TECHNISCHE UNIVERSITÄT MÜNCHEN

TUM School of Computation, Information and Technology

Statistical learning with vine copulas in regression settings

Marija Tepegjozova

Vollständiger Abdruck der von der TUM School of Computation, Information and Technology der Technischen Universität München zur Erlangung des akademischen Grades einer

Doktorin der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitz:		Prof. Dr. Aleksey Min
Prüfer*innen der Dissertation:	1.	Prof. Claudia Czado, Ph.D.
	2.	Prof. Dr. Luciana Dalla Valle
	3.	Prof. Dr. Ingrid Hobæk Haff

Die Dissertation wurde am 02.05.2023 bei der Technischen Universität München eingereicht und durch die School of Computation, Information and Technology am 19.06.2023 angenommen.

Abstract

Vine copulas allow for separate modeling of marginal distributions and the dependence structure, and can be specified by a sequence of linked trees, together with a set of bivariate copulas, with corresponding copula families and parameters. This thesis extends the existing literature on vine copula based models by several novel aspects: extension of vine based regression to allow for a less greedy forward selection algorithm using nonparametric pair copulas; proposal of a new vine structure, called Y-vine, for bivariate responses, where the conditional density can be specified without integration; development of estimation and prediction methods for Y-vine based regression, new methods for the determination of bivariate (un)conditional level curves, a simulation based approach for the determination of bivariate (un)conditional quantile curves; and proposal and estimation of new risk measures derived from the Y-vine model.

Quantile regression is a complementary method to linear regression, since computing a range of conditional quantile functions provides more accurate modeling of the stochastic relationship of the response variable given as set of predictor variables, especially in the tails. We introduce a nonrestrictive and highly flexible nonparametric quantile regression approach (for univariate response) based on vine copulas. This way, we obtain a model that overcomes typical issues of quantile regression such as quantile crossings or collinearity, the need for transformations and interactions of variables. We compare two different forward selection methods for predictors, based on maximizing the conditional log-likelihood, while taking into account one- or two-steps ahead in the next tree sequence.

Next, we introduce a novel vine tree sequence, that allows modeling of two response variables in a symmetric manner, so that the aforementioned benefits of vine copula based models are still valid. The main objective is estimating the joint conditional distribution of two response variables, given a set of predictor variables. We develop a forward selection of predictors for the bivariate response regression modeling. Also, we propose prediction and simulation methods for the novel regression model. Then, we dive into the topic of bivariate unconditional and conditional quantiles. We propose a new simulation based method of deriving bivariate quantiles connected with the usage of vine copulas.

Finally, utilizing the regression methods developed, we define univariate and bivariate vine copula based conditional probability risk measures, that are applied to a large

data set, involving climatological measurements in Bavaria over a sequence of years. Our focus is modeling the univariate frost and drought risks, and their associated bivariate joint risk, given a set of possible predictor variables.

Zusammenfassung

Vine Copulas ermöglichen eine separate Modellierung von Randverteilungen und der Abhängigkeitsstruktur und können durch eine Folge verknüpfter Bäume zusammen mit einem Satz bivariater Copulas mit entsprechenden Copula-Familien und Parametern spezifiziert werden. Diese Dissertation erweitert die bestehende Literatur zu Vine Copula basierten Modellen um mehrere neue Aspekte: Erweiterung der Vine basierten Regression, um einen weniger gierigen Vorwärtsselektionsalgorithmus mit nichtparametrischen bivariaten Copulas zu ermöglichen; Vorschlag einer neuen Vinestruktur namens Y-Vine für bivariate Zielvariablen, bei der die bedingte Dichte ohne Integration angegeben werden kann; Entwicklung von Schätz- und Vorhersagemethoden für die Y-Vine basierte Regression, neue Methoden zur Bestimmung bivariater (nicht) bedingter Niveaukurven, ein simulationsbasierter Ansatz zur Bestimmung bivariater (nicht) bedingter Quantilkurven; und Vorschlag und Schätzung neuer Risikomaße, die aus dem Y-Vine Modell abgeleitet werden.

Die Quantilregression ist eine ergänzende Methode zur linearen Regression. Die Berechnung einer Reihe von bedingten Quantilfunktionen ermöglicht eine genauere Modellierung der stochastischen Beziehung der Zielvariablen gegeben einen Satz von Prädiktorvariablen, insbesondere in den tails. Wir stellen einen nicht restriktiven und hochflexiblen nichtparametrischen Quantil-Regressionsansatz (für eine univariate Zielvariable) basierend auf Vine-Copulas vor. Auf diese Weise erhalten wir ein Modell, das typische Probleme der Quantilregression wie Quantilkreuzungen oder Kollinearität, die Notwendigkeit von Transformationen und Wechselwirkungen von Variablen überwindet. Wir vergleichen zwei verschiedene Vorwärtsauswahlmethoden für Prädiktoren, basierend auf der Maximierung der bedingten Log-Likelihood, während wir in der nächsten Baumsequenz einen oder zwei Schritte voraus berücksichtigen.

Als nächstes führen wir eine neuartige Vinestruktur ein, welche die symmetrische Modellierung von zwei Zielvariablen ermöglicht, sodass die oben genannten Vorteile von Vine Copula basierten Modellen weiterhin gültig sind. Das Hauptziel ist die Schätzung der gemeinsamen bedingten Verteilung zweier Zielvariablen, gegeben einen Satz von Prädiktorvariablen. Wir entwickeln einen Vorwärtsselektionsalgorithmus von Prädiktoren für die bivariate Regressionsmodellierung. Außerdem schlagen wir Vorhersage- und Simulationsmethoden für das neuartige Regressionsmodell vor. Dann tauchen wir in das Thema der bivariaten nicht bedingten und bedingten Quantile ein.

Wir schlagen eine neue simulationsbasierte Methode zur Herleitung bivariater Quantile vor, die mit der Verwendung von Vine Copulas verbunden ist.

Schließlich definieren wir unter Verwendung der entwickelten Regressionsmethoden univariate und bivariate Vine Copula basierte bedingte Wahrscheinlichkeitsrisikomaße, die auf einen großen Datensatz angewendet werden, der klimatologische Messungen in Bayern über eine Folge von Jahren beinhaltet. Unser Fokus liegt auf der Modellierung der univariaten Frost- und Dürrerisiken und des damit verbundenen gemeinsamen Risikos, wenn eine Reihe möglicher Prädiktorvariablen gegeben ist.

Acknowledgements

First and foremost, I would like to express my deepest gratitude to Prof. Claudia Czado for giving me the opportunity to pursue a PhD title and develop more my research skills. Our collaboration started with my master thesis, and I was more than happy to continue my research journey with vine copulas. I am very thankful for the supervision and guidance during my doctoral candidacy.

I would also like to express my gratitude to my mentor Prof. Matthias Scherer and the TUM ISAM Graduate school, especially Dr. Isabella Wiegand for the support and guidance during my time at TUM. Also, I would like to thank Prof. Giampiero Marra and his PhD student M.Sc. Alessia Eletti, for the amazing week I spend with them in London and all the nice discussions and research ideas we had, it was my real pleasure. Special thanks also to my coauthors, Prof. Gerda Claeskens, Prof. Christian Zang, Prof. Anja Rammig, Dr. Jing Zhou and M.Sc. Benjamin Meyer for the inspiring work we did together and all the fruitful discussions we had. Also, special thanks to Dr. Ricardo Acevedo Cabra and all involved in the TUM Data Innovation Lab, I learned so much by supervising the DI Lab projects.

Special thanks must also go to the other fellow PhD students for all the nice moments we had in the office, and all the fun we had at the conferences we went to. It was a pleasure working with you, Özge Sahin, Ariane Hanebeck, Hassan Alnasser and Alex Kreuzer. Also, I would like to thank all the master students that were part of the exercise sessions I taught and the master students who I co-supervised. I enjoyed sharing my knowledge with you and answering the really creative questions you asked, that deepened my knowledge and I was more than happy to learn from you as well.

Also, I would like to thank all my friends for all the support they provided. If it wasn't for your support during the hard corona times, and all the fun times we had, this journey would not have been the same.

Finally, and most importantly, I want to sincerely thank my family and my boyfriend for their continuous and unconditional love and encouragement. Thank you for everything you did for me and for always bringing me strength and joy in my life.

Contents

1. Introduction

"Data is the new oil" is a quote from the British mathematician Clive Humby, who coined the phrase in 2006. Later this quote was expanded on as *"Data is just like crude. It's valuable, but if unrefined it cannot really be used. It has to be changed into gas, plastic, chemicals, etc. to create a valuable entity that drives profitable activity; so must data be broken down, analyzed for it to have value."* [1] However, it was not until 2017 that this idea gained significant traction when "The Economist" published an article titled "The world's most valuable resource is no longer oil, but data" [2]. This article sparked widespread discussions and became a well-adopted tagline for the upcoming Fourth Industrial Revolution, mainly based on data driven solutions and data analytics. These statements lay out the foundation why it is imperative to be able to analyze data in a proper manner, by making less assumptions (about the data generating process, the underlying distribution and similar) and introduce more flexible approaches.

Nowadays, given the huge computational capabilities we have, data collection and data mining is very cheap and easy. This results in the creation of data sets that have a significant amount of dependencies among the collected variables. However, the usual statistical tools for data analysis are not able nor are flexible enough to capture and explain these dependencies. Thus, models that can deal with dependencies and extract insights from it are of paramount importance. A possible solution to this, are copula based models. The copula approach is a multivariate modeling approach that can handle complex dependence structures. It allows for separate modeling of the copula function and its arguments, the univariate marginal distributions functions. As a result, a wide range of dependence structures can be modeled by utilizing different functional forms for both, the copula function and the marginal distribution functions.

However, the number of parameters required to estimate a copula function increases in many cases quadratically with the number of variables. Thus, generalization to higher dimensions of copulas is rather complicated. Multivariate Gaussian and Student-t copulas have been widely used, for example in Lasmar and Berthoumieu (2014) for texture image retrieval or in Renard and Lang (2007) for multivariate extreme value analysis in hydrology using a multivariate Gaussian copula; in Liang et al. (2013) for

[1] https://www.theguardian.com/technology/2013/aug/23/tech-giants-data

[2] https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data

independent vector analysis in non-stationary signal processing or in W. Sun et al. (2008) for modeling the comovement of indexes in the German equity markets using a multivariate Student-t copula. However, they are not suited for asymmetric dependence structures. Copulas based on generator functions such as the class of Archimedean copulas, have only one or two parameters and thus, are very restrictive. They however can deal with asymmetry. High-dimensional hierarchical Archimedean copulas are more flexible, (see Savu and Trede (2010) for portofolio risk management), however they impose constraints as well (Berg and Aas, 2009).

Nevertheless, it is possible to overcome all of these aforementioned limitations by using vine copulas. Using the fact that there is an abundance of bivariate copulas or pair copulas (more details in Czado (2019, Chapter 3)), Joe (1996) and later Bedford and Cooke (2002) pioneered a pair copula construction (PCC) method. PCC is a flexible way of high-dimensional copula construction that uses a set of pair copulas. These pair copulas are chosen independent of each other, based on the (conditional) dependence characteristics of each pair of variables. The resulting multivariate copula of the PCC method is a so-called regular vine copula. A regular (R-)vine copula is characterized by a sequence of trees which define the unconditional and conditional pairwise dependencies that are considered in the model. This flexible model can handle high-dimensional data with asymmetric dependencies and tail-dependencies, which is the main benefit of their usage. Later, Aas et al. (2009) introduced statistical inference techniques, such as maximum likelihood estimation for R-vine copulas and Dissmann et al. (2013) introduced a sequential top-down approach for selecting and fitting an R-vine copula to given data. Ever since then, the theory of R-vine copulas has undergone constant refinement.

Another important advance is also the development of the R statistical software package `rvinecopulib` (Nagler and Vatter, 2021), which has made R-vine copula-based modeling accessible to a diverse range of statisticians and practitioners. More recently, vine copulas have been investigated in the areas of regression modeling (Kraus and Czado, 2017; Chang and Joe, 2019; Zhu et al., 2021), clustering and mixture models (Kim et al., 2013; M. Sun et al., 2016; Sahin and Czado, 2022), time-series modeling (Vatter and Nagler, 2018; Kreuzer and Czado, 2021; Nagler et al., 2022), structural equation models and Bayesian networks (Haff et al., 2016; Cooke et al., 2022; Czado and Scharl, 2021) and other areas. Further, R-vine copulas have been applied in a variety of research areas, such as climatology and environmental sciences (T. Wu et al., 2022; Tao et al., 2021; H. Li et al., 2021; Niemierko et al., 2019; Bevacqua et al., 2017; Ansell and Valle, 2021), health sciences (D'Urso et al., 2022; Barthel et al., 2018; Ye et al., 2022), finance and economics (Kielmann et al., 2022; Czado et al., 2022) or engineering (Torre et al., 2019; Cheng et al., 2020; Qian and Dong, 2022).

In this thesis we extend the current research in statistical learning using vine copulas

in a regression setting. Our main goal is adapting and developing new regression frameworks, which implies modeling a continuous response/s, using all the benefits a vine copula model offers. Comprehending the dependence of response variable/s and determining their statistical properties in connection to a set of predictor variables is an very important topic in statistical learning. Being able to do so using a very flexible framework that does not make strict distributional assumptions and can handle complex dependencies is a major advantage.

The assumption made by the most widely used quantile regression method, the linear quantile regression (Koenker and Bassett, 1978), that the dependence between a normally distributed response and predictors is a Gaussian copula, is overly restrictive and is rarely satisfied in real-world applications, thus resulting in quantile crossings and model-misspecification (Bernard and Czado, 2015). Thus, our main motivation for the first new contribution is the fact that flexible models without restrictive assumptions are of interest in practise. It is centered around introducing a fully nonparametric quantile regression framework based on two different R-vines structures: C-vine copulas and D-vine copulas. The class of C-vine copulas are characterized by a sequence of stars (a tree with one node having edges with all the other nodes) and D-vine copulas are characterized by a sequence of paths (a tree with a sequence of edges, each one incident to the next). In this approach both the marginal distribution functions and the pair copulas are estimated nonparametrically, to reduce possible bias introduced by parametric assumptions. Also, quantile crossings and collinearity are avoided by construction, and there is no need for transformations or interactions of variables.

The vine copula models are constructed in a way that the conditional distribution function of the response, given the predictors, or the conditional quantile function, can be calculated in an analytic manner, which quarantees precise and computationally inexpensive results. It is also constructed in a sequential manner, adding one predictor at a time, using two different forward selection approaches. One is a greedy approach by only considering the benefit of adding a predictor one-step ahead in the model (Kraus and Czado, 2017) and the other proposal considers two-steps ahead in the model construction. This way we also order the predictors by their influence on the response variable. The nonparametric estimation approaches we use are very fast for estimation, and we test the different approaches introduced in both low- and high-dimensional data (with the help of a newly introduced variable selection reduction for the two-step ahead approach).

The next contribution introduces a vine based regression framework that is able to handle two response regression modeling. It is motivated by a large number of data applications where the main interest is modeling the joint conditional distribution function of two responses given a set of any number of predictors (Frees et al., 2016; Singh et al., 2022; Bevacqua et al., 2017). In many such data studies there exist

dependence, not only among the predictor variables, but also between the two response variables. Ignoring this dependence between the response variables is not advisable, as it produces biased results. Thus, we introduce a novel vine tree sequence, called a Y-vine tree sequence that has the ability for a symmetric treatment of two response variables in a regression setting. The resulting Y-vine copula has the usual benefits of vine copula modeling and in addition, the joint conditional density of the responses given the predictors is analytically expressible using only pair copulas that are part of the Y-vine copula specification, making it computationally inexpensive. For bivariate response regression modeling, we also introduce an appropriate fit measure to be used for an automatic forward selection algorithm. It is able to construct the Y-vine in a sequential manner and exclude non-influential predictors. This results in a predictor order based on their influence on the two response variables. In addition, we develop a method to predict the bivariate conditional distribution given a set of predictor values, and show how to simulate bivariate response data from a Y-vine copula model.

Next, we utilize the proposed Y-vine regression framework to analyse and estimate conditional level curves and bivariate quantile curves based on a bivariate conditional distribution function. The bivariate conditional level curves and quantiles are especially important for modeling joint risks of events while adjusting for tail and asymmetric dependencies. We develop a numerical method for the estimation of bivariate (un)conditional level curves and a simulation based method to adjust them to bivariate quantile curves, so that the probability coverage below and above the quantile curves is exact. This way we are able to construct bivariate confidence regions, which are generalizations of confidence intervals in the univariate case. They can be used to identify bivariate outliers, visualize trends and identify parts of the conditional distribution having high density values, given observations of the predictors. This allows us to observe how the joint conditional distribution changes as the conditioning values of the predictors change.

The final contribution in this thesis, is a large data application of the vine copula based regression methods suggested. We use a data set, containing annual data for the period between 1952-2020 having values for 26 variables on a fine 5km by 5km grid of Bavaria, Germany. These 26 variables include, drought and late-frost indices, and bioclimatic and topographical (terrain) variables that are influencing the occurrence of extreme drought and frost events. Motivated by the issues the changing climate might impose on the forest ecosystems, mainly influenced by extreme drought and late-frost occurrences, we do a historical data analysis on the univariate drought and late-frost risks and a joint risk analysis of these extremes. For the univariate analyses, we fit a D-vine regression model, for which we propose a conditional probability risk measure of extreme drought or late-frost risk conditioned on a set of predictors. For the bivariate analysis, we fit a Y-vine regression model to both responses, and propose a

corresponding joint conditional probability risk measure conditioned on a set of chosen predictor values. We also develop methods to obtain temporal and spatial "at-risk" regions, so that forest management recommendations can be suggested.

Outline of the thesis
The content in the thesis is based on 3 manuscripts:

- Tepegjozova, M., J. Zhou, G. Claeskens, and C. Czado (2022).
  "Nonparametric C- and D-vine based quantile regression."
  Dependence Modeling 10.1, pp. 1–21.

- Tepegjozova, M. and C. Czado (2022).
  "Bivariate vine copula based quantile regression."
  Under revision at Computational Statistics & Data Analysis, arXiv:2205.02557.

- Tepegjozova, M., B. Meyer, A. Rammig, C. Zang and C. Czado (2023).
  "Univariate and bivariate risk analysis of late-frost and drought conditions using vine copulas in Bavaria."
  To be submitted to the Journal of the Royal Statistical Society, Series C (Applied Statistics).

For the thesis some content of these papers has been revised and extended in various sections, including additional methodology, illustrations or explanations.

We start with a brief introduction of the foundational concepts of copulas and vine copulas in Chapter 2. Section 2.1 deals with the basic properties of copulas, the dependence measures used in the thesis and the introduction to some parametric and nonparametric bivariate (or pair) copulas that we use. Section 2.2 introduces the concept of a pair copula construction (PCC) and vine copulas.

Chapter 3 is mainly based on Tepegjozova et al. (2022) with some modifications. Motivated by a thorough literature review in Section 3.1, we propose a nonparametric vine copula based quantile regression framework in Section 3.2. The algorithms that we use for forward selection of predictors, in order to construct the vine copula model, are explained in Section 3.3. Section 3.3.1 deals with the one-step ahead approach, while Section 3.3.2 deals with the two-step ahead approach. We explain the algorithms in detail, in a way that the difference between the two different forward selection approaches is evident. Also, the algorithms are presented in a uniform manner, so that either a C-vine or a D-vine structure can be constructed, based on the data at hand. Additionally, we extend the two-step ahead approach with a variable selection reduction in Section 3.3.2, so that the two-step ahead models are applicable to many

predictors which would be impossible otherwise, due to the increased computational cost of the two-step ahead approach.

The finite sample performance of the nonparametric vine copula based conditional quantile estimator is evaluated in Section 3.4 by several performance measures suitable for quantiles in various low- and high-dimensional simulation settings. We explore the performance of the 4 newly introduced models in both low- and high-dimensional real data application in Section 3.5. We dive deeper especially in the data analysis of the Concrete data set (Yeh, 1998) in Section 3.5.1, so that we can analyse and compare how the 4 different approaches choose different orders of the predictors. We also explore what is the optimal order of the predictors (ordered by their influence on the response variable) that each model finds after 100 replications of data splitting and model fitting. In a similar manner, using the optimal orders we find the most influential predictors in a high-dimensional setting, faced in the Riboflavin data set in Section 3.5.2.

In Section 3.6, we explore a possible stopping criteria for the nonparametric case, which we use to recompute the models on the Concrete data set and compare the out-of-sample statistics with the usage of the stopping criteria. However, the results are ambiguous and more research is needed in this area. Finally, in Section 3.7 we conclude and discuss possible directions of future research.

Chapters 4 and 5 are mainly based on Tepegjozova and Czado, 2022 with major modifications. We start by motivating the problem of conditional distribution function estimation of multiple responses given a set of predictor variables in Section 4.1. Then, in Section 4.2 we introduce the Y-vine copula based regression model for bivariate responses. Here we introduce a novel vine tree structure, specifically designed for a bivariate response regression. In Section 4.3, we propose an automatic forward selection algorithm of predictors. It is based on a novel adjusted conditional log-likelihood fit measure, that sequentially adds predictors to the model based on quantifying the influence of the predictors on the two responses. For application purposes, in Section 4.4 we present a prediction method for Y-vine copulas. Section 4.5 deals with the simulation of a bivariate response data from a specified Y-vine copula.

The implementation of the Y-vine regression together with all the other tools discussed in this whole chapter is discussed in Section 4.6. For demonstration of the usefulness of this novel method we include a real data example in Section 4.7 that contains dependent bivariate responses, the minimum and maximum daily temperatures. Finally, in Section 4.8 we give conclusions and areas of future research.

Section 5.1 motivates the problem of construction of multivariate quantiles and the lack of a consensus of the generalization of univariate quantiles to the multivariate case. It also deals with applications where multivariate quantiles based on copulas have been used. Next, in Section 5.2 we define bivariate (un)conditional level curves

associated with (Y-vine) copula derived distribution functions. Section 5.3 develops a numerical method used for the construction of the bivariate (un)conditional level curves. In Section 5.3.2, we employ and test this numerical procedure for known parametric pair copulas and a 3-dimensional vine copula model.

For the Y-vine regression model, we present a possible adjustment of the bivariate level curves to provide bivariate quantile curves, with exact coverage probabilities. We propose a simulation based method for the estimation of the bivariate quantile curves and also a construction of bivariate confidence regions. We continue the data analysis started in Section 4.7, in Section 5.5. Here we illustrate the bivariate (un)conditional level curves for the temperature data set, the bivariate (un)conditional quantile curves and the corresponding confidence regions. We also highlight the advantages of bivariate response modelling over standard univariate models that assume (conditional) independence between the two responses.

Chapter 6 forms the basis of Tepegjozova et al. (2023). It contains a vine copula based analysis of a large real data set in the area of climatology. Section 6.1 discusses the need and importance of statistical methods to be able to analyse the joint occurrence of two extremes, drought and frost, given a set of possible predictor variables, among which there is high dependence. Next, Section 6.2 describes the data set utilized. Section 6.3 introduces the data modeling approaches we use, by employing the D-vine and Y-vine copula regression previously introduced. It contains an exploratory dependence analysis of the data at hand and an exploration of the fitted models, by studying the pair copula families that are fitted and the orders of the predictors the models choose.

Section 6.4 suggests novel conditional risk probability measures for the D-vine and Y-vine regression models. These vine copula based risk measures are used to identify high risk years and regions, for both univariate and bivariate responses. We also estimate corresponding survival probabilities and analyse how these conditional probabilities vary over all locations in Section 6.5. Based on the survival probabilities, we also estimate return periods for each extreme in Section 6.6. All these measures are used for finding temporal and spatial "at-risk" regions. Finally, we conclude and propose possible areas of future research in Section 6.7.

Finally, in Chapter 7 we summarize the overall contributions of the thesis, and propose future areas of research.

2. Preliminaries

2.1. Copulas

Let $\mathbf{X}$ be a continuous d-dimensional random vector $\mathbf{X} = (X_1, \ldots, X_d)^T$ with observed values $\mathbf{x} = (x_1, \ldots, x_d)^T$. Assume $\mathbf{X}$ have joint distribution function F, joint density f and marginal distributions F_{X_i}, $i = 1, \ldots, d$. The fundamental representation theorem for multivariate distributions in terms of their marginal distributions and a corresponding d-dimensional copula C, by Sklar (1959) states the following

$$F(x_1, \ldots, x_d) = C(F_{X_1}(x_1), \ldots, F_{X_d}(x_d)). \tag{2.1}$$

The copula $C : [0,1]^d \mapsto [0,1]$ corresponds to the distribution of the random vector $\mathbf{U} = (U_1, \ldots, U_d)^T$, where the components of $\mathbf{U}$ (u-scale) are the probability integral transforms (PITs) of the components of $\mathbf{X}$ (x-scale), $U_i = F_{X_i}(X_i)$ for $i = 1, \ldots, d$. Every U_i is uniformly distributed and their joint distribution function C is the copula associated with $\mathbf{X}$. If all marginal distributions F_{X_i} are continuous, then Sklar's Theorem implies that C is unique. If derivatives of the marginal distributions F_{X_i} exist, then the density f can be derived as

$$f(x_1, \ldots, x_d) = c(F_{X_1}(x_1), \ldots, F_{X_d}(x_d)) \cdot \prod_{i=1}^{d} f_{X_i}(x_i), \tag{2.2}$$

where c is the d-dimensional density corresponding to the copula C and $f_{X_1}, \ldots, f_{X_d}$ are the univariate marginal densities. (see more in Nelsen (2007)). Throughout this thesis, we assume that all considered random variables to be continuous.

2.1.1. Dependence measures

To quantify and characterize the dependence between random variables measures of dependence are needed. In the thesis, we will use the Kendall's τ and the

Kendall's τ is a rank based dependence measure with range of values in the interval $[-1, 1]$ defined in M. G. Kendall (1938). It is defined as the probability of concordance minus the probability of discordance of two continuous random variables, say X_1 and X_2

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

where (X_{11}, X_{12}) and (X_{21}, X_{22}) are independent and identically distributed copies of the pair (X_1, X_2).

Closer values to the boundaries of the interval $[-1, 1]$ mean greater dependence. Positive values indicate positive dependence, while negative values of the Kendall's τ indicate negative dependence between two random variables. A value of Kendall's $\tau = 0$ indicated independence. Also, Kendall's τ being a rank-based dependence measure is invariant with respect to monotone transformations of the margins.

The partial correlation is a dependence measure between two variables after the linear effect of the remaining variables is been removed. Let $X_1, \ldots, X_d$ be random variables with zero mean and variance σ_i^2 for $i = 1, \ldots, d$. Let $I^d_{-(i,j)}$ be the set $\{1, \ldots, d\}$ with indices i and j removed, for $i \neq j$. Following Udny Yule, M. Kendall, et al. (1950), define partial regression coefficients $b_{i,j;I^d_{-(i,j)}}$ for $i < j$ as the quantities that minimize

$$E[(X_i - \sum_{j=2, j\neq i}^{d} a_{i,j;I^d_{-(i,j)}} X_j)^2].$$

Then, the partial correlation $\rho_{i,j;I^d_{-(i,j)}}$ is defined as

$$\rho_{i,j;I^d_{-(i,j)}} := sign(b_{i,j;I^d_{-(i,j)}}) \times \sqrt{b_{i,j;I^d_{-(i,j)}} \times b_{i,j;I^d_{-(i,j)}}}.$$

2.1.2. Bivariate (pair) copulas

To adequately model various types of dependencies, there is a wide range of parametric and nonparametric bivariate or pair copulas. For example, a Clayton copula can characterize lower tail dependence, while a Gumbel copula characterizes upper tail dependence. Gaussian copula characterizes symmetric dependence with no tail preference, while Student-t characterizes symmetric lower and upper tail dependence, governed by the same parameter (Czado, 2019).

Parametric pair copulas

The parametric pair copulas are characterized by the copula family and corresponding parameters. Depending on their construction, there is a distinction between elliptical copulas, which are based on elliptical distributions and are constructed by applying the inverse statement of Sklar's Theorem (Sklar, 1959), such as Gaussian and Student-t copula, and Archemedian copulas, constructed using a generator function ϕ.

Examples of elliptical copulas and the corresponding bivariate distribution functions are listed below.

- Bivariate Gaussian copula
  Let $\Phi_1(\cdot)$ be the distribution function of a univariate standard normal distribution (with zero mean and unit variance) denoted as $N(0,1)$ and let $\Phi_2(\cdot)$ be the distribution function of a bivariate standard normal distribution, denoted as $N_2\left((0,0)^T, \Sigma\right)$, with zero mean vector and where Σ is a symmetric positive definite 2×2 correlation matrix with unit variance. Then, by applying the inverse Sklar's theorem (Sklar, 1959) the bivariate Gaussian copula distribution function is given as
  $$C(u_1, u_2) = \Phi_2\left(\Phi_1^{-1}(u_1), \Phi_1^{-1}(u_2)\right). \tag{2.3}$$

- Bivariate Student-t copula
  Let $T_{1,v}(\cdot)$ be the distribution function of a univariate standard Student-t distribution with $v > 0$ degrees of freedom, zero mean, and unit scale parameter, denoted as $t_1(v, 0, 1)$ and let $T_{2,v}(\cdot)$ be the distribution function of a bivariate standard Student-t distribution, denoted as $t_2(v, (0,0)^T, \Sigma)$, with $v > 0$ degrees of freedom, zero mean vector, and $\Sigma \in [-1,1]^{2\times 2}$ is a scale parameter matrix. By applying the inverse Sklar's theorem the bivariate Student-t copula distribution function is given as
  $$C(u_1, u_2) = T_{2,v}\left(T_{1,v}^{-1}(u_1), T_{1,v}^{-1}(u_2)\right). \tag{2.4}$$

Examples of Archimedean bivariate copulas with a single parameter and the corresponding bivariate distribution functions are listed below.

- Clayton copula
  $$C(u_1, u_2) = (u_1^{-\theta} + u_2^{-\theta} - 1)^{-\frac{1}{\theta}}, \tag{2.5}$$
  where $0 < \theta < \infty$ is the parameter controlling the degree of dependence. Independence correspond to $\theta \to 0$, and full dependence to $\theta \to \infty$.

- Gumbel copula
  $$C(u_1, u_2) = exp[-\{(-\ln u_1)^\theta + (-\ln u_2)^\theta\}^{\frac{1}{\theta}}], \tag{2.6}$$
  where $\theta \geq 1$ is the parameter controlling the degree of dependence. Independence correspond to $\theta = 1$, and full dependence to $\theta \to \infty$.

- Frank copula
  $$C(u_1, u_2) = -\frac{1}{\theta} \ln\left(\frac{1}{1 - e^{-\theta}}[(1 - e^{-\theta}) - (1 - e^{-\theta u_1})(1 - e^{-\theta u_2})]\right),$$
  where $\theta \in [-\infty, \infty] \setminus \{0\}$.

- Joe copula

$$C(u_1, u_2) = 1 - \Big((1-u_1)^\theta + (1-u_2)^\theta - (1-u_1)^\theta(1-u_2)^\theta\Big)^{\frac{1}{\theta}},$$

  where the parameter is $\theta \geq 1$.

Further, there is a one-to-one correspondence between the copula parameter and the Kendal's tau for elliptical copulas (Embrechts et al., 2003), and for the one-parameter Archimedean copulas (Hürlimann, 2003). There are also Archimedean copulas with two parameters, such as the BB copulas introduced in (Joe, 1997).

Nonparametric pair copulas

There are many approaches how to estimate bivariate copula densities in a nonparametric manner. Examples include the mirror-reflection estimator (Gijbels and Mielniczuk, 1990), the transformation estimator (Charpentier et al., 2007), the transformation local likelihood estimator (Geenens et al., 2017), the tapered transformation estimator (Wen and X. Wu, 2015) and the beta kernel estimator (Charpentier et al., 2007). Among the above-mentioned estimators, the transformation local likelihood estimator (Geenens et al., 2017) was found by Nagler et al. (2017) to have an overall best performance. Therefore, following Nagler et al. (2017) and Tepegjozova et al. (2022) we review shortly the construction of the transformation local likelihood estimator.

Let the $N \times 2$ transformed sample matrix be $D = (S, T)$, where the transformed samples $D_n = \big(S_n = \Phi^{-1}(U_i^{(n)}), T_n = \Phi^{-1}(U_j^{(n)})\big), n = 1, \ldots, N$, and Φ denotes the cumulative distribution function of a standard Gaussian distribution. The logarithm of the density $f_{S,T}$ of the transformed samples $(S_n, T_n), n = 1, \ldots, N$ is approximated locally by a bivariate polynomial expansion $P_{\boldsymbol{a}_m}$ of order m with intercept $\tilde{a}_{m,0}$ such that the approximation is given by

$$\tilde{f}_{S,T}(\Phi^{-1}(u_i^{(n)}), \Phi^{-1}(u_j^{(n)})) = \exp\left\{\tilde{a}_{m,0}(\Phi^{-1}(u_i^{(n)}), \Phi^{-1}(u_j^{(n)}))\right\}.$$

The transformation local likelihood estimator for the copula density is then defined as

$$\tilde{c}(u_i^{(n)}, u_j^{(n)}) = \frac{\tilde{f}_{S,T}(\Phi^{-1}(u_i^{(n)}), \Phi^{-1}(u_j^{(n)}))}{\phi(\Phi^{-1}(u_i^{(n)}))\phi(\Phi^{-1}(u_j^{(n)}))}.$$

To get the local polynomial approximation, we need a kernel function $\boldsymbol{K}$ with 2×2 bandwidth matrix $\boldsymbol{B}_N$. For some pair $(\check{s}, \check{t})$ close to (s, t), $\log f_{ST}(\check{s}, \check{t})$ is assumed to be well approximated, locally, by for instance a polynomial with $m = 1$ (log-linear)

$$P_{\boldsymbol{a}_1}(\check{s} - s, \check{t} - t) = a_{1,0}(s, t) + a_{1,1}(s, t)(\check{s} - s) + a_{1,2}(s, t)(\check{t} - t),$$

or $m = 2$ (log-quadratic)

$$P_{\boldsymbol{a}_2}(\check{s}-s,\check{t}-t) = a_{2,0}(s,t) + a_{2,1}(s,t)(\check{s}-s) + a_{2,2}(s,t)(\check{t}-t) \\ +a_{2,3}(s,t)(\check{s}-s)^2 + a_{2,4}(s,t)(\check{t}-t)^2 + a_{2,5}(s,t)(\check{s}-s)(\check{t}-t).$$

The coefficient vector of the polynomial expansion P_{a_m} is denoted by $\mathbf{a}_m(s,t)$, where $\mathbf{a}_1(s,t) = (a_{1,0}(s,t), a_{1,1}(s,t), a_{1,2}(s,t))$ for the log-linear approximation and $\mathbf{a}_2(s,t) = (a_{2,0}(s,t), \ldots, a_{2,5}(s,t))$ for the log-quadratic. The estimated coefficient vector $\tilde{\boldsymbol{a}}_m(s,t)$ is obtained by a maximization problem in

$$\tilde{\boldsymbol{a}}_m(s,t) = \arg\max_{a_m} \Bigg\{ \sum_{n=1}^{N} \boldsymbol{K}\Bigg(\boldsymbol{B}_N^{-1/2} \begin{pmatrix} s-S_n \\ t-T_n \end{pmatrix} \Bigg) P_{\boldsymbol{a}_m}(S_n - s, T_n - t) \\ -N\Big\{ \iint_{\mathbb{R}^2} \boldsymbol{K}\Bigg(\boldsymbol{B}_N^{-1/2} \begin{pmatrix} s-\check{s} \\ t-\check{t} \end{pmatrix} \Bigg) \exp\Big(P_{\boldsymbol{a}_m}(\check{s}-s, t-t) \Big) \mathrm{d}\check{s}\mathrm{d}\check{t} \Big\} \Bigg\}.$$

Also, note that even though it is well-known that kernel estimators suffer from the curse of dimensionality (more in Scott (2008) for example), in our case only two-dimensional functions need to be estimated, thus problems with high-dimensionality are avoided.

Rotated pair copulas

Some pair copulas as the Clayton and Gumbel for example, only allow for positive dependence. To overcome this drawback, also counterclockwise rotated versions of the copula density $c(\cdot,\cdot)$ are considered. These are:

- 90 degrees rotation: $c_{90}(u_1,u_2) := c(1-u_1,u_2)$,
- 180 degrees rotation: $c_{180}(u_1,u_2) := c(1-u_1,1-u_2)$,
- 270 degrees rotation: $c_{270}(u_1,u_2) := c(u_1,1-u_2)$.

2.2. Vine copulas

Joe (1996) has shown that a d-dimensional copula density can be decomposed into $d\,(d-1)\,/2$ bivariate copula densities. However, the decomposition is not unique. A graphical model introduced by Bedford and Cooke (2002) called regular vine copulas (R-vines), organizes all such decompositions that lead to a valid density. Thus, the estimation of any d-dimensional copula density can be divided into the estimation of $d(d-1)/2$ two-dimensional pair copula densities, which can be chosen completely independent of each other.

A regular vine copula consists of a regular vine tree sequence (or tree structure), denoted by $\mathcal{V}$, a set of bivariate copula families (also known as pair copulas) $\mathcal{B}(\mathcal{V})$, and a set of parameters corresponding to the bivariate copula families $\Theta(\mathcal{B}(\mathcal{V}))$. Given d uniformly distributed random variables $U_1, \ldots, U_d$, the vine tree sequence $\mathcal{V}$ consists of a sequence of $d-1$ linked trees, $T_k = (N_k, E_k)$, $k = 1, \ldots, d-1$, satisfying the following conditions:

(i) T_1 is a tree with node set $N_1 = \{U_1, \ldots, U_d\}$ and edge set E_1.

(ii) For $k \geq 2$, T_k is a tree with node set $N_k = E_{k-1}$ and edge set E_k.

(iii) (Proximity condition) For $k \geq 2$, two nodes of the tree T_k can be connected by an edge if the corresponding edges of T_{k-1} have a common node.

The tree sequence uniquely specifies which bivariate (conditional) copula densities occur in the decomposition. Each edge $e \in E_k$ for $k = 1, \ldots, d-1$ is associated with a bivariate copula family $c_{U_{j_e}, U_{k_e}; \mathbf{U}_{D_e}} \in \mathcal{B}(\mathcal{V})$, and a corresponding set of parameters $\boldsymbol{\theta}_{j_e, k_e; D_e} \in \Theta(\mathcal{B}(\mathcal{V}))$. U_{j_e} and U_{k_e} are the conditioned variables and $\mathbf{U}_{D_e}$ represents the conditioning set corresponding to edge e, $\mathbf{U}_{D_e} = (U_i)_{i \in D_e}$. Denote the conditional distribution of $U_{j_e} | \mathbf{U}_{D_e} = \mathbf{u}_{D_e}$ as $C_{U_{j_e}|\mathbf{U}_{D_e}}\left(u_{j_e}|\mathbf{u}_{D_e}\right)$. In a similar manner, $C_{U_{k_e}|\mathbf{U}_{D_e}}\left(u_{k_e}|\mathbf{u}_{D_e}\right)$ is defined. Then $c_{U_{j_e}, U_{k_e}; \mathbf{U}_{D_e}}\left(C_{U_{j_e}|\mathbf{U}_{D_e}}\left(u_{j_e}|\mathbf{u}_{D_e}\right), C_{U_{k_e}|\mathbf{U}_{D_e}}\left(u_{k_e}|\mathbf{u}_{D_e}\right)|\mathbf{u}_{D_e}\right)$ corresponds to the copula density of (U_{j_e}, U_{k_e}) given $\mathbf{U}_{D_e} = \mathbf{u}_{D_e}$ evaluated at $U_{j_e} = u_{j_e}$, $U_{k_e} = u_{k_e}$ and $\mathbf{U}_{D_e} = \mathbf{u}_{D_e}$. The corresponding copula distribution function is denoted as $C_{U_{j_e}, U_{k_e}; \mathbf{U}_{D_e}}$.

However, the pair copulas $c_{U_{j_e}, U_{k_e}; \mathbf{U}_{D_e}}$ dependent on the value of $\mathbf{u}_{D_e}$. This represents the different conditional dependencies between U_{j_e} and U_{k_e} for different conditioning values of $\mathbf{u}_{D_e}$. To allow for computational tractability, it is customary to ignore this influence and simplify it to $c_{U_{j_e}, U_{k_e}; \mathbf{U}_{D_e}}\left(C_{U_{j_e}|\mathbf{U}_{D_e}}\left(u_{j_e}|\mathbf{u}_{D_e}\right), C_{U_{k_e}|\mathbf{U}_{D_e}}\left(u_{k_e}|\mathbf{u}_{D_e}\right)\right)$. This simplification is known as the simplifying assumption (more in Haff et al. (2010), Stoeber et al. (2013), Spanhel and Kurz (2015), and Kurz and Spanhel (2022)).
In this simplified case, we talk about pair copula constructions (PCC) of multivariate densities instead of decompositions. Bedford and Cooke (2002) have shown that regular vines lead to a construction of the joint density using the pair copulas defined via the tree sequence as

$$c(u_1, \ldots, u_d) = \prod_{k=1}^{d-1} \prod_{e \in E_k} c_{U_{j_e}, U_{k_e}; \mathbf{U}_{D_e}}\left(C_{U_{j_e}|\mathbf{U}_{D_e}}\left(u_{j_e}|\mathbf{u}_{D_e}\right), C_{U_{k_e}|\mathbf{U}_{D_e}}\left(u_{k_e}|\mathbf{u}_{D_e}\right)\right). \tag{2.7}$$

To derive the conditional distributions in Equation (2.7), we use the recursion formula from Joe (1996). It defines a recursion for conditional distributions of a regular vine

over its tree sequence. Let $l \in D_e$ and $D_{-l} \coloneqq D_e \setminus \{l\}$. Further, let $h_{U_{j_e}|U_l;\mathbf{U}_{D_{-l}}}(\cdot|\cdot)$ denote the so-called h-function associated with the pair copula $c_{U_{j_e},U_l;\mathbf{U}_{D_{-l}}}$, defined as $h_{U_{j_e}|U_l;\mathbf{U}_{D_{-l}}}(u_{j_e}|u_l) \coloneqq \frac{\partial}{\partial u_l} C_{U_{j_e},U_l;\boldsymbol{U}_{D_{-l}}}(u_{j_e},u_l)$. Then the following recursion is valid

$$C_{U_{j_e}|\mathbf{U}_{D_e}}\left(u_{j_e}|\mathbf{u}_{D_e}\right) = h_{U_{j_e}|U_l;\mathbf{U}_{D_{-l}}}\left(C_{U_{j_e}|\mathbf{U}_{D_{-l}}}\left(u_{j_e}|\mathbf{u}_{D_{-l}}\right)|C_{U_l|\mathbf{u}_{D_{-l}}}\left(u_l|\mathbf{U}_{D_{-l}}\right)\right). \tag{2.8}$$

There are few subclasses of vine copulas that are widely used, as canonical (C-) and drawable (D-) vine copulas (Aas et al., 2009). C-vine copulas are characterised by a sequence of stars (a tree with one node having edges with all the other nodes) and D-vine copulas are characterised by a sequence of paths (a tree with sequence of edges, each one incident to the next). An illustration of a C-vine and a D-vine copula in 4 dimensions is given in Figure 2.1 and 2.2, respectively. More details on vine copula estimation can be found in Czado (2019, Chapter 7).

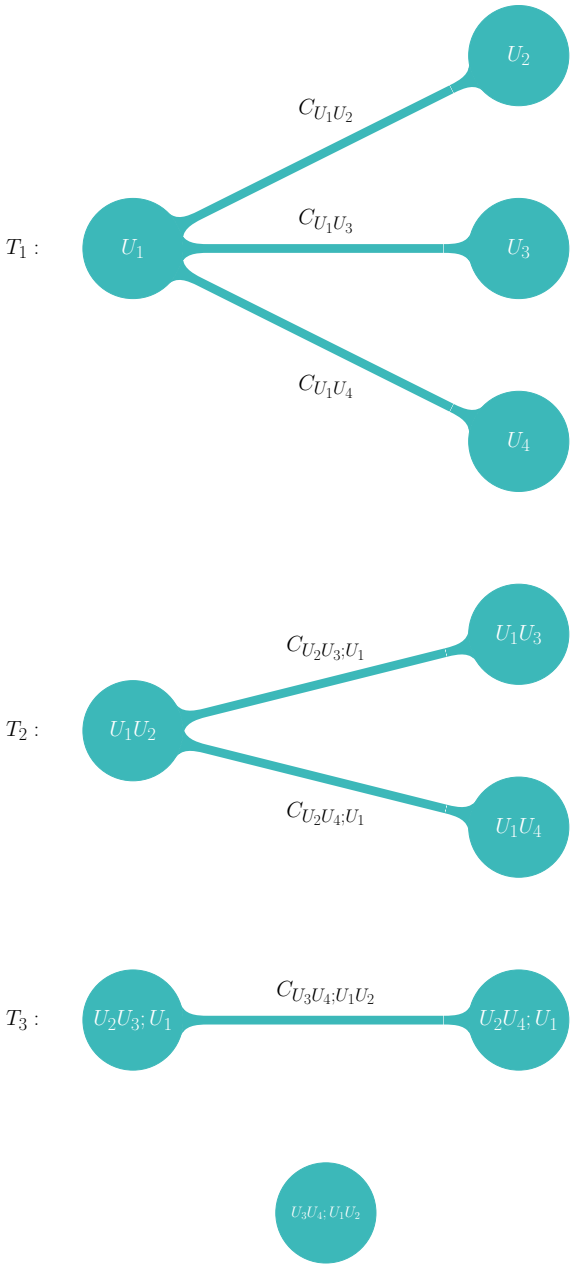


Figure 2.1.: A C-vine copula in 4 dimensions.

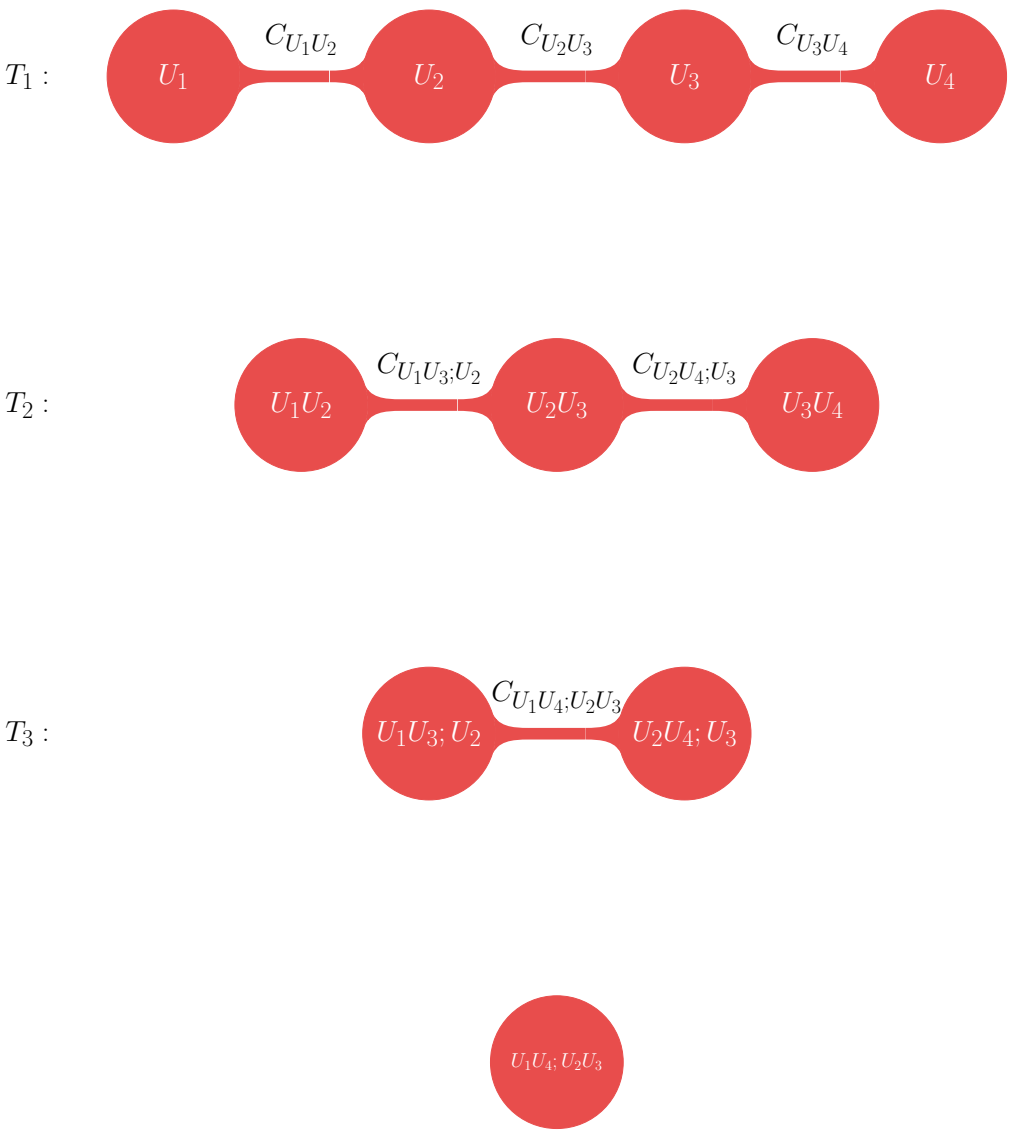


Figure 2.2.: A D-vine copula in 4 dimensions.

3. Univariate response vine copula based regression

Parts of Chapter 3 are very similar to the publication Tepegjozova et al. (2022). However, Sections 3.5.1 and 3.6 contain new material.

3.1. Introduction

The topic of predicting quantiles of a response variable conditioned on a set of predictor variables taking on fixed values, continuously attracts interest. Since the introduction of the linear quantile regression by Koenker and Bassett (1978) many extensions have been developed for the case of a univariate response variable. A short summary of developments in quantile regression modelling is given in Koenker (2017).

The pioneer literature by Koenker (2005) investigated linear quantile regression systematically. It presented properties of the estimators including asymptotic normality and consistency, under various assumptions such as independence of the observations, independent and identically distributed (i.i.d.) errors with continuous distribution, and predictors having bounded second moment. Subsequent extensions of linear quantile regression have been intensively studied, see for example adapting quantile regression in the Bayesian framework (Yu and Moyeed, 2001), for longitudinal data (Koenker, 2004), time-series models (Xiao and Koenker, 2009), high-dimensional models with l_1-regularizer (Belloni and Chernozhukov, 2011), nonparametric estimation by kernel weighted local linear fitting (Yu and M. Jones, 1998), and by additive models (Koenker, 2011; Fenske et al., 2011), etc. The theoretical analysis of the above-mentioned extensions is based on imposing additional assumptions such as samples that are i.i.d. (see for example Yu and M. Jones (1998) and Belloni and Chernozhukov (2011)), or that are generated by a known additive function (see for example Koenker (2011) and Koenker (2004)). Such assumptions, which guarantee the performance of the proposed methods for certain data structures, cause concerns in applications due to the uncertainty of the real-world data structures.

Bernard and Czado (2015) addressed other potential concerns such as quantile crossings and model-misspecification, when the dependence structure of the response variables and the predictors does not follow a Gaussian copula. Flexible models

without assuming homoscedasticity, or a linear relationship between the response and the predictors are of interest. Recent research on dealing with this issue includes quantile forests (Meinshausen, 2006; Hanbo Li and Martin, 2017; Athey et al., 2019) inspired by the earlier work of random forests (Breiman, 2001) and modeling conditional quantiles using copulas (see also Noh et al. (2013), Noh et al. (2015), and Chen et al. (2009)).

One of the most recent approaches for quantile regression are vine copula based quantile regression methods (Kraus and Czado, 2017; Chang and Joe, 2019; Zhu et al., 2021). Copulas allow for separate modelling of the marginal distributions and the dependence structure in the data, while vine copulas allow the multivariate copula to be constructed using bivariate building blocks only, a so-called pair copula construction. This way, a very flexible model, without assuming homoscedasticity, or a linear relationship between the response and the predictors, is constructed. Thus, vine based quantile regression methods overcome two drawbacks of the standard quantile regression methods. First, by construction quantile crossings and collinearity are avoided, and second, there is no need for transformations or interactions of variables Kraus and Czado (2017).

There are several different vine copula tree structures that can be considered, resulting in the most general regular (R-)vine copulas, or its subsets as drawable (D-vines whose tree structure is a sequence of paths) and canonical (C-vines whose tree structure is a sequence of stars) vines. Kraus and Czado (2017) developed a parametric one-step ahead D-vine based quantile regression method by optimizing the conditional log-likelihood and adding predictors until there is no improvement, thus introducing an automatic forward variable selection method. This approach was extended in Tepegjozova (2019) where a C-vine copula based quantile regression was introduced. They also follow the approach to maximize the conditional log likelihood, but introduce an additional step to check for future improvement of the conditional log likelihood, a so called two-step ahead approach. Chang and Joe (2019) introduced an R-vine based quantile regression by first finding the optimal R-vine structure among all predictors and then adding the response variable to each tree in the vine structure as a leaf node. Another R-vine based regression was introduced in Zhu et al. (2021) by optimizing the R-vine structure which gives the largest sum of the absolute value of the partial correlations in each step of the forward extension with predictor variables, while keeping the response as a leaf node. This approach is motivated by the algorithm and results from Zhu et al. (2020). All these selected structures allow to express the conditional density of the response given the predictors without integration.

The main contribution of this chapter is the adaptation of the one-step (Kraus and Czado, 2017) and the two-step ahead forward selection algorithms (Tepegjozova, 2019), from a parametric set up to a fully nonparametric set up, where both the marginal

distributions and the pair copulas are estimated in a nonparametric manner. This approach allows more flexibility than parametric specifications, and Kraus and Czado (2017) addressed the necessity and possible benefit of a nonparametric estimation of bivariate copulas in the quantile regression framework. We implement the D-vine and C-vine one- and two-step ahead algorithms in the nonparametric setting in the R programming language. Also, we extend the two-step ahead algorithm to allow for a variable selection reduction, so that the models can be used when there are many possible predictors available. We present an extensive simulation study in both low and high-dimensional data, which compares the above mentioned algorithms to a benchmark, which is the nonparametric version of the D-vine one step-ahead algorithm by Kraus and Czado (2017).

3.2. Vine based quantile regression

3.2.1. General framework

In the general regression framework the predictive ability of a set of variables $\boldsymbol{X} = (X_1, \dots, X_p)^T$ for the response $Y \in \mathbb{R}$ is studied. The main interest of vine based quantile regression is to predict the $\alpha \in (0,1)$ quantile $q_\alpha(x_1, \dots, x_p) = F^{-1}_{Y|X_1,\dots,X_p}(\alpha|x_1, \dots, x_p)$ of the response variable Y given $\boldsymbol{X}$ by using a copula based model on $(Y, \boldsymbol{X})^T$. As shown in Kraus and Czado (2017), the quantile function q_α can be expressed in terms of a conditional univariate distribution derived from a joint copula as

$$F^{-1}_{Y|X_1,\dots,X_p}(\alpha|x_1, \dots, x_p) = F^{-1}_Y\big(C^{-1}_{V|U_1,\dots,U_p}(\alpha|F_{X_1}(x_1), \dots, F_{X_p}(x_p))\big), \tag{3.1}$$

where $C_{V|U_1,\dots,U_p}$ is the conditional distribution function of $V = F_Y(Y)$ given $U_j = F_{X_j}(X_j) = u_j$ for $j = 1, \dots, p$ with corresponding density $c_{V|U_1,\dots,U_p}$. The conditional distribution function in Equation (3.1) can be obtained from the $(p+1)$-dimensional copula $C_{V,U_1,\dots,U_p}$ associated with the joint distribution of $(Y, \boldsymbol{X})^T$.

In general $C_{V,U_1,\dots,U_p}$ can be any $(p+1)$-dimensional multivariate copula, however only for certain vine structures the corresponding conditional distribution function $C_{V|U_1,\dots,U_p}$ can be obtained in a closed form not requiring numerical integration. For D-vine structures this has been already utilized in Kraus and Czado (2017). Tepegjozova (2019) showed that this is also the case for certain C-vine structures. More precisely, the copula $C_{V,U_1,\dots,U_p}$ with D-vine structure allows to express $C_{V|U_1,\dots,U_p}$ in a closed form if and only if the response V is a leaf node in the first tree of the tree sequence. For a C-vine structure we require, that the node containing the response variable V in the conditioned set, is not a root node in any tree. Additional flexibility in using such D- and C-vine structures can be achieved by allowing for nonparametric pair copulas as building blocks.

The order of the predictors within the tree sequences itself is a free parameter with direct impact on the target function $C_{V|U_1,\dots,U_p}$ and thus, on the corresponding prediction performance of $q_\alpha(x_1,\dots,x_p)$. For this we recall the concept of a node order for C- and D-vine copulas introduced in Tepegjozova (2019). A D-vine copula denoted by $\mathcal{C}_D$ has order $\mathcal{O}_D(\mathcal{C}_D) = (V, U_{i_1},\dots,U_{i_p})$, if the response V is the first node of the first tree T_1 and U_{i_k} is the $(k+1)$-th node of T_1, for $k = 1,\dots,p$. For example, the D-vine copula in Figure 2.2 has an order (U_1, U_2, U_3, U_4). A C-vine copula $\mathcal{C}_C$ has order $\mathcal{O}_C(\mathcal{C}_C) = (V, U_{i_1},\dots,U_{i_p})$, if U_{i_1} is the root node in the first tree T_1, $U_{i_2}U_{i_1}$ is the root node in the second tree T_2, and $U_{i_k}U_{i_{k-1}};U_{i_1},\dots,U_{i_{k-2}}$ is the root node in the k-th tree T_k for $k = 3,\dots,p-1$. For example, the D-vine copula in Figure 2.1 has an order (U_1, U_2, U_3, U_4).

In order to find an optimal order of D- or C-vine copula model a fit measure is required. This measure has to quantify the explanatory power of a model. One such measure is the estimated conditional copula log-likelihood function. For N i.i.d. observations $\boldsymbol{v} \coloneqq (v^{(1)},\dots,v^{(N)})^T$ and $\boldsymbol{u}_j \coloneqq (u_j^{(1)},\dots,u_j^{(N)})^T$, for $j = 1,\dots,p$ of the random vector $(V, U_1,\dots,U_p)^T$ we fit a C- or D-vine copula with order $(V, U_1,\dots,U_p)$. We denote this vine copula by $\hat{\mathcal{C}}$, and then the fitted conditional log-likelihood can be determined as

$$cll(\hat{\mathcal{C}}, \boldsymbol{v}, (\boldsymbol{u}_1,\dots,\boldsymbol{u}_p)) = \sum_{n=1}^{N} \ln \hat{c}_{V|U_1,\dots,U_p}(v^{(n)}|u_1^{(n)},\dots,u_p^{(n)}) = \sum_{n=1}^{N} \Big[\ln \hat{c}_{V,U_1}(v^{(n)}, u_1^{(n)}) + \sum_{j=2}^{p} \ln \hat{c}_{V,U_j|U_1,\dots,U_{j-1}}(\hat{C}_{V|U_1,\dots,U_{j-1}}(v^{(n)}|u_1^{(n)},\dots,u_{j-1}^{(n)}), \hat{C}_{U_j|U_1,\dots,U_{j-1}}(u_j^{(n)}|u_1^{(n)},\dots,u_{j-1}^{(n)})) \Big].$$

Penalizations for model complexity when parametric pair copulas are used, can be added as shown in Tepegjozova (2019). To define the appropriate penalty in the case of using nonparametric pair copulas, is an open research question, which we shortly discuss in Section 3.6.

3.2.2. Nonparametric estimation of marginals and bivariate copulas

One of the simplest nonparametric estimation methods for marginal distributions is the empirical distribution function, but due to its discrete nature and the fact that we need inverses for calculating q_α, we opt against it. Instead we use the univariate local polynomial kernel density estimators. Given a sample $(x_1,\dots,x_N)$ from a random variable X, the univariate local polynomial kernel density estimator is defined as

$$\hat{F}(x) = \frac{1}{Nb} \sum_{i=1}^{N} K\left(\frac{x - x_i}{b}\right), \quad x \in \mathbb{R},$$

where $K(x) := \int_{-\infty}^{x} k(t)\,dt$ with $k(\cdot)$ being a symmetric probability density function and $b > 0$ is a bandwidth parameter. In the following we use the Gaussian kernel defined as

$$K(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2},$$

with the optimal bandwidth parameter b developed in Sheather and M. C. Jones (1991) can be used. The bandwidth b controls the smoothness or complexity of the estimate $\hat{F}(x)$. Thus, b plays the main role in the bias-variance trade off of the nonparametric estimator. This univariate local polynomial kernel density estimator is implemented in the R library `kde1d` (Nagler and Vatter, 2020) and we use it for the marginal distribution estimation.

The nonparametric pair copula densities are estimated using the transformation local likelihood estimator, as discussed in Section 2.1.2. The choice of the bandwidth parameter follows Geenens et al. (2017, Section 4). For the estimation, we use the R package `rvinecopulib` (Nagler and Vatter, 2021).

3.3. Forward selection algorithms

Having a set of p predictors, there are $p!$ different orders that uniquely determine $p!$ C-vines and $p!$ D-vines. Fitting and comparing all of them is computationally inefficient. Thus, the idea is to have an algorithm that will sequentially choose the elements of the order, so that at every step the resulting model for the prediction of the conditional quantiles has the highest conditional log-likelihood. For the C-vine one-step ahead algorithm, we use the forward selection algorithm based on the conditional log-likelihood by Kraus and Czado (2017) and for the D- and C-vine two-step ahead algorithms we use the two-step ahead approach from Tepegjozova (2019). The algorithms build the C- or D-vine sequentially, starting with an order consisting of only the response variable V. Each step adds one of the predictors to the order based on the improvement of the conditional log-likelihood, and the two-step ahead approach takes into account the possibility of future improvement, i.e. extending our view two steps ahead in the order.

We present the implementation for both C-vine and D-vine in a single algorithm, in which the user decides whether to fit a C-vine or D-vine model based on the background knowledge of dependency structures in the data.

