Technische Universität München
TUM School of Computation, Information and Technology

Statistical learning based on vine copulas with societal applications

Özge Şahin

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

Doktorin der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitz:		Prof. Dr. Matthias Scherer
Prüfer*innen der Dissertation:	1.	Prof. Claudia Czado, Ph.D.
	2.	Prof. Dr. Hans Manner
	3.	Prof. Harry Joe, Ph.D.

Die Dissertation wurde am 23.02.2023 bei der Technische Universität München eingereicht und durch die TUM School of Computation, Information and Technology am 16.05.2023 angenommen.

Zusammenfassung

Die Literatur über statistisches Lernen ist in den letzten Jahren enorm gewachsen. Es wurden verschiedene Methoden und Algorithmen vorgeschlagen und auf reale Daten angewandt. In der Literatur besteht jedoch eine Lücke in der expliziten Modellierung komplexer Abhängigkeitsstrukturen und der gleichzeitigen Interpretation der Modellergebnisse für verwertbare Erkenntnisse. Vine Copulas sind flexible und interpretierbare multivariate Verteilungsfunktionen, die komplexe Datenabhängigkeitsstrukturen modellieren. Um diese Lücke also zu schließen, entwickeln wir statistische Lernalgorithmen, die Vine Copulas verwenden. Darüber hinaus zeigen wir ihre gesellschaftlichen Anwendungen in den Bereichen Finanzen und Genetik.

Nach einem Überblick über die theoretischen Grundlagen von Vine Copulas und Optimierung entwickeln wir überwachte und unüberwachte Lernalgorithmen unter Verwendung von Vine Copulas, insbesondere ein modellbasiertes Clustering und eine hochdimensionale dünnbesetze Regression. Anschließend diskutieren wir die Modellauswahl, die Parameterschätzung und Komplexitätsprobleme mit effizienten Lösungen. Danach vergleichen wir unsere Algorithmen mit den in der Literatur vorhandenen Ansätzen und demonstrieren die Nützlichkeit von Vine Copulas für verschiedene statistische Lernaufgaben.

Wir präsentieren eine Fallstudie zum modellbasierten Clustering mit Vine Copulas bei der Analyse von Finanzdaten unter Verwendung sogenannter ESG Scores (Environmental, Social und Governance). Darüber hinaus diskutieren wir die Nachteile der ESG-Scores und schlagen eine neue Technik zur Behandlung fehlender Daten vor. Schließlich machen wir als Anwendung der hochdimensionalen dünnbesetzen Vine Copula basierten Regression den ersten Schritt in der genomischen Vorhersage von phänotypischen Merkmalen von Mais. Nach der Entwicklung von Methoden zur Extraktion genomischer Merkmale zeigen wir die Vorteile unserer Methoden gegenüber bestehenden Ansätzen.

Abstract

Statistical learning literature has been tremendously growing in recent years. Various methods and algorithms have been proposed and applied to real data. However, there is a gap in the literature that modeling complex dependence structures explicitly and interpreting model results for actionable insights simultaneously have not yet been handled well. *Vine copulas* are flexible and interpretable multivariate distribution functions that model complex data dependence structures. Thus, to bridge the gap, we develop statistical learning algorithms using vine copulas. Moreover, we show some societal applications of the proposed models in finance and genetics.

After reviewing the theoretical foundations of vine copulas and optimization, we develop supervised and unsupervised learning algorithms using vine copulas, mainly a model-based clustering and a high-dimensional sparse regression. Next, we discuss model selection, parameter estimation, and complexity problems with efficient solutions. Later, we compare our algorithms with the existing approaches in the literature and demonstrate the usefulness of vine copulas for various statistical learning tasks.

We present a case study of model-based clustering with vine copulas in financial data analyses, using so-called Environmental, Social, and Governance (ESG) scores. We further discuss the pitfalls of the ESG scores and propose a new missing data handling technique. Finally, as an application of high-dimensional sparse vine copula based regression, we take the first step in the genomic prediction of maize traits. After developing genomic feature extraction methods, we show the advantage of our methods over existing approaches.

Acknowledgments

Many people made my time fun, educational, and, most importantly, a unique experience.

I owe a debt of gratitude to my advisor Claudia Czado, who was also my advisor for my master's thesis, whose many lectures I enjoyed attending, and who always supported my freedom, which is my most indispensable quality. I have developed myself and learned much from you in every meeting, daily conversation, and social event. I am glad that you offered me a position many years ago.

One of the highlights of my doctorate was the opportunity to work with Sandra Paterlini and Karoline Bax. It was a pleasure to study financial applications with you, learn new concepts, and discuss many different ideas. Thank you for all.

My other valuable collaboration was with Harry Joe and his research group at the University of British Columbia. He welcomed me to Vancouver with great warmth and sincerity. I enjoyed listening to him and being inspired by his ideas and experiences. Thank you very much for hosting me and being a member of my thesis committee.

I am also grateful to Hans Manner for being on my thesis committee and sharing his valuable insights. I would also like to thank Chris-Carolin Schön and Munich Data Science Institute for supporting our project through Seed funds. I would also like to thank my former colleague Feyyaz and my professor Yaman Barlas, with whom I continued to work for a while after my bachelor studies and who guided my initial research experience.

Other thanks also to my colleagues, especially Ariane and Marija, who make my office time fun. Our talks about everything, including and beyond the research, are special.

I am grateful to the Global Challenges for Women in Math Science program of the Technical University of Munich for the research grant. I am also grateful to German Research Foundation (DFG) for funding the doctorate project (DFG grant CZ 86/6-1).

Big love and thanks to my dearest Emre, whose love and support I always feel. Special thanks to my dear elders Nevriye and Şenol Türkoğlu, with whom I felt at home in Munich. I would also like to thank my dear friend Ekin, who witnessed many moments of my doctorate time despite the distance between us. Finally, I am grateful to my dear friends Begüm, Cemre, Gizem, Gökçe, Gülbükre, Nazlican, Öykü, and Sofia, whom I feel the same as before every time we talk, no matter where we are in the world.

Lastly, I would like to thank my dear mother, father, and brother, who support me in every decision I make. If I have contributed to science, it is all with your love and open-mindedness.

Contents

Chapter 1

Introduction

As data dimensionality increases and data analysis improves the decision-making process, the need for models to learn from data is huge. Data consists of many observations and features. For example, consider the data collection process of companies' sustainability levels. A company's sustainability level corresponds to a score for its responsibility for environmental, social, and governmental issues. Thus, a company is an observation having values for an issue being a feature. Alternatively, consider the data collection process of songs listened to by people on an online platform. A person is an observation, and what they listen to has features such as danceability and instrumentality. From observations and features in data, useful conclusions can be drawn. For instance, identifying groups of people based on their musical taste improves music recommendation systems. Alternatively, consider portfolio risk management. If which companies depend on each other regarding their sustainability levels is known, sustainable portfolio risks can be predicted and managed. What is common in these examples from diverse areas is statistical learning, that is, the process of learning from data and making inferences (Hastie et al. 2009).

Supervised learning is conducted when data is in the main charge of the learning process, guiding the statistical model. One of the core elements of supervised learning is to make predictive inferences, e.g., a regression. It has a variety of applications: predicting a trait from genomes (Li et al. 2018), a company's return values from its peers (Kraus and Czado 2017), and an object class from its properties (Nguyen and Wu 2011). On the other hand, unsupervised learning comes into play when the aim is to extract knowledge from data without any advance knowledge. For example, a task of revealing groups of observations using features in data, also known as clustering, belongs to unsupervised learning. Its real-life applications include identifying groups of proteins resulting in disease (Zhang and Shi 2017), groups of periods with different asset prices (Weiß and Scheffer 2015), and estimating stations' different dependence structures based on the different amount of precipitations (Kim et al. 2013).

To evaluate a performance of a supervised learning algorithm, data can be partitioned into two sets: training and test. A training set is used to make inferences, whereas a test set helps assess the models' performance in unseen observations. Since data supervises the model in

supervised learning, using all observations to make inferences may provide an excellent fit for the observed data. However, it does not guarantee that the model still performs very well for new observations that are needed to be predicted. This phenomenon is known as overfitting. Thus, data partition prevents overfitting.

On the other hand, we might not have any labels or supervision in data to assess how well an unsupervised learning algorithm performs. Therefore, statistical model criteria, such as the Bayesian (Schwarz 1978) or the Akaike (Akaike 1998) information criterion, are widely used (Scrucca et al. 2016) when applicable.

Current supervised learning approaches for regression include linear mean/quantile regression and regression trees. The advantages of linear models are that they are easy to fit and interpret results. However, they do not explicitly consider feature interactions or skewness (Hastie et al. 2009). Moreover, even though regression tree and forest approaches are nonlinear models, one of their drawbacks is to suffer from dependent features (Strobl et al. 2008).

Likewise, popular unsupervised learning models for clustering are distance-based, like k-means, hierarchical clustering, and model-based, like Gaussian mixture models (Hastie et al. 2009). However, their performance differs by analyzed data (Hennig 2022). While distance-based methods are computationally efficient, they do not consider the dependence structure in data. Similarly, mixture models usually assume that each component follows the same statistical distribution family, reducing the model flexibility (McLachlan and Peel 2000).

These are critical issues since, in real-life problems, data will likely be high-dimensional, nonlinear, skewed, and dependent. Moreover, the models mentioned above might not be adequate when the focus is to model dependence structures explicitly and have actionable insights. Even though they may achieve good prediction or clustering results, the interpretation and reasoning behind the performance are not yet well understood. For example, consider portfolio risk management again. To predict the likelihood of a market crash and understand the dependence among portfolio assets, explicit dependence modeling, which can work with dependent features, here assets, is needed. Alternatively, consider finding people groups when they listen to pop or rock songs. A clustering model capturing complex dependencies is needed since different dependencies exist between the pop or rock songs' danceability and instrumentality. Here the use of copulas comes in handy.

Copulas are powerful dependence modeling tools, explicitly capturing complex dependence structures between multiple random variables. They go beyond correlation, a linear dependence measure. Copulas glue variables' marginal distributions and standardized dependence structure and form their joint distribution (Sklar 1959). They provide a full distribution of variables of interest and can also help assess the probability of extreme events.

Despite copulas' power in modeling data analyses, standard copula families, such as Gaussian, t, and Archimedean, do not express different dependence structures between different pairs of variables. However, pair-copula constructions extend their usage, applying sequentially mixing conditional distributions involving a set of bivariate copulas (Aas et al. 2009;

Joe 1996). The set of bivariate copulas is usually applied to pairs of univariate conditional distributions. Since such a construction is not unique, an undirected graph structure, called vines, is used to organize them (Bedford and Cooke 2002). Accordingly, vine copulas provide flexible and nonlinear models in high dimensions.

In this thesis, we answer how we can model nonlinear dependence in supervised and unsupervised learning while explicitly modeling relationships between data features. Moreover, we provide how dependence modeling in supervised and unsupervised learning may aid decision-making processes in real-life applications. Accordingly, we will delve into *using vine copulas for statistical learning*, developing new methods and algorithms for their implementation. First, we will discuss vine copulas' main statistical learning applications, such as clustering and prediction. We will then present real-life case studies applying our methods and algorithms.

The thesis consists of two parts: the first part presents the development of statistical learning algorithms based on vine copulas, and the main content of Chapters 3 and 4 are based on the following research papers.

- Sahin, Ö., & Czado, C. (2022). Vine copula mixture models and clustering for non-Gaussian data. *Econometrics and Statistics, 22*, 136-158.
- Sahin, Ö., & Czado, C. (2022). High-dimensional sparse vine copula regression with application to genomic prediction. *In revision for Biometrics, arXiv:2208.12383*.

In the second part, we focus on the societal applications of the algorithms we developed. The following research papers form the main content of Chapters 5 and 6.

- Sahin, Ö., Bax, K., Czado, C., & Paterlini, S. (2022). Environmental, Social, Governance scores and the Missing pillar—Why does missing information matter?. *Corporate Social Responsibility and Environmental Management, 29(5)*, 1782-1798.
- Sahin, Ö., Bax, K., Paterlini, S., & Czado, C. (2023). The pitfalls of (non-definitive) Environmental, Social, and Governance scoring methodology. *Global Finance Journal, 56*, 100780.
- Sahin, Ö., & Czado, C. (2022). High-dimensional sparse vine copula regression with application to genomic prediction. *In revision for Biometrics, arXiv:2208.12383*.

We discuss the content of each paper and provide additional data analyses, materials, and discussion points.

We review the theoretical foundations of copulas, vine copulas, and optimization in Chapter 2. It provides main concepts, including copula construction methods and bivariate copulas with their density and conditional distribution. It also preliminarily shows the core ideas behind numerical optimization.

In Chapter 3, based on Sahin and Czado (2022b), we specify a vine copula mixture model formulation and use the model for clustering tasks. We treat the number of mixture

components fixed and provide efficient solutions for model selection and parameter estimation problems, including vine tree and pair-copula family selections. We compare its performance with existing methods using simulation studies and real-world data. In addition, we provide its software implementation, making a big step forward in developing explainable and effective clustering algorithms for complex data.

We will then focus on selecting features for conditional quantile estimates, i.e., quantile regression, in high dimensions using vine copulas in Chapter 4 based on Sahin and Czado (2022a). High-dimensional data likely includes irrelevant features for the outcome of interest and dependent features with each other. Thus, not all features are needed for prediction. We define and identify features for quantile regression and propose new feature selection approaches for vine copula based regression, which are shown to be computationally efficient. We compare their performance with linear and nonlinear models based on random forests using training and test data sets.

As data collection increases in various fields, the pitfalls of collection processes and data handling increase too. Chapter 5 focuses on one of the core data sets in sustainable finance: Environmental, Social, and Governance (ESG) scores. They show how responsible a company is compared to its peers in sustainability matters. As a case study of vine copula mixture models, we quantify different dependence structures between good and bad score indices depending on the stock price movements. More importantly, we discuss how data handling, including missing scores, results in different outcomes reported in Sahin et al. (2023). We propose a new method to account for missing scores in data sets and show its performance in training and test data sets reported in Sahin et al. (2022).

Another rapid advancement in data generation and collection is in the field of genomics. Lately, genotype data has been getting high-dimensional, and predicting complex phenotypes from genotypes has been the main goal for various purposes, including adopting maize genotypes for future breeding. In Chapter 6, based on Sahin and Czado (2022a), we apply our prediction algorithm for high-dimensional data using vine copulas to the genomic prediction of maize traits. Since genotype data consists of binary features, we propose new feature extraction methods. We demonstrate significant improvements in the accuracy and efficiency of our genomic prediction results compared to the existing approaches. Our work presents how modeling complex dependencies in genomic prediction by vine copulas guides new insights into our understanding of genotype and phenotype relationships.

Finally, we overview and conclude our findings in Chapter 7 and provide further research directions that may increase the potential of vine copulas for handling complex and high-dimensional data for various statistical learning tasks.

Part I

Statistical learning with vine copulas

Chapter 2

Preliminaries

In this chapter, we will briefly review the main concepts in copulas, vine copulas, and numerical optimization. We will mainly follow the reference books by Wright and Nocedal (2006), Joe (2014), and Czado (2019).

2.1 Copulas

As large amounts of data have been collected, the need for statistical models to make inferences about data quantities, such as group densities, tail probabilities, and conditional expectations, has increased. Especially the models that can accommodate various dependence structures and tail behavior are highly needed; for example, copulas.

Copulas are a class of multivariate non-normal distributions and a powerful tool in dependence modeling. For continuous multivariate distributions, after univariate margins, which can be of different types, are modeled, a copula represents the dependence structure separated from such margins. Such a representation results from Sklar's theorem (Sklar 1959). Formally, a d-dimensional copula C is a multivariate distribution function $C : [0,1]^d \to [0,1]$ with $U(0,1)$ margins and is given by $C(u_1, \ldots, u_d) = P(U_1 \leq u_1, \ldots, U_d \leq u_d)$, where $U_1, U_2, \ldots, U_d$ are uniformly distributed in the interval $[0,1]$. In the remainder of the chapter, we will use these notations.

Since the d-dimensional copula input (data) is uniformly distributed in $[0,1]^d$, to get it, we apply *the probability integral transform*, i.e., if the continuous random variable $X \sim F_X$, then $F_X(X) \sim U(0,1)$. To see that, it holds for $0 < u < 1$, $P(F_X(X) \leq u) = P(X \leq F_X^{-1}(u)) = F_X(F_X^{-1}(u)) = u$, where F_X is continuous.

One can use univariate margins from parametric families with one, two, or more parameters with different modalities, tail weight, and asymmetry properties to apply the probability integral transform. Some parametric univariate families are detailed in Table 2.1, and their densities are shown in Figure 2.1. For more continuous univariate distribution families and their definitions, we refer to Appendix A in Klugman et al. (2019).

Theorem 2.1: Sklar's Theorem

Let $\boldsymbol{X} = (X_1, \ldots, X_d)^\top \in \mathbb{R}^d$ be a d-dimensional random vector following a joint distribution F with the univariate marginal distributions $F_1, \ldots F_d$, then the copula associated with F is a d-dimensional distribution function $C : [0,1]^d \to [0,1]$ with $U(0,1)$ margins and satisfies

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

In the remainder, assume all random variables to be absolutely continuous. Then the copula corresponding to F with quantile functions F_p^{-1} for $p = 1, \ldots, d$ is unique:

$$C(u_1, \ldots, u_d) = F(F_1^{-1}(u_1), \ldots, F_d^{-1}(u_d)). \tag{2.2}$$

Moreover, the d dimensional joint density f can be written as

$$f(\boldsymbol{x}) = c\big(F_1(x_1), \ldots, F_d(x_d)\big) \cdot f_1(x_1) \cdots f_d(x_d), \quad \boldsymbol{x} \in \mathbb{R}^d, \tag{2.3}$$

where $f_1, \ldots, f_d$ are univariate marginal distributions associated with $F_1, \ldots F_d$, and c is a copula density of the the random vector $\boldsymbol{F} = \big(F_1(X_1), \ldots F_d(X_d)\big)^\top \in [0,1]^d$.

Proof. See Theorem 1.1 of Joe (2014).

Example 2.1 *(Univariate normal distribution). A unimodal and symmetric distribution with the support $(-\infty, \infty)$ is the univariate normal distribution. If a random variable $X \in \mathbb{R}$ follows the univariate normal distribution, its density function in x is*

$$f\left(x; \mu, \sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right),$$

where $\mu \in \mathbb{R}$ is the mean, and $\sigma \in \mathbb{R}^+$ is the variance, denoting $X \sim \mathcal{N}\left(\mu, \sigma^2\right)$. Its cumulative distribution function (cdf) is

$$F(x; \mu, \sigma^2) = \frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{x-\mu}{\sqrt{2}\sigma}\right)\right],$$

where $\operatorname{erf}(x)$ is the error function defined as

$$\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x \exp\left(-t^2\right) \mathrm{d}t.$$

Example 2.2 *(Univariate Student's t distribution). Another unimodal and symmetric distribution with the support $(-\infty, \infty)$ is the univariate Student's t distribution. Compared to the univariate normal distribution, the t distribution has heavy tails, changing the measure*

based on its degrees of freedom parameter. If a random variable $X \in \mathbb{R}$ follows the univariate Student's t distribution, its density in x is

$$f(x;\mu,\nu,\sigma) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)\sigma^{-1}}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\nu\pi}}\left(1+\frac{(x-\mu)^2}{\nu\sigma^2}\right)^{-(\nu+1)/2},$$

where μ is the location, $\nu \in \mathbb{R}^+$ is the degrees of freedom, $\sigma \in \mathbb{R}^+$ is the scale, and Γ is the gamma function, denoting $X \sim t(\mu,\nu,\sigma)$. Its cumulative distribution function is

$$F(x;\mu,\nu,\sigma) = \int_{-\infty}^{x} f(t;\mu,\nu,\sigma)\mathrm{d}t.$$

Example 2.3 *(Logistic distribution). The logistic distribution is also unimodal and symmetric with the support $(-\infty,\infty)$. Like the Student's t distribution, compared to the univariate normal distribution, it has heavy tails but independent of its parameters (Pingel 2014). If a random variable $X \in \mathbb{R}$ follows a logistic distribution, its density in x is*

$$f(x;\mu,s) = \frac{e^{-(x-\mu)/s}}{s\left(1+e^{-(x-\mu)/s}\right)^2}$$

where $\mu \in \mathbb{R}$ is the mean, $s \in \mathbb{R}^+$, is the scale denoting $X \sim logis(\mu,s)$. Its cumulative distribution function is

$$F(x;\mu,s) = \frac{1}{1+e^{-\frac{x-\mu}{s}}}.$$

Example 2.4 *(Univariate skew t distribution). The univariate skew t distribution is unimodal with the support $(-\infty,\infty)$ but asymmetric. Compared to the univariate normal distribution, it has heavy tails. According to Fernández and Steel (1998), if a random variable $X \in \mathbb{R}$ follows the skew t distribution, its density in x is*

$$\begin{aligned} f(x;\mu,\sigma,\nu,\gamma) =& \frac{\Gamma\left(\frac{\nu+1}{2}\right)\sigma^{-1}}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\pi\nu}}\frac{2}{\gamma+\frac{1}{\gamma}} \\ & \cdot\left[1+\frac{(x-\mu)^2}{\nu\sigma^2}\left\{\gamma^2\mathbb{I}_{(-\infty,0)}(x-\mu)+\frac{1}{\gamma^2}\mathbb{I}_{[0,\infty)}(x-\mu)\right\}\right]^{-(\nu+1)/2}, \end{aligned}$$

where $\mu \in \mathbb{R}$ is the location, $\sigma \in \mathbb{R}^+$ is the scale, $\nu \in \mathbb{R}^+$ is the degrees of freedom, and $\gamma \in \mathbb{R}^+$ is the skewness, denoting $X \sim st(\mu,\sigma,\nu,\gamma)$. Its cumulative distribution function is

$$F(x;\mu,\sigma,\nu,\gamma) = \int_{-\infty}^{x} f(t;\mu,\sigma,\nu,\gamma)\mathrm{d}t.$$

We remark that there are other definitions of the skew t distribution as given in Azzalini and Capitanio (2003) and Jones and Faddy (2003).

Example 2.5 *(Exponential distribution). A unimodal but asymmetric distribution with the support* $[0, \infty)$ *is the exponential distribution. It has exponential tails. If a random variable* $X \in \mathbb{R}$ *follows the exponential distribution, its density in* x *is*

$$f(x; \lambda) = \lambda \exp(-\lambda x),$$

where $\alpha \in \mathbb{R}^+$ *is the rate, denoting* $X \sim \exp(\lambda)$*. Its cumulative distribution function is*

$$F(x; \lambda) = 1 - \exp(-\lambda x).$$

Example 2.6 *(Gamma distribution). Another unimodal but asymmetric distribution with the support* $(0, \infty)$ *is the gamma distribution. It has exponential tails. If a random variable* $X \in \mathbb{R}$ *follows the gamma distribution, its density in* x *is*

$$f(x; \alpha, \beta) = \frac{(x\beta)^\alpha \exp(-x\beta)}{x\Gamma(\alpha)},$$

where $\alpha \in \mathbb{R}^+$ *is the shape, and* $\beta \in \mathbb{R}^+$ *is the rate, denoting* $X \sim \Gamma(\alpha, \beta)$*.* Γ *is the gamma function defined by*

$$\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} \exp(-t) \mathrm{d}t.$$

The incomplete gamma function is defined by

$$\Gamma(\alpha; x) = \frac{1}{\Gamma(\alpha)} \int_0^x t^{\alpha-1} \exp(-t) \mathrm{d}t.$$

The cumulative distribution function is

$$F(x; \alpha, \beta) = \Gamma(\alpha; x\beta).$$

Example 2.7 *(Log-normal distribution). Like the gamma distribution, the log-normal distribution is also unimodal and asymmetric with the support* $(0, \infty)$*. However, it has subexponential tails. A random variable* $X \in \mathbb{R}$ *follows a lognormal distribution if* $Y \sim \mathcal{N}(\mu, \sigma^2)$ *for* $Y = \log X$*. Accordingly, the density of* X *in* x *is*

$$f(x; \mu, \sigma) = \frac{1}{x\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\log x - \mu)^2}{2\sigma^2}\right).$$

We denote $X \sim lnorm(\mu, \sigma^2)$*. Its cumulative distribution function is*

$$F(x; \mu, \sigma) = \Phi\left(\frac{\log x - \mu}{\sigma}\right),$$

where $\Phi(\cdot)$ *denotes the univariate normal distribution function with zero mean and unit variance.*

Example 2.8 *(Log-logistic distribution). Another asymmetric distribution with the support* $(0, \infty)$ *is the log-logistic distribution. However, it has heavier tails than the univariate log-normal distribution (Lemonte 2014) and is unimodal as long as its shape parameter is bigger than one. If a random variable* $X \in \mathbb{R}$ *follows a log-logistic distribution, the density of* X *in* x *is*

$$f(x; \alpha, \beta) = \frac{\beta \left(x/\alpha\right)^{\beta}}{x \left(1 + \left(x/\alpha\right)^{\beta}\right)^{2}},$$

where $\alpha \in \mathbb{R}^{+}$ *is the scale, and* $\beta \in \mathbb{R}^{+}$ *is the shape, denoting* $X \sim llogis\left(\alpha, \beta\right)$*. Its cumulative distribution function is*

$$F(x; \alpha, \beta) = \frac{\left(x/\alpha\right)^{\beta}}{1 + \left(x/\alpha\right)^{\beta}}.$$

Table 2.1: Parametric univariate families with the first (mean) and second (variance) cumulants.

	Mean ($\mathbb{E}(X)$)	Variance (Var(X))
$X \sim \mathcal{N}\left(\mu, \sigma^2\right)$	μ	σ^2
$X \sim t\left(\mu, \nu, \sigma\right)$	μ for $\nu > 1$	$\sigma^2\nu/(\nu - 2)$, for $\nu > 2$
$X \sim logis\left(\mu, s\right)$	μ	$s^2\pi^2/3$
$X \sim st\left(\mu, \sigma, \nu, \gamma\right)$	$M_1\left(\gamma - \frac{1}{\gamma}\right)$	$M_2(\gamma^2 - 1 + \gamma^{-2}) - M_1^2(\gamma - \frac{1}{\gamma})^2$ [a]
$X \sim \exp\left(\lambda\right)$	$1/\lambda$	$1/\lambda^2$
$X \sim \Gamma\left(\alpha, \beta\right)$	α/β	α/β^2
$X \sim lnorm\left(\mu, \sigma^2\right)$	$\exp(\mu + \sigma^2/2)$	$(\exp(\sigma^2) - 1)\exp(2\mu + \sigma^2)$
$X \sim llogis\left(\alpha, \beta\right)$	$\frac{\alpha\pi/\beta}{sin\ \pi/\beta}$, for $\beta > 1$	$\alpha^2\left(\frac{2\pi/\beta}{sin\ 2\pi/\beta} - \frac{\pi^2/\beta^2}{\sin^2 \pi/\beta}\right)$, for $\beta > 2$

[a] $M_c = 2\int_0^{\infty} s^c f(s)\mathrm{d}s = \mathbb{E}(|s|^c)$, and $f(s)$ is the corresponding density of the univariate Student's t distribution being symmetric about zero.

Alternatively, a non-parametric approach, such as one based on kernels (Parzen 1962), can be applied to use the probability integral transform. An example is shown in Figure 2.2.

Example 2.9 *(Kernel smoother). Let* $(x_1, \ldots, x_n)$ *be independent continuous observations of the random variable* X*. If X follows a distribution with an unknown density* f*, its kernel density estimate in x is*

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{x - x_i}{h}\right),$$

where $k(.)$ *is a symmetric probability density function like a univariate Gaussian distribution, and* $h > 0$ *is a bandwidth parameter that can be chosen by a plug-in method as proposed in*

Sheather and Jones (1991). Its cumulative distribution function estimate is

$$\hat{F}(x) = \int_{-\infty}^{x} \hat{f}(t)\mathrm{d}t.$$

Figure 2.1: Densities of the univariate families listed in Table 2.1, where $\mathbb{E}(X) = 1$ and $\mathrm{Var}(X) = 1$, $\mathbb{E}(X) = 2$ and $\mathrm{Var}(X) = 4$, and $\mathbb{E}(X) = 2$ and $\mathrm{Var}(X) = 1$ are fixed for red, orange, and black curves, respectively, if relevant. The degrees of freedom for the t distributions are 3, and the skewness parameter is 2.

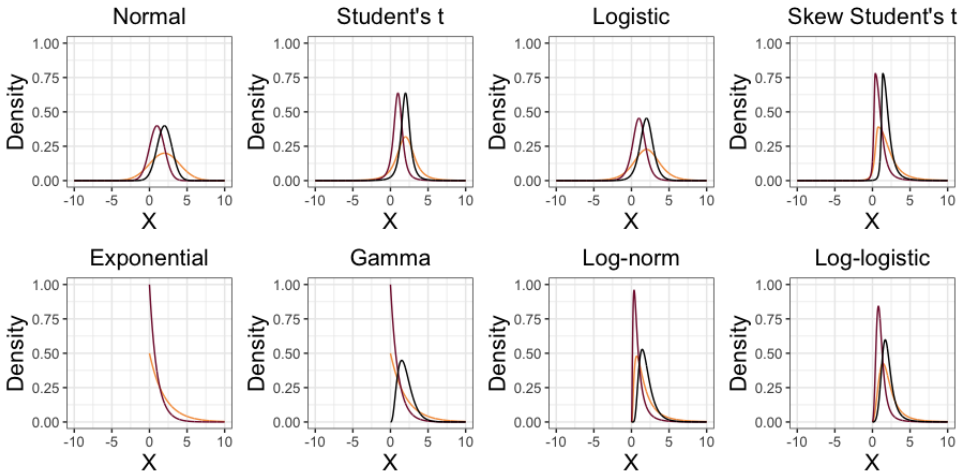
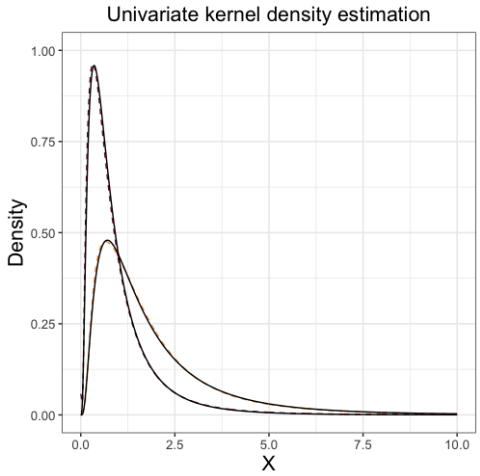


Figure 2.2: True densities (black lines) and estimated univariate kernel densities for 1000 observations (dashed) from the log-normal distribution with $\mathbb{E}(X) = 2$, $\mathrm{Var}(X) = 4$ (orange) and $\mathbb{E}(X) = 2$, $\mathrm{Var}(X) = 1$ (red). The kernel is Gaussian with the bandwidth 0.6.



Even though one can construct the copula data, i.e., u-data, by applying the probability integral transform, the important question is how to build suitable copulas. There are different copula construction methods (see Chapter 3 of Joe (2014)); however, simple constructions may not result in flexible dependence in high dimensions. Still, there are alternative and powerful approaches in dependence modeling, such as pair-copula constructions discussed in Section 2.2, that use a set of bivariate copulas to construct multivariate copulas. Thus, we first focus on some bivariate parametric copula families and their construction.

Bivariate copula construction allows the modeling of various dependence structures like asymmetric dependence, conditional independence, and tail dependence between two random variables. The tail dependence measures the strength of the dependence in the joint lower or joint upper tail of a bivariate copula. For instance, for a strong upper tail dependence between random variables X_1 and X_2, one can expect that when the value of X_1 is very large, very likely that the value of X_2 is very large too.

Definition 2.1: Upper and lower tail dependence coefficients

Suppose $(U_1, U_2) \sim C$, where C is a bivariate copula. Then the upper tail dependence coefficient of C is

$$\lambda^{upper} = \lim_{u \to 1^-} P(U_2 > u | U_1 > u) = \lim_{u \to 1^-} \frac{1 - 2u + C(u, u)}{1 - u}.$$

The lower tail dependence coefficient of C is

$$\lambda^{lower} = \lim_{u \to 0^+} P(U_2 \leq u | U_1 \leq u) = \lim_{u \to 0^+} \frac{C(u, u)}{u}.$$

We remark that the limits prevent us from having an empirical estimate of tail dependence coefficients. However, as Krupskii and Joe (2015) proposed, alternative tail-weighted dependence measures can be useful for data analyses. Moreover, a non-empirical estimate of tail dependence is proposed by Lee et al. (2018).

Our first bivariate copula example is independence, where margins are independent. Next, one can easily construct bivariate Gaussian copula and bivariate Student's t copula using the inverse of Sklar's Theorem in Equation (2.2). Both capture a wide range of dependence and have the closure property under marginalization, as well as a closed-form density. Nevertheless, the former cannot model tail dependence, whereas the latter lacks modeling tail asymmetries as shown in Figure 2.3.

