 359(6381):1251–1255.

Ho, Y. H. S. and Lee, S. M. S. (2005). Iterated smoothed bootstrap confidence intervals for population quantiles. *The Annals of Statistics*, 33(1):437–462.

Hobæk Haff, I. (2012). Comparison of estimators for pair-copula constructions. *Journal of Multivariate Analysis*, 110(0047):91–105.

Hobæk Haff, I. (2013). Parameter estimation for pair-copula constructions. *Bernoulli*, 19(2):462–491.

Hobæk Haff, I., Aas, K., and Frigessi, A. (2010). On the simplified pair-copula construction - Simply useful or too simplistic? *Journal of Multivariate Analysis*, 101(5):1296–1310.

Hobæk Haff, I., Aas, K., Frigessi, A., and Lacal, V. (2016). Structure learning in Bayesian Networks using regular vines. *Computational Statistics & Data Analysis*, 101:186–208.

Hobæk Haff, I. and Segers, J. (2015). Nonparametric estimation of pair-copula constructions with the empirical pair-copula. *Computational Statistics & Data Analysis*, 84:1–13.

Hochberg, Y. (1988). A sharper bonferroni procedure for multiple tests of significance. *Biometrika*, 75(4):800–802.

Hofert, M. (2008). Sampling Archimedean copulas. *Computational Statistics and Data Analysis*, 52(12):5163–5174.

Holm, S. (1979). A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics*, 6(2):65–70.

Holmes, M., Kojadinovic, I., and Quessy, J. F. (2013). Nonparametric tests for change-point detection à la Gombay and Horváth. *Journal of Multivariate Analysis*, 115:16–32.

Hyndman, R. J. (1996). Computing and Graphing Highest Density Regions. *The American Statistician*, 50(2):120–126.

Janssen, A. and Pauls, T. (2003). How Do Bootstrap and Permutation Tests Work? *The Annals of Statistics*, 31(3):768–806.

Joe, H. (1996). Families of m-Variate Distrbutions with Given Margins and m(m-1)/2 Bivariate Dependence Parameters. *IMS Lecture Notes - Monograph Series*, 28:120–141.

Joe, H. (2005). Asymptotic efficiency of the two-stage estimation method for copula-based models. *Journal of Multivariate Analysis*, 94(2):401–419.

Joe, H. (2011a). Dependence Comparisons of Vine Copulae with Four or More Variables. In Kurowicka, D. and Joe, H., editors, *Dependence Modeling: Vine Copula Handbook*, pages 139–164, Singapore. World Scientific.

Joe, H. (2011b). Tail Dependence in Vine Copulae. In Kurowicka, D. and Joe, H., editors, *Dependence Modeling: Vine Copula Handbook*, pages 165–187, Singapore. World Scientific.

Joe, H. (2015). *Dependence Modeling with Copulas*. Chapman & Hall, Boca Raton, FL.

Joe, H., Li, H., and Nikoloulopoulos, A. K. (2010). Tail dependence functions and vine copulas. *Journal of Multivariate Analysis*, 101(1):252–270.

Kallenberg, O. (1997). *Foundations of modern probability.* Probability and its applications. Springer, New York.

Killiches, M., Kraus, D., and Czado, C. (2016). Using model distances to investigate the simplifying assumption, goodness-of-fit and truncation levels for vine copulas. Working Paper, arXiv:1602.05795v2.

Killiches, M., Kraus, D., and Czado, C. (2017). Examination and visualization of the simplifying assumption for vine copulas in three dimensions. *Australian & New Zealand Journal of Statistics*, 59(1):95–117.

Kraus, D. and Czado, C. (2017). Growing simplified vine copula trees: improving Dißmann's algorithm. Working Paper, arxiv:1703.05203.

Kurowicka, D. (2011). Optimal Truncation of Vines. In Kurowicka, D. and Joe, H., editors, *Dependence Modeling: Vine Copula Handbook*, pages 233–247, Singapore. World Scientific.

Kurz, M. S. and Spanhel, F. (2018). Testing the simplifying assumption in high-dimensional vine copulas. Working Paper, arXiv:1706.02338v2.

Lebrun, R. and Dutfoy, A. (2009). An innovating analysis of the Nataf transformation from the copula viewpoint. *Probabilistic Engineering Mechanics*, 24(3):312–320.

Levi, E. and Craiu, R. V. (2018). Bayesian inference for conditional copulas using Gaussian Process single index models. *Computational Statistics and Data Analysis*, 122:115–134.

Li, Q. and Racine, J. S. (2007). *Nonparametric Econometrics: Theory and Practice.* Princeton University Press, Princeton, N.J.