Input and data preprocessing: Consider N i.i.d observations $\boldsymbol{y} := (y^{(1)}, \dots, y^{(N)})$ and $\boldsymbol{x}_j := (x_j^{(1)}, \dots, x_j^{(N)})$ for $j = 1, \dots, p$, from the random vector $(Y, X_1, \dots, X_p)^T$. The input data is on the x-scale, but in order to fit bivariate copulas we need to transform it to the u-scale using the probability integral transform. Since the marginal distributions

are unknown we estimate them using a univariate nonparametric kernel density estimator from the R package `kde1d` (Nagler and Vatter, 2020). This results in the pseudo copula data $\hat{v}^{(n)} := \hat{F}_Y(y^{(n)})$ and $\hat{u}_j^{(n)} := \hat{F}_{X_j}(x_j^{(n)})$, for $n = 1, \dots, N, \quad j = 1, \dots, p$. The normalized marginals (z-scale) are defined as $Z_j := \Phi^{-1}(U_j)$ for $j = 1, \dots, p$, and $Z_V := \Phi^{-1}(V)$, where Φ denotes the standard normal distribution function.

3.3.1. One-step ahead algorithm

Step 1: The k candidate predictors and the corresponding candidate index set of step 1 are defined as $U_{q_1}, \dots, U_{q_k}$ and $K_1 = \{q_1, \dots, q_k\}$, respectively. In the first step $k = p$. For all $c \in K_1$ the bivariate copulas $\mathcal{C}^1_{V,U_c}$ are estimated and the predictor that maximizes the log-likelihood of this copula is chosen as the first predictor in the order. The conditional log-likelihood is just the density of this copula in the first step

$$cll\left(\mathcal{C}^1_{V,U_c}, \hat{\boldsymbol{v}}, \hat{\boldsymbol{u}}_c\right) = \sum_{n=1}^{N} \left[\log \hat{c}_{V,U_c}(\hat{v}^{(n)}, \hat{u}_c^{(n)})\right].$$

The maximal conditional log-likelihood at step 1, cll_c^1 is defined as $cll_c^1 := \max_{c \in K_1} cll\left(\mathcal{C}^1_{V,U_c}, \hat{\boldsymbol{v}}, \hat{\boldsymbol{u}}_c\right)$. Based on the maximal conditional log-likelihood at step 1, cll_c^1, the index t_1 is chosen as $t_1 := \arg\max_{c \in K_1} cll_c^1$, and the corresponding candidate predictor U_{t_1} is selected as the first predictor to be added to the order. Figure 3.1 illustrates the first step in the one-step ahead algorithm.

Step r: After $r-1$ steps, the current optimal fit is the C- or D-vine copula $\mathcal{C}_{r-1}$ with order $\mathcal{O}(\mathcal{C}_{r-1}) = (V, U_{t_1}, \dots, U_{t_{r-1}})$. At each previous step i, the order of the current optimal fit is sequentially updated with the predictor U_{t_i} for $i = 1, \dots, r-1$. The remaining $k = p - (r-1)$ candidate predictors and the corresponding candidate index set of step r are defined as $U_{q_1}, \dots, U_{q_k}$ and the set $K_r = \{q_1, \dots, q_k\}$, respectively. For all $c \in K_r$, the vine copulas $\mathcal{C}_c^r$ with order $\mathcal{O}(\mathcal{C}_c^r) = (V, U_{t_1}, \dots, U_{t_{r-1}}, U_c)$ are estimated. Their corresponding conditional log-likelihood functions are given as

$$cll\left(\mathcal{C}_c^r,\ \hat{\boldsymbol{v}}, (\hat{\boldsymbol{u}}_{t_1} \dots \hat{\boldsymbol{u}}_{t_{r-1}}, \hat{\boldsymbol{u}}_c)\right) = cll\left(\mathcal{C}_{r-1}, \hat{\boldsymbol{v}}, (\hat{\boldsymbol{u}}_{t_1} \dots \hat{\boldsymbol{u}}_{t_{r-1}})\right) +$$
$$\sum_{n=1}^{N} \log \hat{c}_{VU_c;U_{t_1},\dots,U_{t_{r-1}}}\left(\hat{C}_{V|U_{t_1},\dots,U_{t_{r-1}}}(\hat{v}^{(n)}|\hat{u}_{t_1}^{(n)}, \dots, \hat{u}_{t_{r-1}}^{(n)}), \hat{C}_{U_c|U_{t_1},\dots,U_{t_{r-1}}}(\hat{u}_c^{(n)}|\hat{u}_{t_1}^{(n)}, \dots, \hat{u}_{t_{r-1}}^{(n)})\right).$$

The r-th predictor is then added to the order based on the maximal conditional log-likelihood at Step r, cll_c^r, defined as

$$cll_c^r := \max_{c \in K_r} cll\left(\mathcal{C}_c^r, \hat{\boldsymbol{v}}, (\hat{\boldsymbol{u}}_{t_1} \dots \hat{\boldsymbol{u}}_{t_{r-1}}, \hat{\boldsymbol{u}}_c,)\right).$$

The index t_r is chosen as $t_r := \arg\max_{c \in K_r} cll_c^r$, and the predictor U_{t_r} is selected as the $r-$th predictor of the order. At this step, the current optimal fit is the C-vine or D-vine

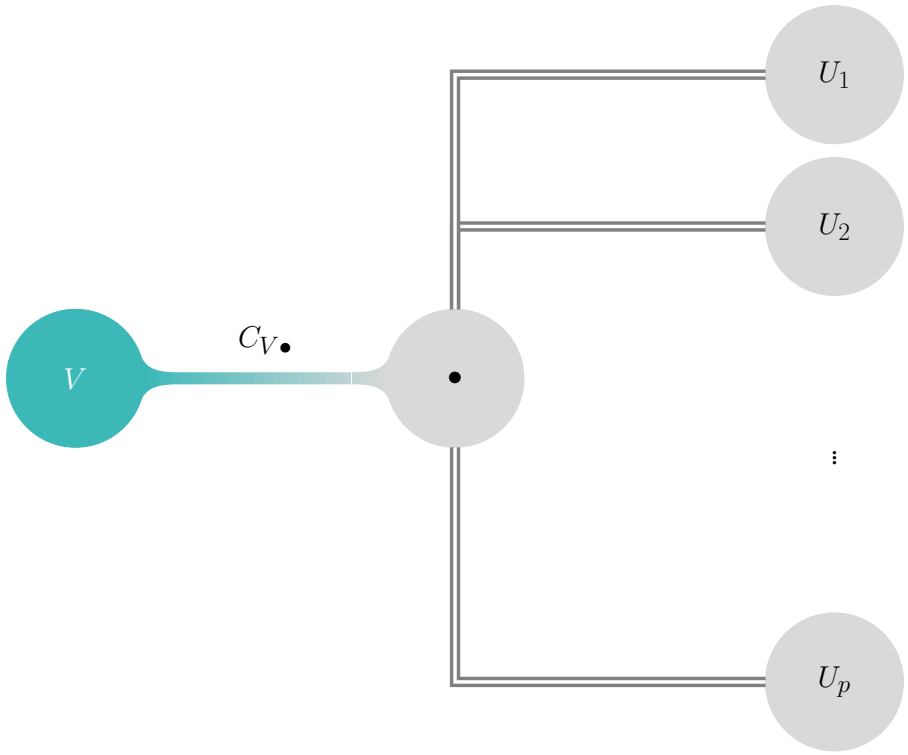


Figure 3.1.: Step 1 for the one-step ahead algorithm.

copula $\mathcal{C}_r$, with order $\mathcal{O}(\mathcal{C}_r) = (V, U_{t_1}, \dots U_{t_r})$. The iterative procedure is repeated until all predictors are included in the order of the C- or D-vine copula model.

3.3.2. Two-step ahead algorithm

Step 1: To reduce computational complexity, we perform a pre-selection of the predictors based on Kendall's τ. This is motivated by the fact that Kendall's τ is rank-based, therefore invariant with respect to monotone transformations of the marginals and can be expressed in terms of pair copulas. Using the pseudo copula data $(\hat{\boldsymbol{v}}, \hat{\boldsymbol{u}}_j) = \{\hat{v}^{(n)}, \hat{u}_j^{(n)} | n = 1, \dots, N\}$, estimates $\hat{\tau}_{VU_j}$ of the Kendall's τ values between the response V, and all possible predictors U_j for $j = 1, \dots, p$, are obtained. For a given $k \leq p$, the k largest estimates of $|\hat{\tau}_{VU_j}|$ are selected and the corresponding indices $q_1, \dots, q_k$ are identified such that $|\hat{\tau}_{VU_{q_1}}| \geq |\hat{\tau}_{VU_{q_2}}| \geq \dots \geq |\hat{\tau}_{VU_{q_k}}| \geq |\hat{\tau}_{VU_{q_{k+1}}}| \geq \dots \geq |\hat{\tau}_{VU_{q_p}}|$. The parameter k is a hyper-parameter, chosen by the user in advance and subject to tuning. To obtain a parsimonious model, we suggest a k corresponding to 5% - 20% of the total number of predictors. The k candidate predictors and the corresponding candidate index set of step 1 are defined as $U_{q_1}, \dots, U_{q_k}$ and $K_1 = \{q_1, \dots, q_k\}$, respectively. For all $c \in K_1$ and $j \in \{1, \dots, p\} \setminus \{c\}$ the candidate two-step ahead C- or D-vine copulas are defined as the 3-dimensional copulas $\mathcal{C}^1_{c,j}$ with order $\mathcal{O}(\mathcal{C}^1_{c,j}) = (V, U_c, U_j)$. The first predictor is added to the order based on the conditional log-likelihood of the candidate two-step ahead C- or D-vine copulas, $\mathcal{C}^1_{c,j}$ given as

$$cll\left(\mathcal{C}^1_{c,j}, \hat{\boldsymbol{v}}, (\hat{\boldsymbol{u}}_c, \hat{\boldsymbol{u}}_j)\right) = \sum_{n=1}^{N} \Big[\log \hat{c}_{V,U_c}(\hat{v}^{(n)}, \hat{u}_c^{(n)}) + \log \hat{c}_{V,U_j|U_c}\big(\hat{h}_{V|U_c}(\hat{v}^{(n)}|\hat{u}_c^{(n)}), \hat{h}_{U_j|U_c}(\hat{u}_j^{(n)}|\hat{u}_c^{(n)})\big)\Big].$$

For each candidate predictor U_c, the maximal two-step ahead conditional log-likelihood at step 1, cll^1_c, is defined as $cll^1_c := \max_{j \in \{1,\dots,p\} \setminus \{c\}} cll\left(\mathcal{C}^1_{c,j}, \hat{\boldsymbol{v}}, (\hat{\boldsymbol{u}}_c, \hat{\boldsymbol{u}}_j)\right),\ \forall c \in K_1$.

Finally, based on the maximal two-step ahead conditional log-likelihood at step 1, cll^1_c, the index t_1 is chosen as $t_1 := \arg\max_{c \in K_1} cll^1_c$, and the corresponding candidate predictor U_{t_1} is selected as the first predictor added to the order. An illustration of the two-step ahead forward selection algorithm is given in Figure 3.2. Comparing it to the first step of the one-step ahead algorithm in Figure 3.1 we can easily see the difference in these two approaches. Finally, the current optimal fit after the first step is the C-vine or D-vine copula, $\mathcal{C}_1$ with order $\mathcal{O}(\mathcal{C}_1) = (V, U_{t_1})$.

Step *r*: After $r-1$ steps, the current optimal fit is the C- or D-vine copula $\mathcal{C}_{r-1}$ with order $\mathcal{O}(\mathcal{C}_{r-1}) = (V, U_{t_1}, \dots, U_{t_{r-1}})$. At each previous step i, the order of the current optimal fit is sequentially updated with the predictor U_{t_i} for $i = 1, \dots, r-1$. At the r-th step the next predictor candidate is to be included. To do so, the set of potential candidates is narrowed based on a partial correlation measure. Defining a partial Kendall's τ is not straightforward and requires the notion of a partial copula, which is the average over the conditional copula given the values of the conditioning values (for example see Gijbels and Matterne (2021) and the references given there). In

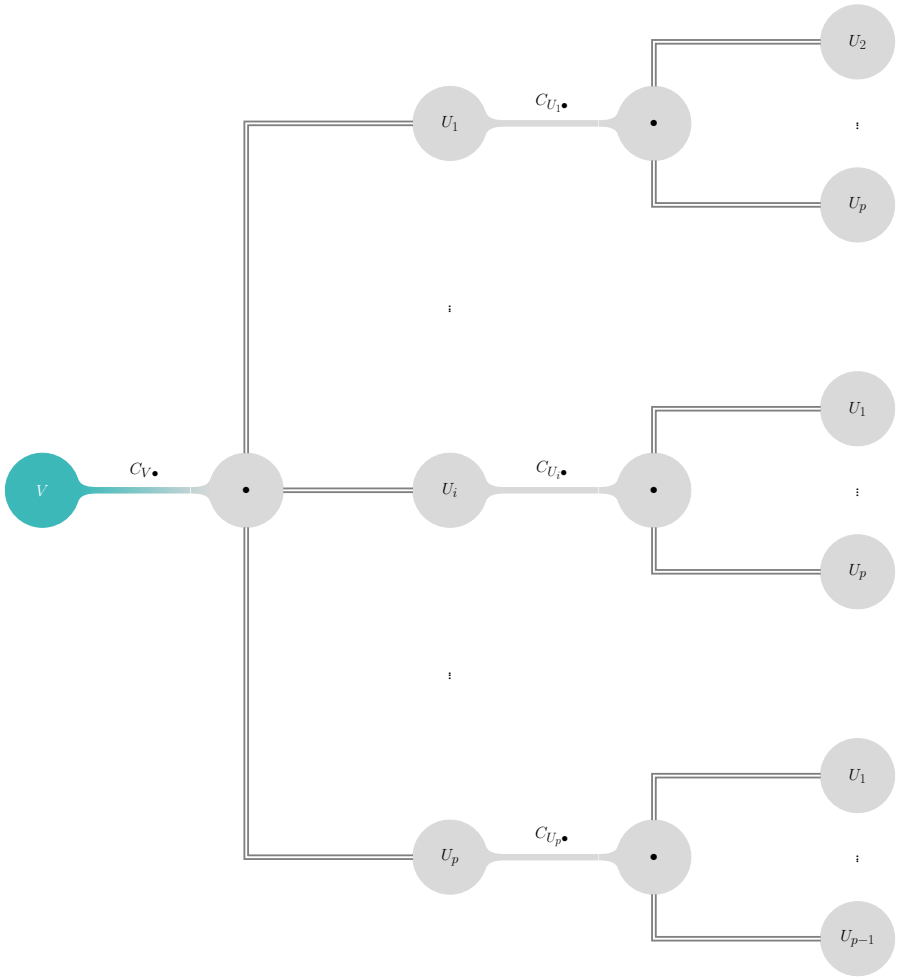


Figure 3.2.: Step 1 for the two-step ahead algorithm.

addition, the computation in the case of multivariate conditioning is very demanding and still an open research problem. Therefore we took a pragmatic view and base our candidate selection on partial correlation. Due to the assumption of Gaussian margins inherited to the Pearson's partial correlation, the estimates are computed on the z-scale. Estimates of the empirical Pearson's partial correlation, $\hat{\rho}_{Z_V,Z_j;Z_{t_1},\ldots,Z_{t_{r-1}}}$, between the normalized response variable $Z_V = \Phi^{-1}(V)$ and available predictors $Z_j = \Phi^{-1}(U_j)$ for $j \in \{1,2,\ldots,p\} \setminus \{t_1,\ldots,t_{r-1}\}$ are obtained. Similar to the first step, a set of candidate predictors of size k is selected based on the largest values of $|\hat{\rho}_{Z_V,Z_j;Z_{t_1},\ldots,Z_{t_{r-1}}}|$ and the corresponding indices $q_1,\ldots,q_k$. The k candidate predictors and the corresponding candidate index set of step r are defined as $U_{q_1},\ldots,U_{q_k}$ and the set $K_r = \{q_1,\ldots,q_k\}$, respectively. For all $c \in K_r$ and $j \in \{1,2,\ldots,p\} \setminus \{t_1,\ldots,t_{r-1},c\}$ the candidate two-step ahead C- or D-vine copulas are defined as the copulas $\mathcal{C}^r_{c,j}$ with order $\mathcal{O}(\mathcal{C}^r_{c,j}) = (V,U_{t_1},\ldots,U_{t_{r-1}},U_c,U_j)$. There are $k(p-r)$ different candidate two-step ahead C- or D-vine copulas $\mathcal{C}^r_{c,j}$ (since we have k candidates for the one-step ahead extension U_c, and for each, $p-(r-1)-1$ two step ahead extensions U_j). Their corresponding conditional log-likelihood functions are given as

$$
\begin{aligned}
&cll\left(\mathcal{C}^r_{c,j},\ \hat{\boldsymbol{v}},(\hat{\boldsymbol{u}}_{t_1}\ldots\hat{\boldsymbol{u}}_{t_{r-1}},\hat{\boldsymbol{u}}_c,\hat{\boldsymbol{u}}_j)\right) = \ cll\left(\mathcal{C}_{r-1},\hat{\boldsymbol{v}},(\hat{\boldsymbol{u}}_{t_1}\ldots\hat{\boldsymbol{u}}_{t_{r-1}})\right) + \\
&\sum_{n=1}^{N}\log\hat{c}_{VU_c;U_{t_1},\ldots,U_{t_{r-1}}}\left(\hat{C}_{V|U_{t_1},\ldots,U_{t_{r-1}}}(\hat{v}^{(n)}|\hat{u}^{(n)}_{t_1},\ldots,\hat{u}^{(n)}_{t_{r-1}}),\hat{C}_{U_c|U_{t_1},\ldots,U_{t_{r-1}}}(\hat{u}^{(n)}_c|\hat{u}^{(n)}_{t_1},\ldots,\hat{u}^{(n)}_{t_{r-1}})\right) \\
&+\sum_{n=1}^{N}\log\hat{c}_{VU_j;U_{t_1},\ldots,U_{t_{r-1}},U_c}\left(\hat{C}_{V|U_{t_1},\ldots,U_{t_{r-1}},U_c}(\hat{v}^{(n)}|\hat{u}^{(n)}_{t_1},\ldots,\hat{u}^{(n)}_{t_{r-1}},\hat{u}^{(n)}_c),\right. \\
&\qquad\left.\hat{C}_{U_j|U_{t_1},\ldots,U_{t_{r-1}},U_c}(\hat{u}^{(n)}_j|\hat{u}^{(n)}_{t_1},\ldots,\hat{u}^{(n)}_{t_{r-1}},\hat{u}^{(n)}_c)\right).
\end{aligned}
$$

The r-th predictor is then added to the order based on the maximal two-step ahead conditional log-likelihood at Step r, cll^r_c, defined as

$$
cll^r_c := \max_{j\in\{1,2,\ldots,p\}\setminus\{t_1,\ldots,t_{r-1},c\}} cll\left(\mathcal{C}^r_{c,j},\hat{\boldsymbol{v}},(\hat{\boldsymbol{u}}_{t_1}\ldots\hat{\boldsymbol{u}}_{t_{r-1}},\hat{\boldsymbol{u}}_c,\hat{\boldsymbol{u}}_j)\right),\ \forall c\in K_r. \tag{3.2}
$$

The index t_r is chosen as $t_r := \arg\max_{c\in K_r} cll^r_c$, and the predictor U_{t_r} is selected as the $r-$th predictor of the order. At this step, the current optimal fit is the C-vine or D-vine copula $\mathcal{C}_r$, with order $\mathcal{O}(\mathcal{C}_r) = (V,U_{t_1},\ldots U_{t_r})$. The iterative procedure is repeated until all predictors are included in the order of the C- or D-vine copula model.

Additional variable reduction in higher dimensions

In order to be able to decrease computational intensity, a variable reduction is possible. The two-step ahead algorithm requires calculating $p-r$ conditional log-likelihoods for each candidate predictor at a given step r. This leads to calculating a total of

$(p-r)k$ conditional log-likelihoods, where k is the number of candidates. For p large, this procedure would cause a heavy computational burden. Hence, the solution is to reduce the number of conditional log-likelihoods calculated for each candidate predictor. This is achieved by reducing the size of the set, over which the maximal two-step ahead conditional log-likelihood cll_c^r in (3.2), is computed. Instead of over the set $\{1, 2, \ldots, p\} \setminus \{t_1, \ldots, t_{r-1}, c\}$, the maximum can be taken over an appropriate subset. This subset can be then chosen either based on the largest Pearson's partial correlations in absolute value denoted as $|\hat{\rho}_{Z_V, Z_j; Z_{t_1}, \ldots, Z_{t_{r-1}}, Z_c}|$, by random selection, or a combination of the two. The selection method and the size of reduction are user-decided.

Implementation

The two-step ahead forward selection algorithms for C- and D-vine based quantile regression, from Section 3.3.2, and the additional variable selection we implement in the statistical language R (R Core Team, 2020). We also implement the C-vine one-step ahead algorithm in the statistical language R, while the D-vine one-step ahead algorithm is already implemented and available in the R package `vinereg` (Nagler, 2022).

3.4. Simulation study

In the simulation study from Kraus and Czado (2017), it is shown that the D-vine one-step ahead forward selection algorithm performs better or similar, compared to other state of the art quantile methods, such as boosting additive quantile regression (Koenker, 2005; Fenske et al., 2011), nonparametric quantile regression (Q. Li et al., 2013), semi-parametric quantile regression (Noh et al., 2015), and the linear quantile regression (Koenker and Bassett, 1978). Thus we use the one-step ahead algorithm as the benchmark competitive method in the simulation study.

We set up the following simulation settings given below. Each setting is replicated for $R = 100$ times. In each simulation replication, we randomly generate N_{train} samples used for fitting the appropriate nonparametric vine based quantile regression models. Additionally, another $N_{eval} = \frac{1}{2} N_{train}$ samples for Settings (a) – (f) and $N_{eval} = N_{train}$ for Settings (g) and (h) are generated for predicting conditional quantiles from the models. Settings (a) – (f) are designed to test quantile prediction accuracy of nonparametric C- or D-vine quantile regression in cases where $p \leq N$; hence, we set $N_{train} = 1000$ or 300. Settings (g) and (h) test quantile prediction accuracy in cases where $p > N$; hence, we set $N_{train} = 100$.

(a) Simulation Setting M5 from Kraus and Czado (2017):

$$Y = \sqrt{|2X_1 - X_2 + 0.5|} + (-0.5X_3 + 1)(0.1X_4^3) + \sigma\varepsilon,$$

with $\varepsilon \sim N(0,1), \sigma \in \{0.1, 1\}$, $(X_1, X_2, X_3, X_4)^T \sim N_4(0, \Sigma)$, and the (i,j)th component of the covariance matrix given as $(\Sigma)_{i,j} = 0.5^{|i-j|}$.

(b) $(Y, X_1, \ldots, X_5)^T$ follows a mixture of two 6-dimensional t copulas with degrees of freedom equal to 3 and mixture probabilities 0.3 and 0.7. Association matrices R_1, R_2 and marginal distributions are recorded in Table 3.1.

$$R_1 = \begin{pmatrix} 1 & 0.6 & 0.5 & 0.6 & 0.7 & 0.1 \\ 0.6 & 1 & 0.5 & 0.5 & 0.5 & 0.5 \\ 0.5 & 0.5 & 1 & 0.5 & 0.5 & 0.5 \\ 0.6 & 0.5 & 0.5 & 1 & 0.5 & 0.5 \\ 0.7 & 0.5 & 0.5 & 0.5 & 1 & 0.5 \\ 0.1 & 0.5 & 0.5 & 0.5 & 0.5 & 1 \end{pmatrix} \quad R_2 = \begin{pmatrix} 1 & -0.3 & -0.5 & -0.4 & -0.5 & -0.1 \\ -0.3 & 1 & 0.5 & 0.5 & 0.5 & 0.5 \\ -0.5 & 0.5 & 1 & 0.5 & 0.5 & 0.5 \\ -0.4 & 0.5 & 0.5 & 1 & 0.5 & 0.5 \\ -0.5 & 0.5 & 0.5 & 0.5 & 1 & 0.5 \\ -0.1 & 0.5 & 0.5 & 0.5 & 0.5 & 1 \end{pmatrix}$$

Y	X_1	X_2	X_3	X_4	X_5
$N(0,1)$	t_4	$N(1,4)$	t_4	$N(1,4)$	t_4

Table 3.1.: Association matrices of the multivariate t-copula and marginal distributions for Setting (b).

.

(c) Linear and heteroscedastic (Chang and Joe, 2019):
$Y = 5(X_1 + X_2 + X_3 + X_4) + 10(U_1 + U_2 + U_3 + U_4)\varepsilon$, where $(X_1, X_2, X_3, X_4)^T \sim N(0, \Sigma)$, $\Sigma_{i,j} = 0.5^{I\{i \neq j\}}$, $\varepsilon \sim N_4(0, 0.5)$, and U_j, $j = 1, \ldots, 4$ are obtained from the X_j's by the probability integral transform.

(d) Nonlinear and heteroscedastic (Chang and Joe, 2019):
$Y = U_1 U_2 e^{1.8U_3U_4} + 0.5(U_1 + U_2 + U_3 + U_4)\varepsilon$, where $U_j, j = 1, \ldots, 4$ are probability integral transformed from $N_4(0, \Sigma)$, $\Sigma_{i,j} = 0.5^{I\{i \neq j\}}$, and $\varepsilon \sim N(0, 0.5)$.

(e) R-vine copula (Czado, 2019): $(V, U_1, \ldots, U_4)^T$ follows an R-vine distribution with pair copulas given in Table 3.2.

(f) D-vine copula (Tepegjozova, 2019): $(V, U_1, \ldots, U_5)^T$ follows a D-vine distribution with pair copulas given in Table 3.3.

(g) Similar to Setting (a),

$$Y = \sqrt{|2X_1 - X_2 + 0.5|} + (-0.5X_3 + 1)(0.1X_4^3) + (X_5, \ldots, X_{110})(0, \ldots, 0)^T + \sigma\varepsilon,$$

Tree	Edge	Conditioned	;	Conditioning	Family	Parameter	Kendall's τ
1	1	U_1, U_3	;		Gumbel	3.9	0.74
1	2	U_2, U_3	;		Gauss	0.9	0.71
1	3	V, U_3	;		Gauss	0.5	0.33
1	4	V, U_4	;		Clayton	4.8	0.71
2	1	V, U_1	;	U_3	Gumbel(90)	6.5	-0.85
2	2	V, U_2	;	U_3	Gumbel(90)	2.6	-0.62
2	3	U_3, U_4	;	V	Gumbel	1.9	0.48
3	1	U_1, U_2	;	V, U_3	Clayton	0.9	0.31
3	2	U_2, U_4	;	V, U_3	Clayton(90)	5.1	-0.72
4	1	U_1, U_4	;	V, U_2, U_3	Gauss	0.2	0.13

Table 3.2.: Pair copulas of the R-vine C_{V,U_1,U_2,U_3,U_4}, with their family parameter (rotation) and Kendall's τ for Setting (e).

where $(X_1, \ldots, X_{110})^T \sim N_{110}(0, \Sigma)$ with the (i,j)th component of the covariance matrix $(\Sigma)_{i,j} = 0.5^{|i-j|}$, $\varepsilon \sim N(0,1)$, and $\sigma \in \{0.1, 1\}$.

(h) Similar to (g),
$Y = (X_1^3, \ldots, X_{110}^3)\boldsymbol{\beta} + \varepsilon$, where $(X_1, \ldots, X_{10})^T \sim N_{10}(0, \Sigma_A)$ with the (i,j)th component of the covariance matrix $(\Sigma_A)_{i,j} = 0.8^{|i-j|}$, $(X_{11}, \ldots, X_{110})^T \sim N_{100}(0, \Sigma_B)$ with $(\Sigma_B)_{i,j} = 0.4^{|i-j|}$. The first 10 entries of $\boldsymbol{\beta}$ are a descending sequence between $(2, 1.1)$ with increment of 0.1 respectively, and the rest are equal to 0. We assume $\varepsilon \sim N(0, \sigma)$ and $\sigma \in \{0.1, 1\}$.

Since the true regression quantiles are difficult to obtain in most settings, we consider the averaged check loss (Komunjer, 2013) and the interval score (Chang and Joe, 2019; Gneiting and Raftery, 2007), instead of the out-of-sample mean averaged square error used in Kraus and Czado (2017), to evaluate the performance of the estimation methods. For a chosen $\alpha \in (0,1)$, the averaged check loss is defined as

$$\widehat{\mathrm{CL}}_\alpha = \frac{1}{R}\sum_{r=1}^{R}\left\{\frac{1}{N_{eval}}\sum_{n=1}^{N_{eval}}\left\{\gamma_\alpha\left(Y_{r,n}^{eval} - \hat{q}_\alpha(X_{r,n}^{eval})\right)\right\}\right\}, \tag{3.3}$$

where γ_α is the check loss function.

Tree	Edge	Conditioned	;	Conditioning	Family	Parameter	Kendall's τ
1	1	V, U_1	;		Clayton	3.00	0.60
1	2	U_1, U_2	;		Joe	8.77	0.80
1	3	U_2, U_3	;		Gumbel	2.00	0.50
1	4	U_3, U_4	;		Gauss	0.20	0.13
1	5	U_4, U_5	;		Indep.	0.00	0.00
2	1	V, U_2	;	U_1	Gumbel	5.00	0.80
2	2	U_1, U_3	;	U_2	Frank	9.44	0.65
2	3	U_2, U_4	;	U_3	Joe	2.78	0.49
2	4	U_3, U_5	;	U_4	Gauss	0.20	0.13
3	1	V, U_3	;	U_1, U_2	Joe	3.83	0.60
3	2	U_1, U_4	;	U_2, U_3	Frank	6.73	0.55
3	3	U_2, U_5	;	U_3, U_4	Gauss	0.29	0.19
4	1	V, U_4	;	U_1, U_2, U_3	Clayton	2.00	0.50
4	2	U_1, U_5	;	U_2, U_3, U_4	Gauss	0.09	0.06
5	1	V, U_5	;	U_1, U_2, U_3, U_4	Indep.	0.00	0.00

Table 3.3.: Pair copulas of the D-vine $C_{V,U_1,U_2,U_3,U_4,U_5}$, with their family parameter and Kendall's τ for Setting (f).

The interval score, for the $(1-\alpha)\times 100\%$ prediction interval, is defined as

$$\begin{aligned}\widehat{\mathrm{IS}}_\alpha = \frac{1}{R}\sum_{r=1}^{R}\Bigg\{\frac{1}{N_{eval}}\sum_{n=1}^{N_{eval}}\Big\{&\big(\hat{q}_{\alpha/2}(X_{r,n}^{eval}) - \hat{q}_{1-\alpha/2}(X_{r,n}^{eval})\big) \\ &+\frac{2}{\alpha}\big(\hat{q}_{1-\alpha/2}(X_{r,n}^{eval}) - Y_{r,n}^{eval}\big) I\{Y_{r,n}^{eval} \le \hat{q}_{1-\alpha/2}(X_{r,n}^{eval})\} \\ &+\frac{2}{\alpha}\big(Y_{r,n}^{eval} - \hat{q}_{\alpha/2}(X_{r,n}^{eval})\big) I\{Y_{r,n}^{eval} > \hat{q}_{\alpha/2}(X_{r,n}^{eval})\}\Big\}\Bigg\},\end{aligned} \tag{3.4}$$

and smaller interval scores are better.

Setting	Model	$\widehat{\text{IS}}_{0.05}$	$\widehat{\text{CL}}_{0.05}$	$\widehat{\text{CL}}_{0.5}$	$\widehat{\text{CL}}_{0.95}$	$\widehat{\text{IS}}_{0.05}$	$\widehat{\text{CL}}_{0.05}$	$\widehat{\text{CL}}_{0.5}$	$\widehat{\text{CL}}_{0.95}$
		$N_{train} = 300$				$N_{train} = 1000$			
(a) $\sigma = 0.1$ **	D-vine One-step	55.54	0.66	0.16	0.51	55.89	0.67	0.15	0.50
	D-vine Two-step	43.33	0.47	0.10	0.41	40.74	0.45	0.09	0.37
	C-vine One-step	53.51	0.64	0.16	0.49	54.52	0.66	0.15	0.49
	C-vine Two-step	42.01	0.45	0.10	0.40	40.04	0.44	0.09	0.37
(a) $\sigma = 1$ **	D-vine One-step	154.35	1.63	0.45	1.62	162.12	1.70	0.43	1.66
	D-vine Two-step	148.53	1.57	0.45	1.56	156.77	1.63	0.42	1.62
	C-vine One-step	151.60	1.61	0.45	1.60	160.78	1.68	0.43	1.65
	C-vine Two-step	148.41	1.56	0.45	1.56	156.79	1.63	0.42	1.62
(b) *	D-vine One-step	118.75	1.29	0.42	1.30	125.33	1.37	0.40	1.36
	D-vine Two-step	119.10	1.30	0.42	1.30	125.24	1.36	0.40	1.36
	C-vine One-step	119.08	1.30	0.41	1.30	125.12	1.36	0.40	1.36
	C-vine Two-step	118.90	1.30	0.42	1.30	125.30	1.36	0.40	1.36
(c) **	D-vine One-step	2908.90	30.54	8.55	30.42	3064.78	31.69	8.15	31.47
	D-vine Two-step	2853.52	30.21	8.70	29.95	3041.95	31.61	8.20	31.26
	C-vine One-step	2859.23	30.24	8.59	29.95	3046.52	31.64	8.18	31.25
	C-vine Two-step	2850.10	30.19	8.64	29.84	3042.46	31.62	8.20	31.23
(d) **	D-vine One-step	86.40	0.92	0.24	0.91	91.11	0.96	0.22	0.95
	D-vine Two-step	83.54	0.90	0.24	0.88	89.56	0.96	0.22	0.92
	C-vine One-step	84.99	0.91	0.24	0.90	90.40	0.96	0.22	0.94
	C-vine Two-step	83.33	0.90	0.24	0.87	89.47	0.96	0.22	0.92
(e) *	D-vine One-step	10.59	0.11	0.03	0.11	10.49	0.11	0.03	0.11
	D-vine Two-step	10.32	0.10	0.03	0.11	10.26	0.09	0.02	0.11
	C-vine One-step	10.23	0.11	0.03	0.10	10.02	0.10	0.02	0.10
	C-vine Two-step	10.35	0.10	0.03	0.11	10.33	0.10	0.02	0.11
(f) **	D-vine One-step	13.79	0.16	0.04	0.14	13.70	0.16	0.04	0.14
	D-vine Two-step	8.44	0.09	0.02	0.08	8.28	0.09	0.02	0.08
	C-vine One-step	12.62	0.14	0.04	0.13	12.23	0.13	0.04	0.13
	C-vine Two-step	9.09	0.10	0.02	0.09	8.93	0.09	0.02	0.08

Table 3.4.: Out-of-sample predictions $\widehat{\text{IS}}_{0.5}$, $\widehat{\text{CL}}_{0.05}$, $\widehat{\text{CL}}_{0.5}$, $\widehat{\text{CL}}_{0.95}$ for Settings (a) – (f) with $N_{train} = 300$ and $N_{train} = 1000$. Lower values, indicating better performance, are highlighted in gray. With ** we denote the scenarios in which there is an improvement through the second step and with * we denote scenarios in which the models perform similar.

Model	$\widehat{\text{IS}}_{0.05}$	$\widehat{\text{CL}}_{0.05}$	$\widehat{\text{CL}}_{0.5}$	$\widehat{\text{CL}}_{0.95}$	$\widehat{\text{IS}}_{0.05}$	$\widehat{\text{CL}}_{0.05}$	$\widehat{\text{CL}}_{0.5}$	$\widehat{\text{CL}}_{0.95}$
	(g), $\sigma = 0.1$ *				(g), $\sigma = 1$ **			
D-vine One-step	19.63	0.26	0.25	0.23	53.38	0.69	0.67	0.65
D-vine Two-step	20.48	0.26	0.26	0.25	52.17	0.68	0.65	0.63
C-vine One-step	19.73	0.25	0.25	0.24	53.62	0.69	0.67	0.65
C-vine Two-step	19.79	0.25	0.25	0.25	52.35	0.67	0.65	0.64
	(h), $\sigma = 0.1$ **				(h), $\sigma = 1$ **			
D-vine One-step	558.36	6.92	6.98	7.04	554.18	6.87	6.93	6.99
D-vine Two-step	529.51	6.46	6.62	6.78	531.30	6.64	6.64	6.64
C-vine One-step	514.08	6.05	6.43	6.81	512.96	6.39	6.41	6.44
C-vine Two-step	479.66	5.87	6.00	6.12	483.92	6.05	6.05	6.05

Table 3.5.: Out-of-sample predictions $\widehat{\text{IS}}_{0.5}$, $\widehat{\text{CL}}_{0.05}$, $\widehat{\text{CL}}_{0.5}$, $\widehat{\text{CL}}_{0.95}$ for Settings (g) – (h) with $N_{train} = 100$. Lower values, indicating better performance, are highlighted in gray. With ** we denote the scenarios in which there is an improvement through the second step and with * we denote scenarios in which the models perform similar.

For Settings (a) – (f), the estimation procedure for the two-step ahead C- or D-vine quantile regression follows exactly Section 3.3.2 where the candidate sets at each step include all possible remaining predictors. The additional variable reduction described in Section 3.3.2 is not applied; thus, we calculate all possible conditional log-likelihoods in each step. On the contrary, due to computational burden in Settings (g) and (h), we set the number of candidates to be $k = 5$ and the additional variable reduction from Section 3.3.2 is applied. The chosen subset contains 20% of all possible choices, where 10% are predictors having the highest Pearson's partial correlation with the response and the remaining 10% are chosen randomly from the remaining predictors. Performance of the C- and D-vine two-step ahead quantile regression is compared with the C- and D-vine one-step ahead quantile regression. The performance of the competitive methods, evaluated by the averaged check loss at 5%, 50%, 95% quantile levels and interval score for the 95% prediction interval, are recorded in Tables 3.4 and 3.5. All densities are estimated nonparametrically for a fair comparison. Table 3.4 shows that the C- and D-vine two-step ahead regression models outperform the C- and D-vine one-step ahead regression models in five out of seven settings, except Settings (b) and (e), in which all models perform quite similarly to each other. Again, when comparing regression models within the same vine copula class, the C-vine two-step ahead regression models outperform the C-vine one-step ahead models in five out of seven settings. Similarly, the D-vine two-step ahead models outperform the D-vine one-step ahead models in six out of seven scenarios, except Setting (b) only.

In scenarios where there is no significant improvement through the second step, both one-step and two-step ahead approaches perform very similar. All of that implies that the two-step ahead vine based quantile regression improves the performance of the one-step ahead quantile regression. Table 3.5 indicates that in the high-dimensional settings, where the two-step ahead quantile regression was used in combination with the additional variable selection from Section 3.3.2, in three out of four simulation settings, the two-step ahead models outperform the one-step ahead models. In Setting (g), we can see that all models show similar performance. In Setting (g) with standard deviation $\sigma = 0.1$, the D-vine one-step ahead model outperforms the other models, while in Setting (g) with $\sigma = 1$, the D-vine two-step ahead model shows a better performance. In Setting (h), we see a significant improvement in the two-step ahead models compared to the one-step ahead models. For both $\sigma = 0.1$ and $\sigma = 1$, the best performing model is the C-vine two-step ahead model. These results indicate that the two-step method improves the accuracy of the one-step ahead quantile regression in high dimensions, even with an attempt to ease the computational complexity of the two-step ahead model with a low number of candidates, compared to the number of predictors.

The proposed two-step algorithms, as compared to the one-step algorithms are computationally more intensive. We present the averaged computation time over $R = 100$ replications on 100 paralleled cores (Xeon Gold 6140 CPUSs@2.6 GHz) in Settings (g), (h) where $p > N_{train}$, for the one step ahead and the two-step ahead approach. The high-dimensional settings have similar computational times since the computational intensity depends on the number of pair copula estimations and the number of candidates, which are the same for Settings (g), (h). Hence, we only report the averaged computational times for Settings (g), (h). The average computation time in minutes for the one-step ahead (C- and D-vine) approach is 83.01, in contrast to 200.28 by the two-step ahead (C- and D-vine) approach. With the variable reduction from Section 3.3.2, the two-step algorithms double the time consumption of the one-step algorithms in exchange for prediction accuracy.

3.5. Data application

We test the proposed methods on two real data sets, i.e., the Concrete data set from Yeh (1998) corresponding to $p \leq N$, and the Riboflavin data set from Bühlmann and Geer (2011) corresponding to $p > N$. For both, performance of the four competitive algorithms is evaluated by the averaged check loss defined in (3.3) at 5%, 50% and 95% quantile levels, and the 95% prediction interval score defined in (3.4), by randomly splitting the data set into training and evaluation sets 100 times.

3.5.1. Concrete data set

The concrete dataset was originaly used in Yeh (1998) , and is available at the UCI Machine Learning Repository (Dua and Graff, 2017). The dataset has in total 1030 samples. Our objective is quantile predictions of the concrete compressive strength, which is a highly nonlinear function of age and ingredients. The predictors are age (`AgeDay`, counted in days) and 7 physical measurements of the concrete ingredients (given in kg in a m^3 mixture): cement (`CementComp`), blast furnace slag (`BlastFur`), fly ash (`FlyAsh`), water (`WaterComp`), superplastizer (`Superplastizer`), coarse aggregate (`CoarseAggre`) and fine aggregate (`FineAggre`).

We randomly split the dataset into training set with 830 samples and evaluation set with 200 samples; the random splitting is repeated for 100 times. Performance of the proposed C- and D-vine two-step ahead quantile regression, compared with the C- and D-vine one-step ahead quantile regression, is evaluated by several measurements reported in Table 3.6 after 100 repetitions of fitting the models. The results are quite close to each other, but given the small number of predictors, that is what one would expect of the forward sequential algorithm. However, there is an improvement in the performance of the two-step ahead approach compared to the one-step ahead approach for both C- and D-vine based models. Also, the C-vine model seems more appropriate for modelling the dependency structure in the dataset, because the C-vine based models show better results. Finally, out of all models, the C-vine two-step ahead algorithm is the best performing algorithm in terms of out-of-sample predictions $\widehat{\text{IS}}_{0.5}$, $\widehat{\text{CL}}_{0.05}$, $\widehat{\text{CL}}_{0.5}$, $\widehat{\text{CL}}_{0.95}$ on the concrete dataset, as seen in Table 3.6 .

Model	$\widehat{\text{IS}}_{0.05}$	$\widehat{\text{CL}}_{0.05}$	$\widehat{\text{CL}}_{0.5}$	$\widehat{\text{CL}}_{0.95}$
D-vine One-step	1032.32	10.75	2.76	10.52
D-vine Two-step	987.10	10.54	2.78	9.82
C-vine One-step	976.75	10.65	2.70	9.45
C-vine Two-step	967.00	10.52	2.64	9.45

Table 3.6.: Concrete data set: Out-of-sample predictions $\widehat{\text{IS}}_{0.5}$, $\widehat{\text{CL}}_{0.05}$, $\widehat{\text{CL}}_{0.5}$, $\widehat{\text{CL}}_{0.95}$. The best performing model is highlighted in gray.

Order analysis

In order to explain the different approaches of the one-step ahead and the two-step ahead, we consider the order of the predictors which the algorithms provide. The order of the predictors that enter the model is based on maximising the conditional log likelihood, thus it provides a descending order of influence of the predictors on the

conditional quantile function of the response.

Figure 3.3 shows the individual distribution of positions for each predictor in the four different models, i.e. each individual bar plot has all the possible positions in the order on the x-axis and the counts of how many times the given predictor appeared on a specific position of the order on the y-axis out of the 100 repeated model fits.
From Figure 3.3 we can indeed see the greedy approach of the one-step algorithms. Both one-step ahead C- and D- vine models always choose the same predictor as the first predictor to enter the model. That is because the pair copula between the response and the predictor `AgeDay` has the biggest likelihood, out of the possible pair copula between the response and each of the predictors. Next, as second predictor both one-step ahead algorithms always choose the predictor `CementComp`, because, similarly as before, the bivariate copula between the response and `CementComp` conditioned on the already chosen `AgeDay` has the biggest likelihood out of the possible pair copula between the response and the other possible predictors conditioned on `AgeDay`. Note that in only 3 dimensions the models for both C- and D-vines are equivalent (in 3 dimensions a path is also a star, and vice versa). On other side, the two-step ahead approaches do not make such a uniform decision about the first predictor to be included. Instead of choosing `AgeDay` as a first predictor, the algorithms consider the future possible improvement and based on that, they choose the predictor `CementComp` as the first to enter the model in 70 cases out of 100. The most influential predictor from the one-step ahead models, the `AgeDay`, is chosen to be second or third predictor of the two-step ahead models, more precisely is it second in 24 cases and third in 37 for the C-vine two-step ahead, and it is second in 41 cases and third in 27 for the D-vine two-step ahead, which turns out to be better ordering.

Next, we look at the so-called optimal orders of each algorithm. The optimal order is defined as the order in which the first element corresponds to the predictor that appeared the most in the first place over the 100 iteration, then the second element is defined as the element that appeared the most in the second position among the elements not chosen as first and so on. Figure 3.4 shows the optimal orders of the four different algorithms. Each plot shows the predictors on the x-axis in the optimal order and the corresponding counts of each predictor.

There is almost no difference in the orders of the D-vine one-step ahead algorithm and the C-vine one-step ahead algorithm. The first five predictors, i.e. `AgeDay`, `CementComp`, `WaterComp`, `BlastFur` and `FlyAsh`, have almost identical distributions of order position in both algorithms. Further, the optimal orders of the one-step ahead algorithms coincide in the first five places. The other three predictors do differ in the distribution. On the other hand the two-step ahead algorithms show more difference in the distributions of all predictors. The optimal orders of the two-step ahead algorithms do coincide on the first, fourth and fifth place but they differ on the second and third place.

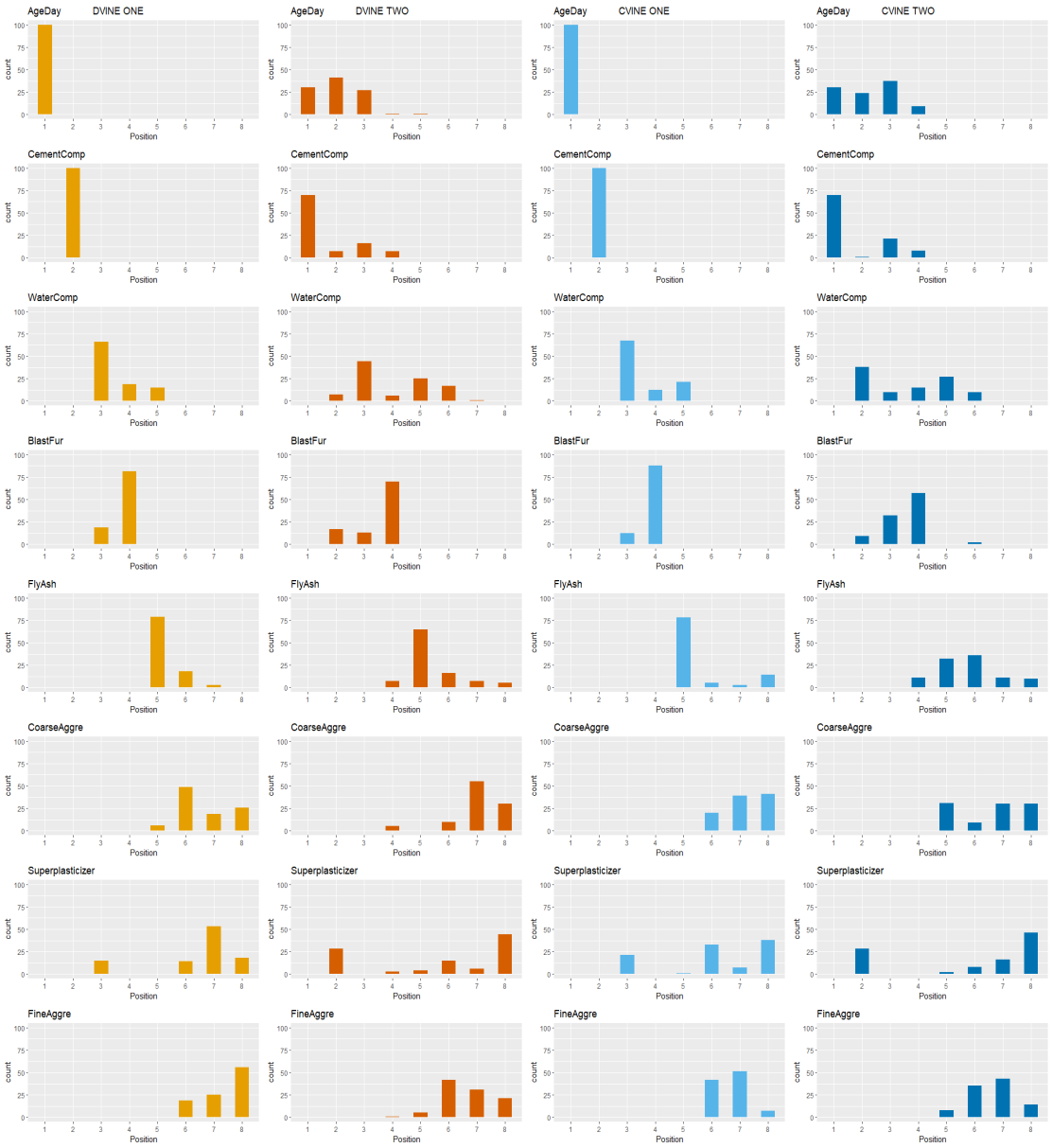


Figure 3.3.: Individual distribution plots. Column 1: D-vine one-step ahead, Column 2: D-vine two-step ahead, Column 3: C-vine one-step ahead, Column 4: C-vine two-step ahead.

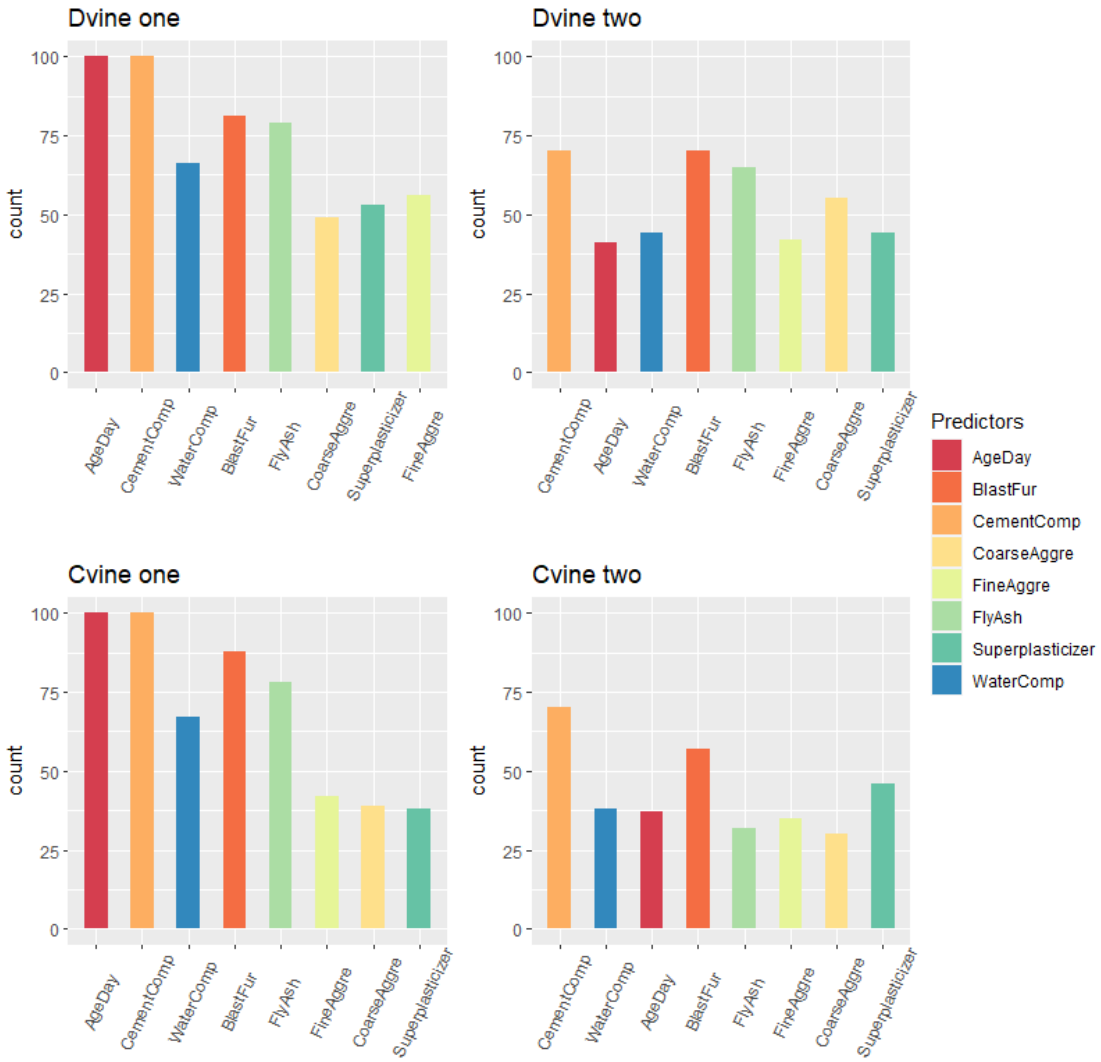


Figure 3.4.: Optimal orders of the algorithms. x-axis shows the optimal order, y-axis shows the count of how many times the corresponding predictor appeared in the 100 iterations at the position the predictor occupies in the optimal order.

3.5.2. Riboflavin data set

The Riboflavin data set, available in the R package `hdi`, aims at quantile predictions of the log-transformed production rate of Bacillus subtilis using log-transformed expression levels of 4088 genes. To reduce the computational burden, we perform a pre-selection of the top 100 genes with the highest variance (Bühlmann and Geer, 2011), resulting in a subset with $p = 100$ log-transformed gene expressions and $N = 71$ samples. Random splitting of the subset into training set with 61 samples and evaluation set with 10 samples, is repeated for 100 times. For the C- and D-vine two-step ahead quantile regression the number of candidates is set to $k = 10$. Additionally, to further reduce the computational burden the additional variable selection from Section 3.3.2 is applied with the chosen subset containing 25% of all possible choices, where 15% are predictors having the highest partial correlation with the log-transformed Bacillus subtilis production rate and the remaining 10% are chosen randomly from the remaining predictors. Performance of competitive quantile regression models is reported in Table 3.7, where we see that the proposed C-vine two-step ahead quantile regression is the best performing model and outperforms both the D-vine one-step ahead quantile regression from Kraus and Czado (2017) and the C-vine one-step ahead quantile regression to a large extent. Further, the second best performing method is the D-vine two-step ahead model which, while performing slightly worse than the C-vine two-step ahead model, also significantly outperforms both the C-vine and D-vine one-step ahead models. Since the predictors entering the C- and D-vine models yield a descending

Model	$\widehat{\text{IS}}_{0.05}$	$\widehat{\text{CL}}_{0.05}$	$\widehat{\text{CL}}_{0.5}$	$\widehat{\text{CL}}_{0.95}$
D-vine One-step	33.83	0.44	0.42	0.41
D-vine Two-step	30.57	0.44	0.38	0.33
C-vine One-step	34.52	0.49	0.43	0.38
C-vine Two-step	28.59	0.41	0.36	0.30

Table 3.7.: Out-of-sample predictions $\widehat{\text{IS}}_{0.5}$, $\widehat{\text{CL}}_{0.05}$, $\widehat{\text{CL}}_{0.5}$, $\widehat{\text{CL}}_{0.95}$. The best performing model is highlighted in gray.

order of the predictors contributing to maximizing the conditional log-likelihood, the order indicates the influence of the predictors to the response variable. It is often of practical interest to know which gene expressions are of the highest importance for prediction. Since we repeat the random splitting of the subset for $R = 100$ times, the importance of the gene expressions is ranked sequentially by choosing the one with the

highest frequency of each element in the order excluding the gene expressions chosen in the previous steps. For instance, the most important gene expression is chosen as the one most frequently ranked first; the second most important gene is chosen as the one most frequently chosen as the second element in the order, excluding the most important gene selected in the previous step. The top ten most influential gene expressions using the C- and D-vine one- or two-step ahead models are recorded in Table 3.8.

Model/Position	1	2	3	4	5	6	7	8	9	10
D-vine One-step	GGT	YCIC	MTA	RPSE	YVAK	THIK	ANSB	SPOVB	YVZB	YQJB
D-vine Two-step	MTA	RPSE	THIK	YMFE	YCIC	sigM	PGM	YACC	YVQF	YKPB
C-vine One-step	GGT	YCIC	MTA	RPSE	HIT	BFMBAB	PHRC	YBAE	PGM	YHEF
C-vine Two-step	MTA	RPSE	THIK	YCIC	YURU	PGM	sigM	YACC	YKRM	ASNB

Table 3.8.: The 10 most influential gene expressions on the conditional quantile function, ranked based on their position in the order.

3.6. On the stopping criteria

In the parametric set up of the D-vine regression (Kraus and Czado, 2017), there is a possibility of having a selection criteria penalization based on the number of parameters used for the fitted vine copula model. They utilize the selection criteria of Akaike information criterion (AIC) (Akaike, 1973b) and the Bayesian information criterion (BIC) (Schwarz, 1978). However, in the nonparametric setting the number of parameters is an open research question.

Usually in this case the *effective degree of freedom* (or edf) is used as a selection criteria. One possibility for its estimation is following Loader (2006, Section 5.3.2), using a so-called influence function and method of infinitesimal perturbations. This approach is implemented in the `rvinecopulib` library, so we do a short study on the possible advantages or disadvantages of its usage. Because of the low number of variables and the data sample size, we investigate adjusted AIC and BIC penalized conditional log-likelihoods as a selection criteria for the nonparametric case on the real life Concrete data set from Section 3.5.1.

3.6.1. AIC and BIC penalization

Let $\mathcal{C}$ be a C- or D-vine copula with order $\mathcal{O}(\mathcal{C}) = (V, U_1, \ldots, U_p)$. Additionally, assume that we are given N observations $\mathbf{v}$ and $\mathbf{u}_j$ for $j = 1, \ldots, p$. Let the effective degree of freedom be denoted as $|\Theta|$. To be able to properly compare these values for

this nonparametric approach, we consider slightly different versions of the AIC and BIC criterion. We define the adjusted AIC- and BIC-penalized conditional log likelihood functions as

$$cll^a_{AIC}\left(\mathcal{C}, \mathbf{v}, \left(\mathbf{u}_1, \ldots, \mathbf{u}_p\right)\right) = 2cll\left(\mathcal{C}, \mathbf{v}, \left(\mathbf{u}_1, \ldots, \mathbf{u}_p\right)\right) - 2|\Theta|,$$
$$cll^a_{BIC}\left(\mathcal{C}, \mathbf{v}, \left(\mathbf{u}_1, \ldots, \mathbf{u}_p\right)\right) = 2cll\left(\mathcal{C}, \mathbf{v}, \left(\mathbf{u}_1, \ldots, \mathbf{u}_p\right)\right) - log(N)|\Theta|. \tag{3.5}$$

The more standard way to define these values is to change the signs in the above equations, but for the sake of the analysis, we chose to consider these values, so that greater values of all 3 selection criteria indicate a better fit than lower values. In Table 3.9 we show the log-likelihoods and effective degrees of freedom of the pair copulas contributing to the conditional log-likelihood, the conditional log-likelihood cll, the penalized conditional log-likelihoods cll^a_{AIC} and cll^a_{BIC} at each step of the algorithm for each of the 4 considered algorithms with the optimal orders as shown in Figure 3.4.

Next, in Figure 3.5 we show how the conditional log-likelihood cll, the cll^a_{AIC} and the cll^a_{BIC} change for each model as a predictor is being included in the model at each step. On one side, in all 4 panels we can see that both the cll and penalized conditional log-likelihood cll^a_{AIC} have an increasing trend at the beginning, somewhere until the 5-th predictor is included in the model, and then they level off at a certain value. On the other side, the cll^a_{BIC} curve has a type of upside down u-shape. First, there is an increasing trend, and after 5 predictors are included in the model, the cll^a_{BIC} starts to decrease. This is the case as the BIC selection criteria usually penalizes more than the AIC criteria. Thus, if one would use the penalized cll^a_{BIC} or cll^a_{AIC} as a selection criteria, the algorithm would include only the first 5 predictors, since there is no significant improvement in the conditional log-likelihoods to justify including more predictors.

3.6.2. Recomputed models with a cut-off

Having the optimal orders of each algorithm and seeing that using a penalized cll^a_{AIC} or cll^a_{BIC} only the first 5 predictors in each model will be chosen, we can recompute the statistics of Table 3.6 such that the 100 iterations of splitting the data set remain the same (830 points training sample and 200 points evaluation sample), but at each iteration instead of searching the optimal order on the training set, the models are fitted with the first five elements of the optimal orders provided in Figure 3.4 (note that for every algorithm the order was taken as the first five elements of the orders in Figure 3.4 and this is done because the last three predictors were not influential based on both cll^a_{BIC} and cll^a_{AIC}). Table 3.10 shows the out-of-sample statistics for the models fitted with the optimal order.