Example 2.10 *(Bivariate independence copula). Assuming random variables X_1, X_2 are independent, i.e., $F(x_1, x_2) = F(x_1)F(x_2)$, Equation (2.1) yields the bivariate independence copula as*

$$C(u_1, u_2) = u_1 u_2, \quad 0 < u_1, u_2 < 1.$$

Its conditional distribution function and density are

$$C_{2|1}(u_2|u_1) = u_2, \quad c(u_1, u_2) = 1, \quad 0 < u_1, u_2 < 1.$$

Its upper and lower tail dependence coefficients are given by

$$\lambda^{upper} = 0, \qquad \lambda^{lower} = 0.$$

Example 2.11 *(Bivariate Gaussian copula). Let $\Phi(\cdot)$ denote the univariate normal distribution function with zero mean and unit variance and let $\Phi_2(\cdot, \cdot; \rho)$ the bivariate normal distribution function, where the random vector has the zero mean vector and unit variances, and the correlation among the variables is ρ. Likewise, let ϕ and ϕ_2 be the corresponding densities. Then Equation (2.2) yields the bivariate Gaussian copula as*

$$C(u_1, u_2; \rho) = \Phi_2(\Phi^{-1}(u_1), \Phi^{-1}(u_2); \rho), \quad 0 < u_1, u_2 < 1.$$

Its conditional distribution function and density are

$$C_{2|1}(u_2|u_1; \rho) = \Phi\left(\frac{\Phi^{-1}(u_2) - \rho\Phi^{-1}(u_1)}{\sqrt{1-\rho^2}}\right),$$

$$c(u_1, u_2; \rho) = \frac{\phi_2\left(\Phi^{-1}(u_1), \Phi^{-1}(u_2); \rho\right)}{\phi\left(\Phi^{-1}(u_1)\right)\phi\left(\Phi^{-1}(u_2)\right)}, \quad 0 < u_1, u_2 < 1.$$

Its upper and lower tail dependence coefficients are given by

$$\lambda^{upper} = 0, \qquad \lambda^{lower} = 0.$$

Example 2.12 *(Bivariate Student's t copula). Let T_ν denote the univariate Student's t distribution function with the degrees of freedom ν and let $T_{2,\nu}$ the bivariate Student's t distribution function, where the random vector has unit variances, and the correlation among the variables is ρ. Likewise, let t_ν and $t_{2,\nu}$ be the corresponding densities. Then Equation (2.2) yields the bivariate Student's t copula as*

$$C(u_1, u_2; \nu, \rho) = T_{2,\nu}(T_\nu^{-1}(u_1), T_\nu^{-1}(u_2); \nu, \rho), \quad 0 < u_1, u_2 < 1.$$

Its conditional distribution function and density are

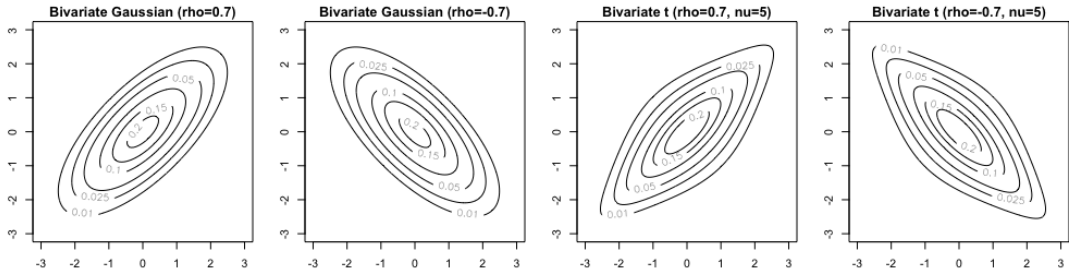
$$C_{2|1}(u_2|u_1; \rho, \nu) = T_{\nu+1}\left(\frac{T_\nu^{-1}(u_2) - \rho T_\nu^{-1}(u_1)}{\sqrt{(1-\rho^2)(\nu + [T_\nu^{-1}(u_1)]^2)/\nu + 1}}\right),$$

$$c(u_1, u_2; \nu, \rho) = \frac{t_{2,\nu}\left(T_\nu^{-1}(u_1), T_\nu^{-1}(u_2); \rho\right)}{t_\nu\left(T_\nu^{-1}(u_1)\right) t_\nu\left(T_\nu^{-1}(u_2)\right)}, \quad 0 < u_1, u_2 < 1.$$

Its upper and lower tail dependence coefficients are given by

$$\lambda^{upper} = \lambda^{lower} = 2T_{v+1}\left(-\sqrt{v+1}\sqrt{\frac{1-\rho}{1+\rho}}\right).$$

Figure 2.3: Marginally normalized contour plots of bivariate Gaussian and Student's t copulas based on van der Waerden scores (Waerden 1953), i.e., x-axis and y-axis correspond to $\Phi^{-1}(U_1)$ and $\Phi^{-1}(U_2)$, respectively, in each plot.



Another (bivariate) copula construction method is to use generator functions, resulting in the class of Archimedean copulas (for more details, see Chapter 4 of Nelsen (2007)).

Definition 2.2: Bivariate Archimedean copulas

Let φ be a generator being continuous, convex, and strictly decreasing $\varphi : [0,1] \to [0,\infty)$ with $\varphi(0) = \infty$, $\varphi(1) = 0$. Then the bivariate Archimedean copula is

$$C_\varphi(u_1, u_2) = \varphi^{-1}(\varphi(u_1) + \varphi(u_2)), \quad 0 \leq u_1, u_2 \leq 1,$$

where $\varphi^{-1} : [0,\infty) \to [0,1]$ is the pseudo-inverse of φ given by

$$\varphi^{-1}(t) = \begin{cases} \varphi^{-1}(t) & , 0 \leq t \leq \varphi(0), \\ 0 & , \varphi(0) \leq t < \infty. \end{cases}$$

By choosing a suitable generator, bivariate Archimedean copulas can be constructed, such as Clayton, Gumbel, Frank, Joe, BB1, BB6, BB7, and BB8. Like the bivariate Gaussian and t copulas, they have a closed-form density. Moreover, the bivariate Archimedean copulas can model tail asymmetries. However, some of them, such as Clayton, Gumbel, Joe, BB1, BB6, BB7, and BB8, copulas can only capture a positive dependence in contrast to the bivariate Gaussian and t copulas. Still, their range of dependence can be extended by using their their densities' counterclockwise rotations for $0 < u_1, u_2 < 1$ (Czado 2019):

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

The marginally normalized contour plots of the bivariate Archimedean copulas and bivariate Archimedean copulas' $90°$ rotations are shown in Figures 2.4 and 2.5, respectively. We remark that we do not refer to rotations of the random vector (U_1, U_2) but those of copula densities (Czado 2019).

Example 2.13 *(Bivariate Clayton copula). The generator* $\varphi(x;\delta) = \frac{1}{\delta}(x^{-\delta} - 1)$ *gives the bivariate Clayton copula for* $\delta \in [0, \infty)$*. Its cdf is*

$$C(u_1, u_2; \delta) = (u_1^{-\delta} + u_2^{-\delta} - 1)^{-1/\delta}, \quad 0 \leq u_1, u_2 \leq 1.$$

Its conditional distribution function is

$$C_{2|1}(u_2|u_1;\delta) = \left[1 + u_1^{\delta}\left(u_2^{-\delta} - 1\right)\right]^{-1-1/\delta}.$$

Its density is

$$c(u_1, u_2; \delta) = (1+\delta)[u_1 u_2]^{-\delta-1}\left(u_1^{-\delta} + u_2^{-\delta} - 1\right)^{-2-1/\delta}, \quad 0 < u, v < 1.$$

Its upper and lower tail dependence coefficients are given by

$$\lambda^{upper} = 0, \qquad \lambda^{lower} = 2^{-1/\delta}.$$

Example 2.14 *(Bivariate Gumbel copula). For* $\delta \in [1, \infty]$*, the generator* $\varphi(x;\delta) = (-\log x)^{\delta}$ *yields the bivariate Gumbel copula. Its cdf is*

$$C(u_1, u_2; \delta) = \exp\left\{-\left[(-\log u_1)^{\delta} + (-\log u_2)^{\delta}\right]^{1/\delta}\right\}, \quad 0 \leq u_1, u_2 \leq 1.$$

Let $x = -\log u$*,* $y = -\log v$*. Then its conditional distribution function is*

$$C_{2|1}(u_2|u_1;\delta) = u_1^{-1}\exp\left[-\left(x^{\delta} + y^{\delta}\right)^{1/\delta}\right]\left[1 + (y/x)^{\delta}\right]^{1/\delta - 1}.$$

Its density for $0 < u, v < 1$ *is*

$$c(u_1, u_2; \delta) = \exp\left[-\left(x^{\delta} + y^{\delta}\right)^{1/\delta}\right]\left[\left(x^{\delta} + y^{\delta}\right)^{1/\delta} + \delta - 1\right]\left[x^{\delta} + y^{\delta}\right]^{1/\delta - 2}(xy)^{\delta-1}(u_1 u_2)^{-1}.$$

Its upper and lower tail dependence coefficients are given by

$$\lambda^{upper} = 2 - 2^{1/\delta}, \qquad \lambda^{lower} = 0.$$

Example 2.15 *(Bivariate Frank copula). The generator* $\varphi(x;\delta) = -\log\left({}^{(e^{-\delta x}-1)}/_{(e^{-\delta}-1)}\right)$ *yields the bivariate Frank copula for* $\delta \in \mathbb{R}/\{0\}$*. Its cdf is*

$$C(u_1, u_2; \delta) = -\frac{1}{\delta}\log\left(1 + \frac{(e^{-\delta u_1} - 1)(e^{-\delta u_2} - 1)}{e^{-\delta} - 1}\right), \quad 0 \leq u_1, u_2 \leq 1.$$

Its conditional distribution function is

$$C_{2|1}(u_2|u_1;\delta) = e^{-\delta u_1}\left[\left(1-e^{-\delta}\right)\left(1-e^{-\delta u_2}\right)^{-1} - \left(1-e^{-\delta u_1}\right)\right]^{-1}.$$

Its density is

$$c(u_1,u_2;\delta) = \frac{\delta\left(1-e^{-\delta}\right)e^{-\delta(u_1+u_2)}}{\left[1-e^{-\delta}-\left(1-e^{-\delta u_1}\right)\left(1-e^{-\delta u_2}\right)\right]^2}, \quad 0 < u_1, u_2 < 1.$$

Its upper and lower tail dependence coefficients are given by

$$\lambda^{upper} = 0, \qquad \lambda^{lower} = 0.$$

Example 2.16 *(Bivariate Joe copula). For $\delta \in [1,\infty]$, the generator $\varphi(x;\delta) = -\log\left(1-(1-x)^{\delta}\right)$ yields the bivariate Joe copula. Its cdf is*

$$C(u_1,u_2;\delta) = 1 - \left[(1-u_1)^{\delta} + (1-u_2)^{\delta} - (1-u_1)^{\delta}(1-u_2)^{\delta}\right]^{1/\delta}, \quad 0 \le u_1, u_2 \le 1.$$

Its conditional distribution function is

$$C_{2|1}(u_2|u_1;\delta) = \left[1+(1-u_2)^{\delta}(1-u_1)^{-\delta} - (1-u_2)^{\delta}\right]^{-1+1/\delta}\left[1-(1-u_2)^{\delta}\right].$$

Its density for $\bar{u}_1 = 1-u_1, \bar{u}_2 = 1-u_2$ and $0 < u_1, u_2 < 1$ is

$$c(u_1,u_2;\delta) = \left(\bar{u}_1^{\delta} + \bar{u}_2^{\delta} - \bar{u}_1^{\delta}\bar{u}_2^{\delta}\right)^{-2+1/\delta}\bar{u}_1^{\delta-1}\bar{u}_2^{\delta-1}\left[\delta - 1 + \bar{u}_1^{\delta} + \bar{u}_2^{\delta} - \bar{u}_1^{\delta}\bar{u}_2^{\delta}\right].$$

Its upper and lower tail dependence coefficients are given by

$$\lambda^{upper} = 2 - 2^{1/\delta}, \qquad \lambda^{lower} = 0.$$

Example 2.17 *(Bivariate BB1 copula). For $\delta \in [1,\infty]$ and $\theta \in (0,\infty)$, the generator $\varphi(x;\delta,\theta) = \left(x^{-\theta}-1\right)^{\delta}$ yields the bivariate BB1 copula. Its cdf is*

$$C(u_1,u_2;\theta,\delta) = \left\{1 + \left[\left(u_1^{-\theta}-1\right)^{\delta} + \left(u_2^{-\theta}-1\right)^{\delta}\right]^{1/\delta}\right\}^{-1/\theta}, \quad 0 \le u_1, u_2 \le 1.$$

Its conditional distribution function for $a = (u_1^{-\theta}-1)^{\delta}$, $b = (u_2^{-\theta}-1)^{\delta}$, and $0 < u_1, u_2 < 1$ is

$$C_{2|1}(u_2|u_1;\theta,\delta) = \left(1+(a+b)^{1/\delta}\right)^{-1/\theta-1}(a+b)^{1/\delta-1}a^{1-1/\delta}u_1^{-\theta-1}.$$

Its density is

$$\begin{aligned} c(u_1,u_2;\theta,\delta) = {} & \left(1+(a+b)^{1/\delta}\right)^{-1/\theta-2}(a+b)^{1/\delta-2}\left[\theta(\delta-1) + (\theta\delta+1)(a+b)^{1/\delta}\right] \\ & \cdot (ab)^{1-1/\delta}(u_1u_2)^{-\theta-1}. \end{aligned}$$

Its upper and lower tail dependence coefficients are given by

$$\lambda^{upper} = 2 - 2^{1/\delta}, \qquad \lambda^{lower} = 2^{1/(\delta\theta)}.$$

Example 2.18 *(Bivariate BB6 copula). For $\delta \in [1,\infty]$ and $\theta \in [1,\infty]$, the generator $\varphi(x;\delta,\theta) = \left\{-\log\left[1-(1-x)^{\theta}\right]\right\}^{\delta}$ yields the bivariate BB6 copula. Its cdf with $\bar{u}_1 = 1-u_1$, $\bar{u}_2 = 1-u_2$, and $0 \leq u_1, u_2 \leq 1$ is*

$$C(u_1,u_2;\theta,\delta) = 1 - \left(1 - \exp\left\{-\left[\left(-\log\left(1-\bar{u}_1^{\theta}\right)\right)^{\delta} + \left(-\log\left(1-\bar{u}_2^{\theta}\right)\right)^{\delta}\right]^{1/\delta}\right\}\right)^{1/\theta}.$$

Its conditional distribution function with $a = -\log\left(1-(1-u_1^{\theta})\right)$, $b = -\log\left(1-(1-u_2^{\theta})\right)$, and $w = \exp\left(-\left(a^{\delta}+b^{\delta}\right)^{1/\delta}\right)$ is

$$C_{2|1}(u_2|u_1;\theta,\delta) = (1-w)^{1/\theta-1} w \left(a^{\delta}+b^{\delta}\right)^{1/\delta-1} a^{\delta-1} e^{a} \left(1-e^{-a}\right)^{1-1/\theta}.$$

Its density is

$$\begin{aligned} c(u_1,u_2;\theta,\delta) =& (1-w)^{1/\theta-2} w \left(a^{\delta}+b^{\delta}\right)^{1/\delta-2} \left[(\theta-w)\left(a^{\delta}+b^{\delta}\right)^{1/\delta} + \theta(\delta-1)(1-w)\right] \\ & \cdot (ab)^{\delta-1}\left(1-\bar{u}^{\theta}\right)^{-1}\left(1-\bar{v}^{\theta}\right)^{-1}(\bar{u}\bar{v})^{\theta-1}. \end{aligned}$$

Its upper and lower tail dependence coefficients are given by

$$\lambda^{upper} = 2 - 2^{1/(\delta\theta)}, \qquad \lambda^{lower} = 0.$$

Example 2.19 *(Bivariate BB7 copula). For $\delta \in (0,\infty]$ and $\theta \in [1,\infty]$, the generator $\varphi(x;\delta,\theta) = \left[1-(1-x)^{\theta}\right]^{-\delta} - 1$ yields the bivariate BB7 copula. Its cdf with $\bar{u}_1 = 1-u_1$, $\bar{u}_2 = 1-u_2$, and $0 \leq u_1, u_2 \leq 1$ is*

$$C(u_1,u_2;\theta,\delta) = 1 - \left(1 - \left[\left(1-\bar{u}_1^{\theta}\right)^{-\delta} + \left(1-\bar{u}_2^{\theta}\right)^{-\delta} - 1\right]^{-1/\delta}\right)^{1/\theta}.$$

Its conditional distribution function with $a = -\left(1-(1-u_1^{\theta})\right)^{-\delta}$, $b = -\left(1-(1-u_2^{\theta})\right)^{-\delta}$ is

$$C_{2|1}(u_2|u_1;\theta,\delta) = \left[1-(a+b+1)^{-1/\delta}\right]^{1/\theta-1}(a+b+1)^{-1/\delta-1}(a+1)^{1+1/\delta}(1-u_1)^{\theta-1}.$$

Its density is

$$\begin{aligned} c(u_1,u_2;\theta,\delta) =& \left[1-(a+b+1)^{-1/\delta}\right]^{1/\theta-2}(a+b+1)^{-1/\delta-2}[(a+1)(b+1)]^{1+1/\delta} \\ & \cdot \left[\theta(\delta+1) - (\theta\delta+1)(a+b+1)^{-1/\delta}\right][(1-u_1)(1-u_2)]^{\theta-1}. \end{aligned}$$

Its upper and lower tail dependence coefficients are given by

$$\lambda^{upper} = 2 - 2^{1/\theta}, \qquad \lambda^{lower} = 2^{-1/\delta}.$$

Example 2.20 *(Bivariate BB8 copula). For $\delta \in (0,1]$, $\theta \in [1,\infty]$, and $\eta = 1-(1-\delta)^{\vartheta}$, the generator $\varphi(x;\delta,\theta) = -\log\left\{\left[1-(1-\delta x)^{\theta}/\eta\right]\right\}$ yields the bivariate BB8 copula whose cdf is*

$$C(u_1,u_2;\vartheta,\delta) = \delta^{-1}\left(1-\left(1-\eta^{-1}\left[1-(1-\delta u_1)^{\vartheta}\right]\left[1-(1-\delta u_2)^{\vartheta}\right]\right)^{1/\vartheta}\right), \quad 0 \le u,v \le 1.$$

Its conditional distribution function and density with $a = 1-(1-\delta u_1)^{\vartheta}$, $b = 1-(1-\delta u_2)^{\vartheta}$:

$$C_{2|1}(u_2|u_1;\vartheta,\delta) = \frac{\eta^{-1}b\left(1-\eta^{-1}ab\right)^{1/\vartheta-1}}{(1-a)^{1/\vartheta-1}}.$$

$$c(u_1,u_2;\vartheta,\delta) = \eta^{-1}\delta\left(1-\eta^{-1}ab\right)^{1/\vartheta-2}\left(\vartheta-\eta^{-1}ab\right)(1-\delta u_1)^{\vartheta-1}(1-\delta u_2)^{\vartheta-1}.$$

Its upper and lower tail dependence coefficients are given by

$$\lambda^{upper} = 0, \ \textit{if } \delta \neq 1, \quad \lambda^{upper} = 2-2^{1/\theta}, \ \textit{if } \delta = 1, \qquad \lambda^{lower} = 0.$$

As seen in the above examples, bivariate Archimedean copulas may have one or two parameters that determine their shape, dependence, and tail behavior. Even though the bivariate Gaussian and t copulas can be parametrized by the correlation ρ, the same does not apply to bivariate Archimedean copulas. Moreover, the correlation changes by monotone transformations on the variables. Hence, to make the dependence consistent and comparable, we will use a monotone association measure invariant to monotone transformations on the variables: Kendall's τ (Kendall 1938). It quantifies the amount of increase tendency in one variable given that other increases or decreases. It is between -1 and 1, where $\tau = 1/-1$ refers to a perfect positive/negative monotonous association of two variables. Further, if two variables are independent, $\tau = 0$. In addition, there is a one-to-one relationship between Kendall's τ and parameters of bivariate Gaussian, t, Clayton, Gumbel, Frank, Joe, BB1, and BB7 copulas (see Table 3.2 of Czado (2019)).

Definition 2.3: Kendall's tau (τ)

Let (X_1,X_2) and $(\tilde{X}_1,\tilde{X}_2)$ be independent random pairs following a continuous distribution. The two pairs are concordant if $(X_1-\tilde{X}_1)(X_2-\tilde{X}_2) > 0$, discordant if $(X_1-\tilde{X}_1)(X_2-\tilde{X}_2) < 0$, and extra $X_1\&X_2$ pair if $X_1 = \tilde{X}_1 \& X_2 = \tilde{X}_2$. Then the Kendall's τ between random variables X_1 and X_2 is

$$\tau = P[(X_1-\tilde{X}_1)(X_2-\tilde{X}_2) > 0] - P[(X_1-\tilde{X}_1)(X_2-\tilde{X}_2) < 0].$$

Suppose that $(x_{i1},x_{i2}), i = 1,\ldots,n$, is a random sample from the joint distribution of (X_1,X_2). In the sample, let N_c be the number of concordant pairs, N_d be the number of discordant pairs, N_1 be the number of extra x_1 pairs, and N_2 be the number of extra x_2 pairs. Then the Kendall's τ estimate is

$$\hat{\tau} = \frac{N_c - N_d}{\sqrt{N_c+N_d+N_1}\times\sqrt{N_c+N_d+N_2}}.$$

Figure 2.4: Marginally normalized contour plots of bivariate Archimedean copulas. The empirical Kendall's τ is 0.50 for the copulas with one parameter and is ≈0.50 for the copulas with two parameters.

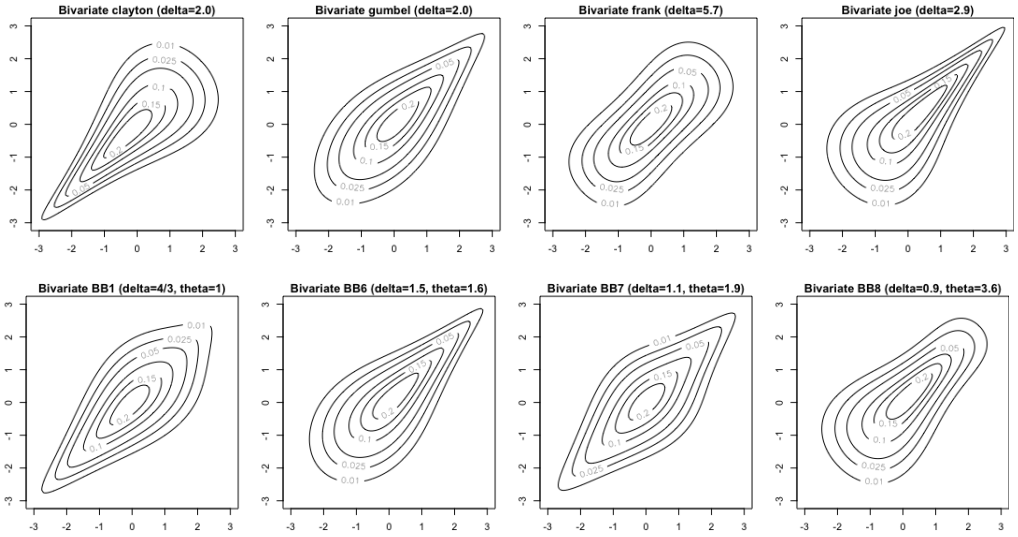
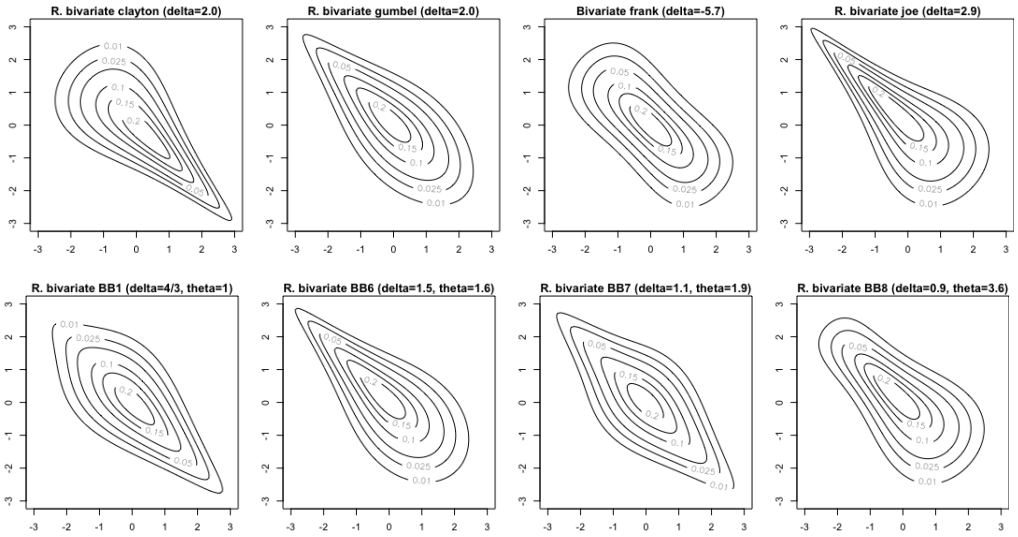


Figure 2.5: Marginally normalized contour plots of the $90°$ rotated bivariate Archimedean copulas, except Frank. The empirical Kendall's τ is 0.50 for the copulas with one parameter and is ≈0.50 for the copulas with two parameters.



2.2 Vine copulas

Another important question in copula construction is how to extend it from the bivariate case to the multivariate case to get even more flexible dependence and high-dimensional modeling. The pair-copula construction or the vine copula approach is an ad-hoc solution. It applies sequentially mixing conditional distributions (Joe 1996) involving a set of bivariate copulas. The set of bivariate copulas is mostly applied to pairs of univariate conditional distributions. As a result, vine copulas capture a wide range of dependence, including tail asymmetries. However, unlike the bivariate Gaussian and t copulas, they do not have the closure property under marginalization. Furthermore, even though vine copulas do not have a closed-form cumulative distribution function, they have a closed-form density (Bedford and Cooke 2001) and allow any d-dimensional copula and its density to be expressed by $\frac{d\cdot(d-1)}{2}$ bivariate copulas and their densities. Nevertheless, such an expression is not unique, and accounting for different expressions and organizing them is important. Such an organization tool based on an undirected graph structure was developed by Bedford and Cooke (2002) and denoted as a *regular vine (R-vine)*. In this section, we will introduce R-vines and then discuss the corresponding pair-copulas with their statistical properties, such as density functions, mainly based on Sahin and Czado (2022a) and Sahin and Czado (2022b). However, for more details about the pair-copula construction, we refer to Aas et al. (2009) and Joe (2014).

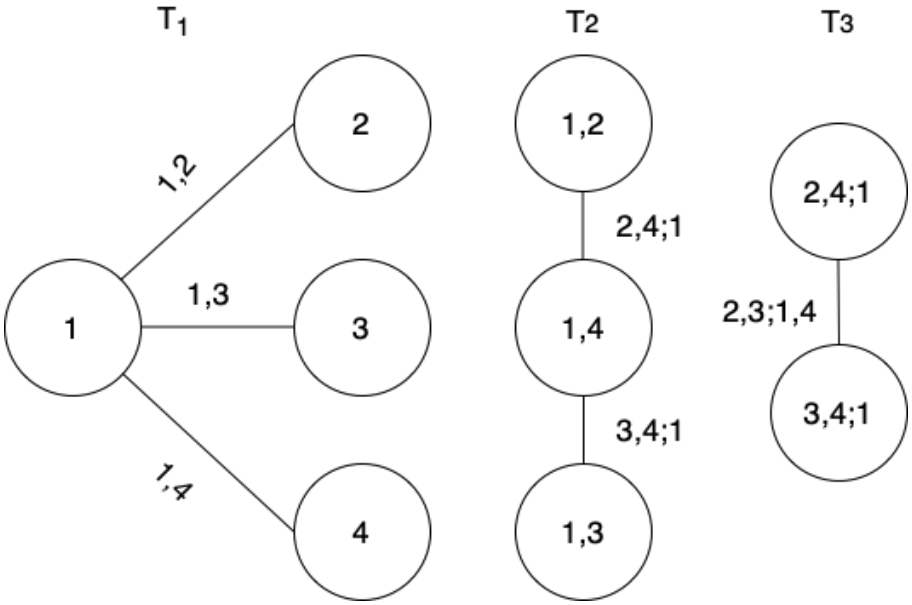
Suppose the interest is d-dimensional dependence modeling. Then a d-dimensional pair-copula is identified by an undirected graph. First, the graph, i.e., a d-dimensional R-vine, consists of $d-1$ nested trees, each tree T_m has a node set V_m, and an edge set E_m for $m = 1, \ldots, d-1$. The edges in tree level m are the nodes in tree level $m+1$, and a pair of nodes in tree T_{m+1} are allowed to be connected if they have a common node in tree T_m. A formal definition of Bedford and Cooke (2002) is that a structure $\mathcal{V} = (T_1, \ldots, T_{d-1})$ is a regular vine on d elements if it meets the following conditions:

1. T_1 is a tree with node set $V_1 = \{1, \ldots, d\}$ and edge set E_1.
2. T_m is a tree with node set $V_m = E_{m-1}$ for $m = 2, \ldots, d-1$.
3. *(Proximity)* If an edge connects two nodes in T_{m+1}, their associated edges in T_m must have a shared node in T_m.

The number of edges attached to a node gives the corresponding node's degree. If there is one node of degree $d-m$ in tree T_m for $m = 1, \ldots, d-2$, a regular vine is called a canonical (C-) vine as shown in Figure 2.6. Another special class of R-vines are D-vines, whose graph structure is a path, i.e., all nodes' degree in the graph is smaller than three, as shown in Figure 2.7. The node having a degree of one is called a leaf node.

In the representation of a vine copula or pair-copula construction by a vine, there is a bivariate (pair) copula associated with each edge in the vine. A node in a vine represents a variable, and an edge between a pair of nodes corresponds to dependence among the variables

Figure 2.6: Example of a 4-dimensional C-vine.



of the respective nodes expressed by a pair-copula. For instance, the nodes and edges in the first tree represent d variables and unconditional dependence for $d-1$ pairs of variables, respectively. In the higher trees, the conditional dependence of a pair of variables conditioning on other variables is modeled.

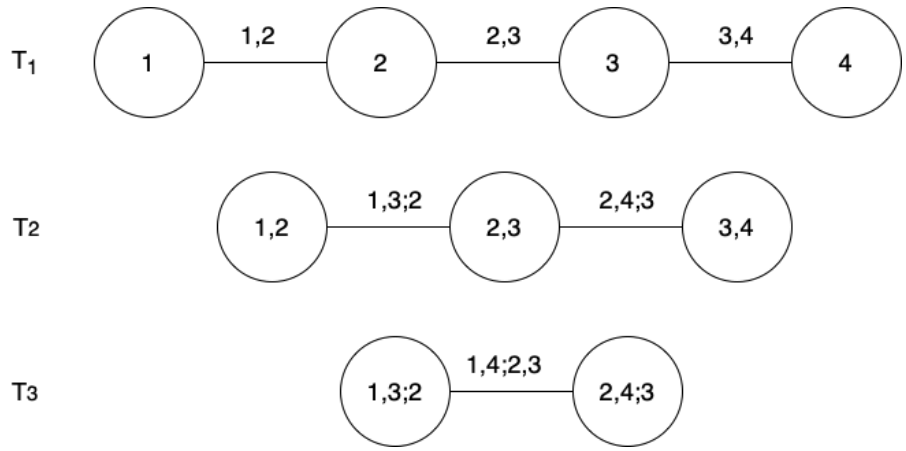
The structure $\mathcal{V} = (T_1, \ldots, T_{d-1})$ is also called a *vine tree structure* and truncating a vine after the first tree is equivalent to a Markov tree model. Thus, vines are extending Markov trees in the sense that they allow conditional dependencies. Moreover, if the pair-copulas in the higher tree levels than t are independence, where $t < d$ and $t \geq 1$, representing conditional independence, a t-truncated vine copula is obtained (Brechmann 2010).

Further, selecting the parameters of the edges in the first tree level as correlations and the parameters of the edges in the next tree levels as partial correlations, where the partial correlations are conditioned on $m-1$ variables in tree level m, a vine can represent a multivariate Gaussian distribution when margins follow a univariate normal distribution, and pair-copulas are Gaussian.

Example 2.21 *Following the notation in Chapter 5 of Czado (2019), Figure 2.7 shows the graphical specification for a 4-dimensional D-vine in the form of a set of linked trees. The D-vine with four variables has three trees, and the jth tree has $5-j$ nodes and $4-j$ edges, $j = 1, 2, 3$. The first tree level T_1 consists of the node set $V_1 = \{1, 2, 3, 4\}$ and the edge set $E_1 = \{\{1, 2\}, \{2, 3\}, \{3, 4\}\}$. In the second tree level, the node set is $V_2 = E_2$, and the edge set is $E_2 = \{\{1, 2; 3\}, \{2, 4; 3\}\}$. In the last tree level, two elements exist in V_3, and one edge connects two nodes.*

The first and fourth variables are leaf nodes in the first tree, and the pairs (1,2) and (3,4) are the leaf nodes in the second tree in Figure 2.7. Each edge is associated with a pair-copula to model the dependence between a pair of variables. Thus, in the first tree, the pair-copulas (their densities), $C_{1,2}$, $C_{2,3}$, and $C_{3,4}$ ($c_{1,2}$, $c_{2,3}$, and $c_{3,4}$) model the dependence of the three pairs of variables, (1,2), (2,3), and (3,4), respectively. In the second tree, two conditional dependencies are modeled: between the first and third variables given the second variable is modeled by the edge (1,3;2) with the associated pair-copula (density) $C_{1,3;2}$ ($c_{1,3;2}$), and between the second and fourth variables given the third variable is modeled by the edge (2,4;3) with the associated pair-copula (density) $C_{2,4;3}$ ($c_{2,4;3}$). The two nodes in T_2 are joined since they share the common nodes 2 or 3 in T_1. Likewise, the third tree models the conditional dependence between the first and fourth variables given the second and third variables using the associated pair-copula (density) $C_{1,4;2,3}$($c_{1,4;2,3}$). In addition, if $c_{1,4;2,3}$ is one everywhere in the domain, i.e., independence copula, the resulting D-vine is a 2-truncated vine.

Figure 2.7: Example of a 4-dimensional D-vine.