Li, X., Mikusiński, P., and Taylor, M. (1998). Strong Approximation of Copulas. *Journal of Mathematical Analysis and Applications*, 225(2):608–623.

Loomis, L. H. and Sternberg, S. (1980). *Advanced calculus.* Addison-Wesley series in mathematics. Addison-Wesley Publ., Reading, Mass.

Lopez-Paz, D., Hernández-Lobato, J. M., and Ghahramani, Z. (2013). Gaussian Process Vine Copulas for Multivariate Dependence. *Proceedings of the 30th International Conference on Machine Learning.*

Mammen, E. and Polonik, W. (2013). Confidence regions for level sets. *Journal of Multivariate Analysis*, 122:202–214.

Mikusiński, P. and Taylor, M. D. (2010). Some approximations of n-copulas. *Metrika*, 72(3):385–414.

Molchanov, I. S. (2005). *Theory of random sets*. Probability and its applications. Springer, London.

Morales-Nápoles, O. (2011). Counting Vines. In Kurowicka, D. and Joe, H., editors, *Dependence Modeling: Vine Copula Handbook*, pages 189–218, Singapore. World Scientific.

Nagler, T. and Czado, C. (2016). Evading the curse of dimensionality in nonparametric density estimation with simplified vine copulas. *Journal of Multivariate Analysis*, 151:69–89.

Nelsen, R. B. (2006). *An Introduction to Copulas*. Springer Series in Statistics. Springer New York, New York, NY.

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

Omelka, M., Gijbels, I., and Veraverbeke, N. (2009). Improved kernel estimation of copulas: Weak convergence and goodness-of-fit testing. *Annals of Statistics*, 37(5B):3023–3058.

Omelka, M. and Pauly, M. (2012). Testing equality of correlation coefficients in two populations via permutation methods. *Journal of Statistical Planning and Inference*, 142(6):1396–1406.

Pappadà, R., Durante, F., and Salvadori, G. (2017). Quantification of the environmental structural risk with spoiling ties: is randomization worthwhile? *Stochastic Environmental Research and Risk Assessment*, 31(10):2483–2497.

Patton, A. J. (2006). Modelling Asymmetric Exchange Rate Dependence. *International Economic Review*, 47(2):527–556.

Prim, R. (1957). Shortest Connection Networks And Some Generalizations. *Bell System Technical Journal*, 36(6):1389–1401.

Remillard, B. and Scaillet, O. (2009). Testing for equality between two copulas. *Journal of Multivariate Analysis*, 100(3):377–386.

Requena, A. I., Mediero, L., and Garrote, L. (2013). A bivariate return period based on copulas for hydrologic dam design: Accounting for reservoir routing in risk estimation. *Hydrology and Earth System Sciences*, 17(8):3023–3038.

Riesz, F. (1928). Sur la convergence en moyenne. *Acta Scientiarum Mathematicarum*, 4(1/2):58–64.

Rockafellar, R. T. (1970). *Convex Analysis*. Princeton mathematical series; 28. Princeton Univ. Press, Princeton, NJ.

Rockafellar, R. T. and Wets, R. J.-B. (1998). *Variational analysis*. Die Grundlehren der mathematischen Wissenschaften in Einzeldarstellungen; 317. Springer, Berlin.

Rosenblatt, M. (1952). Remarks on a Multivariate Transformation. *The Annals of Mathematical Statistics*, 23(3):470–472.

Salvadori, G. (2004). Bivariate return periods via 2-Copulas. *Statistical Methodology*, 1(1-2):129–144.

Salvadori, G. and De Michele, C. (2004). Frequency analysis via copulas: Theoretical aspects and applications to hydrological events. *Water Resources Research*, 40(12):1–17.

Salvadori, G. and De Michele, C. (2007). On the Use of Copulas in Hydrology: Theory and Practice. *Journal of Hydrologic Engineering*, 12(4):369–380.

Salvadori, G., De Michele, C., and Durante, F. (2011). On the return period and design in a multivariate framework. *Hydrology and Earth System Sciences*, 15(11):3293–3305.

Salvadori, G., Durante, F., De Michele, C., Bernardi, M., and Petrella, L. (2016). A multivariate copula-based framework for dealing with hazard scenarios and failure probabilities. *Water Resources Research*, 52:3701–3721.

Salvadori, G., Durante, F., and Perrone, E. (2013). Semi-parametric approximation of Kendall's distribution function and multivariate Return. *Journal de la Société Francaise de Statistique*, 154(1):151–173.