Comparing Tables 3.6 and 3.10 we can see that both one-step ahead algorithms and the two-step ahead D-vine did not manage to improve the out-of-sample statistics

Predictor	Pos	Loglik_i	Edf_i	cll	cll^a_{AIC}	cll^a_{BIC}
D-vine One-step						
AgeDay	1	255.38	22.45	255.38	465.86	359.86
CementComp	2	285.70	25.11	541.08	987.04	762.49
WaterComp	3	163.88	23.43	704.96	1267.94	932.77
BlastFur	4	174.52	18.34	879.48	1580.30	1158.54
FlyAsh	5	79.41	16.92	958.89	1705.28	1203.63
CoarseAggre	6	36.19	18.73	995.08	1740.20	1150.12
Superplasticizer	7	28.51	20.89	1023.59	1755.44	1066.73
FineAggre	8	19.10	18.54	1042.69	1756.56	980.31
D-vine Two-step						
CementComp	1	179.19	24.51	179.19	309.36	193.64
AgeDay	2	365.88	26.61	545.07	987.90	746.54
WaterComp	3	170.94	27.63	716.01	1274.52	902.71
BlastFur	4	219.90	20.21	935.90	1673.90	1206.67
FlyAsh	5	78.52	20.76	1014.42	1789.42	1224.17
FineAggre	6	41.19	41.27	1055.61	1789.26	1029.16
CoarseAggre	7	26.17	24.10	1081.79	1793.40	919.51
Superplasticizer	8	38.94	23.86	1120.73	1823.56	837.02
C-vine One-step						
AgeDay	1	255.28	22.92	255.28	464.72	356.50
CementComp	2	285.65	25.41	540.93	985.20	757.01
WaterComp	3	165.83	25.69	706.76	1265.48	916.00
BlastFur	4	201.88	20.45	908.64	1628.34	1182.31
FlyAsh	5	79.90	19.73	988.63	1748.68	1209.49
FineAggre	6	28.07	52.75	1016.70	1699.32	911.08
CoarseAggre	7	41.30	28.72	1058.00	1724.48	800.64
Superplasticizer	8	30.92	18.98	1088.93	1748.36	734.91
C-vine Two-step						
CementComp	1	179.20	24.36	179.20	309.68	194.67
WaterComp	2	118.71	28.75	297.90	489.60	238.85
AgeDay	3	428.35	25.63	726.26	1295.04	923.27
BlastFur	4	227.42	25.24	953.67	1699.40	1208.47
FlyAsh	5	99.21	21.07	1052.88	1855.68	1265.27
FineAggre	6	36.45	22.25	1089.32	1884.08	1188.61
CoarseAggre	7	37.98	23.63	1127.31	1912.78	1105.75
Superplasticizer	8	21.74	16.30	1149.04	1923.66	1039.67

Table 3.9.: Log-likelihood, edf for the pair copulas contributing to the cll, the conditional log-likelihood cll, the penalized cll^a_{AIC} and cll^a_{BIC} .

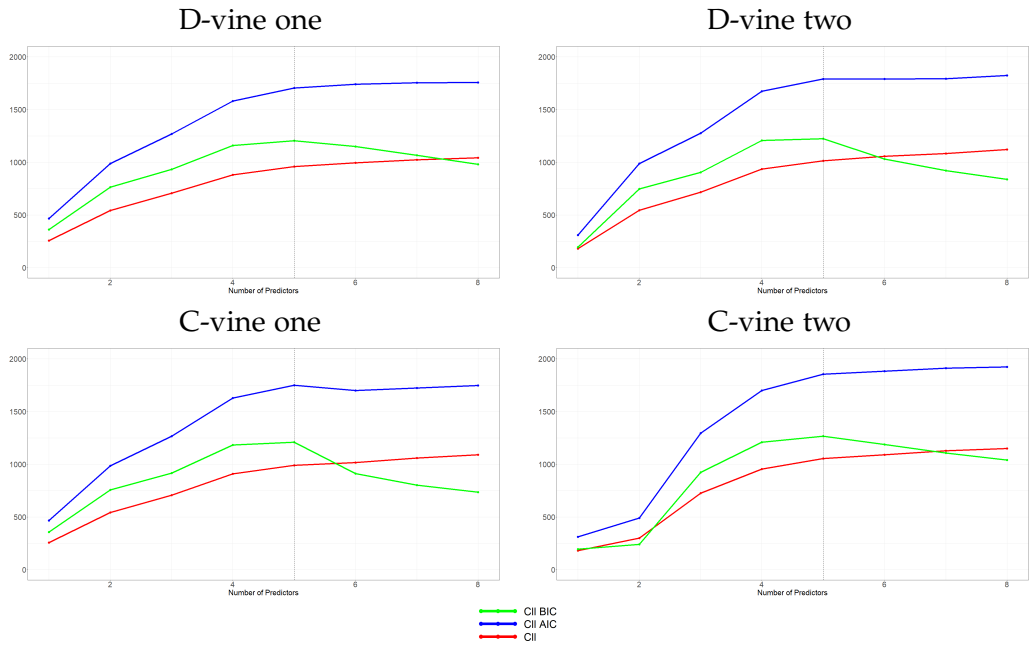


Figure 3.5.: Conditional log-likelihood cll and AIC/BIC penalized cll^a_{AIC}, cll^a_{BIC} plot.

with the new order. However, the two step-ahead C-vine algorithm showed a slight improvement in all four statistics.

Model	$\widehat{\text{IS}}_{0.05}$	$\widehat{\text{CL}}_{0.05}$	$\widehat{\text{CL}}_{0.5}$	$\widehat{\text{CL}}_{0.95}$
D-vine One-step ahead	1078.73	11.27	2.80	10.91
D-vine Two-step ahead	1018.72	10.98	2.77	9.98
C-vine One-step ahead	1038.93	11.18	2.69	9.95
C-vine Two-step ahead	949.42	10.49	2.57	9.09

Table 3.10.: Statistics on the models computed with the optimal fit using only the first 5 predictors.

3.7. Conclusion and outlook

In this chapter, we introduce a fully nonparametric vine based quantile regression framework. We suggest the usage of either one-step ahead or two-step ahead forward selection of predictors, which automatically orders predictors by their influence on the

target quantile function. Also, two possibilities for the selection of the tree structure can be used in such framework, either D- or C-vines. We compare these 4 models in an extensive simulation study including several different settings and data sets with different dimensions, strengths of dependence and tail dependencies.

Based on the simulation study, inclusion of future information, obtained through considering the next tree in the two-step ahead algorithm, yields a significantly less greedy sequential selection procedure in comparison to the already existing one-step ahead algorithm for D-vine based quantile regression in Kraus and Czado (2017). This is especially visible in the real data Concrete analysis, where we compare the orders of the predictors of the 4 models selected. However, the cost of this advantage is that the two-step ahead approach is more computationally expensive compared to the one-step ahead approach. But the usage of the variable reduction aids the two-step ahead approach to be less computationally expensive and work good in both low- and high-dimensional settings.

Further, for the first time, nonparametric bivariate copulas are used to construct vine copula-based quantile regression models. The nonparametric estimation overcomes the problem of possible family misspecification in the parametric estimation of bivariate copulas and allows for even more flexibility in dependence estimation. Additionally, under mild regularity conditions, the nonparametric conditional quantile estimator is shown to be consistent (Tepegjozova et al., 2022).

A further open research area is developing similar forward selection algorithms for R-vine tree structures while optimising the conditional log-likelihood (or shortly cll). However, in this case obtaining the conditional log-likelihood and estimating the quantile function can be computationally expensive, as numerical integration will be required. The standard lack of ability of copula based models to include discrete variables is also an ongoing research topic. Some results from Schallhorn et al. (2017) are available, however more research on this topic is required.

Also, there is no single 'best' model between the different vine structures, but it depends on the data at hand. Therefore, it would be useful to be able to choose between a C-vine, D-vine or R-vine information criteria. When maximum likelihood estimation is employed, the selection criteria by Akaike (AIC) (Akaike, 1973b), the Bayesian information criterion (BIC) (Schwarz, 1978) and the focussed information criterion (FIC) (Claeskens and N. Hjort, 2003) might be used immediately. Ko et al. (2019) studied FIC and AIC specifically for the selection of parametric copulas. The copula information criterion in the spirit of the Akaike information criterion by Grønneberg and N. L. Hjort (2014) can be used for selection among copula models with empirically estimated margins, while Ko and N. L. Hjort (2019) studied such a criterion for parametric copula models only. Another open research area is the information criteria for nonparametrically estimated copulas and for vines in particular, as our

small study in Section 3.6.2 did not show advantages in using the edf from Loader (2006) as a stopping criteria.

Nonparametrically estimated vines are offering considerable flexibility. Their parametric counterparts, on the other hand, are enjoying simplicity. An interesting route for further research is to combine parametric and nonparametric components in the construction of the vines in an efficient way to bring the most benefit, which should be made tangible through some criterion such that guidance can be provided about which components should be modeled nonparametrically and which others are best modeled parametrically. For some types of models, such choice between a parametric and a nonparametric model has been investigated by Jullum and N. L. Hjort (2017) via the focussed information criterion. This and alternative methods taking the effective degrees of freedom into account are worth further investigating for vine copula models.

4. Bivariate response vine copula based regression

Parts of Chapter 4 are similar to the publication Tepegjozova and Czado (2022). However, Sections 4.1 and 4.5 contain new material.

4.1. Introduction

Comprehending the dependence of a set of response variables and determining their statistical properties in relation to a set of predictor variables is the mathematical basis for many practical usages. For example, when it comes to analyzing insurance data, insurance providers keep track of the quantity of claims (frequency) over a subset of policyholders and the mean amount of claim sizes made (severity). Investigating the joint conditional distribution of these variables, with consideration of the policyholders' attributes, is a crucial aspect of insurance companies decision-making and risk assessment procedures. In Frees et al. (2016), the authors use a multivariate frequency-severity regression modeling for each of the response variables, and a joint copula for modeling the dependence among these outcomes. In advance, utilizing the joint conditional distribution in relation to a specific set of predictors can establish a strong statistical foundation for calculating and optimizing conditional risk metrics, such as Value-at-Risks and Expected Shortfall (Noyan and Rudolf, 2013).

Also, in climatological applications there has been an increased need for joint modeling of more than one response variable. For example, Singh et al. (2022) model the joint distribution of temperature and precipitation using a hierarchical Bayesian framework to analyse the bias in different weather simulations. Bevacqua et al. (2017) studied compound floods in the coastal region of Ravenna, Italy based on the joint occurrence of storm surges and high river levels using vine copulas. In Shiau and Modarres (2009), Sarhadi et al. (2016), and Kwon and Lall (2016) copulas are employed to construct the joint distribution function of drought severity and duration. Copula based methods are used for downscaling climate simulations and bias correction, based on the copula between the observed and the simulated values of precipitation measurement (Laux et al., 2011).

However, usually in the applications using copula modeling, either the marginal distributions are modeled first using univariate regression models, and then a copula is used to model their joint dependence, or the joint dependence between two variables is modeled with a copula, ignoring any predictor variables that may influence this dependence. There is a need for a joint modeling of more than one response variable, given a set of predictor variables, especially if there is dependence in the data (either among the responses, among the predictors, or between the responses and the predictors). The one response vine based regressions, discussed in Chapter 3, are able to handle modeling dependence among the predictors and between a response variable and a set of predictors, but we would like to introduce a model that can handle modeling the dependence between more than one response variable. We start by introducing a vine copula based regression model that can handle modeling of two response variables and a set of predictor variables.

The first heuristic for a vine based regression with multiple responses is given in Zhu et al. (2021). However, this approach has an asymmetric treatment of the response variables. This might lead to different performance of the regression methods when the order of the response variables is exchanged. Further, the suggested heuristic for the bivariate response case is limited to modelling only, but not prediction. Therefore, we tackle the problem of proposing a vine copula based regression framework that can handle two responses in a symmetric manner and for which prediction methods would be obtainable.

We propose a novel vine tree structure or sequence, called Y-vine tree sequence, which is a member of the set of regular vine tree sequences. It is designed to allow for a symmetric treatment of the responses. Moreover, we show that using the Y-vine tree sequence the associated bivariate conditional density is analytically expressible as a product of all the pair copula terms involving one or both of the response variables. In the case of more than one conditioning variable (predictor) we develop a forward selection method. For this we propose an appropriate fit measure for the predictors to prevent overfitting and remove non-significant predictors.

Further, for applicability of the proposed method we develop a prediction method. We extend the methodology with a simulation method of bivariate copula data as well. Finally, we give an application involving a data set with minimal and maximal daily temperatures together with other weather variables. For this application we show that the conditional dependence cannot be ignored between the response variables and that it is non-Gaussian dependence structure, thus requiring the full class of pair copula families.

4.2. Vine copula based bivariate regression

4.2.1. General framework

Consider the variables $(Y_1, Y_2)^T$ as the 2-dimensional response vector and $\mathbf{X} = (X_1, \ldots, X_p)^T$ as the p-dimensional predictor vector. The main interest of the bivariate regression is to model the joint conditional distribution function of the response variables $\mathbf{Y} = (Y_1, Y_2)^T$ given the outcome of some predictor variables $\mathbf{X} = \mathbf{x}$, denoted as $F_{Y_1,Y_2|\mathbf{X}}(y_1, y_2|\mathbf{x})$.

This can be achieved by joint modelling of $(\mathbf{Y}, \mathbf{X})^T$ and subsequently estimating the conditional distribution of the bivariate response vector $\mathbf{Y}$ given $\mathbf{X} = \mathbf{x}$. The same can be achieved by joint modelling of the PIT values of the responses $\mathbf{V} = (V_1, V_2)^T$, the predictors $\mathbf{U} = (U_1, \ldots, U_p)^T$, and the corresponding conditional distribution function of $\mathbf{V}$ given $\mathbf{U} = \mathbf{u}$, denoted as $C_{V_1,V_2|\mathbf{U}}(v_1, v_2|\mathbf{u})$. The connection between these two approaches for the joint conditional distribution, on the x- and u-scale, is derived in Proposition 1.

Proposition 1 *The conditional distribution of* $\mathbf{Y} = (Y_1, Y_2)^T$ *given* $\mathbf{X} = (X_1, \ldots, X_p)^T$, *with corresponding PITs* $V_j := F_{Y_j}(Y_j)$, $j = 1, 2$ *and* $U_i := F_{X_i}(X_i)$, $i = 1, \ldots, p$ *can be expressed in terms of a conditional distribution function associated with a copula as*

$$F_{Y_1,Y_2|\mathbf{X}}(y_1, y_2|\mathbf{x}) = C_{V_1,V_2|\mathbf{U}}\left(F_{Y_1}(y_1), F_{Y_2}(y_2)|F_{X_1}(x_1), \ldots, F_{X_p}(x_p)\right).$$

Proof of Proposition 1 is given in Appendix A.1.1. Here $C_{V_1,V_2|\mathbf{U}}$ denotes the bivariate conditional distribution associated with the $p+2$ dimensional copula $C_{V_1,V_2,\mathbf{U}}$ and does not need to have uniform margins. In general, $C_{V_1,V_2|\mathbf{U}}$ is different than $C_{V_1,V_2;\mathbf{U}}$, as $C_{V_1,V_2;\mathbf{U}}$ is a bivariate copula with uniform marginal distributions and corresponds to the copula associated with the bivariate conditional distribution of (Y_1, Y_2) given $\mathbf{X} = \mathbf{x}$.

From Proposition 1, in order to model the bivariate conditional distribution function, we need to estimate the marginal distributions F_{Y_j}, F_{X_i} for $j = 1, 2$, $i = 1, \ldots, p$, and the bivariate conditional distribution $C_{V_1,V_2|\mathbf{U}}$. To obtain the later, we need to estimate the $p+2$ dimensional copula $C_{V_1,V_2,\mathbf{U}}$ describing the joint distribution of $(V_1, V_2, \mathbf{U})$. Following Kraus and Czado (2017) and Noh et al. (2013), we estimate the marginal distributions nonparametericallly to reduce the bias caused by model misspecification. The same nonparametric estimation of the marginal distribution functions is used in Chapter 3 as well. A more complex task is estimating the $p+2$ dimensional copula $C_{V_1,V_2,\mathbf{U}}$ and subsequently, deriving the bivariate conditional distribution from this copula. We propose to model the copula $C_{V_1,V_2,\mathbf{U}}$ using regular vine copulas. However, we also have to take care that deriving the bivariate conditional distribution $C_{V_1,V_2|\mathbf{U}}$

remains numerically tractable. Thus, to obtain the joint conditional distribution of the response variables using only pair copulas estimated in the vine copula model, additional constraints are required.

The constraint for a univariate vine regression is that the node containing the response in the conditioned set is a leaf node in each tree of the tree sequence, as discussed in Section 3.2. Following these results, the constraint for the bivariate vine regression model is that the two response variables are exactly the conditioned set of the edge of the last tree in the vine tree sequence, as also used by Zhu et al. (2021). However, in their approach there is no symmetric treatment of the two responses, which is a drawback. Therefore, we propose a novel vine tree sequence specifically designed for bivariate regression modelling allowing for a symmetric treatment of the responses.

4.2.2. Y-vine copula model

Let $\mathbf{X}_{-i}$ be a $(p-1)$-dimensional vector defined as $\mathbf{X}_{-i} \coloneqq (X_1, \ldots, X_{i-1}, X_{i+1}, \ldots X_p)^T$ and let $\mathbf{X}_{i:i+k}$ be a $(k+1)$-dimensional vector defined as $\mathbf{X}_{i:i+k} \coloneqq (X_i, \ldots, X_{i+k})^T$. Similar definitions hold for the vectors $\mathbf{x}_{-i}, \mathbf{U}_{-i}$, $\mathbf{u}_{-i}$, and for $\mathbf{x}_{i:i+k}, \mathbf{U}_{i:i+k}$ $\mathbf{u}_{i:i+k}$, respectively.

Definition 1 *Given the marginal PIT transformed response variables V_1, V_2 and predictor variables $U_1, \ldots, U_p$, we define the $p+1$ trees of the Y-vine tree sequence for bivariate regression as the following:*

$\mathbf{T_1}$ *with* $N_1 = \{V_1, V_2, U_1, \ldots, U_p\}$ *and* $E_1 = \{(V_1, U_1),\ (V_2, U_1)\} \bigcup_{i=1}^{p-1} (U_i, U_{i+1})$.

$\mathbf{T_2}$ *with* $N_2 = \{V_1U_1, V_2U_1, U_1U_2, \ldots, U_{p-1}U_p\}$ *and*

$E_2 = \{(V_1U_1,\ U_1U_2),\ (V_2U_1,\ U_1U_2)\} \bigcup_{i=1}^{p-2} (U_iU_{i+1},\ U_{i+1}U_{i+2})$.

$\mathbf{T_k}$ *for* $3 \leq k \leq p$ *with* $N_k = \bigcup_{j=1,2} \{V_jU_{k-1}; \mathbf{U}_{1:k-2}\} \bigcup_{i=1}^{p-k+1} \{U_iU_{i+k-1}; \mathbf{U}_{i+1:i+k-2}\}$

and $E_k = \bigcup_{j=1,2} \{(V_jU_{k-1}; \mathbf{U}_{1:k-2},\ \ U_1U_k; \mathbf{U}_{2:k-1})\}$

$\bigcup_{i=1}^{p-k} \{(U_iU_{i+k-1}; \mathbf{U}_{i+1:i+k-2},\ \ U_{i+1}U_{i+k}; \mathbf{U}_{i+2:i+k-1})\}$.

$\mathbf{T_{p+1}}$ *with* $N_{p+1} = \bigcup_{j=1,2} \{V_jU_p; \mathbf{U}_{1:p-1}\}$ *and* $E_{p+1} = \{(V_1U_p; \mathbf{U}_{1:p-1},\ \ V_2U_p; \mathbf{U}_{1:p-1})\}$.

The newly proposed Y-vine tree sequence is graphically illustrated in Figure 4.1. In each tree of the vine tree sequence the nodes containing the predictor variables in the conditioned set are arranged in a path, while the nodes containing the response variables in the conditioned set are added as leafs of the path on one end. The subset of the sequence that contains a single response and all predictors forms a D-vine tree sequence. In addition, this tree structure allows for symmetric treatment of the response variables, especially important since an asymmetric treatment might lead to different performances of the regression models on different responses.

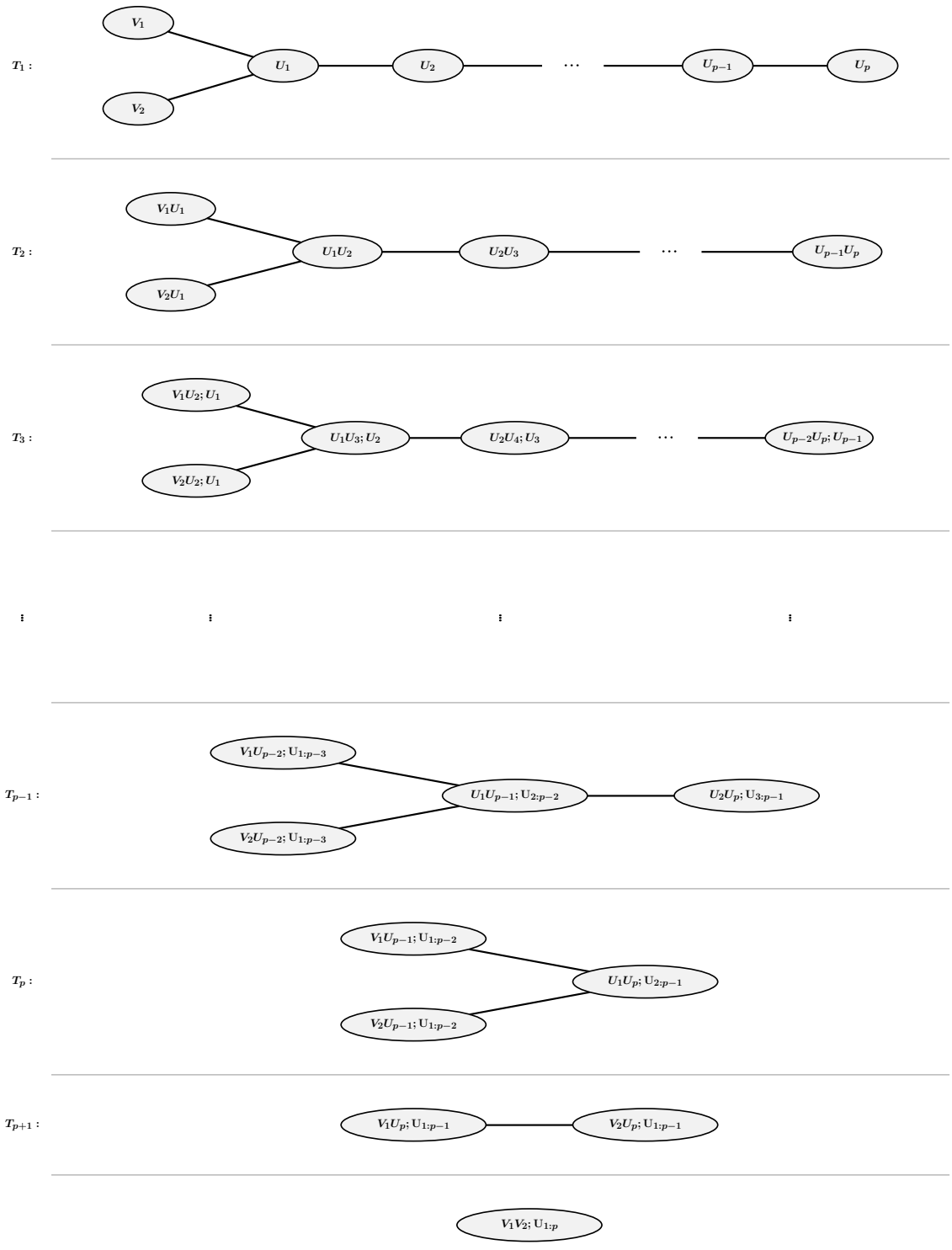


Figure 4.1.: Y-vine tree sequence on the u-scale.

Proposition 2 *The Y-vine tree sequence from Definition 1, satisfies the regular vine tree sequence conditions (i)-(iii) from Section 2.2 and thus, represents a valid regular vine tree sequence.*

Proof of Proposition 2 is given in Appendix A.1.2.

A regular vine copula associated with a Y-vine tree sequence, denoted as $\mathcal{V}$, together with a set of bivariate copulas $\mathcal{B}(\mathcal{V})$ and the corresponding pair copula parameters $\Theta(\mathcal{B}(\mathcal{V}))$ is called a Y-vine copula and we denote it by $\mathcal{Y}$. The joint density $f_{Y_1,Y_2,\mathbf{X}}$ using a Y-vine tree sequence can be expressed by Equation(4.1) as

$$\begin{aligned} f_{Y_1,Y_2,\mathbf{X}}(y_1,y_2,\mathbf{x}) = \prod_{k=1}^{p-1}\Bigg[\prod_{i=1}^{p-k} c_{U_i,U_{i+k};\mathbf{U}_{i+1:i+k-1}}\Big(F_{X_i|\mathbf{X}_{i+1:i+k-1}}(x_i|\mathbf{x}_{i+1:i+k-1}), \\ F_{X_{i+k}|\mathbf{X}_{i+1:i+k-1}}(x_{i+k}|\mathbf{x}_{i+1:i+k-1})\Big)\Bigg] \\ \cdot \prod_{i=1}^{p}\Bigg[\prod_{j=1,2} c_{V_j,U_i;\mathbf{U}_{1:i-1}}\Big(F_{Y_j|\mathbf{X}_{1:i-1}}(y_j|\mathbf{x}_{1:i-1}),\ F_{\mathbf{X}_i|X_{1:i-1}}(x_i|\mathbf{x}_{1:i-1})\Big)\Bigg] \\ \cdot c_{V_1,V_2;\mathbf{U}}\left(F_{Y_1|\mathbf{X}}(y_1|\mathbf{x}), F_{Y_2|\mathbf{X}}(y_2|\mathbf{x})\right)\cdot\prod_{i=1}^{p} f_{X_i}(x_i)\cdot\prod_{j=1,2} f_{Y_j}(y_j). \end{aligned} \tag{4.1}$$

Theorem 1 *The joint conditional density of (Y_1,Y_2) given the predictors $\mathbf{X}=(X_1,\ldots,X_p)^T$ denoted by $f_{Y_1,Y_2|\mathbf{X}}$ in a Y-vine copula is given as*

$$\begin{aligned} f_{Y_1,Y_2|\mathbf{X}}(y_1,y_2|\mathbf{x}) = \prod_{i=1}^{p}\Bigg[\prod_{j=1,2} c_{V_j,U_i;\mathbf{U}_{1:i-1}}\Big(F_{Y_j|\mathbf{X}_{1:i-1}}(y_j|\mathbf{x}_{1:i-1}), F_{X_i|\mathbf{X}_{1:i-1}}(x_i|\mathbf{x}_{1:i-1})\Big)\Bigg] \\ \cdot c_{V_1,V_2;\mathbf{U}}\left(F_{Y_1|\mathbf{X}}(y_1|\mathbf{x}), F_{Y_2|\mathbf{X}}(y_2|\mathbf{x})\right)\cdot\prod_{j=1,2} f_{Y_j}(y_j). \end{aligned} \tag{4.2}$$

Proof of Theorem 1 is given in Appendix A.1.3.

In order to determine the joint and the bivariate conditional density, $c_{V_1,V_2,\mathbf{U}}$ and $c_{V_1,V_2|\mathbf{U}}$, we only need to set the marginals to uniform densities, i.e. $f_{Y_j}(y_j)=1$, $j=1,2$ and $f_{X_i}(x_i)=1$, $i=1,\ldots,p$ in Equation (4.1) and Equation (4.2) respectively. Thus, with the proposed Y-vine copula we can express the conditional bivariate density as a product of pair copula densities occurring in the Y-vine tree sequence that contain a response in the conditioned set, and the marginal densities of the responses. No integration is needed.

In addition to the analytic form of the joint conditional density $f_{Y_1,Y_2|\mathbf{X}}$, from the Y-vine we can also derive other conditional densities in an analytic form.

Corollary 1 *From the Y-vine copula associated with the Y-vine tree sequence of Definition 1, we can derive the following conditional densities:*

a. for $j = 1, 2$ it holds

$$f_{Y_j|\mathbf{X}}\left(y_j|\mathbf{x}\right) = f_{Y_j}(y_j) \cdot \prod_{i=1}^{p} c_{V_j,U_i;\mathbf{U}_{1:i-1}}\left(F_{Y_j|\mathbf{X}_{1:i-1}}\left(y_j|\mathbf{x}_{1:i-1}\right), F_{X_i|\mathbf{X}_{1:i-1}}\left(x_i|\mathbf{x}_{1:i-1}\right)\right); \tag{4.3}$$

b. for $j, k \in \{1, 2\}$ with $j \neq k$, it holds

$$\begin{aligned} f_{Y_k|\mathbf{X},Y_j}\left(y_k|\mathbf{x}, y_j\right) = &\prod_{i=1}^{p}\left[c_{V_k,U_i;\mathbf{U}_{1:i-1}}\left(F_{Y_k|\mathbf{X}_{1:i-1}}\left(y_k|\mathbf{x}_{1:i-1}\right), F_{X_i|\mathbf{X}_{1:i-1}}\left(x_i|\mathbf{x}_{1:i-1}\right)\right)\right] \\ &\cdot c_{V_1,V_2;\mathbf{U}_{1:p}}\left(F_{Y_1|\mathbf{X}_{1:p}}\left(y_1|\mathbf{x}\right), F_{Y_2|\mathbf{X}}\left(y_2|\mathbf{x}\right)\right) \cdot f_{Y_k}(y_k). \end{aligned} \tag{4.4}$$

Proof of Corollary 1 is given in Appendix A.1.4. For the associated univariate conditional densities $c_{V_1|\mathbf{U}}\left(v_1|\mathbf{u}\right)$, $c_{V_2|\mathbf{U}}\left(v_2|\mathbf{u}\right)$, and $c_{V_1|\mathbf{U},V_2}\left(v_1|\mathbf{u}, v_2\right)$, $c_{V_2|\mathbf{U},V_1}\left(v_2|\mathbf{u}, v_1\right)$, we set $f_{Y_j}(y_j) = 1$, $j = 1, 2$, in Equation (4.3) and Equation (4.4) respectively. The univariate conditional distribution functions $C_{V_1|\mathbf{U}}$, $C_{V_2|\mathbf{U}}$, and $C_{V_1|\mathbf{U},V_2}$, $C_{V_2|\mathbf{U},V_1}$ can be obtained through integration of these associated conditional densities. The bivariate conditional distribution function $C_{V_1,V_2|\mathbf{U}_{1:p}}$ is obtained as:

$$\begin{aligned} C_{V_1,V_2|\mathbf{U}_{1:p}}\left(v_1, v_2|\mathbf{u}_{1:p}\right) &= \int_0^{v_1}\int_0^{v_2} c_{V_1,V_2|\mathbf{U}_{1:p}}\left(v_1', v_2'|\mathbf{u}_{1:p}\right) dv_2' dv_1' \\ &= \int_0^{v_1}\int_0^{v_2} c_{V_2|\mathbf{U}_{1:p}}\left(v_2'|\mathbf{u}_{1:p}\right) \cdot c_{V_1|V_2,\mathbf{U}_{1:p}}\left(v_1'|v_2', \mathbf{u}_{1:p}\right) dv_2' dv_1' \\ &= \int_0^{v_2} c_{V_2|\mathbf{U}_{1:p}}\left(v_2'|\mathbf{u}_{1:p}\right) \cdot \left[\int_0^{v_1} c_{V_1|V_2,\mathbf{U}_{1:p}}\left(v_1'|v_2', \mathbf{u}_{1:p}\right) dv_1'\right] dv_2' \\ &= \int_0^{v_2} c_{V_2,\mathbf{U}_{1:p}}\left(v_2', \mathbf{u}_{1:p}\right) \cdot C_{V_1|V_2,\mathbf{U}_{1:p}}\left(v_1|v_2', \mathbf{u}_{1:p}\right) dv_2'. \end{aligned} \tag{4.5}$$

One can also condition on V_1 instead of V_2 in Equation (4.5).

4.3. Sequential forward selection of predictors

Until now, we ordered the predictors as U_1 to U_p, however other permutations are possible. Let's denote the associated permutation of the Y-vine $\mathcal{Y}$ from Figure 4.1 by

$\mathcal{O}_{\mathcal{Y}}(\mathcal{Y}) := (U_1, U_2, \ldots, U_{p-1}, U_p)$. It is the order in which the predictors appear in T_1 of the tree sequence. Compared to the univariate case in Section 3.2 for the order of C- and D-vines, here we don't include the responses in the order as they have symmetric treatment, so we do not order them too.

Similar to the univariate case, one can choose the order of the predictors randomly, but the predictive power of the fit greatly depends on the chosen order. Different orders will produce different Y-vine fits, as the influence over the two responses varies with the predictors. There are $p!$ possible permutations of this order, computing and comparing each of them is not feasible and the optimal permutation is in general unknown. Thus, we propose an algorithm that automatically constructs a Y-vine by sequentially ordering predictors. In addition, we apply a stopping criteria to prevent overfitting, meaning that the least influential predictors will not be considered in the model. This way we obtain an automatic forward selection of predictors for the bivariate regression model. Other approaches can be developed based, for example, on the two-step ahead approach from Section 3.3.2 or approaches based on background knowledge specifying a predefined order, or different fit measures and selection criteria.

Joint conditional log-likelihood

The goal is to find the order of the predictors that has the greatest explanatory power. To compare and quantify the explanatory power of different bivariate regression models again we propose a log-likelihood approach. Inspired by the one response vine based regression in Chapter 3, we would like to associate the fit measure with the target function of the bivariate vine based regression. A suitable choice is the log-likelihood of $c_{V_1,V_2|\mathbf{U}_{1:p}}$, since $c_{V_1,V_2|\mathbf{U}_{1:p}}$ is the corresponding density of the target function. However, before deciding on the fit measure we take a more precise look at the proposed log-likelihood.

Following Killiches et al. (2018), the conditional copula density $c_{V_j|\mathbf{U}_{1:p}}$ can be rewritten as a product of all pair copulas that contain the response V_j in a D-vine copula. In the bivariate response case using Y-vines, we can express $c_{V_1,V_2|\mathbf{U}_{1:p}}$ as a product of all pair copulas that contain the responses V_1 and V_2, as shown in Equation (4.2) by setting the marginals to uniform densities. Thus, the log-likelihood of $c_{V_1,V_2|\mathbf{U}_{1:p}}$ associated with a Y-vine, can be written as

$$\begin{aligned}\ell\left(c_{V_1,V_2|\mathbf{U}_{1:p}}\right) &= \ell\left(c_{V_1,V_2;\mathbf{U}_{1:p}}\right) + \ell\left(c_{V_1|\mathbf{U}_{1:p}}\right) + \ell\left(c_{V_2|\mathbf{U}_{1:p}}\right) \\ &= \ell\left(c_{V_1,V_2;\mathbf{U}_{1:p}}\right) + \sum_{j=1,2}\left[\ell\left(c_{V_j,U_1}\right) + \sum_{k=2}^{p}\ell\left(c_{V_j,U_k;\mathbf{U}_{1:k-1}}\right)\right],\end{aligned}$$

where $\ell(f)$ denotes the log-likelihood associated to a statistical model with density

f and a given independent and identically distributed sample. Here we used the predictor order as given in Figure 4.1. The pair copula density $c_{V_j,U_k;\mathbf{U}_{1:k-1}}$ represents the behaviour between U_k and V_j given that the effects of the conditioning values $U_1,\ldots,U_{k-1}$ are adjusted. Therefore, a large value of the log-likelihood $\ell\left(c_{V_j,U_k;\mathbf{U}_{1:k-1}}\right)$ indicates an influence of U_k on the response V_j after $\mathbf{U}_{1:k-1}$ are already in the model. This implies that the log-likelihoods associated with the pair copulas $c_{V_j,U_k;\mathbf{U}_{1:k-1}}$ are suitable for a fit measure since we can interpret an increase in the fit measure as an increase in influence from a certain predictor. But what importance does the copula between the responses given the predictors $c_{V_1,V_2;\mathbf{U}_{1:k}}$ have on the predictive power of the model is a valid question for $k=2,\ldots,p$. The term $c_{V_1,V_2;\mathbf{U}_{1:k}}$ represents the behaviour between V_1 and V_2 given that the effects of $U_1,\ldots,U_k$ are adjusted. This implies that neither an increase nor a decrease in the log-likelihood $\ell\left(c_{V_1,V_2;\mathbf{U}_{1:k}}\right)$ can be interpreted as an increase in influence for a single predictor. Thus, $c_{V_1,V_2;\mathbf{U}_{1:k}}$ for $k=2,\ldots,p$ fails to quantify the marginal effect of any predictor on the responses and we exclude it from our proposed fit measure. Finally, we formally introduce the *adjusted conditional log-likelihood* as our fit measure.

Definition 2 *The adjusted conditional log-likelihood of a bivariate Y-vine based regression model, denoted by ac$\ell\ell$, with PIT transformed response and predictor variables $V_1, V_2, U_1,\ldots,U_p$, is defined as*

$$\begin{aligned} acll\left(\mathcal{Y}\right) &:= \ell\left(c_{V_1,V_2|\mathbf{U}_{1:p}}\right) - \ell\left(c_{V_1,V_2;\mathbf{U}_{1:p}}\right) \\ &= \sum_{j=1,2}\left[\ell\left(c_{V_j,U_1}\right) + \sum_{k=2}^{p}\ell\left(c_{V_j,U_k;\mathbf{U}_{1:k-1}}\right)\right]. \end{aligned} \tag{4.6}$$

Since we are interested in forward selection of predictors, we need to easily compare nested models with one predictor difference. Let $\mathcal{Y}_{p-1}$ and $\mathcal{Y}_p$ be two nested Y-vine based regression models with response variables V_1, V_2, where $\mathcal{Y}_{p-1}$ includes the predictors $U_1,\ldots,U_{p-1}$ in that order and $\mathcal{Y}_p$ includes the predictors $U_1,\ldots,U_{p-1},U_p$. Then the connection between the adjusted conditional log-likelihoods of those nested models is given as

$$acll\left(\mathcal{Y}_p\right) = \ell\left(c_{V_1,U_p;\mathbf{U}_{1:p-1}}\right) + \ell\left(c_{V_2,U_p;\mathbf{U}_{1:p-1}}\right) + acll\left(\mathcal{Y}_{p-1}\right), \tag{4.7}$$

and we use this result for forward selection of predictors.

Automatic forward selection algorithm

Assume we start with the PIT transformed response and predictors $V_1, V_2, U_1,\ldots,U_p$, and their observations $\mathbf{v}_n = \left(v_1^n, v_2^n\right)^T$, $\mathbf{u}_n = \left(u_1^n,\ldots,u_p^n\right)^T$, for $n=1,\ldots,N$. We would

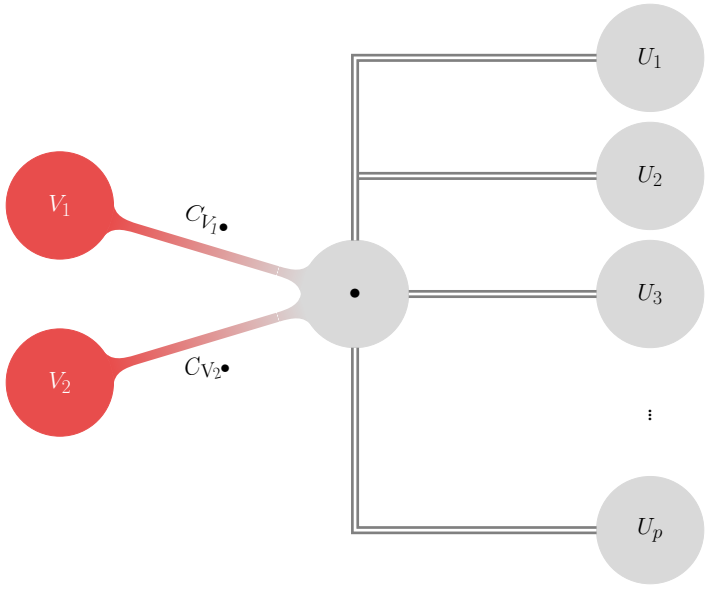


Figure 4.2.: Step 1 for the forward selection algorithm.

like to fit a Y-vine copula model to the data, given that V_1, V_2 are the responses. First, we build a Y-vine copula model with one predictor only. To see which predictor needs to be on the first place in the order, we fit all possible one-predictor Y-vines. See Figure 4.2 for an illustration of the step 1 of the algorithm. We derive their adjusted conditional log-likelihoods using Equation (4.6), and the predictor that maximizes it, say U_1 becomes the first predictor in the order of the Y-vine model. Let's denote the fitted Y-vine model with one predictor as $\hat{\mathcal{Y}}_1$ with order $\mathcal{O}_{\mathcal{Y}}(\hat{\mathcal{Y}}_1) = (U_1)$.

In the next step, we need to choose the second predictor to be added to the model. See Figure 4.3 for an illustration of the step 2 of the algorithm. We fit the additional pair copulas that need to be estimated for the adjusted conditional log-likelihood. Following Equation (4.7), we need to estimate two more copulas for each of the remaining predictors, derive the adjusted conditional log-likelihoods and the predictor that maximizes it, say U_2 becomes the second predictor in the order. Thus, at the end of the second step we have a fitted Y-vine model with two predictors denoted as $\hat{\mathcal{Y}}_2$ with order $\mathcal{O}_{\mathcal{Y}}(\hat{\mathcal{Y}}_2) = (U_1, U_2)$. We continue this forward selection algorithm until we order all predictors or if none of the remaining predictors is able to increase the conditional log-likelihood of the model. Other options are available, based on the AIC/BIC penalized conditional log-likelihood, similar as in Kraus and Czado (2017) and defined in Equation (3.5) in Section 3.6, where $|\Theta|$ is the number of parameters of

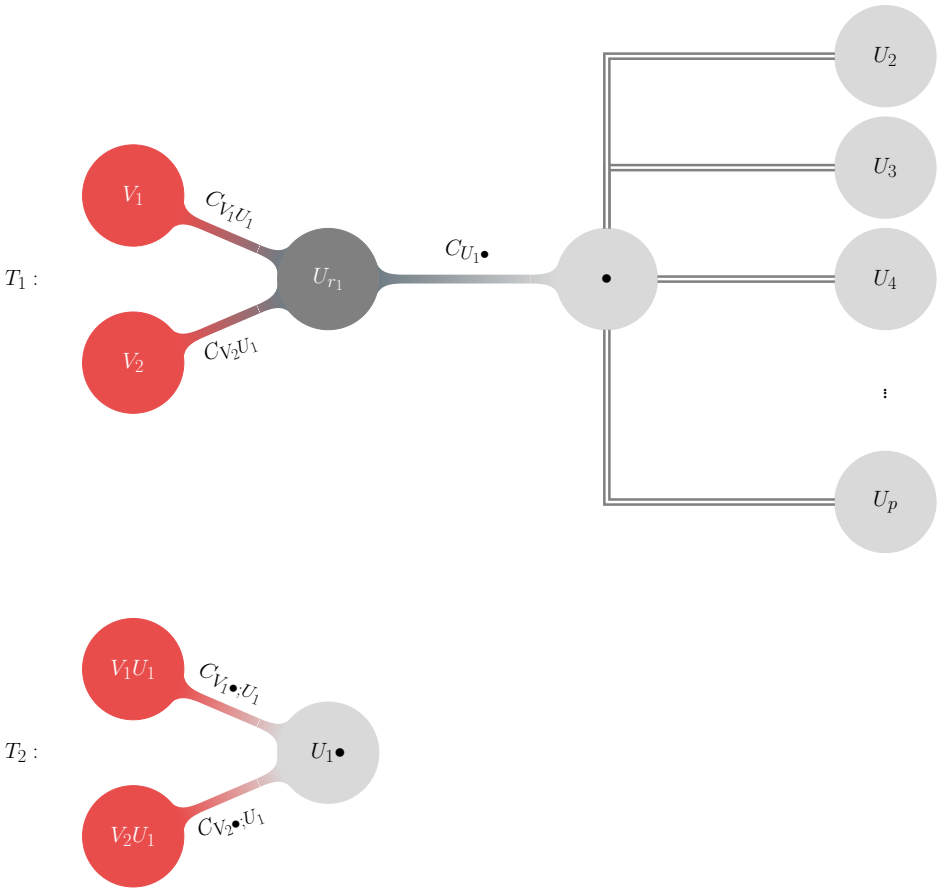


Figure 4.3.: Step 2 for the forward selection algorithm.

the pair copulas that are used to calculate the adjusted conditional log-likelihood. In the Y-vine automatic forward selection algorithm, we use a semi-parametric approach, since a stopping criteria for a fully nonparametric approach is still an open question, as discussed in Section 3.7. This implies that we estimate the marginal distribution function in a nonparametric way, as in Section 3.2.2 and use parametric pair copulas, discussed in Section 2.1.2 to construct the Y-vine, to be able to utilize a stopping criteria. The full estimation procedure and the pseudo code for the algorithm is given in Appendix A.2.

4.4. Prediction method for bivariate regression

Assume we have fitted a bivariate Y-vine regression model $\hat{\mathcal{Y}}$ on a bivariate response vector $(V_1, V_2)^T$ with order of predictors $\mathcal{O}_{\mathcal{Y}}(\hat{\mathcal{Y}}) = (U_1, \ldots, U_p)$. The fitted vine has a tree sequence and pair copula family sets denoted by $\hat{\mathcal{V}}$ and $\hat{\mathcal{B}}(\hat{\mathcal{V}})$, respectively. Given a new realization $\mathbf{u}^{new} = (u_1^{new}, \ldots, u_p^{new})^T$, our target is to evaluate the function $C_{V_1,V_2|\mathbf{U}}$ at every integration point $\mathbf{v}^{inp} = (v_1^{inp}, v_2^{inp})^T \in [0,1]^2$ and determine the integral given in Equation (4.5).

We apply the chosen adaptive quadrature algorithm for integration (see more in Piessens et al. (2012)), which requires the ability to evaluate the function under the integral at all points of the integration interval. Therefore, given a point $\mathbf{v}^{inp} = (v_1^{inp}, v_2^{inp})^T$ we define the integrand associated with Equation (4.5) and denoted by $IN(z)$ for any $0 < z < v_2^{inp}$, as

$$IN(z) \coloneqq c_{V_2|\mathbf{U}}\left(z|\mathbf{u}^{new}\right) \cdot C_{V_1|V_2,\mathbf{U}}\left(v_1^{inp}|z, \mathbf{u}^{new}\right). \tag{4.8}$$

The integration is carried out over the interval $(0, v_2^{inp})$. While the first term in Equation (4.8) is available analytically since it is the conditional density associated with the D-vine with order $(V_2, U_1, \ldots, U_p)$, the second term needs further consideration. For this we define the pseudo copula data for $\mathbf{u}^{new}$ as the following $u_{i|i-1}^{new} = h_{U_i|U_{i-1}}\left(u_i^{new}|u_{i-1}^{new}\right)$, $\quad u_{i-1|i}^{new} = h_{U_{i-1}|U_i}\left(u_{i-1}^{new}|u_i^{new}\right) \quad \forall i = 2,, \ldots, p$, where the h-functions $h_{U_i|U_{i-1}}$ and $h_{U_{i-1}|U_i}$ are obtained from the pair copula $c_{U_i,U_{i-1}} \in \hat{\mathcal{B}}(\hat{\mathcal{V}})$, as defined in Section 2.2.

For any $k = 2, \ldots, p-1$ it holds $u_{i|i-k:i-1}^{new} = h_{U_i|U_{i-k};\mathbf{U}_{i-k+1:i-1}}\left(u_{i|i-k+1:i-1}^{new}|u_{i-k|i-k+1:i-1}^{new}\right)$, and similarly $u_{i-k|i-k+1:i}^{new} = h_{U_{i-k}|U_i;\mathbf{U}_{i-k+1:i-1}}\left(u_{i-k|i-k+1:i-1}^{new}|u_{i|i-k+1:i-1}^{new}\right) \ \forall i = 2,, \ldots, p$, where the h-functions $h_{U_i|U_{i-k};\mathbf{U}_{i-k+1:i-1}}$ and $h_{U_{i-k}|U_i;\mathbf{U}_{i-k+1:i-1}}$ are determined from the pair copula $c_{U_i,U_{i-k};\mathbf{U}_{i-k+1:i-1}} \in \hat{\mathcal{B}}(\hat{\mathcal{V}})$. In addition, based on this pseudo-copula data estimated from the fitted Y-vine we introduce the following two matrices, $W \in [0,1]^{p\times p}$ and $W' \in [0,1]^{p\times p}$, as

$$W\left(\mathbf{u}^{new}; \hat{\mathcal{B}}(\hat{\mathcal{V}})\right) \coloneqq \begin{pmatrix} u_1^{new} & u_2^{new} & u_3^{new} & \ldots & u_{p-1}^{new} & u_p^{new} \\ u_{2|1}^{new} & u_{3|2}^{new} & u_{4|3}^{new} & \ldots & u_{p|p-1}^{new} & \\ \vdots & \vdots & \vdots & \ldots & & \\ u_{p-1|1:p-2}^{new} & u_{p-2|2:p-3}^{new} & & & & \\ u_{p|1:p-1}^{new} & & & & & \end{pmatrix}$$

$$W'\left(\mathbf{u}^{new};\hat{\mathcal{B}}(\hat{\mathcal{V}})\right) := \begin{pmatrix} u_1^{new} & u_2^{new} & u_3^{new} & \dots & u_{p-1}^{new} & u_p^{new} \\ u_{1|2}^{new} & u_{2|3}^{new} & u_{3|4}^{new} & \dots & u_{p-1|p}^{new} & \\ \vdots & \vdots & \vdots & \dots & & \\ u_{1|2:p-1}^{new} & u_{2|3:p-2}^{new} & & & & \\ u_{1|2:p}^{new} & & & & & \end{pmatrix}.$$

Using matrices W and W' , we define the following pseudo copula data for $j = 1, 2$, $u_{v_j|1} = h_{V_j|U_1}\left(w|u_1^{new}\right)$ and $u_{1|v_j} = h_{U_1|V_j}\left(u_1^{new}|w\right)$, where $h_{V_j|U_1}$ and $h_{U_1|V_j}$ are estimated from the pair copula $c_{V_j,U_1} \in \hat{\mathcal{B}}$. Further, for $i = 2, \dots, p$, define $u_{v_j|1:i} = h_{V_j|U_i;\mathbf{U}_{1:i-1}}\left(u_{v_j|1:i-1}^{new}|u_{i|1:i-1}^{new}\right)$ and $u_{i|v_j1:i-1} = h_{U_i|V_j;\mathbf{U}_{1:i-1}}\left(u_{i|1:i-1}^{new}|u_{v_j|1:i-1}^{new}\right)$. These h-function are estimated from the pair copula $c_{V_j,U_i;\mathbf{U}_{1:i-1}} \in \hat{\mathcal{B}}$. Then, we also define the matrix $W^2 \in [0,1]^{(p+1)\times 2}$ with $j \in \{1, 2\}$ as

$$W^2\left(w, j; W, W'\right) := \begin{pmatrix} w & w \\ u_{v_j|1} & u_{1|v_j} \\ u_{v_j|12} & u_{2|v_j1} \\ u_{v_j|1:3} & u_{3|v_i12} \\ \vdots & \vdots \\ u_{v_j|1:p} & u_{p|v_j1:p-1} \end{pmatrix}.$$

For a fixed input v_1^{inp} we can evaluate

$$C_{V_1|V_2,\mathbf{U}}\left(v_1^{inp}|z^{new},\mathbf{u}^{new}\right) = h_{V_1|V_2;\mathbf{U}}\left(u_{v_1|1:p}|u_{v_2|1:p}\right), \tag{4.9}$$

at $z = z^{new}$, such that $u_{v_1|1:p}$ is obtained from $W^2\left(w = v_1^{inp}, j = 1; W, W'\right)$ and $u_{v_2|1:p}$ is obtained from $W^2\left(w = z^{new}, j = 2; W, W'\right)$. The h-function $h_{V_1|V_2;\mathbf{U}}$ is estimated from the pair copula $c_{V_1,V_2;\mathbf{U}} \in \hat{\mathcal{B}}(\hat{\mathcal{V}})$.

$c_{V_2|\mathbf{U}}$ is evaluated as

$$c_{V_2|\mathbf{U}}\left(z^{new}|\mathbf{u}^{new}\right) = \frac{c_{V_2,\mathbf{U}}}{c_{\mathbf{U}}} = c_{V_2,U_1}\left(z^{new}, u_1^{new}\right)\prod_{i=2}^{p} c_{V_2,U_i;\mathbf{U}_{1:i-1}}\left(u_{v_2|1:i-1}, u_{i:i-1}^{new}\right), \tag{4.10}$$

where $c_{V_2,U_1}, c_{V_2,U_i;\mathbf{U}_{1:i-1}} \in \hat{\mathcal{B}}(\hat{\mathcal{V}})$ for $i = 1 \dots, p$. Therefore, the integrand in Equation (4.8) can be evaluated with no further calculations from the Y-vine copula, by combining

Equations (4.9) and (4.10), as

$$\begin{aligned} IN\left(z^{new}\right) =& c_{V_2,U_1}\left(z^{new}, u_1^{new}\right) \prod_{i=2}^{p} c_{V_2,U_i;\mathbf{U}_{1:i-1}}\left(u_{v_2|1:i-1}, u_{i|1:i-1}^{new}\right) \\ & \cdot h_{V_1|V_2;\mathbf{U}}\left(u_{v_1|1:p}|u_{v_2|1:p}\right). \end{aligned} \tag{4.11}$$

To summarize, given the integration point $\mathbf{v}^{inp} = (v_1^{inp}, v_2^{inp})^T$, the integrand $IN\left(z^{new}\right)$ at a point $z^{new} \in (0, v_2^{inp})$ conditioned on $\mathbf{u}^{new}$, can be computed using the matrices W, W', $W^2\left(w = z^{new}, j = 2; W, W'\right)$, $W^2\left(w = v_1^{inp}, j = 1; W, W'\right)$ and h-functions obtained from the pair copulas defined by $\hat{\mathcal{B}}(\hat{\mathcal{V}})$. This implies that we can efficiently evaluate the function $C_{V_1,V_2|\mathbf{U}}$ using Equation (4.11).

4.5. Simulation of bivariate data in a Y-vine copula

Since the joint bivariate density $c_{V_1,V_2|\mathbf{U}}\left(v_1, v_2|\mathbf{u}\right)$ can be expressed as the product of $c_{V_1|\mathbf{U}}\left(v_1|\mathbf{u}\right)$ and $c_{V_2|V_1,\mathbf{U}}\left(v_2|v_1, \mathbf{u}\right)$, bivariate samples $\left(v_1(\mathbf{u}), v_2(\mathbf{u})\right) \sim C_{V_1,V_2|\mathbf{U}}\left(\cdot,\cdot|\mathbf{u}\right)$ can be generated as follows.

For the first term, the conditional distribution function $C_{V_1|\mathbf{U}}\left(v_1|\mathbf{u}\right)$ is available analytically, since it is the conditional distribution associated with the D-vine with order $(V_1, U_1, \ldots, U_p)$ from Section 3.2. Its inverse $C_{V_1|\mathbf{U}}^{-1}\left(v_1|\mathbf{u}\right)$ corresponds to the quantile function in Equation (3.1), which is also analytically available. This allows us to get a sample $v_1(\mathbf{u})$, by setting $v_1(\mathbf{u}) = C_{V_1|\mathbf{U}}^{-1}(a_1|\mathbf{u})$ for a value a_1 sampled from a uniform distribution on $[0, 1]$.

For the second term, the conditional density $c_{V_2|V_1,\mathbf{U}}\left(v_2|v_1, \mathbf{u}\right)$ can be obtained using Equation (4.4), by setting the marginal density to 1, i.e. $f_{Y_2}(y_2) = 1$. However, the conditional distribution $C_{V_2|V_1,\mathbf{U}}\left(v_2|v_1, \mathbf{u}\right)$ can be obtained only numerically using integration, i.e.

$$C_{V_2|V_1,\mathbf{U}}\left(v_2|v_1, \mathbf{u}\right) = \int_0^{v_2} c_{V_2|V_1,\mathbf{U}}\left(v_2'|v_1, \mathbf{u}\right) dv_2'.$$

Thus, also the inverse $C_{V_2|V_1,\mathbf{U}}^{-1}\left(v_2|v_1, \mathbf{u}\right)$ can be obtained numerically. Then, a sampled value $v_2(\mathbf{u})$ from $C_{V_2|V_1,\mathbf{U}}\left(\cdot|v_1, \mathbf{u}\right)$ is obtained by setting $v_2(\mathbf{u}) = C_{V_2|V_1,\mathbf{U}}^{-1}\left(a_2|v_1, \mathbf{u}\right)$ for a uniform $[0, 1]$ sampled value a_2. This allows us to get the desired sample $\left(v_1(\mathbf{u}), v_2(\mathbf{u})\right)$ from $C_{V_1,V_2|\mathbf{U}}\left(\cdot,\cdot|\mathbf{u}\right)$ in a step wise fashion shown in Algorithm 1. More details on simulation from general R-vines can be found in Dißmann (2010, Chapter 5).

Algorithm 1: Algorithm for simulating a bivariate sample $(v_1(\mathbf{u}), v_2(\mathbf{u}))$ from $C_{V_1,V_2|\mathbf{U}}(\cdot,\cdot|\mathbf{u})$

Input: N - sample size,
$\mathbf{u}$ - conditioning value of the vector $\mathbf{u}$,
$\mathcal{Y}$ - Y-vine model

for $n = 1, \ldots, N$ **do**

Sample independently a_1^n and a_2^n from $U(0,1)$.

For fixed $\mathbf{U} = \mathbf{u}$, set

$$v_1^n(\mathbf{u}) = C^{-1}_{V_1|\mathbf{U}}(a_1^n|\mathbf{u}),$$

using h-functions from the Y-vine $\mathcal{Y}$.

For fixed $\mathbf{U} = \mathbf{u}$ and $V_1 = v_1^n(\mathbf{u})$, set

$$v_2^n(\mathbf{u}) = C^{-1}_{V_2|V_1,\mathbf{U}}\left(a_2^n|v_1^n, \mathbf{u}\right),$$

using numerical integration and h-functions from the Y-vine $\mathcal{Y}$.

end

return Bivariate samples $(v_1^n(\mathbf{u}), v_2^n(\mathbf{u})) \sim C_{V_1,V_2|\mathbf{U}}(\cdot,\cdot|\mathbf{u})$ for $n = 1, \ldots, N$.

4.6. Implementation

The implementation of the Y-vine regression, together with all the other tools discussed in this chapter is done in the statistical software `R` (R Core Team, 2022). In the estimation of our Y-vine regression model we model the marginals distributions using a nonparametric approach, while we model the pair copulas in a parametric approach, resulting in a semiparametric model. Modeling the marginals as well as the copulas parametrically might cause the resulting fully parametric estimator to be biased and inconsistent if one of the parametric models is misspecified (Noh et al., 2013). Modeling them both using a nonparametric approach leads to a fully nonparametric approach that might overfit the data, because penalization is still an open research topic in the nonparametric case, as noted in Section 3.7. Thus, we opt for a semiparametric approach. The marginals are estimated using a univariate nonparametric kernel density estimator implemented in the `R` package `kde1d` (Nagler and Vatter, 2020), and the pair copulas are fitted using a parametric maximum-likelihood approach with the Akaike Information Criterion penalization (Akaike, 1973a) (AIC) implemented in the `R` package `rvinecopulib` (Nagler and Vatter, 2021). However, a user can also specify to use parametric marginal distributions, nonparametric pair copulas, different penalizations on the family selection, as the BIC penalization (Schwarz, 1978) and so on. Also, instead of the adjusted conditional log-likelihood, we can use the AIC/BIC penalized adjusted conditional log-likelihood as selection criteria in the forward selection of predictors.

4.7. Data application

As an application to real data we consider the Seoul weather data set, which contains two dependent responses, daily minimum and maximum air temperature. The data originates from the UCI machine learning repository (Dua and Graff, 2019), it can be downloaded from https://archive.ics.uci.edu/ml/datasets/Bias+correction+of+numerical+prediction+model+temperature+forecast and was first studied by Cho et al. (2020). It contains daily data for 25 weather stations in Seoul, South Korea between June 30th and August 30th in the period 2013-2017. Cho et al. (2020) use it for enhancing next-day maximum and minimum air temperature forecasts based on the Local Data Assimilation and Prediction System (LDAPS) model.

To illustrate the proposed vine based bivariate regression model, we consider the station located in central Seoul (station 25) and we model the temporal dependence in the responses, by considering the present minimum and maximum air temperature (including two lagged variables into the regression model) when modeling next day values. Disregarding geographical markers and precipitation measurements, we are left

with a data set containing two response variables and 13 continuous predictors, with 307 data points representing summer days of the years 2013 to 2017. Table 4.1 gives a variable description, the unit of measurement and the range of possible values for the 2 predictors Next_Tmax or T_max, Next_Tmin or T_min and the 13 possible continuous predictors we consider. We divide the data set into a training and testing set, consisting of 246 data points from 2013-2016, and 61 data points from 2017, respectively.

In Figure 4.4 shown are the empirical marginally normalized contour plots for pairs of variables from the training set. On this lower diagonal, any deviance from elliptical shapes indicates a non-Gaussian dependence structure in the data (see Section 3.8 of Czado (2019) for a precise definition). This is the case in almost all marginally normalized contour plots and it supports our non-Gaussian approach with flexible vine copulas over any other modeling approach that assumes Gaussianity. On the upper diagonal, we see a scatter plot of the estimated u-data together with the corresponding pairwise empirical Kendall's $\hat{\tau}$. There are many large values of the Kendall's tau estimated, between pairs of the possible predictors, like a value of 0.6 between LDAPS_CC1 and LDAPS_CC2 or a value of 0.59 between LDAPS_RHmin and LDAPS_CC2; and also between the responses and the possible predictors.

The highest estimated pairwise dependence for the response Next_Tmax is with the predictor LDAPS_Tmax_lapse with a Kendall's tau value of 0.68. The highest estimated dependence for the response Next_Tmin is with the predictor LDAPS_Tmin_lapse with a Kendall's tau value of 0.75. This is an expected result and these predictors are suggested to be included the model for enhancing the forecast using the LDAPS models. Also, the response variables, minimum and maximum air temperature, are not independent of each other either and are expected to rise and fall together. This dependence is emphasized by an estimated pairwise Kendall's τ value of 0.46. Thus, using vine copulas we can efficiently model and capture these high non-Gaussian dependencies between pairs of variables.

In the estimation of the Y-vine regression model we model the marginals distributions using a nonparametric approach, while we model the pair copulas in a parametric way using the AIC penalization criteria for pair copula family selection, resulting in a semiparametric model. Further, the selection criteria for the forward selection of predictors is a AIC penalized adjusted conditional log likelihood.