Moreover, we can give the 4-dimensional joint density g of the specified D-vine. Let F_1, F_2, F_3, F_4 denote parametric, univariate marginal distribution functions with the corresponding parameters $\gamma_1, \gamma_2, \gamma_3, \gamma_4$, and f_1, f_2, f_3, f_4 indicate their densities. Then the 4-dimensional joint density is

$$
\begin{aligned}
g(x_1, x_2, x_3, x_4; \boldsymbol{\psi}) = & c_{1,2}\big(F_1(x_1;\boldsymbol{\gamma}_1), F_2(x_2;\boldsymbol{\gamma}_2); \boldsymbol{\theta}_{1,2}\big) c_{2,3}\big(F_2(x_2;\boldsymbol{\gamma}_2), F_3(x_3;\boldsymbol{\gamma}_3); \boldsymbol{\theta}_{2,3}\big) \\
& \cdot c_{3,4}\big(F_3(x_3;\boldsymbol{\gamma}_3), F_4(x_4;\boldsymbol{\gamma}_4); \boldsymbol{\theta}_{3,4}\big) \\
& \cdot c_{1,3;2}\big(F_{1|2}(x_1|x_2;\boldsymbol{\gamma}_1,\boldsymbol{\gamma}_2,\boldsymbol{\theta}_{1,2}), F_{3|2}(x_3|x_2;\boldsymbol{\gamma}_2,\boldsymbol{\gamma}_3,\boldsymbol{\theta}_{2,3}); x_2, \boldsymbol{\theta}_{1,3;2}\big) \\
& \cdot c_{2,4;3}\big(F_{2|3}(x_2|x_3;\boldsymbol{\gamma}_2,\boldsymbol{\gamma}_3,\boldsymbol{\theta}_{2,3}), F_{4|3}(x_4|x_3;\boldsymbol{\gamma}_3,\boldsymbol{\gamma}_4,\boldsymbol{\theta}_{3,4}); x_3, \boldsymbol{\theta}_{2,4;3}\big) \\
& \cdot c_{1,4;2,3}\big(F_{1|2,3}(x_1|x_2,x_3;\boldsymbol{\gamma}_1,\boldsymbol{\gamma}_2,\boldsymbol{\gamma}_3,\boldsymbol{\theta}_{2,3},\boldsymbol{\theta}_{1,2}), \\
& F_{4|2,3}(x_4|x_2,x_3;\boldsymbol{\gamma}_2,\boldsymbol{\gamma}_3,\boldsymbol{\gamma}_4,\boldsymbol{\theta}_{2,3},\boldsymbol{\theta}_{3,4}); x_2, x_3, \boldsymbol{\theta}_{1,4;2,3}\big) \\
& \cdot f_1(x_1;\boldsymbol{\gamma}_1) f_2(x_2;\boldsymbol{\gamma}_2) f_3(x_3;\boldsymbol{\gamma}_3) f_4(x_4;\boldsymbol{\gamma}_4),
\end{aligned}
$$

where the vector $\boldsymbol{\psi}$ contains the marginal and pair-copula parameters. As an example of pair-copulas, $c_{1,3;2}\big(F_{1|2}(x_1|x_2), F_{3|2}(x_3|x_2); x_2\big)$ is the joint (copula) density corresponding to the random vector $\big(F_{1|2}(X_1|X_2), F_{3|2}(X_3|X_2)\big)^\top$ given $X_2 = x_2$. The similar notation applies to other pair-copulas. Moreover, $F_{1|2}$ is the conditional distribution of the random variable $X_1|X_2 = x_2$. Using the definition of a conditional distribution and the Sklar's Theorem, it can be shown that $F_{1|2} = \frac{\partial C_{1,2}(F_1(x_1;\boldsymbol{\gamma}_1),u_2;\boldsymbol{\theta}_{1,2})}{\partial u_2}|_{u_2=F_2(x_2;\boldsymbol{\gamma}_2)}$. For more details, we refer to Joe (1996).

In Example 2.21, it is assumed that all marginal distributions are parametric, which corresponds to the *inference for margins* (IFM) approach (Joe and Xu 1996). Alternatively, the copula data can be obtained empirically, such as using kernels, called a *semiparametric* approach (Genest et al. 1995).

In addition, observe that the copula density $c_{1,3;2}$ depends on the specific value x_2 of the conditioning variable X_2 in Example 2.21. A similar case of the dependence on the specific value of the conditioning variables can be seen for the pair-copulas $C_{2,4;3}$ and $C_{1,4;2,3}$. However, we will ignore this dependence to reduce model complexity, i.e., make the *simplifying assumption*. Under the simplifying assumption, the copula density does not have any conditional dependence on the specific value and is a 2-dimensional copula density. For instance, $c_{1,3;2}$ is independent of the specific value of x_2 and is a bivariate copula density. Nevertheless, the pair-copulas still depend on the conditioning value through their arguments. We refer to Stöber et al. (2013) for more details.

Example 2.22 *The 4-dimensional joint density g of Example 2.21 under the simplifying assumption is*

$$
\begin{aligned}
g(x_1,x_2,x_3,x_4;\boldsymbol{\psi}) =& c_{1,2}\big(F_1(x_1;\boldsymbol{\gamma}_1), F_2(x_2;\boldsymbol{\gamma}_2);\boldsymbol{\theta}_{1,2}\big)\, c_{2,3}\big(F_2(x_2;\boldsymbol{\gamma}_2), F_3(x_3;\boldsymbol{\gamma}_3);\boldsymbol{\theta}_{2,3}\big)\\
&\cdot c_{3,4}\big(F_3(x_3;\boldsymbol{\gamma}_3), F_4(x_4;\boldsymbol{\gamma}_4);\boldsymbol{\theta}_{3,4}\big)\\
&\cdot c_{1,3;2}\big(F_{1|2}(x_1|x_2;\boldsymbol{\gamma}_1,\boldsymbol{\gamma}_2,\boldsymbol{\theta}_{1,2}), F_{3|2}(x_3|x_2;\boldsymbol{\gamma}_2,\boldsymbol{\gamma}_3,\boldsymbol{\theta}_{2,3});\boldsymbol{\theta}_{1,3;2}\big)\\
&\cdot c_{2,4;3}\big(F_{2|3}(x_2|x_3;\boldsymbol{\gamma}_2,\boldsymbol{\gamma}_3,\boldsymbol{\theta}_{2,3}), F_{4|3}(x_4|x_3;\boldsymbol{\gamma}_3,\boldsymbol{\gamma}_4,\boldsymbol{\theta}_{3,4});\boldsymbol{\theta}_{2,4;3}\big)\\
&\cdot c_{1,4;2,3}\big(F_{1|2,3}(x_1|x_2,x_3;\boldsymbol{\gamma}_1,\boldsymbol{\gamma}_2,\boldsymbol{\gamma}_3,\boldsymbol{\theta}_{2,3},\boldsymbol{\theta}_{1,2}),\\
&F_{4|2,3}(x_4|x_2,x_3;\boldsymbol{\gamma}_2,\boldsymbol{\gamma}_3,\boldsymbol{\gamma}_4,\boldsymbol{\theta}_{2,3},\boldsymbol{\theta}_{3,4});\boldsymbol{\theta}_{1,4;2,3}\big)\\
&\cdot f_1(x_1;\boldsymbol{\gamma}_1) f_2(x_2;\boldsymbol{\gamma}_2) f_3(x_3;\boldsymbol{\gamma}_3) f_4(x_4;\boldsymbol{\gamma}_4).
\end{aligned}
$$

For a general dimension d and R-vines, a d-dimensional joint density is similarly constructed using d univariate marginal densities and $\frac{d\cdot(d-1)}{2}$ associated pair-copula densities. Let $c_{e_a,e_b;D_e}$ be a parametric pair-copula density associated with an edge e in vine tree structure $\mathcal{V}$ and $\boldsymbol{\theta}_{e_a,e_b;D_e}$ be its parameters ($e \in E_m, \text{ for } m = 1,\dots,d-1$). Let f_p denote a parametric, univariate marginal density with the parameters $\boldsymbol{\gamma}_p$ for $p = 1,\dots,d$. Then a d-dimensional

joint density g under the simplifying assumption can be constructed as follows:

$$g(\boldsymbol{x};\boldsymbol{\psi}) = \prod_{m=1}^{d-1}\prod_{e\in E_m} c_{e_a,e_b;D_e}\big(F_{e_a|D_e}(x_{e_a}|\boldsymbol{x}_{D_e};\boldsymbol{\gamma}_{e_a|D_e},\boldsymbol{\theta}_{e_a|D_e}), \\ F_{e_b|D_e}\big(x_{e_b}|\boldsymbol{x}_{D_e};\boldsymbol{\gamma}_{e_b|D_e},\boldsymbol{\theta}_{e_b|D_e});\boldsymbol{\theta}_{e_a,e_b;D_e}\big)\prod_{p=1}^{d} f_p(x_p;\boldsymbol{\gamma}_p), \tag{2.4}$$

where $\boldsymbol{x}_{D_e} = (x_z)_{z\in D_e}$ is a subvector of $\boldsymbol{x} = (x_1,\dots,x_d)^\top \in \mathbb{R}^d$, the parameter vector $\boldsymbol{\psi}$ contains the marginal and pair-copula parameters, $F_{e_a|D_e}$ is the conditional distribution function of the random variable $X_{e_a}|\boldsymbol{X}_{D_e} = \boldsymbol{x}_{D_e}$. It can be calculated recursively using only pair-copula terms occuring in the previous tree levels (Joe 1996). The marginal $\boldsymbol{\gamma}_{e_a|D_e}$ and pair-copula $\boldsymbol{\theta}_{e_a|D_e}$ parameters, which are based on the previous tree levels, are used to determine $F_{e_a|D_e}$. The set D_e is called the *conditioning set*, and the indices e_a, e_b form the *conditioned set*. D_e has $m-1$ elements in tree level m; therefore, it is empty in the first tree.

D-vine copulas also allow to express the conditional density of a leaf node in the first tree in a closed form. Assume that $(x_{i,1},\dots,x_{i,d})$, $i = 1,\dots,n$, are realizations of the random vector $(X_1,\dots,X_d)$ with their marginal distributions $F_1,\dots,F_d$. For the D-vine copula with the node order $1-\ldots-d$ corresponding to the variables $X_1-\ldots-X_d$, $d\geq 2$, the conditional density $f_{1|2,\dots,d}$ of X_1 given the others is

$$f_{1|2\dots,d}(x_1|x_2,\dots,x_d) = \Big[\prod_{j=3}^{d} c_{1,j;2,\dots,j-1}\big(F_{1|2,\dots,j-1}(x_1|x_2,\dots,x_{j-1}), F_{j|2,\dots,j-1}(x_j|x_2,\dots,x_{j-1})\big)\Big] \\ \cdot c_{1,2}\big(F_1(x_1),F_2(x_2)\big) f_1(x_1).$$

Further, we can calculate the conditional log-likelihood $cll_{1|2,\dots,d}(F_1(x_1),F_2(x_2)\dots,F_d(x_d))$ based on the D-vine copula as

$$cll_{1|2,\dots,d}(F_1(x_1),\dots,F_d(x_d)) = \sum_{j=3}^{d}\Big[\log c_{1,j;2,\dots,j-1}\big(F_{1|2,\dots,j-1}(x_1|x_2,\dots,x_{j-1}), F_{j|2,\dots,j-1}(x_j|x_2,\dots,x_{j-1})\big)\Big] \\ + \log c_{1,2}\big(F_1(x_1),F_2(x_2)\big) + \log f_1(x_1). \tag{2.5}$$

Additionally, as stated in Kraus and Czado (2017), the conditional quantile function $F^{-1}_{1|2,\dots,d}$ at quantile α can be expressed in terms of the inverse marginal distribution function F_1^{-1} of X_1 and the conditional D-vine copula quantile function $C^{-1}_{1|2,\dots,d}$ at quantile α:

$$F^{-1}_{1|2,\dots,d}(\alpha|x_2,\dots,x_d) = F_1^{-1}\big(C^{-1}_{1|2,\dots,d}(\alpha|F_2(x_2),\dots,F_d(x_d)\big). \tag{2.6}$$

Once the node order in the D-vines' first tree is determined, the associated D-vine copula structure is also determined uniquely. However, finding the node order in the first tree is still a question. For R-vines, the selection of a vine tree structure is even more challenging than

for D-vines since each tree level needs to be determined. In addition, one needs to select the pair-copula families and estimate their parameters associated with a vine. For more details and heuristics approaches to be used, we refer to Chapter 8 of Czado (2019).

2.3 Numerical optimization

This section briefly introduces numerical optimization used in Chapters 3 and 5. Numerical optimization is an important tool for estimating the parameters of statistical models. It aims to find the values of the parameters that optimize the given objective function. The objective function can be in different forms, and the parameters can be constrained. For example, consider statistical maximum likelihood estimation: the objective is to maximize the likelihood of a given statistical model, which depends on parameters.

Let $\boldsymbol{x} \in \mathbb{R}^n$ denote the parameter vector of interest. Assume the objective is expressed by its scalar function $f(\boldsymbol{x}) : \mathbb{R}^n \to \mathbb{R}$ to be minimized or maximized. The vector valued function $g(\boldsymbol{x}) : \mathbb{R}^n \to \mathbb{R}^m$ contains scalar constraint functions that define inequalities the parameter vector needs to satisfy. Likewise, the equations the parameter vector must comply with is expressed by the scalar constraint functions $h(\boldsymbol{x}) : \mathbb{R}^n \to \mathbb{R}^p$. Then the optimization problem is written as

$$\max_{\boldsymbol{x}} \quad f(\boldsymbol{x}) \tag{2.7a}$$

$$\text{subject to } h(\boldsymbol{x}) = 0, \tag{2.7b}$$

$$g(\boldsymbol{x}) \leq 0. \tag{2.7c}$$

Such problems convey different characteristics according to the nature of the objective function, constraints, the number of parameters, and the smoothness of the functions. For instance, if constraints exist/do not exist, the problems are classified as constrained/unconstrained optimization. A derivative-free optimization tool can be used if the functions' derivatives are unavailable (Larson et al. 2019; Powell 2003) as in Section 5. Based on the optimization problem characteristics, different numerical optimization techniques exist.

While finding optimal parameters of a statistical model, one usually has an unconstrained optimization problem, or the given problem can be converted into it as in Section 3. Thus, we focus on solving it. An optimization algorithm starts with an initial point $\boldsymbol{x}_0$ and generates iterative sequences $\{\boldsymbol{x}_k\}_{k=0}^{\infty}$ until a solution is sufficiently accurate or the objective cannot be improved. One of the main differences between algorithms to create such sequences lies in their search mechanism about moving from one point $\boldsymbol{x}_k$ to the next. Two main strategies are line search (Lemaréchal 2005) and trust region methods (Sorensen 1981).

In line search methods, the algorithm chooses a direction $\boldsymbol{p}_k$, the negative (approximate) gradient at the current point, to search for a good solution point. Then the main idea is to

find the step size in the direction that optimizes the objective function. The search for an optimal step size can be expressed as follows

$$\min_{\alpha>0} \quad f(\boldsymbol{x}_k + \alpha \boldsymbol{p}_k).$$

Even though line search methods are simple and efficient to implement, they can have slow convergence rates and get stuck in narrow valleys, giving only a local optimum point.

Another method, i.e., trust region, constructs trustworthy regions to look for a good solution point. After identifying a region, the method searches for steps within the region, optimizing the objective function. A model function m_k using the information in the objective function can be constructed at each iteration as

$$\min_{\boldsymbol{s}} \quad m_k(\boldsymbol{x}_k + \boldsymbol{s}),$$

where $\boldsymbol{x}_k + \boldsymbol{s}$ is in the trust region. Let ∇f_k and B_k denote the gradient of f_k and a matrix at the current point $\boldsymbol{x}_k$, respectively. B_k can be the (approximate) Hessian matrix, e.g., $\nabla^2 f_k$ or $\approx \nabla^2 f_k$. Then m_k is usually a quadratic function, e.g., the second order Taylor approximation of the objective function, in the following form

$$m_k(\boldsymbol{x}_k + \boldsymbol{s}) = f_k + \boldsymbol{s}^\top \nabla f_k + \boldsymbol{s}^\top B_k \boldsymbol{s}.$$

Therefore, a good solution point is searched within the trust region and optimizes the model function that approximates the objective function. If improvement is acceptable/unsatisfactory in a given iteration of the trust regions method, the region gets expanded/shrank. Trust region methods balance the exploration and exploitation of the objective function; however, initial trust region size may significantly impact the results and need fine-tuning.

Example 2.23 *Consider the following optimization problem*

$$\min_{\boldsymbol{x}} \quad (x_1 - 1)^2 + (x_2 - 1)^2.$$

where $f(\boldsymbol{x}) = (x_1 - 1)^2 + (x_2 - 1)^2$. *The optimal solution is* $\boldsymbol{x}^* = (1, 1)^\top$.

Suppose a line search and a trust region method are at the point $\boldsymbol{x}_k = (0, 0)^\top$. *At* $\boldsymbol{x}_k$, *the gradient vector and the Hessian matrix of the objective is given by* $\nabla f_k = (-2, -2)^\top$ *and* $\nabla^2 f_k = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$, *respectively.*

In the next iteration, a trust region can be defined as a ball centered at the current point $\boldsymbol{x}_k$ *with a radius* R. *Then R corresponds to the size of the trust region. On the other hand, the line search looks for the best step size in the direction* $-\nabla f_k = -(-2, -2)^\top = (2, 2)^\top$.

To sum up, a line search method fixes the step direction and looks for an optimal step size, whereas a trust region method identifies a maximum distance that the method can search for a good point and then look for a direction and step size simultaneously. For more details, we refer to the book by Wright and Nocedal (2006).

Chapter 3

Vine copula mixture models

This chapter is mainly based on Sahin and Czado (2022b), but Section 3.7 introduces its software implementation, and Section 3.8 discusses further open research problems.

3.1 Motivation

Finite mixture models are convenient statistical tools for model-based clustering. They assume that observations in the multivariate data can be clustered using k components. Each component has its density, and each observation is assigned to a component with a probability. They have many applications in finance, genetics, and marketing (e.g., Hu (2006); Gambacciani and Paolella (2017); Sun et al. (2017); Zhang and Shi (2017)). McLachlan and Peel (2000) provides more details about finite mixture models. Bouveyron and Brunet-Saumard (2014) and McNicholas (2016) review recent model-based clustering methods.

One of the main questions to be addressed in the finite mixture models is *how to select the density of each component*. An early answer to this question is to assume a symmetric distribution such as multivariate normal distribution (e.g., Celeux and Govaert (1995); Fraley and Raftery (1998)) or multivariate t distribution (e.g., Peel and McLachlan (2000); Andrews and McNicholas (2011)). However, these models cannot accommodate the shape of asymmetric components. Hennig (2010) showed one such data example. Their skewed formulations and factor analyzers, therefore, have been extensively studied (e.g., Lin et al. (2007); Lee and McLachlan (2014); Murray et al. (2017)).

Additionally, the models based on other distributions, for example, shifted asymmetric Laplace distributions (Franczak et al. 2014), multivariate power exponential distributions (Dang et al. 2015), and generalized hyperbolic distributions (Browne and McNicholas 2015) have been proposed for the past few years. Since one of the main interests of copulas is to relax the normality assumption both in marginal distributions and dependence structure, the finite mixture models with copulas have also been studied (e.g., Diday and Vrac (2005); Kosmidis and Karlis (2016); Zhuang et al. (2021)). Cuvelier and Noirhomme-Fraiture (2005)

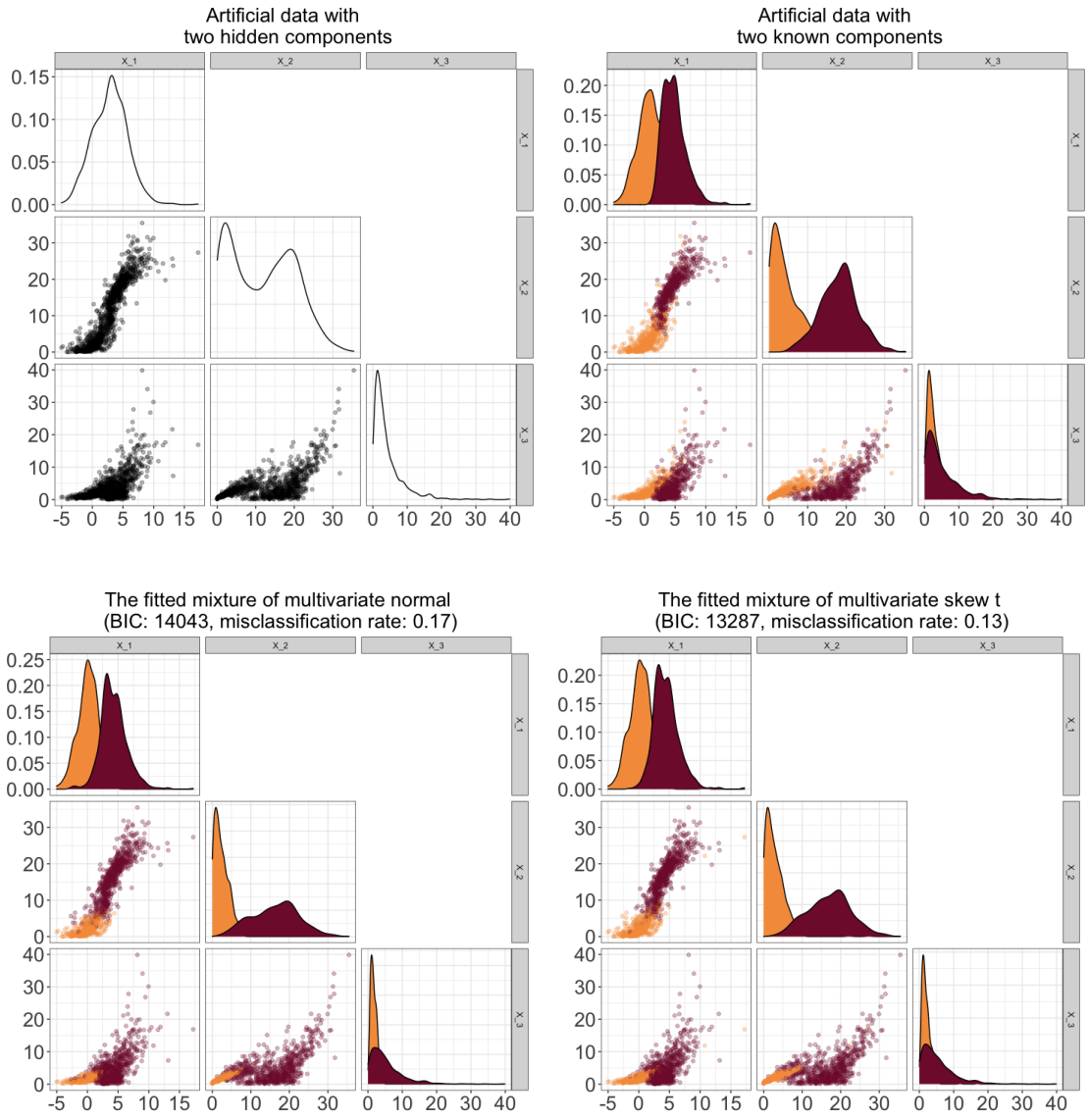
worked with the Clayton copula to represent lower tail dependence within the components, while Vrac et al. (2005) applied the Frank copula to have non-Gaussian but symmetric dependence. Nevertheless, these methods raise another question in the finite mixture models: *how to select flexible densities of each component* so the model can represent different asymmetric or/and tail dependencies for different pairs of variables. For this question, the vine copula or pair-copula construction is a flexible and efficient tool in high-dimensional dependence modeling (Aas et al. 2009; Joe 1996).

To motivate, consider a data set shown in Figure 3.1 simulated from a mixture of two three-dimensional vine copulas. Its data generating process includes asymmetric tail and non-Gaussian dependencies in the components. Most univariate margins are chosen to be non-Gaussian and heavy-tailed as listed in Table 3.3. As seen in the top left panel of Figure 3.1, the components are well separated in two out of three bivariate scatter plots. Marginal multimodality can be seen for the second variable. The resulting components are non-elliptical, similar to a banana shape, as shown in the top right. After fitting a mixture of multivariate normal distributions with two components, its challenge to capture the true shape of the components can be seen in the bottom left panel, where the associated Bayesian Information Criterion (BIC) (Schwarz 1978) and misclassification rate, the number of misclassified observations in the sample divided by the total number of observations, are provided. Even though fitting a mixture of multivariate skew t distributions with two components in the bottom right provides a better fit than the mixture of multivariate normal distributions fit, it can also not reveal the true characteristics of the data set. The mixture of multivariate normal distributions and multivariate skew t distributions is fitted using the `R` packages `mclust` (Scrucca et al. 2016) and `mixsmsn` (Prates et al. 2013), respectively. The fitting procedures apply 100 different seeds and report the best performance in terms of the misclassification rate.

To capture such behavior, applying vine copulas in finite mixture models has been explored before (e.g., Kim et al. (2013); Roy and Parui (2014); Weiß and Scheffer (2015); Sun et al. (2017)). However, they only worked with a subclass of vine tree structures and a small number of pair-copula families. Therefore, a vine copula mixture model allowing for all classes of vine tree structures and many different pair-copula families is needed. Since it provides flexible densities, formulating its model-based clustering algorithm overcomes the drawbacks mentioned above, especially for non-Gaussian data.

In this chapter, we formulate a vine copula mixture model for continuous data allowing all types of vine tree structures, parametric pair-copulas, and margins. For simplicity, we treat the number of components as known and present well-performing solutions for the remaining model selection problems. We adopt the expectation conditional maximum algorithm for parameter estimation (Meng and Rubin 1993). We present the first study in the finite mixture models literature that works with the full class of vine tree structures and a wide range of pair-copula families to the best of our knowledge. It combines the flexibility of vine copulas and finite mixture models to capture complex and diverse dependence structures in multivariate data. Another contribution is a new model-based clustering algorithm, called *VCMM*,

Figure 3.1: Pairwise scatter plot of a simulated data set (500 observations per component) under the scenario specified in Table 3.3 (top left). The orange and red points show the observations of one cluster and the other cluster (top right), respectively. The bottom left and bottom right plots display the fitted mixture of multivariate normal and multivariate skew t distributions, respectively. The diagonal of the plots gives the fitted marginal density function for each component.



that incorporates realistic interdependence structures of clusters. The proposed algorithm is interpretable and allows for various shapes of the clusters. For instance, it shows how the dependence structure varies within clusters of the data.

The remainder of the chapter is organized as follows. Sections 3.2, 3.3, and 3.4 describe the vine copula mixture model and discuss the model selection and parameter estimation problems. Section 3.5 provides the new model-based clustering algorithm based on it, and simulation studies are presented in Section 3.6. The corresponding software is detailed in Section 3.7. Section 3.8 discusses open problems, and Section 3.9 concludes the chapter.

3.2 Vine copula mixture model formulation

We formulate a vine copula mixture model, which is fully parametric, i.e., it works with parametric pair-copulas and univariate marginal distributions.

Suppose data consists of n observations, where an observation $\boldsymbol{x}_i = (x_{i,1}, \ldots, x_{i,d})^\top$ is an independent realization of a d-dimensional random vector $\boldsymbol{X} = (X_1, \ldots, X_d)^\top$ for $i = 1, \ldots, n$. Assume that a mixture of k components ($k \in \mathbb{R}^+$) generates the data, and the density g_j of the jth component for $j = 1, \ldots, k$ can be stated as given in Equation (2.4). Assume that, additionally, the jth component has a mixture weight π_j with $\pi_j \in (0, 1)$ for $j = 1, \ldots, k$ and $\sum_{j}^{k} \pi_j = 1$. Then the density of the vine copula mixture model for $\boldsymbol{X} = (X_1, \ldots, X_d)^\top$ at $\boldsymbol{x} = (x_1, \ldots, x_d)^\top$ can be written as

$$g(\boldsymbol{x}; \boldsymbol{\eta}) = \sum_{j=1}^{k} \pi_j \cdot g_j(\boldsymbol{x}; \boldsymbol{\psi}_j). \tag{3.1}$$

Here the vector $\boldsymbol{\psi}_j$ contains the marginal and pair-copula parameters of the jth component, $\boldsymbol{\eta}$ denotes all model parameters, i.e., $\boldsymbol{\eta} = (\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_k)^\top$ and $\boldsymbol{\eta}_j = (\pi_j, \boldsymbol{\psi}_j)^\top$ for $j = 1, \ldots, k$.

Example 3.1 *(Vine copula mixture model formulation in three dimensions with two components). Assume data, where an observation* $\boldsymbol{x}_i = \left(x_{i,1}, x_{i,2}, x_{i,3}\right)^\top$ *for* $i = 1, \ldots, n$ *is given, and there are two components generating the data with the mixture weights* $\pi_1 > 0$, $\pi_2 > 0$, *and* $\pi_1 + \pi_2 = 1$. *An observation of the first and second component is an independent realization of a* 3*-dimensional random vector* $\boldsymbol{X}_{(1)} = (X_{1(1)}, X_{2(1)}, X_{3(1)})^\top$ *and* $\boldsymbol{X}_{(2)} = (X_{1(2)}, X_{2(2)}, X_{3(2)})^\top$, *respectively.*

The vine tree structure's edge sets of the first component are $E_{(1)1} = \{\{1, 2\}, \{2, 3\}\}$ *for the first tree level,* $E_{(1)2} = \{1, 3; 2\}$ *for the second tree level and that of the second component are* $E_{(2)1} = \{\{1, 2\}, \{1, 3\}\}$, $E_{(2)2} = \{2, 3; 1\}$, *respectively. The pair-copula families and marginal distributions of both components are parametric. Accordingly,* $C_{(1)1,2}$, $C_{(1)2,3}$, $C_{(1)1,3;2}$ *denote pair-copula families of the first component associated with the edges in* $E_{1(1)}$ *and* $E_{2(1)}$, *while* $\boldsymbol{\theta}_{(1)1,2}$, $\boldsymbol{\theta}_{(1)2,3}$, $\boldsymbol{\theta}_{(1)1,3;2}$ *show the corresponding parameters. For the second component, we use notations* $C_{(2)1,2}$, $C_{(2)1,3}$, $C_{(2)2,3;1}$ *for pair-copula families associated with the edges in* $E_{1(2)}$ *and* $E_{2(2)}$, *and* $\boldsymbol{\theta}_{(2)1,2}$, $\boldsymbol{\theta}_{(2)1,3}$, $\boldsymbol{\theta}_{(2)2,3;1}$ *refer to the corresponding parameters. A copula density is denoted by the small* c *letter.*

$F_{1(1/2)}, F_{2(1/2)}, F_{3(1/2)}$ *refer to the marginal distributions of the random variables* $X_{1(1/2)}$, $X_{2(1/2)}$, $X_{3(1/2)}$ *with the corresponding parameters* $\boldsymbol{\gamma}_{1(1/2)}, \boldsymbol{\gamma}_{2(1/2)}, \boldsymbol{\gamma}_{3(1/2)}$ *for the first/second component. The small* f *letter denotes a marginal density.*

We can now write the **density of the first component** *at* $\boldsymbol{x} = (x_1, x_2, x_3)^\top$*:*

$$\begin{aligned} g_1(\boldsymbol{x}; \boldsymbol{\psi}_1) = & c_{(1)1,2}\big(F_{1(1)}(x_1; \boldsymbol{\gamma}_{1(1)}), F_{2(1)}(x_2; \boldsymbol{\gamma}_{2(1)}); \boldsymbol{\theta}_{(1)1,2}\big) \\ & \cdot c_{(1)2,3}\big(F_{2(1)}(x_2; \boldsymbol{\gamma}_{2(1)}), F_{3(1)}(x_3; \boldsymbol{\gamma}_{3(1)}); \boldsymbol{\theta}_{(1)2,3}\big) \\ & \cdot c_{(1)1,3;2}\big(F_{(1)1|2}(x_1|x_2; \boldsymbol{\gamma}_{1(1)}, \boldsymbol{\gamma}_{2(1)}, \boldsymbol{\theta}_{(1)1,2}), F_{(1)3|2}(x_3|x_2; \boldsymbol{\gamma}_{3(1)}, \boldsymbol{\gamma}_{2(1)}, \boldsymbol{\theta}_{(1)2,3}); \boldsymbol{\theta}_{(1)1,3;2}\big) \\ & \cdot f_{1(1)}(x_1; \boldsymbol{\gamma}_{1(1)}) \cdot f_{2(1)}(x_2; \boldsymbol{\gamma}_{2(1)}) \cdot f_{3(1)}(x_3; \boldsymbol{\gamma}_{3(1)}), \end{aligned} \tag{3.2}$$

where the pair-copula parameters used to determine the conditional distribution functions $F_{(1)1|2}$ *and* $F_{(1)3|2}$ *are given by* $\boldsymbol{\theta}_{(1)1|2} = \boldsymbol{\theta}_{(1)1,2}$ *and* $\boldsymbol{\theta}_{(1)3|2} = \boldsymbol{\theta}_{(1)2,3}$*, respectively. The marginal parameters needed for the same calculation are denoted by* $\boldsymbol{\gamma}_{(1)1|2} = (\boldsymbol{\gamma}_{1(1)}, \boldsymbol{\gamma}_{2(1)})^\top$ *and* $\boldsymbol{\gamma}_{(1)3|2} = (\boldsymbol{\gamma}_{2(1)}, \boldsymbol{\gamma}_{3(1)})^\top$*. We show the marginal and pair-copula parameters of the first component by* $\boldsymbol{\psi}_1 = (\boldsymbol{\gamma}_1, \boldsymbol{\theta}_1)^\top$*, where* $\boldsymbol{\gamma}_1 = (\boldsymbol{\gamma}_{1(1)}, \boldsymbol{\gamma}_{2(1)}, \boldsymbol{\gamma}_{3(1)})^\top$ *and* $\boldsymbol{\theta}_1 = (\boldsymbol{\theta}_{(1)1,2}, \boldsymbol{\theta}_{(1)2,3}, \boldsymbol{\theta}_{(1)1,3;2})^\top$*. As in the first component, we can define the parameters and write the* **density of the second component** *at* $\boldsymbol{x} = (x_1, x_2, x_3)^\top$*:*

$$\begin{aligned} g_2(\boldsymbol{x}; \boldsymbol{\psi}_2) = & c_{(2)1,2}\big(F_{1(2)}(x_1; \boldsymbol{\gamma}_{1(2)}), F_{2(2)}(x_2; \boldsymbol{\gamma}_{2(2)}); \boldsymbol{\theta}_{(2)1,2}\big) \\ & \cdot c_{(2)1,3}\big(F_{1(2)}(x_1; \boldsymbol{\gamma}_{1(2)}), F_{3(2)}(x_3; \boldsymbol{\gamma}_{3(2)}); \boldsymbol{\theta}_{(2)1,3}\big) \\ & \cdot c_{(2)2,3;1}\big(F_{(2)2|1}(x_2|x_1; \boldsymbol{\gamma}_{2(2)}, \boldsymbol{\gamma}_{1(2)}, \boldsymbol{\theta}_{(2)1,2}), F_{(2)3|1}(x_3|x_1; \boldsymbol{\gamma}_{3(2)}, \boldsymbol{\gamma}_{1(2)}, \boldsymbol{\theta}_{(2)1,3}); \boldsymbol{\theta}_{(2)2,3;1}\big) \\ & \cdot f_{1(2)}(x_1; \boldsymbol{\gamma}_{1(2)}) \cdot f_{2(2)}(x_2; \boldsymbol{\gamma}_{2(2)}) \cdot f_{3(2)}(x_3; \boldsymbol{\gamma}_{3(2)}), \end{aligned} \tag{3.3}$$

As a result, the **vine copula mixture model density** *at* $\boldsymbol{x} = (x_1, x_2, x_3)^\top$ *is given by*

$$g(\boldsymbol{x}; \boldsymbol{\eta}) = \pi_1 g_1(\boldsymbol{x}; \boldsymbol{\psi}_1) + \pi_2 g_2(\boldsymbol{x}; \boldsymbol{\psi}_2), \tag{3.4}$$

where $\boldsymbol{\eta}_1 = (\pi_1, \boldsymbol{\psi}_1)^\top$ *and* $\boldsymbol{\eta}_2 = (\pi_2, \boldsymbol{\psi}_2)^\top$ *indicate the model parameters of the first and second component. All model parameters are given by* $\boldsymbol{\eta} = (\boldsymbol{\eta}_1, \boldsymbol{\eta}_2)^\top$*.*

3.3 Model selection

Vine copula mixture models inherit the problem of estimating the total number of components k hidden in the data from finite mixture models. Moreover, due to its formulation in Equations (2.4) and (3.1), the vine tree structure $\mathcal{V}_j$, pair-copula families $\mathcal{B}_j(\mathcal{V}_j)$, and marginal distributions $\mathcal{F}_j = \{F_{1(j)}, \ldots, F_{d(j)}\}$ of the jth component need to be chosen for $j = 1, \ldots, k$. Accordingly, pair-copula parameters $\boldsymbol{\theta}_j(\mathcal{B}_j(\mathcal{V}_j))$ and marginal parameters $\boldsymbol{\gamma}_j(\mathcal{F}_j)$ should be estimated for $j = 1, \ldots, k$. To simplify, we will assume that the total number of components generating the data is known. If a priori information about k with complete certainty does not

exist, methods that estimate it from the data need to be developed for vine copula mixture models. Even though we will propose a first approach for the number of clusters selection problem in Section 3.8, it is currently not our focus. Instead, we will explain the approaches to the remaining model selection problems.