Salvadori, G., Durante, F., Tomasicchio, G. R., and D'Alessandro, F. (2015). Practical guidelines for the multivariate assessment of the structural risk in coastal and off-shore engineering. *Coastal Engineering*, 95:77–83.

Salvadori, G., Tomasicchio, G. R., and D'Alessandro, F. (2014). Practical guidelines for multivariate analysis and design in coastal and off-shore engineering. *Coastal Engineering*, 88:1–14.

Scarsini, M. (1984). On Measures of Concordance. *Stochastica*, 8(3):201–218.

Scheffé, H. (1947). A Useful Convergence Theorem for Probability Distributions. *The Annals of Mathematical Statistics*, 18(3):434–438.

Schellhase, C. and Spanhel, F. (2018). Estimating non-simplified vine copulas using penalized splines. *Statistics and Computing*, 28(2):387–409.

Schmid, F. and Schmidt, R. (2007). Multivariate conditional versions of Spearman's rho and related measures of tail dependence. *Journal of Multivariate Analysis*, 98(6):1123–1140.

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, 6(2):461–464.

Schweizer, B. and Wolff, E. F. (1981). On Nonparametric Measures of Dependence for Random Variables. *The Annals of Statistics*, 9(4):879–885.

Serfling, R. (1980). *Approximation Theorems of Mathematical Statistics.* John Wiley & Sons, New York.

Serfling, R. (2002). Quantile functions for multivariate analysis: Approaches and applications. *Statistica Neerlandica*, 56(2):214–232.

Serinaldi, F. (2013). An uncertain journey around the tails of multivariate hydrological distributions. *Water Resources Research*, 49(10):6527–6547.

Serinaldi, F. (2016). Can we tell more than we can know? The limits of bivariate drought analyses in the United States. *Stochastic Environmental Research and Risk Assessment*, 30(6):1691–1704.

Šidák, Z. (1967). Rectangular Confidence Regions for the Means of Multivariate Normal Distributions. *Journal of the Americal Statistical Association*, 62(318):626–633.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis.* Chapman & Hall, London.

Silverman, B. W. and Young, G. A. (1987). The bootstrap: To smooth or not to smooth? *Biometrika*, 74(3):469–479.

Sklar, A. (1959). Fonctions de répartition à n dimensions et leurs marges. *Publications de L'institut de Statistique de L'université de Paris*, 8:229–231.

Stöber, J., Joe, H., and Czado, C. (2013). Simplified pair copula constructions-Limitations and extensions. *Journal of Multivariate Analysis*, 119:101–118.

Tibiletti, L. (1993). On a new notion of multidimensional quantile. *Metron*, 51:77–83.

Trutschnig, W. (2011). On a strong metric on the space of copulas and its induced dependence measure. *Journal of Mathematical Analysis and Applications*, 384(2):690–705.

Trutschnig, W. (2012). Some results on the convergence of (quasi-) copulas. *Fuzzy Sets and Systems*, 191:113–121.

Trutschnig, W. and Fernández Sánchez, J. (2013). Some results on shuffles of two-dimensional copulas. *Journal of Statistical Planning and Inference*, 143(2):251–260.

Vatter, T. and Chavez-Demoulin, V. (2015). Generalized additive models for conditional dependence structures. *Journal of Multivariate Analysis*, 141:147–167.

Vatter, T. and Nagler, T. (2018). Generalized Additive Models for Pair-Copula Constructions. *Journal of Computational and Graphical Statistics*, pages 1–34.

Veraverbeke, N., Omelka, M., and Gijbels, I. (2011). Estimation of a conditional copula and association measures. *Scandinavian Journal of Statistics*, 38(4):766–780.

Vostrikova, L. (1981). Detecting ”disorder” in multidimensional random processes. *Soviet Mathematics Doklady*, 24(1):55–59.

Vuong, Q. H. (1989). Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses. *Econometrica*, 57(2):307–333.

Wilson, E. B. (1958). *Advanced calculus: a text upon select parts of differential calculus, differential equations, integral calculus, theory of functions, with numerous exercises.* Dover Publ., New York.

Yue, S., Ouarda, T., Bobée, B., Legendre, P., and Bruneau, P. (1999). The Gumbel mixed model for flood frequency analysis. *Journal of Hydrology*, 226(1-2):88–100.

Yue, S. and Rasmussen, P. (2002). Bivariate frequency analysis: Discussion of some useful concepts in hydrological application. *Hydrological Processes*, 16(14):2881–2898.

Zilko, A. A. and Kurowicka, D. (2016). Copula in a multivariate mixed discrete-continuous model. *Computational Statistics & Data Analysis*, 103:28–55.