The automatically chosen order of the predictors in the fitted Y-vine regression model $\hat{\mathcal{Y}}$ is given by

$$O(\hat{\mathcal{Y}}) = (\texttt{ LDAPS\_Tmin\_lapse, LDAPS\_Tmax\_lapse, LDAPS\_CC1, LDAPS\_WS,}$$
$$\texttt{Present\_Tmin, LDAPS\_RHmax, LDAPS\_CC3, LDAPS\_LH, Present\_Tmax}).$$

It orders the predictors by their influence over the two responses. Also, only 9 out of 13 possible predictors are chosen to be in the model. The 4 non-influential predictors,

based on the Y-vine model are LDAPS_CC2, LDAPS_CC4, LDAPS_RHmax and solar radiation. More details on the fitted pair copulas selected by the Y-vine regression model is given in the Appendix A.3.2.

The fitted pair copula between the responses given the 9 chosen predictors, $\hat{c}_{V_1,V_2;\mathbf{U}}$ is a Joe copula with an estimated Kendall's τ of 0.09. This implies that after the effect of the predictors is adjusted in the model, there is almost no dependence between the responses. All of this, implies the benefits of having a model able to consider two dependent responses and model their joint conditional distribution using a set of highly dependent variables, with non-Gaussian dependence structure. We dive deeper into the analysis on this data set in the next Chapter 5, Section 5.5.

Table 4.1.: Variable description, the unit of measurement and the range of possible values the considered variables can take.

Variable name	Description(unit)	Range
Next_Tmax	The next-day maximum air temperature $(^{\circ}C)$	17.4 to 38.9
Next_Tmin	The next-day minimum air temperature $(^{\circ}C)$	11.3 to 29.8
Present_Tmax	Maximum air temperature between 0 and 21 h on the present day $(^{\circ}C)$	20 to 37.6
Present_Tmin	Minimum air temperature between 0 and 21 h on the present day $(^{\circ}C)$	11.3 to 29.9
LDAPS_RHmin	LDAPS model forecast of next-day minimum relative humidity (%)	19.8 to 98.5
LDAPS_RHmax	LDAPS model forecast of next-day maximum relative humidity (%)	58.9 to 100
LDAPS_Tmax_lapse	LDAPS model forecast of next-day maximum air temperature applied lapse rate $(^{\circ}C)$	17.6 to 38.5
LDAPS_Tmin_lapse	LDAPS model forecast of next-day minimum air temperature applied lapse rate $(^{\circ}C)$	14.3 to 29.6
LDAPS_WS	LDAPS model forecast of next-day average wind speed (m/s)	2.9 to 21.9
LDAPS_LH	LDAPS model forecast of next-day average latent heat flux (W/m^2)	-13.6 to 213.4
LDAPS_CC1	LDAPS model forecast of next-day 1st 6-hour split average cloud cover (0-5 h) (%)	0 to 0.97
LDAPS_CC2	LDAPS model forecast of next-day 2nd 6-hour split average cloud cover (6-11 h) (%)	0 to 0.97
LDAPS_CC3	LDAPS model forecast of next-day 3rd 6-hour split average cloud cover (12-17 h) (%)	0 to 0.98
LDAPS_CC4	LDAPS model forecast of next-day 4th 6-hour split average cloud cover (18-23 h) (%)	0 to 0.97
Solar radiation	Daily incoming solar radiation (wh/m^2)	4329.5 to 5992.9

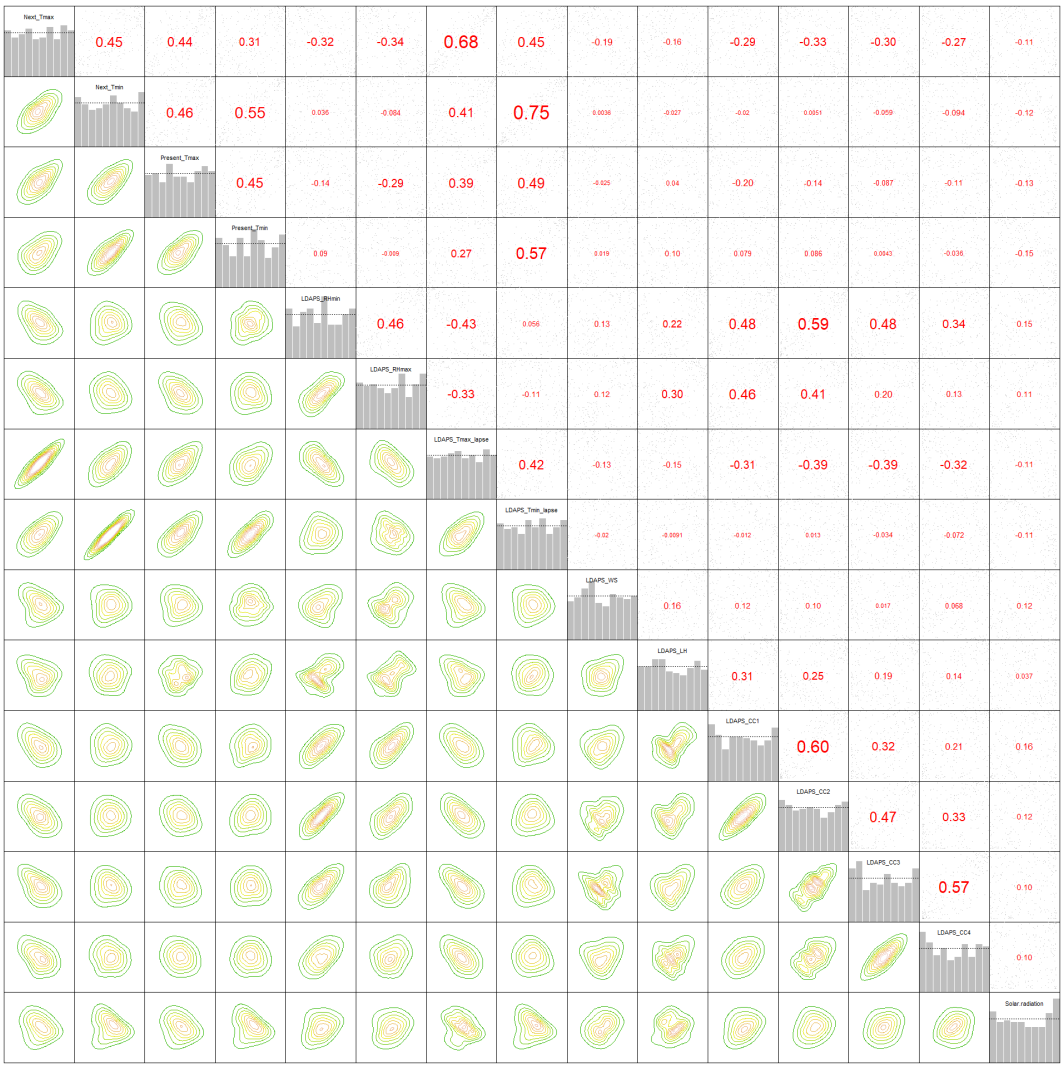


Figure 4.4.: Lower diagonal: marginally normalized contour plots, upper diagonal: pairwise scatter plots with the associated empirical Kendall's $\hat{\tau}$ values and on the diagonal: histograms of the u-data.

4.8. Conclusion and outlook

We studied the problem of two response joint conditional distribution function estimation using a very flexible class of models, vine copulas. They are multivariate distributions constructed from bivariate blocks (pair copulas) using conditioning. We develop a novel vine tree structure, the Y-vine tree structure, that is suitable for a regression problem containing bivariate response variables. Also, a forward selection of predictors gives the best suitable fitted Y-vine, by ordering the predictors based on their joint influence over the two responses. In addition, the Y-vine tree structure enables an easy way of obtaining the bivariate conditional density. This way a joint analysis of the dependence structure of the responses given the predictors is possible. This is a significant result especially when dealing with responses that are not (conditionally) independent. We also develop a prediction method for the joint conditional distribution function using the Y-vine regression. This enables us to not only jointly model, but also predict the joint conditional distribution, given a set of observed predictors. This way, we can study how the conditional distribution changes for different observations of the predictors and do a corresponding risk analysis (more details on this topic are discussed later in Section 5.5).

Additionally, simulation of bivariate data from a Y-vine copula model is available. This is one of the very few methods that can generate bivariate data, conditioned on a specific value of the predictors, with possible tail dependence and asymmetric dependence structures. We apply our proposed model to a real life data set containing a bivariate response, minimal and maximal daily temperature. We analyse the data with our new approach for dependent responses and provide a joint vine copula model for the two responses.

For future possible applications we think of adding a spatial and/or temporal component to our Y-vine based regression. It would be interesting to see how the response dependence changes when the spacial and/or temporal dependence component is also accounted for, but that is out of the scope of this thesis. The standard lack of ability of copula based models to include discrete variables is also an ongoing research topic, as also discussed in Section 3.7. Further, it would be an area of possible extension to include a mix of continuous and discrete response variable. Also, applications of different vine structures and variable selection methods, and subsequent comparisons of the performance, are left for further investigation and are expected to be heavily data specific problems.

In addition, we can use the Y-vine tree structure for testing of conditional independence between two variables given a set of conditioning variables. The Y-vines provide a very symmetric treatment of the two variables whose conditional independence is being tested. Using this way of testing for conditional independence we do not

need any asymptotic results, as we use a very flexible modeling approach. A similar approach was proposed in Bauer and Czado (2016) using R-vines, for non-Gaussian conditional independence testing in continuous Bayesian networks. However, their approach needed, possibly high dimensional, integration for determining the required conditional distribution function and thus, is not applicable for large network problems. In contrast, we expect our approach to remain tractable in large networks.

A possible further extension is to develop a similar new tree structure that will be suitable for more than two responses. The Y-vine tree structure, resembles a C-vine tree structure between the responses and a D-vine tree structure between the predictor variables. Thus, we can extend the C-vine part between the responses to include any number of responses and still the conditional distribution of the responses given the predictors should be available in no computationally expensive manner. However, usually there is no need for more than two responses in real life data, but it is an interesting further research area, if a need for such modeling is to appear.

5. Bivariate unconditional and conditional level curves and quantile curves

Parts of Chapter 5 are similar to the publication Tepegjozova and Czado (2022). However, Sections 5.4 and 5.5 contain new material.

5.1. Introduction

Despite the great attention univariate quantiles have received, the extension to multivariate response quantiles is not trivial nor well-defined. Several theoretical notions of multivariate quantiles have been introduced, but there is no consensus which one is the corresponding generalization of the univariate quantiles. These include geometric quantiles based on halfspace depth contours with different concepts of statistical depth (e.g. see Tukey (1975), Chaudhuri (1996), Hallin et al. (2010), Chernozhukov et al. (2017)), vector quantiles (see Carlier et al. (2016) and Carlier et al. (2017)), spatial quantiles (see Abdous and Theodorescu (1992)).

Copula based models are known to be excelling at modeling tail events and asymmetric dependencies. Bivariate quantiles arising from copula based models are advantageous in assessing the joint risk of failure or occurrence of two events. One example of specifically bivariate quantiles being applied is the work of Chebana and Ouarda (2011) in flood analysis. The risk of flood is studied through a joint analysis of flood peaks and flood volumes, using level sets of bivariate copulas. Requena et al. (2013) present a similar analysis of the risk of flood, focusing on hydrologic dam designs. Additionally, the authors use the same approach for a joint analysis of reservoir volume and spillway crest length as indicators for the risk of dam overtopping. A multivariate risk of failure analysis based on copulas is presented in Salvadori et al. (2015) for structural failure assessment in engineering. An application in financial mathematics can be found in Di Bernardino and Prieur (2014). The authors propose estimation of a tail event risk measure based on multivariate level sets of copulas. However, in all these approaches there is modelling of two responses with a copula, disregarding any explanatory variables or predictors. It would be even more beneficial to jointly model two response variables, taking into account the influence of a set of possible predictors.

Thus, to fill in this gap, we focus on bivariate conditional quantiles and their estimation using a very flexible vine copula model, the Y-vine regression, introduced in Chapter 4.

In the examples we provided, the notion of multivariate quantiles is linked to the notion of level curves of multivariate distribution functions. However, in Belzunce et al. (2007) it is noted that for some $\alpha \in (0,1)$, the α-th level curve of a bivariate distribution function does not provide exact coverage probability. Thus, we look at the extension of the univariate quantiles (Koenker and Bassett, 1978), defined as cut points that divide the range of a univariate distribution into intervals with given probabilities, to bivariate quantiles having this exact property.

Thus, there is still a need for: $(1.)$ a valid definition of (unconditional and conditional) level curves of (vine) copula based derived distributions and their connection to level curves defined on the x-scale; $(2.)$ a numerical method for obtaining the bivariate (unconditional and conditional) level curves from the estimated (vine) copula based model; $(3.)$ a valid definition of (unconditional and conditional) bivariate quantiles linked to the usage of (vine) copulas.

For $(1.)$ we extend the definitions of bivariate unconditional and conditional level curves in terms of a (vine) copula based distribution function. For $(2.)$ we develop a numerical method to evaluate the bivariate unconditional and conditional level curves. Further, we illustrate the bivariate unconditional level curves for known bivariate copula distributions and the bivariate conditional level curves for a 3-dimensional vine copula distribution and analyse the performance of our numerical estimation algorithm. For $(3.)$, we show results about the coverage probability of the level curves and suggest a new definition for bivariate quantiles, as adjusted level curves that have exact coverage probabilities. Based on the estimated quantiles, we can construct bivariate confidence regions, which are a generalization of univariate confidence intervals. They are used to locate parts of a distribution with high density values. Such confidence regions can be also effective for visualizing trends, patterns and outliers (Korpela et al., 2014; Korpela et al., 2017; Guilbaud, 2008). In addition, we illustrate our confidence regions and compare them to alternative ways of construction of confidence regions, assuming independence between the responses. We highlight the advantages of our modeling approach and its usability in data analysis.

5.2. Bivariate level curves

The notion of multivariate quantiles is linked to the notion of multivariate level curves, so we start by defining and exploring the level curves of bivariate unconditional and conditional distribution functions.

5.2.1. Bivariate unconditional level curves

Let Y_1 and Y_2 be two continuous random variables with observed values y_1, y_2 and a joint distribution function $F_{Y_1,Y_2}(y_1, y_2)$.

Definition 3 *The bivariate level curve for continuous random variables Y_1, Y_2 at level $\alpha \in (0,1)$ is a curve in $\mathbb{R}^2$ defined by the set*

$$\begin{aligned} Q_\alpha^Y :=& \{(y_1, y_2) \in \mathbb{R}^2 \; ; \; F_{Y_1,Y_2}(y_1, y_2) = \alpha\} \\ =& \{(y_1, y_2) \in \mathbb{R}^2 \; ; \; \mathbb{P}_{Y_1,Y_2}(Y_1 \leq y_1, Y_2 \leq y_2) = \alpha\}. \end{aligned}$$

Define the probability integral transforms of the random variable Y_j as $V_j := F_{Y_j}(Y_j)$, with corresponding observed values $v_j := F_{Y_j}(y_j)$ for $j = 1, 2$. Applying Sklar's Theorem (Equation (2.1)) to the joint distribution function of Y_1, Y_2, we obtain $F_{Y_1,Y_2}(y_1, y_2) = C(F_{Y_1}(y_1), F_{Y_2}(y_2)) = C_{V_1,V_2}(v_1, v_2)$. So we can rewrite the bivariate level curves from Definition 3 in terms of copulas as, $Q_\alpha^Y = \{(F_{Y_1}^{-1}(v_1), F_{Y_2}^{-1}(v_2)) \in \mathbb{R}^2 \; ; \; C_{V_1,V_2}(v_1, v_2) = \alpha, \quad v_1, v_2 \in (0,1)\}$. We can also define the bivariate level curves of the probability integral transformed variables on the unit square $[0,1]^2$.

The bivariate level curves at $\alpha \in (0,1)$ for the continuous random variables Y_1, Y_2 with random PITs V_1, V_2 is a curve in $[0,1]^2$ defined by the set

$$\begin{aligned} Q_\alpha^V :=& \{(v_1, v_2) \in [0,1]^2 \; ; \; C_{V_1,V_2}(v_1, v_2) = \alpha\} \\ =& \{(v_1, v_2) \in [0,1]^2 \; ; \; \mathbb{P}(V_1 \leq v_1, V_2 \leq v_2) = \alpha\}. \end{aligned} \tag{5.1}$$

The difference between Q_α^Y and Q_α^V is that Q_α^Y is defined on $\mathbb{R}^2$, while Q_α^V is defined on $[0,1]^2$. The connection between the two is given as $Q_\alpha^Y = \{(F_{Y_1}^{-1}(v_1), F_{Y_2}^{-1}(v_2)) \in \mathbb{R}^2 \; ; \; (v_1, v_2) \in Q_\alpha^V\}$. Sklar's Theorem implies that a transformation of the bivariate level curves between the x- and u-scale is obtained using inverses of the univariate marginal distributions $F_{Y_1}^{-1}, F_{Y_2}^{-1}$, rather that the bivariate joint distribution F_{Y_1,Y_2}.

5.2.2. Bivariate conditional level curves

Let $\mathbf{Y} = (Y_1, Y_2)^T$ and $\mathbf{X} = (X_1, \ldots, X_p)^T$, $p \geq 1$, be two continuous random vectors, with the corresponding marginal distribution functions $Y_j \sim F_{Y_j}$, for $j = 1, 2$ and $X_i \sim F_{X_i}$, for $i = 1, \ldots, p$. Our interest is the bivariate conditional level curves of $\mathbf{Y}$ given $\mathbf{X} = \mathbf{x}$. Denote the conditional distribution function of $Y_1, Y_2 | \mathbf{X} = \mathbf{x}$ as $F_{Y_1,Y_2|\mathbf{X}}(y_1, y_2|\mathbf{x})$.

Definition 4 *The bivariate conditional level curves for a continuous bivariate vector $\mathbf{Y} = (Y_1, Y_2)^T$ given the outcome of a p-dimensional random vector ($p \geq 1$), $\mathbf{X} = \mathbf{x}$ at level $\alpha \in (0,1)$ is a curve in $\mathbb{R}^2$ defined by the set*

$$\begin{aligned} Q_\alpha^Y(\mathbf{x}) :=& \{(y_1, y_2) \in \mathbb{R}^2 \; ; \; F_{Y_1,Y_2|\mathbf{X}}(y_1, y_2|\mathbf{x}) = \alpha\} \\ =& \{(y_1, y_2) \in \mathbb{R}^2 \; ; \; \mathbb{P}_{Y_1,Y_2|\mathbf{X}}(Y_1 \leq y_1, Y_2 \leq y_2|\mathbf{X} = \mathbf{x}) = \alpha\}. \end{aligned}$$

In order to derive the level curves in terms of copulas, we need to express the conditional distribution of $Y_1, Y_2|\mathbf{X}$ in terms of a copula distribution function. For this we use the results from Section 4, Proposition 1. Thus, the bivariate level curve $Q_\alpha^Y(\mathbf{x})$ can be rewritten as $Q_\alpha^Y(\mathbf{x}) = \{(F_{Y_1}^{-1}(v_1), F_{Y_2}^{-1}(v_2)) \in \mathbb{R}^2 \; ; \; C_{V_1,V_2|\mathbf{U}}(v_1, v_2|\mathbf{u}) = \alpha, \quad v_1, v_2 \in (0,1)\}$, where $\mathbf{u} = \left(u_1, \ldots, u_p\right)^T$ are realizations of the random vector $\mathbf{U} = \left(U_1, \ldots, U_p\right)^T$.

Similarly, we define the bivariate conditional level curves of the probability integral transformed variables on the unit square $[0,1]^2$. The bivariate conditional level curves at $\alpha \in (0,1)$ for the continuous random variables Y_1, Y_2 with random PITs V_1, V_2 given the outcome of the random vector $\mathbf{X} = \mathbf{x}$, with PITs $\mathbf{U} = \mathbf{u}$ is a curve in $[0,1]^2$ defined by the set

$$\begin{aligned} Q_\alpha^V(\mathbf{u}) :=& \{(v_1, v_2) \in [0,1]^2 \; ; \; C_{V_1,V_2|\mathbf{U}}(v_1, v_2|\mathbf{u}) = \alpha\} \\ =& \{(v_1, v_2) \in [0,1]^2 \; ; \; \mathbb{P}_{V_1,V_2|\mathbf{U}}(V_1 \leq v_1, V_2 \leq v_2|\mathbf{U} = \mathbf{u}) = \alpha\}. \end{aligned} \tag{5.2}$$

However, note that the bivariate conditional and unconditional level curves as defined in Equation (5.1) and (5.2), do not have the property that the α-th level curve separates the lowest $\alpha \times 100$ percent of the observations from the remaining $(1-\alpha) \times 100$ percent of the observations, as discussed later in Section 5.4.2.

5.3. Numerical evaluation of bivariate level curves

5.3.1. Algorithms

Let $C(a,b)$ be a bivariate (conditional) distribution defined on the unit square $[0,1]^2$ with no closed form solution for the bivariate level curve. Assume that $C(a,b)$ can be evaluated at all points $(a,b) \in [0,1]^2$. The goal is to obtain a numerical estimate of the set defining the (conditional) bivariate level curves, given in Equation (5.1) (or Equation (5.2) for the conditional case). Given a granularity parameter $m \in \mathbb{N}^+$ and $\alpha \in (0,1)$ we employ the following procedure:

1. The set $M = \{w_1, \ldots, w_m\}$ is initialized as m equidistant points in the interval $[0,1]$.

2. We define the set of lines L as follows:

$$L = \{((0,0), (w_i, 1)) \,|\forall w_i \in M\} \cup \{((0,0), (1, w_i)) \,|\forall w_i \in M\}.$$

3. Each line $l_s \in L$ is treated as a separate optimization problem and a line search procedure is employed to obtain the point $(a_s, b_s) \in l_s$ for which $C(a_s, b_s) = \alpha$. Consider any line l_s and two points on the line, denoted as (a_1, a_2) and (b_1, b_2),

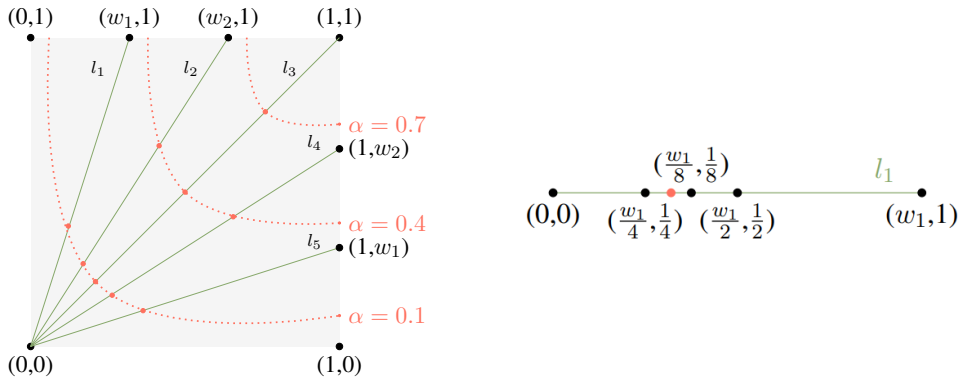


Figure 5.1.: Graphical representation of the numerical estimation procedure.

such that $a_1 \leq b_1$ and $a_2 \leq b_2$. Then, $P(V_1 \leq a_1, V_2 \leq a_2) \leq P(V_1 \leq b_1, V_2 \leq b_2)$ holds. This follows since C is a bivariate distribution function and is continuous. Given the condition that the margins of $C(a_s, b_s)$ are monotonically increasing, this implies that values of $C(\cdot,\cdot)$, along any line l_s starting from $(0,0)$, are increasing.

4. For each $l_s \in L$ a line search is guaranteed to converge to a solution, if $C(w_{s_1}, w_{s_2}) \geq \alpha$, where (w_{s_1}, w_{s_2}) is the endpoint of line l_s. In the case $C(w_{s_1}, w_{s_2}) < \alpha$ there is no solution on the line l_s. (The same arguments hold for the conditional copula distribution function as well.)

5. Finally, the remaining points (a_s, b_s) for $s = 1, \ldots, 2m$ for which a solution exists, are smoothed to obtain a curve representing an estimate of the (conditional) bivariate level curve for a given α.

The algorithms used for this numerical evaluation of bivariate level curves are given in Algorithm 2 and 3. The bivariate distribution function $C(a, b)$ is equivalent to $C_{V_1,V_2}(v_1, v_2)$ (or $C_{V_1,V_2|\mathbf{U}}(v_1, v_2|\mathbf{u})$) if unconditional (or conditional) bivariate level curves are evaluated.

In Figure 5.1 we show a graphical representation of the numerical procedure for evaluating bivariate level curves. In the left panel, on the unit square $[0,1]^2$ shown are 5 exemplary lines, $l_1 = ((0,0),(w_1,1))$, $l_2 = ((0,0),(w_2,1))$, $l_3 = ((0,0),(1,1))$, $l_4 = ((0,0),(1,w_2))$, $l_5 = ((0,0),(1,w_1))$ on which a line search is employed to find the pair (a^*, b^*) such that $C(a^*, b^*) = \alpha$ holds. The dotted lines represent the solution of the line search, in our case, the bivariate level curves for $\alpha = 0.1, 0.4, 0.7$. In the right panel, we illustrate the binary line search for an exemplary line, say line $l_1 = ((0,0),(w_1,1))$. First, the desired function is evaluated at the middle point of the line l_1, at $C(\frac{w_1}{2},\frac{1}{2})$. Here

Algorithm 2: PseudoInverse

Input: **m** - granularity parameter (default = 1000),
eq - function based on which $C(u,v)$ is to be evaluated,
err - accuracy of algorithm,
α - alpha level

Initialization:

$$M = \bigcup_{i=1}^{m} \left\{ \frac{i}{m} \right\},$$

$$L = \{line((0,0),(q_1,q_2))\}, \quad (q_1,q_2) \in \{(w_i,1) \,|\forall w_i \in M\} \cup \{(1,w_i) \,|\forall w_i \in M\},$$

$$Points = \emptyset.$$

for $l_s \in L$ **do**
  if $C(q_1,q_2) \geq \alpha$ **then**
    $point = \text{BinaryLineSearch}(l_s, \mathbf{eq}, err, \alpha)$
    $Points = Points \cup point$
  else
    $Points = Points$
  end
end
return *Points*

Algorithm 3: BinaryLineSearch

Input: **l** - line defined by two coordinates,
eq - function based on which $C(u, v)$ is to be evaluated,
err - accuracy of algorithm,
α - alpha level

Initialization:

Introduce notation $\mathbf{l} = line\left(p_{start} = (0,0), p_{end} = (w_i, 1)\right)$.

$$evl = \mathbf{eq}\left(\frac{p_{start} + p_{end}}{2}\right)$$
$$diff = \alpha - evl$$

while $diff > err$ **do**
  if $diff > 0$ **then**
    $p_{start} = \frac{p_{start}+p_{end}}{2}$
  else
    $p_{end} = \frac{p_{start}+p_{end}}{2}$
  end
  $evl = \mathbf{eq}\left(\frac{p_{start}+p_{end}}{2}\right)$
end
return $\frac{p_{start}+p_{end}}{2}$

it holds $C(\frac{w_1}{2},\frac{1}{2}) > \alpha$, so the middle point of the line $((0,0),(\frac{w_1}{2},\frac{1}{2}))$ is evaluated next, $C(\frac{w_1}{4},\frac{1}{4})$. Then, it holds $C(\frac{w_1}{4},\frac{1}{4}) < \alpha$, so the middle point of the line $((\frac{w_1}{4},\frac{1}{4}),(\frac{w_1}{2},\frac{1}{2}))$ is evaluated next, $C(\frac{3w_1}{8},\frac{3}{8})$. Here $C(\frac{3w_1}{8},\frac{3}{8}) > \alpha$, so we consider the middle point of the line $((\frac{w_1}{4},\frac{1}{4}),(\frac{3w_1}{8},\frac{3}{8}))$ next and iteratively continue until the algorithm converges to a solution. The red dot (star), say (a^*, b^*) is the point at which $C(a^*, b^*) = \alpha$.

5.3.2. Illustration of bivariate level curves on the unit square

We illustrate the bivariate unconditional level curves for known pair copula distributions and the bivariate conditional level curves for a 3-dimensional vine structure. They correspond to the case of no predictors or 1 predictor in a regression setting, respectively.

In Figure 5.2 we explore plots of the unconditional level curves on the unit square for the bivariate Gauss, Student-t, Clayton and Gumbel copulas (rows) with different strengths of dependency, expressed through Kendall's τ (defined in Section 2.1.1), with $\tau = 0.25, 0.5, 0.75$ (columns). The level curves can be obtained in an analogous way for any other copula family, given that the margins are strictly monotonic, othervise level sets can be obtained. The theoretical level curves of a bivariate random vector $(V_1, V_2)^T$ with bivariate distribution function $C_{V_1,V_2}(v_1, v_2; \theta)$ and a parameter θ are derived using Equation (5.1) for a given α and are depicted with thick black lines.

Further, we estimate the bivariate level curves for the given pair copulas. For this we simulate data from the given copula and based on the simulated data, a pair copula is estimated. The gray points are 300 data points simulated from the given copulas. Subsequently, level curves are evaluated and plotted. The coloured lines represent the corresponding estimated level curves. In the Appendix B.1, we give a detailed description on how the theoretical and the estimated level curves are obtained for each of the four copula families. The panels of Figure 5.2 showcase bivariate level curves at $\alpha = 0.05, 0.1, 0.25, 0.5, 0.75, 0.90, 0.95$.

Differences can be spotted between estimated and theoretical level curves only for the Gumbel level curves, in the case when Kendall's $\tau = 0.25$. In all other cases, differences between the theoretical and estimated level curves are not visible. When it comes to differences in the level curves for different copula families, the Clayton copula level curve has a significantly smaller surface below the $\alpha = 0.05$ level curve caused by its heavy lower tail (expected realizations are closer to the lower diagonal as compared to a lighter lower tail copula) compared to the other copula families at the $\alpha = 0.05$ level curve. On the other hand, the heavy upper tail of the Gumbel copula is causing a bigger surface above the $\alpha = 0.95$ level curve compared to the Clayton copula. In contrast, the Gaussian copula has no tails at all and the Student-t copula has a symmetric tail dependence governed by a single parameter. Their surface below the $\alpha = 0.05$ level curve is greater than the corresponding surface in the lower heavy-tailed

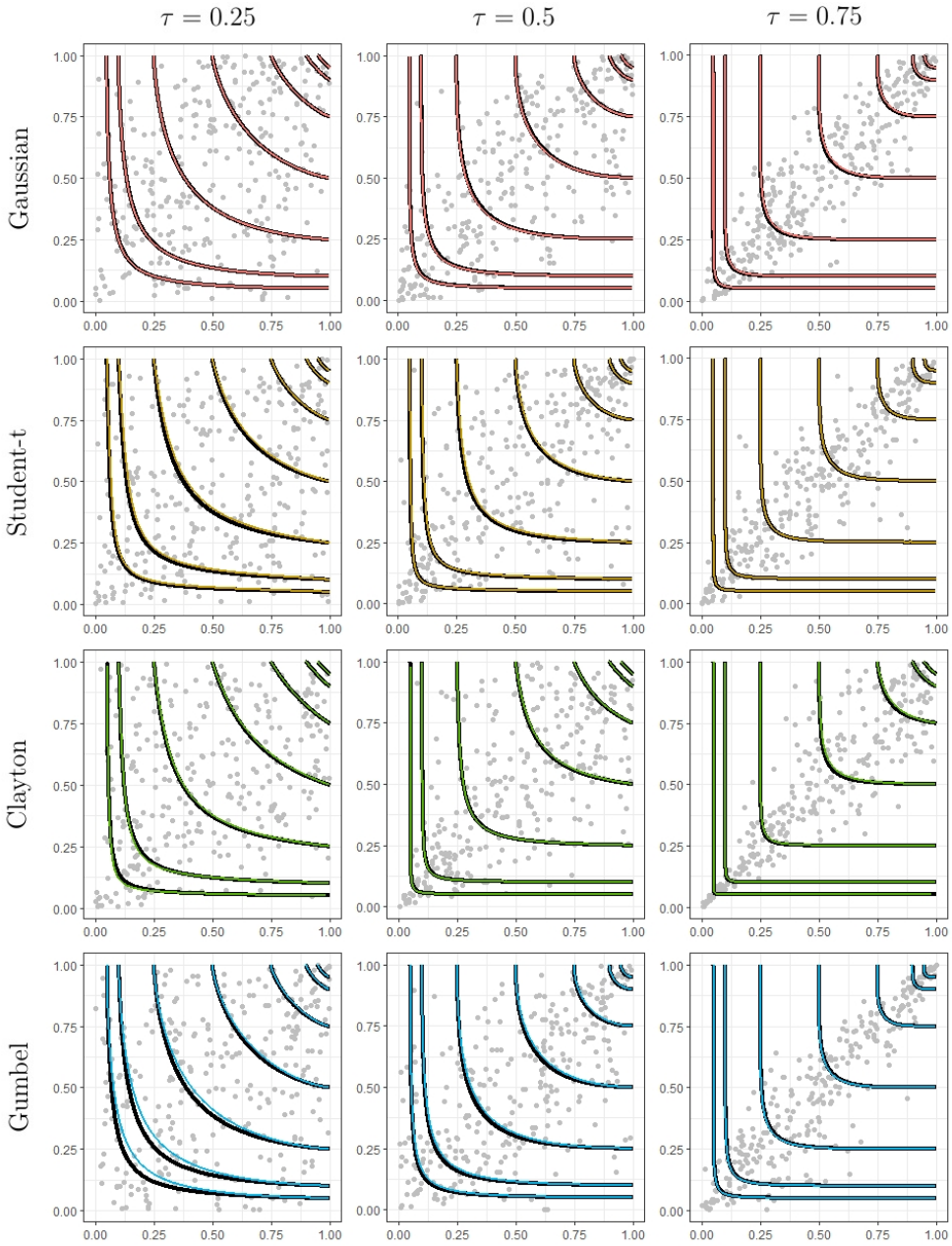


Figure 5.2.: x-axis: V_1, y-axis: V_2. Gray points: simulated data from copula (n=300). Black curves: theoretical level curves. Colored curves: estimated level curves. Depicted are level curves at $\alpha = 0.05, 0.1, 0.25, 0.5, 0.75, 0.90, 0.95$ (left bottom to right top in each panel) for Gaussian, Student-t ($df = 5$), Clayton and Gumbel copulas (top to bottom) and $\tau = 0.25, 0.5, 0.75$ (left to right).

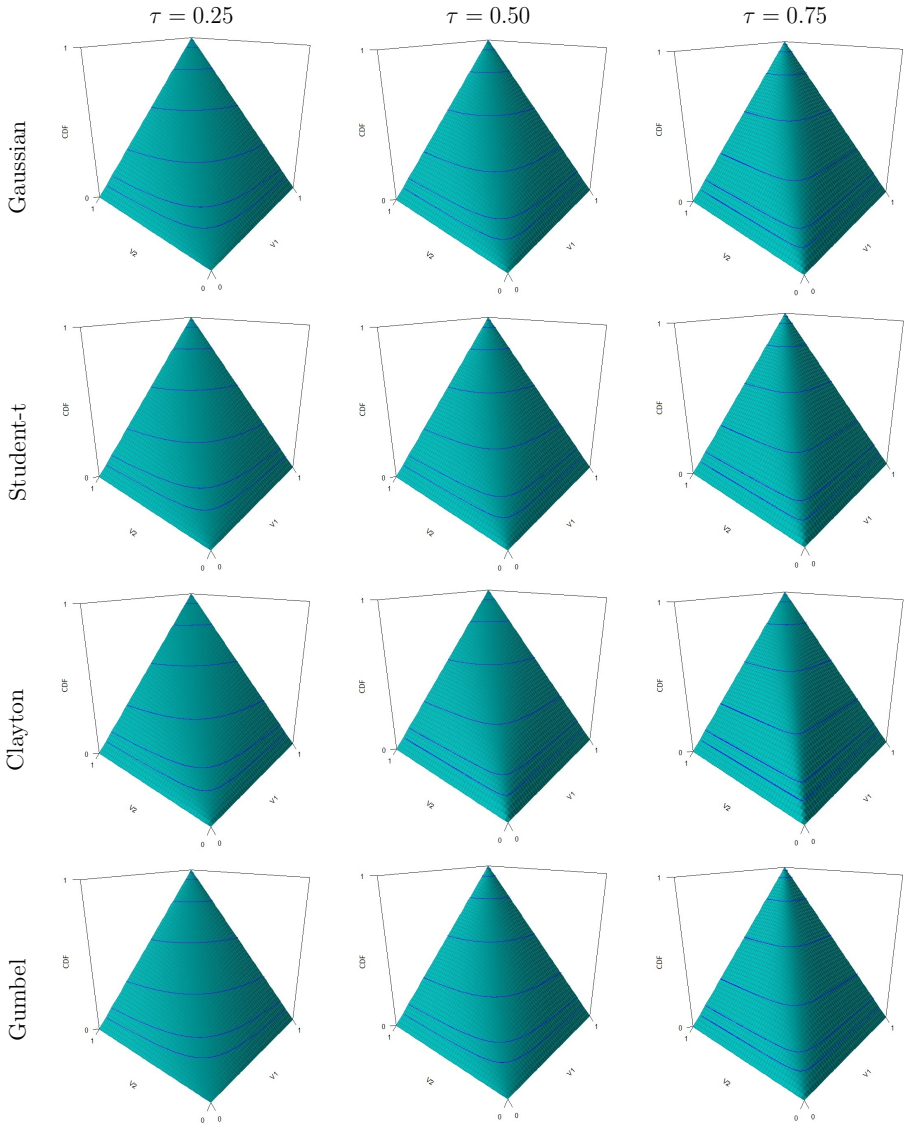


Figure 5.3.: A 3-dimensional plot of bivariate copula distributions with theoretical level curves at $\alpha = 0.05, 0.1, 0.25, 0.5, 0.75, 0.90, 0.95$. Shown are Gaussian, Student-t ($df = 5$), Clayton and Gumbel copulas (top to bottom) and $\tau = 0.25, 0.5, 0.75$ (left to right).

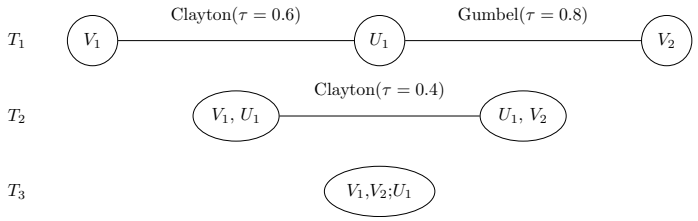


Figure 5.4.: Vine tree sequence of $\mathcal{D}_3$ with the pair copula families and Kendall's τ corresponding to parameters.

Clayton copula, and the surface above the $\alpha = 0.95$ level curve is smaller than the upper heavy-tailed Gumbel copula. Considering the $\alpha = 0.5$ level curve, the greatest surface below it has the Gumbel copula, due to it upper heavy tail, and the smallest surface below the $\alpha = 0.5$ level curve has the Clayton copula, again due to the heavy lower tail. This holds for all Kendall's τ values. Also, as the dependence between the variables increases, the data is more centered around the diagonal, so the curves have sharper curvature around the diagonal. Further, in Figure 5.3 we show the associated bivariate distribution functions in a 3-dimensional plot in which the theoretical level curves are shown at given α levels.

Next we consider conditional bivariate level curves arising from a 3-dimensional regular vine distribution $\mathcal{D}_3$. Let $(V_1, V_2, U_1)^T \sim \mathcal{D}_3$ with vine tree sequence and pair copulas of $\mathcal{D}_3$ given by Figure 5.4. The corresponding parameters to the copulas are $\theta_{V_1,U_1} = 3$ $(\tau = 0.6), \theta_{U_1,V_2} = 5$ $(\tau = 0.8$)and $\theta_{V_1,V_1;U_1} = 1.33$ $(\tau = 0.4)$. To obtain theoretical level curves from $\mathcal{D}_3$ we employ the following procedure. First, to evaluate $C_{V_1,V_2|U_1}$ at a specific point $(\tilde{v}_1, \tilde{v}_2)$ conditioned on $U_1 = \tilde{u}_1$ we use

$$\begin{aligned} C_{V_1,V_2|U_1}\left(\tilde{v}_1, \tilde{v}_2|\tilde{u}_1\right) &= \int_0^{\tilde{v}_1}\int_0^{\tilde{v}_2} c_{V_1,V_2|U_1}\left(v_1', v_2'|\tilde{u}_1\right) dv_2' dv_1' \\ &= \int_0^{\tilde{v}_1}\int_0^{\tilde{v}_2} c_{V_2|U_1}\left(v_2'|\tilde{u}_1\right) \cdot c_{V_1|V_2,U_1}\left(v_1'|v_2', \tilde{u}_1\right) dv_2' dv_1' \\ &= \int_0^{\tilde{v}_2} c_{V_2|U_1}\left(v_2'|\tilde{u}_1\right) \left[\int_0^{\tilde{v}_1} c_{V_1|V_2,U_1}\left(v_1'|v_2', \tilde{u}_1\right) dv_1'\right] dv_2' \\ &= \int_0^{\tilde{v}_2} c_{V_2,U_1}\left(v_2', \tilde{u}_1\right) C_{V_1|V_2,U_1}\left(\tilde{v}_1|v_2', \tilde{u}_1\right) dv_2'. \end{aligned} \tag{5.3}$$

We can also condition on V_1 instead of V_2. The corresponding conditional level curve is evaluated using the numerical evaluation procedure from Section 5.3 and Equation (5.3).

We are also interested in the estimated conditional level curves. To obtain them,

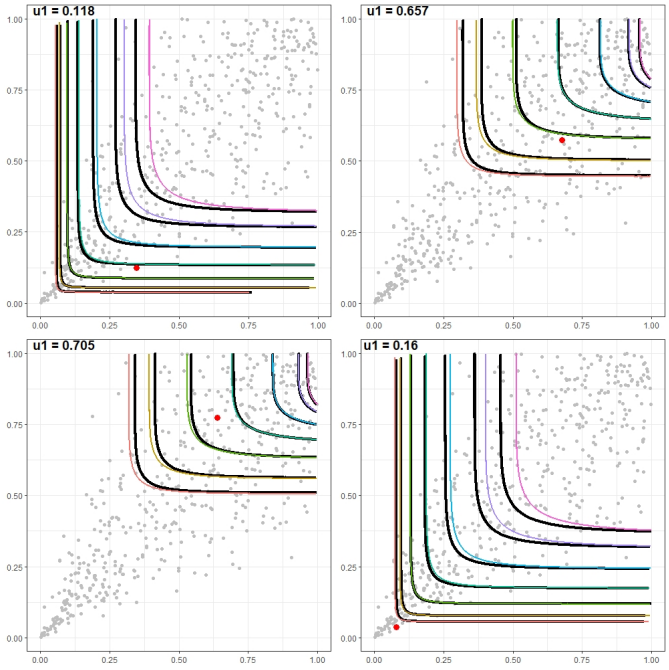


Figure 5.5.: x-axis: V_1, y-axis: V_2. Gray points: simulated data from vine distribution (n=500). Black curves: theoretical conditional level curves. Colored curves: estimated conditional level curves. Depicted are level curves at $\alpha = 0.05, 0.1, 0.25, 0.5, 0.75, 0.90, 0.95$ (left bottom to right top in each panel). Red dot: associated values of (v_1, v_2) with u_1 as conditioning value.

we simulate a data set $\mathbf{W} \in [0,1]^{504\times 3}$ from $\mathcal{D}_3$ and split $\mathbf{W}$ into $\mathbf{W}_{train} \in \mathbb{R}^{500\times 3}$ and $\mathbf{W}_{test} \in \mathbb{R}^{4\times 3}$. On the training set $\mathbf{W}_{train}$ we fit a vine model $\hat{\mathcal{D}}_3$ with the same vine tree structure and order of the variables as the data generator $\mathcal{D}_3$. In 3 dimensions, a C- and a D-vine tree structure coincide, so by order we mean the order from left to right in which the variables appear in the first tree of the sequence, as defined for a general D-vine copula (or C-vine copula in 3 dimensions) is Section 3.2. The estimated pair copulas are $\hat{C}_{V_1,U_1} \sim \mathit{Clayton}\left(\hat{\tau} = 0.57,\ \hat{\theta}_{V_1,U_1} = 2.65\right)$, $\hat{C}_{U_1,V_2} \sim \mathit{Gumbel}\left(\hat{\tau} = 0.79,\ \hat{\theta}_{U_1,V_2} = 4.92\right)$, $\hat{C}_{V_1,V2;U_1} \sim \mathit{Clayton}\left(\hat{\tau} = 0.40,\ \hat{\theta}_{V_1,V_2;U_1} = 1.34\right)$.

The corresponding conditional level curves of $\hat{\mathcal{D}}_3$ are obtained using the numerical evaluation procedure from Section 5.3 and evaluating $\hat{C}_{V_1,V2;U_1}$ in a similar manner as in Equation (5.3), using the estimates of each term. Note that the estimated and the data-generating vine are approximately very close, due to the use of the same tree structure in both data generation and estimation. But in practice this is not the case, as the underlying tree structure is unknown. However, we can use the Y-vine regression

model developed in Chapter 4 to model the tree structure and the pair copulas, in a way that the joint conditional distribution is easy to be estimated.

Figure 5.5 shows the theoretical and the estimated level curves for 4 conditioning values of u_1. The values of u_1 are chosen from $\mathbf{W}_{test}$. The level curves depend on the conditioning value. If the value of u_1 is low (top-left and bottom-right plot) the level curves are more restricted to the lower left corner. For greater values of u_1 (top-right and bottom-left plot) the level curves are more restricted to the top right corner. These occurrences can be explained by the high positive dependence of the pairs (V_1, U_1) and (V_2, U_1) in the first tree of the vine structure, meaning that low values of u_1 correspond to low values of both v_1, v_2.

Thus, Figures 5.2 and 5.5 show that the numerical procedure for obtaining both unconditional and conditional level curves is properly estimating the bivariate level curves and we will employ it to estimate conditional level curves in the case of more than 1 conditioning value (corresponding to more than one predictor in a regression setting).

5.4. Bivariate quantile curves

The notion of multivariate quantiles is not trivial nor well-defined. Usually, in literature the level sets or curves of a multivariate distribution are considered as multivariate quantile, however in this case the coverage probability is not exact. For example, in Fernández-Ponce and Suárez-Lloréns (2002) the bivariate unconditional quantiles are defined as the level sets of a bivariate distribution function. The authors state that this definition is a natural generalization of the univariate quantile sets (Lewis and Thompson, 1981), however later in Belzunce et al. (2007) it is shown that the level curves do not have the property that the α-th level curve separates the lowest $\alpha \times 100$ percent of the observations from the remaining $(1 - \alpha) \times 100$ percent of the observations. Thus, we suggest to define the bivariate quantile curves as adjusted level curves where the coverage probability is exact. Since we are interested in a regression setting, the following study is done on the conditional case, however, the same methodology can be applied for the unconditional case. Also, we define the bivariate quantiles on the u-scale (copula level), and to transform the bivariate quantiles on the x-scale, we use the same analogy as for the level curves.

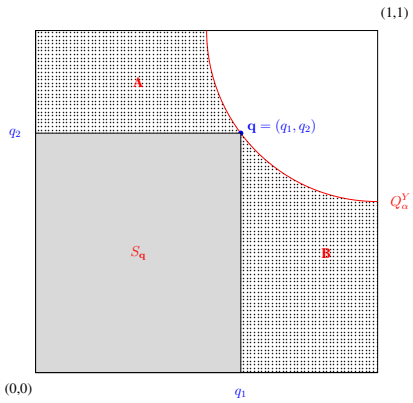


Figure 5.6.: A randomly chosen $\mathbf{q} = (q_1, q_2)$ vector, its corresponding $S_\mathbf{q}$ and S_α^{lower}, where $S_\alpha^{lower} = S_\mathbf{q} \bigcup A \bigcup B$.

5.4.1. General framework

Consider any bivariate vector $\mathbf{q} = (q_1, q_2) \in Q_\alpha^V(\mathbf{u})$ that lies on the level set $Q_\alpha^V(\mathbf{u})$ for some $\alpha \in (0,1)$. Further, let $S_\mathbf{q}$ be a set of bivariate vectors defined by

$$S_\mathbf{q} := \{(v_1, v_2) \in [0,1]^2 \; ; \; v_1 \leq q_1, v_2 \leq q_2\}.$$

Then for any random vector $\mathbf{W} = (W_1, W_2)^T \sim C_{V_1,V_2|\mathbf{U}}(\cdot, \cdot|\mathbf{u})$ it holds that

$$P\left(\mathbf{W} \in S_\mathbf{q}\right) = \alpha, \tag{5.4}$$

by following Equation (5.2). In Figure 5.6 we can see an exemplary illustration for the set $S_\mathbf{q}$. Let the region S_α^{lower} be defined as the set of bivariate vectors below the level curve $Q_\alpha^V(\mathbf{u})$,

$$S_\alpha^{lower} := \bigcup_{\forall (q_1,q_2) \in Q_\alpha^V(\mathbf{u})} \left\{(v_1, v_2) \in [0,1]^2; v_1 < q_1, v_2 < q_2\right\}.$$

Then for the random vector $\mathbf{W} = (w_1, w_2)^T$ it holds that

$$P\left(\mathbf{W} \in S_\alpha^{lower}\right) > \alpha, \tag{5.5}$$

since $S_\mathbf{q} \subset S_\alpha^{lower}$ for all $\mathbf{q} = (q_1, q_2) \in Q_\alpha^V(\mathbf{u})$, as also noted in Fernández-Ponce and Suárez-Lloréns (2002). (See Figure 5.6 to observe the S_α^{lower} region.)

This implies that the level curve $Q_\alpha^V(\mathbf{u})$ divides the $[0,1]^2$ square into a region for which it holds that $P\left(\mathbf{W} \in S_\alpha^{lower}\right) \neq \alpha$. It also follows that $P\left(\mathbf{W} \notin S_\alpha^{lower}\right) < 1 - \alpha$.

Thus, for the definition of a $\alpha-$ quantile curve we want to find an adjusted $\beta(\alpha)$ level curve which will divide the observation space into α and $1-\alpha$ percent, i.e. for which $P\left(\mathbf{W} \in S^{lower}_{\beta(\alpha)}\right) = \alpha$ holds.

Definition 5 *The bivariate conditional quantile for $\alpha \in (0,1)$, a transformation $\beta : (0,1) \longmapsto (0,1)$ and continuous random variables Y_1, Y_2 with random PITs V_1, V_2 given the outcome of the random vector $\mathbf{X} = \mathbf{x}$, with PITs $\mathbf{U} = \mathbf{u}$ is a curve in $[0,1]^2$ defined by the set*

$$q^V_\alpha(\mathbf{u}) \coloneqq \{(v_1, v_2) \in [0,1]^2 \,;\, C_{V_1,V_2|\mathbf{U}}(v_1, v_2|\mathbf{u}) = \beta(\alpha)\}, \tag{5.6}$$

so that the observation space is divided into α and $1-\alpha$ percent regions, i.e. $P\left(\mathbf{W} \in S^{lower}_{\beta(\alpha)}\right) = \alpha$ holds.

The estimation of the transformation $\beta(\alpha)$ is discussed later in Section 5.4.2. Following Definition 5, we can also define the exact $100 \times (1-\alpha)\,\%$ confidence region arising from the quantile curves $q^V_{\alpha/2}(\mathbf{u})$ and $q^V_{1-\alpha/2}(\mathbf{u})$.

Definition 6 *The $100 \times (1-\alpha)\,\%$ bivariate confidence region for $\alpha, \beta \in (0,1)$ and a continuous bivariate vector continuous random variables Y_1, Y_2 with random PITs V_1, V_2 given the outcome of the random vector $\mathbf{X} = \mathbf{x}$, with PITs $\mathbf{U} = \mathbf{u}$, is set of points in $[0,1]^2$ enclosed by the quantile curves $q^V_{\alpha/2}(\mathbf{u})$ and $q^V_{1-\alpha/2}(\mathbf{u})$, i.e.*

$$CI^{V_1,V_2|\mathbf{U}}_\alpha \coloneqq \Bigg\{ (w_1^*, w_2^*) \in [0,1]^2 \mid \exists \ \left(v_1^1, v_2^1\right) \in q^V_{\alpha/2}(\mathbf{u}),\ \left(v_1^2, v_2^2\right) \in q^V_{1-\alpha/2}(\mathbf{u}) \ \textit{such that}: \\ v_1^1 \le w_1^* \le v_1^2 \ \textit{and} \ v_2^1 \le w_2^* \le v_2^2 \Bigg\}.$$

In this case,

$$P(\mathbf{W} \in CI^{V_1,V_2|\mathbf{U}}_\alpha) = P(\mathbf{W} \in S^{lower}_{\beta(\alpha/2)}) - P(\mathbf{W} \in S^{lower}_{\beta(1-\alpha/2)}) = \alpha/2 - (1-\alpha/2) = 1-\alpha,$$

implying that $CI^{V_1,V_2|\mathbf{U}}_\alpha$ is an exact $100 \times (1-\alpha)\,\%$ confidence region.

5.4.2. From level curves to quantile curves

Returning back to the problem of estimating the transformation $\beta(\alpha)$, for $\beta : (0,1) \longmapsto (0,1)$, so that the quantile curves $q^V_\alpha(\mathbf{u})$ are estimated, we suggest a numerical procedure. Basically, we need to change the α-level curve to a new $\beta(\alpha)$- level curve so that $P(\mathbf{W} \in S^{lower}_{\beta(\alpha)}) = \alpha$ holds true. To achieve this, we define the function

$$\begin{aligned} G(\beta) &\coloneqq P\left(\mathbf{W} \in S^{lower}_{\beta}\right) \\ &= P\left(C_{V_1,V_2|\mathbf{U}}(\cdot,\cdot|\mathbf{u}) \le \beta\right) \quad \forall\, \beta \in (0,1). \end{aligned} \tag{5.7}$$

From Equation (5.5) we can see that $G(\alpha) > \alpha$. However, we are interested to find the value $\beta(\alpha)$ so that it holds that $G(\beta(\alpha)) = \alpha$, thus $\beta(\alpha) = G^{-1}(\alpha)$. To do so, we suggest a numerical procedure. As the function $G(\beta)$ is difficult to evaluate (more details follow below in this section), we suggest to estimate it using a simulated sample from the Y-vine copula with $\mathbf{U} = \mathbf{u}$ fixed. For $n = 1, \cdots, N$ we simulate observations $(v_1^n(\mathbf{u}), v_2^n(\mathbf{u})) \sim C_{V_1,V_2|\mathbf{U}}(\cdot, \cdot|\mathbf{u})$, as described in Section 4.5. Then, we estimate $G(\beta)$ as the proportion of the simulated data below the α-quantile over the sample size N, i.e.

$$\hat{G}(\beta) = \frac{1}{N} \sum_{n=1}^{N} \mathbb{I}\left((v_1^n(\mathbf{u}), v_2^n(\mathbf{u})) \in S_\beta^{lower}\right),$$

where $\mathbb{I}$ is an indicator function, being equal to 1 when the condition $(v_1^n(\mathbf{u}), v_2^n(\mathbf{u})) \in S_\beta^{lower}$ is satisfied, and equal to 0, otherwise. To find the desired $\beta(\alpha)$ we use a line search algorithm on the $(0, 1)$ interval and obtain the estimated $\hat{\beta}(\alpha)$ such that $\hat{G}(\hat{\beta}(\alpha)) = \alpha$. This way the suggested methodology from Section 5.2 can be extended to find the bivariate quantiles $q_\alpha^V(\mathbf{u})$ such that $\hat{G}(\beta(\alpha)) = P\left(\mathbf{W} \in S_{\beta(\alpha)}^{lower}\right) = \alpha$, holds, i.e. the $\beta(\alpha)$-th level set separates the lowest $\alpha \times 100$ percent of the observations from the remaining $(1 - \alpha) \times 100$ percent of the observations.

Theoretical formulation

The theoretical derivation of $P\left(C_{V_1,V_2|\mathbf{U}}(v_1, v_2|\mathbf{u}) \leq \beta\right) \quad \forall\ \beta \in (0, 1)$ from Equation (5.7) is the following. Consider the case when we look for a solution of the equation $C_{V_1,V_2|\mathbf{U}}(v_1, 1|\mathbf{u}) = \beta$. This can be written as

$$\begin{aligned} \beta =& C_{V_1,V_2|\mathbf{U}}(v_1, 1|\mathbf{u}) \\ =& P(V_1 \leq v_1, V_2 \leq 1|\mathbf{U} = \mathbf{u}) \\ =& P(V_1 \leq v_1|\mathbf{U} = \mathbf{u}) \\ =& C_{V_1|\mathbf{U}}(v_1|\mathbf{u}). \end{aligned} \tag{5.8}$$

Then, denote the solution of Equation (5.8) for v_1 as $C_{V_1|\mathbf{U}}^{-1}(\beta|\mathbf{u})$. Now consider the case when we look for a solution of the equation $C_{V_1,V_2|\mathbf{U}}(v_1, 1|\mathbf{u}) = \beta$ for v_2, when β, v_1 and $\mathbf{u}$ are fixed. Denote the solution for v_2 for fixed β, v_1 and $\mathbf{u}$ as $C_{V_1,V_2|\mathbf{U}}^{-1}(\beta|v_1, \mathbf{u})$. Then, we have

$$
\begin{aligned}
P\left(C_{V_1,V_2|\mathbf{U}}(v_1,v_2|\mathbf{u}) \leq \beta\right) =& 1 - P\left(C_{V_1,V_2|\mathbf{U}}(v_1,v_2|\mathbf{u}) > \beta\right) \\
=& 1 - \int_{C^{-1}_{V_1|\mathbf{U}(\beta|\mathbf{u})}}^{1} P\left((v_1,v_2) \notin S_\beta^{lower}|\mathbf{U}=\mathbf{u}, V_1 = v_1\right) \cdot c_{V_1|\mathbf{U}}(v_1|\mathbf{u}) dv_1 \\
=& 1 - \int_{C^{-1}_{V_1|\mathbf{U}(\beta|\mathbf{u})}}^{1} \left[\int_{C^{-1}_{V_1,V_2|\mathbf{U}(\beta|\mathbf{u},v_1)}}^{1} c_{V_1,V_2|\mathbf{U}}(v_1,v_2|\mathbf{u}) dv_2\right] \cdot c_{V_1|\mathbf{U}}(v_1|\mathbf{u}) dv_1.
\end{aligned} \tag{5.9}
$$

However, to numerically evaluate in stable fashion the integrals needed in Equation (5.9) is very difficult and it remains an open task.

Kendall distribution function extension

Another concept for the construction of exact confidence regions is been developed in Coblenz et al. (2018). The authors propose, to construct an exact confidence region for unconditional bivariate copula distribution functions. They use the Kendall distribution function of a bivariate copula C at a level $\alpha \in (0,1)$, $K(C,\alpha)$ defined as

$$K(C,\alpha) \coloneqq P\left(C(U,V) \leq \alpha,\ (U,V) \sim C\right),$$

in Genest and Rivest (1993) and Barbe et al. (1996). In comparison to our methodology it holds that $G(\beta(\alpha)) = K(C,\alpha)$ in the unconditional case, as shown in Chakak and Ezzerg (2000). For bivariate copula distribution functions computing the Kendall distribution function is possible and certain approaches are available (Chakak and Ezzerg, 2000; Ezzerg et al., 1999), however it is very computationally expensive (Brechmann, 2013). Once K is estimated, $\beta(\alpha)$ can be obtained as the inverse of the Kendall distribution function evaluated at α, i.e. $\beta(\alpha) = K^{-1}(C,\alpha)$.

However, estimating the Kendall distribution functions in the conditional case is difficult in general and computationally expensive. We shortly provide an idea for a future research topic using the non-simplified conditional copula (Gijbels and Matterne, 2021). We are interested in the joint conditional distribution function of Y_1, Y_2 given $\mathbf{X} = \mathbf{x}$, i.e. $F_{Y_1,Y_2|\mathbf{X}}(y_1,y_2|\mathbf{x})$ with conditional marginal distributions $F_{Y_1|\mathbf{X}}(y_1|\mathbf{x})$ and $F_{Y_2|\mathbf{X}}(y_2|\mathbf{x})$. If the conditional marginal distributions are continuous, Sklar's Theorem (2.1) applied to the conditional marginals ensures the existence of a unique copula $C^{V_1,V_2|\mathbf{U}=\mathbf{u}}$ for any fixed $\mathbf{U} = \mathbf{u}$. In particular, we have for $\mathbf{u} = (F_{X_1}(x_1), \ldots, F_{X_p}(x_p))$ that

$$F_{Y_1,Y_2|\mathbf{X}}(y_1,y_2|\mathbf{x}) = C^{V_1,V_2|\mathbf{U}=\mathbf{u}}\left(F_{Y_1|\mathbf{X}}(y_1|\mathbf{x}), F_{Y_2|\mathbf{X}}(y_2|\mathbf{x})\right).$$

This is a so-called conditional copula of (Y_1, Y_2) given $\mathbf{X} = \mathbf{x}$. The conditional copula function fully describes the conditional dependence of Y_1 and Y_2 given the observed

vector $\mathbf{X} = \mathbf{x}$ (see more in Gijbels et al. (2011) and Veraverbeke et al. (2011)). Then, the corresponding conditional Kendall's tau can be determined by

$$\tau_{C^{V_1,V_2|\mathbf{U}=\mathbf{u}}} = 4 \int_0^1 \int_0^1 C^{V_1,V_2|\mathbf{U}=\mathbf{u}} \left(F_{Y_1|\mathbf{X}}(y_1|\mathbf{x}), F_{Y_2|\mathbf{X}}(y_2|\mathbf{x}) \right) dC^{V_1,V_2|\mathbf{U}=\mathbf{u}} - 1.$$

However, estimation of this conditional copula is, up to this moment, a very unexplored topic and estimation of the conditional Kendall's tau is very sensitive to bias properties of the underlying copula estimator (Gijbels et al., 2011), so further research is needed in this area. Further, estimation of the conditional Kendall's function associated $K(C^{V_1,V_2|\mathbf{U}=\mathbf{u}}, \alpha) := P\left(C^{V_1,V_2|\mathbf{U}=\mathbf{u}} \leq \alpha\right)$, is an even more involved topic, and based on our knowledge, a completely unexplored topic.