Assume an observation $\boldsymbol{x}_i = (x_{i,1}, \ldots, x_{i,d})^\top$ is assigned to a component for $i = 1, \ldots, n$. We will learn model components with the assignment for which approaches are presented in Section 3.5. We will discuss a modification when it is hard to learn model components, e.g., components have non-negligible overlaps, in Section 3.6.

Assume the total number of observations assigned to the jth component is n_j, and the observations belonging to the jth component are given by $\boldsymbol{x}_{(j)i_j} = (x_{(j)i_j,1}, \ldots, x_{(j)i_j,d})^\top$ for $i_j = 1, \ldots n_j$ and $j = 1, \ldots, k$. It holds that $\sum_{j=1}^{k} n_j = n$ and $\bigcup_{\forall(j,i_j)} \boldsymbol{x}_{(j)i_j} = \bigcup_{\forall i} \boldsymbol{x}_i$. We denote the pth variable in the jth component by $\boldsymbol{x}_{p(j)} = (x_{(j)1,p}, \ldots, x_{(j)n_j,p})^\top$ for $p = 1, \ldots d$ and $j = 1, \ldots, k$. With the given assignment, we first select the marginal distributions of each cluster. Their candidate set could be prespecified or chosen by a data analysis, such as using a histogram and quantile-quantile (QQ) plot. In addition, the marginal distribution family $F_{p(j)}$ for the variable $\boldsymbol{x}_{p(j)}$ can be determined using a model selection criteria. More precisely:

- **Marginal distribution selection** $\mathcal{F}_j$: For $p = 1, \ldots, d$ and $j = 1, \ldots, k$, the log-likelihood of the marginal distribution $F_{p(j)}$ with the density $f_{p(j)}$ and parameters $\boldsymbol{\gamma}_{p(j)}$ on the variable $\boldsymbol{x}_{p(j)}$ is

$$\ell(\boldsymbol{\gamma}_{p(j)}) = \sum_{i=1}^{n_j} \log\left(f_{p(j)}(x_{(j)i,p}; \boldsymbol{\gamma}_{p(j)})\right) \text{ for } \quad p = 1, \ldots d \quad \text{and} \quad j = 1, \ldots k. \tag{3.5}$$

  The BIC, a commonly used model selection criteria, is given by

$$BIC(\boldsymbol{\gamma}_{p(j)}) = -2 \cdot \ell(\boldsymbol{\gamma}_{p(j)}) + |\boldsymbol{\gamma}_{p(j)}| \cdot \log(n_j) \text{ for } p = 1, \ldots d \text{ and } j = 1, \ldots k, \tag{3.6}$$

  where $|\boldsymbol{\gamma}_{p(j)}|$ refers to the number of marginal parameters in $F_{p(j)}$, and n_j denotes the total number of observations in $\boldsymbol{x}_{p(j)}$. For each candidate for the marginal distribution on the variable $\boldsymbol{x}_{p(j)}$, first, the parameters that maximize the log-likelihood $\ell(\boldsymbol{\gamma}_{p(j)})$ are estimated, then the marginal distribution family with the lowest BIC is selected: $\hat{F}_{p(j)}$.

Since the joint density can be decomposed into univariate marginal densities and a vine copula density, we now estimate the u-data for a vine copula model by applying the probability integral transformation using the estimated margins $\hat{F}_{p(j)}$ for each cluster: $\hat{\boldsymbol{u}}_{p(j)} = \hat{F}_{p(j)}(\boldsymbol{x}_{p(j)}; \hat{\boldsymbol{\gamma}}_{p(j)})$ and set $\hat{\boldsymbol{u}}_{p(j)} = (\hat{u}_{(j)1,p}, \ldots, \hat{u}_{(j)n_j,p})^\top$. After obtaining the u-data of the jth component, the best selection of $\mathcal{V}_j$ would be the true structure selection, but the total number of vine tree structures on d variables is $\frac{d!}{2} \cdot 2^{\binom{d-2}{2}}$ (Morales-Nápoles 2010). If d is small, it is possible to enumerate all possible structures. However, it is usually not a feasible approach, even with

small dimensions, as one also needs to select pair-copula families for each scenario. Therefore, a greedy algorithm proposed for vine tree structure and pair-copula family selections might be used for each component like the greedy algorithm of Dißmann et al. (2013). Briefly, it works as follows:

- **Vine tree structure selection** $\mathcal{V}_j$: For $j = 1, \ldots, k$, it proceeds sequentially tree by tree, starting from tree one, and finds the maximum spanning tree at each tree among all edges allowed by proximity. Edge weight is the absolute empirical Kendall's τ value between the pair of nodes forming the edge.

- **Pair-copula family selection** $\mathcal{B}_j(\mathcal{V}_j)$: For $j = 1, \ldots, k$, after learning the vine tree structure, pair-copula families of the given structure are also estimated sequentially tree by tree. For a parametric pair-copula $C_{(j)e_a,e_b;D_e}$ associated with an edge e in $\mathcal{V}_j$ with the density $c_{(j)e_a,e_b;D_e}$ and parameters $\boldsymbol{\theta}_{(j)e_a,e_b;D_e}$, one first estimates the parameters that maximize the log-likelihood $\ell(\boldsymbol{\theta}_{(j)e_a,e_b;D_e})$. Later the copula family with the lowest Akaike Information Criterion (AIC) (Akaike 1998) is chosen:

$$AIC(\boldsymbol{\theta}_{(j)e_a,e_b;D_e}) = -2 \cdot \ell(\boldsymbol{\theta}_{(j)e_a,e_b;D_e}) + 2 \cdot |\boldsymbol{\theta}_{(j)e_a,e_b;D_e}| \text{ for } j = 1, \ldots k \text{ and } e \in \mathcal{V}_j, \quad (3.7)$$

  where $|\boldsymbol{\theta}_{(j)e_a,e_b;D_e}|$ denotes the number of copula parameters in $c_{(j)e_a,e_b;D_e}$. One does not need an alternative selection criterion like the BIC to induce model sparsity for the pair-copula family selection when the fitted pair-copula families have one or two parameters as later applied in Section 3.5.

3.4 Parameter estimation

Given that the marginal distributions, vine tree structure, and associated pair-copula families of each component are selected and known, another task in vine copula mixture models is to estimate the model parameters $\boldsymbol{\eta}$ in Equation (3.1). We remark that while selecting marginal distributions and pair-copula families, their parameters are also estimated. However, they are just the initial values. The optimal parameter estimates would be the values that maximize the log-likelihood of the given data defined in Equation (3.8):

$$\ell(\boldsymbol{\eta}) = \log \prod_{i=1}^{n} g(\boldsymbol{x}_i; \boldsymbol{\psi}) = \log \prod_{i=1}^{n} \sum_{j=1}^{k} \pi_j \cdot g_j(\boldsymbol{x}_i; \boldsymbol{\psi}_j). \quad (3.8)$$

Nevertheless, the true assignment of the observations to each component is unknown, and the parameter estimates would change depending on the component to which an observation belongs. As a solution to this problem, the expectation-maximization (EM) algorithm (Dempster et al. 1977) views the observations $\boldsymbol{x}_i = (x_{i,1}, \ldots, x_{i,d})^\top$ as incomplete and introduces

latent variables $\boldsymbol{z}_i = (z_{i,1}, \ldots, z_{i,k})^\top$, where each element $z_{i,j}$ is a binary variable defined as

$$z_{i,j} = \begin{cases} 1, & \text{if } \boldsymbol{x}_i \text{ belongs to the } j\text{th component,} \\ 0, & \text{otherwise,} \end{cases} \tag{3.9}$$

and $\sum_{j=1}^{k} z_{i,j} = 1$. The random vector $\boldsymbol{Z}_i$ corresponding to $\boldsymbol{z}_i$ follows a multinomial distribution with one trial and probabilities $\pi_1, \ldots, \pi_k$, that is, $\boldsymbol{Z}_i \sim Mult\big(1, \boldsymbol{\pi} = (\pi_1, \ldots, \pi_k)\big)$. Then we can write the *complete data log-likelihood* $\ell_c(\boldsymbol{\eta}; \boldsymbol{z}, \boldsymbol{x})$ of the complete data $\boldsymbol{y}_i = (\boldsymbol{x}_i, \boldsymbol{z}_i)^\top$ from Equation (3.1) as

$$\ell_c(\boldsymbol{\eta}; \boldsymbol{z}, \boldsymbol{x}) = \log \prod_{i=1}^{n} \prod_{j=1}^{k} [\pi_j \cdot g_j(\boldsymbol{x}_i; \boldsymbol{\psi}_j)]^{z_{i,j}} = \sum_{i=1}^{n} \sum_{j=1}^{k} z_{i,j} \cdot \log\, \pi_j + \sum_{i=1}^{n} \sum_{j=1}^{k} z_{i,j} \cdot \log\, g_j(\boldsymbol{x}_i; \boldsymbol{\psi}_j), \tag{3.10}$$

where $g_j(\boldsymbol{x}_i; \boldsymbol{\psi}_j)$ is given in Equation (2.4). Hence, we can write:

$$\begin{aligned} \ell_c(\boldsymbol{\eta}; \boldsymbol{z}, \boldsymbol{x}) = & \sum_{i=1}^{n} \sum_{j=1}^{k} [z_{i,j} \cdot \log\, \pi_j] + \sum_{i=1}^{n} \sum_{j=1}^{k} \sum_{p=1}^{d} [z_{i,j} \cdot \log\, f_{p(j)}(x_{i,p}; \boldsymbol{\gamma}_{p(j)})] \\ & + \sum_{i=1}^{n} \sum_{j=1}^{k} \sum_{m=1}^{d-1} \sum_{e \in E_{(j)m}} \big[z_{i,j} \cdot \log\, c_{(j)e_a, e_b; \boldsymbol{D_e}} \big(F_{(j)e_a|\boldsymbol{D_e}}(x_{i,e_a} | \boldsymbol{x}_{i,\boldsymbol{D_e}}; \boldsymbol{\gamma}_{(j)e_a|\boldsymbol{D_e}}, \boldsymbol{\theta}_{(j)e_a|\boldsymbol{D_e}}), \\ & F_{(j)e_b|\boldsymbol{D_e}}\big(x_{i,e_b} | \boldsymbol{x}_{i,\boldsymbol{D_e}}; \boldsymbol{\gamma}_{(j)e_b|\boldsymbol{D_e}}, \boldsymbol{\theta}_{(j)e_b|\boldsymbol{D_e}}); \boldsymbol{\theta}_{(j)e_a, e_b; \boldsymbol{D_e}} \big) \big]. \end{aligned} \tag{3.11}$$

Note that we use $\boldsymbol{\theta}_j$ for the pair-copula parameters instead of $\boldsymbol{\theta}_j(\mathcal{B}_j(\mathcal{V}_j))$ and $\boldsymbol{\gamma}_j$ for the marginal parameters instead of $\boldsymbol{\gamma}_j(\mathcal{F}_j)$ to simplify notation.

The EM algorithm alternates the E and M steps, increasing the data log-likelihood at each iteration (Dempster et al. 1977). The E-step requires calculating the conditional expectation of the complete-data log likelihood, given the observed data and current parameter estimates. The M-step maximizes the expected complete data log-likelihood from the E-step over all parameters. We need to estimate marginal parameters $\boldsymbol{\gamma}_j$, pair-copula parameters $\boldsymbol{\theta}_j$, and mixture weight π_j of the jth component. Since their joint estimation is not tractable and efficient, we use the expectation conditional maximum (ECM) algorithm (Meng and Rubin 1993). Here, the M-step in the EM is replaced by three lower dimensional maximization problems called CM-steps. The vine tree structure, associated pair-copula families, and marginal distributions remain fixed at the given choice for each component in the ECM iterations. Then the $(t+1)$th iteration steps are

1. **E-step** *(Posterior probabilities)*: Calculate the posterior probability that an observation $\boldsymbol{x}_i$ belongs to the jth component given the current values of the model parameters $\pi_j^{(t)}$

and $\boldsymbol{\psi}_j^{(t)} = (\boldsymbol{\gamma}_j^{(t)}, \boldsymbol{\theta}_j^{(t)})^\top$:

$$r_{i,j}^{(t+1)} = \frac{\pi_j^{(t)} g_j(\boldsymbol{x}_i; \boldsymbol{\psi}_j^{(t)})}{\sum_{j'=1}^{k} \pi_{j'}^{(t)} g_{j'}(\boldsymbol{x}_i; \boldsymbol{\psi}_{j'}^{(t)})} \quad \text{for} \quad i = 1, \dots n \quad \text{and} \quad j = 1, \dots k. \tag{3.12}$$

2. **CM-step 1** *(Mixture weights)*: Maximize $\ell_c(\boldsymbol{\eta}; \boldsymbol{z}, \boldsymbol{x})$ over π_j given the updated posterior probabilities $r_{i,j}^{(t+1)}$:

$$\pi_j^{(t+1)} = \arg\max_{\pi_j} \sum_{i=1}^{n} r_{i,j}^{(t+1)} \cdot \log \pi_j \quad \text{for} \quad j = 1, \dots k. \tag{3.13}$$

A closed form solution exists and is given by

$$\pi_j^{(t+1)} = \frac{\sum_{i=1}^{n} r_{i,j}^{(t+1)}}{n} \quad \text{for} \quad j = 1, \dots k. \tag{3.14}$$

3. **CM-step 2** *(Marginal parameters)*: Optimal marginal parameter estimates $\boldsymbol{\gamma}_j^*$ of the jth component would maximize $\ell_c(\boldsymbol{\eta}; \boldsymbol{z}, \boldsymbol{x})$ over $\boldsymbol{\gamma}_j$ given the current values of the pair-copula parameters $\boldsymbol{\theta}_j^{(t)}$ and updated posterior probabilities $r_{i,j}^{(t+1)}$:

$$\boldsymbol{\gamma}_j^* = \arg\max_{\boldsymbol{\gamma}_j} \sum_{i=1}^{n} r_{i,j}^{(t+1)} \cdot \log\, g_j(\boldsymbol{x}_i; \boldsymbol{\gamma}_j, \boldsymbol{\theta}_j^{(t)}) \quad \text{for} \quad j = 1, \dots k. \tag{3.15}$$

However, since a closed form solution does not exist, we numerically maximize $\ell_c(\boldsymbol{\eta}; \boldsymbol{z}, \boldsymbol{x})$ over $\boldsymbol{\gamma}_j$

$$\max_{\boldsymbol{\gamma}_j} \sum_{i=1}^{n} r_{i,j}^{(t+1)} \cdot \log\, g_j(\boldsymbol{x}_i; \boldsymbol{\gamma}_j, \boldsymbol{\theta}_j^{(t)}) \quad \text{for} \quad j = 1, \dots k \tag{3.16}$$

to find the updated values of the marginal parameters $\boldsymbol{\gamma}_j^{(t+1)}$.

4. **CM-step 3** *(Pair-copula parameters)*: Again, a closed form solution that maximizes $\ell_c(\boldsymbol{\eta}; \boldsymbol{z}, \boldsymbol{x})$ over $\boldsymbol{\theta}_j$ given the current values of the marginal parameters $\boldsymbol{\gamma}_j^{(t+1)}$ and updated posterior probabilities $r_{i,j}^{(t+1)}$ does not exist. Hence, we numerically maximize $\ell_c(\boldsymbol{\eta}; \boldsymbol{z}, \boldsymbol{x})$ over $\boldsymbol{\theta}_j$

$$\max_{\boldsymbol{\theta}_j} \sum_{i=1}^{n} r_{i,j}^{(t+1)} \cdot \log\, g_j(\boldsymbol{x}_i; \boldsymbol{\gamma}_j^{(t+1)}, \boldsymbol{\theta}_j) \quad \text{for} \quad j = 1, \dots k \tag{3.17}$$

to find the updated values of the pair-copula parameters $\boldsymbol{\theta}_j^{(t+1)}$. In the case of a d-dimensional vine copula with parametric pair-copulas, the total number of pair-copula parameters to be estimated grows quadratically in dimension d. However, truncating a vine tree structure at tree level 1, i.e., obtaining a Markov tree, reduces the total number of parameter estimates linear in dimension d, one of the main motivations for the third step's formulation with Markov trees in our Algorithm in Section 3.5.

Starting values: For the ECM algorithm to run, we require initial parameters $\boldsymbol{\eta}_j^{(0)}$ of the jth component for $j = 1, \ldots, k$. Since the marginal distributions, vine tree structure, and associated pair-copula families of each component stay fixed in the ECM iterations, we, additionally, need to select them. One method is to select an initial partition in advance. A more general alternative is to use quick clustering algorithms with possible weights of observations $\boldsymbol{x}_i = (x_{i,1}, \ldots, x_{i,d})^\top$ for $i = 1, \ldots, n$ in different components to have a starting partition. Assume that the total number of observations assigned to the jth component at the 0th iteration is $n_j^{(0)}$, and the observations belonging to the jth component at the 0th iteration are given by $\boldsymbol{x}_{(j)i_j}^{(0)} = (x_{(j)i_j,1}^{(0)}, \ldots, x_{(j)i_j,d}^{(0)})^\top$ for $i_j = 1, \ldots, n_j^{(0)}$ and $j = 1, \ldots, k$. It holds that $\sum_{j=1}^{k} n_j^{(0)} = n$ and $\bigcup_{\forall(j,i_j)} \boldsymbol{x}_{(j)i_j}^{(0)} = \bigcup_{\forall i} \boldsymbol{x}_i$. After specifying an initial set for parametric marginal distributions, vine tree structures and associated parametric pair-copula families, the starting values can be obtained as follows:

1. Initial marginal distributions $F_{p(j)}^{(0)}$ and marginal parameters $\boldsymbol{\gamma}_{p(j)}^{(0)}$: For $j = 1, \ldots, k$ and $p = 1, \ldots, d$, the marginal parameters maximize the log-likelihood of a variable $\boldsymbol{x}_{p(j)}^{(0)} = (x_{(j)1,p}^{(0)}, \ldots, x_{(j)n_j^{(0)},p}^{(0)})^\top$ given in Equation (3.5), then the marginal distribution with the lowest BIC given in Equation (3.6) is chosen as described in Section 3.3.

2. Initial vine tree structure $\mathcal{V}_j^{(0)}$, pair-copula families $\mathcal{V}_j^{(0)}(\mathcal{B}_j^{(0)})$ and its parameters $\boldsymbol{\theta}_{(j)}^{(0)}$: For $j = 1, \ldots, k$ and $p = 1, \ldots, d$, first, the cumulative distribution function of the chosen marginal distribution $F_{p(j)}^{(0)}$ with parameters $\boldsymbol{\gamma}_{p(j)}^{(0)}$ on the variable $\boldsymbol{x}_{p(j)}^{(0)}$ is fitted to obtain u-data of the jth component, $\boldsymbol{u}_{p(j)}^{(0)} = F_{p(j)}^{(0)}(\boldsymbol{x}_{p(j)}^{(0)}; \boldsymbol{\gamma}_{p(j)}^{(0)})$. The estimated u-data is then used to select a vine or Markov tree and the associated pair-copula families with their parameters as discussed in Section 3.3.

3. Initial mixture weights $\pi_{(j)}^{(0)}$: For $j = 1, \ldots, k$, they are proportional to the total number of observations belonging to the jth component and given by $\pi_{(j)}^{(0)} = \frac{n_j^{(0)}}{n}$.

The ECM is sensitive to the starting values, and a poor choice of them may result in convergence to a local maximum as an EM-type algorithm (Karlis and Xekalaki 2003). However, our starting values optimize an associated model selection criterion based on the data log-likelihood. Even though there is no guarantee that the vine copula mixture model initialized

by a fast clustering algorithm will select the correct model components, and its ECM iterations will converge to the global optimum, we show promising results of our current setup and discuss modifications with simulation studies in Section 3.6.

Stopping condition: A stopping criterion terminates the ECM algorithm when the relative difference in the data log-likelihood between two successive iterations is less than the desired tolerance as follows:

$$\frac{\ell(\boldsymbol{\eta}^{(t+1)}) - \ell(\boldsymbol{\eta}^{(t)})}{\ell(\boldsymbol{\eta}^{(t)})} < tol \quad \text{for} \quad t = 1, \ldots, (s-1). \tag{3.18}$$

We use the tolerance level (tol) 0.00001 in our simulation studies in Section 3.6 and real data analysis in Section 3.8.

3.5 Model-based clustering algorithm: VCMM

After discussing the vine copula mixture model formulation, model selection and parameter estimation problems, we will formulate a model-based clustering algorithm (VCMM) with vine copula mixture models. We provide our pseudo-code in Algorithm 1, which consists of six primary building blocks. We implement these blocks as a `R` software package, called `vineclust` (Sahin 2021), explained in Section 3.7.

The first step (initial clustering assignment via a fast clustering algorithm) partitions observations $\boldsymbol{x}_i = (x_{i,1}, \ldots, x_{i,d})^\top$ for $i = 1, \ldots, n$ into k components (clusters) using a quick clustering algorithm. Our analyses will mostly use the distance-based clustering, k-means (Hartigan and Wong 1979) algorithm, with default specifications used in the package `stats`. Alternative clustering algorithms or partitions could be specified based on the analyzed data set. We will present an exemplary scenario in Section 3.6. For clustering algorithms, which are sensitive to the variables' scale, like k-means, one can apply a standardization for each variable; for instance, using the function `scale`. Since these algorithms might depend on the seed, we will present a real data analysis in Section 3.7 to guide choosing a good option.

As a second step, we select an initial VCMM model. The marginal distributions are determined among a candidate set of univariate parametric distributions by a model selection criteria. As discussed in Section 3.3, the parametric families in the candidate set could be chosen after initial graphical data analysis as given in the top left of Figure 3.1. Using the chosen parametric marginal distributions, we obtain the u-data. Then we first truncate a vine tree structure at tree level one for the initial selection of vine copula models, thereby obtaining a Markov tree model. As discussed in Section 3.4, working with Markov trees allows us to decrease the optimization problem's size (CM-step 3) from quadratic to linear in dimension d, thereby reducing our computational effort. However, the performance of using different truncation levels might be investigated in the future. A wide range of parametric pair-copula families is applicable. For this part, our software is based on the package `VineCopula` and mainly its function `RVineStructureSelect` (Nagler et al. 2021). The VCMM selects

$d \cdot k$ marginal distributions, k Markov tree structures, and $(d-1) \cdot k$ pair-copula families at this initial model selection step.

The third step (parameter estimation via the ECM algorithm allowing for Markov (vine) tree structures) updates the VCMM parameters while keeping the marginal distributions, Markov tree structures, and pair-copula families fixed until a stopping condition holds. For each cluster, its mixture weight, pair-copula parameters, and marginal parameters are updated per iteration. The total number of updated parameters per iteration is the sum of the total number of pair-copula parameters, marginal parameters, and mixture weights over k clusters. Assume the ECM stops after s iterations. The total number of updated parameters up to this point is linear in dimension d, the total number of clusters k, and iterations s.

Next, we partition the observations into k clusters with the updated posterior probabilities as a temporary clustering assignment. An observation is a member of the cluster, where its posterior probability is highest.

The marginal distributions and dependence structure of the clusters can change due to the successive ECM steps. Moreover, dependencies can exist in higher tree levels, and accounting for those can improve model power. Thus, we perform a final model selection, and it is based on a full vine specification. The fifth step estimates the model components, including all possible vine tree levels and their parameters, with the fourth step's clustering assignment. The VCMM chooses $d \cdot k$ marginal distributions, k vine tree structures, and $\frac{d \cdot (d-1)}{2} \cdot k$ pair-copula families. The total number of updated parameters with this final step is linear in the total number of clusters k and iterations s, as in the third step, but is quadratic in dimension d due to the estimation of all possible vine tree levels. Using a full vine specification introduces additional parameter estimates; however, it increases model power as shown in Section 3.7.

The last step (final clustering assignment based on the full vine specification) assigns the observations to the clusters with the final model's posterior probabilities.

Algorithm 1 Vine copula mixture models clustering: VCMM

Input: d-dimensional n observations to cluster $\boldsymbol{x}_i = (x_{i,1}, \ldots, x_{i,d})^\top \in \mathbb{R}^d$ for $i = 1, \ldots, n$ and total number of clusters k.
Output: A clustering partition of the observations $\mathcal{C} = \{\mathcal{C}_1, \ldots, \mathcal{C}_k\}$, estimated model components and parameters of the jth cluster $\hat{\mathcal{F}}_j$, $\hat{\boldsymbol{\gamma}}_j$, $\hat{\mathcal{V}}_j$, $\hat{\mathcal{B}}_j(\hat{\mathcal{V}}_j)$, $\hat{\boldsymbol{\theta}}_j\big(\hat{\mathcal{B}}_j(\hat{\mathcal{V}}_j)\big)$, $\hat{\pi}_j$ and final posterior probabilities $r_{i,j}^{(final)}$ for $i = 1, \ldots, n$ and $j = 1, \ldots, k$.
for $j = 1, \ldots, k$ **do**

Step I: Initial clustering assignment via a fast clustering algorithm

$\boldsymbol{x}_{(j)i_j}^{(0)} \leftarrow (x_{(j)i_j,1}^{(0)}, \ldots, x_{(j)i_j,d}^{(0)})^\top$ for $i_j = 1, \ldots, n_j^{(0)}$, $\sum_{j=1}^{k} n_j^{(0)} = n$ and $\bigcup_{\forall(j,i_j)} \boldsymbol{x}_{(j)i_j}^{(0)} = \bigcup_{\forall i} \boldsymbol{x}_i$.

Step II: Initial model selection based on the Step I assignment

for $p = 1, \ldots, d$ **do**

$\boldsymbol{x}_{p(j)}^{(0)} \leftarrow (x_{(j)1,p}^{(0)}, \ldots, x_{(j)n_j^{(0)},p}^{(0)})^\top$,

$(F_{p(j)}^{(0)}, \boldsymbol{\gamma}_{p(j)}^{(0)}) \leftarrow \underset{mar,\boldsymbol{\gamma}}{\arg\min} \; -2 \cdot \ell^{mar}(\boldsymbol{\gamma}; \boldsymbol{x}_{p(j)}^{(0)}) + |\boldsymbol{\gamma}| \cdot \log(n_j^{(0)})$, where ℓ^{mar} is the log-likelihood of the univariate distribution mar for $mar \in$ {candidate univariate parametric marginal distributions},

$\boldsymbol{u}_{p(j)}^{(0)} \leftarrow F_{p(j)}^{(0)}(\boldsymbol{x}_{p(j)}^{(0)}; \boldsymbol{\gamma}_{p(j)}^{(0)})$,

end for

$\boldsymbol{u}_{(j)}^{(0)} \leftarrow (\boldsymbol{u}_{1(j)}^{(0)}, \ldots, \boldsymbol{u}_{d(j)}^{(0)})^\top$,

$\mathcal{F}_j^{(0)} \leftarrow \{F_{1(j)}^{(0)}, \ldots, F_{d(j)}^{(0)}\}$,

bicop $\in$ {candidate parametric pair-copula families},

$\pi_j^{(0)} \leftarrow \frac{n_j^{(0)}}{n}$,

$model_{markov} \leftarrow$ RVineStructureSelect ($\boldsymbol{u}_{(j)}^{(0)}$, familyset=bicop, trunclevel=1), where RVineStructureSelect determines the vine tree structure truncated at tree level 1, associated pair-copula families among the set bicop and its parameters for the component j as described in Section 3.3,

$\mathcal{V}_j^{(0)} \leftarrow model_{markov}\$Matrix$, $\mathcal{B}_j^{(0)} \leftarrow model_{markov}\$Family$, $\boldsymbol{\theta}_j^{(0)} \leftarrow model_{markov}\$pars$.

end for

Step III: Parameter estimation via the ECM algorithm allowing for Markov (vine) tree structures

while stop=FALSE **do**

$t \leftarrow 0$,

for $j = 1, \ldots, k$ **do**

$\mathcal{F}_j^{(t+1)} \leftarrow \mathcal{F}_j^{(0)}$, $\mathcal{V}_j^{(t+1)} \leftarrow \mathcal{V}_j^{(0)}$, $\mathcal{B}_j^{(t+1)} \leftarrow \mathcal{B}_j^{(0)}$,

Update $r_{i,j}^{(t+1)}$ as in Equation (3.12) for $i = 1, \ldots, n$,

Update $\pi_j^{(t+1)}$, $\boldsymbol{\gamma}_j^{(t+1)}$ and $\boldsymbol{\theta}_j^{(t+1)}$ sequentially as in Equations (3.14), (3.16) and (3.17), respectively,

end for

$t \leftarrow t + 1$,

if the termination criterion like in Equation (3.18) holds **then**

$s \leftarrow t$, stop=TRUE and **break**.

end if

end while

Step IV: Temporary clustering assignment based on the Markov (vine) tree structure

$\boldsymbol{x}_i \in \mathcal{C}_{j^*} \iff j^* = \underset{j=1,\ldots,k}{\arg\max}\, r_{i,j}^{(s)}$ for $i = 1, \ldots, n$.

for $j = 1, \ldots, k$ **do**

$\boldsymbol{x}_{(j)i_j}^{(s)} \leftarrow (x_{(j)i_j,1}^{(s)}, \ldots, x_{(j)i_j,d}^{(s)})^\top$ for $i_j = 1, \ldots, n_j^{(s)}$, $\sum_{j=1}^{k} n_j^{(s)} = n$, $\bigcup_{\forall(j,i_j)} \boldsymbol{x}_{(j)i_j}^{(s)} = \bigcup_{\forall i} \boldsymbol{x}_i$.

Step V: Final model selection based on a full vine specification

Perform lines 7– 12 with the new assignment $\boldsymbol{x}_{(j)i_j}^{(s)}$ and change the iteration index from (0) to (s),

$m_{VCMM} \leftarrow$ RVineStructureSelect ($\boldsymbol{u}_{(j)}^{(s)}$, familyset=bicop, trunclevel=d-1),

$\mathcal{V}_j^{(s)} \leftarrow m_{VCMM}\$Matrix$, $\mathcal{B}_j^{(s)} \leftarrow m_{VCMM}\$Family$ and $\boldsymbol{\theta}_j^{(s)} \leftarrow m_{VCMM}\$pars$.

end for

Step VI: Final clustering assignment based on the full vine specification

$\boldsymbol{x}_i \in \mathcal{C}_{j^*} \iff j^* = \underset{j=1,\ldots,k}{\arg\max}\, r_{i,j}^{(s+1)}$ for $i = 1, \ldots, n$, where $r_{i,j}^{(s+1)}$ are the posterior probabilities calculated from the final model,

$\hat{\mathcal{F}}_j \leftarrow \{F_{1(j)}^{(s)}, \ldots, F_{d(j)}^{(s)}\}$, $\hat{\boldsymbol{\gamma}}_j \leftarrow \{\boldsymbol{\gamma}_{1(j)}^{(s)}, \ldots, \boldsymbol{\gamma}_{d(j)}^{(s)}\}$, $\hat{\mathcal{V}}_j \leftarrow \mathcal{V}_j^{(s)}$, $\hat{\mathcal{B}}_j(\hat{\mathcal{V}}_j) \leftarrow \mathcal{B}_j^{(s)}$, $\hat{\boldsymbol{\theta}}_j\big(\hat{\mathcal{B}}_j(\hat{\mathcal{V}}_j)\big) \leftarrow \boldsymbol{\theta}_j^{(s)}$, $\hat{\pi}_j \leftarrow \pi_j^{(s)}$, $r_{i,j}^{(final)} \leftarrow r_{i,j}^{(s+1)}$.