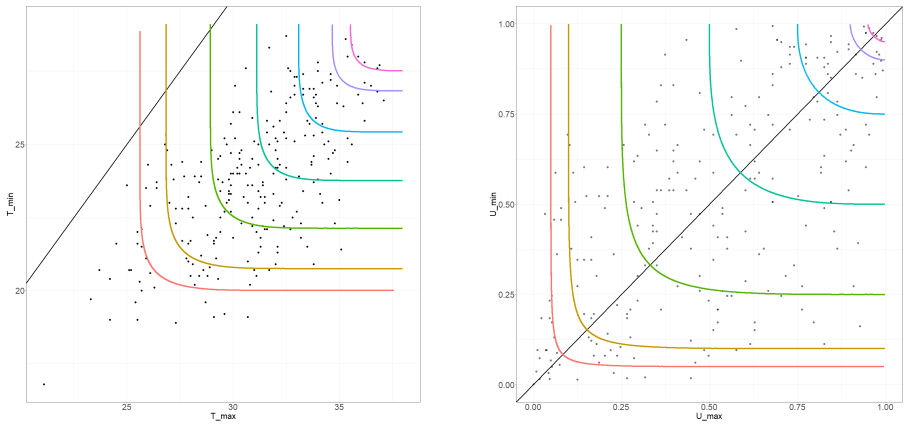
5.5. Data application

We continue the data analysis on the data set introduced in Section 4.7. Here we explore the unconditional and conditional level curves, quantiles curves and corresponding confidence regions.

5.5.1. Bivariate level curves

For illustrating the unconditional level curves of the joint unconditional bivariate distribution of the two responses, U_max and U_min we fit a pair copula between them. The estimated pair copula is the Gaussian copula with a parameter of 0.66. The unconditional quantile curves are defined as in Definition 5 by using the pair copula distribution function between the responses C_{V_1,V_2}, instead of the bivariate conditional distribution $C_{V_1,V_2|\mathbf{U}}$, and we denote them as q_α^V for $\alpha \in (0,1)$. The level curves of this copula, on both the x- and the u-scale are given in Figure 5.7. Due to constraints of the weather system in question, the maximum temperature is always required to be greater than the minimum temperature. However, this ordering constraint does not imply an ordering constraint on the PITs on the u-scale (as the marginal distributions are separately and independently modeled). For illustration see Figure 5.7, where the ordering is visible in panel (a), as all the data is below the diagonal, while this ordering is lost in panel (b).

Next, we show conditional level curves for 3 chosen days from the testing set, estimated using the fitted Y-vine model $\hat{\mathcal{Y}}$. The estimated level curves for the chosen 3 days of the testing set are given in Figure 5.8. The top row are estimates on the x-scale and the bottom row are on the u-scale. The ranges of the x-scale plots are the ranges of the minimum and maximum possible temperatures, which are $(22, 38)$ for the maximum temperature and $(16, 29)$ for the minimum temperature.



(a) x-scale (b) u-scale

Figure 5.7.: Black points: data from 2013-2016 (n=246). Colored curves: estimated unconditional level curves at $\alpha = 0.05, 0.1, 0.25, 0.5, 0.75, 0.90, 0.95$ (left bottom to right top).

Comparing the 3 days shown in Figure 5.8, we can see very different level curves depending on the observed conditioning variables, i.e. predictors. For the day 10.08.2017 we observe that higher minimum and maximum temperatures are observed compared to the other days. The estimated level curves on the u-scale for 10.08.2017 are more skewed to the upper tail, compared to the estimates for 25.08.2017, which are skewed towards the lower tail. Thus, the extreme case of very high maximum and minimum temperatures, is very probable on 10.08.2017, opposite of 25.08.2017 when the probability is lower. Opposite to this, the extreme case of a very small minimum temperature and small to moderate maximum temperatures is very probable on 25.08.2017 and highly not probable on 10.08.2017. The estimates obtained for 18.08.2017 are very moderate and extreme values for both responses have very low probabilities on this date. The different shapes of the level curves for the three chosen days, coming from conditional distributions with different conditioning values, also show that the dependence structure between the response variables is not static and it changes based on the conditioning variables. Ignoring this dependence can lead to a significant underestimation of extreme events, encoded in the tail dependencies of the joint conditional distribution of the responses.

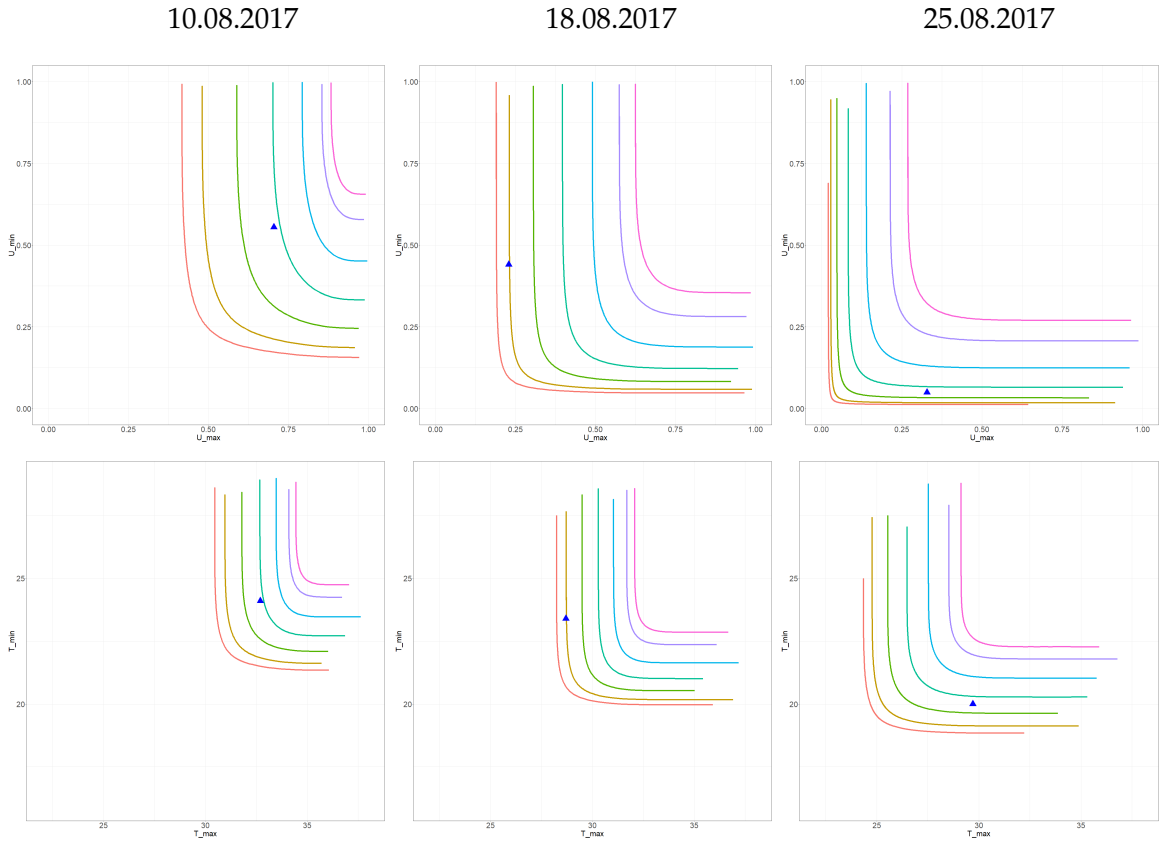


Figure 5.8.: The plots correspond to the days 10.08.2017, 18.08.2017 and 25.08.2017 (left to right). Shown are estimated conditional level curves at $\alpha = 0.05, 0.1, 0.25, 0.5, 0.75, 0.90, 0.95$ (left bottom to right top). Row 1 are estimates on the x-scale and row 2 is on the u-scale. The blue triangle is the true observed value.

5.5.2. Bivariate quantile curves, confidence regions and advantages of joint modeling of dependent responses

Unconditional case

First, we consider the unconditional quantile curves and the corresponding confidence region obtained from fitting a bivariate pair copula between the responses U_{max} and U_{min}, and the confidence region obtained by assuming dependence between the responses. The unconditional quantile curves are defined as in Definition 5 by using the pair copula between the responses C_{V_1,V_2}, instead of the bivariate conditional distribution $C_{V_1,V_2|\mathbf{U}}$, and are denoted as q_α^V for $\alpha \in (0,1)$. Using Definition 6, by substituting the conditional quantile curves with the unconditional ones, we can define the corresponding unconditional confidence region $CI_\alpha^{V_1,V_2}$ as set of points in $[0,1]^2$ enclosed by the quantile curves $q_{\alpha/2}^V$ and $q_{1-\alpha/2}^V$ for some $\alpha \in (0,1)$, i.e.

$$CI_\alpha^{V_1,V_2} := \Bigg\{ (w_1^*, w_2^*) \in [0,1]^2 \mid \exists \ \left(v_1^1, v_2^1\right) \in q_{\alpha/2}^V, \ \left(v_1^2, v_2^2\right) \in q_{1-\alpha/2}^V \textit{ such that :} \\ v_1^1 \leq w_1^* \leq v_1^2 \ \textit{ and } \ v_2^1 \leq w_2^* \leq v_2^2 \Bigg\}.$$

The case when we assume independence between the responses, we construct a bivariate quantile region from the univariate empirical quantiles, denoted as $q_{\alpha,emp}^{V_1}$ for $\alpha \in (0,1)$, using the Bonferroni correction for multiple testing (Bonferroni, 1936). We are interested in the bivariate quantile region with coverage probability at 50%, $\alpha = 0.50$ meaning that the two univariate empirical quantiles, from which we construct the bivariate quantile region, need to be evaluated at $\frac{\alpha}{4} = 0.125$ and $1 - \frac{\alpha}{4} = 0.875$, and we denote the corresponding confidence region of the univariate empirical quantiles as $CI_{0.50}^{V_1 \perp V_2}$, i.e.

$$CI_{0.50}^{V_1 \perp V_2} := \left[q_{\frac{0.50}{4},emp}^{V_1}, q_{1-\frac{0.50}{4},emp}^{V_1} \right] \times \left[q_{\frac{0.50}{4},emp}^{V_2}, q_{1-\frac{0.50}{4},emp}^{V_2} \right].$$

Also, we are interested in the bivariate quantile region with coverage probability of 90%, $\alpha = 0.10$, meaning that the two univariate empirical quantiles, from which we construct the bivariate quantiles, need to be evaluated at $\frac{\alpha}{4} = 0.025$ and $1 - \frac{\alpha}{4} = 0.975$, and we denote the corresponding confidence region of the univariate empirical quantiles as $CI_{0.90}^{V_1 \perp V_2}$, i.e.

$$CI_{0.90}^{V_1 \perp V_2} := \left[q_{\frac{0.10}{4},emp}^{V_1}, q_{1-\frac{0.10}{4},emp}^{V_1} \right] \times \left[q_{\frac{0.10}{4},emp}^{V_2}, q_{1-\frac{0.10}{4},emp}^{V_2} \right].$$

The first row of Figure 5.9, shows the bivariate unconditional level curves (solid lines) and quantile curves (dashed lines) (left panel is on the u-scale, right panel on the x-scale). The adjusted level curves, the bivariate quantiles are estimated using the

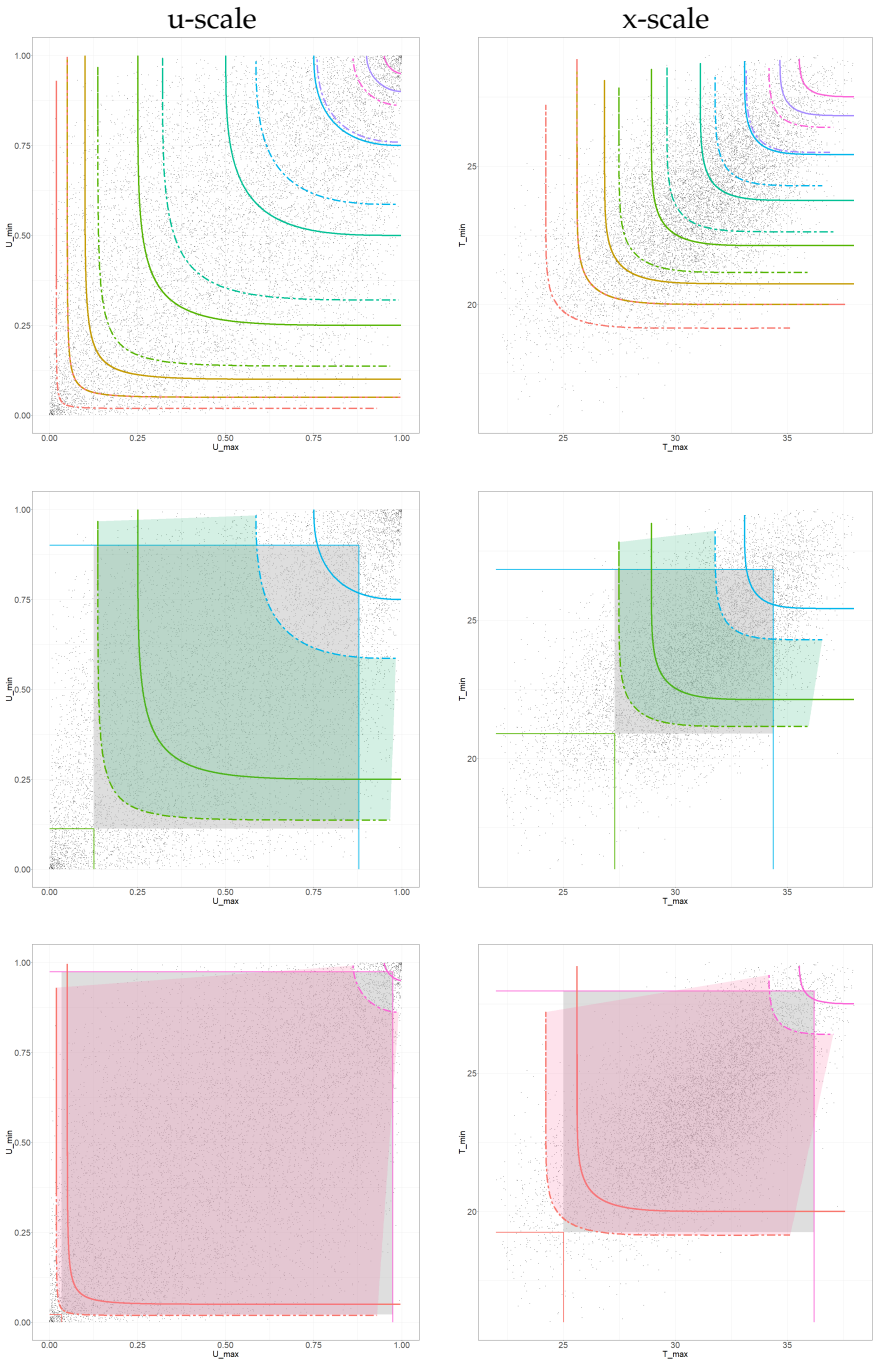


Figure 5.9.: First row: **unconditional** level curves (solid lines) and corresponding quantile curves (dashed lines) at $\alpha = 0.05, 0.1, 0.25, 0.5, 0.75, 0.90, 0.95$. Second row: $CI_{0.50}^{V_1,V_2}$ (green region) and $CI_{0.50}^{V_1 \perp V_2}$ (gray region). Third row: $CI_{0.90}^{V_1,V_2}$ (red region) and $CI_{0.90}^{V_1 \perp V_2}$ (gray region).

proposed method introduced in Section 5.4.2. From the fitted pair copula, we simulate 10 000 data points from which the unconditional quantile curves are estimated and the simulated points are shown as well. Table 5.1 shows for all α levels, the estimated coverage probabilities $\hat{G}(\alpha)$ and the estimated adjustment $\hat{\beta}(\alpha)$ for the corresponding unconditional quantile levels.

α	0.05	0.10	0.25	0.50	0.75	0.90	0.95
$\hat{G}(\alpha)$	0.10	0.20	0.41	0.67	0.89	0.97	0.99
$\hat{\beta}(\alpha)$	0.02	0.05	0.14	0.32	0.59	0.76	0.86

Table 5.1.: For all α levels, estimated coverage probabilities $\hat{G}(\alpha)$ and estimated adjustment $\hat{\beta}(\alpha)$ for the corresponding **unconditional** quantile levels.

The second row shows the $CI_{0.50}^{V_1,V_2}$ (green shaded region) and $CI_{0.50}^{V_1\perp V_2}$ (gray shaded region). The estimated values for the adjustment to quantile curves (dashed lines) are $\hat{\beta}(0.25) = 0.14$ and $\hat{\beta}(0.75) = 0.59$. Using these values, we construct the confidence region $CI_{0.50}^{V_1,V_2}$, shown in the green shaded region between the bivariate quantile curves. The coverage probability for the confidence region is 0.50, while the coverage probability below the level curve at $\alpha = 0.25$ is 0.41, and below the level curve at $\alpha = 0.75$ is 0.89. Also, we show the difference between the confidence interval constructed from the fitted pair copula model, which jointly models the responses and the Bonferroni corrected confidence regions, constructed from the empirical quantiles. The gray shaded region is the $CI_{0.50}^{V_1\perp V_2}$ region. The panels additionally contain horizontal lines whose y-intercepts correspond to the 0.125 and 0.875 univariate empirical quantiles for the minimum temperature. Moreover, it contains vertical lines with x-intercepts corresponding to the $0.125, 0.875$ empirical quantiles for the maximal temperature.

The last row shows the $CI_{0.90}^{V_1,V_2}$ (red shaded region) and $CI_{0.90}^{V_1\perp V_2}$ (gray shaded region). The estimated values for the adjustment to quantile curves (dashed lines) are $\hat{\beta}(0.05) = 0.02$ and $\hat{\beta}(0.95) = 0.86$. Using these values, we construct the confidence region $CI_{0.90}^{V_1,V_2}$, shown in the red shaded region between the bivariate quantile curves. The coverage probability for the confidence region is 0.90, while the coverage probability below the level curve at $\alpha = 0.05$ is 0.10, below the level curve at $\alpha = 0.95$ is 0.99. Also, we show the difference between the confidence interval constructed from the fitted pair copula model, which jointly models the responses and the Bonferroni corrected confidence regions, constructed from the empirical quantiles. The gray shaded region is the $CI_{0.90}^{V_1\perp V_2}$ region. The panels additionally contain horizontal lines whose y-intercepts correspond to the 0.025 and 0.975 univariate empirical quantiles for the minimum temperature. Moreover, it contains vertical lines with x-intercepts corresponding to the $0.025, 0.975$ empirical quantiles for the maximal temperature.

Note that in Figure 5.9, all the left panels are given on the u-scale. However, using the transformations of the level curves between the u-scale and the x-scale, explained in Section 5.2.1, we also provide all the unconditional level curves, unconditional quantile curves and the corresponding confidence regions on the transformed x-scale in the corresponding right panel of Figure 5.9.

Conditional case

Next we compare the bivariate conditional quantile curves obtained from the Y-vine regression and its corresponding confidence regions to differently obtained quantiles and regions.

For comparison purposes, we treat the response variables as conditionally independent given a set of predictors. Basically, the tasks of predicting maximal and minimal temperatures are treated as completely independent problems and univariate conditional quantiles are estimated for both response variables. For this purpose, two univariate D-vine regression models with the same predictor order as the Y-vine regression are used. This way we can construct a bivariate quantile region from the univariate quantiles using the Bonferroni correction for multiple testing (Bonferroni, 1936), similar as before for the unconditional case. We denote these univariate D-vine based quantiles as, $q^{V}_{\alpha,Dvine}(\mathbf{u})$ for $\alpha \in (0,1)$. We are interested in the bivariate quantile region with coverage probability of 50%, $\alpha = 0.50$ meaning that the two univariate quantiles, need to be evaluated at $\frac{\alpha}{4} = 0.125$ and $1 - \frac{\alpha}{4} = 0.875$, and we denote the corresponding confidence region using the univariate conditional quantiles as $CI^{V_1 \perp V_2|\mathbf{U}}_{0.50}$, i.e.

$$CI^{V_1 \perp V_2|\mathbf{U}}_{0.50} := \left[q^{V_1}_{\frac{0.50}{4},Dvine}(\mathbf{u}), q^{V_1}_{1-\frac{0.50}{4},Dvine}(\mathbf{u})\right] \times \left[q^{V_2}_{\frac{0.50}{4},Dvine}(\mathbf{u}), q^{V_2}_{1-\frac{0.50}{4},Dvine}(\mathbf{u})\right].$$

Also, we are interested in the bivariate quantile region with coverage probability of 90%, at $\alpha = 0.10$ meaning that the two univariate quantiles, need to be evaluated at $\frac{\alpha}{4} = 0.025$ and $1 - \frac{\alpha}{4} = 0.975$, and we denote the corresponding confidence region of the univariate conditional quantiles as $CI^{V_1 \perp V_2|\mathbf{U}}_{0.90}$, i.e.

$$CI^{V_1 \perp V_2|\mathbf{U}}_{0.90} := \left[q^{V_1}_{\frac{0.10}{4},Dvine}(\mathbf{u}), q^{V_1}_{1-\frac{0.10}{4},Dvine}(\mathbf{u})\right] \times \left[q^{V_2}_{\frac{0.10}{4},Dvine}(\mathbf{u}), q^{V_2}_{1-\frac{0.10}{4},Dvine}(\mathbf{u})\right].$$

The first row of Figure 5.10, shows the bivariate conditional level curves (solid lines) and quantile curves (dashed lines) for 2 chosen dates from the testing set, 02.07.2017 and 21.08.2017. The adjusted level curves, the bivariate quantiles are estimated using the proposed method introduced in Section 5.4.2. The simulated 10 000 data points from which the quantiles are estimated are shown as well. The blue triangle dot is the observed value on that day. Tables 5.2 and 5.3 show for all α levels, the estimated

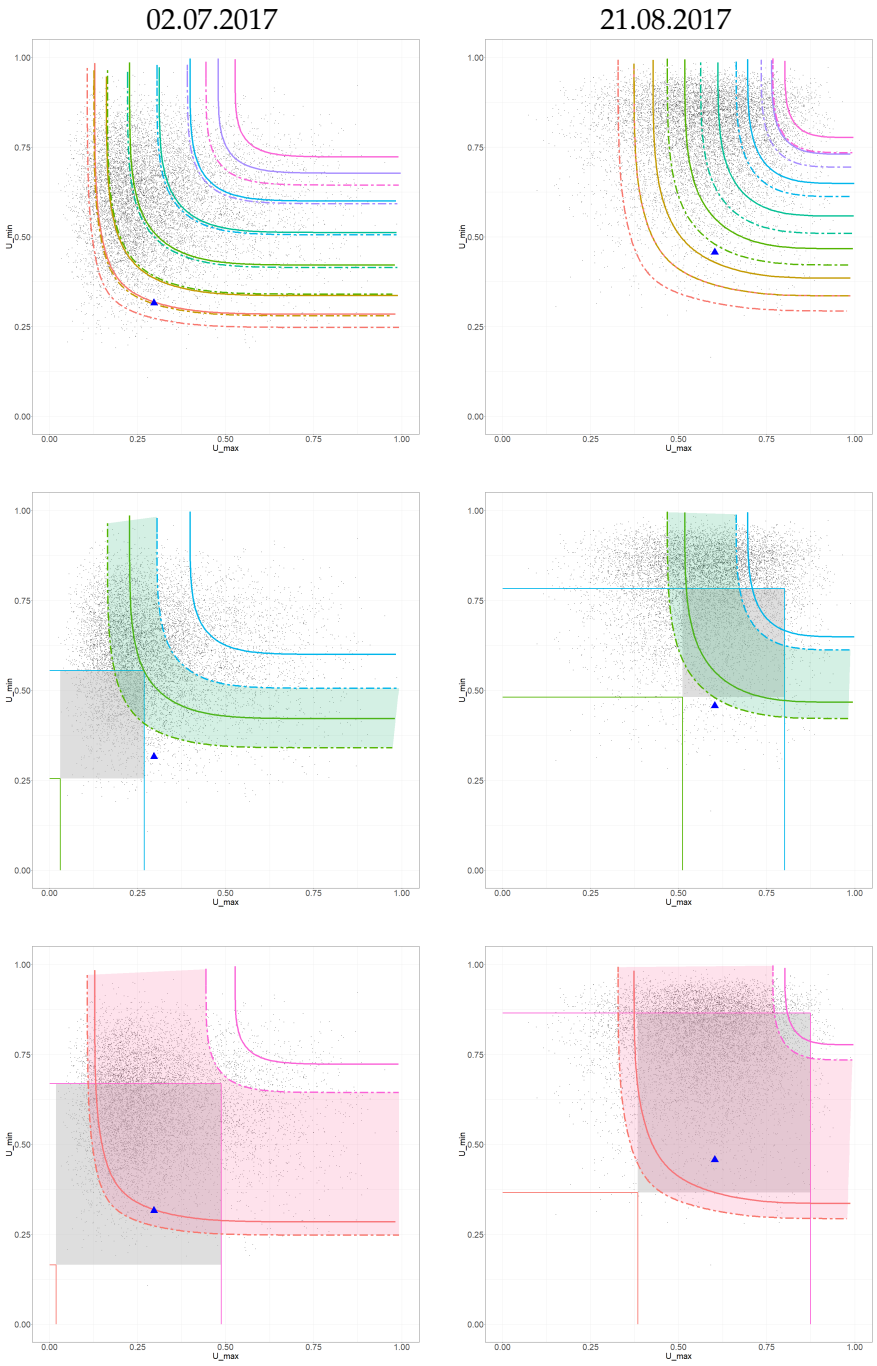


Figure 5.10.: First row: **conditional** level curves (solid lines) and corresponding quantile curves (dashed lines) at $\alpha = 0.05, 0.1, 0.25, 0.5, 0.75, 0.90, 0.95$. Second row: $CI_{0.50}^{V_1,V_2|\mathbf{U}}$ (green region) and $CI_{0.50}^{V_1 \perp V_2|\mathbf{U}}$ (gray region). Third row: $CI_{0.90}^{V_1,V_2|\mathbf{U}}$ (red region) and $CI_{0.90}^{V_1 \perp V_2|\mathbf{U}}$ (gray region). (All panels on **u-scale**.)

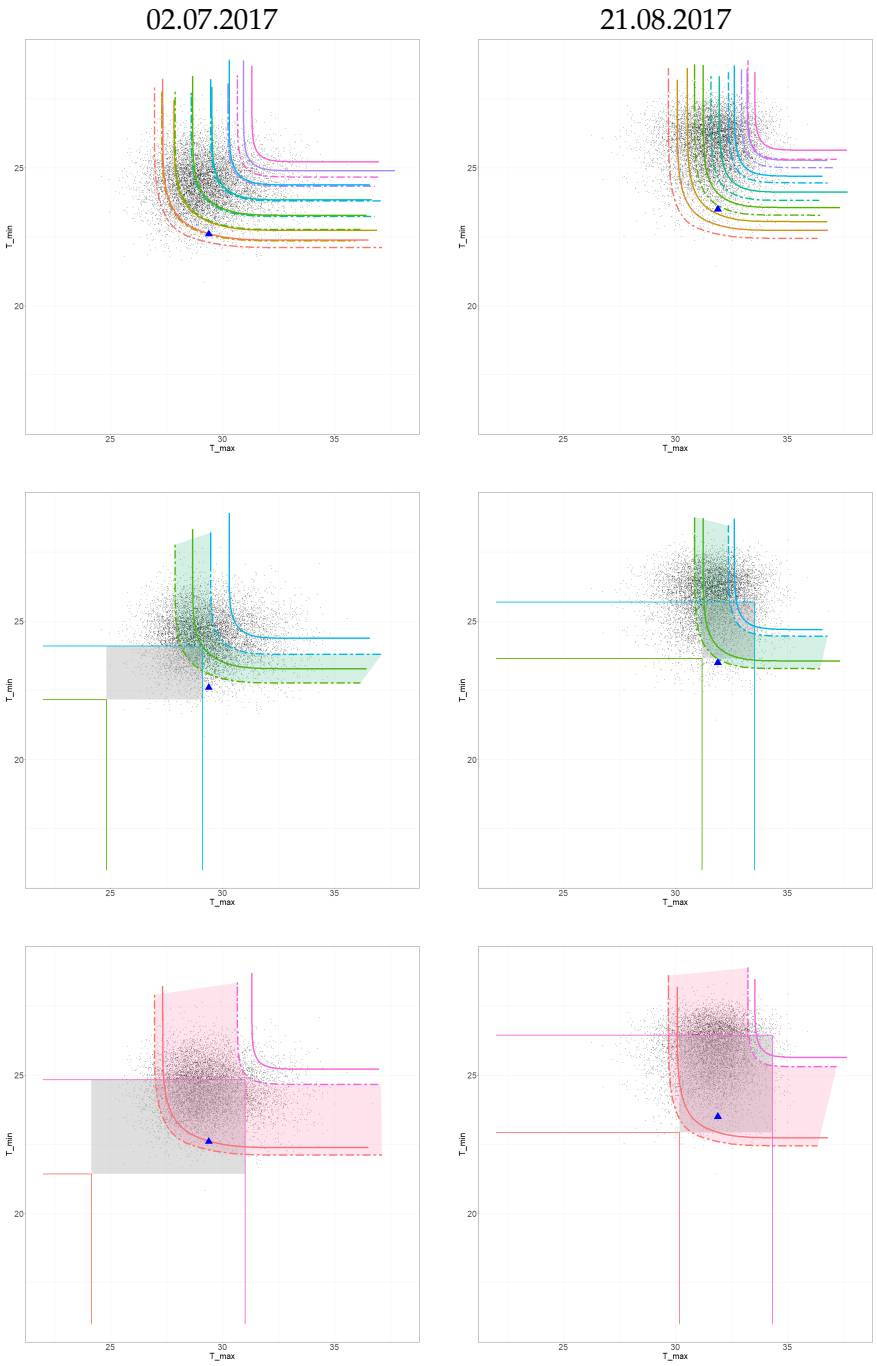


Figure 5.11.: First row: **conditional** level curves (lines) and corresponding quantile curves (dashed lines) at $\alpha = 0.05, 0.1, 0.25, 0.5, 0.75, 0.90, 0.95$. Second row: $CI_{0.50}^{Y_1,Y_2|\mathbf{X}}$ (green region) and $CI_{0.50}^{Y_1 \perp Y_2|\mathbf{X}}$ (gray region). Third row: $CI_{0.90}^{Y_1,Y_2|\mathbf{X}}$ (red region) and $CI_{0.90}^{Y_1 \perp Y_2|\mathbf{X}}$ (gray region). (All panels on **x-scale**.)

coverage probabilities $\hat{G}(\alpha)$ and the estimated adjustment $\hat{\beta}(\alpha)$ for the corresponding conditional quantile levels for 02.07.2017 and 21.08.2021, respectively.

α	0.05	0.10	0.25	0.50	0.75	0.90	0.95
$\hat{G}(\alpha)$	0.12	0.24	0.52	0.76	0.91	0.97	0.99
$\hat{\beta}(\alpha)$	0.03	0.05	0.11	0.23	0.48	0.73	0.85

Table 5.2.: For all α levels, estimated coverage probabilities $\hat{G}(\alpha)$ and estimated adjustment $\hat{\beta}(\alpha)$ for the corresponding conditional quantile levels for **02.07.2017**.

α	0.05	0.10	0.25	0.50	0.75	0.90	0.95
$\hat{G}(\alpha)$	0.10	0.17	0.37	0.64	0.83	0.95	0.98
$\hat{\beta}(\alpha)$	0.025	0.05	0.16	0.36	0.66	0.84	0.91

Table 5.3.: For all α levels, estimated coverage probabilities $\hat{G}(\alpha)$ and estimated adjustment $\hat{\beta}(\alpha)$ for the corresponding conditional quantile levels for **21.08.2017**.

In the second row, we show the bivariate conditional level curves at levels $\alpha = 0.25, 0.75$ (solid lines). The estimated values for the adjustment to quantile curves (dashed lines) are $\hat{\beta}(0.25) = 0.11$ and $\hat{\beta}(0.75) = 0.48$ for the date 02.07.2017 and $\beta(0.25) = 0.16$ and $\beta(0.75) = 0.66$ for the date 21.08.2017, respectively. Using these values, we construct the confidence region $CI_{0.50}^{V_1,V_2|\mathbf{U}}$, shown in the green shaded region between the bivariate quantile curves. The coverage probability for the confidence region is 0.50, while the coverage probability below the level curve at $\alpha = 0.25$ is 0.52, below the level curve at $\alpha = 0.75$ is 0.91 for date 02.07.2017. For 21.08.2017 the coverage probability below the level curve at $\alpha = 0.25$ is 0.37, below the level curve at $\alpha = 0.75$ is 0.83. Also, we show the difference between the confidence interval constructed from the Y-vine regression model, which jointly models the responses and the Bonferroni corrected confidence regions, constructed from the D-vine univariate regression models, which models the two responses conditionally independent of each other. The gray shaded region is the $CI_{0.50}^{V_1 \perp V_2|\mathbf{U}}$ region. The panels additionally contain horizontal lines whose y-intercepts correspond to the 0.125 and 0.875 univariate quantiles obtained from the univariate D-vine regression model with the minimal temperature as the response variable. Moreover, it contains vertical lines with x-intercepts corresponding to $0.125, 0.875$ quantiles of the univariate D-vine regression with the maximal temperature as the response variable.

In the third row, we show the level curves at levels $\alpha = 0.05, 0.95$ (solid lines). The estimated values for the adjustment to quantile curves (dashed lines) are $\hat{\beta}(0.05) = 0.03$

and $\hat{\beta}(0.95) = 0.85$ for the date 02.07.2017 and $\hat{\beta}(0.05) = 0.025$ and $\hat{\beta}(0.95) = 0.91$ for the date 21.08.2017, respectively. Using these values, we also construct the confidence region $CI_{0.90}^{V_1,V_2|\mathbf{U}}$, shown in the red shaded region between the bivariate quantile curves. The coverage probability for the confidence region is 0.90, while the coverage probability below the level curve at $\alpha = 0.05$ is 0.12, below the level curve at $\alpha = 0.95$ is 0.99 for date 02.07.2017. For 21.08.2017 the coverage probability below the level curve at $\alpha = 0.05$ is 0.10, below the level curve at $\alpha = 0.95$ is 0.98. Similarly, the gray shaded region is the $CI_{0.90}^{V_1 \perp V_2|\mathbf{U}}$ region. The panels additionally contain horizontal lines whose y-intercepts correspond to the 0.025 and 0.975 univariate quantiles obtained from the univariate D-vine regression model with the minimal temperature as the response variable. Moreover, it contains vertical lines with x-intercepts corresponding to $0.025, 0.975$ quantiles of the univariate D-vine regression with the maximal temperature as the response variable. Note that in Figure 5.10, all the plots are given on the u-scale. However, using the transformations of the level curves between the u-scale and the x-scale, explained in Section 5.2, we also provide all the conditional level curves, conditional quantile curves and the corresponding confidence regions on the transformed x-scale in Figure 5.11.

First, note the obvious difference in the obtained shapes of confidence regions arising from bivariate quantiles (dependent responses) and the univariate quantiles based regions (conditionally independent responses). While the bivariate confidence regions are free to vary in shape, the regions obtained by the univariate confidence intervals are bound to be rectangles. Also, for 21.08.2017 the univariate quantiles based confidence regions are subsets of the bivariate confidence regions obtained from the Y-vine regression. So, there is a whole range of points that are excluded from the confidence region constructed from the univariate quantiles. For 02.07.2017, the $CI_{0.50}^{V_1 \perp V_2|\mathbf{U}}$ is not even a subset of the bivariate confidence regions obtained from the Y-vine regression $CI_{0.50}^{V_1,V_2|\mathbf{U}}$, while for the $CI_{0.90}^{V_1 \perp V_2|\mathbf{U}}$ it is a partial subset of the bivariate $CI_{0.90}^{V_1,V_2|\mathbf{U}}$. Also, the univariate conditional quantiles based confidence regions are quite low in volume and don't capture any dependence between the responses. They also tend to underestimate the risk of extreme high values, which is a major drawback. Thus, using univariate quantiles and the corresponding confidence regions fails to capture, not only the dependence, but also the multidimensional nature of the problem.

5.6. Conclusion and outlook

The topic of bivariate response conditional quantiles is been tackled with the usage of a flexible vine copula model, the Y-vine copula regression model. This way we extend the Y-vine regression to bivariate quantile regression, by extending the notion of level curves, which are curves where the bivariate (conditional) distribution reaches

a specified level to adjusted level curves so that the coverage probability below and above the level curve is exact. This way we construct bivariate (conditional) quantile curves that divide the range of a bivariate (conditional) distribution into regions with given probabilities.

One of the contributions of this chapter is the numerical method that can be used to estimate the level curves of bivariate (conditional) distribution functions defined on the unit square. Also, we provide a method to transform the estimated level curves from the u-scale back onto the x-scale. Using this numerical method, we explore the unconditional level curves of bivariate pair copula distributions and the conditional level curves of a known 3-dimensional vine copula distribution and compare estimated and theoretical level curves. We also suggest a simulation based method for finding the adjusted bivariate level curves that are used to define the bivariate quantile curves. Using these bivariate quantile curves we construct confidence regions with exact coverage probabilities. We apply the proposed methodology on the minimum and maximum temperature data introduced in Section 4.7 and explore (conditional) level curves, (conditional) quantile curves and the corresponding confidence regions for different conditioning values of the predictors.

A possible future outlook is trying the other methodologies discussed in Section 5.4.2 on how the find the adjusted level curves that corresponds to bivariate quantile curves. Based on our knowledge, the other methods are very hard to be computed numerically, but in the future with more numerical methods and more research on this topic, it might become easier to estimate them.

Further, multivariate quantile curves and confidence regions are of increasing interest in hydrology, so as a possible future aspect we suggest, is the usage of our suggested methodology in a hydrological application (Coblenz et al., 2018). Especially interesting might be to analyse the behaviour of the quantile curves and confidence regions for some extreme values of the conditioning predictors, since the more standard methods might not be able to capture this behaviour on the tails of a distribution.

6. Univariate and bivariate risk analysis of late-frost and drought conditions in Bavaria

Chapter 6 is based on Tepegjozova et al. (2023).

6.1. Introduction

Since the end of the Industrial Revolution, carbon emissions caused by human activity have increased the concentration of carbon dioxide in the atmosphere by nearly 150% (Friedlingstein et al., 2022). The direct result of this are the shifts in long-term weather patterns more commonly referred to as anthropogenic climate change. In the simplest terms, only two strategies exist to combat climate change: eliminating the excessive discharge of carbon dioxide into the atmosphere and sequestering previously emitted carbon. Abetting humanity in the struggle to contain the concentration of carbon dioxide are the metaphoric lungs of our planet – forests. In an ironic twist of fate, the role of these forests in the context of climate change is a double-edged sword. While they are one of humanities greatest assets, they are also particularly threatened by the shifting climate.

In Central Europe, natural forests are dominated by European beech (*Fagus sylvatica* L.) (Leuschner and Ellenberg, 2017). Naturally, this tree species would cover more than 60% of the land surface area of Germany (Bohn and Welß, 2003), and it is also widespread across Europe with its distribution ranging from Sicily in the South up to Bergen in Southern Norway, covering approx. $140,000 km^2$ of forested area in total (Durrant et al., 2016). European beech has been promoted as a tree species well adapted to the future climate and as the most efficient broad-leaved tree species for climate change mitigation (Yousefpour et al., 2018).

However, recent evidence points to increased susceptibility of beech forests to increasingly dry and hot summers which have been the main effect of climate change in Central Europe in the past 20 years (Spinoni et al., 2017). In the absence of ample water supply, beech forests are susceptible to growth declines, large-scale damage and mortality (Scharnweber et al., 2011; Meyer et al., 2020). Most recently, this has been

observed in the wake of two successive drought events in 2018 and 2019 (Buras et al., 2019; Schuldt et al., 2020). Although these conditions are extreme outliers in the current climate, as climate change progresses they will likely become the new norm.

In addition to the adverse effects of increasing frequency and intensity of drought, beech ecosystems are also affected by another climate extreme: late-spring frost. Below freezing temperatures in spring, after trees have begun unfurling their leaves, can result in late-frost damage, defoliating large parts of the canopy (Dittmar et al., 2006; Menzel et al., 2015). Consequently, affected trees must expend carbohydrate reserves to grow a second canopy before the physiological processes necessary for photosynthesis can resume (D'Andrea et al., 2019). Somewhat counter intuitively, increasing temperatures may exacerbate spring late-frost risk: as (mean) temperature rises, the timing of leaf-out shifts – instead of leaves unfurling near the beginning of May, they can develop as early as the beginning of April when the probability of sub-zero minimum temperatures is higher (Zohner et al., 2020).

Both types of disturbance through climate extremes inhibit the regular functioning of beech and force the trees to expend stored resources to recuperate at the cost of forest vitality and productivity. Consequently, the joint occurrence of spring late-frost and drought poses a significant threat to forest health, multiplying the detrimental effects in comparison to the isolated effect of one of these climate extremes alone. However, we currently lack basic understanding of the statistical coupling between drought and spring late-frost as the necessary underpinning for risk assessment and associated forest management recommendations. Thus, our main objective is to quantify the joint probability of drought and spring late-frost in the historic domain and identify regions that exhibit the highest risk of extreme late-frost and drought conditions.

We propose to approach this topic using dependence modelling with copulas, which have become more popular in ecological analysis in recent years due to their ability to deal with non-Gaussian data. Climate data and indices derived from climate data often fall into this category, as they frequently belong to bounded or skewed distributions (Schölzel and Friederichs, 2008). Our proposed work is a step change from previous applications of copulas in ecology, which so far have focused only on jointly modelling multiple components of the same climate extreme, for example drought severity and drought duration (Sarhadi et al., 2016; Kwon and Lall, 2016) or frost severity and duration (Chatrabgoun et al., 2020). In our case however, we are interested in joint modeling of two extremes given a set of possible predictors. We use the Standardized Precipitation Evapotranspiration Index (SPEI) to quantify drought conditions. This index is based on temperature and precipitation and is standardized on a log-logistic distribution (Beguería and Vicente-Serrano, 2017). To identify frost occurrence we use a phenological model to project the timing of leaf-out and intersect this with a threshold minimum temperature (Kramer et al., 2017).

When quantifying the joint probability of drought and late spring-frost occurrences, and especially when relying on predictions from this quantification over longer periods, one has to account for extreme case weather events. To properly quantify these tail events we propose to use joint regression modeling of drought and late spring-frost based on a specific R-vine copula, able to jointly model two responses with a symmetric treatment, the Y-vine copula based regression, introduced in Chapter 4. In addition, we model drought and late spring-frost separately with a different R-vine model, which can handle a single response regression, the D-vine regression model (Kraus and Czado, 2017). This way we are able to compare the marginal and joint effects of change in drought and late spring-frost. Further, the ability to separately model marginal distributions and the dependence structure, where the later is modeled using graphical trees and non-Gaussian bivariate building blocks (copulas), allows R-vines to capture asymmetric and heavy tailed dependencies.

In addition, we propose novel risk measures from the vine copula based regression models, which we use to identify spatial and temporal "at-risk" regions for forest ecosystems. We also suggest corresponding survival probabilities, that can identify "at-risk" spatial regions over longer periods of time and a corresponding return periods of extreme events, that can identify temporal "at-risk" regions. To our knowledge, vine copulas have not yet been investigated in such a climatological application.

6.2. Data description

To quantify changing drought and frost risk we use a late-frost index and a drought index rather than raw climate variables. We calculate these indices using the BayObs product, a multivariate, gridded climate data set covering Bavaria at a spatial resolution of 5km by 5km, provided by the Bavarian Environment Agency (LfU). The dataset contains daily minimum air temperature, daily maximum air temperature, daily mean air temperature, and daily precipitation sum from 1952 until 2020 (Bayerisches Landesamt für Umwelt [Hrsg.], 2020).

Late-frost index

To quantify frost risk we use a modified version of the Frost Index in April (FI4) proposed by Sangüesa-Barreda et al. (2021). The original FI4 takes into account mean and minimum temperatures between mid-April and mid-May, a time period which generally marks the beginning of leaf-unfolding in European beech. In contrast, our modified index, Frost Index at Leaf-Out (FILO) uses a phenological model to more accurately pinpoint the begin of leaf-unfolding. We use the phenological model outlined

in Kramer et al. (2017). A frost index having a value of 0 indicates average conditions (i.e. average frost risk), positive values indicate a lower frost risk, and negative values indicate a higher frost risk.

Drought index

To determine drought risk we use the Standardized Precipitation Evapotranspiration Index (SPEI). This index describes the relative water availability at a given site and time as a function of precipitation and potential evapotranspiration, i.e. the difference between water supply and water demand (Vicente-Serrano et al., 2010; Beguería et al., 2014). Negative SPEI values indicate drier-than-average conditions while positive values indicate wetter-than-average conditions. The SPEI is standardized across the entire period of historical climate data available in the BayObs data set (1952-2020). Here, we focus on the SPEI-6 in August, that is, the SPEI integrated over August and the preceding 5 months. This allows us to take into account medium-term droughts spanning from early spring to the height of summer which are critical in capturing the drought signal of European beech (Bhuyan et al., 2017).

Climatic and topographic predictors

Previous studies have identified the possible effect of factors such as elevation, aspect, annual precipitation, and mean annual temperature on the spatial incidence of late-frost events (Olano et al., 2021). Similarly, topography (elevation, aspect) and a combination of short- and long-term temperature and precipitation conditions have been shown to cause a deficit of water i.e. drought (Bhuyan et al., 2017; Van Loon, 2015). To identify factors which influence both late-frost and drought we utilized a set of bioclimatic indices as well as a set of topographic indices. The bioclimatic indices are based on the bioclimatic variables derived from the WordClim database (Fick and Hijmans, 2017; Hijmans et al., 2005). Since we are interested in intra-annual fluctuations of precipitation and temperature patterns, we derive these indices on a yearly basis. We first aggregated our daily climate data (precipitation, min. temperature, max. temperature, mean temperature) to monthly values. Subsequently, we calculated the annual bioclimatic variables using the R package `dismo` (Hijmans et al., 2021).

For the topographic predictors (elevation, slope, aspect), we extracted relevant terrain information from the digital surface model (DSM) EU-DEM v1.0 provided by the European Environment Agency (EEA) under the Copernicus program (publicly available at http://land.copernicus.eu/pan-european/satellite-derived-products/eu-dem/eu-dem-v1-0-and-derived-products/eu-dem-v1.0/view. We reprojected the EU-DEM from its native resolution of 25 m and ETRS89 reference system to a resolution of 5 km and

a WGS84 reference system to match our climate data. We then extracted slope and aspect information from the DSM using the R package `terra` (Hijmans, 2022). Also, we include the gridcell specific location by including latitude and longitude in the model.

Data summary

Overall, we have produced a data set containing annual data for 69 years (1952-2020) for each of the 2867 gridcells considered in the region of Bavaria, Germany. For each year and location (or gridcell) there are 26 available variables in total (2 responses, 19 bioclimatic predictors and 5 topographic predictors). Thus, in total the data set has a size of $197823 (= 2867 \cdot 69)$ data points. In Table 6.1 we give a short description of the variables used in our data analysis. Further, in Figure C.1 given in Appendix C.1, panels for each variable (apart from the topographic variables) in which the data is summarized are shown. In each left panel for each variable, shown are the annual mean observation over all gridcells per each year in the period 1952-2020. The smoothed line is the fitted moving averages model and the shaded area is the corresponding 95% confidence interval (CI) for each variable. Using these plots we can observe how the mean values change over the historical period and what is the trend for each variable. For example, we can observe the increasing trend for the `temp_mean` or `temp_warm` over the years or the clearly decreasing trend of `preci` over the last 20 years (2000-2020). Also, since our main goal is understanding the drought and frost indices, we can easily see the decreasing trend for both of these variables in the last 20 years (2000-2020), which implies worse frost and drought conditions. Also, we can identify outliers in these variables, years which had low average values of frost and drought indices. For example, the lowest value of the annual average frost index is achieved in the year 2011, while the lowest value of the drought index is in the year 2003.

The right panels of Figure C.1 for each variable, shows the annual mean averages over all gridcells, the corresponding 95% and 50 % confidence interval per year over all gridcells. This plot gives more information about the variability of each variable over all the possible gridcells for each year. For example, for `temp_wet` and `temp_dry` we can see very big variability in the observations for some years, while for others the majority of the observations are very close to each other. This implies that there are years in which these two variables vary over the locations we have considered, while for others, almost all the locations exhibit similar conditions.

In addition, Appendix C.2 shows marginally normalized contour plots, where the marginal distributions are fitted in a nonparametric manner, using kernel density smoothing, for two randomly chosen years, at the beginning of our analysis, year 1953 (first 3 plots) and at the end of the analysis, year 2011 (last 3 plots). Each plot is based on all 2867 locations for the two responses and a subset of the predictors. On the lower

diagonal, any deviance from elliptical shapes indicates a non-Gaussian dependence structure in the data (see Section 3.8 of Czado (2019) for a precise definition) and we see that almost all panels imply non-Gaussian dependence structures. In addition, on the upper diagonal, we see a scatter plot of the estimated u-data together with the corresponding estimated empirical pairwise Kendall's $\hat{\tau}$. In all 6 figures, we can see many high values of the pairwise Kendall's $\hat{\tau}$. Similar results follow for the other years, i.e. majority of non-Gaussian dependence between pairs of variables and high estimated empirical pairwise Kendall's $\hat{\tau}$ are detectable for all years considered. However, using vine copulas we can efficiently model and capture these high non-Gaussian dependencies between pairs of variables.

6.3. Data modeling

In order to examine what is the effect of the possible predictors on the late-frost and drought indices and how it changes over time, we fit a D-vine regression model for each of the two responses (introduced in Chapter 3, with one-step ahead forward selection of predictors) and a Y-vine regression model for their joint behavior (introduced in Chapter 4). The models are fitted for each year separately, using parametric bivariate copula families with a single parameter (Section 2.1.2), an AIC-penalized log likelihood selection criteria on the choice of the copula family and the marginal distributions are fitted in a nonparametric manner, using kernel density smoothing. The use of the parametric copula families is due to the need for quantifying and analyzing the tail dependence in the models. Each model is set to find the 5 most influential predictors for each year. This is done because of computational limitations, due to the large size of the data set and the number of models to be fitted (in total $69 \cdot 3 = 207$ vine copula regression models are fitted). Thus, we have:

- **Data periods:** 69 years, $t \in [1952 - 2020]$.
- **Locations:** Per year there are 2867 gridcells where the climatological variables are evaluated, $l \in [1, \ldots, 2867]$. We model the spatial effect in the data by including as possible predictors the gridcell spatial coordinates, i.e. the `latitude` and `longitude`.
- **Univariate D-vine models:** For each year, two D-vine regression models are fitted on all 2867 grid points. The ones where the response is the frost index we denote as $\hat{\mathcal{D}}_{frost_{1952}}, \ldots, \hat{\mathcal{D}}_{frost_{2020}}$. The drought index is a response variable in the models $\hat{\mathcal{D}}_{drought_{1952}}, \ldots, \hat{\mathcal{D}}_{drought_{2020}}$.

- **Bivariate Y-vine models:** For each year, we fit a joint response Y-vine model over all grid points. The fitted Y-vine models we denote as $\hat{\mathcal{Y}}_{1952}, \ldots, \hat{\mathcal{Y}}_{2020}$.

Variable name	Description
Responses	
frost	late-frost index at leaf out
drought	drought index
Bioclimatic variables	
temp_mean	Annual mean temperature
temp_diu	Mean diurnal range (mean of monthly max temp - min temp)
isotherm	Isothermality
temp_season	Temperature seasonality
temp_max	Temperature of warmest month
temp_min	Temperature of coldest month
temp_season	Temperature annual range
temp_wet	Mean temperature of wettest quarter
temp_dry	Mean temperature of driest quarter
temp_warm	Mean temperature of warmest quarter
temp_cold	Mean temperature of coldest quarter
preci	Total (annual) precipitation
preci_wet_m	Precipitation of wettest month
preci_dry_m	Precipitation of driest month
preci_season	Precipitation seasonality (coefficient of variation)
preci_wet_q	Precipitation of wettest quarter
preci_dry_q	Precipitation of driest quarter
preci_warm	Precipitation of warmest quarter
preci_cold	Precipitation of coldest quarter
Topographic variables	
elevation	Average elevation above sea level (in meters)
aspect	Aspect of each gridcell in degrees (0° = north)
slope	Average slope of each gridcell in degrees
latitude	latitude of gridcell
longitude	longitude of gridcell

Table 6.1.: Variable description.

6.3.1. Dependence analysis

Unconditional dependence analysis between drought and late-frost indices

In the top panel of Figure 6.1 we analyze the unconditional dependence between the late-frost index *(f)* and the drought index *(d)* using the Kendall's τ measure of dependence (introduced in Section 2.1.1), denoted as $\tau_{f,d}$. Shown are the estimated $\tau_{f,d}$ values over all gridcells per each year in the period 1952-2020. The smoothed line is the fitted moving average model and the shaded area is the corresponding 95% confidence interval. We observe how this dependence evolves over the years. At the beginning, in the period approximately 1952-1969 there is an increasing trend in the dependence of the late-frost and drought index, until it reaches the maximal value of the dependence, Kendall's $\hat{\tau}_{f,d} = 0.62$ in year 1969. This values is the maximal Kendall's $\hat{\tau}_{f,d}$ value in absolute value as well. Then there is decreasing trend for a short period, and then, in the next period around 1980-2000, this trend stabilizes, as there are similar smaller dependencies, but in opposite directions, some being positive, while others negative.

In the next period, the dependence has a decreasing trend and reaches the minimal value of dependence, Kendall's $\hat{\tau}_{f,d} = -0.33$ in year 2007. The minimal Kendall's $\hat{\tau}$ in absolute value is reached in year 1970 and it is approximately zero $|\hat{\tau}_{f,d}| = 0.0004$. Finally, in the last 5 years there is an increasing trend. Note that the blue vertical ribbons represents years identified as extreme only by the frost D-vine model, the apricot colored vertical ribbon represents years identified only by the drought D-vine model, purple ribbons represents years identified by the joint Y-vine model and light gray (only year 1953) ribbon is where both the univariate models identifies risks, but not the joint Y-vine model. More details on how we identify these extreme risk years is explained later in Section 6.4.

Next, we fit a bivariate copula model on the late-frost index and the drought index over all average annual observations in the period 1952-2020 to analyze the overall dependence between the two responses. The fitting is done using a parametric pair copula family with a single parameter with an AIC-penalized log likelihood selection criteria. The selected pair copula is a Gaussian copula with parameter value of 0.34 and estimated Kendall's $\hat{\tau}_{f,d} = 0.22$. This implies that there is relatively small overall dependence between the late-frost and drought index, and there is no tail dependence modeled between these two variables (without considering predictor variables).

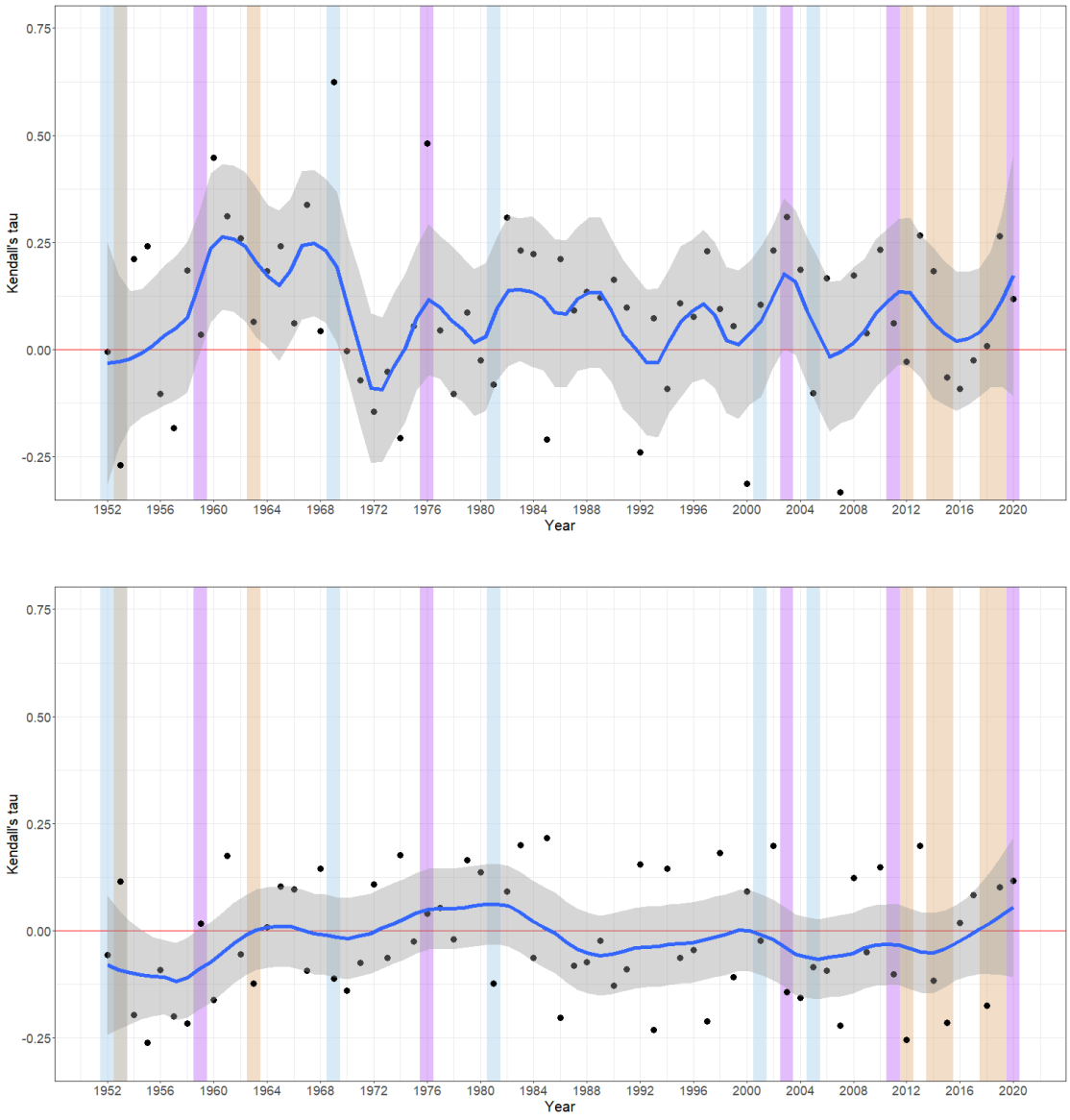


Figure 6.1.: x-axis: 1952-2020 year. y-axis: Top row: unconditional Kendall's $\hat{\tau}_{f,d}$ value, bottom row: conditional $\hat{\tau}_{f,d;\mathbf{u}}$. (The black points denote the estimated values at each year, the red horizontal line denotes when Kendall's $\hat{\tau} = 0$, which indicates independence. The blue line is the smoothed regression line and it's 95% confidence interval. The vertical ribbons denote extreme years identified for frost risk (blue), drought risk (apricot), joint frost and drought risk (purple), and marginal drought and frost risk, but not joint risk identified(light gray).)

Conditional dependence analysis between drought and late-frost indices

We take a look at the conditional dependence between the frost and drought indices in the bottom panel of Figure 6.1. After fitting the Y-vine models of Chapter 4 to the data, for each year there is a pair copula fitted between the frost and drought indices, conditioned on the chosen 5 predictors. This corresponds to the last pair copula fitted in the Y-vine model. For each of these fitted pair copulas, we extract the estimated value of the Kendall's τ, denoted as $\tau_{f,d;\mathbf{u}}$ and we plot it for each year. Here we see a different trend in the conditional dependence, than in the unconditional dependence between the frost and drought indices. In the period 1952-1985 there is an overall increasing trend in the conditional dependence, reaching a maximal value in year 1985 of $\hat{\tau}_{f,d;\mathbf{u}} = 0.22$. Afterwards, there is a decreasing trend in their dependence until around the 2000s, after which an increasing trend follows again. The maximal absolute value is reached in year 1955 of $|\hat{\tau}_{f,d;\mathbf{u}}| = 0.26$, which is the minimal overall value as well. The minimal absolute value is reached in year 1964 and it is $|\hat{\tau}_{f,d;\mathbf{u}}| = 0.008$.

Pair copula families

To amplify the benefits of the usage of vine copula models on the data, whose main advantage is modeling non-Gaussian relationships with tail and asymmetric dependencies, we analyse how many of the selected pair copula families are Gaussian pair copulas and how many are non-Gaussian. In the fitted Y-vine models $\hat{\mathcal{Y}}_t$ there are in total 21 fitted pair copulas, and in the fitted D-vine models, $\hat{\mathcal{D}}_{frost_t}$ and $\hat{\mathcal{D}}_{drought_t}$ there are 15 fitted pair copulas for $t \in [1952 - 2020]$. In Figure 6.2 we show how many times, in each model for each year, Gaussian copula (rotations included) is been fitted and how many time a non-Gaussian copula is fitted (the choices are: Clayton, Gumbel, Frank, Joe, and their rotations as well). The non-Gaussian fitted pair copulas are shown with red color, while the Gaussian pair copulas are shown with blue color. We observe that the red color is much more pronounced in all 3 models.

On average for the frost D-vine models $\hat{\mathcal{D}}_{frost_t}$ for all $t \in [1952 - 2020]$, there are 11% of Gaussian pair copulas, and 89% of non-Gaussian pair copulas fitted. For the drought D-vine models $\hat{\mathcal{D}}_{drought_t}$ for all $t \in [1952 - 2020]$, there are 15% of Gaussian pair copulas, and 85% of non-Gaussian pair copulas fitted. For the joint model of frost and drought, the Y-vine model $\hat{\mathcal{Y}}_t$ for $t \in [1952 - 2020]$, there are 12% of Gaussian pair copulas, and 88% of non-Gaussian pair copulas fitted. Thus, due to the fact that the majority of the fitted pair copulas are non-Gaussian, it follows that the overall dependence structure in the data is non-Gaussian and vine copulas succeed to capture this dependence.

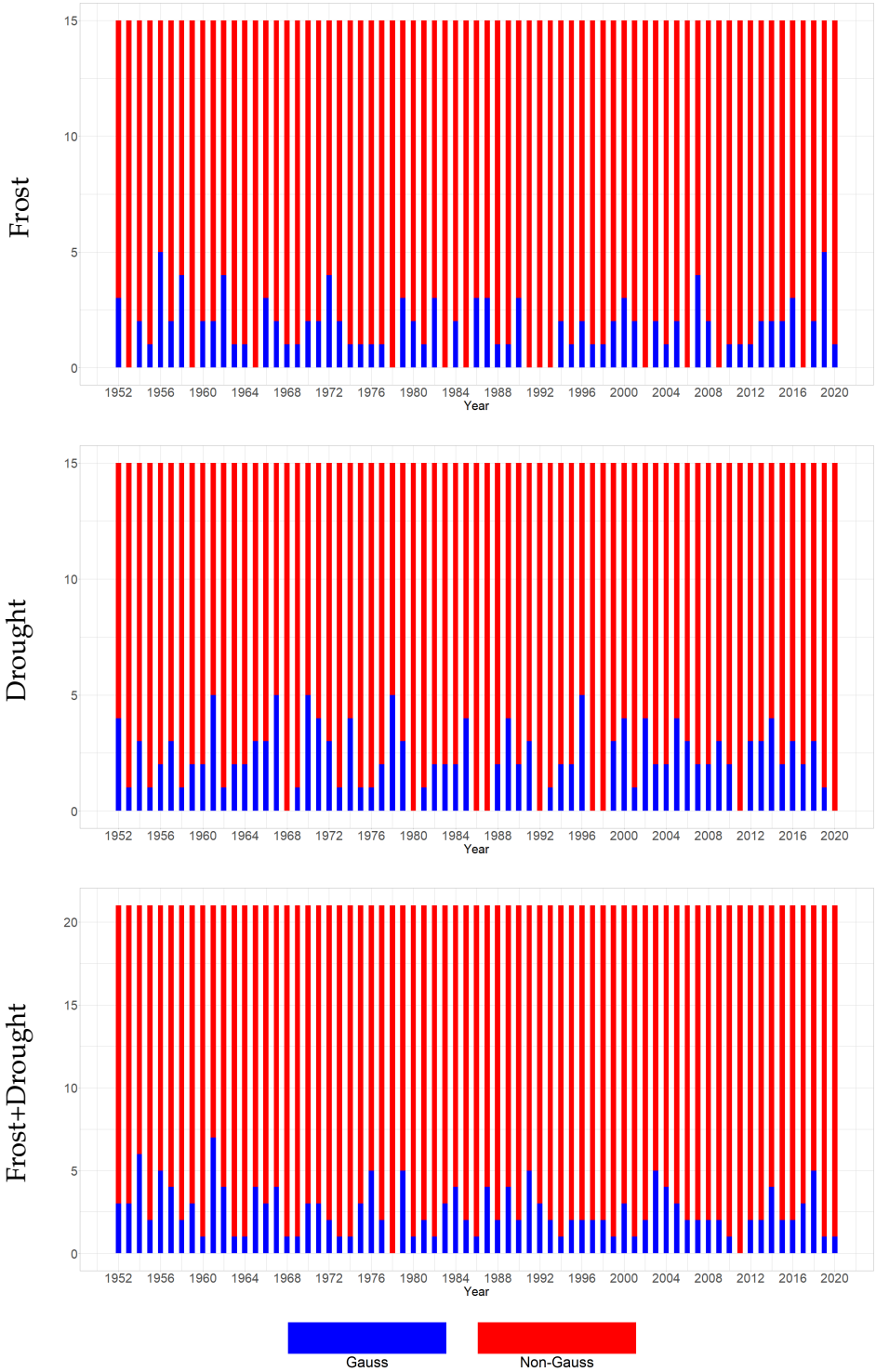


Figure 6.2.: Count of the fitted Gaussian pair copulas, shown in blue and non-Gaussian pair copulas (including rotations), shown in red.

6.3.2. Order analysis of selected predictors

Each of the fitted models selects the 5 most influential predictors for the frost and drought indices, and the 5 most influential predictors for the joint modeling of them. To analyse how the influence of the predictors vary over the years and which of the possible predictors are chosen the most over the historical period from 1952 to 2020, we show in Figure 6.3 for each year which 5 predictors are chosen by each model. Their influence depends on the position in the order. The first predictor in the order is the most influential one on the response/s, the second in the second most influential one and so on.

The 5 predictors that are chosen the most by each of the 69 models, not taking into account the position they are chosen in, are the following:

- $\hat{\mathcal{D}}_{\mathbf{frost_t}}$ **for all t** $\in [\mathbf{1952 - 2020}]$: `longitude` (43 times), `latitude` (35 times), `temp_min` (23 times), `isotherm` (each 22 times) and `temp_mean` (21 times);
- $\hat{\mathcal{D}}_{\mathbf{drought_t}}$ **for all t** $\in [\mathbf{1952 - 2020}]$: `latitude` (44 times), `longitude` and `preci_warm` (36 times each), `elevation` (30 times), `preci_wet_q` (23 times), `preci_season` (22 times);
- $\hat{\mathcal{Y}}_{\mathbf{t}}$ **for all t** $\in [\mathbf{1952 - 2020}]$: `latitude` (41 times), `longitude` (39 times), `preci_warm` (30 times), `temp_min` (23 times), `preci_wet_q` and `isotherm` (each 19 times).

The optimal order for each model is defined as the order in which the first element of the order corresponds to the predictor that appeared the most in the first position over the 69 models, then the second element is defined as the element that appeared the most in the second position among the elements not chosen as first and so on (defined and used in Section 3.5.1). The optimal orders for each model are given in Table 6.2, together with how many times the chosen predictor is selected to be in a given position for all 5 possible positions in the order. Note that for some positions in the order 2 predictors appear the same number of times in a particular position, for example in the optimal order for $\hat{\mathcal{D}}_{frost_t}$ both `temp_range` and `temp_season` appear in the fourth position in the order 7 times each out of the 69 possible orders.

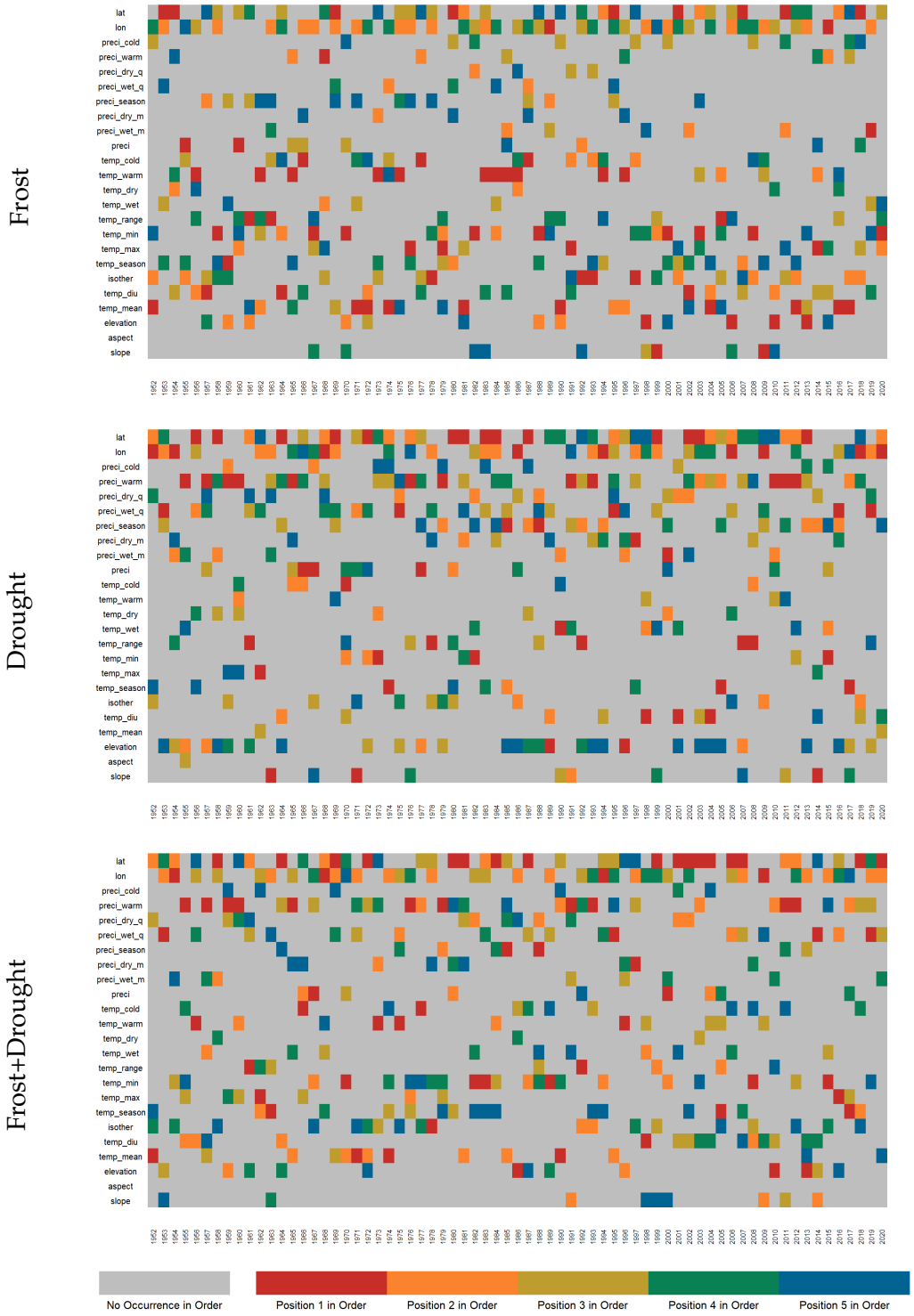


Figure 6.3.: Orders of the fitted annual models.

Model	1	2	3	4	5
$\hat{\mathcal{D}}_{frost_t}$	temp_warm(11)	lon(17)	lat(10)	temp_range, temp_season (7)	preci_season(7)
$\hat{\mathcal{D}}_{drought_t}$	lat(14)	lon (11)	preci_warm(10)	preci_wet_q(7)	elevation(13)
$\hat{\mathcal{Y}}_t$	lat (17)	lon (14)	preci_warm, preci_wet_q(6)	isotherm(6)	temp_season (8)

Table 6.2.: The optimal orders for each model over all years, together with the count of appearances of the predictor in a certain position in the order.

Out of the chosen predictors for all 3 models, we can conclude that the spatial effects have very influential role, as `latitude` and `longitude` are chosen by all 3 models in the optimal orders. Further, for the $\hat{\mathcal{D}}_{frost_t}$ influential predictors are also the temperature based predictors, as `temp_min`, `temp_warm`, `temp_range`. For the $\hat{\mathcal{D}}_{drought_t}$ influential predictors are the precipitation based predictors, as `preci_warm`, `preci_wet_q`, and also the `elevation` spatial predictor. For the joint Y-vine regression models $\hat{\mathcal{Y}}_t$ influential predictors are both temperature based predictors, such as `temp_min`, `isotherm`, but also precipitation based predictors, as `preci_wet_q`, `preci_warm`.

6.4. Univariate and bivariate conditional probability risk measures of extreme events

For the fitted vine models, we propose a probability risk measure, which is defined as the conditional probability of the random variable/s to be less that a specified threshold given the predictors. Denote the frost index random variable as $Y_{f,t,l}$ and the drought index random variable as $Y_{d,t,l}$ for year $t \in [1952, \ldots, 2020]$ at gridcell (location) $l \in [1, \ldots, 2867]$. The corresponding marginal distribution functions are denoted as $F_{Y_{f,t,l}}$ and $F_{Y_{d,t,l}}$ respectively. Denote the 5 ordered predictors chosen by each model as the vector $\mathbf{X}_{t,l} = (X_{1,t,l}, \ldots, X_{5,t,l})^T$ with corresponding marginal distribution functions $F_{1,t,l}(X_{1,t,l}), \ldots, F_{5,t,l}(X_{5,t,l})$ for $t \in [1952, \ldots, 2020]$ and $l \in [1, \ldots, 2867]$.

6.4.1. General framework

Given a threshold vector $\mathbf{p} = (y_f, y_d)^T$, we define the conditional probability of the occurrence of an average observation of $Y_{f,t,l} \leq y_f$ given a set of predictors $\mathbf{X}_{t,l}$, i.e. $P(Y_{f,t,l} \leq y_f | \mathbf{X}_{t,l})$, within a time period t for all l, as a risk measure for the occurrence of frost. This conditional probability is estimated as

$$\hat{P}(Y_{f,t,l} \leq y_f | \mathbf{X}_{t,l}) := C_{\hat{\mathcal{D}}_{frost_t}}\left(V_{f,t,l} \leq F_{Y_{f,t,l}}(y_f) \,|\mathbf{U}_{t,l} = \mathbf{u}_{t,l}\right), \tag{6.1}$$

where $V_{f,t,l} = F_{Y_{f,t,l}}\left(Y_{f,t,l}\right)$, and $\mathbf{U}_{t,l} = \left(F_{1,t,l}\left(X_{1,t,l}\right), \ldots, F_{5,t,l}\left(X_{5,t,l}\right)\right)^T$ coming from the estimated $\hat{\mathcal{D}}_{frost_t}$ model in the chosen order. We denote the right hand side of Equation (6.1) as $\hat{P}_{\hat{\mathcal{D}}_{frost_t}}\left(y_f|\mathbf{x}_{t,l}\right)$. Following the same analogy, a risk measure for the occurrence of drought, $\hat{P}_{\hat{\mathcal{D}}_{drought_t}}\left(y_d|\mathbf{x}_{t,l}\right)$ is defined, where $\mathbf{x}_{t,l}$ contains the observations of the 5 chosen predictors in the order of the $\hat{\mathcal{D}}_{drought_t}$ model for each time t and location l.

The joint risk measure for the joint occurrence of frost and drought, given a threshold vector $\mathbf{p} = \left(y_f, y_d\right)^T$, is the conditional probability of the joint occurrence of an average observation of $Y_{f,t,l} \leq y_f$ and $Y_{d,t,l} \leq y_d$ given a set of predictors, within a time period t. We estimate it as

$$\hat{P}\left(Y_{f,t,l} \leq y_f, Y_{d,t,l} \leq y_d|\mathbf{X}_{t,l}\right) \coloneqq C_{\hat{\mathcal{Y}}_t}\left(V_{f,t,l} \leq F_{Y_{f,t,l}}\left(y_f\right), V_{d,t,l} \leq F_{Y_{d,t,l}}\left(y_d\right)|\mathbf{U}_{t,l} = \mathbf{u}_{t,l}\right), \tag{6.2}$$

where $V_{d,t,l} = F_{Y_{d,t,l}}\left(Y_{d,t,l}\right)$ and $\mathbf{u}_{t,l}$ contains the observations of the 5 chosen predictors in the order of the $\hat{\mathcal{Y}}$ model for each time t at location l. We denote the right hand side of Equation (6.2) as $\hat{P}_{\hat{\mathcal{Y}}_t}\left(\mathbf{p}|\mathbf{x}_{t,l}\right)$.

In order to evaluate the proposed conditional probabilities, we chose the threshold to be $\mathbf{p} = \left(y_f, y_d\right) = (-2, -1.5)$. Thus, we determine the estimates of $\hat{P}_{\hat{\mathcal{D}}_{frost_t}}\left(-2|\mathbf{x}_{t,l}\right)$, $\hat{P}_{\hat{\mathcal{D}}_{drought_t}}\left(-1.5|\mathbf{x}_{t,l}\right)$ and $\hat{P}_{\hat{\mathcal{Y}}_t}\left(-2, -1.5|\mathbf{x}_{t,l}\right)$, where for each model the conditioning values $\mathbf{u}_{t,l}$ differ. For the drought index, we set the threshold at -1.5 which represents a commonly accepted threshold beyond which drought conditions are classified as severely to extremely dry (Slette et al., 2019). Since such a commonly accepted threshold is not available for the frost index, we set the threshold at -2, signifying two standard deviations below the mean of the frost index. The greater these conditional probabilities are, the greater the chances are of 'extreme' drought, frost or joint frost and drought risk.

Thus, to summarize we have the following risk measures estimated for all gridcells and all years :

- $\hat{P}_{\hat{\mathcal{D}}_{frost_t}}\left(-2|\mathbf{x}_{t,l}\right)$: estimated conditional risk probability of extreme frost;
- $\hat{P}_{\hat{\mathcal{D}}_{drought_t}}\left(-1.5|\mathbf{x}_{t,l}\right)$: estimated conditional risk probability of extreme drought;
- $\hat{P}_{\hat{\mathcal{Y}}_t}\left(-2, -1.5|\mathbf{x}_{t,l}\right)$: estimated conditional risk probability of joint extreme frost and drought.

6.4.2. Results

The estimated conditional probabilities from the 3 fitted vine regression models are given in Figures 6.4, 6.5 and 6.6. Per year the conditional probabilities are estimated for each location, and subsequently their 90% and 50% confidence intervals and means are shown in the left panel of each figure. Additionally, the 98% and 90% confidence intervals and means (top-right panels) are shown for three selected time periods, i.e. years 1952-1974, 1975-1997 and 1998-2020. Further, a year is considered to have an **"annual extreme event occurrence" if the 0.95 empirical quantile of the estimated conditional risk probability for that year over all gridcells is greater than 0.2**, in which case at least 5% of the locations have a 20% or higher chance of an extreme event occurring. The number of years which exhibit such extreme occurrences, within each of the considered time periods, are shown in the bottom-right panels of each figure.

Extreme frost events

Figure 6.4 shows a summary of the conditional probabilities associated with an extreme frost event. It indicates that in approximately 23 years out of 69 there is a non zero probability of extreme events happening in at least 5% of the locations in Bavaria. As the means of extreme events are fairly constant between the three 23 year periods, there seems not to be a significantly increased risk of extreme frost between the first and third period. Also, there is a total of 4 annual extreme events in the first period, 2 such in the second and 4 in the third period, resulting in a total of 10 annual extreme events for the frost risk. These identified risky years for the frost are: 1952, 1953, 1959, 1969, 1976, 1981, 2001, 2003, 2005, 2011.

Extreme drought events

Next, Figure 6.5 shows a summary of the conditional probabilities associated with an extreme drought event. Again, quite often there is a non-zero probability of extreme drought occurs in at least 5% of the locations in Bavaria, in approximately 26 years out of 69. However, in contrast to frost, there is a clear increase in the mean conditional probability of an extreme drought event occurring, as well as the frequency of occurrence. In the period between 1998 and 2020 there is almost every year a high probability of an extreme drought event occurring over all locations. Further, the frequency of the annual extreme events increases as well, with 3 extreme events in the period 1952-1974 and a single extreme event in the period 1975-1997, to a total of 7 extreme events in 1998-2020. These identified risky years for the drought index are: 1953, 1959, 1963, 1976, 2003, 2011, 2012, 2014, 2015, 2018, 2019.

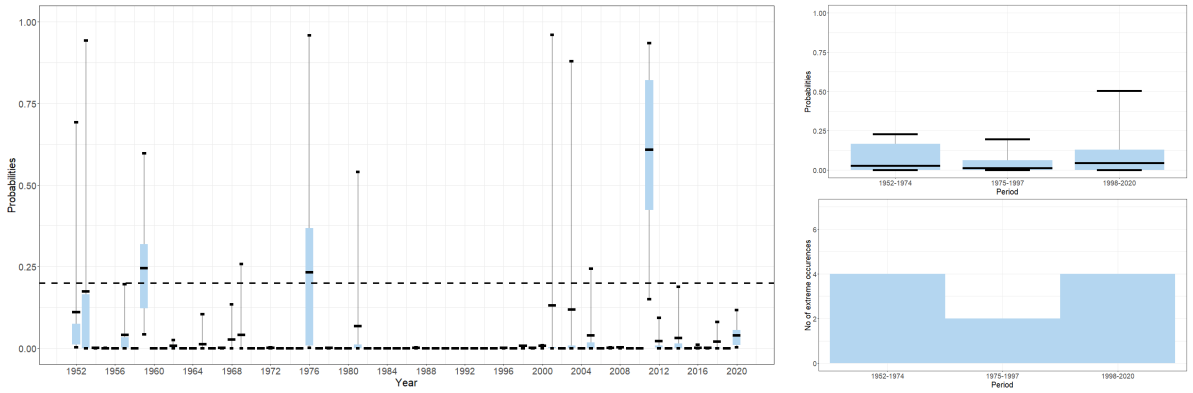


Figure 6.4.: Conditional probabilities of **annual extreme frost** occurring, $\hat{P}_{\hat{\mathcal{D}}_{frost_t}}(-2|\mathbf{x}_{t,l})$.

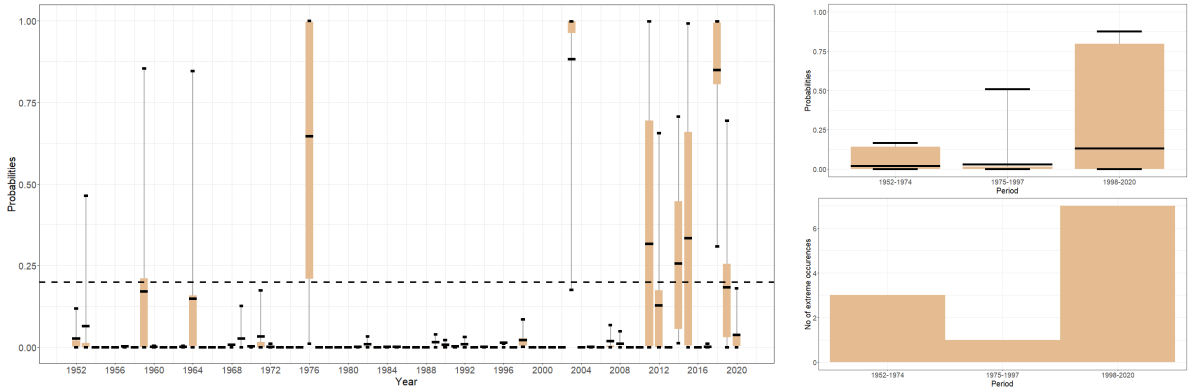


Figure 6.5.: Conditional probabilities of **annual extreme drought** occurring, $\hat{P}_{\hat{\mathcal{D}}_{drought_t}}(-1.5|\mathbf{x}_{t,l})$.

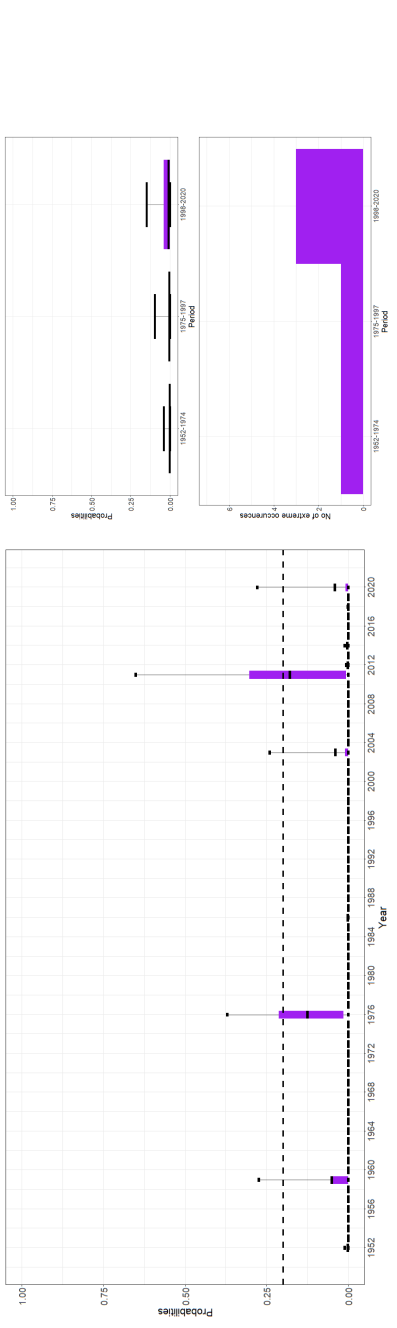


Figure 6.6.: Conditional probabilities of **annual jointly extreme frost and drought** occurring, $\hat{P}_{\hat{y}_t}(-2, -1.5|\mathbf{x}_{t,l})$.

Extreme frost and drought events

Figure 6.6 gives a summary of the joint conditional probabilities of both, extreme frost and extreme drought events, occurring. There is a clear increase in the mean of the joint conditional probabilities, approximately by a factor of three, as well as a significant increase in occurrence frequencies between the last period (1998-2020) and the first two (1952-1974 and 1975-1997). A significantly high risk of annual jointly extreme events is present in the years 1959, 1976, 2003, 2011 and 2020. Furthermore, in the years that there is an increased joint risk, there are also increased marginal risks of frost and drought. Basically, the joint Y-vine model identified 5 very extreme joint events, which were also identified by the univariate D-vine models for both the frost and drought risks, except for the year 2020, where both the conditional frost and drought risks are just below the threshold for an extreme event, but it is a quite high non-zero value.

The years that are identified by each model are also highlighted with a vertical ribbon in the background of Figures 6.1 and all figures in Appendix C.1. The blue ribbon represents years identified only by the frost D-vine model, the apricot colored ribbon represents years identified only by the drought D-vine model, purple ribbons represents years identified by the joint Y-vine model and light gray (only year 1953) ribbon is where both the univariate models identifies risks, but not the joint Y-vine model.

Figure 6.7 shows the estimated conditional probabilities of frost $\hat{P}_{\hat{\mathcal{D}}_{frost_t}}(-2|\mathbf{x}_{t,l})$, drought $\hat{P}_{\hat{\mathcal{D}}_{drought_t}}(-1.5|\mathbf{x}_{t,l})$ and joint events $\hat{P}_{\hat{\mathcal{Y}}_t}(-2,-1.5|\mathbf{x}_{t,l})$, for the years 1959, 1976, 2003, 2011 and 2020, for each of the considered locations or gridcells in Bavaria. It is interesting to note that, at a given location, if there are high univariate conditional probabilities of frost and drought, there is not necessarily a high bivariate conditional probability of a joint event. An example is the north of Bavaria in the year 2003. Despite having high chances of frost and drought individually, there is almost no chance of a joint event occurring at those locations. This indicates that separate assessment of risks, interpreted together will likely overestimate the joint risk and fail to detect true regions of interest.

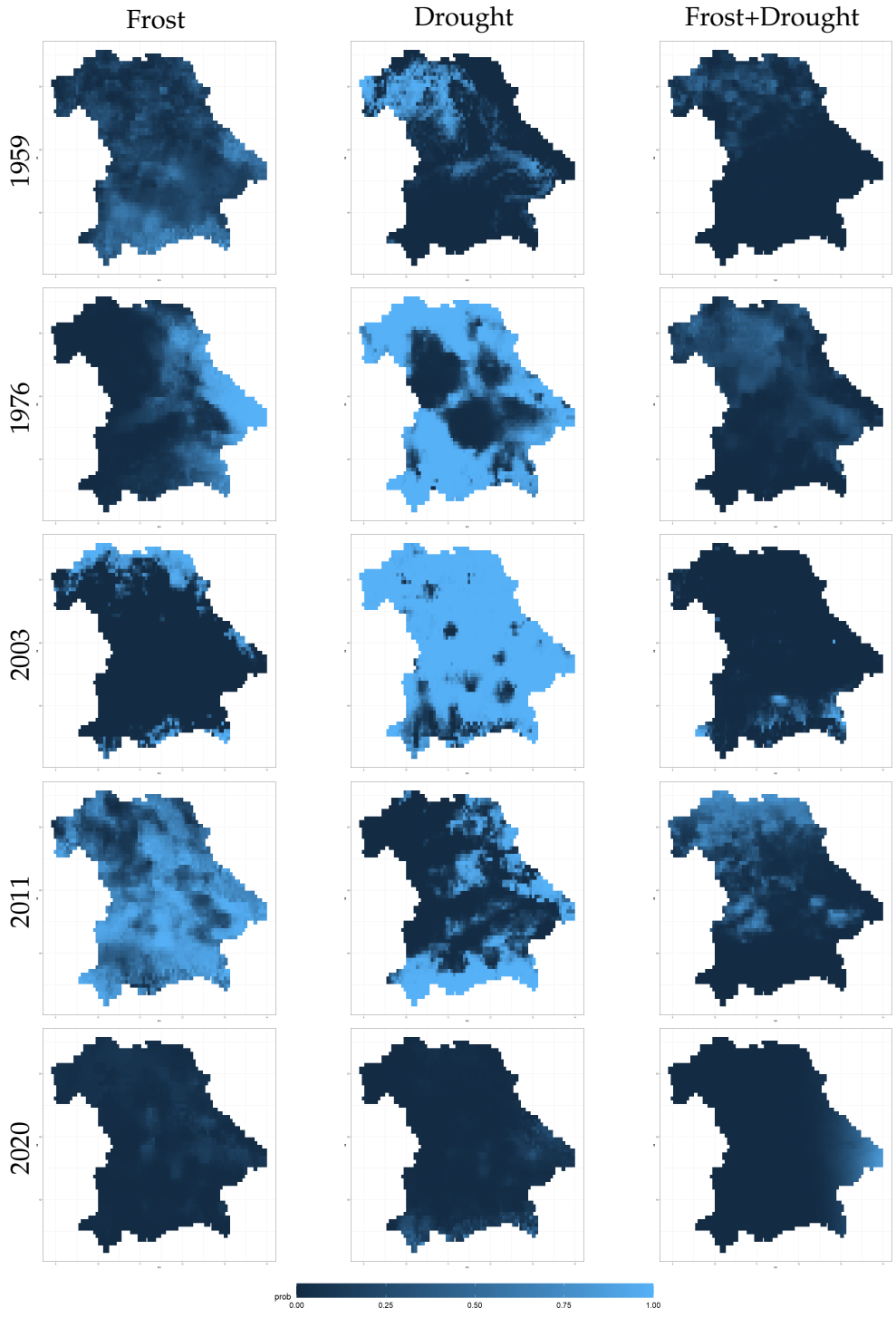


Figure 6.7.: $\hat{P}_{\hat{\mathcal{D}}_{frost_t}}(-2|\mathbf{x}_{t,l})$, $\hat{P}_{\hat{\mathcal{D}}_{drought_t}}(-1.5|\mathbf{x}_{t,l})$ and $\hat{P}_{\hat{y}_t}(-2,-1.5|\mathbf{x}_{t,l})$, for the years identified as extreme by the joint Y-vine model for all gridcells.

6.5. Survival probabilities

The survival function is the complement of the cumulative distribution function, and it gives the probability that an event will not occur. We determine the survival function of the event starting from time s until the time period T by subtracting from 1 the sum of the estimated conditional probabilities occurring in each year between time s and $T \leq 2020$ (more in Klein et al., 1997, Chapter 2). Then, the survival probability for the frost index is estimated as

$$\hat{S}_{\hat{\mathcal{D}}_{frost}}(s,T) := 1 - \sum_{t=s}^{T} \hat{P}_{\hat{\mathcal{D}}_{frost}}\left(-2|\mathbf{x}_{t,l}\right). \tag{6.3}$$

In the same manner, the survival probability for the drought index $\hat{S}_{\hat{\mathcal{D}}_{drought}}(s,T)$ is defined. For the joint bivariate case, the survival probability is estimated as

$$\hat{S}_{\hat{\mathcal{Y}}}(s,T) := 1 - \sum_{t=s}^{T} \hat{P}_{\hat{\mathcal{Y}}_t}\left(-2,-1.5|\mathbf{x}_{t,l}\right). \tag{6.4}$$

Figure 6.8 shows the estimated survival probabilities for all considered locations for $s = 1952$ and $T = 1975, 1998, 2020$, going from top to bottom. The survival probabilities can be interpreted as the probability of a location not experiencing an extreme event in the time periods 1952-1975 (top row), 1952-1998 (middle row) and 1952-2020 (bottom row). Yellow regions have survival probabilities close to 0, and those regions have a high risk of an extreme event happening beyond the year T and purple regions have survival probabilities close to 1, meaning that those regions have a low risk of an extreme event happening beyond a given year T. For example, purple coloured locations (i.e survival probability ≥ 0.75) in the middle plot indicate that the estimated probability of an extreme drought event not occurring in the time period 1952-1998 is above 75%. With the increase in size of the time interval considered, the survival probability is expected to decrease, i.e. the longer the period of observation the higher the chances of a death event. Therefore, the bottom three plots are expected to exhibit lower estimated survival probabilities. However, it is interesting that despite the estimated survival probabilities of almost all locations are close to 0 for both frost and drought indices in the last row, the joint estimated survival probability is above 0.75 for more then half of the locations in Bavaria, with only the Northern and Eastern regions exhibiting a close to 0 joint survival probability. This implies that considering only the marginal models for frost and drought risks, we would not have been able to properly locate "at-risk" regions for the joint occurrence of these two extremes.

For a better comparison, we also consider 3 equal length time periods (1952-1974, 1975-1997, 1998-2020). In Equations (6.3) and (6.4) the corresponding pair (s, T) is evaluated at the pairs $(1952, 1974), (1975, 1997)$, and $(1998, 2020)$. Figure 6.9 shows the estimated survival probabilities between those specific periods. In contrast to Figure 6.8, the considered periods are of equal length and give rise to survival probabilities that are comparable between each other. Also, from these plots we can see the influence each period has on the associated overall survival risk.

The period 1952-1974 has low estimated survival probabilities, i.e. high risks of extreme even occurrences, in approximately half of the considered locations, while the other half has quite high estimated survival probabilities for both frost and drought. There are almost no chances of a joint occurrence in the majority of locations and the extreme joint events seem to be located in the norther borders of Bavaria. In the next period, between 1975-1997, there is a small risk for a frost event, appart from the east regions of Bavaria. The estimated survival probabilities for the drought are quite low in the majority of the locations considered, indicating that this period had higher drought risks associated than the previous period considered. The joint estimated survival probabilities are quite high for almost all locations, indication low risk of a joint event. Also, in this period there is only one joint extreme event in 1976, in the same region as identified in the second row, third column of Figure 6.7, indicating that the period afterwards was quite a stable one in terms of extreme events.

On the other hand, the last time period considered, 1998-2020, has almost exclusively 0 estimated survival probabilities for both frost and drought throughout all locations considered in Bavaria. Joint extreme events are also very likely in for example, the north and east regions. The estimated survival probabilities derived in all three cases, both univariate and the joint, indicate a significant increase in risk for frost, drought and their joint occurrence in the last 20 years compared to the other two considered periods.

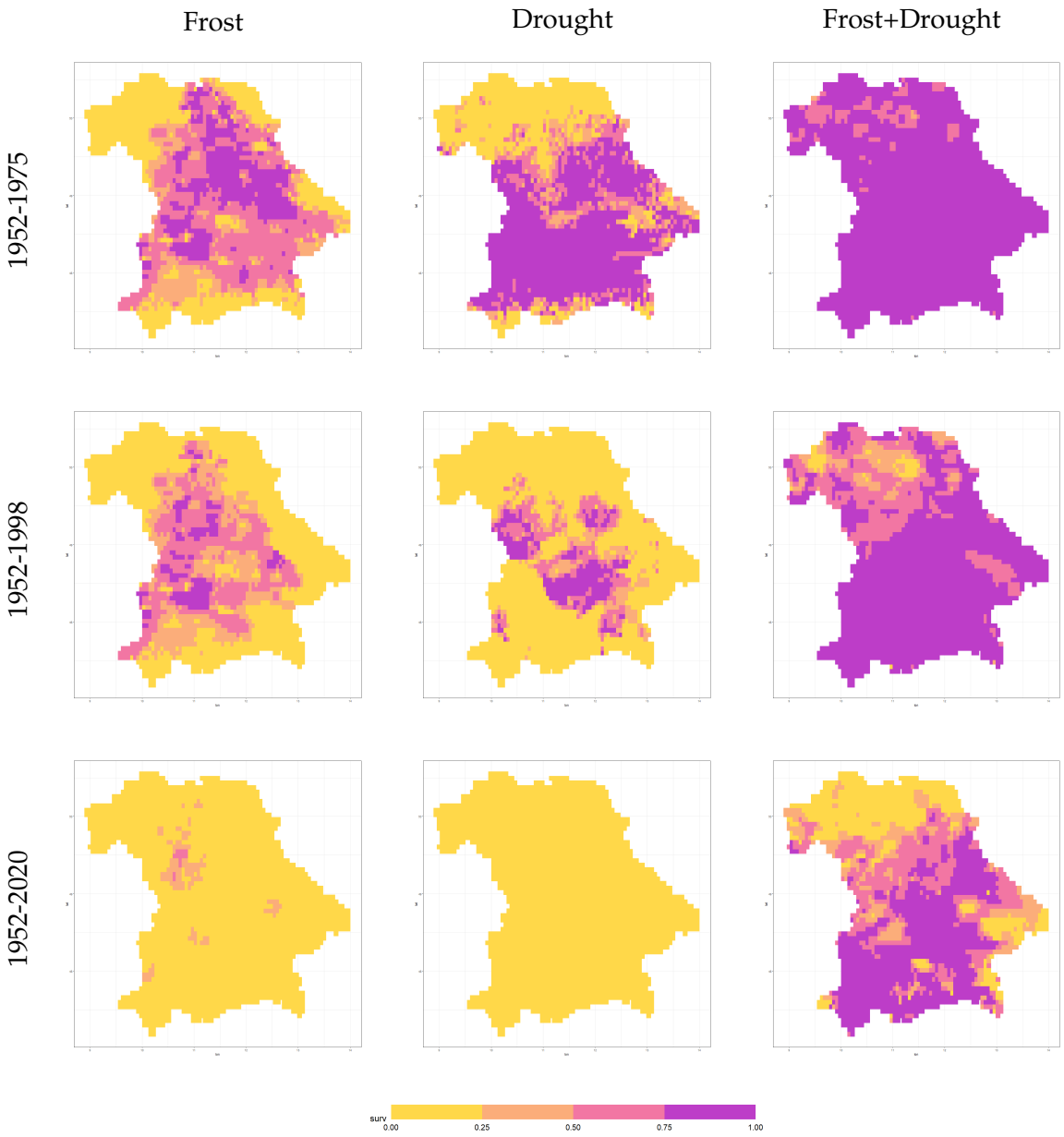


Figure 6.8.: Estimated survival probabilites $\hat{S}_{\hat{\mathcal{D}}_{frost}}(s,T)$, $\hat{S}_{\hat{\mathcal{D}}_{drought}}(s,T)$, and $\hat{S}_{\hat{\mathcal{Y}}}(s,T)$ for periods $(s,T) = \{(1952,1974),(1952,1998),(1952,2020)\}$ (from top to bottom row).

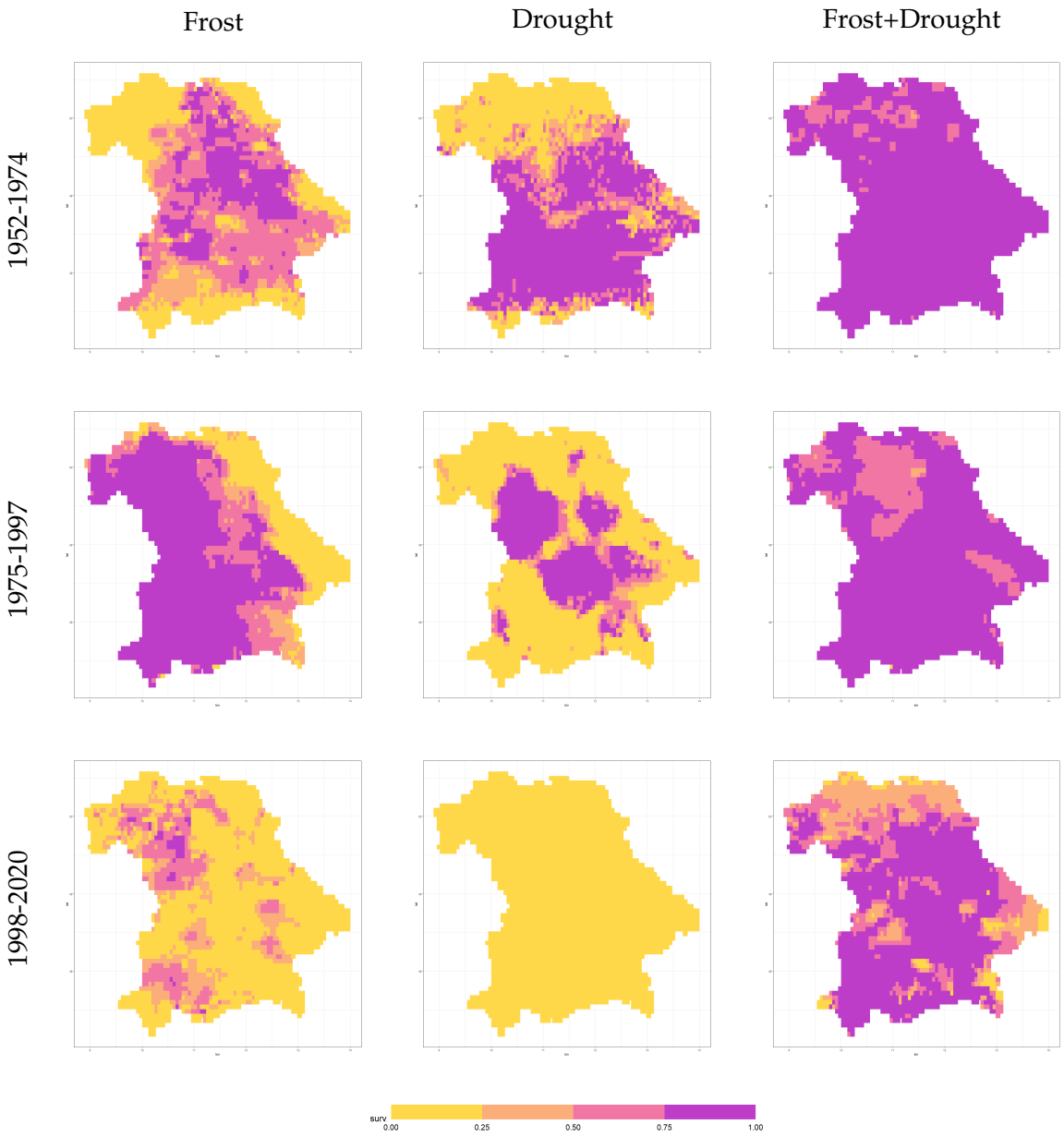


Figure 6.9.: Estimated survival probabilities $\hat{S}_{\hat{\mathcal{D}}_{frost}}(s,T)$, $\hat{S}_{\hat{\mathcal{D}}_{drought}}(s,T)$, and $\hat{S}_{\hat{\mathcal{Y}}}(s,T)$ for periods $(s,T) = \{(1952,1974),(1975,1997),(1998,2020)\}$ (from top to bottom row).

6.6. Return periods

Return period of an event is the expected time until the event reoccurs. Depending on the usage goal and data at hand, it can be defined as the expected time interval at which an event of a given magnitude is exceeded for the first time or the average of the time intervals between two exceedances of a given threshold (Volpi et al., 2015). Motivated by the goal of our study, we use the first definition based on the waiting time until an event happens. We define the **event happened if the estimated survival probability hits a threshold of 0.5**, which indicates the first time there will be a greater chance of the event to happen than not to, i.e. the return period is the number of years at which the probability of surviving is equal to or greater than 0.5, or 50%. The return period for the frost index is then defined as

$$R_{\hat{\mathcal{D}}_{frost}} := \inf_{t \in [1952,2020]} \left\{ t | \hat{S}_{\hat{\mathcal{D}}_{frost}}(1952, t) \leq 0.5 \right\}.$$

Similarly, the return periods for the univariate D-vine model for the drought index are defined $R_{\hat{\mathcal{D}}_{drought}}$. In the bivariate case, we define the return period as

$$R_{\hat{\mathcal{Y}}} := \inf_{t \in [1952,2020]} \left\{ t | \hat{S}_{\hat{\mathcal{Y}}}(1952, t) \leq 0.5 \right\}.$$

We iterate over increasing values of t, evaluating $\hat{S}_{\hat{\mathcal{Y}}}(1952, t)$ and the first time $\hat{S}_{\hat{\mathcal{Y}}}(1952, t) \leq 0.5$ happens, we obtain the waiting time until the event occurs. Since this approach assumes that the survival function is continuous and strictly decreasing, we interpolated between the discrete values of our estimated survival function $\hat{S}_{\hat{\mathcal{Y}}}(1952, t)$. In Figure 6.10 we plot the estimated survival functions for each of the 3 estimated conditional probabilites for a randomly chosen location. The dotted line represent the threshold 0.5, and its intersection with the survival functions indicates the value of the return period. For example, the return period for the drought event is 8 years, the return period for the frost risk cannot be precisely determined, as the threshold has not been reached for this location and the return period for the joint extreme event is 60 years.

Figure 6.11 shows the return periods for all considered locations for the 3 different models considered, $R_{\hat{\mathcal{D}}_{frost}}$, $R_{\hat{\mathcal{D}}_{drought}}$, $R_{\hat{\mathcal{Y}}}$. Regions are coloured based on the value of the estimated return period. We distinguish between regions with a return periods of 0-20 years (orange), 21-40 years (light green), 41-60 years (dark green) and more than 60 years (blue). Gray regions show regions for which the threshold of 0.50 has not been reached in the 69 considered years. Based on this plots we can distinguish between temporal "at-risk regions" based on the estimated return periods. The highest risk regions have the lowest return times and the lowest risk regions have the highest return

times or return times that are greater than 69 years. For the return periods associated with the univariate regression models, the highest risk is shown approximately in the northern, central-east border regions and southern regions of Bavaria. For the return period of the joint vine regression model we see lower risks and the estimated highest risks are in the northern and central-east border regions.

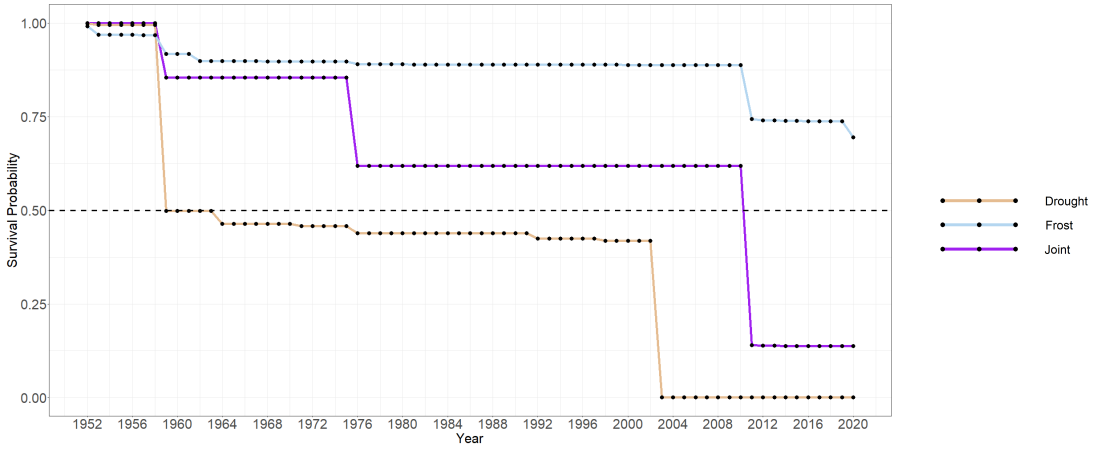


Figure 6.10.: Plotted are $\hat{S}_{\hat{\mathcal{D}}_{frost}}(1952, 2020)$, $\hat{S}_{\hat{\mathcal{D}}_{drought}}(1952, 2020)$ and $\hat{S}_{\hat{\mathcal{Y}}}(1952, 2020)$ for randomly chosen location with latitude coordinate 49.4874 and longitude coordinate 10.809.

6.7. Conclusion and outlook

We fitted the models developed for vine based regression to a large real data set, in one of the most important topics in today's society, changing climate. The utilized data set shows majorly non-Gaussian dependencies, especially between the possible predictor variables, thus making vine copulas suitable for modeling the data at hand. We utilize the D-vine copula model for modeling the drought and frost indices separately, and the Y-vine copula for their joint modeling. Based on annual fitted models we propose conditional risk measures, which quantify the univariate and bivariate risks of extreme annual events. This way, we identify years which are extreme for both the univariate and bivariate risks. We also suggest a survival probability analysis and return times analysis, based on the models fitted, so that we can identify "at-risk" spatial and temporal regions. Further, up to our knowledge this is one of the biggest scale data modeling application based on vine copula models.

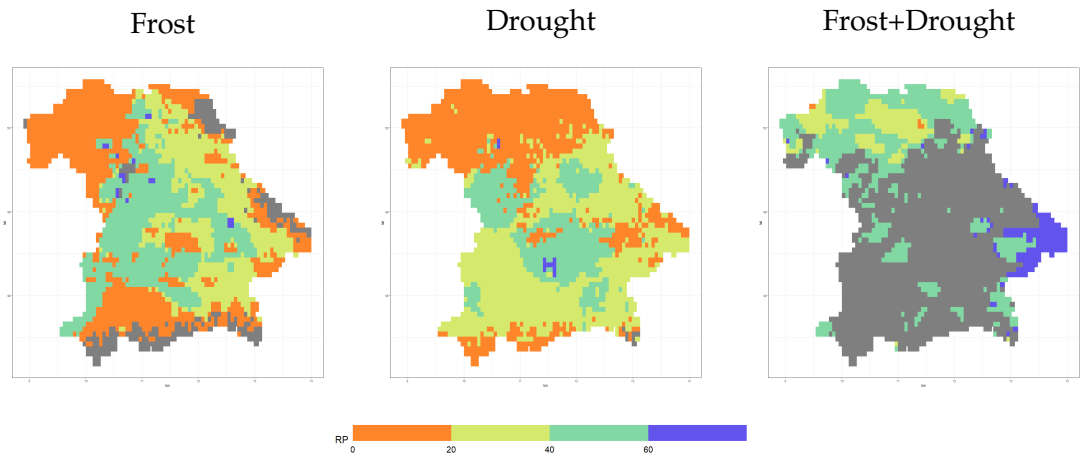


Figure 6.11.: Return periods for $R_{\hat{\mathcal{D}}_{frost}}$, $R_{\hat{\mathcal{D}}_{drought}}$, $R_{\hat{\mathcal{Y}}}$. Regions are coloured based on the estimated return period. Distinguished are return periods of 0-20 years (orange), 21-40 (light green), 41-60 (dark green) and > 60 (blue). Gray regions denote regions where the threshold has not been reached.

After establishing the statistical tools required to properly model the frost and drought risks, conditioned on a set of predictors, we are currently extending the data set by including projected data for the next 80 years, from 2020-2100 under different climate scenarios. Our main goal is to answer the questions: how will the joint probability of drought and frost events and the frequency of co-occurring frost and drought, shift under various climate change scenarios. This is extremely important for future forest management recommendations on the region of Bavaria.

Further, a possible outlook is widening the spatial scope from Bavaria to Europe and integrating different tree species. Currently, such analysis is limited by the (non-) availability of Europe wide high-resolution climate projections, which adequately capture the extremes in the primary climate variables associated with drought and especially spring late-frost. By building on recent advances in climate downscaling, which use deep convolutional neural networks with batch normalization and residual networks, it will soon become possible to produce high-resolution climate data sets for the European domain which will accurately represent such extremes. Further, this can allow development of associated forest management recommendations on the European scale.

7. Overall conclusion and future outlook

The main topic of this thesis was to deepen the understanding and to extend the applicability of vine copula based models in regression settings. We extend the univariate response vine copula based regression to a fully nonparametric setup, with a less greedy forward selection approach than the standard approach suggest in Kraus and Czado (2017). Then, we develop a novel bivariate response vine copula based regression, utilizing a newly developed Y-vine tree sequence. We suggest many novel methods and extensions around the Y-vine copula, concerning simulation, prediction, numerical estimation of bivariate conditional level curves and simulation based bivariate conditional quantile curves. Based on the Y-vine, we also suggest and estimate bivariate conditional risk measures. Also, all the methods developed are coded in the `R` statistical software, and can be further used for other applications or utilized for other extension methods.

One possible promising future research area is the connection of the developed methods proposed in the thesis with Bayesian networks (BNs) (more details on BNs in Lauritzen (1996) and Cowell et al. (2007)). The first connection between copulas and Bayesian networks was made in Elidan (2010) and Elidan (2012). It turns out that if the density of a multivariate distribution decomposes into conditional densities, then the underlying copula density decomposes into corresponding conditional copula densities. Copula Bayesian networks (CBNs) exploit the graph, to encode conditional independences in a parsimonious way and copulas, to model marginal distributions and conditional copula densities separately. However, only a limited number of families of multivariate copulas with one parameter are considered (Liebscher, 2006). An improvement is seen when using CBNs over Gaussian BNs in Elidan (2010), while further results on copula based Bayesian networks can be found in Liu et al. (2009), Kirshner (2008), and Darsow et al. (1992).

First connections between vine copulas and BNs were made in Kurowicka and Cooke (2002), (and later in Kurowicka and Cooke (2005) and Kurowicka and Cooke (2006)) where it is shown that every continuous multivariate distribution associated with a directed acyclic graph (DAG) can be decomposed into a family of bivariate (conditional) distributions, which correspond to the edges of the underlying graph. Further, Bauer and Czado (2016) introduce Pair copula Bayesian Networks (PCBNs) together with al-

gorithms for evaluations of the joint density arising from a PCBN, maximum likelihood estimation, simulation from PCBNs and a model selection algorithm inspired by the widely used PC algorithm by Spirtes and Glymour (1991). The flexibility of PCBNs allows for capturing of a wide range of distributional features to be modelled, such as heavy tails, tail dependencies, and non-linear, asymmetric dependencies. However, the methodology given in Bauer and Czado (2016) in limited in terms of the number of variables, the computational cost and the estimation procedure needed. Thus, it would be beneficial to expand the idea of Bauer and Czado (2016) to a full computationally inexpensive procedure, without losing flexibility.

There are two separate problems in BN estimation procedure: structure learning and conditional density estimation. For the structure learning procedure of a DAG one can either use expert knowledge or data driven approaches, usually either constraint-based or score-and-search-based approaches (Koller and Friedman, 2009). The constraint-based approach is based on a series of conditional independence tests, while the score-and-search-based approaches optimize a given scoring function.

Bauer and Czado (2016) suggest a test for conditional independence between two variables by fitting a regular vine copula model on the data, with a constraint that the variables in the conditioned set are always leaf nodes in the vine copula. This way the marginal conditional distributions required for testing conditional independence are convenient to estimate, with no integration required. However, Bauer and Czado (2016) fit the best vine model with the given leaf node constraint, but they don't consider nor optimize the joint conditional distribution of the two variables given the rest, needed for the conditional independence test. They fit the best vine model on the set of variables, with no restrictions towards optimizing the joint conditional distribution. Also, there is an asymmetric treatment of the two variables being tested. To overcome all of this, we propose to do the conditional independence test utilizing the Y-vine copula model. It is designed in a way that it optimizes the joint conditional log-likelihood of two variables given the rest. Also, it provide symmetric treatment of the two variables, have an automatic forwards selection of variables in the conditioning set and are numerically tractable even in higher dimensions. Thus, due to the convenience by design, the Y-vine copula models are interesting to be explored for structure learning and conditional independence testing in non-Gaussian DAGs. Following Bauer and Czado (2016), we suggest the widely used constraint-based PC algorithm (Spirtes and Glymour, 1991) for the case of learning non-Gaussian directed acyclic graphs (DAGs), with a novel conditional independence test based on Y-vine copulas (introduced in Chapter 4).

To model the conditional densities specified in a BN, Bauer and Czado (2016) used a broader class of vine copula models for which computationally expensive integration might be required for some node. However, we propose to model the conditional densities using less expensive models, that involves automatic forward selection of

parent nodes and no integration is needed, as already suggested and used in Czado and Scharl (2021). We propose to model the conditional densities using a D-vine regression model (Kraus and Czado, 2017) or a C-vine regression model (Chapter 3), where the node whose conditional density we model, will be the response variable and the parents are the predictor variables in these regression settings. This way we will model the conditional density for each node using a flexible class of D- and C-vine copulas, with an included automatic forward selection of parents and non-expensive estimation. Even more, these approaches will be able to identify edges that are not supported by the data, thus allowing for potential reduction of edges in the resulting Bayesian network.

A. Appendix to Chapter 4

A.1. Proofs

A.1.1. Proof of Proposition 1

Proof 1

$$
\begin{aligned}
F_{Y_1,Y_2|\mathbf{X}}\left(y_1,y_2|\mathbf{x}\right) &= \int_{-\infty}^{y_1}\int_{-\infty}^{y_2} f_{Y_1,Y_2|\mathbf{X}}\left(y_1',y_2'|\mathbf{x}\right)dy_2'\,dy_1' \\
&= \int_{-\infty}^{y_1}\int_{-\infty}^{y_2} \frac{f_{Y_1,Y_2,\mathbf{X}}\left(y_1',y_2',\mathbf{x}\right)}{f_{\mathbf{X}}\left(\mathbf{x}\right)}dy_2'\,dy_1' \\
&= \frac{1}{f_{\mathbf{X}}\left(\mathbf{x}\right)}\int_{-\infty}^{y_1}\int_{-\infty}^{y_2} \frac{\partial^{p+2}}{\partial y_1\,\partial y_2\,\partial x_1\ldots\partial x_p}F_{Y_1,Y_2,\mathbf{X}}\left(y_1,y_2,\mathbf{x}\right)\Big|_{y_1=y_1',y_2=y_2'}dy_2'\,dy_1' \\
&= \frac{1}{f_{\mathbf{X}}\left(\mathbf{x}\right)}\cdot\frac{\partial^p}{\partial x_1\ldots\partial x_p}F_{Y_1,Y_2,\mathbf{X}}\left(y_1,y_2,\mathbf{x}\right) \\
(\textit{by Sklar's theorem}) &= \frac{1}{f_{\mathbf{X}}\left(\mathbf{x}\right)}\cdot\frac{\partial^p}{\partial x_1\ldots\partial x_p}C_{V_1,V_2,\mathbf{U}}\left(F_{Y_1}(y_1),F_{Y_2}(y_2),F_{X_1}(x_1),\ldots,F_{X_p}(x_p)\right) \\
&= \frac{1}{f_{\mathbf{X}}\left(\mathbf{x}\right)}\cdot\frac{\partial^p}{\partial u_1\ldots\partial u_p}C_{V_1,V_2,\mathbf{U}}\left(v_1,v_2,u_1,\ldots,u_p\right)\Big|_{v_j=F_{Y_j}(y_j),u_i=F_{X_i}(x_i)}\frac{\partial u_1\ldots\partial u_p}{\partial x_1\ldots\partial x_p} \\
\left(\frac{\partial u_1\ldots\partial u_p}{\partial x_1\ldots\partial x_p}=\prod_{i=1}^{p}f_{X_i}\left(x_i\right)\right) &= \frac{\partial^p}{\partial u_1\ldots\partial u_p}C_{V_1,V_2,\mathbf{U}}\left(v_1,v_2,u_1,\ldots,u_p\right)\Big|_{v_j=F_{Y_j}(y_j),u_i=F_{X_i}(x_i)}\cdot\frac{\prod_{i=1}^{p}f_{X_i}\left(x_i\right)}{f_{\mathbf{X}}\left(\mathbf{x}\right)} \\
\left(\frac{\partial u_1\ldots\partial u_p}{\partial x_1\ldots\partial x_p}=\prod_{i=1}^{p}f_{X_i}\left(x_i\right)\right) &= \frac{\partial^p}{\partial u_1\ldots\partial u_p}C_{V_1,V_2,\mathbf{U}}\left(v_1,v_2,u_1,\ldots,u_p\right)\Big|_{v_j=F_{Y_j}(y_j),u_i=F_{X_i}(x_i)}\cdot\frac{1}{c_{\mathbf{U}}\left(\mathbf{u}\right)} \\
&= C_{V_1,V_2|\mathbf{U}}\left(F_{Y_1}(y_1),F_{Y_2}(y_2)|F_{X_1}(x_1),\ldots,F_{X_p}(x_p)\right),
\end{aligned}
$$

where $C_{V_1,V_2|\mathbf{U}}\left(F_{Y_1}(y_1),F_{Y_2}(y_2)|F_{X_1}(x_1),\ldots,F_{X_p}(x_p)\right)$ *or shortly* $C_{V_1,V_2|\mathbf{U}}\left(v_1,v_2|\mathbf{u}\right)$ *is the conditional distribution of* V_1,V_2 *given* $\mathbf{U}=\mathbf{u}$ *and the joint copula distribution of* $Y_1,Y_2,\mathbf{X}$ *is denoted by* $C_{V_1,V_2,\mathbf{U}}$.