3.6 Simulation studies

This section will demonstrate the remarkable and promising results of our clustering algorithm, VCMM, using simulated data. We compare its performance with the initial partition from k-means to some well-known model-based clustering algorithms: the mixture of multivariate normal, skew normal, t, and skew t distributions, i.e., we perform various benchmarking analyses in clustering. We fit the mixture of multivariate normal distributions using the package `mclust` and the others using the package `mixsmsn`. The latter fits scale mixtures of skew normal distributions and works with the initial partition of k-means like our algorithm. Therefore, the comparison of the performance of the VCMM and the chosen competing algorithms is fairer. We work with the default specifications of the packages but specify the total number of clusters and seed. For the clustering performance evaluation, we use the BIC value and misclassification rate when the true labels are available. The lower the BIC value or misclassification rate, the better the clustering assignment.

The BIC criterion compares only the models studied; thus, it selects a better model among the evaluated ones. However, it does not imply that it selects the best model. Since unsupervised learning problems do not contain the true labels of the observations, a separation of the data in training and test sets is not feasible. Therefore, the misclassification rate we consider can be regarded as the in-sample misclassification rate of supervised learning.

Our candidate set for margins is normal, Student's t with degrees of freedom 3, logistic, log-normal, log-logistic, and gamma distribution whose abbreviations are listed in Table 3.1. Thus, we allow for a heavy-tailed marginal distribution. Our pair-copula families are parametric: Gaussian (N), t (t), Clayton (C), Gumbel (G), Frank (F), Joe (J), BB1, BB6, and BB8 copula. Since we apply their possible $90°$, $270°$, and $180°$ (S) rotations, the total number of pair-copula families utilized is 27. Chapter 3 of Czado (2019) provides more details about them. Thus, we have a flexibility to capture different dependence structures in data. As a stopping criterion given in Equation (3.18), we use the tolerance level (tol) 0.00001.

Table 3.1: Abbreviation for univariate marginal distributions used in Chapter 3.

$lnorm(\mu, \sigma)$	log-normal distribution with mean parameter μ, standard deviation parameter σ on the logarithmic scale,
$\exp(\lambda)$	exponential distribution with rate parameter λ,
$llogis(\alpha, \beta)$	log-logistic distribution with shape parameter α, scale parameter β,
$logis(l, s)$	logistic distribution with location parameter l, scale parameter s,
$\Gamma(\alpha, \beta)$	gamma distribution with shape parameter α, rate parameter β,
$\mathcal{N}(\mu, \sigma)$	normal distribution with mean parameter μ, standard deviation σ,
$t_3(\mu, \sigma)$	Student's t distribution with mean parameter μ, standard deviation parameter σ, degrees of freedom 3.

In the first two scenarios, the simulated data has three variables, two clusters, and either 100 or 500 observations in each cluster, and we replicate their data generating process 100 times. Now $X_{p(1)} = F^{-1}_{p(1)}(U_{p(1)}; \boldsymbol{\gamma}_{p(1)})$ and $X_{p(2)} = F^{-1}_{p(2)}(U_{p(2)}; \boldsymbol{\gamma}_{p(2)})$ for $p = 1, 2, 3$ give the variables of the first and second clusters, respectively. We simulate $U_{p(1)}$ and $U_{p(2)}$ from a specified vine copula model and specify $F_{p(1)}, F_{p(2)}, \boldsymbol{\gamma}_{p(1)}, \boldsymbol{\gamma}_{p(2)}$. A mixture of vine copulas generates the first scenario with one/two parameter pair-copula families and non-Gaussian margins. The second scenario simulates data from a mixture of vine copulas with single parameter pair-copulas and Gaussian/non-Gaussian margins. In the first scenario, we aim to analyze how well the VCMM improves the clustering compared to its starting partition from k-means. In the second scenario, we aim to analyze how well the VCMM and other model-based clustering algorithms capture different dependence structures and shapes hidden in the multivariate data with different numbers of observations.

In the third scenario, we discuss the effect of different initial clustering techniques at Step I in Algorithm 1 on the VCMM using the artificial data from a mixture of vine copulas with Gaussian copulas and Gaussian margins. In the last scenario, we aim to analyze the performance of the VCMM when the data generating process is misspecified, e.g., we simulate the data from a mixture of multivariate skew t distributions.

The mixture of vine copulas with one/two parameter pair-copulas and non-Gaussian margins

We simulate data $U_{p(1)}$ and $U_{p(2)}$ for $p = 1, 2, 3$ from a vine copula model, where pair-copula families have either single or two parameters, as shown in Table 3.2. Both clusters include positive as well as high, medium, and low strength dependencies. Then we transform the data from u-scale $U_{p(1)}, U_{p(2)}$ to x-scale $X_{p(1)}, X_{p(2)}$. Table 3.2 presents the marginal distributions with the parameters. The clusters are non-Gaussian.

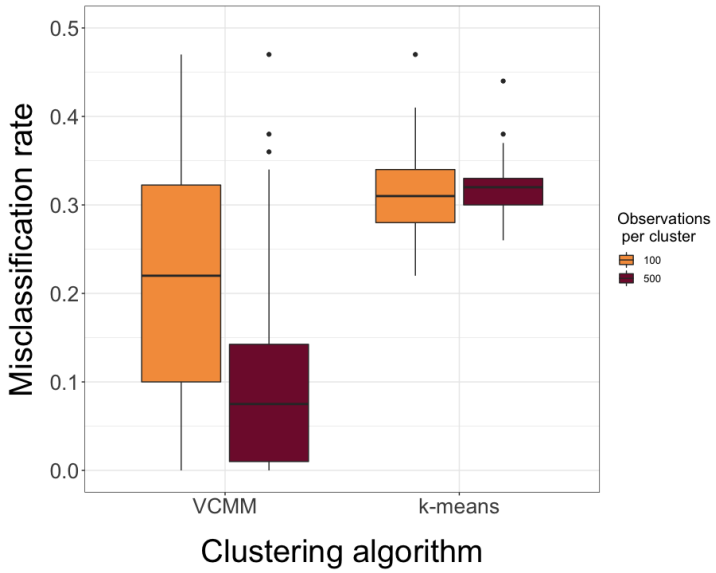
Table 3.2: The pair-copula families and univariate marginal distributions with true parameter values of each cluster in the first simulation scenario.

Pair-copula(parameters/Kendall's τ)		Marginal distribution(parameters)	
$C_{(1)1,2}$	BB1(3,2/0.8)	$F_{1(1)}(\boldsymbol{\gamma}_{1(1)})$	$llogis(1.5, 1.25)$
$C_{(1)2,3}$	F(8.0/0.6)	$F_{2(1)}(\gamma_{2(1)})$	$\exp(0.1)$
$C_{(1)1,3;2}$	G(1.3/0.2)	$F_{3(1)}(\gamma_{3(1)})$	$lnorm(0.1, 1.3)$
$C_{(2)1,3}$	C(4.7/0.7)	$F_{1(2)}(\boldsymbol{\gamma}_{1(2)})$	$lnorm(2.5, 0.5)$
$C_{(2)2,3}$	BB1(2,1/0.5)	$F_{2(2)}(\boldsymbol{\gamma}_{2(2)})$	$logis(5, 3)$
$C_{(2)1,2;3}$	BB1(0.5,1/0.2)	$F_{3(2)}(\gamma_{3(2)})$	$\exp(0.05)$

We evaluate the clustering performance of the VCMM and k-means, visualizing the misclassification rate per simulation replication in box plots. For the larger number of observations (500 observations per cluster), Figure 3.2 shows that the VCMM provides a noticeably better

fit than k-means. On average, it improves the accuracy by 22% compared to its initial partition. For the small number of observations (100 observations per cluster), the VCMM usually has lower misclassification rates than k-means. Its average misclassification rate is 10% less than k-means. However, the VCMM variance in the accuracy increases as the number of observations gets lower. Moreover, the VCMM requires, on average, 25 and 17 ECM iterations for the large and small numbers of observations, respectively, i.e., the convergence takes more iterations with a larger data set in this simulation.

Figure 3.2: Comparison of the clustering performance of the VCMM and its initial partition algorithm k-means over 100 replications under the scenario specified in Table 3.2 for the different number of observations per cluster.



The mixture of vine copulas with one parameter pair-copulas and Gaussian/non-Gaussian margins

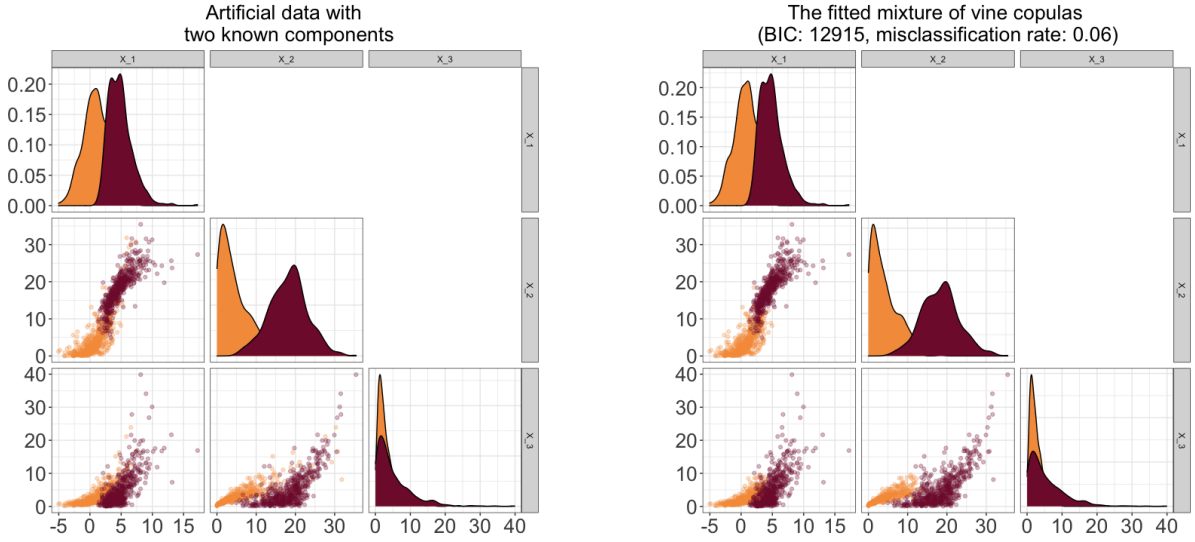
In this scenario, we work with single parameter pair-copulas, add Gaussian margins within the clusters and simulate data on u-scale $U_{p(1)}$, $U_{p(2)}$ for $p = 1, 2, 3$ as specified in Table 3.3. Both clusters share the same vine tree structure and include medium sized asymmetric tail dependencies. In contrast, the second cluster has a symmetric, non-Gaussian dependency between $U_{1(2)}$ and $U_{2(2)}$. The next step is to obtain the data on the x-scale $X_{p(1)}, X_{p(2)}$ as described before. Marginal distributions and the parameters used are listed in Table 3.3.

Table 3.3: The pair-copula families and univariate marginal distributions with true parameter values of each cluster in the second simulation scenario.

Pair-copula(parameters/Kendall's τ)		Marginal distribution(parameters)	
$C_{(1)1,2}$	G(2.5/0.6)	$F_{1(1)}(\gamma_{1(1)})$	$\mathcal{N}(1,2)$
$C_{(1)2,3}$	SG(5.0/0.8)	$F_{2(1)}(\gamma_{2(1)})$	$\exp(0.2)$
$C_{(1)1,3;2}$	C(0.9/0.3)	$F_{3(1)}(\gamma_{3(1)})$	$lnorm(0.8, 0.8)$
$C_{(2)1,2}$	F(11.4/0.7)	$F_{1(2)}(\gamma_{1(2)})$	$lnorm(1.5, 0.4)$
$C_{(2)2,3}$	SC(2.0/0.5)	$F_{2(2)}(\gamma_{2(2)})$	$\mathcal{N}(18,5)$
$C_{(2)1,3;2}$	J(1.4/0.2)	$F_{3(2)}(\gamma_{3(2)})$	$\exp(0.2)$

We illustrate the simulated data on the x-scale in the left panel of Figure 3.3. The generated clusters are non-elliptical. The fitted VCMM detects the true shape of the clusters as seen in the right panel of Figure 3.3 as opposed to other algorithms given in Figure 3.1.

Figure 3.3: Pairwise scatter plot of simulated data (500 observations per cluster) on x-scale under the scenario specified in Table 3.3 (left). The right plot shows the fitted VCMM. The orange and red points show the observations of one cluster and the other cluster, respectively. The diagonal of the plots shows each cluster's associated variable's marginal density function.



Figures 3.4 and 3.5 visualize the misclassification rate and BIC value per simulation replication in box plots for 100 and 500 observations per cluster to compare the algorithms' clustering performance. For 500 observations per cluster, the VCMM is superior to other model-based clustering algorithms regarding both the misclassification rate and the BIC value. It separates two non-elliptical clusters adequately as expected, whereas others tend to find it challenging. Even though they cannot model two clusters as the VCMM does, the mixtures of multivariate

skew distributions show a better fit than elliptical distributions in terms of the BIC value. The mean misclassification rate of the mixtures of multivariate skew t distributions and that of the VCMM are 4% and 8% less than that of multivariate normal distributions.

Figure 3.4 shows that the misclassification rates of the VCMM do not change enormously for 100 observations per cluster and are still lower than the others. Its variance in the accuracy increases with the smaller number of observations as in the first simulation scenario. The BIC values favor the VCMM, and after the VCMM, the mixtures of multivariate skew distributions provide better fits than the others as shown in Figure 3.5.

Figure 3.4: Comparison of the model-based clustering algorithms' misclassification rate over 100 replications with 100 and 500 observations per cluster under the scenario specified in Table 3.3.

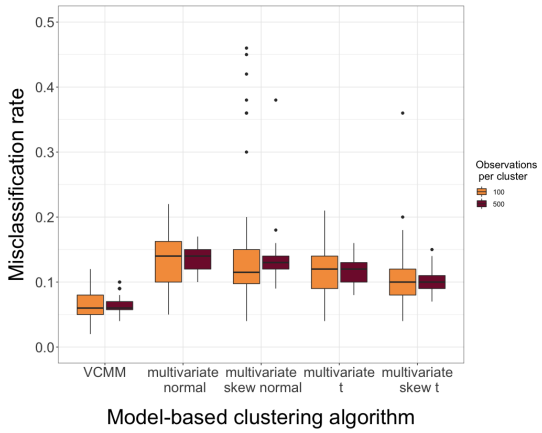
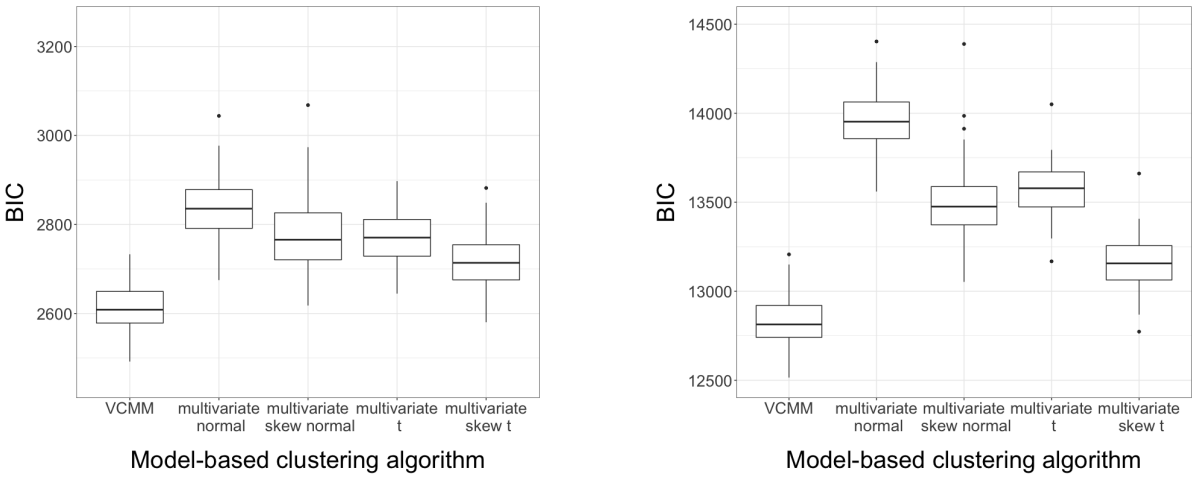


Figure 3.5: Comparison of the model-based clustering algorithms' BIC over 100 replications with 100 (left) and 500 (right) observations per cluster under the scenario given in Table 3.3.



The mixture of vine copulas with Gaussian copulas and Gaussian margins: significant overlaps

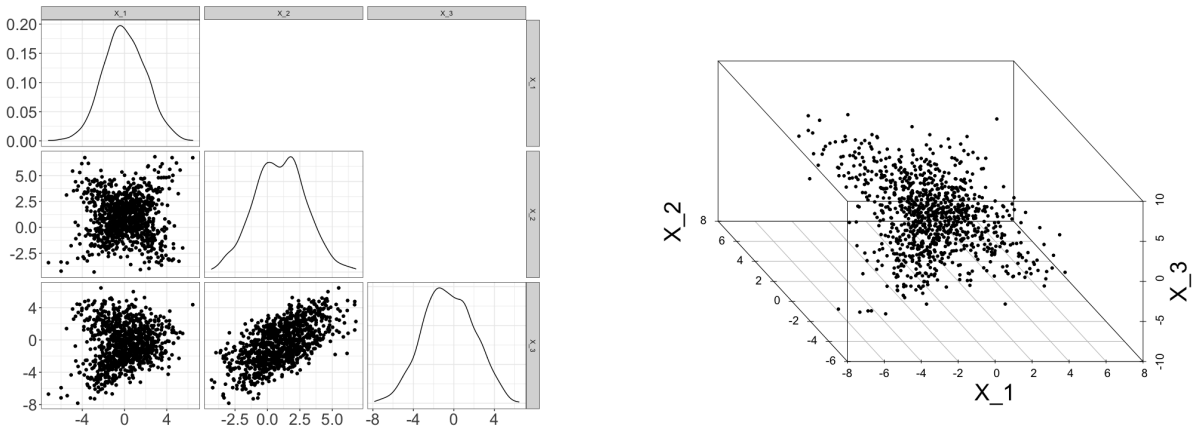
We generate artificial data from a mixture of two vine copulas as specified in Table 3.4. They have Gaussian copulas between the pairs of variables, and each margin follows a univariate normal distribution, i.e., the scenario represents a mixture of multivariate normal distributions.

Table 3.4: The pair-copula families and univariate marginal distributions with true parameter values of each cluster in the third simulation scenario.

Pair-copula(parameters/Kendall's τ)		Marginal distribution(parameters)	
$C_{(1)1,2}$	N(0.7/0.5)	$F_{1(1)}(\gamma_{1(1)})$	$\mathcal{N}(0,2)$
$C_{(1)2,3}$	N(0.8/0.6)	$F_{2(1)}(\gamma_{2(1)})$	$\mathcal{N}(1,2)$
$C_{(1)1,3;2}$	N(0.3/0.2)	$F_{3(1)}(\gamma_{3(1)})$	$\mathcal{N}(1,2)$
$C_{(2)1,2}$	N(-0.7/-0.5)	$F_{1(2)}(\gamma_{1(2)})$	$\mathcal{N}(0,2)$
$C_{(2)2,3}$	N(0.8/0.6)	$F_{2(2)}(\gamma_{2(2)})$	$\mathcal{N}(1,2)$
$C_{(2)1,3;2}$	N(0.3/0.2)	$F_{3(2)}(\gamma_{3(2)})$	$\mathcal{N}(-2,2)$

We do not observe strong multimodality of variables in the diagonal of the left panel in Figure 3.6. However, the data has two obvious clusters with many overlaps, creating an X-shape in three dimensions, in the right panel of Figure 3.6.

Figure 3.6: Pairwise (left) and 3-dimensional scatter plot (right) of artificial data (500 observations per cluster) on x-scale under the scenario specified in Table 3.4. The diagonal of the plot on the left shows the corresponding variable's marginal density function.



In an attempt to determine the true groups, we fit the VCMM as in previous simulation studies. Since k-means assumes spherical shapes of clusters, we suspect that the VCMM's model selection using the initial partition from k-means could be satisfactory in such a scenario.

The VCMM results in 65% classification accuracy with the BIC of 12447. As we suggest at Step I in Algorithm 1, other partition strategies can be used before fitting the VCMM. For instance, we partition the data by the model-based hierarchical clustering using the function `hcVVV` with its default specifications in the package `mclust` and then run our algorithm. As a result, the VCMM has 95% classification accuracy with the BIC of 12039. Therefore, using different starting partitions in the VCMM and selecting a final model with the lowest BIC are suggested. Also, fitting the mixture of multivariate normal distributions gives 96% classification accuracy with the BIC of 11987.

The mixture of multivariate skew t distributions: misspecification

As a misspecification scenario, we simulate data in three dimensions with two components from the mixture of multivariate skew t distributions expressed as a class of skew normal independent distributions (Cabral et al. 2012). The total number of observations is 1000. Denoting the density of a multivariate skew t distribution by $ST(.;\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\lambda},\nu)$ with location vector $\boldsymbol{\mu}$, scale matrix $\boldsymbol{\Sigma}$, skewness vector $\boldsymbol{\lambda}$, and degrees of freedom ν, assume a random vector $\boldsymbol{X}$ has this mixture distribution. Then its density at $\boldsymbol{x}$ is given by

$$\pi_1 \cdot ST(\boldsymbol{x};\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1,\boldsymbol{\lambda}_1,\nu_1) + (1-\pi_1)\cdot ST(\boldsymbol{x};\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_2,\boldsymbol{\lambda}_2,\nu_2), \tag{3.19}$$

where the true values of the parameters used in the simulation are $\boldsymbol{\mu}_1 = (1,1,0)^\top$, $\boldsymbol{\mu}_2 = (-2,-2,-2)^\top$, $\pi_1 = 0.6$, $\boldsymbol{\lambda}_1 = (4,-4,4)^\top$, $\boldsymbol{\lambda}_2 = (-4,4,4)^\top$, $\nu_1 = 8$, $\nu_2 = 10$, and $\boldsymbol{\alpha}_1 = \boldsymbol{\alpha}_2 = (2,1,1,2,1,2)$ vectorizing the upper triangular matrix of symmetric matrices $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$. We replicate the data generating process 100 times. The mixture of multivariate skew t distributions fits well with the data as expected with an average misclassification rate of 0.06 and an average BIC of 10676. The mixture of multivariate normal distributions has more difficulty in capturing non-elliptical components than the VCMM. The former and latter have the average misclassification rate of 0.13 and 0.09 and the average BIC of 10881 and 10822, respectively. All in all, the VCMM has good credibility also under this scenario. We remark that adding the bivariate skew t copula to the candidate list at Step II in Algorithm 1 could increase the performance of the VCMM in this last scenario.

To conclude, the VCMM does show its effectiveness and flexibility in clustering multivariate non-Gaussian and Gaussian data in our simulation studies. Even though its starting partition cannot reasonably identify the clusters, the VCMM often deals with them.

3.7 Software: vineclust & model adequacy

We illustrate how to use our Algorithm 1 by focusing on improvements it provides thanks to its steps via a real data set analysis. We also guide how to assess the adequacy of the fitted marginal distributions by the VCMM. Our implementation is available with its `R` package called `vineclust` (Sahin 2021). The candidate set for margins is normal, Student's t with degrees

of freedom 3, logistic, log-normal, log-logistic, gamma distribution, and that for pair-copula families are Gaussian, t, Clayton, Gumbel, Frank, Joe, BB1, BB6, BB8 copula with their possible $90°$, $180°$, $270°$ rotations. Still, we are incorporating further marginal distributions into `vineclust`. The tolerance level (tol) in Equation (3.18) is 0.00001.

Additionally, we will discuss the computational effort of the VCMM. All computations in Sections 3.7 and 3.8 are run on a MacBook Air (2018) with a $1,6$ GHz Dual-Core Intel Core i5 and 8 GB of RAM, running R version $4.0.3$.

A well-analyzed Australian Institute of Sport (AIS) data consists of 13 measurements made on 102 male and 100 female athletes. Our objective of clustering this data is to see if the VCMM can find two clusters for females and males and analyze its steps' performance as mentioned above. For our analysis, we select a subset of five variables: lean body mass (LBM), weight (Wt), body mass index (BMI), white blood cell count (WBC), and percentage of body fat (PBF). This sample appears non-Gaussian and has asymmetric dependence patterns shown on the bottom panels in Figure 3.7. To obtain u-data, we fit the empirical cumulative distribution function of each variable in each class. The lower panels in pairs plots show marginally normalized contour plots, and we often do not observe Gaussian dependence since most contours are non-elliptical. The pairwise dependence between the same pair of variables usually is the same in females and males, but its strength is different (e.g., Wt and BMI). The marginal density function of BMI and WBC for both classes is similar to each other as shown in the diagonal of the top left panel in Figure 3.7.

Fitting k-means with 10000 different seeds leads to two different partitions of the data set. We fit the VCMM using both partitions but present the result for the best VCMM, whose BIC value is lower than the other. However, both models have the same final accuracy.

Since we know the gender of each observation, we can evaluate the misclassification rate of the binary classification for the clustering algorithm and associate the final clusters with the classes. The VCMM improves its clustering power with its steps in Algorithm 1 as shown in Table 3.5. The starting partition obtained from k-means assigns almost one-fourth of males to females (Step I). Then Step II fits the initial VCMM model using Markov trees, thereby returning a log-likelihood value. The accuracy is eight percent higher than the starting partition. After the ECM iterations and temporary clustering assignment (Step III and Step IV), the VCMM reduces the misclassification rate by 12% compared to k-means. Selecting the final VCMM with a full vine specification (Step V and Step VI) provides a crucial gain in the log-likelihood. In the end, the VCMM identifies almost all females correctly except one female. Overall it provides 12% higher accuracy in revealing females and males than k-means. The misclassified observations by the VCMM lie on the boundary of the classes as shown in the top right panel of Figure 3.7.

Figure 3.7: Pairwise scatter plot of the subset of `AIS` data (top left) with orange/red points for observations of females/males, diagonal: marginal density function of each class's corresponding variable, and that of the misclassified observations by the VCMM shown by magenta (top right). Pairs plots of females (bottom left) and males (bottom right), where upper: pairs plots of copula data, diagonal: histogram of copula margins, lower: marginally normalized contour plots.

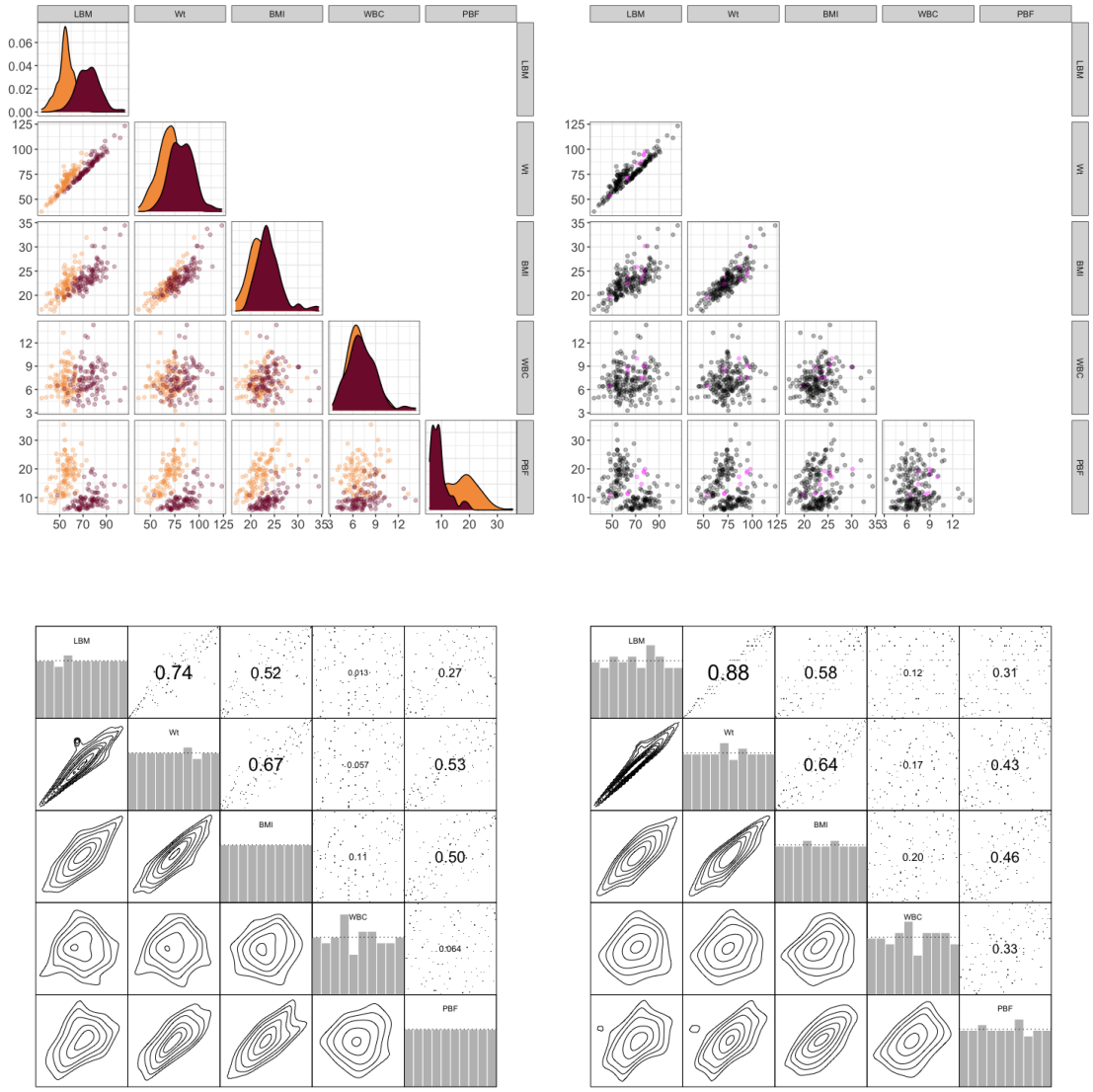


Table 3.5: Comparison of the VCMM's steps' clustering performance on the subset of `AIS`.

	Gender	Partition		Misclassification rate	Log-likelihood
		1	2		
k-means	F	93	7	0.16	-
	M	26	76		
VCMM after Step II	F	95	9	0.08	-2688
(with Markov trees)	M	12	90		
VCMM after Step IV	F	100	0	0.04	-2603
(ECM using Markov trees)	M	8	94		
Final VCMM (Full vine	F	99	1	0.04	-2326
using Step IV's assignment)	M	7	95		

The final VCMM runs for 2.28 minutes, and the ECM iterations stop after 15 iterations. Its estimated mixture weight of females and males is 0.54 and 0.46, respectively. We see in Table 3.6 that the VCMM fits a log-logistic distribution for BMI in both clusters, while other variables' marginal distributions are different in males and females.

Table 3.6: Selected marginal distributions and estimated parameters of of females and males. The variable encoding is given as follows: 1: LBM, 2: Wt, 3: BMI, 4: WBC, and 5: PBF. The cluster index (1) refers to females, whereas the index (2) denotes males.

Marginal distribution(parameters)		Marginal distribution(parameters)	
$F_{1(1)}(\boldsymbol{\gamma}_{1(1)})$	$llogis(12.4, 55.6)$	$F_{1(2)}(\boldsymbol{\gamma}_{1(2)})$	$lnorm(4.3, 0.1)$
$F_{2(1)}(\gamma_{2(1)})$	$\mathcal{N}(68.6, 12.1)$	$F_{2(2)}(\boldsymbol{\gamma}_{2(2)})$	$lnorm(4.4, 0.1)$
$F_{3(1)}(\gamma_{3(1)})$	$llogis(14.4, 22.0)$	$F_{3(2)}(\boldsymbol{\gamma}_{3(2)})$	$llogis(17.9, 23.5)$
$F_{4(1)}(\gamma_{4(1)})$	$\Gamma(18.0, 2.5)$	$F_{4(2)}(\boldsymbol{\gamma}_{4(2)})$	$lnorm(1.9, 0.3)$
$F_{5(1)}(\gamma_{5(1)})$	$\Gamma(10.8, 0.6)$	$F_{5(2)}(\boldsymbol{\gamma}_{5(2)})$	$lnorm(2.1, 0.2)$

The selected vine tree structure is not the same for females and males as seen in Figure 3.8. It is a path, i.e., D-vine, for males. The estimated pairwise dependence between pairs of the variables is non-Gaussian with diverse dependence strengths in females and males, except the pair of Wt and BMI. For instance, the pair of the variables, PBF and Wt, shows high strength and non-Gaussian dependence (Survival BB8 copula) in females. In higher tree levels, the highest (absolute value) estimated Kendall's τ for females is 0.74 and exists in the second tree, while 0.76 in the third tree for males, i.e., high strength conditional dependence exists in the higher tree levels.

Figure 3.8: Estimated vine copula models for males (left) and females (right). A letter at an edge with numbers inside the parenthesis refers to its bivariate copula family with its parameter(s)/Kendall's $\hat{\tau}$. The variable encoding is given as follows: 1: LBM, 2: Wt, 3: BMI, 4: WBC, and 5: PBF.

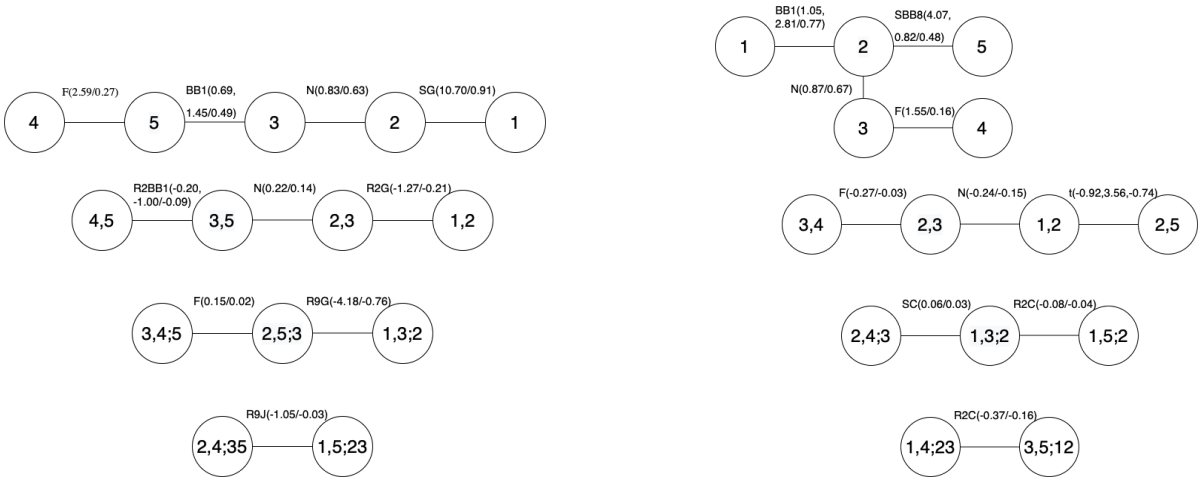
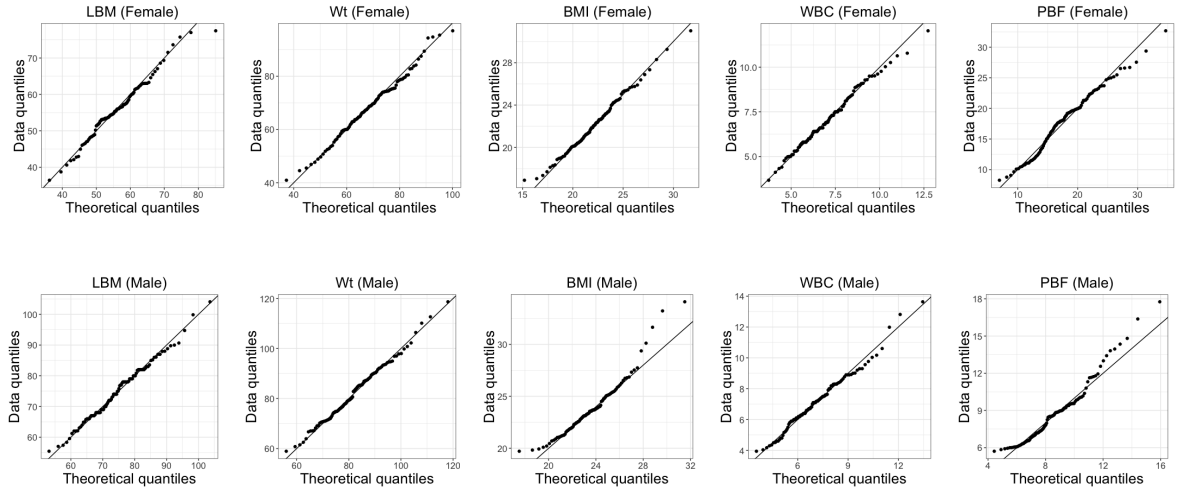


Figure 3.9 shows the QQ plots of the variables in both clusters. The fitted marginal distributions by the VCMM are adequate for the female cluster. However, the margin selection could be improved for the variables BMI and PBF in the male cluster, e.g., by adding the univariate skew t distribution into the list of univariate marginal distributions in Algorithm 1.

Figure 3.9: QQ plots of the variables in female (top) and male (bottom) clusters by the VCMM.



3.8 Open problems

We discuss open problems for vine copula mixture models, and some are shown by analyzing and clustering multivariate non-Gaussian real data. The candidate set for margins and pair-copula families is as given in Section 3.7. We run all clustering algorithms using the same random seed for a fair comparison, thereby using the same starting partitions.

Selection of vine tree structure for clusters

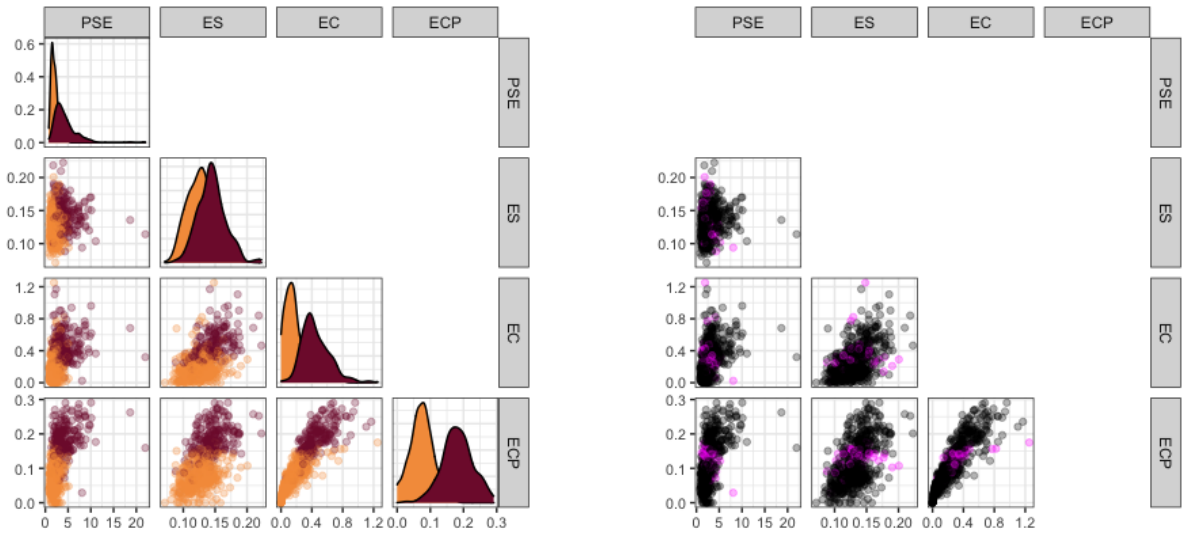
In Section 3.3, we describe a greedy approach for the selection of clusters' vine tree structure in the VCMM. We now compare its performance with a fixed vine tree structure for clusters in the VCMM. We specify each cluster's vine tree structure as a star, i.e., C-vine, in the VCMM, selecting their root node with our approach explained in Section 3.3, and denote this model by VCMM(C-vine). We assess their performance on the Breast Cancer Wisconsin (Diagnostic) data obtained from the UCI Machine Learning data repository (Dheeru and Karra Taniskidou 2017), where a digitized image of a fine needle aspirate (FNA) of a breast mass (Mangasarian et al. 1995) is used to have ten features from 569 patients. The mean value, extreme value (mean of the three largest values) and standard error of each feature are calculated, returning 30 continuous variables. The data contains two types of diagnosis: benign (352 patients) and malignant (212 patients), enabling us to assess the misclassification rate of the binary classification for the algorithms. We limit this illustration to a subset of four variables: perimeter standard error (PSE), extreme smoothness (ES), extreme concavity (EC), and extreme concave points (ECP) shown on the left panel in Figure 3.10. The data has non-Gaussian variables like PSE. We fit two-component clustering models to the data, and k-means is not sensitive to seeds.

Imposing a fixed vine structure in the VCMM, VCMM(C-vine), has the BIC of -4014 and the misclassification rate of 0.18, whereas the VCMM has that of -3970 and 0.10, respectively. Thus, imposing a C-vine tree structure decreases the model accuracy, where the selected root node is the variable ECP for both clusters. The VCMM(C-vine) appears to be the best vine copula mixture model regarding the BIC value. Since its misclassification rate is higher than the VCMM, we would need to construct a better model comparison criterion than the BIC value in the vine copula mixture model context in the future. The VCMM(C-vine) takes 7.14 minutes and 57 ECM iterations.

The selected vine tree structure by the VCMM is a path for the malignant cluster and a star with the root node of the variable ECP for the benign cluster. It runs for 1.07 minutes, and the ECM iterations stop after nine iterations. The right panel in Figure 3.10 shows that the observations whose estimated posterior probability is smaller than 0.9 in their assigned cluster are at the border of the support regions of benign and malignant.

Hence, allowing a flexible vine tree structure selection for clusters rather than imposing it improves the accuracy here. However, big simulation studies can be further conducted.

Figure 3.10: Pairwise scatter plots of the subset of `Breast Cancer Wisconsin(Diagnostic)`, where orange/red points denote observations of benign/malignant on the left. On the right, magenta points show observations whose estimated posterior probability by the VCMM is smaller than 0.9 in their assigned cluster.

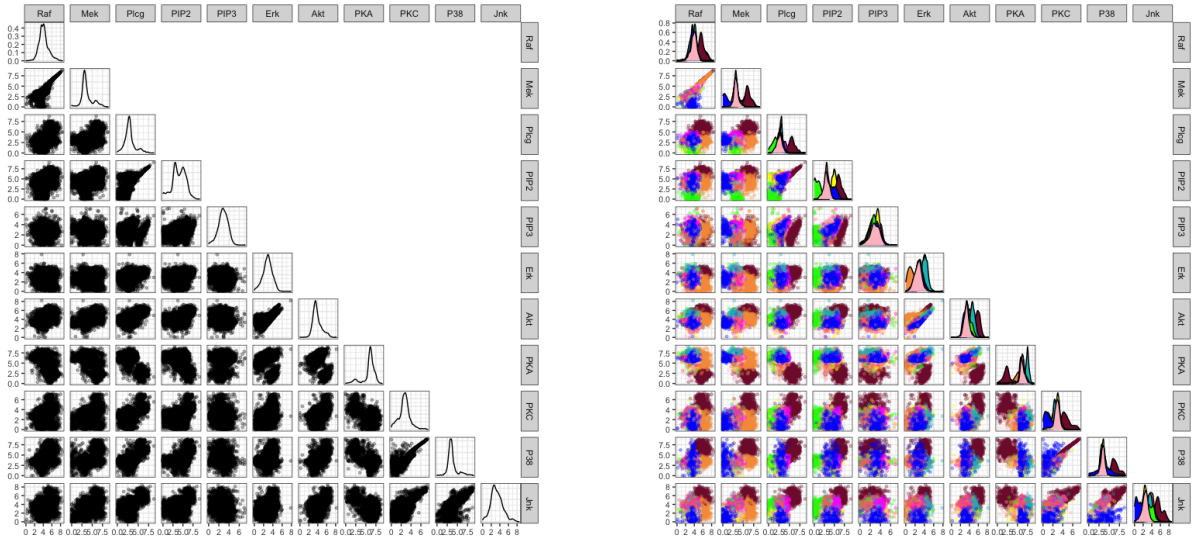


Selection of the number of components and model selection criteria

To study the problem of the selection of the number of components in the VCMM, we consider the Sachs Protein data analyzed by Sachs et al. (2005). It consists of logarithmized levels of 11 phosphorylated proteins and phospholipids in individual cells, subjected to general and specific molecular interventions. The original goal is to learn the causal pathways linking a set of 11 proteins and compare them to the known links in the literature, thereby validating the important model tools in genetics, Bayesian networks. The data is continuous as seen in the left panel of Figure 3.11, where we work with 6161 observations from nine experiments ($b2camp$, $cd3cd28$, $cd3cd28 + aktinhib$, $cd3cd28 + g007$, $cd3cd28 + ly$, $cd3cd28 + psitect$, $cd3cd28 + u0126$, $cd3cd28icam2$, pma) after removing 1305 observations with zero values to avoid dealing with zero inflation. The standard approach in the literature is to assume that the data follows a multivariate Gaussian distribution, thereby formulating a Gaussian Bayesian network (GBN). However, Figure 3.11 shows that the data has non-Gaussian univariate distributions and different components. Thus, Zhang and Shi (2017) work with the two-component Gaussian mixture copula Bayesian network and detect the underlying causal links better than the GBN.

Even though bimodality exists for some variables, such as PIP2 and mek, and there are two obvious groups on some scatter plots like the one of Akt and PKA in Figure 3.11, it is unclear how many hidden components exist in the data. Hence, following our suggestion in the previous sections, we fit the VCMM using a starting partition of k-means and model-

Figure 3.11: Pairwise scatter plots of the `Sachs Protein` data (left) and its partition into nine clusters by the VCMM (right). The diagonal of the plots shows the corresponding variable's marginal density function in each cluster.

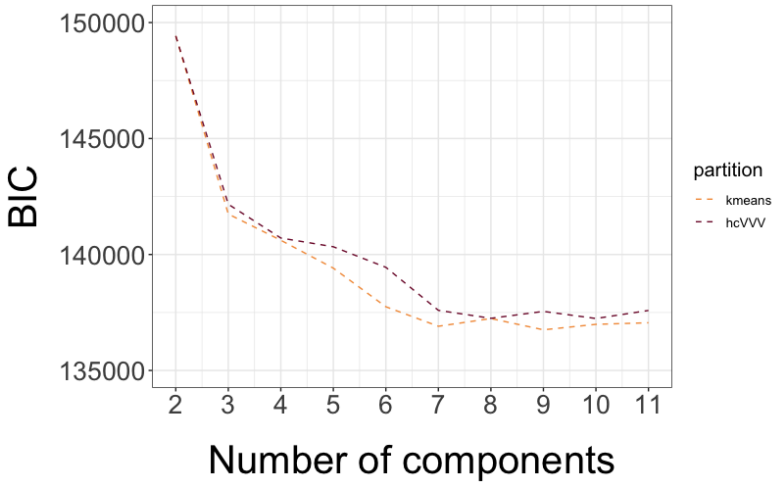


based hierarchical clustering with two to 11 components. Figure 3.12 shows that using the initial partition from model-based hierarchical clustering in the VCMM suggests having ten components since it is the global minimum point of its plot. However, starting with the assignment of k-means points out nine components in the experiments, giving the lowest BIC value among the evaluated ones. Thus, we select it as our final VCMM model. Nevertheless, we remark that the VCMM's BIC values from seven to eleven components are very close to each other using the initial partition from model-based hierarchical clustering and k-means.

Using the VCMM that starts with the assignment of k-means with nine components reveals the most observations of the experiments $b2camp$, $cd3cd28 + g007$, $cd3cd28 + psitect$, and $cd3cd28 + u0126$ as a cluster. The remaining five experiments' observations usually belong to the other four clusters. The partition is given in the right panel of Figure 3.11. The fitted univariate marginal distributions and bivariate copula families are mostly non-Gaussian in the clusters, showing the need for a non-Gaussian model. The fitted VCMMs are available upon request.

To provide guidelines on the number of components to consider, we give a first analysis based on the BIC criterion. It turns out that the initial partition impacts the final model. Hence, one possible future research direction is the selection of the number of components in vine copula mixture models. It can be combined with the construction of a model comparison criterion. The ideas for sparse model selection in Nagler et al. (2019) can be a starting point.

Figure 3.12: BIC values for the VCMM with k-means and model-based hierarchical clustering for the different number of components on the `Sachs Protein` data.



EM-type algorithms for vine copula mixture models

As a remark of a reviewer of Sahin and Czado (2022b), we note that the CM-step 2 in Equation (3.16), i.e., the optimization of univariate marginal parameters, can be done separately by variable $p \in \{1, \dots, d\}$ within cluster k. In high dimensions, it decreases the number of parameters to optimize per optimization problem. In this case, following our notation in the chapter, CM-step 2 at the $(t+1)$th iteration could numerically maximize Equation (3.20) to obtain updated marginal parameter estimates $\boldsymbol{\gamma}_{p(j)}^{(t+1)}$ of the pth variable in the jth component. Assume $\boldsymbol{\gamma}_{p(j)}^{(t+1)}$ is obtained. It is then used to update the u-data of the pth variable in the jth component by applying probability integral transformation: $\boldsymbol{u}_{p(j)}^{(t+1)} = \hat{F}_{p(j)}(\boldsymbol{x}_{p(j)}; \boldsymbol{\gamma}_{p(j)}^{(t+1)})$. However, this scenario would need more estimation steps within CM-step 2, thereby more computational effort, than CM-step 2 we presented in this chapter.

$$\begin{aligned}
&\max_{\boldsymbol{\gamma}_{p(j)}} \sum_{i=1}^{n} [r_{i,j}^{(t+1)} \cdot \log\, f_{p(j)}(x_{i,p}; \boldsymbol{\gamma}_{p(j)})] \\
&\quad + \sum_{i=1}^{n} \sum_{m=1}^{d-1} \sum_{e \in E_{(j)m}} \Big[r_{i,j}^{(t+1)} \cdot \log\, c_{(j)e_a,e_b;\boldsymbol{D_e}} \big(F_{(j)e_a|\boldsymbol{D_e}}(x_{i,e_a}|\boldsymbol{x}_{i,\boldsymbol{D_e}}; \boldsymbol{\gamma}_{(j)e_a|\boldsymbol{D_e}}, \boldsymbol{\theta}_{(j)e_a|\boldsymbol{D_e}}^{(t)}), \\
&\quad F_{(j)e_b|\boldsymbol{D_e}} \big(x_{i,e_b}|\boldsymbol{x}_{i,\boldsymbol{D_e}}; \boldsymbol{\gamma}_{(j)e_b|\boldsymbol{D_e}}, \boldsymbol{\theta}_{(j)e_b|\boldsymbol{D_e}}^{(t)} \big); \boldsymbol{\theta}_{(j)e_a,e_b;\boldsymbol{D_e}}^{(t)} \big) \Big] \text{ for } p = 1, \dots d, j = 1, \dots k.
\end{aligned} \tag{3.20}$$

Moreover, the order of the CM steps can be changed, and the performance of the ECM algorithm's extensions, such as the expectation conditional maximization of either algorithm (Liu and Rubin 1994), can be analyzed for our framework.

Computationally efficient vine copula mixture models

A potential drawback of the proposed method is the computational cost for high-dimensional data. We provide an initial framework for using vine copulas with finite mixture models and clustering. Therefore, another future research direction is to handle variable selection and dimensionality reduction for vine copula based clustering. The traditional variable selection methods for other model-based clustering algorithms have to be reviewed and adjusted (e.g., Raftery and Dean (2006), Maugis et al. (2009)).

Parsimonious vine copula mixture models

While the current algorithm has significant advancement in revealing non-elliptical components, further development would modify the proposed method to obtain parsimonious vine copula mixture models and then use them for clustering. On the one hand, the optimal truncation level can be studied. On the other hand, factor models explain the dependence structure among the observed variables using one or several common factors. Krupskii and Joe (2013) combine factor models with copula models to capture the complex dependence structure of the data with relatively few parameters. An extension of Krupskii and Joe (2013) in the context of mixture models can be considered for parsimony.

Initial model selection

We show that different quick clustering methods can initialize the algorithm, and the final model can differ accordingly. Since running the algorithm until the stopping condition holds takes time with high-dimensional data, initialization approaches for the vine copula mixture models need to be further improved in the future. Scrucca and Raftery (2015) study a similar problem with regard to the initial parameter values in multivariate normal mixture models. Initializing the algorithm could be further studied for harder situations, as there are significant overlaps among the components with an unknown number of components.

Mixed discrete/continuous variables and missing data

The proposed method can be extended to deal with mixed discrete/continuous variables and missing data as other future research directions. The construction defined in Panagiotelis et al. (2012) and further studied in Panagiotelis et al. (2017) can be a starting point for the former. Wang and Lin (2015) discuss how to handle missing data in multivariate skew t mixture models.

3.9 Conclusion

We propose a vine copula mixture model that works with continuous data and fits all classes of vine tree structures. It uses parametric marginal distributions and pair-copula families. It applies a wide range of pair-copula families; thus, it accommodates diverse tail dependence and asymmetries within the components. Due to its parametric nature, it nicely interprets the structure of the data. Assuming the number of components in the data is known, we follow a data-driven approach for the remaining model selection problems. We work with the ECM algorithm for parameter estimation. With the proposed method, we formulate a new model-based clustering algorithm called VCMM.

We evaluate the performance of the algorithm on simulated and real data. Our simulation studies illustrate that the vine copula based clustering has greater flexibility than various model-based clustering algorithms available in the literature and hence captures the non-Gaussian components hidden in the data better than others, especially when the data has heavy-tailed margins and tail dependence between pairs of variables. The real data analysis supports it and the better clustering assignment thanks to allowing all types of vine tree structures. Due to its flexibility in the formulation, it can also capture Gaussian components. Additionally, it can answer whether the dependence between a pair of variables changes with the variables' different values or differs among the clusters. It is an appealing clustering approach since the era of big data comes with different data characteristics. We also discuss open problems for vine copula mixture models.

Chapter 4

High-dimensional sparse vine copula regression

This chapter contains the materials from Sahin and Czado (2022a), but it extends simulation studies in Section 4.7 and discusses open research problems in Section 4.8.

4.1 Motivation

Vine copula based (quantile) regression is a significant tool for modeling the nonlinear relationship between explanatory variables and response. It considers higher-order explanatory variables and their interaction. In addition, it can deal with unknown functional error forms. However, the current vine copula based regression methods' computational complexity makes them infeasible to be applied in high-dimensional data sets (Chang and Joe 2019; Kraus and Czado 2017; Tepegjozova et al. 2022; Zhu et al. 2021). We refer to high- and ultra high-dimensional data sets when the number of explanatory variables is between ten and 1000 and larger than 1000, respectively.

In this chapter, we propose two vine copula based regression methods that perform well in analyzing high-dimensional sparse data sets, where sparsity means that many explanatory variables are not related to the response. We show that their computational complexity is significantly less than the existing methods. We define relevant, redundant, and irrelevant explanatory variables for quantile regression and assess the methods' prediction power in high-dimensional sparse simulated data sets. Our analyses regarding the inclusion of relevant variables and exclusion of irrelevant variables show our methods' capability to provide sparse models. Overall, our methods are novel and powerful and can resolve data dimensionality issues in vine copula based regression. To the best of our knowledge, there has not yet been any study assessing vine copula regression methods' performance in the presence of redundant and irrelevant variables.

Alternative quantile regression models include linear models (Koenker and Bassett Jr

1978), generalized additive models (GAM) (Koenker 2011; Wood 2017), quantile regression forests (QRF) (Meinshausen 2006), and quantile regression neural networks ($QRNN$) (Cannon 2011). The models GAM, QRF, and $QRNN$ are nonlinear models like vine copulas. However, $QRNN$ may suffer from quantile crossing, which does not exist in vine copula based approaches by construction, and need some strategies for avoiding it (Cannon 2018). Kraus and Czado (2017) show a better performance of their vine copula based approach than GAM. Hence, among the nonlinear models, we compare our models with quantile regression forests and show our advantages, especially in the presence of correlated variables. Additionally, despite the quantile crossing problem, we analyze the performance of linear models incorporating variable selection in nonlinear situations with many redundant and irrelevant variables.

The chapter is organized as follows: Section 4.2 introduces sparse vine copula regression methods, and Section 4.3 discusses model selection problems. The complexity of the proposed methods is analyzed in Section 4.4, and a variable classification as relevant, redundant, and irrelevant is defined in Section 4.5. We present an illustrative example of the methods in Section 4.6, provide simulation studies in Section 4.7, discuss open problems in Section 4.8, and conclude in Section 4.9.

4.2 Sparse vine copula regression formulation

In the remainder of the chapter, assume that $(y_i, x_{i,1}, \ldots, x_{i,p})$, $i = 1, \ldots, n$, are realizations of the random vector $(Y, X_1, \ldots, X_p)$, and Y denotes a response variable with its marginal distribution F_Y and the others correspond to explanatory variables with their marginal distributions $F_1, \ldots, F_p$. Our analyses are based on D-vines introduced in Chapter 2. Since p-dimensional D-vine copula's input is the marginally uniform data on $[0,1]^p$, the estimation of the D-vine copula follows a two-step approach called the inference for margins (Joe and Xu 1996). First, each marginal distribution is estimated. Then the data is converted to copula data by applying probability integral transformation (PIT). Univariate parametric margins can be used to apply the PIT; however, they might not fit the data to be modeled well. Alternatively, the PIT could be applied using the marginal distributions' kernel based density estimate. However, since vine copula based regression methods use the estimated inverse of the response's marginal distribution function for quantile predictions as discussed in Equation (2.6), such an application results in the predictions being equal to the observed response values. To overcome these difficulties, we estimate the marginal distributions using a univariate non-parametric kernel density estimator with the `R` package `kde1d` (Nagler and Vatter 2022a), i.e., $\hat{F}_Y$ and $\hat{F}_d$, $d = 1, \ldots, p$. Next, we have the copula data: $(v_i, u_{i,1}, \ldots, u_{i,p}) = (\hat{F}_Y(y_i), \hat{F}_1(x_{i,1}), \ldots, \hat{F}_p(x_{i,p}))$, $i = 1, \ldots, n$, being realizations of the random vector $(V, U_1, \ldots, U_p)$.

vinereg

Kraus and Czado (2017) proposed a D-vine copula based quantile regression with a forward variable selection, and the `R` package `vinereg` (Nagler 2022) provides its implementation. We summarize the main steps of `vinereg` in the following.

Assume the response variable's index is denoted by 0, and the response variable is a leaf node in the first tree of a D-vine.

Step 1 (initialization): For the given data $(y_i, x_{i,1}, \ldots, x_{i,p})$ and $(v_i, u_{i,1}, \ldots, u_{i,p})$, the initial D-vine order $\mathcal{D}^{(1)} = (0)$, the initial chosen variable index set $\mathcal{I}_{var}^{(1)} = \emptyset$, and the initial set of candidate explanatory variables $p_{cand}^{(2)} = \{1, \ldots, p\}$.

For $s = 1, 2, \ldots,$

Step 2 (variable selection): Extend the D-vine structure order by adding a variable whose index is d to have a D-vine structure order $\mathcal{D}^{(s+1)} = (\mathcal{D}^{(s)}, d)$. Fit a parametric D-vine copula having the structure $\mathcal{D}^{(s+1)}$ and denote the vine copula, its density, and its estimated parameters by $\widehat{CD}^{d_{(s+1)}}$, $\widehat{cd}^{d_{(s+1)}}$, and $\hat{\boldsymbol{\theta}}^{d_{(s+1)}}$, respectively. Then find the variable with the index $d_{(s+1)}^*$ for which the conditional log-likelihood of the D-vine copula $\widehat{CD}^{d_{(s+1)}^*}$ is maximized, i.e.,

$$d_{(s+1)}^* = \underset{d_{(s+1)} \in p_{cand}^{(s+1)}}{\arg\max} \sum_{i=1}^{n} \ln \widehat{cd}_{0|d_{(0)},\ldots,d_{(s+1)}}^{d_{(s+1)}}(v_i | u_{i,d_{(0)}}, \ldots, u_{i,d_{(s+1)}}; \hat{\boldsymbol{\theta}}^{d_{(s+1)}}). \tag{4.1}$$

Step 3 (D-vine extension): Extend the D-vine structure order by adding the variable whose index is $d_{(s+1)}^*$ to have a D-vine structure order $\mathcal{D}^{(s+1)} = (\mathcal{D}^{(s)}, d_{(s+1)}^*)$. Select the new parametric pair-copula families and estimate their parameters in the extended D-vine structure. Denote the associated D-vine copula by $\hat{C}^{(s+1)}$ and its estimated parameters by $\hat{\boldsymbol{\theta}}^{(s+1)}$.

Step 4 (Chosen variable indices and hyperparameter updates): Extend the chosen variable indices $\mathcal{I}_{var}^{(s+1)} = \mathcal{I}_{var}^{(s)} \cup d_{(s+1)}^*$ and update $p_{cand}^{(s+2)} = p_{cand}^{(s+1)} \setminus d_{(s+1)}^*$.

However, *vinereg*'s computational complexity makes it infeasible to run in high dimensions as will be discussed in Section 4.4. Thus, we propose two methods to perform a D-vine copula regression on high-dimensional sparse data sets: *vineregRes* and *vineregParCor*.

vineregRes

The method *vineregRes* performs the variable selection at a given iteration based on the residuals of the previous iteration, i.e., the pseudo-response. It finds the variable among the candidates, which provides the best bivariate copula conditional log-likelihood conditioned on the variable and conditioning the pseudo-response of the previous iteration. Assume $\tilde{y}_i^{(s)}$ and $\tilde{v}_i^{(s)}$ $i = 1, \ldots, n$, denote the pseudo-response and its copula data in the sth iteration, respectively, which are realizations of the random variable $Y^{(s)}$ and $V^{(s)}$, respectively. $V^{(0)}$ and $V^{(s)}$ have the indices 0 and $0^{(s)}$, respectively, and are always a leaf node in the first tree.

Step 1 (initialization): For the given data $(y_i, x_{i,1}, \ldots, x_{i,p})$ and $(v_i, u_{i,1}, \ldots, u_{i,p})$, define the initial pseudo-response $\tilde{y}_i^{(1)} = y_i$ with its copula scale $\tilde{v}_i^{(1)}$, $i = 1, \ldots, n$, the initial D-vine order $\mathcal{D}^{(1)} = (0)$, the initial chosen variable index set $\mathcal{I}_{var}^{(1)} = \emptyset$, and the initial set of candidate explanatory variables $p_{cand}^{(1)} = \{1, \ldots, p\}$.

For $s = 1, 2, \ldots,$

Step 2 (variable selection): Fit a parametric bivariate copula to data $\{(\tilde{v}_i^{(s)}, u_{i,d}),\ i = 1, \ldots, n\}$ for $d \in p_{cand}^{(s)}$ and denote the copula, copula density, and its estimated parameters by $\widehat{CR}^{d_{(s)}}$, $\widehat{cr}^{d_{(s)}}$ and $\hat{\boldsymbol{\theta}}^{d_{(s)}}$, respectively. Then find the variable for which the conditional log-likelihood of the copula $\widehat{CR}^{d^*_{(s)}}$ is maximized, i.e.,

$d^*_{(s+1)} = \arg\max_{d_{(s)} \in p_{cand}^{(s)}} \sum_{i=1}^{n} \ln \widehat{cr}^{d_{(s)}}_{0^{(s)}|d_{(s)}}(\tilde{v}_i^{(s)} | u_{i,d_{(s)}}; \hat{\boldsymbol{\theta}}^{d_{(s)}})$.

Step 3 (D-vine extension): Extend the D-vine order by adding the variable with index $d^*_{(s+1)}$ to get a D-vine order $\mathcal{D}^{(s+1)} = (\mathcal{D}^{(s)}, d^*_{(s+1)})$. Select the parametric pair-copula families and estimate the parameters in the extended D-vine structure, where the associated D-vine copula and its estimated parameters are denoted by $\hat{C}^{(s+1)}$ and $\hat{\boldsymbol{\theta}}^{(s+1)}$, respectively.

Step 4 (Chosen variable indices and hyperparameter updates): Extend the chosen variable set, adding the new variable, $\mathcal{I}_{var}^{(s+1)} = \mathcal{I}_{var}^{(s)} \cup d^*_{(s+1)}$ and update $p_{cand}^{(s+1)} = p_{cand}^{(s)} \setminus d^*_{(s+1)}$.

Step 5 (Pseudo-response update or stop): If a stopping condition (see Section 4.3) does not hold, estimate the median of the response variable based on the D-vine copula $\hat{C}^{(s+1)}$ and update the pseudo-response using Equation (2.6), i.e.,

$\tilde{y}_i^{(s+1)} = y_i - \hat{F}_Y^{-1}(\hat{C}^{-1(s+1)}_{0|\boldsymbol{I}_{var}^{(s+1)}}(0.50 | u_{i,p_1}, \ldots, u_{i,p_{d_{(s+1)}}}; \hat{\boldsymbol{\theta}}^{(s+1)}))$, $\quad \tilde{v}_i^{(s+1)} = \hat{F}_{Y^{(s+1)}}(\tilde{y}_i^{(s+1)})$,

where $i = 1, \ldots, n$ and $\{p_1, \ldots, p_{d_{(s+1)}}\} \subseteq \mathcal{I}_{var}^{(s+1)}$.

vineregParCor

Another method to perform an efficient D-vine copula regression for high-dimensional data is to use the partial correlation between the response and a candidate explanatory variable given the chosen variables at each iteration based on their empirical normal scores. Such scores are calculated as detailed in Section 1 of Joe (2014): order the dth variable $(x_{1,d}, \ldots, x_{n,d})^\top$ non-decreasingly such that $x_{1_o,d} \leq \ldots \leq x_{n_o,d}$ and $\{x_{1_o,d}, \ldots, x_{n_o,d}\} = \{x_{1,d}, \ldots, x_{n,d}\}$ for $d = 1, \ldots, p$. Then the ith ordered value $x_{i_o,d}$ is transformed from i to its empirical normal score $z_{i_o,d}$ via $z_{i_o,d} = \Phi^{-1}\big({}^{(i+a)}/_{(n+1+2a)}\big)$ with $a \approx -0.50$. Next, *vineregParCor* follows:

Step 1 (initialization): As given in *vinereg* and the data's empirical normal scores.

For $s = 1, 2, \ldots,$

Step 2 (variable selection): $d^*_{(s+1)} = \arg\max_{d_{(s)} \in p_{cand}^{(s)}} |\rho_{0,d_{(s)};\mathcal{I}_{var}^{(s)}}|$, where $\rho_{j,k;S}$ is the partial correlation of variables j, k given those indexed in the set S based normal scores.

Step 3 (D-vine extension): As given in *vinereg*.

Step 4 (Chosen variable indices and hyperparameter updates): As given in *vinereg*.

4.3 Model selection

Important model selection problems in $vineregRes$ and $vineregParCor$ include the selection of pair-copula families and when to stop adding explanatory variables into the model.

Bivariate copula selection

Step 3 of $vineregRes$ and $vineregParCor$ selects parametric pair-copulas and estimates their parameters associated with the extension of the D-vine structure. Also, Step 2 of $vineregRes$ fits a parametric bivariate copula to the pseudo-response and a candidate explanatory variable. The selection of pair-copulas can be decided by analyzing the expected behavior of conditional quantiles in the extreme values of the variable's space. Alternatively, a model selection criterion can select a good fit among the candidate bivariate copula families. First, we estimate the parameters that maximize the log-likelihood of a candidate bivariate copula family. We work with the following bivariate copula families and their 90, 180, and 270 degree rotations: BB1, BB6, BB7, BB8, Clayton, Frank, Gaussian, Gumbel, Joe, t, and independence copula. Later we can select the one with the lowest AIC or the Bayesian information criterion. While extending the D-vine structure and adding new trees at Step 3 of the methods, the fit of parametric pair-copulas can be performed sequentially from the lowest to the highest trees (Brechmann 2010).