A.1.2. Proof of Proposition 2

Proof 2 *We prove that a Y-vine tree sequence,* $\left\{T_1,\ldots,T_{p+1}\right\}$*, satisfies conditions* (i)*-*(iii) *from Section 2. The first condition* (i) *is trivial and follows by definition of* T_1*. The next condition requires that* $N_k=E_{k-1}\ \forall\ k\geq 2$*. For* $k=2$*,* $N_2=\left\{V_1U_1,V_2U_1,U_1U_2,\ldots,U_{p-1}U_p\right\}=E_1$

follows directly from Definition 4.1. To prove the statement for $k > 2$, we start with the edge set of tree T_{k-1}, E_{k-1} given as

$$E_{k-1} = \bigcup_{j=1,2} \left\{ \left(V_j U_{k-2}; \mathbf{U}_{1:k-3},\ U_1 U_{k-1}; \mathbf{U}_{2:k-2} \right) \right\} \bigcup_{i=1}^{p-k+1} \left\{ \left(U_i U_{i+k-2}; \mathbf{U}_{i+1:i+k-3},\ U_{i+1} U_{i+k-1}; \mathbf{U}_{i+2:i+k-2} \right) \right\}.$$

Edge $\left(V_j U_{k-2}; \mathbf{U}_{1:k-3},\ U_1 U_{k-1}; \mathbf{U}_{2:k-2} \right)$ is associated with node $V_j U_{k-1}; \mathbf{U}_{1:k-2}$ in T_k for $j = 1, 2$ and edge $(U_i U_{i+k-2}; \mathbf{U}_{i+1:i+k-3},\ U_{i+1} U_{i+k-1}; \mathbf{U}_{i+2:i+k-2})$ is associated with node $U_i U_{i+k-1}; \mathbf{U}_{i+1:i+k-2}$ for $i = 1, , \ldots, p-k+1$ in T_k. Therefore, by Definition 4.1, $N_k = E_{k-1}$ holds for all k in the Y-vine tree sequence. The last condition, the proximity condition, states that for $k \geq 2$ two nodes can be connected in T_k only if the corresponding edges in the previous tree T_{k-1} share a common node. Consider the part of the tree sequence that only contains the predictors $(X_1, \ldots, X_p)$. By definition of the Y-vine tree sequence, the predictors are arranged in a D-vine tree sequence, which is a known regular vine tree sequence subset, implying that for the nodes containing only the predictors the proximity condition is satisfied. So, we consider the remaining nodes that contain the response variables in the conditioned set and the node that connects them to the D-vine of the predictors. For T_2, nodes $V_1 U_1$, $V_2 U_1$ are both connected to $U_1 U_2$. For $V_j U_1$, $j = 1, 2$ the corresponding edge in T_1 is $(V_j U_1)$ which shares the node U_1 with the corresponding edge of node $U_1 U_2$, edge $(U_1 U_2)$. For $k > 2$ in T_k the nodes $V_1 U_{k-1}; \mathbf{U}_{1:k-2}$ and $V_2 U_{k-1}; \mathbf{U}_{1:k-2}$ are connected to $U_1 U_k; \mathbf{U}_{2:k-1}$. In T_{k-1} the corresponding edge of node $V_j U_{k-1}; \mathbf{U}_{1:k-2}$ for $j = 1, 2$ is the edge $(V_j U_{k-2}; \mathbf{U}_{1:k-3},\ U_1 U_{k-2}; \mathbf{U}_{1:k-3})$ and for node $U_1 U_k; \mathbf{U}_{2:k-1}$ the corresponding edge is $(U_1 U_{k-2}; \mathbf{U}_{2:k-3},\ U_2 U_k; \mathbf{U}_{3:k-1})$. They share a common node $U_1 U_{k-2}; \mathbf{U}_{2:k-3}$ in T_{k-1}, thus the proximity condition is satisfied.

A.1.3. Proof of Theorem 1

Proof 3 *By definition of a conditional density it follows that $f_{Y_1, Y_2|\mathbf{X}} = \frac{f_{Y_1, Y_2, \mathbf{X}}}{f_{\mathbf{X}}}$. The numerator $f_{Y_1, Y_2, \mathbf{X}}$ is expressed in Equation (4.1), and we need to derive the denominator $f_{\mathbf{X}}$ in terms of copulas. Consider the part of the Y-vine tree sequence after removing the PITs of the responses V_1 and V_2, i.e., the tree sequence consisting of only the PITs of the predictors $(U_1, \ldots, U_p)^T$. By definition of the Y-vine tree structure, the predictors are arranged in a D-vine tree sequence with a specific order. Thus, the density of a D-vine with this given order (see more in Czado, 2010) can be expressed as*

$$f_{\mathbf{X}}(\mathbf{x}) = \prod_{k=1}^{p} f_{X_k}(x_k) \cdot \prod_{k=1}^{p-1} \prod_{i=1}^{p-k} c_{U_i, U_{i+k}; \mathbf{U}_{i+1:i+k-1}} \left(F_{X_i|\mathbf{X}_{i+1:i+k-1}}(x_i|\mathbf{x}_{i+1:i+k-1}), F_{X_{i+k}|\mathbf{X}_{i+1:i+k-1}}(\mathbf{x}_{i+k}|x_{i+1:i+k-1}) \right). \tag{A.1}$$

Canceling out all common terms in the expansions of the numerator and the denominator, given in Equation (4.1) *and* (A.1) *respectively, we are left with the expression in Equation* (4.2). *All the required copulas in Equation* (4.2) *are already derived in the Y-vine tree sequence,* $c_{V_j,U_i;\mathbf{U}_{1:i-1}} \in \mathcal{B}(\mathcal{V})$ *for* $j = 1,2,\ i = 1,\ldots,p$ *and* $c_{V_1,V_2;\mathbf{U}} \in \mathcal{B}(\mathcal{V})$ *(these copulas can be seen as the copulas on the furthest left side of each tree in Figure 4.1).*

A.1.4. Proof of Corollary 1

Proof 4 *Let's prove part a.) for* $j = 1$. *Due to symmetry the same proof follows for* $j = 2$. *By definition of a conditional density it follows that* $f_{Y_1|\mathbf{X}} = \frac{f_{Y_1,\mathbf{X}}}{f_{\mathbf{X}}}$. *The denominator is expressed in Equation* (A.1), *while the numerator needs to be expanded. Consider the random vector* $(V_1, \mathbf{U})^T$ *in the tree sequence of the Y-vine, i.e. remove the node of the PIT of the response* V_2 *from the first tree* T_1 *and all the nodes in the further trees that will disappear by removing the variable* V_2. *By definition of the Y-vine, the variables* $(V_1, U_1, \ldots, U_p)$ *are arranged in a D-vine tree sequence with a specific order. Thus, the density of a D-vine with this given order (see more in Czado, 2010) is given as*

$$f_{Y_1,\mathbf{X}}(y_1,y_2,\mathbf{x}) = \prod_{k=1}^{p} f_{X_k}(x_k)\cdot f_{Y_1}(y_1)\cdot \prod_{k=1}^{p-1}\left[\prod_{i=1}^{p-k} c_{U_i,U_{i+k};\mathbf{U}_{i+1:i+k-1}}\left(F_{X_i|\mathbf{X}_{i+1:i+k-1}}(x_i|\mathbf{x}_{i+1:i+k-1}),\right.\right.$$
$$\left.\left. F_{X_{i+k}|\mathbf{X}_{i+1:i+k-1}}(x_{i+k}|\mathbf{x}_{i+1:i+k-1})\right)\right]$$
$$\prod_{i=1}^{p}\left[c_{V_1,U_i;\mathbf{U}_{1:i-1}}\left(F_{Y_1|\mathbf{X}_{1:i-1}}(y_1|\mathbf{x}_{1:i-1}),\ \ F_{\mathbf{X}_i|X_{1:i-1}}(x_i|\mathbf{x}_{1:i-1})\right)\right]. \tag{A.2}$$

Cancelling common terms of the numerator, Equation (A.2), *and the denominator, Equation* (A.1), *we are left with Equation* (4.3) *for* $j = 1$.

Now let's prove part b.) for $(j,k) = (1,2)$. *Due to symmetry the same proof follows for* $(j,k) = (2,1)$. *Use that* $f_{Y_2|\mathbf{X},Y_1} = \frac{f_{Y_1,Y_2,\mathbf{X}}}{f_{Y_1,\mathbf{X}}}$ *holds. The numerator is expressed in Equation* (4.1), *and the denominator is expressed as in the part a.) Equation* (A.2). *Considering the associated ratio and cancelling all common terms, we are left with Equation* (4.4) *for* $(j,k) = (1,2)$. *Again, all the required copulas are already derived in the Y-vine tree sequence,* $c_{V_j,U_i;\mathbf{U}_{1:i-1}} \in \mathcal{B}(\mathcal{V})$ *for* $j = 1,2\ i = 1,\ldots,p$ *and* $c_{V_1,V_2;\mathbf{U}} \in \mathcal{B}(\mathcal{V})$, *which means we don't require any additional calculations.*

A.2. Pseudo-code for the bivariate vine based regression algorithm

Algorithm 4: Bivariate vine based regression algorithm

Input: Data set $\mathbf{y}_n = (y_1^n, y_2^n)^T$, $\mathbf{x}_n = \left(x_1^n, \ldots, x_p^n\right)^T$, for $n = 1, \ldots, N$
Initialization:
$acll_0 = 0$
$NotChosenIndex = \{1, \ldots, p\}$
$ChosenIndex = \emptyset$

1. Estimate marginals F_{Y_j}, F_{X_i}, $j = 1, 2$, $i = 1, \ldots, p$, by a univariate kernel density estimator, implemented in `kde1d`.
2. Obtain pseudo copula data $u_i^n := \hat{F}_{X_i}\left(x_i^n\right)$ for $i = 1, \ldots, p$, $v_1^n := \hat{F}_{Y_1}\left(y_1^n\right)$ and $v_2^n := \hat{F}_{Y_2}\left(y_2^n\right)$.

for $j = 1, \ldots, p$ **do**
  Calculate $acll_1^j$ as

$$acll_1^j = cll_0 + \ell(c_{V_1 U_i}) + \ell(c_{V_2 U_i}) + \ell(c_{V_1 V_2; U_i})$$

end
$r_1 := \arg\max_{j=1,\ldots,p} acll_1^j$
$NotChosenIndex = NotChosenIndex \setminus \{r_1\}$
$ChosenIndex = ChosenIndex \bigcup \{r_1\}$
$acll_1 := acll_1^{r_1}$
for $k = 2, \ldots, p$ **do**
  for $t \in NotChosenIndex$ **do**
    Calculate $acll_k^t$ as

$$acll_k^t = acll_{k-1} + \ell(c_{V_1 U_t; U_{r_1}, \ldots, U_{r_{k-1}}}) + \ell(c_{V_2 U_t; U_{r_1}, \ldots, U_{r_{k-1}}})$$

  end
  $r_k := \arg\max_{t \in NotChosenIndex} acll_k^t$
  $NotChosenIndex = NotChosenIndex \setminus \{r_k\}$
  $ChosenIndex = ChosenIndex \bigcup \{r_k\}$
  $acll_k := acll_k^{r_k}$
end
return $ChosenIndex = \{r_1, \ldots, r_p\}$, i.e. order of the predictors which uniquely determines the fitted bivariate regression model.

A.3. Data application from Section 4.7

A.3.1. Notation

The variables given below are enumerated as follows, the response T_{max} is enumerated with 1, $T_{max} = 1$, the response T_{min} is enumerated with 2, $T_{min} = 2$, then $\texttt{LDAPS\_Tmin\_lapse} = 7$, $\texttt{LDAPS\_Tmax\_lapse} = 6$, $\texttt{LDAPS\_CC1} = 10$, $\texttt{LDAPS\_WS} = 8$, $\texttt{Present\_Tmin} = 4$, $\texttt{LDAPS\_RHmax} = 5$, $\texttt{LDAPS\_CC3} = 11$, $\texttt{LDAPS\_LH} = 9$ and $\texttt{Present\_Tmax} = 3$. Using that enumeration, in Tables A.1 and A.2 we show the parametric pair copulas that were fitted by our Y-vine regression model. In each tree we give the pair copulas conditioned and conditioning sets, the estimated family, the rotation in degrees, the parameters, the degree of freedom (number of parameters) and the Kendall's $\hat{\tau}$ values.

A.3.2. Fitted pair copulas

Table A.1.: For the fitted T_1 to T_5 given are the conditioned and conditioning sets of the pair copulas, the estimated family, the rotation in degrees, the parameters, the degree of freedom and the Kendall's $\hat{\tau}$ values.

tree	edge	conditioned	conditioning	family	rotation	parameters	df	Kendall's $\hat{\tau}$
1	1	1, 7		gaussian	0	0.66	1	0.46
1	2	2, 7		gaussian	0	0.91	1	0.73
1	3	7, 6		gaussian	0	0.64	1	0.44
1	4	6, 10		gaussian	0	-0.45	1	-0.30
1	5	10, 8		clayton	180	0.27	1	0.12
1	6	8, 4		indep	0		0	0.00
1	7	4, 5		indep	0		0	0.00
1	8	5, 11		clayton	180	0.48	1	0.19
1	9	11, 9		bb8	0	1.67 , 0.97	2	0.24
1	10	9, 3		student t	0	0.04, 7.06	2	0.03
2	1	1, 6	7	bb1	180	0.78 , 1.68	2	0.57
2	2	2, 6	7	indep	0		0	0.00
2	3	7, 10	6	clayton	0	0.64	1	0.24
2	4	6, 8	10	joe	90	1.22	1	-0.11
2	5	10, 4	8	clayton	0	0.17	1	0.08
2	6	8, 5	4	bb8	0	1.87, 0.70	2	0.13
2	7	4, 11	5	student t	0	0.01, 5.21	2	0.01
2	8	5, 9	11	gaussian	0	0.34	1	0.22
2	9	11, 3	9	gumbel	270	1.12	1	-0.10
3	1	1, 10	6, 7	joe	90	1.34	1	-0.16
3	2	2, 10	6, 7	indep	0		0	0.00
3	3	7, 8	10, 6	indep	0		0	0.00
3	4	6, 4	8, 10	gaussian	0	0.53	1	0.35
3	5	10, 5	4, 8	bb8	0	3.28, 0.85	2	0.42
3	6	8, 11	5, 4	joe	0	1.17	1	0.09
3	7	4, 9	11, 5	student t	0	0.15, 6.59	2	0.10
3	8	5, 3	9, 11	gumbel	270	1.37	1	-0.27
4	1	1, 8	10, 6, 7	frank	0	-1.77	1	-0.19
4	2	2, 8	10, 6, 7	student t	0	0.10, 7.34	2	0.06
4	3	7, 4	8, 10, 6	gaussian	0	0.65	1	0.45
4	4	6, 5	4, 8, 10	frank	0	-1.37	1	-0.15
4	5	10, 11	5, 4, 8	frank	0	2.21	1	0.23
4	6	8, 9	11, 5, 4	gumbel	0	1.17	1	0.14
4	7	4, 3	9, 11, 5	gaussian	0	0.68	1	0.47
5	1	1, 4	8, 10, 6, 7	student t	0	0.07, 11.79	2	0.05
5	2	2, 4	8, 10, 6, 7	bb8	0	1.44, 0.94	2	0.15
5	3	7, 5	4, 8, 10, 6	indep	0		0	0.00
5	4	6, 11	5, 4, 8, 10	bb8	90	3.00, 0.82	2	-0.36
5	5	10, 9	11, 5, 4, 8	frank	0	1.28	1	0.14
5	6	8, 3	9, 11, 5, 4	indep	0		0	0.00

Table A.2.: For the fitted T_6 to T_{10} given are the conditioned and conditioning sets of the pair copulas, the estimated family, the rotation in degrees, the parameters, the degree of freedom and the Kendall's $\hat{\tau}$ values.

tree	edge	conditioned	conditioning	family	rotation	parameters	df	Kendall's $\hat{\tau}$
6	1	1, 5	4, 8, 10, 6, 7	frank	0	-0.91	1	-0.10
6	2	2, 5	4, 8, 10, 6, 7	frank	0	0.82	1	0.09
6	3	7, 11	5, 4, 8, 10, 6	clayton	0	0.38	1	0.16
6	4	6, 9	11, 5, 4, 8, 10	joe	0	1.09	1	0.05
6	5	10, 3	9, 11, 5, 4, 8	gumbel	90	1.19	1	-0.16
7	1	1, 11	5, 4, 8, 10, 6, 7	clayton	270	0.18	1	-0.08
7	2	2, 11	5, 4, 8, 10, 6, 7	indep	0		0	0.00
7	3	7, 9	11, 5, 4, 8, 10, 6	clayton	90	0.17	1	-0.08
7	4	6, 3	9, 11, 5, 4, 8, 10	bb7	180	1.18, 0.20	2	0.17
8	1	1, 9	11, 5, 4, 8, 10, 6, 7	indep	0		0	0.00
8	2	2, 9	11, 5, 4, 8, 10, 6, 7	gaussian	0	-0.15	1	-0.10
8	3	7, 3	9, 11, 5, 4, 8, 10, 6	bb1	0	0.24, 1.13	2	0.21
9	1	1, 3	9, 11, 5, 4, 8, 10, 6, 7	joe	0	1.08	1	0.04
9	2	2, 3	9, 11, 5, 4, 8, 10, 6, 7	gaussian	0	0.16	1	0.10
10	1	1, 2	3, 9, 11, 5, 4, 8, 10, 6, 7	joe	180	1.18	1	0.09

B. Appendix to Chapter 5

B.1. Theoretical and estimated unconditional level curves

B.1.1. Clayton copula

The distribution function of a bivariate Clayton copula with parameter θ is given in Equation (2.5). By solving $\left(v_1^\theta + v_2^\theta - 1\right)^{-1/\theta} = \alpha$ for v_2, we obtain the α bivariate level curve as

$$Q_\alpha^V := \left\{ \left(v_1, \left(\alpha^{-\theta} - v_1^{-\theta} + 1 \right)^{-1/\theta} \right) \;\middle|\; \forall\, v_1 \in [0,1] \right\}.$$

B.1.2. Gumbel copula

The distribution function of a bivariate Gumbel copula with parameter θ is given in Equation (2.6). By solving $\exp\left\{ -\left[(-\ln v_1)^\theta + (-\ln v_2)^\theta \right]^{1/\theta} \right\} = \alpha$ for v_2, we obtain the α bivariate level curve as

$$Q_\alpha^V := \left\{ \left(v_1, \exp\left\{ -\left[(-\ln \alpha)^\theta - (-\ln v_1)^\theta \right]^{1/\theta} \right\} \right) \;\middle|\; \forall\, v_1 \in [0,1] \right\}.$$

B.1.3. Gaussian copula

In contrast to the Archimedean copulas as Clayton and Gumbel, for which there is a closed form solution of the distribution function for one variable, for the elliptical copulas, such as Gaussian and Student-t copula, there is no closed form solution. Thus, we use a numerical procedure to derive the theoretical level curves.
The distribution function of the Gaussian pair copula (introduced in Equation (2.3)) with a correlation parameter θ is

$$\begin{aligned} C_{V_1,V_2}(v_1, v_2; \theta) &= \Phi_2\left(\Phi_1^{-1}(v_1), \Phi_1^{-1}(v_2) \right) \\ &= \int_{-\infty}^{\Phi_1^{-1}(v_1)} \int_{-\infty}^{\Phi_1^{-1}(v_2)} \frac{1}{2\pi\sqrt{1-\theta^2}} \exp\left(-\frac{a^2 - 2\theta ab + b^2}{2\left(1-\theta^2\right)} \right) da\, db, \end{aligned}$$

where Φ_1 and Φ_2 are the univariate and bivariate standard normal distribution functions, respectively. As already stated, the equation

$$\int_{-\infty}^{\Phi_1^{-1}(v_1)} \int_{-\infty}^{\Phi_1^{-1}(v_2)} \frac{1}{2\pi\sqrt{1-\theta^2}} \exp\left(-\frac{a^2-2\theta ab+b^2}{2\left(1-\theta^2\right)}\right) da\, db = \alpha$$

does not have a closed form solution for one of the variables. Thus, we evaluate C_{V_1,V_2} using the integral of its h-function given as

$$h_{V_1|V_2}\left(\left(v_1, v_2; \theta\right)\right) = \Phi_1\left(\frac{\Phi_1^{-1}\left(v_1\right) - \theta\Phi_1^{-1}\left(v_2\right)}{\sqrt{1-\theta^2}}\right).$$

The distribution function is then evaluated at the point $(\tilde{v_1}, \tilde{v_2})$ as

$$C_{V_1,V_2}\left(\tilde{v_1}, \tilde{v_2}; \theta\right) = \int_0^{\tilde{v_2}} h_{V_1|V_2}\left(\left(\tilde{v_1}, v_2; \theta\right)\right) dv_2. \tag{B.1}$$

Finally, the theoretical bivariate level curve is derived using the numerical evaluation defined in Section 5.3 and Equation (B.1).

B.1.4. Student-t copula

Similarly as with the Gaussian copula, the Equation $C_{V_1,V_2}\left(v_1, v_2; \boldsymbol{\theta} = (\theta, df)\right) = \alpha$ does not have a closed form solution for the bivariate Student-t copula (introduced in Equation (2.4)), where df is the degree of freedom, and θ is the correlation parameter associated with the Student-t copula. Again, we evaluate C_{V_1,V_2} using the integral of its h-function given as

$$h_{V_1|V_2}\left(\left(v_1, v_2; \boldsymbol{\theta} = (\theta, df)\right)\right) = t_{1,df+1}\left(\frac{t_{1,df}^{-1}\left(v_1\right) - \theta t_{1,df}^{-1}\left(v_2\right)}{\sqrt{\frac{\left(df + t_{1,df}^{-1}\left(v_2\right)^2\right)\left(1-\theta^2\right)}{df+1}}}\right),$$

where $t_{1,df}$ is the univariate distribution function of the Student-t distribution with df degrees of freedom. The distribution function is then evaluated at the point $(\tilde{v_1}, \tilde{v_2})$ as

$$C_{V_1,V_2}\left(\tilde{v_1}, \tilde{v_2}; \theta\right) = \int_0^{\tilde{v_2}} h_{V_1|V_2}\left(\left(\tilde{v_1}, v_2; \theta\right)\right) dv_2. \tag{B.2}$$

The theoretical bivariate level curve is derived using the numerical evaluation defined in Section 5.3 and Equation (B.2).

B.1.5. Estimated quantile curves

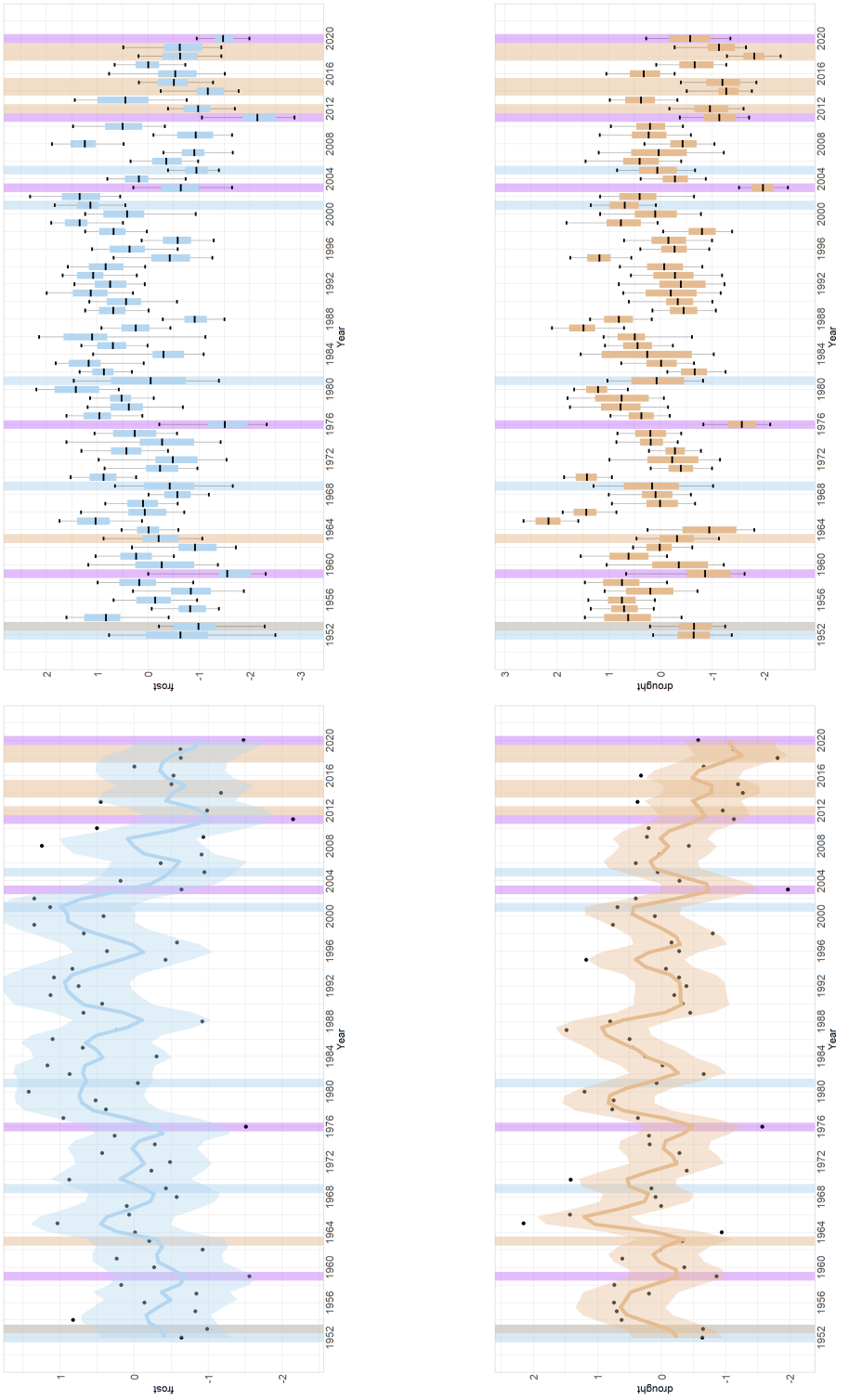
Let $\left\{\left(v_1^i, v_2^i\right)\right\}_{i=1}^n$ be a set of n points randomly drawn from a bivariate copula distribution. Given an estimated parameter $\hat{\theta}$ (together with family) obtained from this set of points we propose to evaluate $\hat{C}_{V_1,V_2}$ at a point $(\tilde{v}_1, \tilde{v}_2)$ as

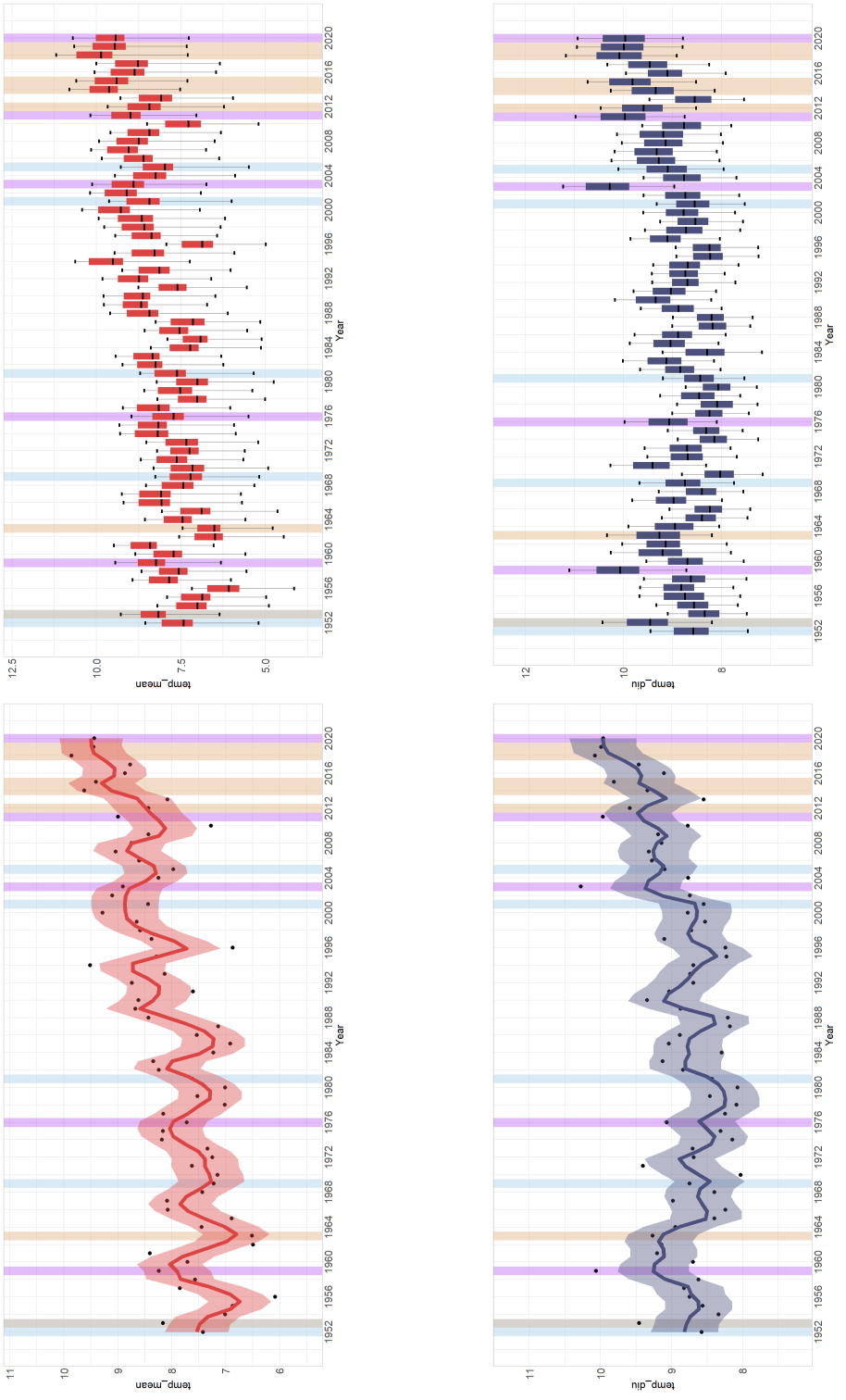
$$\hat{C}_{V_1,V_2}\left(\tilde{v}_1, \tilde{v}_2\right) = \int_0^{\tilde{v}_1} \hat{C}_{V_2|V_1}\left(\tilde{v}_2|v_1'\right) dv_1'. \tag{B.3}$$

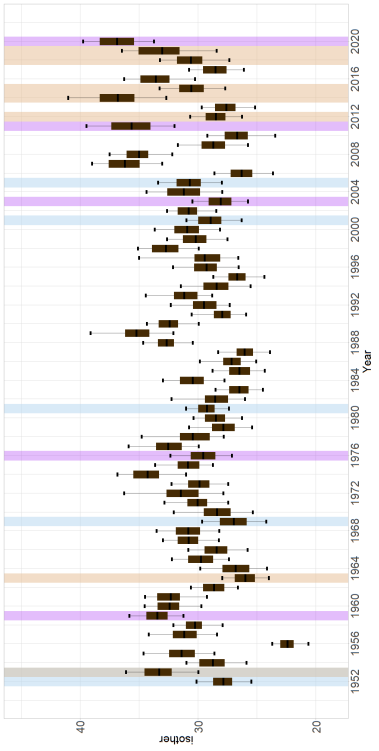
The difference between the estimated and the theoretical level curves for copulas for which the numerical inverse procedure is used is that in the theoretical case we use the theoretical h-function of a copula, while in the estimated case we use the estimated one. Basically, from the simulated data, we estimate a pair-copula, which has an h-function, and that estimated h-function is being used. The estimated bivariate level curves are obtained using the numerical evaluation defined in Section 5.3 and Equation (B.3).

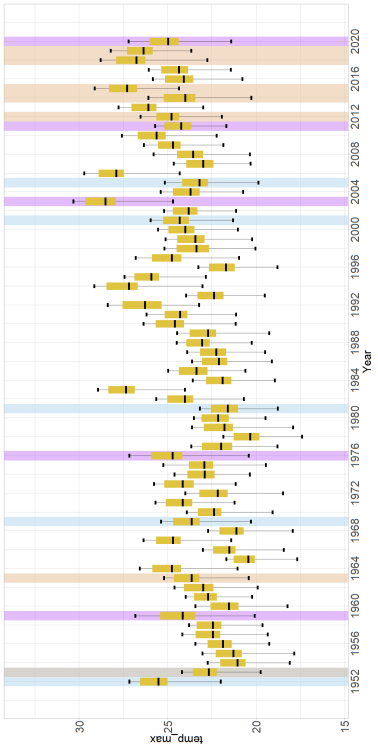
C. Appendix to Chapter 6

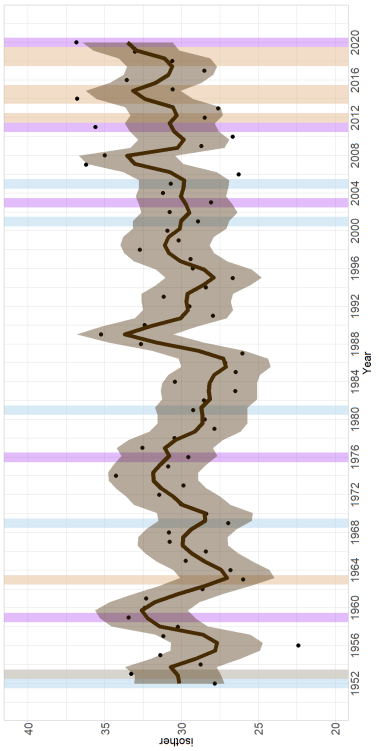
C.1. Exploratory data analysis

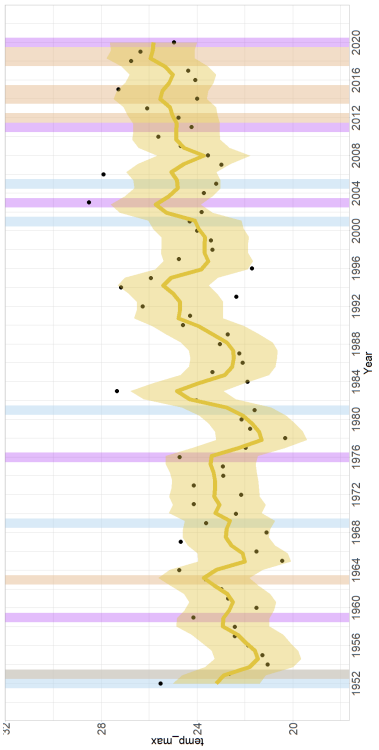

frost
drought
Year
1952
1956
1960
1964
1968
1972
1976
1980
1984
1988
1992
1996
2000
2004
2008
2012
2016
2020

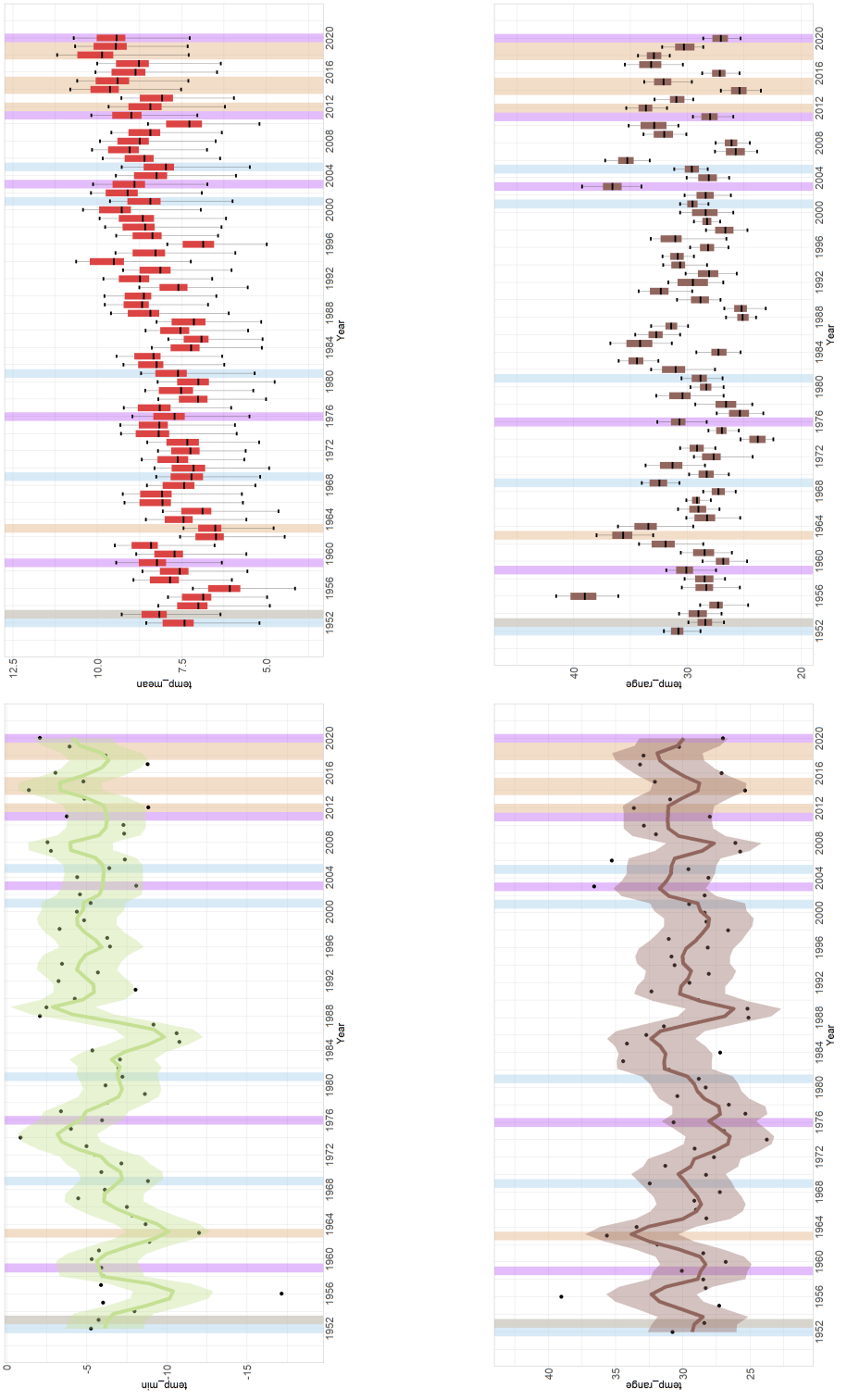

temp_mean
temp_div
Year
1952
1956
1960
1964
1968
1972
1976
1980
1984
1988
1992
1996
2000
2004
2008
2012
2016
2020

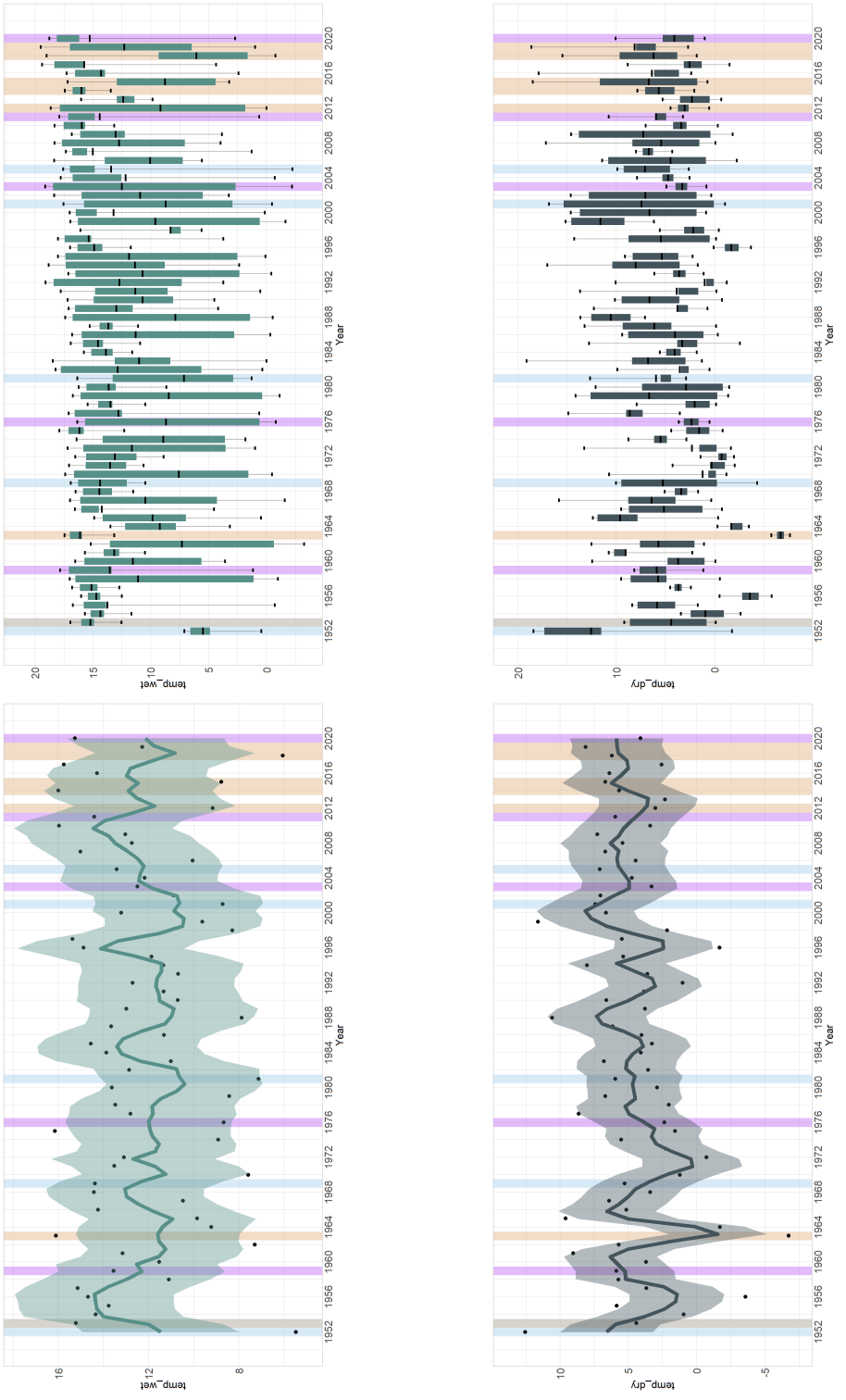

Year
1952
1956
1960
1964
1968
1972
1976
1980
1984
1988
1992
1996
2000
2004
2008
2012
2016
2020
isother
40
30
20

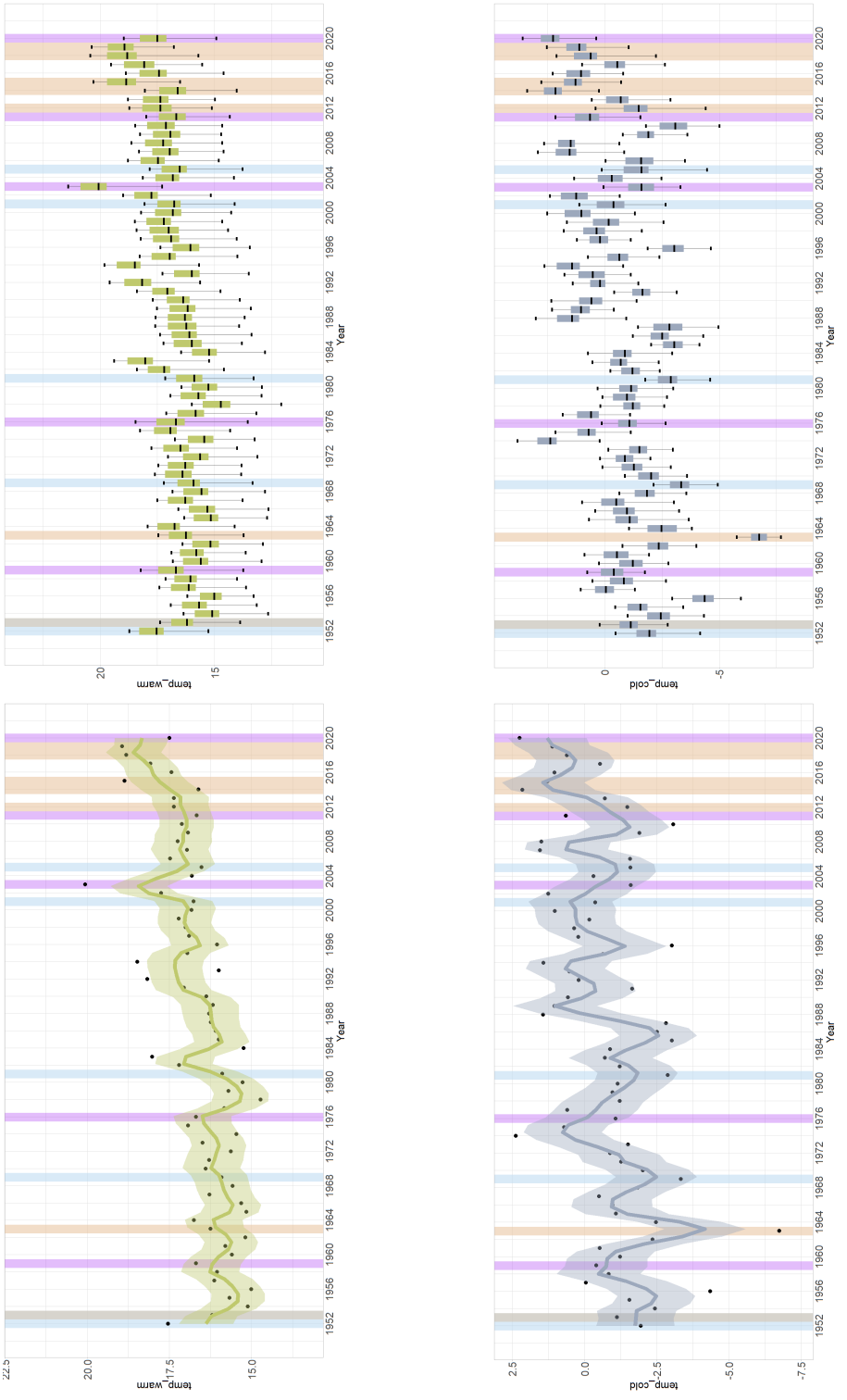

Year
1952
1956
1960
1964
1968
1972
1976
1980
1984
1988
1992
1996
2000
2004
2008
2012
2016
2020
temp_max
30
25
20
15

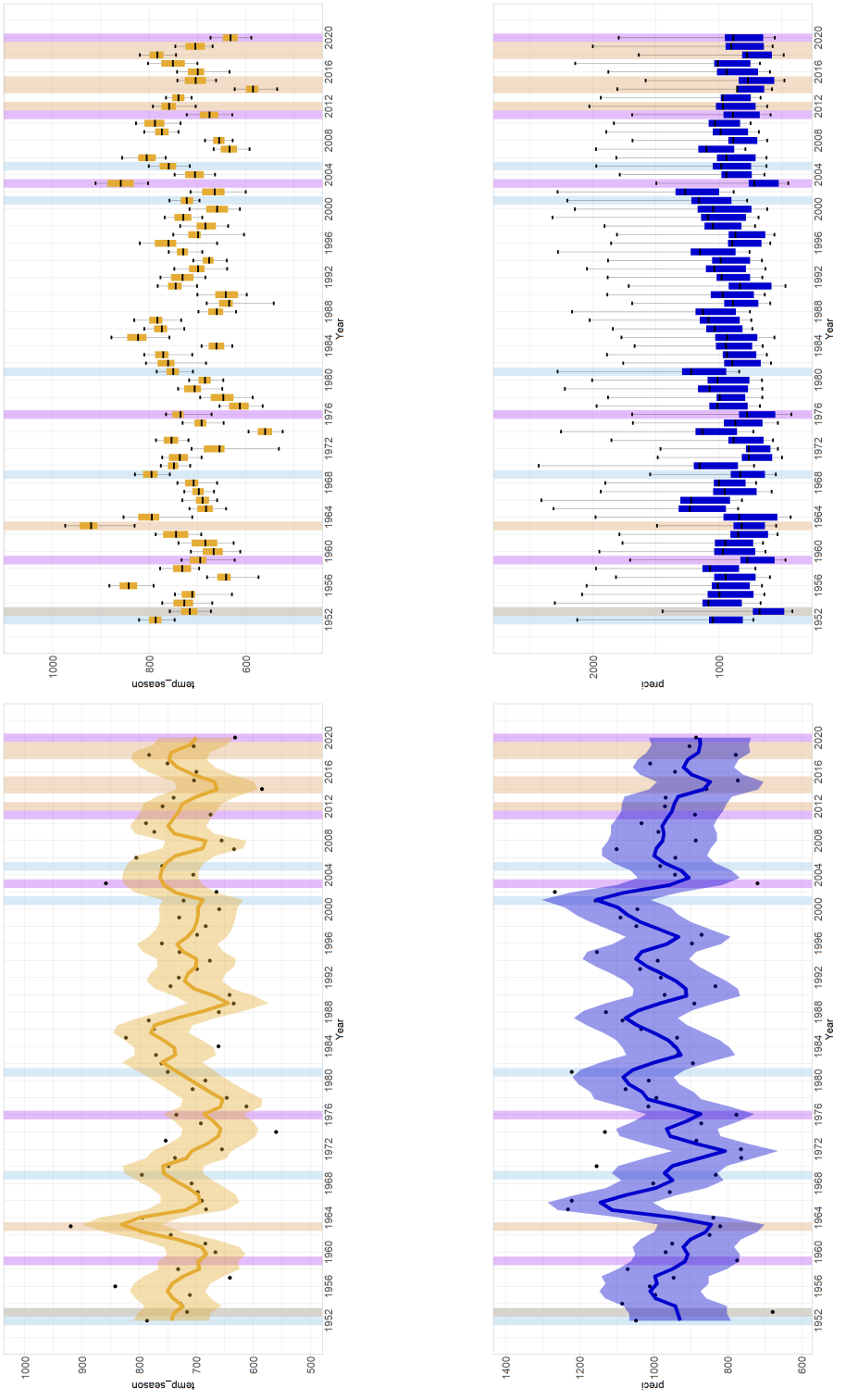

Year
1952
1956
1960
1964
1968
1972
1976
1980
1984
1988
1992
1996
2000
2004
2008
2012
2016
2020
isother
40
35
30
25
20

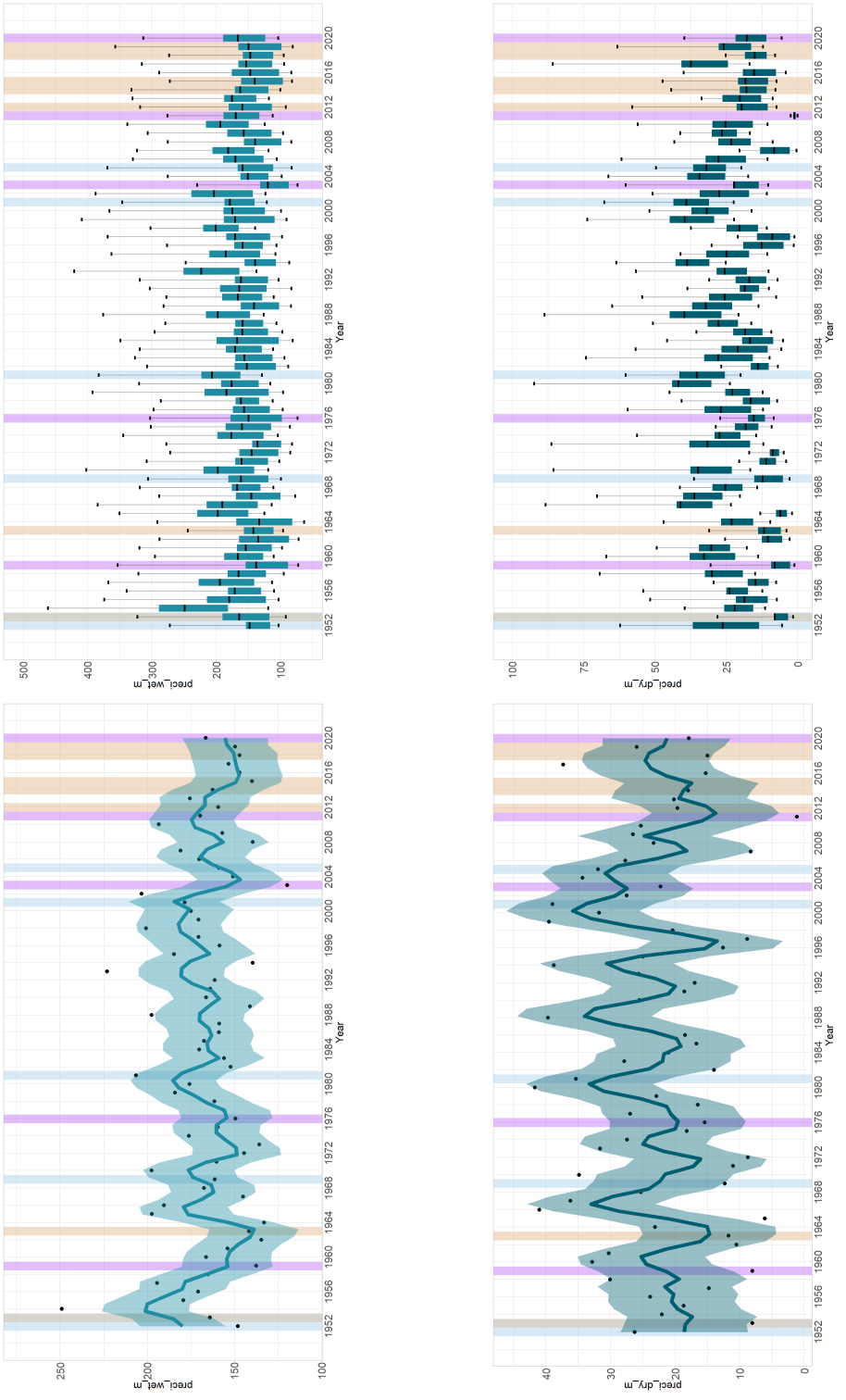

Year
1952
1956
1960
1964
1968
1972
1976
1980
1984
1988
1992
1996
2000
2004
2008
2012
2016
2020
temp_max
28
24
20

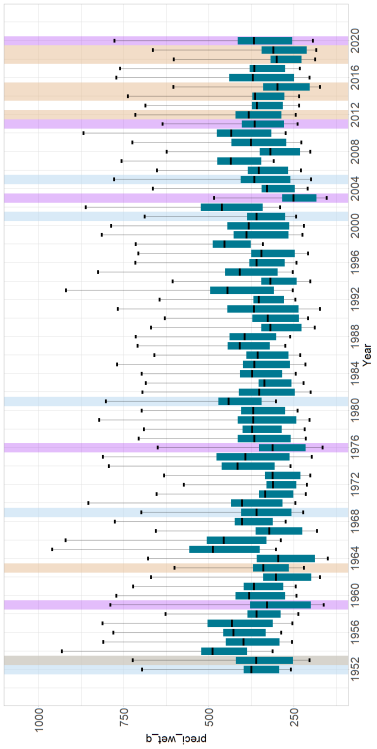

temp_mean
temp_range
temp_min
temp_range
Year
1952
1956
1960
1964
1968
1972
1976
1980
1984
1988
1992
1996
2000
2004
2008
2012
2016
2020

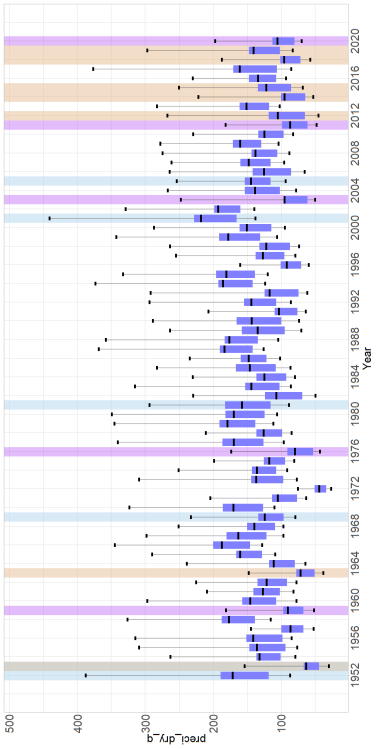

temp_wet
temp_dry
Year
1952
1956
1960
1964
1968
1972
1976
1980
1984
1988
1992
1996
2000
2004
2008
2012
2016
2020

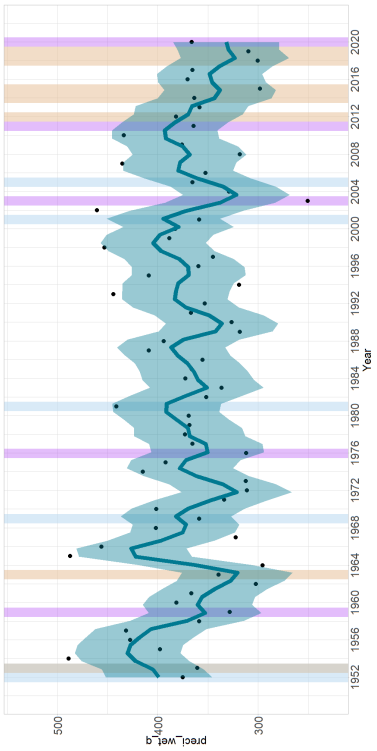

temp_warm
temp_cold
Year
1952
1956
1960
1964
1968
1972
1976
1980
1984
1988
1992
1996
2000
2004
2008
2012
2016
2020
20
15
0
-5
22.5
20.0
17.5
15.0
2.5
0.0
-2.5
-5.0
-7.5

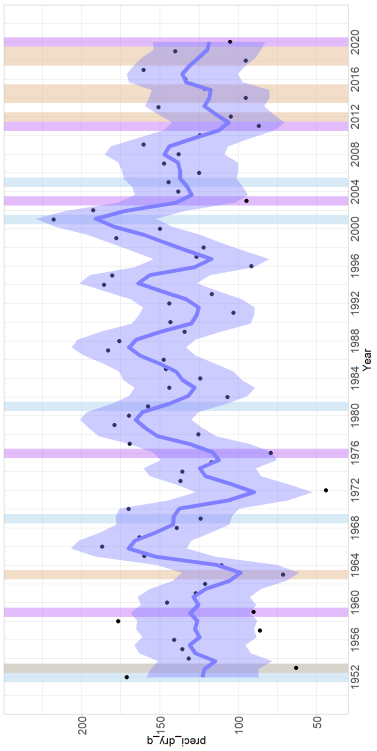

temp_season
1000
800
600
Year
2020
2016
2012
2008
2004
2000
1996
1992
1988
1984
1980
1976
1972
1968
1964
1960
1956
1952
preci
2000
1000
1400
1200
1000
800
600
900
700
500

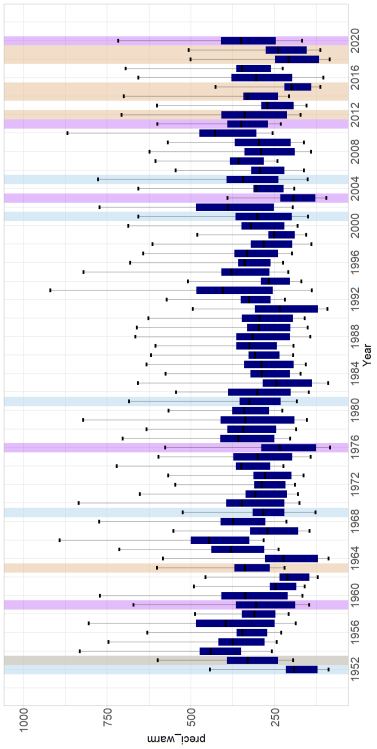

prec_wet_m
prec_dry_m
Year
1952
1956
1960
1964
1968
1972
1976
1980
1984
1988
1992
1996
2000
2004
2008
2012
2016
2020

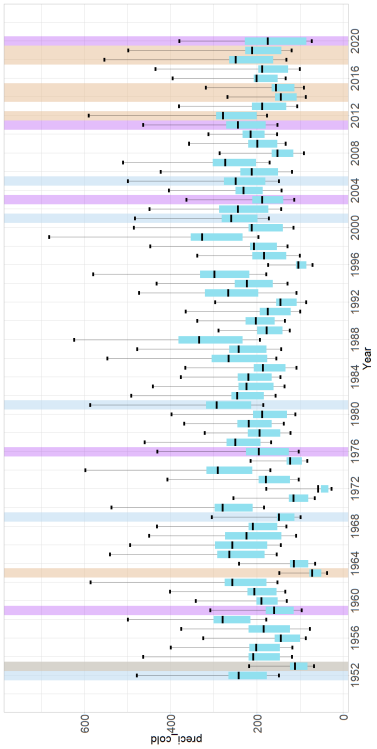

prec_wet_q
1000
750
500
250
Year
2020
2016
2012
2008
2004
2000
1996
1992
1988
1984
1980
1976
1972
1968
1964
1960
1956
1952

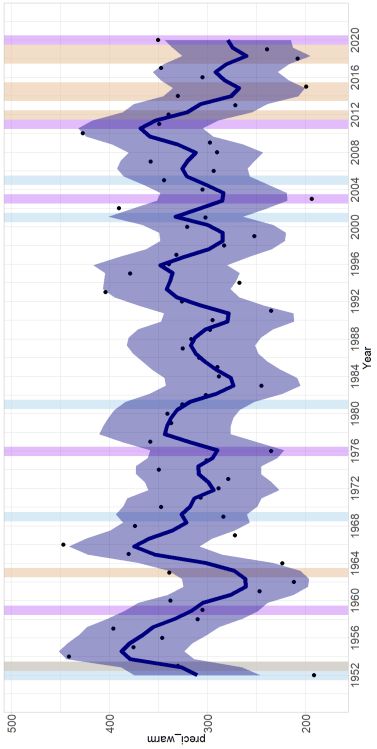

prec_dry_q
500
400
300
200
100
Year
2020
2016
2012
2008
2004
2000
1996
1992
1988
1984
1980
1976
1972
1968
1964
1960
1956
1952

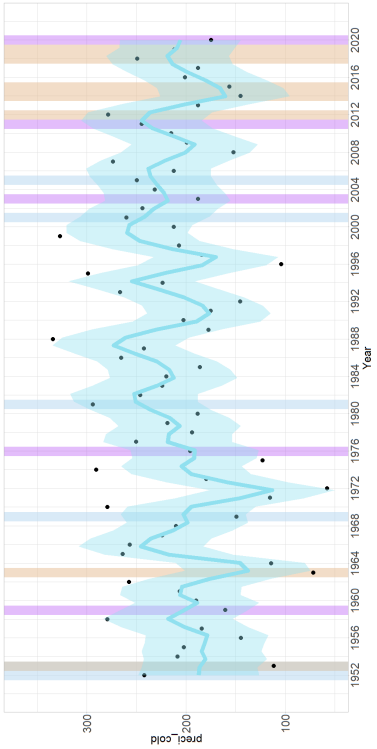

prec_wet_q
500
400
300
Year
2020
2016
2012
2008
2004
2000
1996
1992
1988
1984
1980
1976
1972
1968
1964
1960
1956
1952


prec_dry_q
200
150
100
50
Year
2020
2016
2012
2008
2004
2000
1996
1992
1988
1984
1980
1976
1972
1968
1964
1960
1956
1952


Year
prec_warm


Year
prec_cold


Year
prec_warm


Year
prec_cold

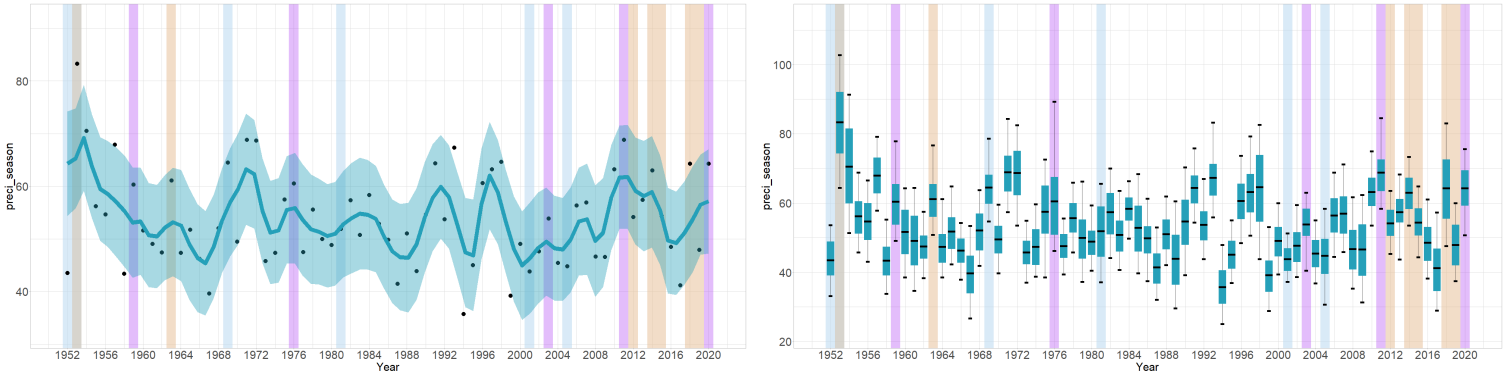


Figure C.1.: *Left column*: the points denote the mean observations per year over all gridcells (1952-2020), the smoothed line is a fitted moving average model, the shaded area is the 95% CI for each variable. *Right column:* the horizontal line represents the 95% CI per year, the box represents the 50% CI, and the horizontal line is the annual mean value over all gridcells. The vertical ribbons denote extreme years identified for frost risk (blue), drought risk (apricot), joint frost and drought risk (purple), and marginal drought and frost risk, but not joint risk identified (light gray).

C.2. Pairs plots

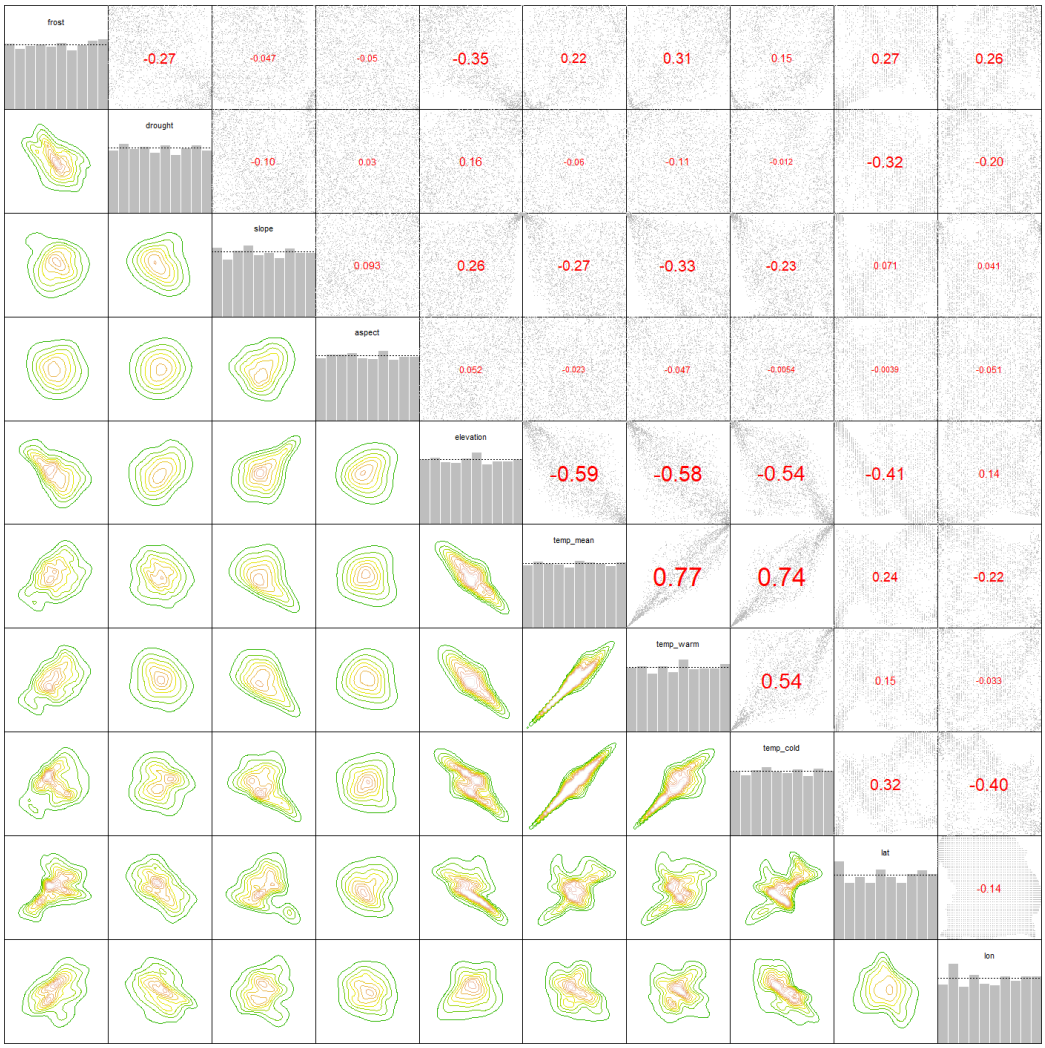


Figure C.2.: Lower diagonal: marginally normalized contour plots, upper diagonal: pairwise scatter plots with the associated empirical Kendall's $\hat{\tau}$ values and on the diagonal: histograms of the u-data, for the 2 responses and a subset of the possible predictor variables for all 2867 locations in year **1953**.

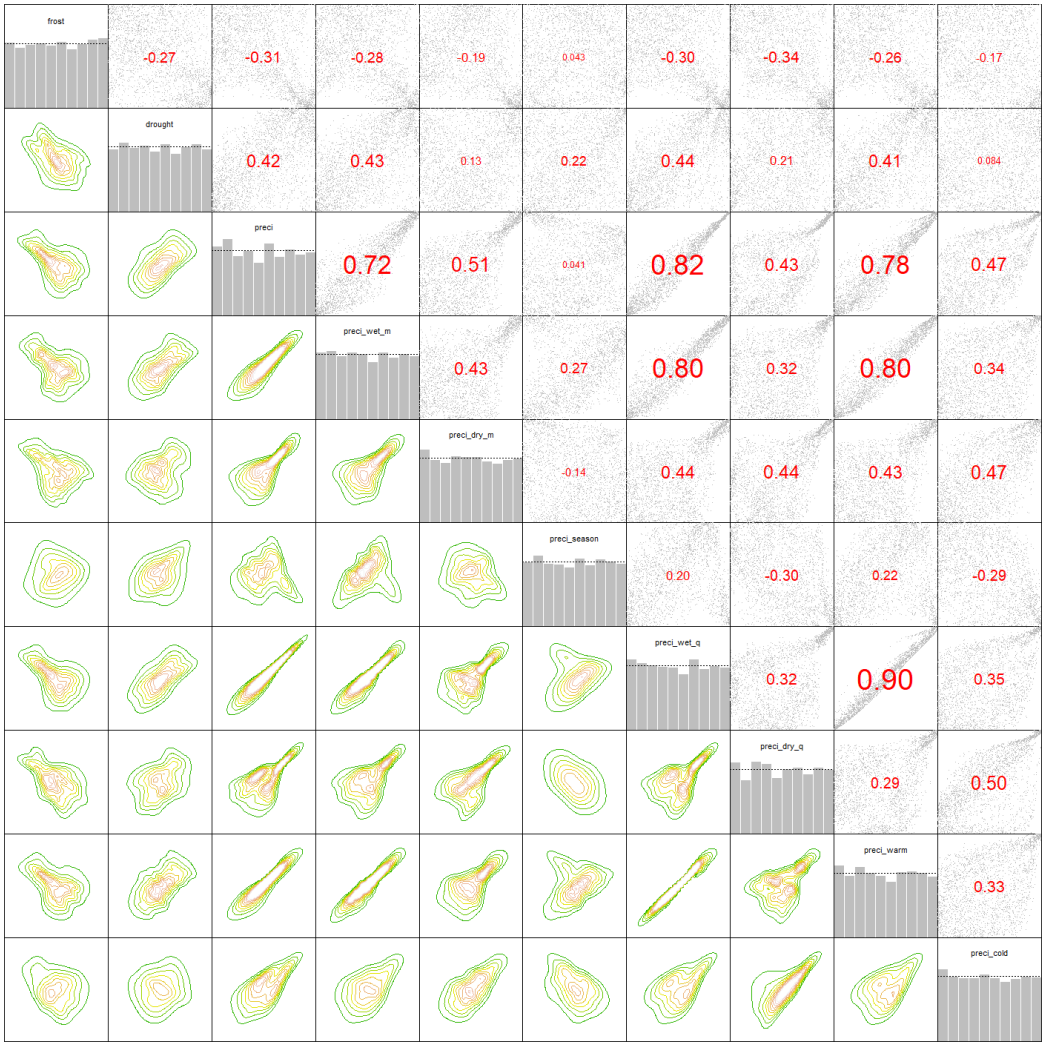


Figure C.3.: Lower diagonal: marginally normalized contour plots, upper diagonal: pairwise scatter plots with the associated empirical Kendall's $\hat{\tau}$ values and on the diagonal: histograms of the u-data, for the 2 responses and a different subset of the possible predictor variables for all 2867 locations in year **1953**.

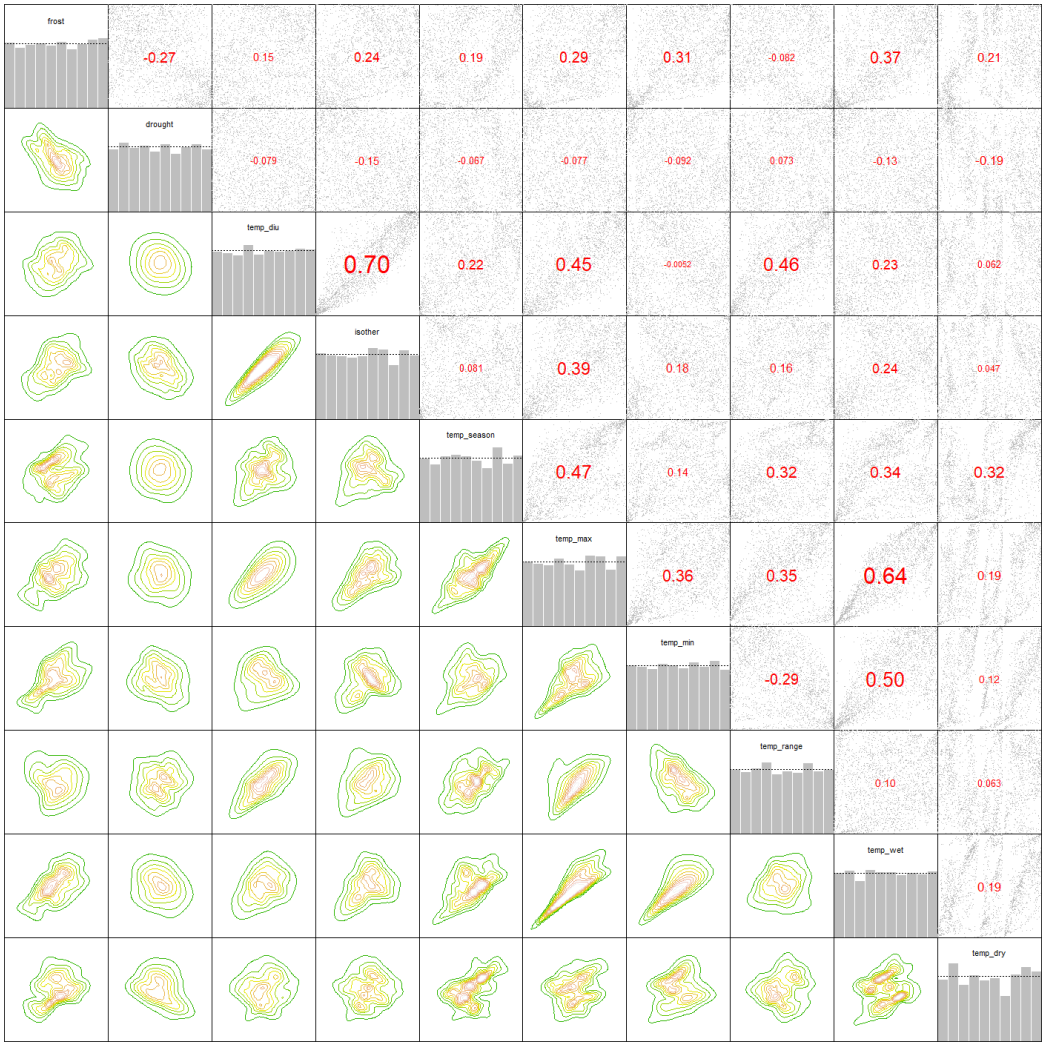


Figure C.4.: Lower diagonal: marginally normalized contour plots, upper diagonal: pairwise scatter plots with the associated empirical Kendall's $\hat{\tau}$ values and on the diagonal: histograms of the u-data, for the 2 responses and a different subset of the possible predictor variables for all 2867 locations in year **1953**.

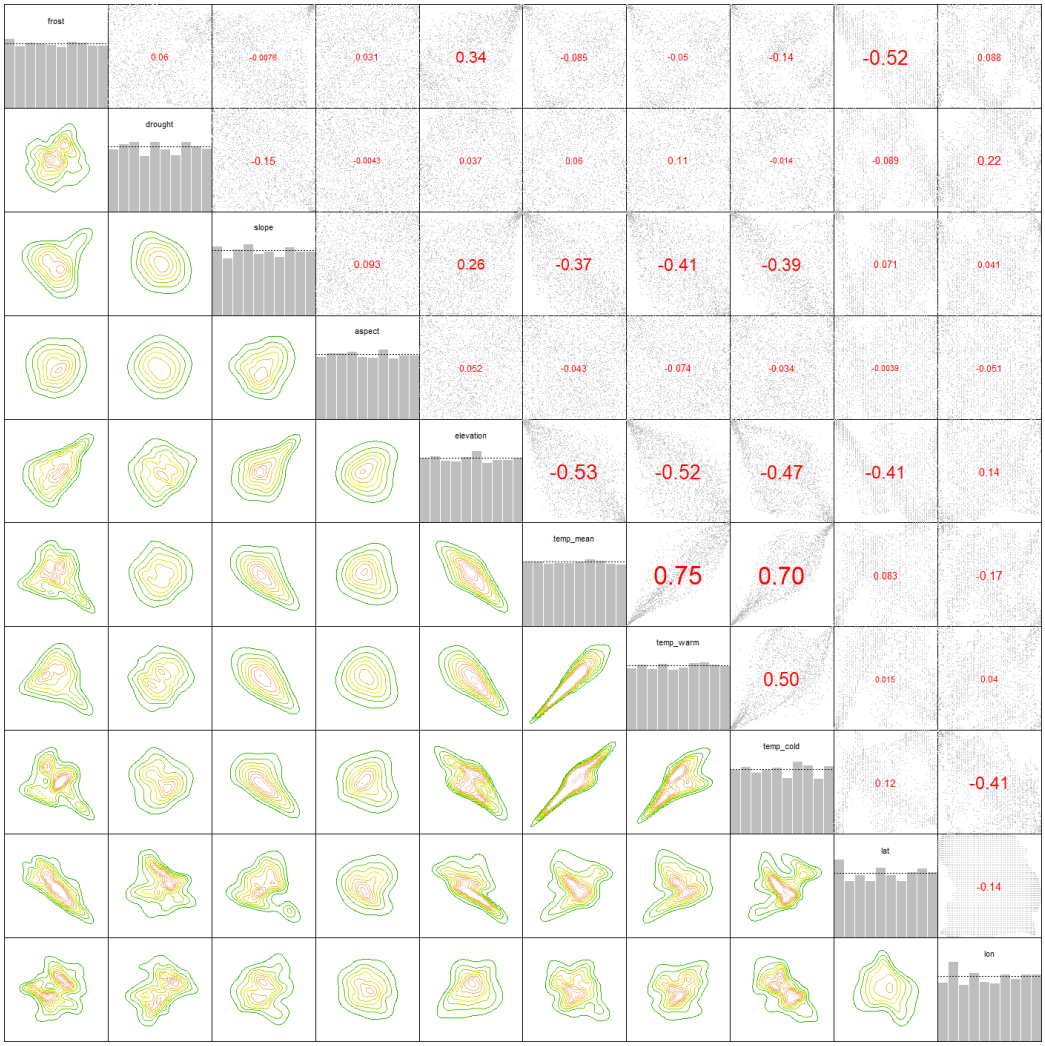


Figure C.5.: Lower diagonal: marginally normalized contour plots, upper diagonal: pairwise scatter plots with the associated empirical Kendall's $\hat{\tau}$ values and on the diagonal: histograms of the u-data, for the 2 responses and a subset of the possible predictor variables for all 2867 locations in year **2011**.

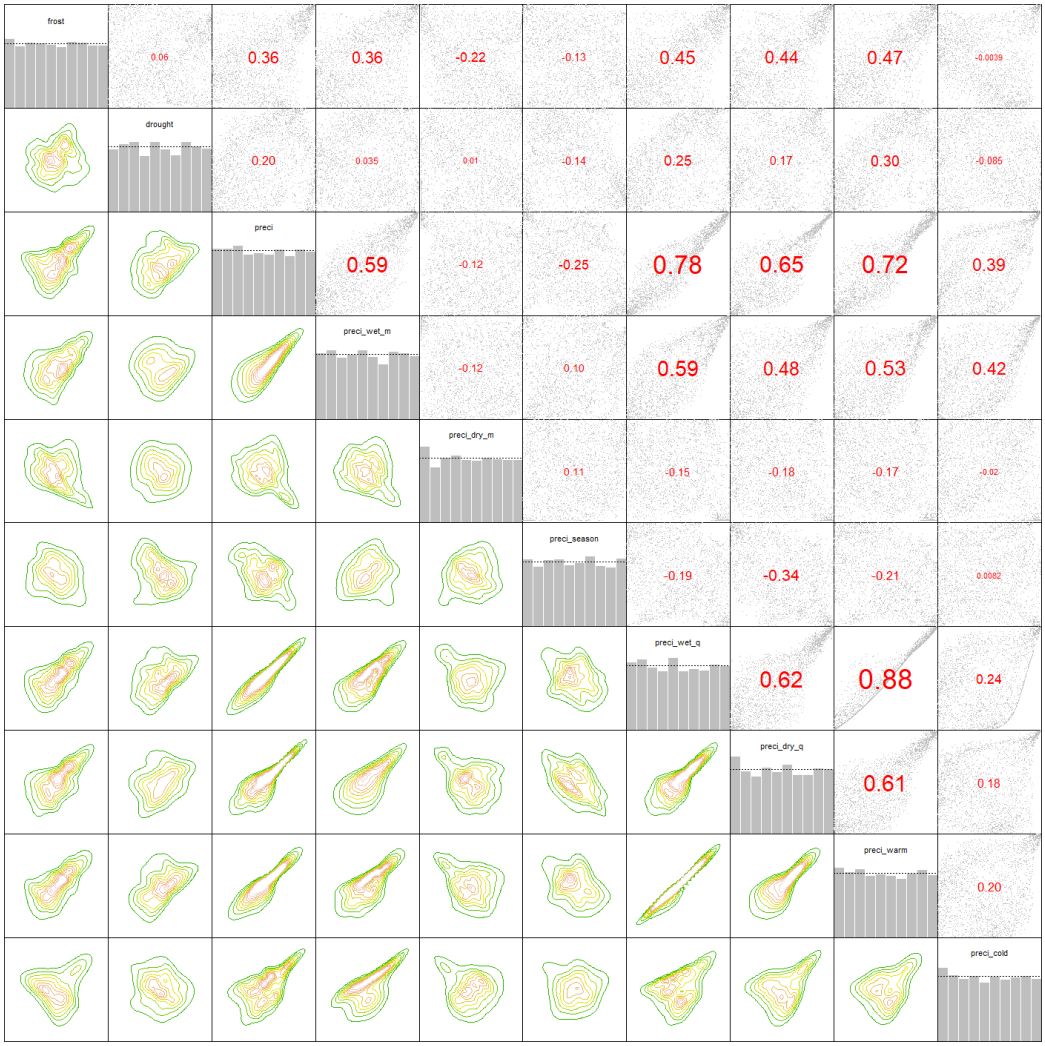


Figure C.6.: Lower diagonal: marginally normalized contour plots, upper diagonal: pairwise scatter plots with the associated empirical Kendall's $\hat{\tau}$ values and on the diagonal: histograms of the u-data, for the 2 responses and a different subset of the possible predictor variables for all 2867 locations in year **2011**.

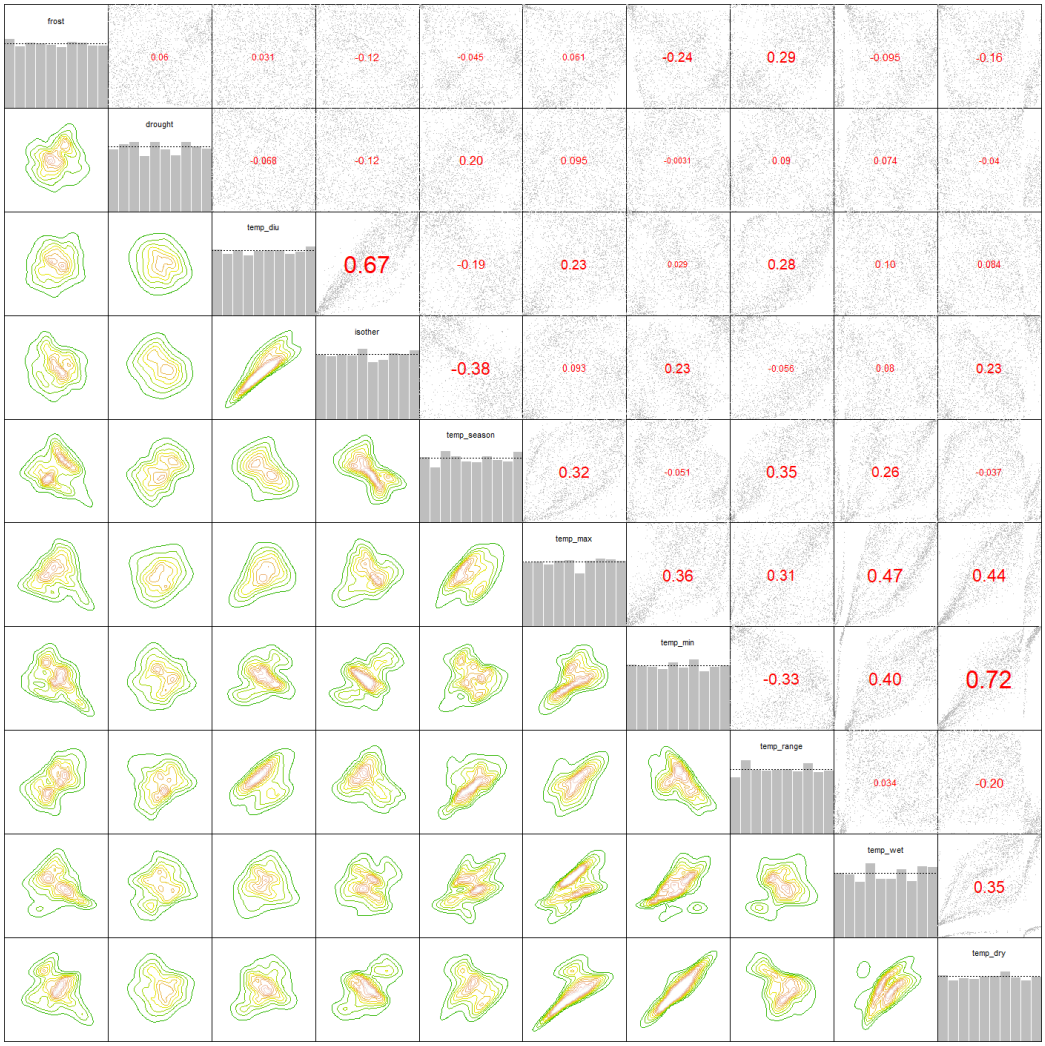


Figure C.7.: Lower diagonal: marginally normalized contour plots, upper diagonal: pairwise scatter plots with the associated empirical Kendall's $\hat{\tau}$ values and on the diagonal: histograms of the u-data, for the 2 responses and a different subset of the possible predictor variables for all 2867 locations in year **2011**.

List of Figures

List of Tables

Bibliography

Aas, K., C. Czado, A. Frigessi, and H. Bakken (2009). "Pair-copula constructions of multiple dependence." In: *Insurance: Mathematics and economics* 44.2, pp. 182–198.

Abdous, B. and R. Theodorescu (1992). "Note on the spatial quantile of a random vector." In: *Statistics & Probability Letters* 13.4, pp. 333–336.

Akaike, H. (1973a). "Theory and an extension of the maximum likelihood principal." In: *International symposium on information theory. Budapest, Hungary: Akademiai Kaiado*.

– (1973b). "Information theory and an extension of the maximum likelihood principle." In: *Second International Symposium on Information Theory*. Ed. by B. Petrov and F. Csáki. Akadémiai Kiadó, Budapest, pp. 267–281.

Ansell, L. and L. D. Valle (2021). "Social media integration of flood data: A vine copula-based approach." In: *arXiv preprint arXiv:2104.01869*.

Athey, S., J. Tibshirani, S. Wager, et al. (2019). "Generalized random forests." In: *The Annals of Statistics* 47.2, pp. 1148–1178.

Barbe, P., C. Genest, K. Ghoudi, and B. Remillard (1996). "On Kendall's process." In: *Journal of multivariate analysis* 58.2, pp. 197–229.

Barthel, N., C. Geerdens, M. Killiches, P. Janssen, and C. Czado (2018). "Vine copula based likelihood estimation of dependence patterns in multivariate event time data." In: *Computational Statistics & Data Analysis* 117, pp. 109–127.

Bauer, A. and C. Czado (2016). "Pair-copula Bayesian networks." In: *Journal of Computational and Graphical Statistics* 25.4, pp. 1248–1271.

Bayerisches Landesamt für Umwelt [Hrsg.] (2020). "Bayerische Klimadaten - Beobachtungsdaten, Klima- Projektionsensemble Und Klimakennwerte Für Bayern." In.

Bedford, T. and R. Cooke (2002). "Vines–a new graphical model for dependent random variables." In: *The Annals of Statistics* 30.4, pp. 1031–1068.

Beguería, S. and S. M. Vicente-Serrano (2017). *SPEI: Calculation of the Standardised Precipitation-Evapotranspiration Index*.

Beguería, S., S. M. Vicente-Serrano, F. Reig, and B. Latorre (2014). "Standardized Precipitation Evapotranspiration Index (SPEI) Revisited: Parameter Fitting, Evapotranspiration Models, Tools, Datasets and Drought Monitoring." In: *International Journal of Climatology* 34.10, pp. 3001–3023. ISSN: 1097-0088. DOI: 10.1002/joc.3887.

Belloni, A. and V. Chernozhukov (2011). "ℓ1-penalized quantile regression in high-dimensional sparse models." In: *The Annals of Statistics* 39.1, pp. 82–130.

Belzunce, F., A. Castaño, A. Olvera-Cervantes, and A. Suárez-Llorens (2007). "Quantile curves and dependence structure for bivariate distributions." In: *Computational Statistics & Data Analysis* 51.10, pp. 5112–5129.

Berg, D. and K. Aas (2009). "Models for construction of multivariate dependence: A comparison study." In: *The European Journal of Finance* 15.7, pp. 639–659.

Bernard, C. and C. Czado (2015). "Conditional quantiles and tail dependence." In: *Journal of Multivariate Analysis* 138, pp. 104–126.

Bevacqua, E., D. Maraun, I. Hobæk Haff, M. Widmann, and M. Vrac (2017). "Multivariate statistical modelling of compound events via pair-copula constructions: analysis of floods in Ravenna (Italy)." In: *Hydrology and Earth System Sciences* 21.6, pp. 2701–2723.

Bhuyan, U., C. Zang, and A. Menzel (June 2017). "Different Responses of Multispecies Tree Ring Growth to Various Drought Indices across Europe." In: *Dendrochronologia* 44, pp. 1–8. ISSN: 1125-7865. DOI: 10.1016/j.dendro.2017.02.002.

Bohn, U. and W. Welß (2003). "Die Potenzielle Natürliche Vegetation." In: *Klima, Pflanzen-Und Tierwelt. In: Leibnitz-Institut für Länderkunde [hrsg.]: Nationalatlas Bundesrepublik Deutschland* 3, pp. 84–87.

Bonferroni, C. (1936). "Teoria statistica delle classi e calcolo delle probabilita." In: *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commericiali di Firenze* 8, pp. 3–62.

Brechmann, E. C. (2013). "Hierarchical Kendall copulas and the modeling of systemic and operational risk." PhD thesis. München, Technische Universität München, Diss., 2013.

Breiman, L. (2001). "Random forests, machine learning 45." In: *J. Clin. Microbiol* 2.30, pp. 199–228.

Bühlmann, P. and S. van de Geer (2011). *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media.

Buras, A., A. Rammig, and C. S. Zang (Aug. 2019). *Quantifying Impacts of the Drought 2018 on European Ecosystems in Comparison to 2003*. Preprint. Earth System Science/Response to Global Change: Climate Change. DOI: 10.5194/bg-2019-286.

Carlier, G., V. Chernozhukov, A. Galichon, et al. (2016). "Vector quantile regression: an optimal transport approach." In: *Annals of Statistics* 44.3, pp. 1165–1192.

Carlier, G., V. Chernozhukov, and A. Galichon (2017). "Vector quantile regression beyond the specified case." In: *Journal of Multivariate Analysis* 161, pp. 96–102.

Chakak, A. and M. Ezzerg (2000). "Bivariate contours of copula." In: *Communications in Statistics-Simulation and Computation* 29.1, pp. 175–185.

Chang, B. and H. Joe (2019). "Prediction based on conditional distributions of vine copulas." In: *Computational Statistics & Data Analysis* 139, pp. 45–63.

Charpentier, A., J.-D. Fermanian, and O. Scaillet (2007). "The estimation of copulas: Theory and practice." In: *Copulas: from theory to application in finance*. Ed. by J. Rank. London : Risk Books, pp. 35–64.

Chatrabgoun, O., R. Karimi, A. Daneshkhah, S. Abolfathi, H. Nouri, and M. Esmaeilbeigi (2020). "Copula-based probabilistic assessment of intensity and duration of cold episodes: A case study of Malayer vineyard region." In: *Agricultural and Forest Meteorology* 295, p. 108150.

Chaudhuri, P. (1996). "On a geometric notion of quantiles for multivariate data." In: *Journal of the American Statistical Association* 91.434, pp. 862–872.

Chebana, F. and T. B. Ouarda (2011). "Multivariate quantiles in hydrological frequency analysis." In: *Environmetrics* 22.1, pp. 63–78.

Chen, X., R. Koenker, and Z. Xiao (2009). "Copula-based nonlinear quantile autoregression." In: *The Econometrics Journal* 12, S50–S67.

Cheng, Y., J. Du, and H. Ji (2020). "Multivariate joint probability function of earthquake ground motion prediction equations based on vine copula approach." In: *Mathematical Problems in Engineering* 2020, pp. 1–12.

Chernozhukov, V., A. Galichon, M. Hallin, M. Henry, et al. (2017). "Monge–Kantorovich depth, quantiles, ranks and signs." In: *Annals of Statistics* 45.1, pp. 223–256.

Cho, D., C. Yoo, J. Im, and D.-H. Cha (2020). "Comparative assessment of various machine learning-based bias correction methods for numerical weather prediction model forecasts of extreme air temperatures in urban areas." In: *Earth and Space Science* 7.4, e2019EA000740.

Claeskens, G. and N. Hjort (2003). "The focused information criterion." In: *Journal of the American Statistical Association* 98. With discussion and a rejoinder by the authors, pp. 900–916.

Coblenz, M., R. Dyckerhoff, and O. Grothe (2018). "Confidence regions for multivariate quantiles." In: *Water* 10.8, p. 996.

Cooke, R., H. Joe, and B. Chang (2022). "Vine regression with Bayes nets: A critical comparison with traditional approaches based on a case study on the effects of breastfeeding on IQ." In: *Risk Analysis* 42.6, pp. 1294–1305.

Cowell, R. G., P. Dawid, S. L. Lauritzen, and D. J. Spiegelhalter (2007). *Probabilistic networks and expert systems: Exact computational methods for Bayesian networks*. Springer Science & Business Media.

Czado, C. (2010). "Pair-copula constructions of multivariate copulas." In: *Copula theory and its applications*. Springer, pp. 93–109.

– (2019). "Analyzing dependent data with vine copulas." In: *Lecture Notes in Statistics, Springer*.

Czado, C., K. Bax, Ö. Sahin, T. Nagler, A. Min, and S. Paterlini (2022). "Vine copula based dependence modeling in sustainable finance." In: *The Journal of Finance and Data Science*.

Czado, C. and S. Scharl (2021). "Analysis of an interventional protein experiment using a vine copula based structural equation model." In: *arXiv preprint arXiv:2111.10113*.

D'Andrea, E., N. Rezaie, A. Battistelli, O. Gavrichkova, I. Kuhlmann, G. Matteucci, S. Moscatello, S. Proietti, A. Scartazza, S. Trumbore, and J. Muhr (Oct. 2019). "Winter's Bite: Beech Trees Survive Complete Defoliation Due to Spring Late-Frost Damage by Mobilizing Old c Reserves." In: *New Phytol* 224.2, pp. 625–631. ISSN: 0028-646X, 1469-8137. DOI: 10.1111/nph.16047.

D'Urso, P., L. De Giovanni, and V. Vitale (2022). "A D-vine copula-based quantile regression model with spatial dependence for COVID-19 infection rate in Italy." In: *Spatial statistics* 47, p. 100586.

Darsow, W. F., B. Nguyen, E. T. Olsen, et al. (1992). "Copulas and Markov processes." In: *Illinois journal of mathematics* 36.4, pp. 600–642.

Di Bernardino, E. and C. Prieur (2014). "Estimation of multivariate conditional-tail-expectation using Kendall's process." In: *Journal of Nonparametric Statistics* 26.2, pp. 241–267.

Dißmann, J. F. (May 2010). "Statistical Inference for Regular Vines and Application." Diplomarbeit. Technische Universität München.

Dissmann, J., E. C. Brechmann, C. Czado, and D. Kurowicka (2013). "Selecting and estimating regular vine copulae and application to financial returns." In: *Computational Statistics & Data Analysis* 59, pp. 52–69.

Dittmar, C., W. Fricke, and W. Elling (2006). "Impact of Late Frost Events on Radial Growth of Common Beech (Fagus Sylvatica l.) in Southern Germany." In: *European Journal of Forest Research* 125.3, pp. 249–259. ISSN: 1612-4669. DOI: 10.1007/s10342-005-0098-y.

Dua, D. and C. Graff (2017). *UCI Machine Learning Repository*.

– (2019). *UCI Machine Learning Repository*.

Durrant, T. H., D. De Rigo, and G. Caudullo (2016). "Fagus Sylvatica in Europe: Distribution, Habitat, Usage and Threats." In: *European Atlas of Forest Tree Species*. Publication Office of the European Union Luxembourg, pp. 94–95.

Elidan, G. (2010). "Copula Bayesian networks." In: *Advances in neural information processing systems*, pp. 559–567.

– (2012). "Inference-less density estimation using Copula Bayesian Networks." In: *arXiv preprint arXiv:1203.3476*.

Embrechts, P., F. Lindskog, and A. McNeil (2003). *Modelling dependence with copulas and applications to risk management. Handbook of Heavy Tailed Distributions in Finance, edited by ST Rachev*.

Ezzerg, M., A. Chakak, and L. Imlahi (1999). "Estimación de la curva mediana de una cópula C (x1,..., xm)." In: *Revista de la Real Academia de Ciencias Exactas, F'ısicas y Naturales* 93.2, pp. 241–250.

Fenske, N., T. Kneib, and T. Hothorn (2011). "Identifying risk factors for severe childhood malnutrition by boosting additive quantile regression." In: *Journal of the American Statistical Association* 106.494, pp. 494–510.

Fernández-Ponce, J. M. and A. Suárez-Lloréns (2002). "Central regions for bivariate distributions." In: *Austrian Journal of Statistics* 31.2&3, pp. 141–156.

Fick, S. E. and R. J. Hijmans (Oct. 2017). "WorldClim 2: New 1-km Spatial Resolution Climate Surfaces for Global Land Areas." In: *International Journal of Climatology* 37.12, pp. 4302–4315. ISSN: 0899-8418, 1097-0088. DOI: 10.1002/joc.5086.

Frees, E. W., G. Lee, and L. Yang (2016). "Multivariate frequency-severity regression models in insurance." In: *Risks* 4.1, p. 4.

Friedlingstein, P. et al. (Nov. 2022). "Global Carbon Budget 2022." In: *Earth System Science Data* 14.11, pp. 4811–4900. ISSN: 1866-3508. DOI: 10.5194/essd-14-4811-2022.

Geenens, G., A. Charpentier, and D. Paindaveine (2017). "Probit transformation for nonparametric kernel estimation of the copula density." In: *Bernoulli* 23.3, pp. 1848–1873.

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

Gijbels, I. and J. Mielniczuk (1990). "Estimating the density of a copula function." In: *Communications in Statistics-Theory and Methods* 19.2, pp. 445–464.

Gijbels, I. and M. Matterne (2021). "Study of partial and average conditional Kendall's tau." In: *Dependence Modeling* 9.1, pp. 82–120.

Gijbels, I., N. Veraverbeke, and M. Omelka (2011). "Conditional copulas, association measures and their applications." In: *Computational Statistics & Data Analysis* 55.5, pp. 1919–1932.

Gneiting, T. and A. E. Raftery (2007). "Strictly proper scoring rules, prediction, and estimation." In: *Journal of the American Statistical Association* 102.477, pp. 359–378.

Grønneberg, S. and N. L. Hjort (2014). "The Copula Information Criteria." In: *Scandinavian Journal of Statistics* 41.2, pp. 436–459. DOI: 10.1111/sjos.12042. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/sjos.12042.

Guilbaud, O. (2008). "Simultaneous confidence regions corresponding to Holm's step-down procedure and other closed-testing procedures." In: *Biometrical Journal: Journal of Mathematical Methods in Biosciences* 50.5, pp. 678–692.

Haff, I. H., K. Aas, and A. Frigessi (2010). "On the simplified pair-copula construction—simply useful or too simplistic?" In: *Journal of Multivariate Analysis* 101.5, pp. 1296–1310.

Haff, I. H., K. Aas, A. Frigessi, and V. Lacal (2016). "Structure learning in Bayesian Networks using regular vines." In: *Computational Statistics & Data Analysis* 101, pp. 186–208.

Hallin, M., D. Paindaveine, M. Šiman, Y. Wei, R. Serfling, Y. Zuo, L. Kong, and I. Mizera (2010). "Multivariate quantiles and multiple-output regression quantiles: from L-1 optimiziation to halfspace depth [with Discussion and Rejoinder]." In: *The Annals of Statistics*, pp. 635–703.

Hanbo Li, A. and A. Martin (2017). "Forest-type regression with general losses and robust forest." In: *International Conference on Machine Learning*, pp. 2091–2100.

Hijmans, R. J. (2022). *Terra: Spatial Data Analysis*. Manual.

Hijmans, R. J., S. E. Cameron, J. L. Parra, P. G. Jones, and A. Jarvis (2005). "Very High Resolution Interpolated Climate Surfaces for Global Land Areas." In: *International Journal of Climatology* 25.15, pp. 1965–1978. ISSN: 1097-0088. DOI: 10.1002/joc.1276.

Hijmans, R. J., S. Phillips, J. Leathwick, and J. Elith (2021). *Dismo: Species Distribution Modeling*. Manual.

Hürlimann, W. (2003). "Hutchinson-Lai's conjecture for bivariate extreme value copulas." In: *Statistics & probability letters* 61.2, pp. 191–198.

Joe, H. (1996). "Families of m-variate distributions with given margins and m (m-1)/2 bivariate dependence parameters." In: *Lecture Notes-Monograph Series*, pp. 120–141.

– (1997). *Multivariate models and multivariate dependence concepts*. CRC press.

Jullum, M. and N. L. Hjort (2017). "Parametric or nonparametric: the FIC approach." In: *Statistica Sinica* 27.3, pp. 951–981.

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

Kielmann, J., H. Manner, and A. Min (2022). "Stock market returns and oil price shocks: A CoVaR analysis based on dynamic vine copula models." In: *Empirical Economics* 62.4, pp. 1543–1574.

Killiches, M., D. Kraus, and C. Czado (2018). "Model distances for vine copulas in high dimensions." In: *Statistics and Computing* 28.2, pp. 323–341.

Kim, D., J.-M. Kim, S.-M. Liao, and Y.-S. Jung (2013). "Mixture of D-vine copulas for modeling dependence." In: *Computational Statistics & Data Analysis* 64, pp. 1–19.

Kirshner, S. (2008). "Learning with tree-averaged densities and distributions." In: *Advances in Neural Information Processing Systems*, pp. 761–768.

Klein, J. P., M. L. Moeschberger, J. P. Klein, and M. L. Moeschberger (1997). "Censoring and truncation." In: *Survival Analysis: Techniques for Censored and Truncated Data*, pp. 55–82.

Ko, V. and N. L. Hjort (2019). "Copula information criterion for model selection with two-stage maximum likelihood estimation." In: *Econometrics and Statistics* 12, pp. 167–180. ISSN: 2452-3062. DOI: https://doi.org/10.1016/j.ecosta.2019.01.001.

Ko, V., N. L. Hjort, and I. Hobæk Haff (2019). "Focused information criteria for copulas." In: *Scandinavian Journal of Statistics* 46.4, pp. 1117–1140. DOI: 10.1111/sjos.12387. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/sjos.12387.

Koenker, R. (2005). *Quantile Regression*. Econometric Society Monographs. Cambridge University Press. ISBN: 9780521608275.

Koenker, R. (2004). "Quantile regression for longitudinal data." In: *Journal of Multivariate Analysis* 91.1, pp. 74–89.

– (2011). "Additive models for quantile regression: Model selection and confidence bandaids." In: *Brazilian Journal of Probability and Statistics* 25.3, pp. 239–262.

– (2017). "Quantile regression: 40 years on." In: *Annual Review of Economics* 9, pp. 155–176.

Koenker, R. and G. Bassett (1978). "Regression quantiles." In: *Econometrica: journal of the Econometric Society*.

Koller, D. and N. Friedman (2009). *Probabilistic graphical models: principles and techniques*. MIT press.

Komunjer, I. (2013). *Quantile Prediction, Chapter 17 in Handbook of Financial Econometrics, edited by Yacine Ait-Sahalia and Lars Peter Hansen*. Elsevier.

Korpela, J., E. Oikarinen, K. Puolamäki, and A. Ukkonen (2017). "Multivariate confidence intervals." In: *Proceedings of the 2017 SIAM International Conference on Data Mining*. SIAM, pp. 696–704.

Korpela, J., K. Puolamäki, and A. Gionis (2014). "Confidence bands for time series data." In: *Data mining and knowledge discovery* 28.5, pp. 1530–1553.

Kramer, K., A. Ducousso, D. Gömöry, J. K. Hansen, L. Ionita, M. Liesebach, A. Lorenţ, S. Schüler, M. Sulkowska, S. de Vries, and G. von Wühlisch (Mar. 2017). "Chilling and Forcing Requirements for Foliage Bud Burst of European Beech (Fagus Sylvatica L.) Differ between Provenances and Are Phenotypically Plastic." In: *Agricultural and Forest Meteorology* 234–235, pp. 172–181. ISSN: 01681923. DOI: 10.1016/j.agrformet.2016.12.002.

Kraus, D. and C. Czado (2017). "D-vine copula based quantile regression." In: *Computational Statistics & Data Analysis* 110, pp. 1–18.

Kreuzer, A. and C. Czado (2021). "Bayesian inference for a single factor copula stochastic volatility model using Hamiltonian Monte Carlo." In: *Econometrics and Statistics* 19, pp. 130–150.

Kurowicka, D. and R. Cooke (2002). "The vine copula method for representing high dimensional dependent distributions: application to continuous belief nets." In: *Proceedings of the Winter Simulation Conference*. Vol. 1. IEEE, pp. 270–278.

– (2005). "Distribution-free continuous Bayesian belief." In: *Modern statistical and mathematical methods in reliability* 10, p. 309.

Kurowicka, D. and R. Cooke (2006). *Uncertainty analysis with high dimensional dependence modelling*. John Wiley & Sons.

Kurz, M. S. and F. Spanhel (2022). "Testing the simplifying assumption in high-dimensional vine copulas." In: *Electronic Journal of Statistics* 16.2, pp. 5226–5276.

Kwon, H.-H. and U. Lall (2016). "A copula-based nonstationary frequency analysis for the 2012–2015 drought in California." In: *Water Resources Research* 52.7, pp. 5662–5675.

Lasmar, N.-E. and Y. Berthoumieu (2014). "Gaussian copula multivariate modeling for texture image retrieval using wavelet transforms." In: *IEEE Transactions on Image Processing* 23.5, pp. 2246–2261.

Lauritzen, S. L. (1996). *Graphical models*. Vol. 17. Oxford University Press.

Laux, P., S. Vogl, W. Qiu, H. R. Knoche, and H. Kunstmann (2011). "Copula-based statistical refinement of precipitation in RCM simulations over complex terrain." In: *Hydrology and Earth System Sciences* 15.7, pp. 2401–2419.

Leuschner, C. and H. Ellenberg (2017). *Ecology of Central European Forests: Vegetation Ecology of Central Europe, Volume I*. Vol. 1. Springer.

Lewis, T. and J. Thompson (1981). "Dispersive distributions, and the connection between dispersivity and strong unimodality." In: *Journal of Applied Probability* 18.1, pp. 76–90.

Li, H., G. Huang, Y. Li, J. Sun, and P. Gao (2021). "A C-vine copula-based quantile regression method for streamflow forecasting in Xiangxi river basin, China." In: *Sustainability* 13.9, p. 4627.

Li, Q., J. Lin, and J. S. Racine (2013). "Optimal bandwidth selection for nonparametric conditional distribution and quantile functions." In: *Journal of Business & Economic Statistics* 31.1, pp. 57–65.

Liang, Y., G. Chen, S. Naqvi, and J. A. Chambers (2013). "Independent vector analysis with multivariate student's t-distribution source prior for speech separation." In: *Electronics Letters* 49.16, pp. 1035–1036.

Liebscher, E. (2006). "Modelling and estimation of multivariate copulas." In: *Working Paper*.

Liu, H., J. Lafferty, and L. Wasserman (2009). "The nonparanormal: Semiparametric estimation of high dimensional undirected graphs." In: *Journal of Machine Learning Research* 10.10.

Loader, C. (2006). *Local regression and likelihood*. Springer Science & Business Media.

Meinshausen, N. (2006). "Quantile regression forests." In: *Journal of Machine Learning Research* 7.Jun, pp. 983–999.

Menzel, A., R. Helm, and C. Zang (Feb. 2015). "Patterns of Late Spring Frost Leaf Damage and Recovery in a European Beech (Fagus Sylvatica l.) Stand in South-Eastern Germany Based on Repeated Digital Photographs." In: *Front. Plant Sci.* 6. ISSN: 1664-462X. DOI: 10.3389/fpls.2015.00110.

Meyer, B. F., A. Buras, A. Rammig, and C. S. Zang (Nov. 2020). "Higher Susceptibility of Beech to Drought in Comparison to Oak." In: *Dendrochronologia*, p. 125780. ISSN: 1125-7865. DOI: 10.1016/j.dendro.2020.125780.

Nagler, T. (2022). *vinereg: D-Vine Quantile Regression*. R package version 0.8.3.

Nagler, T., D. Krüger, and A. Min (2022). "Stationary vine copula models for multivariate time series." In: *Journal of Econometrics* 227.2, pp. 305–324.

Nagler, T., C. Schellhase, and C. Czado (2017). "Nonparametric estimation of simplified vine copula models: comparison of methods." In: *Dependence Modeling* 5.1, pp. 99–120.

Nagler, T. and T. Vatter (2020). *kde1d: Univariate Kernel Density Estimation*. R package version 1.0.3.

– (2021). *rvinecopulib: High Performance Algorithms for Vine Copula Modeling*. R package version 0.6.1.1.1.

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

Niemierko, R., J. Töppel, and T. Tränkler (2019). "A D-vine copula quantile regression approach for the prediction of residential heating energy consumption based on historical data." In: *Applied energy* 233, pp. 691–708.

Noh, H., A. El Ghouch, and T. Bouezmarni (2013). "Copula-based regression estimation and inference." In: *Journal of the American Statistical Association* 108.502, pp. 676–688.

Noh, H., A. El Ghouch, and I. Van Keilegom (2015). "Semiparametric conditional quantile estimation through copula-based multivariate models." In: *Journal of Business & Economic Statistics* 33.2, pp. 167–178.

Noyan, N. and G. Rudolf (2013). "Optimization with multivariate conditional value-at-risk constraints." In: *Operations research* 61.4, pp. 990–1013.

Olano, J. M., A. I. García-Cervigón, G. Sangüesa-Barreda, V. Rozas, D. Muñoz-Garachana, M. García-Hidalgo, and Á. García-Pedrero (Apr. 2021). "Satellite Data and Machine Learning Reveal the Incidence of Late Frost Defoliations on Iberian Beech Forests." In: *Ecological Applications* 31.3. ISSN: 1051-0761, 1939-5582. DOI: 10.1002/eap.2288.

Piessens, R., E. de Doncker-Kapenga, C. W. Überhuber, and D. K. Kahaner (2012). *Quadpack: a subroutine package for automatic integration*. Vol. 1. Springer Science & Business Media.

Qian, J. and Y. Dong (2022). "Surrogate-assisted seismic performance assessment incorporating vine copula captured dependence." In: *Engineering Structures* 257, p. 114073.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria.

– (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria.

Renard, B. and M. Lang (2007). "Use of a Gaussian copula for multivariate extreme value analysis: Some case studies in hydrology." In: *Advances in Water Resources* 30.4, pp. 897–912.

Requena, A., L. Mediero, and L. Garrote (2013). "A bivariate return period based on copulas for hydrologic dam design: accounting for reservoir routing in risk estimation." In: *Hydrology and Earth System Sciences* 17.8, pp. 3023–3038.

Sahin, Ö. and C. Czado (2022). "Vine copula mixture models and clustering for non-Gaussian data." In: *Econometrics and Statistics* 22, pp. 136–158.

Salvadori, G., F. Durante, G. Tomasicchio, and F. D'Alessandro (2015). "Practical guidelines for the multivariate assessment of the structural risk in coastal and off-shore engineering." In: *Coastal Engineering* 95, pp. 77–83.

Sangüesa-Barreda, G., A. Di Filippo, G. Piovesan, V. Rozas, L. Di Fiore, M. García-Hidalgo, A. I. García-Cervigón, D. Muñoz-Garachana, M. Baliva, and J. M. Olano (June 2021). "Warmer Springs Have Increased the Frequency and Extension of Late-Frost Defoliations in Southern European Beech Forests." In: *Science of The Total Environment* 775, p. 145860. ISSN: 00489697. DOI: 10.1016/j.scitotenv.2021.145860.

Sarhadi, A., D. H. Burn, M. Concepcion Ausin, and M. P. Wiper (2016). "Time-varying nonstationary multivariate risk analysis using a dynamic Bayesian copula." In: *Water Resources Research* 52.3, pp. 2327–2349.

Savu, C. and M. Trede (2010). "Hierarchies of Archimedean copulas." In: *Quantitative Finance* 10.3, pp. 295–304.

Schallhorn, N., D. Kraus, T. Nagler, and C. Czado (2017). "D-vine quantile regression with discrete variables." In: *arXiv preprint arXiv:1705.08310*.

Scharnweber, T., M. Manthey, C. Criegee, A. Bauwe, C. Schröder, and M. Wilmking (2011). "Drought Matters – Declining Precipitation Influences Growth of Fagus Sylvatica L. and Quercus Robur L. in North-Eastern Germany." In: *Forest Ecology and Management* 262.6, pp. 947–961. ISSN: 0378-1127. DOI: 10.1016/j.foreco.2011.05.026.

Schölzel, C. and P. Friederichs (Oct. 2008). "Multivariate Non-Normally Distributed Random Variables in Climate Research – Introduction to the Copula Approach." In: *Nonlinear Processes in Geophysics* 15.5, pp. 761–772. ISSN: 1607-7946. DOI: 10.5194/npg-15-761-2008.

Schuldt, B., A. Buras, M. Arend, Y. Vitasse, C. Beierkuhnlein, A. Damm, M. Gharun, T. E. Grams, M. Hauck, P. Hajek, H. Hartmann, E. Hiltbrunner, G. Hoch, M. Holloway-Phillips, C. Körner, E. Larysch, T. Lübbe, D. B. Nelson, A. Rammig, A. Rigling, L. Rose, N. K. Ruehr, K. Schumann, F. Weiser, C. Werner, T. Wohlgemuth, C. S. Zang, and A. Kahmen (2020). "A First Assessment of the Impact of the Extreme 2018 Summer Drought on Central European Forests." In: *Basic and Applied Ecology* 45, pp. 86–103. ISSN: 1439-1791. DOI: 10.1016/j.baae.2020.04.003.

Schwarz, G. (1978). "Estimating the dimension of a model." In: *The Annals of Statistics* 6.2, pp. 461–464.

Scott, D. W. (2008). "The curse of dimensionality and dimension reduction." In: *Multivariate Density Estimation: Theory, Practice, and Visualization* 1, pp. 217–40.

Sheather, S. J. and M. C. Jones (1991). "A reliable data-based bandwidth selection method for kernel density estimation." In: *Journal of the Royal Statistical Society: Series B (Methodological)* 53.3, pp. 683–690.

Shiau, J.-T. and R. Modarres (2009). "Copula-based drought severity-duration-frequency analysis in Iran." In: *Meteorological Applications: A journal of forecasting, practical applications, training techniques and modelling* 16.4, pp. 481–489.

Singh, H., M. R. Najafi, and A. Cannon (2022). "Evaluation and joint projection of temperature and precipitation extremes across Canada based on hierarchical Bayesian modelling and large ensembles of regional climate simulations." In: *Weather and Climate Extremes* 36, p. 100443.

Sklar, M. (1959). "Fonctions de repartition an dimensions et leurs marges." In: *Publ. inst. statist. univ. Paris* 8, pp. 229–231.

Slette, I. J., A. K. Post, M. Awad, T. Even, A. Punzalan, S. Williams, M. D. Smith, and A. K. Knapp (Oct. 2019). "How Ecologists Define Drought, and Why We Should Do Better." In: *Global Change Biology* 25.10, pp. 3193–3200. ISSN: 1354-1013, 1365-2486. DOI: 10.1111/gcb.14747.

Spanhel, F. and M. S. Kurz (2015). "Simplified vine copula models: Approximations based on the simplifying assumption." In: *arXiv preprint arXiv:1510.06971* 715.

Spinoni, J., G. Naumann, and J. V. Vogt (Jan. 2017). "Pan-European Seasonal Trends and Recent Changes of Drought Frequency and Severity." In: *Global and Planetary Change* 148, pp. 113–130. ISSN: 0921-8181. DOI: 10.1016/j.gloplacha.2016.11.013.

Spirtes, P. and C. Glymour (1991). "An algorithm for fast recovery of sparse causal graphs." In: *Social science computer review* 9.1, pp. 62–72.

Stoeber, J., H. Joe, and C. Czado (2013). "Simplified pair copula constructions—limitations and extensions." In: *Journal of Multivariate Analysis* 119, pp. 101–118.

Sun, M., I. Konstantelos, and G. Strbac (2016). "C-vine copula mixture model for clustering of residential electrical load pattern data." In: *IEEE Transactions on Power Systems* 32.3, pp. 2382–2393.

Sun, W., S. Rachev, S. V. Stoyanov, and F. J. Fabozzi (2008). "Multivariate skewed Student's t copula in the analysis of nonlinear and asymmetric dependence in the German equity market." In: *Studies in Nonlinear Dynamics & Econometrics* 12.2.

Tao, Y., Y. Wang, D. Wang, L. Ni, and J. Wu (2021). "A C-vine copula framework to predict daily water temperature in the Yangtze River." In: *Journal of Hydrology* 598, p. 126430.

Tepegjozova, M. (Nov. 2019). "D- and C-vine quantile regression for large data sets." Masterarbeit. Garching b. München: Technische Universität München.

Tepegjozova, M. and C. Czado (2022). "Bivariate vine copula based quantile regression." In: *arXiv preprint arXiv:2205.02557*.

Tepegjozova, M., B. Meyer, A. Rammig, C. Zang, and C. Czado (2023). *"Univariate and bivariate risk analysis of late-frost and drought conditions using vine copulas in Bavaria*. Tech. rep.

Tepegjozova, M., J. Zhou, G. Claeskens, and C. Czado (2022). "Nonparametric C- and D-vine-based quantile regression." In: *Dependence Modeling* 10.1, pp. 1–21. DOI: doi:10.1515/demo-2022-0100.

Torre, E., S. Marelli, P. Embrechts, and B. Sudret (2019). "A general framework for data-driven uncertainty quantification under complex input dependencies using vine copulas." In: *Probabilistic Engineering Mechanics* 55, pp. 1–16.

Tukey, J. W. (1975). "Mathematics and the picturing of data." In: *Proceedings of the International Congress of Mathematicians, Vancouver, 1975*. Vol. 2, pp. 523–531.

Udny Yule, G., M. Kendall, et al. (1950). "An introduction to the theory of statistics." In: *An introduction to the theory of statistics.* 14th ed.

Van Loon, A. F. (2015). "Hydrological Drought Explained." In: *WIREs Water* 2.4, pp. 359–392. ISSN: 2049-1948. DOI: 10.1002/wat2.1085.

Vatter, T. and T. Nagler (2018). "Generalized additive models for pair-copula constructions." In: *Journal of Computational and Graphical Statistics* 27.4, pp. 715–727.

Veraverbeke, N., M. Omelka, and I. Gijbels (2011). "Estimation of a conditional copula and association measures." In: *Scandinavian Journal of Statistics* 38.4, pp. 766–780.

Vicente-Serrano, S. M., S. Beguería, and J. I. López-Moreno (Apr. 2010). "A Multiscalar Drought Index Sensitive to Global Warming: The Standardized Precipitation Evapotranspiration Index." In: *J. Climate* 23.7, pp. 1696–1718. ISSN: 0894-8755, 1520-0442. DOI: 10.1175/2009JCLI2909.1.

Volpi, E., A. Fiori, S. Grimaldi, F. Lombardo, and D. Koutsoyiannis (2015). "One hundred years of return period: Strengths and limitations." In: *Water Resources Research* 51.10, pp. 8570–8585.

Wen, K. and X. Wu (2015). "An improved transformation-based kernel estimator of densities on the unit interval." In: *Journal of the American Statistical Association* 110.510, pp. 773–783.

Wu, T., J. Bai, and H. Han (2022). "Short-Term Agricultural Drought Prediction based on D-vine copula quantile regression in snow-free unfrozen surface area, China." In: *Geocarto International*, pp. 1–19.

Xiao, Z. and R. Koenker (2009). "Conditional quantile estimation for generalized autoregressive conditional heteroscedasticity models." In: *Journal of the American Statistical Association* 104.488, pp. 1696–1712.

Ye, F., J. Ding, K. Chen, and X. Xi (2022). "Investigation of corticomuscular functional coupling during hand movements using vine copula." In: *Brain Sciences* 12.6, p. 754.

Yeh, I.-C. (1998). "Modeling of strength of high-performance concrete using artificial neural networks." In: *Cement and Concrete research* 28.12, pp. 1797–1808.

Yousefpour, R., A. L. D. Augustynczik, C. P. O. Reyer, P. Lasch-Born, F. Suckow, and M. Hanewinkel (Jan. 2018). "Realizing Mitigation Efficiency of European Commercial Forests by Climate Smart Forestry." In: *Scientific Reports* 8.1, p. 345. ISSN: 2045-2322. DOI: 10.1038/s41598-017-18778-w.

Yu, K. and M. Jones (1998). "Local linear quantile regression." In: *Journal of the American statistical Association* 93.441, pp. 228–237.

Yu, K. and R. A. Moyeed (2001). "Bayesian quantile regression." In: *Statistics & Probability Letters* 54.4, pp. 437–447.

Zhu, K., D. Kurowicka, and G. F. Nane (2020). "Common sampling orders of regular vines with application to model selection." In: *Computational Statistics & Data Analysis* 142, p. 106811.

– (2021). "Simplified R-vine based forward regression." In: *Computational Statistics & Data Analysis* 155, p. 107091.

Zohner, C. M., L. Mo, S. S. Renner, J.-C. Svenning, Y. Vitasse, B. M. Benito, A. Ordonez, F. Baumgarten, J.-F. Bastin, V. Sebald, P. B. Reich, J. Liang, G.-J. Nabuurs, S. De-Miguel, G. Alberti, C. Antón-Fernández, R. Balazy, U.-B. Brändli, H. Y. H. Chen, C. Chisholm, E. Cienciala, S. Dayanandan, T. M. Fayle, L. Frizzera, D. Gianelle, A. M. Jagodzinski, B. Jaroszewicz, T. Jucker, S. Kepfer-Rojas, M. L. Khan, H. S. Kim, H. Korjus, V. K. Johannsen, D. Laarmann, M. Lang, T. Zawila-Niedzwiecki, P. A. Niklaus, A. Paquette, H. Pretzsch, P. Saikia, P. Schall, V. Šebeň, M. Svoboda, E. Tikhonova, H. Viana, C. Zhang, X. Zhao, and T. W. Crowther (2020). "Late-Spring Frost Risk between 1959 and 2017 Decreased in North America but Increased in Europe and Asia." In: *Proceedings of the National Academy of Sciences*, p. 201920816. ISSN: 0027-8424. DOI: 10.1073/pnas.1920816117.