Stopping criteria

To decide if a chosen candidate explanatory variable in a given iteration should be in a model, we will consider the conditional AIC, which penalizes the conditional log-likelihood of the model based on the D-vine copula defined in Equation (4.1) by the effective degrees of freedom in the model. Thus, in the sth iteration of $vineregRes$ and $vineregParCor$, the conditional AIC of the model based on the D-vine copula $\hat{C}^{(s)}$ and D-vine copula density $\hat{c}^{(s)}$ with the model's effective degrees of freedom $|\hat{\boldsymbol{\Theta}}^{(s)}|$ is

$$CAIC(\hat{\boldsymbol{\Theta}}^{(s)}) = -2\cdot\sum_{i=1}^{n}\left(\ln\, \hat{c}^{(s)}_{0|\boldsymbol{I}^{(s)}_{var}}(\hat{F}_Y(y_i)|\hat{F}_{p_1}(x_{i,p_1}),\ldots,\hat{F}_{p_{d_{(s)}}}(x_{i,p_{d_{(s)}}})) + \ln\, \hat{f}_Y(y_i)\right) + 2\cdot|\hat{\boldsymbol{\Theta}}^{(s)}|,$$

where it holds $\{p_1,\ldots,p_{d_{(s)}}\} \subseteq \boldsymbol{I}^{(s)}_{var}$.

The effective degrees of freedom is the sum of the number of pair copula parameters that appear in the conditional log-likelihood of the response and the degrees of freedom from the kernel-based margin estimate of the response. The latter is a measure of smoothing, with more details provided in Section 5.3.2 of Loader (2006). However, using the degrees of freedom from the kernel-based margin estimate of the response without considering the degrees of freedom from the explanatory variables in the model might be debatable. Alternatively, one may consider the effective degrees of freedom as the sum of the number of pair copula

parameters in the model and the degrees of freedom from the kernel-based margin estimate of the response and explanatory variables in the model.

We stop adding variables in D-vine copula regressions when the current iteration's conditional AIC is equal to or larger than the previous iteration's conditional AIC. If the conditional AIC always improves in each iteration, we stop after all explanatory variables are included in the model. Considering the computational burden in high dimensions, an alternative is to stop if the selected number of variables reaches a threshold.

4.4 Complexity

Assuming that the data consists of p explanatory variables, the complexity of the existing method $vinereg$ is $\mathcal{O}(p^3)$ in terms of the total number of bivariate copulas to be selected during the algorithm (Tepegjozova 2019). Thus, we evaluate the complexity of $vineregRes$ and $vineregParCor$ using the same criterion. We will consider the worst-case scenario that the algorithms run until all explanatory variables are included in the model. Further, the total number of estimated parameters is linear in terms of the number of bivariate copulas.

In $vineregRes$, the selection of bivariate copulas exists at Steps 2 and 3. Step 2 fits a bivariate copula to the data of the iteration's pseudo-response and a candidate explanatory variable. Therefore, the number of bivariate copulas to be chosen at Step 2 is the total number of candidate explanatory variables at the given iteration. At the sth iteration, there are $p-(s-1)$ candidate explanatory variables. Thus, the number of bivariate copulas to be selected at this step is given by

$$\sum_{s=1}^{p}(p-(s-1)) = \frac{p\cdot(p+1)}{2}. \tag{4.2}$$

In $vineregParCor$, the variable selection is based on partial correlations; thus, its Step 2 does not perform any selection of bivariate copulas.

Step 3 of $vineregRes$ and $vineregParCor$ extends the D-vine structure, adding the selected explanatory variable at Step 2. Thus, at the sth iteration, it results in the selection of s pair-copulas. Hence, Step 3 selects the following number of bivariate copulas:

$$\sum_{s=1}^{p} s = \frac{p\cdot(p+1)}{2}. \tag{4.3}$$

Considering Equations (4.2) and (4.3), we calculate the total complexity of $vineregRes$:

$$\sum_{s=1}^{p} s + \sum_{s=1}^{p}(p-(s-1)) = p\cdot(p+1). \tag{4.4}$$

Considering Equation (4.3), we calculate the total complexity of $vineregParCor$:

$$\sum_{s=1}^{p} s = \frac{p \cdot (p+1)}{2}. \tag{4.5}$$

Equations (4.4) and (4.5) show that the complexity of $vineregParCor$ and $vineregRes$ in terms of the total number of selected bivariate copulas is $\mathcal{O}(p^2)$. As a result, our methods significantly reduce the computational complexity of $vinereg$. Further, setting pair-copulas as independence on low tree levels or applying thresholded vine copulas as investigated by Nagler et al. (2019), where independence copulas are fitted when the dependence strength is less than a threshold, decreases the complexity to less than $\mathcal{O}(p^2)$.

4.5 Relevant, redundant, and irrelevant variables

Now we define relevant, redundant, and irrelevant variables for predicting the conditional quantile of a response variable Y given the index set of explanatory variables $\mathcal{X}$. We will denote the cdf and pdf of the variables with the index $\mathcal{X}$ by $F_{\mathcal{X}}$ and $f_{\mathcal{X}}$, respectively.

Definition 4.1: Relevant variables

The index set of variables $\mathcal{M}$ is called relevant for Y if and only if it holds $F_{Y|\mathcal{M}}(y|\boldsymbol{x}_{\mathcal{M}}) \neq F_Y(y)$, where the vector $\boldsymbol{x}_{\mathcal{M}}$ includes the variables in the set $\mathcal{M}$, $\mathcal{M} \subseteq \mathcal{X}$.

Definition 4.2: Redundant variables

The index set of variables $\mathcal{R}$ is called redundant given the set of variables $\mathcal{M}$ for Y if and only if it holds $F_{Y|\mathcal{M},\mathcal{R}}(y|\boldsymbol{x}_{\mathcal{M}}, \boldsymbol{x}_{\mathcal{R}}) = F_{Y|\mathcal{M}}(y|\boldsymbol{x}_{\mathcal{M}})$ and $F_{\mathcal{M},\mathcal{R}}(\boldsymbol{x}_{\mathcal{M}}, \boldsymbol{x}_{\mathcal{R}}) \neq F_{\mathcal{M}}(\boldsymbol{x}_{\mathcal{M}}) \cdot F_{\mathcal{R}}(\boldsymbol{x}_{\mathcal{R}})$, where the vectors $\boldsymbol{x}_{\mathcal{M}}$ and $\boldsymbol{x}_{\mathcal{R}}$ include the variables in the sets $\mathcal{M}$ and $\mathcal{R}$, respectively, $\mathcal{R} \subseteq \mathcal{X}, \mathcal{M} \subseteq \mathcal{X}, \mathcal{R} \cap \mathcal{M} = \emptyset$.

Let $\mathcal{R}$ and $\mathcal{M}$ be the non-overlapping index sets for variables. $\boldsymbol{X}_{\mathcal{R}}$ is redundant for Y given $\boldsymbol{X}_{\mathcal{M}}$ if $Y \perp\!\!\!\perp \boldsymbol{X}_{\mathcal{R}} | \boldsymbol{X}_{\mathcal{M}}$ and $\boldsymbol{X}_{\mathcal{M}} \not\perp\!\!\!\perp \boldsymbol{X}_{\mathcal{R}}$. Conditional on $\boldsymbol{X}_{\mathcal{M}} = \boldsymbol{x}_{\mathcal{M}}$, for any $\boldsymbol{x}_{\mathcal{M}}$, $\boldsymbol{X}_{\mathcal{R}}$ does not provide any additional information for predicting Y, but $\boldsymbol{X}_{\mathcal{R}}$ without $\boldsymbol{X}_{\mathcal{M}}$ have some predictability for Y.

Example 4.1 *Consider the model* $(Y, X_1, X_2)^\top \sim \mathcal{N}\left(\begin{pmatrix}0\\0\\0\end{pmatrix}, \begin{pmatrix}1 & 0.5 & 0.4\\0.5 & 1 & 0.8\\0.4 & 0.8 & 1\end{pmatrix}\right)$, *where* $\rho_{Y,X_2;X_1} = \frac{\rho_{Y,X_2} - \rho_{Y,X1}\rho_{X_2,X_1}}{\sqrt{(1-\rho^2_{Y,X_1})(1-\rho^2_{X_2,X_1})}} = \frac{0.4-0.5\cdot 0.8}{\sqrt{(1-0.5^2)(1-0.8^2)}} = 0$, *i.e.,* Y *is conditionally independent of* X_2 *given* X_1. *Hence, we have* $f_{Y|X_1,X_2}(y|x_1,x_2) = \frac{f_{Y,X_2|X_1}(y,x_2|x_1)}{f_{X_2|X_1}(x_2|x_1)} = \frac{f_{Y|X_1}(y|x_1)\cdot f_{X_2|X_1}(x_2|x_1)}{f_{X_2|X_1}(x_2|x_1)} = f_{Y|X_1}(y|x_1)$. *Since* $f_{X_1,X_2}(x_1,x_2) \neq f_{X_1}(x_1)\cdot f_{X_2}(x_2)$, X_2 *is redundant given* X_1 *for* Y.

Now consider the case where $\mathcal{M}$ is still relevant for Y, but $\mathcal{R}$ is independent of $\mathcal{M}$.

Definition 4.3: Irrelevant variables

The set of variables $\mathcal{I}$ is called irrelevant given the set of variables $\mathcal{M}$ for Y if and only if it holds $F_{Y|\mathcal{M},\mathcal{I}}(y|\boldsymbol{x}_\mathcal{M}, \boldsymbol{x}_\mathcal{I}) = F_{Y|\mathcal{M}}(y|\boldsymbol{x}_\mathcal{M})$, $F_{\mathcal{M},\mathcal{I}}(\boldsymbol{x}_\mathcal{M}, \boldsymbol{x}_\mathcal{I}) = F_\mathcal{M}(\boldsymbol{x}_\mathcal{M}) \cdot F_\mathcal{I}(\boldsymbol{x}_\mathcal{I})$, and $F_{Y,\mathcal{I}}(y, \boldsymbol{x}_\mathcal{I}) = F_Y(y) \cdot F_\mathcal{I}(\boldsymbol{x}_\mathcal{I})$, where the vectors $\boldsymbol{x}_\mathcal{M}$ and $\boldsymbol{x}_\mathcal{I}$ include the variables in the sets $\mathcal{M}$ and $\mathcal{I}$, respectively, $\mathcal{I} \subseteq \mathcal{X}, \mathcal{M} \subseteq \mathcal{X}, \mathcal{I} \cap \mathcal{M} = \emptyset$.

Let $\mathcal{I}$ and $\mathcal{M}$ be the non-overlapping index sets for variables. $\boldsymbol{X}_\mathcal{I}$ is irrelevant for Y given $\boldsymbol{X}_\mathcal{M}$ if and only if $\boldsymbol{X}_\mathcal{I} \perp\!\!\!\perp (Y, \boldsymbol{X}_\mathcal{M})$. Since $\boldsymbol{X}_\mathcal{I} \perp\!\!\!\perp (Y, \boldsymbol{X}_\mathcal{M})$ if and only if $Y \perp\!\!\!\perp \boldsymbol{X}_\mathcal{I} | \boldsymbol{X}_\mathcal{M}$ and $\boldsymbol{X}_\mathcal{I} \perp\!\!\!\perp \boldsymbol{X}_\mathcal{M}$, it marginally implies $\boldsymbol{X}_\mathcal{I} \perp\!\!\!\perp Y$ so that $\boldsymbol{X}_\mathcal{I}$ has no predictive value for Y unconditionally or conditionally. Given that it holds $\boldsymbol{X}_\mathcal{I} \perp\!\!\!\perp (Y, \boldsymbol{X}_\mathcal{M})$, we can write the following: $f_{YIM} = f_I \cdot f_{YM} = f_I \cdot f_{Y|M} \cdot f_M$. Moreover, we can decompose $f_{YIM} = f_{YI|M} \cdot f_M$. Then it must hold that $f_{YI|M} = f_I \cdot f_{Y|M}$. As a result, we obtain $\boldsymbol{X}_\mathcal{I} \perp\!\!\!\perp (Y, \boldsymbol{X}_\mathcal{M})$ if and only if $Y \perp\!\!\!\perp \boldsymbol{X}_\mathcal{I} | \boldsymbol{X}_\mathcal{M}$ and $\boldsymbol{X}_\mathcal{I} \perp\!\!\!\perp \boldsymbol{X}_\mathcal{M}$.

Example 4.2 *Consider the model* $(Y, X_1, X_2)^\top \sim \mathcal{N}\left(\begin{pmatrix}0\\0\\0\end{pmatrix}, \begin{pmatrix}1 & 0.5 & 0\\0.5 & 1 & 0\\0 & 0 & 1\end{pmatrix}\right)$*, where it holds* $\rho_{Y,X_2;X_1} = 0$*; hence,* $f_{Y|X_1,X_2}(y|x_1, x_2) = f_{Y|X_1}(y|x_1)$*. Also, it holds* $f_{X_1,X_2}(x_1, x_2) = f_{X_1}(x_1) \cdot f_{X_2}(x_2)$ *and* $f_{Y,X_2}(y, x_2) = f_Y(y) \cdot f_{X_2}(x_2)$*; thus,* X_2 *is irrelevant given* X_1 *for* Y*.*

Proposition 4.1: Redundant variables in a D-vine copula

If a p-dimensional D-vine copula is a t-truncated vine, where the response is a leaf node represented by the first node, $t > 1$, $p \geq 3$, $t \leq p - 1$, there are $p - t - 1$ redundant or irrelevant variables.

Proof 4.1 *We can write the conditional density of the response:*

$$
\begin{aligned}
f_{1|2,\ldots,p}(x_1|x_2, \ldots, x_p) = &\Big[\prod_{j=3}^{t} c_{1,j;2,\ldots j-1}\big(F_{1|m+1,\ldots j-1}(x_1|x_2, \ldots, x_{j-1}),\\
&F_{j|m+1,\ldots j-1}(x_j|x_2, \ldots, x_{j-1})\big)\Big]\\
&\cdot \Big[\prod_{k=t+1}^{p} c_{1,j;2,\ldots k-1}\big(F_{1|m+1,\ldots k-1}(x_1|x_2, \ldots, x_{k-1}),\\
&F_{k|m+1,\ldots j-1}(x_k|x_2, \ldots, x_{k-1})\big)\Big]\\
&\cdot c_{1,2}\big(F_1(x_1), F_2(x_2)\big) \cdot f_1(x_1).
\end{aligned}
$$

Due to the independence pair-copulas in the tree levels higher than t in a t-truncated vine:

$$f_{1|2,\ldots,p}(x_1|x_2,\ldots,x_p) = \Big[\prod_{j=3}^{t} c_{1,j;2,\ldots j-1}\big(F_{1|m+1,\ldots j-1}(x_1|x_2,\ldots,x_{j-1}), F_{j|m+1,\ldots j-1}(x_j|x_2,\ldots,x_{j-1})\big)\Big] \cdot c_{1,2}\big(F_1(x_1), F_2(x_2)\big) \cdot f_1(x_1).$$

Thus, we have $f_{1|2,\ldots,p}(x_1|x_2,\ldots,x_p) = f_{1|2,\ldots,t+1}(x_1|x_2,\ldots,x_{t+1})$. If $X_2,\ldots,X_{t+1}$ are dependent (independent) on the variables $X_{t+2},\ldots,X_p$, there are $p-t-1$ redundant (irrelevant) variables for the response.

Example 4.3 *Suppose that the first node and others represent the response and three explanatory variables in the D-vine in Figure 2.7, respectively ($p = 4$). Assume that the D-vine is a 2-truncated vine ($t = 2$) and the copula for the pair (3, 4) is not independence. Then there is one redundant variable for the response, which is the variable represented by the fourth node, given the two relevant variables represented by the second and third nodes.*

Given the set of relevant variables, the redundant and irrelevant variables do not impact the conditional quantile of the response. However, their inclusion in a model decreases the model's interpretation power and, even, the predictive power due to the heuristic variable selection methods as shown in Section 4.7. Their inclusion also increases the model complexity.

4.6 Illustrative example

Now we illustrate how $vinereg$, $vineregRes$, and $vineregParCor$ select variables. We simulate the data with 450 observations from the model $(Y, X_1, X_2, X_3, X_4)^\top \sim \mathcal{N}_5(\mathbf{0}, \Sigma)$ with

$$\Sigma = \begin{pmatrix} 1.00 & 0.40 & 0.70 & 0.00 & 0.00 \\ 0.40 & 1.00 & 0.32 & 0.00 & 0.00 \\ 0.70 & 0.32 & 1.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 1.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 1.00 \end{pmatrix}.$$

Thus, there are a response variable and four explanatory variables in the data, and the third and fourth variables are assumed to be irrelevant for predicting the response.

First, we convert the observations to the copula scale $(v_i, u_{i,1}, u_{i,2}, u_{i,3}, u_{i,4})$, $i = 1,\ldots,450$ using the non-parametric kernel density estimator. Then we define the initial pseudo-response on the copula scale $\tilde{v}_i^{(1)} = v_i$, $i = 1,\ldots,450$, the initial D-vine order $\mathcal{D}^{(1)} = (0)$, the initial chosen variable index set $\mathcal{I}_{var}^{(1)} = \emptyset$, the initial set of candidate explanatory variables $p_{cand}^{(1)} = \{1,2,3,4\}$, and the given data's normal scores $(z_{i,0}, z_{i,1}, z_{i,2}, z_{i,3}, z_{i,4})^\top$, $i = 1,\ldots,450$.

For an illustration purpose, we will show the methods' variable selection step in the third iteration, where it holds that the D-vine order $\mathcal{D}^{(3)} = (0, 2, 1)$ for the D-vine copula $\hat{C}^{(3)}$ with the parameters $\hat{\boldsymbol{\theta}}^{(3)}$, the chosen variable index set $\mathcal{I}_{var}^{(3)} = \{2, 1\}$, and the set of candidate explanatory variables $p_{cand}^{(3)} = \{3, 4\}$. The conditional AIC of the current D-vine copula $\hat{C}^{(3)}$ is 975.30. The next iteration for all methods is to decide which and if the third or fourth variable should be added to the model.

To make the variable selection, $vinereg$ extends the D-vine order as $\mathcal{D}^{(4)} = (0, 2, 1, 3)$ for the third variable and $\mathcal{D}^{(4)} = (0, 2, 1, 4)$ for the fourth variable. Accordingly, three new pair-copula families are selected for each D-vine order, thereby estimating their parameters. Thus, $vinereg$ selects six bivariate copulas in this iteration (one for the first/second/third tree level for each D-vine order) to choose one variable. For more details about $vinereg$, see Example 3.2 in Kraus and Czado (2017).

To perform the variable selection, first, $vineregRes$ estimates the median of the response based on the D-vine copula $\hat{C}^{(3)}$. Then it calculates the pseudo-response and estimates the pseudo-response's marginal distribution nonparametrically. For this, the associated pseudo-response on the copula scale is obtained by the PIT, i.e.,

$$\tilde{y}_i^{(3)} = y_i - \hat{F}_Y^{-1}(\hat{C}_{0|2,1}^{-1(3)}(0.50|u_{i,2}, u_{i,1}; \hat{\boldsymbol{\theta}}^{(3)})), \quad \tilde{v}_i^{(3)} = \hat{F}_{Y^{(3)}}(\tilde{y}_i^{(3)}), \quad i = 1, \ldots, 450.$$

Then it fits a parametric bivariate copula to data $\{(\tilde{v}_i^{(3)}, u_{i,3}),\ i = 1, \ldots, n\}$ for the third variable and $\{(\tilde{v}_i^{(3)}, u_{i,4}),\ i = 1, \ldots, n\}$ for the fourth variable. It holds that the fitted copula for the third variable $\widehat{CR}^{3(3)}$ is a Frank copula with the estimated parameter $\hat{\theta}^{3(3)} = -0.41$. The conditional log-likelihood of $\widehat{CR}^{3(3)}$ conditioned on the third variable is -474.27. Moreover, the fitted copula for the fourth variable $\widehat{CR}^{4(3)}$ is the independence copula, resulting in the conditional log-likelihood of -475.30. Since the conditional log-likelihood of $\widehat{CR}^{3(3)}$ is higher than that of $\widehat{CR}^{4(3)}$, $vineregRes$ identifies the third variable as the candidate explanatory variable to be added to the D-vine copula. Thus, it extends the D-vine order as $\mathcal{D}^{(4)} = (0, 2, 1, 3)$, selecting the new pair-copula families and estimating new parameters. The resulting D-vine copula $\hat{C}^{(4)}$ has the conditional AIC of 975.30. Since $\hat{C}^{(4)}$ does not have a better conditional AIC than $\hat{C}^{(3)}$, $vineregRes$ stops the iterations and returns the final D-vine copula as $\hat{C}^{(3)}$. Thus, $vineregRes$ chooses five bivariate copulas here.

$vineregParCor$ chooses the next variable based on its partial correlation with the response given the chosen explanatory variables using their normal scores. It estimates $\hat{\rho}_{0,3;2,1} = 0.05$ for the third variable and $\hat{\rho}_{0,4;2,1} = 0.03$ for the fourth variable. Since the third variable has a higher (absolute) estimated partial correlation than the other, $vineregParCor$ extends the D-vine order as $\mathcal{D}^{(4)} = (0, 2, 1, 3)$. Next, it selects the new pair-copula families and performs new parameter estimation associated with $\mathcal{D}^{(4)}$. Since the D-vine copula with $\mathcal{D}^{(4)} = (0, 2, 1, 3)$ does not have a better conditional AIC than $\mathcal{D}^{(3)} = (0, 2, 1)$, $vineregParCor$ gives the final model fit $\hat{C}^{(3)}$. Thus, $vineregParCor$ selects three bivariate copulas in this iteration.

4.7 Simulation studies

We show the flexibility and effectiveness of the proposed methods on simulated datasets being nonlinear and having different sparsity. We explore the following questions: (Q1) How do $vineregRes$ and $vineregParCor$ work in situations with nonlinear explanatory variable effects on the response's quantiles in the presence of redundant and irrelevant variables for prediction accuracy and computational complexity? (Q2) How well do $vineregRes$ and $vineregParCor$ identify relevant and irrelevant variables for predicting the response's quantiles? (Q3) How do $vineregRes$ and $vineregParCor$ perform compared to alternative methods?

Data generating process (DGP)

For the questions (Q1, Q2, Q3), we study two nonlinear data generating processes.

DGP1: irrelevant variables

$$\begin{aligned} Y_i^d = X_{i,1} \cdot X_{i,2}^2 \cdot \sqrt{|X_{i,3}| + 0.1} + e^{0.4 \cdot X_{i,4} \cdot X_{i,5}} + \\ (X_{i,6}, \ldots, X_{i,p_d})(0, \ldots, 0)^\top + \epsilon_i \cdot \sigma_i, \quad i = 1, \ldots, n, \quad d = 1, 2, 3, \end{aligned} \tag{4.6}$$

where we sample the irrelevant variables $(X_{i,6}, \ldots, X_{i,p_d})^\top \sim \mathcal{N}_{p_d-5}(\mathbf{0}, \mathbb{I}_{p_d-5})$, relevant variables $(X_{i,1}, \ldots, X_{i,5})^\top \sim \mathcal{N}_5(\mathbf{0}, \Sigma)$, $i = 1, \ldots, n$ with the (a, b)th element of the covariance matrix $\Sigma_{a,b} = 0.75^{|a-b|}$, the random error terms $\epsilon_i \sim \mathcal{N}(0, 1)$ that are independent and identically distributed (iid), independently, and set $\sigma_i \in \{0.5, 1\}$, $i = 1, \ldots, n$. Thus, the relevant and irrelevant variables' variances are one. To analyze the methods' performance concerning the prediction accuracy and relevant variables' selection, we simulate data sets with different number of irrelevant variables and set it to $(p_d - 5)$ in each case $d = 1, 2, 3$: **Case 1** with $p_1 = 10$ (50% of variables are irrelevant), **Case 2** with $p_2 = 20$ (75% of variables are irrelevant), **Case 3** with $p_3 = 50$ (90% of variables are irrelevant).

DGP2: redundant variables

$$\begin{aligned} Y_i^d = \sqrt{|5 \cdot X_{i,1} - 2 \cdot X_{i,9} + 0.5|} + X_{i,8} \cdot (-4 \cdot X_{i,3} + 1) + e^{X_{i,6}} + (2 \cdot X_{i,10}^3 + X_{i,4}^3) \\ + (X_{i,7} + 1) \cdot (\ln\left(|X_{i,2} + X_{i,5}| + 0.01\right)) + \\ (X_{i,11}, \ldots, X_{i,p_d})(0, \ldots, 0)^\top + \epsilon_i \cdot \sigma_i, \quad i = 1, \ldots, n, \quad d = 1, 2, 3, 4, \end{aligned} \tag{4.7}$$

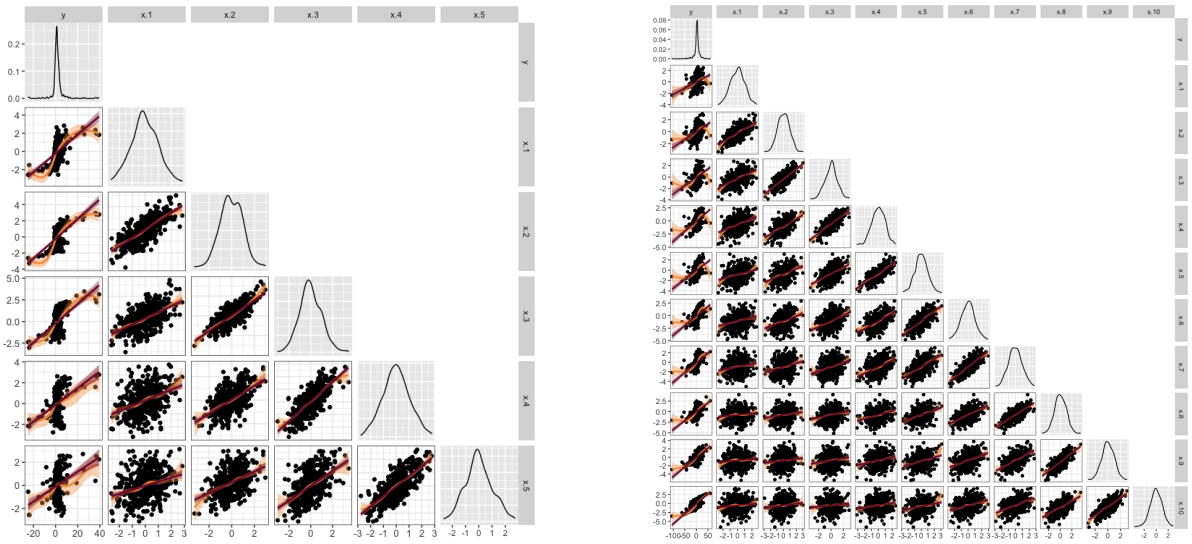
where the samples of explanatory variables are independently generated from a multivariate normal distribution with a Toeplitz correlation structure, i.e., $(X_{i,1}, \ldots, X_{i,p_d})^\top \sim \mathcal{N}_{p_d}(\mathbf{0}, \Sigma)$, $i = 1, \ldots, n$, $j = 1, 2, 3$, with the (a, b)th element of the covariance matrix $\Sigma_{a,b} = \rho^{|a-b|}$. To represent a challenging but realistic scenario, we set $\rho = 0.75$. We sample $\epsilon_i \sim \mathcal{N}(0, 1)$, $i = 1, \ldots, n$ (iid) independently from the explanatory variables and set $\sigma_i \in \{0.5, 1\}$, $i = 1, \ldots, n$. Here, all variables are predicting the response's quantiles. However, given the set of

the first ten variables, the others are redundant. To investigate the impact of the redundant variables on the methods' predictive power and sparsity of the model in terms of the total number of variables, we change the number of redundant variables and set it to $(p_d - 10)$ in each case $d = 1, 2, 3, 4$: **Case 1** with $p_1 = 20$ (50% of variables are redundant given the ten relevant ones), **Case 2** with $p_2 = 40$ (75% of variables are redundant given the ten relevant ones), **Case 3** with $p_3 = 100$ (90% of variables are redundant given the ten relevant ones), **Case 4** with $p_4 = 1000$ (99% of variables are redundant given the ten relevant ones).

Based on the DGPs in Equations (4.6) and (4.7), we simulate samples with size 450 (n=450) each time with a random split of 300/150 observations for a training set/a test set. We replicate our procedure 100 times and average performance measures per sample.

Figure 4.1 shows nonlinear relationships between the response and others for both DGPs.

Figure 4.1: Pairs plots of a simulated data set of the response and relevant explanatory variables with 450 observations under the DGP1 (left) and DGP2 (right) settings, where diagonal: variable's density estimates, lower: pairwise scatter plots with red curve showing a linear model fit and orange curve demonstrating a local polynomial regression fit.



Alternative methods

For comparison, we analyze the following alternative methods performing quantile regression.

Penalized quantile regression with LASSO function (LQRLasso): it extends the linear quantile regression method, adding a LASSO penalty function in the quantile regression objective function to perform the variable selection. Thus, it solves the following optimization

problem to estimate the coefficients at the quantile level α:

$$\underset{\boldsymbol{\beta}\in\mathbb{R}^{p+1}}{\arg\min}\left[\sum_{i=1}^{n}\rho_\alpha\left(y_i-\beta_0-\sum_{j=1}^{p}x_{i,j}\cdot\beta_j\right)+\sum_{j=0}^{p}\lambda\cdot|\beta_j|\right], \tag{4.8}$$

where $\rho_\alpha(r) = r(\alpha - \mathbf{1}(r < 0))$ denotes the check function, $\lambda \in \mathbb{R}$ corresponds to a tuning parameter for the penalization of the coefficients, $(y_i, x_{i,1}, \ldots, x_{i,p})$, $i = 1, \ldots, n$, are realizations of the random vector $(Y, X_1, \ldots, X_p)$, and Y denotes a response variable, and the others correspond to explanatory variables.

We perform a five-fold cross-validation using the training set to optimize λ in each replication and set the associated coefficients below 10^{-8} to zero. The variables with zero coefficients are not relevant for predicting the response's quantiles. Such procedures are implemented in the `R` package `rqPen` (Sherwood and Maidman 2020). Since it is not straightforward to decide which transformation of variables or some interaction effects should be applied in a linear model in the high-dimensional data, we analyze the variables without any transformations. Since a LASSO penalty function is sensitive to the feature scales, we apply a standardization with zero mean and unit variance for each feature before its fit.

Quantile regression forest (QRF): it is a nonlinear quantile regression method built upon random forest ideas (Meinshausen 2006). The core idea is to approximate the conditional distribution function of the response given explanatory variables. It starts with fitting a random forest regression model to data for prediction. Then for each response observation to be predicted, it finds the terminal node, the node without any splits, in each tree. Next, each observation in the terminal node gets a weight inversely proportional to the total number of observations in the terminal node. Such weights are calculated for each tree for the associated observations and then normalized to one. Accordingly, each observation used for training the model gets a weight between zero and one, which sums up to one for all observations. Finally, the weights are used to obtain the empirical conditional distribution function of the response given explanatory variables, where quantile regression estimates are calculated for the given observation to be predicted. The procedure is repeated for each new sample. More details are given by Meinshausen (2006), and the ideas are implemented in the package `quantregForest` (Meinshausen 2022). We work with the package's default specifications; however, we set the minimum number of observations in terminal nodes to one-twentieth of the number of observations. We have not performed any cross-validation.

QRF assesses the feature importance implemented in the package `randomForest` (Liaw and Wiener 2002). The idea is to calculate the decrease in the residual sum of squares (RSS) for each feature, which is the sum of the squares of the difference between the response values and the predicted response values. Specifically, the predicted values, thanks to the split through the feature, are considered in the RSS in a given tree and for a given feature. For instance, if the feature is split two times in a given tree, the decrease in the RSS is evaluated two times for that feature. The process is applied to all features and trees, and then the decrease in the RSS for each feature is divided by the number of trees grown to

calculate the features' importance.

Performance measures

We consider the computation time, the number of chosen variables, the true positive rate (TPR), and the false discovery rate (FDR) as methods' performance measures on the training set. To evaluate the performance of a method on the test set, we apply the pinball loss (PL_α) at $\alpha = 0.05, 0.50, 0.95$.

True positive rate (TPR) is the ratio of the number of chosen relevant variables by a method M to the total number of relevant variables. *False discovery rate (FDR)* is the ratio of the number of chosen irrelevant variables to the total number of chosen variables by a method M. Formally:

$$TPR(M) = \frac{\sum\limits_{\substack{X_j \in \mathcal{X}_{rel} \\ X_j \in \mathcal{X}^M_{chosen} \\ j=1,\ldots,|\mathcal{X}^M_{chosen}|}} 1}{|\mathcal{X}_{rel}|} \quad \text{and} \quad FDR(M) = \frac{\sum\limits_{\substack{X_j \in \mathcal{X}_{irrel} \\ X_j \in \mathcal{X}^M_{chosen} \\ j=1,\ldots,|\mathcal{X}^M_{chosen}|}} 1}{|\mathcal{X}^M_{chosen}|},$$

where $\mathcal{X}_{rel}$ is the set of relevant variables, $\mathcal{X}_{irrel}$ is the set of irrelevant variables, and $\mathcal{X}^M_{chosen}$ denotes the set of chosen variables by a method M. Higher TPR and smaller FDR are better.

The pinball loss (PL_α) measures the accuracy of the quantile predictions $\hat{y}_i^{\alpha,M}$ at the level α by a method M compared to the given response y_i, $i = 1, \ldots, n$ and has the form of

$$PL_\alpha(M) = \frac{\sum_{i=1}^n (\hat{y}_i^{\alpha,M} - y_i)(\mathbb{I}_{[0,\infty)}(\hat{y}_i^{\alpha,M} - y_i) - \alpha)}{n}.$$

The pinball loss is identical to check score, and smaller pinball loss values are better (Steinwart and Christmann 2011).

Results

All computations are run on a single node CPU with Intel Xeon Platinum 8380H Processor with around 25 GB RAM, running `R` version 4.2.2.

Variable selection and computational complexity results on the training set

Table 4.1 compares the average performance of the methods for cases 1—3 and 1—4 specified in Equations (4.6) and (4.7), respectively, concerning the variable selection and computational complexity. We analyze the TPR and FDR only for the DGP1 setting since all variables are relevant in the DGP2 setting. Further, $vinereg$ was not complete within three days per replication for the fourth case of the DGP2 setting, making it computationally infeasible. Likewise,

we did not run $LQRLasso$ for that case since it ran around seven hours per replication and had worse performances than others in other cases.

In addition, since we do not perform any variable selection for quantile regression forests, in all cases, QRF considers all variables in the associated DGP to make predictions. As a result, its TPR is always one, the number of selected variables equals the total number of variables in a sample, and its FDR is the proportion of the irrelevant variables in the associated DGP setting. Still, we rank the variables based on their variable importance given by the model from highest to lowest for each case of both DGPs per replication. Then when we analyze the median rank of variables out of 100 replications, we observe that the five relevant variables in the DGP1 setting have their rank between one and five, and the ten most relevant variables in the DGP2 setting have their rank between one and ten, showing that QRF's variable importance identifies the most important variables correct in both DGPs. However, comparing different variable selection methods for random forests, such as proposed by Genuer et al. (2010) and Conn et al. (2019), would be preferable in future studies. Speiser et al. (2019) study some approaches and conclude that their performance varies regarding computational complexity and accuracy by analyzing different data sets, making it unfair to analyze only one.

Excluding QRF, in all cases of the DGP1 setting, $vinereg$ has a better performance than the others regarding the true positive rate. However, its false discovery rate is also higher than others, adding many irrelevant variables to a model. For $vineregRes$, we observe that it correctly identifies more than 75% of the relevant variables in all cases of the DGP1 setting. Furthermore, the most promising is that its false discovery rate is less than 15% there, making it the best method for the FDR. Further, $vineregParCor$'s TPR is higher than 50% in all cases of the DGP1 setting. Nonetheless, like others, its FDR increases as the number of irrelevant variables increases in the model, reaching more than 50% in the third case of the DGP1 setting. The other method, $LQRLasso$, identifies at least 48% of the relevant variables in all cases. However, its TPR decreases with the increase in the number of irrelevant variables.

Concerning the number of chosen variables by a method in the DGP1 and DGP2 settings: While $vineregRes$ selects the lowest number of variables between four and six as reflected in its high TPR and low FDR in the DGP1 setting, $vinereg$ includes almost half of the total number of variables in the data in each case. This highlights the power of $vineregRes$ regarding the exclusion of irrelevant variables in sparse data sets. $vineregParCor$'s number of chosen variables is between $vinereg$ and $vineregRes$ in all evaluated cases. The same result applies to $LQRLasso$. However, $LQRLasso$ selects more variables for estimating median predictions than other quantiles. Moreover, in an ultra high-dimensional case with 1000 explanatory variables, i.e., the last case in the DGP2 setting, the number of variables chosen by $vineregParCor$ is, on average, 31.72 with the empirical standard error of 1.23 while it is 7.08 with that of 0.54 for $vineregRes$. Hence, our methods are able to perform high-dimensional sparse regression as proposed.

As the number of variables increases, the average running time for all methods increases. Among vine copula based methods, $vineregParCor$ provides the fastest computation as expected from the results in Section 4.4. However, QRF provides the fastest computation among all models considered. In the second case of the DGP2 setting, where 75% of variables are redundant given the ten relevant ones, $vineregRes$ is, on average, 15 times faster than $vinereg$. In the third case of the DGP2 setting with 100 explanatory variables, $vinereg$ takes around three hours, whereas our methods run less than one minute on average. Likewise, $vineregRes$ and $vineregParCor$ run less than 15 minutes in the ultra high-dimensional case. The computation time for different quantile levels does not differ much for $LQRLasso$.

Table 4.1: Comparison of the methods' performance on the training set over 100 replications under the cases 1—3 and 1—4 specified in Equations (4.6) and (4.7), respectively. The numbers in parentheses under a method's name column are the corresponding empirical standard errors. (-) shows computational infeasibility. $LQRLasso$ column corresponds to the quantile levels (0.05, 0.50, 0.95). Chosen Vars. corresponds to the total number of chosen variables. Time is in minutes and per replication.

DGP	Measure	Case	$vinereg$	$vineregRes$	$vineregParCor$	QRF	$LQRLasso$ (0.05, 0.50, 0.95)
1	TPR	1	0.81 (0.01)	0.80 (0.02)	0.67 (0.02)	1.00 (0.00)	0.73 (0.02), 0.69 (0.02), 0.63 (0.02)
		2	0.85 (0.01)	0.79 (0.03)	0.59 (0.02)	1.00 (0.00)	0.68 (0.02), 0.68 (0.02), 0.55 (0.02)
		3	0.89 (0.01)	0.78 (0.02)	0.56 (0.02)	1.00 (0.00)	0.61 (0.02), 0.61 (0.02), 0.48 (0.01)
	FDR	1	0.28 (0.01)	0.08 (0.01)	0.24 (0.02)	0.50 (0.00)	0.28 (0.02), 0.32 (0.02), 0.24 (0.02)
		2	0.55 (0.01)	0.13 (0.02)	0.45 (0.02)	0.75 (0.00)	0.38 (0.02), 0.49 (0.03), 0.34 (0.03)
		3	0.80 (0.00)	0.15 (0.02)	0.65 (0.02)	0.90 (0.00)	0.35 (0.03), 0.58 (0.02), 0.40 (0.03)
	Chosen Vars.	1	5.83 (0.13)	4.56 (0.19)	4.68 (0.16)	10.00 (0.00)	5.24 (0.22), 5.48 (0.21), 4.99 (0.26)
		2	9.76 (0.19)	5.04 (0.26)	6.03 (0.23)	20.00 (0.00)	6.60 (0.41), 8.48 (0.43), 5.30 (0.36)
		3	23.24 (0.42)	5.29 (0.33)	8.74 (0.30)	50.00 (0.00)	5.93 (0.38), 10.20 (0.68), 5.97 (0.47)
	Time	1	0.18 (0.00)	0.17 (0.01)	0.06 (0.00)	0.01 (0.00)	0.02 (0.00), 0.02 (0.00), 0.02 (0.00)
		2	0.97 (0.02)	0.30 (0.02)	0.10 (0.01)	0.01 (0.00)	0.02 (0.00), 0.43 (0.03), 0.02 (0.00)
		3	12.06 (0.31)	0.77 (0.05)	0.34 (0.03)	0.03 (0.00)	0.12 (0.00), 0.14 (0.00), 0.12 (0.00)
2	Chosen Vars.	1	10.94 (0.23)	5.41 (0.19)	6.68 (0.20)	20.00 (0.00)	7.05 (0.32), 10.12 (0.36), 8.95 (0.40)
		2	19.63 (0.54)	5.17 (0.17)	8.25 (0.28)	40.00 (0.00)	7.88 (0.43), 12.36 (0.51), 9.88 (0.48)
		3	62.92 (2.78)	5.83 (0.22)	11.66 (0.42)	100.00 (0.00)	7.35 (0.36), 15.45 (0.78), 9.44 (0.55)
		4	-	7.08 (0.54)	31.72 (1.23)	1000.00 (0.00)	-
	Time	1	1.15 (0.03)	0.20 (0.01)	0.16 (0.01)	0.01 (0.00)	0.09 (0.00), 0.10 (0.00), 0.08 (0.00)
		2	7.30 (0.26)	0.32 (0.01)	0.29 (0.03)	0.02 (0.00)	0.11 (0.00), 0.13 (0.00), 0.11 (0.00)
		3	159.22 (8.66)	0.84 (0.03)	0.72 (0.06)	0.05 (0.00)	0.35 (0.00), 0.35 (0.00), 0.34 (0.00)
		4	-	9.44 (0.69)	12.18 (1.26)	0.44 (0.00)	-

Prediction accuracy results on the test set

We see in Table 4.2 that $vineregRes$ provides the best fit in eight evaluations out of nine, three measures evaluated for three cases, in the DGP1 setting among vine copula based methods. $vinereg$ and $vineregParCor$ have the same accuracy as $vineregRes$ for the first case in the DGP1 setting. However, as the number of irrelevant variables increases, a residual-based variable selection may be better than other vine copula based methods. Moreover, $LQRLasso$ has the lowest accuracy in all cases of the DGP1 setting because of the high nonlinearity in

samples. Even though $vineregParCor$'s performance is better than $LQRLasso$, it provides worse fits than the others in the upper quantile. A likely explanation can be that including irrelevant variables in addition to the most relevant ones in a vine copula may negatively impact the prediction accuracy in addition to an increase in the computational complexity unnecessarily as shown in Table 4.1. However, a similar result does not apply to QRF. Despite considering all irrelevant variables in the model, QRF still performs better than all in seven evaluations out of nine.

Table 4.2 also shows that $vineregRes$ provides the lowest pinball loss at all quantiles in all cases of the DGP2 setting, except the upper quantile in the first case. Since $vineregRes$ gives the most sparse models in the DGP2 setting in Table 4.1, we can infer that including many relevant but potentially redundant variables in $vinereg$, $vineregParCor$, and QRF is worsening the prediction accuracy in the DGP2 setting. Like in the DGP1 setting, $LQRLasso$ suffers from nonlinearity in all cases.

In simulation studies, the relevant, irrelevant, and redundant variables are known. Accordingly, when only the relevant ten variables are used for prediction in the DGP2 setting, QRF has the pinball loss of 0.64, 1.81, and 0.62 at levels $0.05, 0.50$, and 0.95, respectively. Thus, $vineregRes$ would have better accuracy than QRF in most cases of the DGP2 setting, even if the latter selected the most relevant variables.

Thus, a vine copula based prediction method with a variable selection, e.g., $vineregRes$, is more advantageous than quantile regression forests in the presence of many correlated variables in our simulations.

4.8 Open problems

There are open problems for high-dimensional sparse vine copula regression, and we discuss some by proposing their starting point in the following.

Adaptation for more flexible vine tree structures

The flexibility of the proposed variable selection methods can be increased by using more flexible vine tree structures than D-vines, such as R-vines.

A starting point for $vineregParCor$ can be to choose a variable whose inclusion in a candidate vine tree structure maximizes the sum of absolute (partial) correlations based on normal scores, ensuring that a leaf node in each tree level includes the response in its conditioned set. The latter guarantees that the conditional quantile of the response can be obtained analytically from the model. Then the candidate variable with the corresponding vine tree structure can be evaluated for the model fit regarding the CAIC. If it improves the CAIC, then the candidate variable is added to the associated vine tree structure. Later, the next candidate variable can be searched similarly. Otherwise, the algorithm stops.

Table 4.2: Comparison of the average performance of the methods on the test set for the pinball loss (PL_α) at different quantile levels α over 100 replications under the cases 1—3 and 1—4 specified in Equations (4.6) and (4.7), respectively. The best performance for each quantile level and DGP case is highlighted. The numbers in parentheses under a method's name column are the corresponding empirical standard errors. (-) shows computational infeasibility.

DGP	Measure	Case	$vinereg$	$vineregRes$	$vineregParCor$	QRF	$LQRLasso$
1	$PL_{0.05}$	1	0.21 (0.01)	0.21 (0.01)	0.21 (0.01)	0.22 (0.01)	0.34 (0.01)
		2	0.23 (0.01)	0.22 (0.01)	0.22 (0.01)	0.22 (0.01)	0.34 (0.02)
		3	0.24 (0.01)	0.21 (0.01)	0.22 (0.01)	0.22 (0.01)	0.32 (0.01)
	$PL_{0.50}$	1	0.79 (0.04)	0.79 (0.03)	0.81 (0.04)	0.76 (0.04)	0.94 (0.02)
		2	0.84 (0.04)	0.79 (0.03)	0.82 (0.04)	0.69 (0.02)	0.97 (0.04)
		3	0.84 (0.02)	0.76 (0.02)	0.79 (0.02)	0.72 (0.02)	0.90 (0.02)
	$PL_{0.95}$	1	0.43 (0.07)	0.43 (0.06)	0.46 (0.07)	0.41 (0.07)	0.59 (0.04)
		2	0.37 (0.04)	0.39 (0.04)	0.44 (0.04)	0.32 (0.03)	0.64 (0.07)
		3	0.38 (0.03)	0.38 (0.03)	0.41 (0.03)	0.35 (0.03)	0.53 (0.03)
2	$PL_{0.05}$	1	0.53 (0.01)	0.53 (0.01)	0.54 (0.01)	0.70 (0.02)	0.84 (0.02)
		2	0.57 (0.01)	0.54 (0.01)	0.56 (0.01)	0.70 (0.02)	0.85 (0.02)
		3	0.81 (0.04)	0.54 (0.01)	0.58 (0.01)	0.72 (0.01)	0.92 (0.03)
		4	-	0.55 (0.01)	0.89 (0.02)	0.78 (0.02)	-
	$PL_{0.50}$	1	1.87 (0.02)	1.83 (0.02)	1.84 (0.02)	1.91 (0.02)	2.20 (0.02)
		2	1.99 (0.02)	1.84 (0.02)	1.86 (0.02)	1.93 (0.03)	2.26 (0.03)
		3	2.59 (0.09)	1.84 (0.02)	1.94 (0.02)	2.00 (0.03)	2.29 (0.02)
		4	-	1.89 (0.03)	2.42 (0.03)	2.17 (0.03)	-
	$PL_{0.95}$	1	0.53 (0.01)	0.55 (0.01)	0.55 (0.01)	0.65 (0.01)	0.78 (0.02)
		2	0.57 (0.01)	0.56 (0.01)	0.56 (0.01)	0.67 (0.01)	0.82 (0.02)
		3	0.81 (0.05)	0.57 (0.01)	0.61 (0.02)	0.68 (0.01)	0.83 (0.02)
		4	-	0.57 (0.02)	0.88 (0.03)	0.75 (0.02)	-

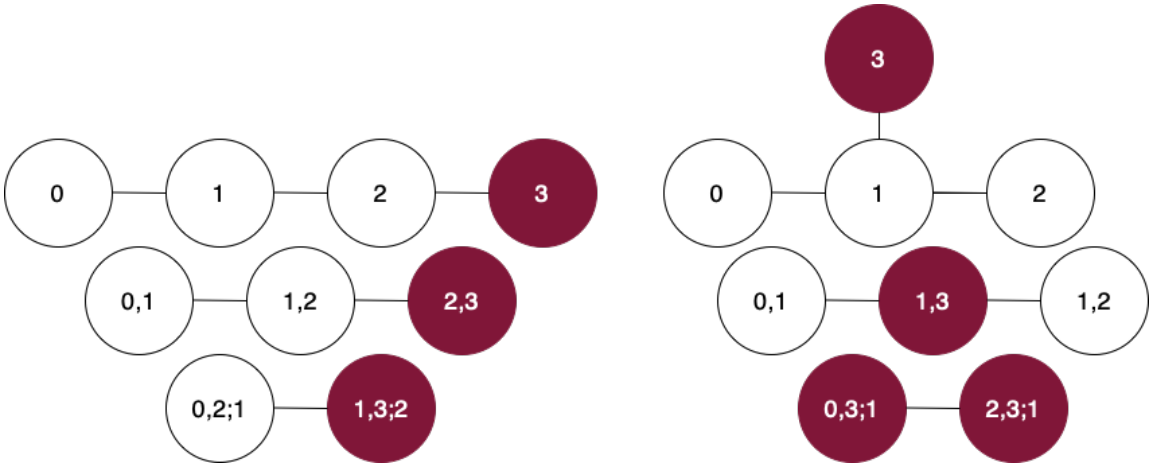
Consider the data example given in Section 4.6. The first and second variables are already added to the model. To allow for flexible vine tree structures and decide on a better one among others using $vineregParCor$, Figure 4.2 shows how the third variable can be incorporated into a vine copula model, where red nodes are the results of such incorporation. To provide guidelines on the idea, we focus on two possible vine structures here; however, one more vine tree structure satisfies the proximity, where a leaf node in each tree level includes the response in its conditioned set. The vine tree structure on the left panel of Figure 4.2 is a D-vine, and $vineregParCor$ estimates the correlation among the response and first variable $\rho_{0,1}$, the first and second variables $\rho_{1,2}$, second and third variables $\rho_{2,3}$ and partial correlations among the response and second variable given the first $\rho_{0,2;1}$, the first and third variables given the second $\rho_{1,3;2}$, and among the response and third variable given the first two $\rho_{0,3;1,2}$ based on the normal scores.

On the other hand, the right panel of Figure 4.2 is a C-vine, where the estimated correlations, including partial ones, are $\rho_{0,1}$, $\rho_{1,2}$, $\rho_{1,3}$ in the first tree, $\rho_{0,3;1}$, $\rho_{2,3;1}$ in the second tree, and $\rho_{0,2;1,3}$ in the last tree level. The sum of absolute (partial) correlation values resulting

from the inclusion of the third variable is 0.59 for the D-vine and 0.72 for the C-vine. Likewise, the candidate variable to be added can be the fourth instead of the third one in Figure 4.2. In this case, the sum of absolute (partial) correlation values due to its inclusion is 0.60 for the D-vine and 0.55 for the C-vine. Since the third variable attains the maximum value of the sum of absolute correlation values for C-vine, the third variable is chosen as the candidate, and the vine structure to consider in the next step is the C-vine shown on the right panel of Figure 4.2. Next, after fitting pair-copulas, the resulting CAIC of the C-vine is compared with the previous model to decide if the algorithm stops.

Even though the number of (partial) correlation calculations is high, it may pay off thanks to allowing flexible vine tree structures. Moreover, the conditional log-likelihood of the response includes the pair-copulas whose conditioned set has the response so that the conditional quantile of the response can be obtained analytically from the model. Thus, one might further reduce computational efforts considering only the correlations modeled by such copulas. For instance, for the inclusion of the third variable in Figure 4.2, the pair copulas $C_{0,1}$, $C_{0,2;1}$, and $C_{0,3;1,2}$ on the left panel and $C_{0,1}$, $C_{0,3;1}$, and $C_{0,2;1,3}$ on the right panel appear on the conditional log-likelihood of the response. Thus, the sum of absolute values of $\rho_{0,1}$, $\rho_{0,2;1}$, and $\rho_{0,3;1,2}$ on the left panel may be compared with the that $\rho_{0,1}$, $\rho_{0,3;1}$, and $\rho_{0,2;1,3}$ on the right panel. Next, the vine copula, where the maximum sum is attained for its tree structure, is fitted and then compared with the previous iteration's vine copula regarding the CAIC.

Figure 4.2: Potential vine tree structures for the inclusion of the third variable, allowing the response denoted by the node 0, to be on the conditioned set of a leaf node in each tree level. The nodes filled by red result from the inclusion of the new variable.



Comparison of vine copula based prediction models

Even though the pinball loss is widely used to assess the quality of quantile regression models, it does not highlight how a vine copula based prediction model is better than the other. An

appropriate asymptotic theory might be derived for vines to construct confidence intervals in predicting a given quantile level. Then the length of the confidence intervals may be used to assess the different vine copula based prediction models. The estimation of standard errors in vine copulas studied by Stöber and Schepsmeier (2013) can be a starting point.

Increase in sparsity in vine copula based prediction models

In a vine copula with p variables, assuming each pair-copula has two parameters, there are $p \cdot (p-1)$ parameters to estimate. As p gets large, its computational burden and the risk of overfitting increase. Even though our methods identify the most relevant variables for prediction well, a huge number of such variables also results in the same problems for us. An approach to deal with such issues is to stop adding variables into the model, e.g., when $p_{model} \approx \sqrt{2n}$, as a rule of thumb, where p_{model} is the number of variables in a vine copula, and n is the number of observations. Further, our methods can be combined with sparse vine copula models proposed by Nagler et al. (2019), and appropriate penalty terms can be proposed and incorporated into the selection of vine copula models tailored to prediction tasks.

4.9 Conclusion

We propose two methods to perform high-dimensional sparse vine copula based regression and analyze their performance in the presence of redundant and irrelevant variables through simulation studies. We show that vine copula based regression methods have better accuracy than linear models in analyzing nonlinear relationships between the response and explanatory variables. Further, our methods perform better than quantile regression forests in the presence of many correlated variables. Finally, we discuss future research directions to extend our methods with high-dimensional data sets to more flexible vine tree structures.

Part II

Societal applications

Chapter 5

Environmental, social, and governance (ESG) data analysis

This chapter includes the materials from Sahin et al. (2022) and Sahin et al. (2023), but Section 5.7 gives a case study characterizing different dependence structures among Energy companies in the Standard&Poor (S&P) 500 based on their sustainability levels and stock prices using VCMM explained in Chapter 3.

5.1 Motivation

As sustainability concerns increase globally, sustainable finance and Environmental, Social, and Governance (ESG) investing strategies gained much interest. According to Bloomberg, the "ESG ETF market had risen over 318% in 2020", indicating the significant interest by investors (Bloomberg 2021). To assess the companies' annual ESG performance and sustainability, investors can use the ESG scores data providers make available, using the publicly available data and voluntary disclosure.

Until now, most scholars have focused on the link between ESG scores and corporate financial performance (Friede et al. 2015). Lately, the European Banking Authority (EBA) has acknowledged the role of ESG scores in impacting companies' riskiness and identified a need to incorporate ESG risks into overall business strategies and risk management frameworks (EBA 2020). ESG scores may affect institutions' financial performance by manifesting themselves in financial risks, such as "credit risk, market risk, operational risk, liquidity, and funding risks," and managing the ESG risk can act as a driver for managing financial risk (Page 10 of EBA (2020)). Accordingly, Kumar et al. (2016) indicate a significant negative correlation between ESG scores and volatility. Diemont et al. (2016) and Verheyden et al. (2016) use the Value-at-Risk (VaR) to measure tail risk and study its link with ESG scores. Bax et al. (2023) point out that ESG risk dependence can be quantified and is not negligible in times of crisis. Nonetheless, Ayton et al. (2022) do not find any statistically significant relationship between

corporate social performance, measured by ESG scores, and the systematic risk of companies.

Recently, ESG data quality issues have raised another big discussion in the ESG literature. Berg et al. (2022) report a large discrepancy between ESG scores from different data providers. Abhayawansa and Tyagi (2021) present the main reason for the divergence as different measurement methods. Gyönyörová et al. (2021) discuss that such a divergence changes from sector to country. Finally, Billio et al. (2021) argue that such discrepancies might make the usage of ESG scores in portfolios difficult for fund managers.

Data methodology and its pitfalls

Another important issue regarding the ESG data is if ESG scores for a certain year within the same provider change over time. For instance, working with the same provider, one could expect companies' ESG scores for 2017 to be the same, independent of the data extraction date. If so, research findings working with the same companies in similar periods should be consistent regarding ESG scores used from the same provider. In this manner, many studies use Thomson Reuters (Refinitiv; previously ASSET4) as a reliable data provider (Berg et al. 2021). Its ESG scores are built on the aggregation of *Environmental*, *Social*, and *Governance* pillar subscores. Nonetheless, Berg et al. (2021) argued that Refinitiv might have a strong incentive to show that their ESG scores exhibit a relationship with financial performance measures, making them more attractive to investors. Moreover, they show that Refinitiv's ESG scores of the previous years are *re-written* thereof. For instance, using Refinitiv's ESG scores and S&P's credit ratings, Aslan et al. (2021) find that the *Social* pillar impacts the probability of credit default for US firms. However, Bannier et al. (2021) do not find the same association for US firms using the same provider, S&P credit ratings, and similar time windows. Thus, although their methodology and number of observations in the studies are slightly different, one might ask if both studies' ESG scores and pillar subscores for the same companies in the same years are consistent with each other.

Indeed, focusing on Refinitiv's ESG scoring methodology described in Refinitiv (2021a), Refinitiv states that its ESG data of the five most recent years are regularly updated, i.e., non-definitive. Refinitiv's methodology to *re-write* its ESG data consists of a weighting scheme to compute pillar subscores and ESG scores, the initial disclosure of ESG information, and an update in the published ESG information. Thus, there are ongoing and unannounced ESG data changes for the five most recent years. It is reasonable for Refinitiv to *re-write* its recent ESG data to provide its customers, most likely investors, with the most updated ESG information. However, Refinitiv's ongoing ESG data modifications for the five most recent years imply that researchers and investors might be using different ESG scores for the same companies in the same years. Naturally, then, it might lead to different research conclusions.

In this chapter, first, we aim to analyze the impact of (non-definitive) ESG scoring methodology on (i) the link between ESG scores and risk and (ii) the potential of manipulative ESG data changes using simple optimization and exploratory data mining approaches. While Berg et al. (2021) document the changes in the Refinitiv ESG data, to the best of our knowledge,

there has not yet been any study discussing the implications of a (non-definitive) ESG scoring methodology for researchers and investors in detail and investigating the potential ESG data mining approaches from a provider perspective. We focus on the risk aspect due to the growing number of recent studies (e.g., Farah et al. (2021), Jarjir et al. (2020), Morelli and D'Ecclesia (2021), and Rehman et al. (2020)), but other financial performance measures can also be considered. More complex optimization schemes, such as in Ahmed et al. (2021) and Pedersen et al. (2021), can be easily adapted.

Working with the top market capitalization companies globally, constituents of the S&P 500, we show that ESG scores from Refinitiv might diverge over time, and ESG research findings using its non-definitive ESG data might not be replicable in the future. More specifically, if the weights given to pillar subscores to compute an ESG score change, the number of sectors providing significant positive dependence between ESG scores and risk might increase. Moreover, it is a common practice in the literature to classify the companies into quartiles based on their ESG scores to assess the impact of high/low levels of social responsibility on financial performance measures and vice versa (e.g., Barko et al. (2021), Díaz et al. (2021), and Lins et al. (2017)). However, we show that such quartiles' companies change by modifying the pillar subscore weights. Therefore, we conjecture that inconsistent relationships between ESG and financial performance measures in the literature might result from the ESG data extracted at different time points and used in the studies. Furthermore, how well socially responsible a company is compared to its peers might depend on its industry group's pillar subscore weights.

Additionally, we discuss that a company's initial disclosure of ESG information results in *rewritten* ESG scores for its peers despite the fact that the peers did not change any underlying ESG information. Hence, such ESG information changes might make some companies more sustainable than before despite nothing has changed. Finally, we show that as companies initially disclose ESG information or update the published ESG information, the relationship between ESG scores and risk changes.

Given the growing literature using the US companies' non-definitive ESG scores (e.g., Bae et al. (2021); Demers et al. (2021); Garel and Petit-Romec (2021); Li et al. (2021); Lööf et al. (2021); Mohr et al. (2022)), we hypothesize that conclusions regarding the data used might depend on sectors and the date and amount of ESG information disclosure. Hence, they need to be evaluated carefully. Also, some results might apply to other countries and providers since they use similar ESG scoring methodologies, but this is open to further research.

The quantification of missing data and its impact

While the above-described debate is still going on, ESG scores still play a crucial role in investors' investment strategies. Based on the global survey conducted by senior investment professionals, Amel-Zadeh and Serafeim (2018) find that the negative screening (only including companies with a high ESG score in a portfolio and excluding companies with a low ESG score (PRI 2021)), either across sectors or within a sector, is still the most used method to

integrate ESG information into portfolios compared to positive screening, active ownership, and full integration. As an outlook on the future, investors argue that they expect positive screening and active ownership to gain importance (Amel-Zadeh and Serafeim 2018). Moreover, Alessandrini and Jondeau (2020) and Alessandrini et al. (2021) discuss that the performance of ESG exclusion strategies varies across geographies and sectors. They find that screening often leads to a better risk profile of the portfolios and often generates protection against credit risks (Alessandrini et al. 2021). Lastly, they recommend screening as the best strategy for passive investors with ESG preferences (Alessandrini et al. 2021). Still, the debate about negative screening has been ongoing since the exclusion strategies based on ESG scores can lead to capital and risk misallocations if ESG scores are not representative of company characteristics (Alessandrini and Jondeau 2020).

Since ESG scores might be subject to changes due to a release of new ESG information, i.e., the release of missing ESG information (Berg et al. 2021), in this chapter, we also focus on studying the role and the amount of missing ESG information as a potential source for a release of new ESG information with impacts on ESG scores in the future. Thus, we introduce a new pillar quantifying the missing ESG information, called the *Missing (M) pillar*, and define new scores: *Environmental, Social, Governance, Missing (ESGM)* scores by simultaneously aggregating the M pillar with the three ESG pillars. The ESGM scores are easily interpretable as a convex combination of the E, S, G, and the newly introduced M pillar subscores. We propose an optimization scheme to link ESGM scores and risk measures and run an in-sample and out-of-sample analysis to ensure robust results. If the amount of missing information is explicitly considered, companies are encouraged to disclose new information by our ESGM scoring construction methodology, as it positively impacts the score. This assumption is reasonable considering the current ESG scoring construction methodology that positively rewards the disclosure of new information and the fact that often missing information is not due to the unwillingness to release such information but its unavailability.

For this part, we work with the Refinitiv ESG data of the constituents of the S&P 500 and EuroStoxx 600 in the period 2017-2019, i.e., when the missing ESG information can still be released, and the ESG scores can be updated. We show that ESG and risk dependence and the amount of missing ESG information change with sectors and geographical regions. We also show that ESGM scores provide better risk profiles for companies than ESG scores.

Moreover, investors and practitioners can benefit from this research for the negative screening since using the widely available Refinitiv ESG data, which do not include the potential of new information disclosure, would mean that possible companies with potentially high scores after new ESG information adoption will be missed. Overall, this could lead to a more risky and less effective portfolio. ESGM scores identify the risky companies better than ESG scores for exclusion strategies. Nonetheless, the impact of missing ESG information on negative screening varies in sectors and regions. We further discuss the implications of our research for researchers, investors, companies, and managers.

The chapter is organized as follows: we introduce the methodology behind the ESG

scoring construction from Refinitiv in Section 5.2 and describe the ESG data in Section 5.3. Section 5.4 explores the data. We discuss the pitfalls of ESG scoring methodology in Section 5.5, while Section 5.6 presents the impact of missing ESG data on risk measures and their quantification. Next, we apply the model-based clustering with vine copulas to evaluate the dependence structure of companies based on their ESG scores in Section 5.7. Finally, we discuss our findings in Section 5.8 and conclude in Section 5.9.

5.2 Review of Refinitiv's ESG scoring methodology

Refinitiv gathers publicly available ESG information of companies and calculates a range of ESG performance indicators. The performance indicators are aggregated with a ranking scheme per Thomson Reuters Business Classifications Industry Group or the respective Country Group (Refinitiv 2021a) such that a category score, which is for *Resource Use*, *Environmental Innovation*, *Emission*, *Workforce*, *Human Rights*, *Community*, *Product Responsibility*, *Management*, *Shareholders*, or *Corporate Social Responsibility*, is based on the company's relative performance compared to its peers. A category score takes values from zero to 100, where a zero category score means that the company has not disclosed any information about the category's indicators. Then the indicator values are denoted by NULL or N/A. We note that Refinitiv shared the information as a reply to our inquiry on 07 May 2021.

Refinitiv pillar subscores are weighted sums of the respective category scores as illustrated in Figure 5.1. The category score weights to find pillar subscores are between zero and one, different per category, sum up to one and can be modified annually (Refinitiv 2021a). A pillar subscore, which is for the *Environmental (E)*, *Social (S)*, or *Governance (G)* pillar, also takes values from zero to 100, where a zero pillar subscore means that the company has not yet disclosed any information with regard to the pillar's and respective categories' indicators.

ESG scores are convex combinations of the three pillar subscores and range from zero to 100. The pillar subscore weights to find ESG scores can be updated annually (Refinitiv 2021a). The higher the ESG score is, the more ESG responsible the company is evaluated.

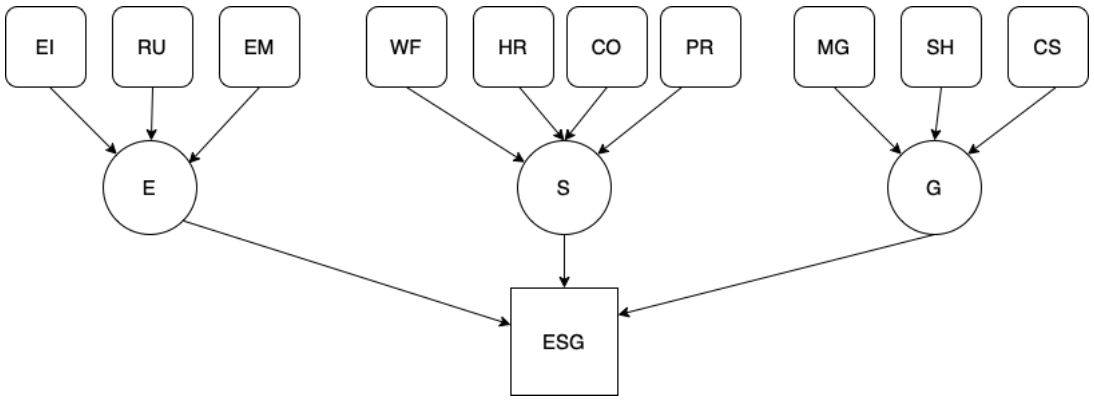
Example 5.1 presents an exemplary ESG score calculation using the fictitious pillar subscore weights for the industry group Household Goods and a generic name for the company.

Example 5.1 *(ESG score calculation of Refinitiv). The ESG pillar subscores of Company F in Household Goods in 2017 are 0.00 (E pillar), 63.01 (S pillar), and 54.77 (G pillar) with weights 0.240, 0.294, and 0.466, respectively. Then its ESG score in 2017 is the weighted sum of all pillar subscores:*

$$x_{ESG,CompanyF,2017} = 0.240 \cdot 0.00 + 0.294 \cdot 63.01 + 0.466 \cdot 54.77 = 44.05.$$

A drawback of the current Refinitiv ESG scoring methodology is that its ESG data for the five most recent years is non-definitive. For example, the ESG scores of 2016 have been marked as definitive since June 2021. The difference between the ESG scoring methodology

Figure 5.1: Aggregation of ten categories and three pillars in Refinitiv's ESG scoring construction methodology, where RU, EI, EM, WF, HR, CO, PR, MG, SH, CS, E, S, and G denote Resource Use, Environmental Innovation, Emission, Workforce, Human Rights, Community, Product Responsibility, Management, Shareholders, Corporate Social Responsibility, Environmental, Social, and Governance.



of definitive and non-definitive ESG scores is that the former's pillar subscore or category score weights are not subject to modifications. For instance, if Refinitiv updates the pillar subscore weights to compute ESG scores in January 2022, as we study in Section 5.5, the ESG scores of 2016 stay the same (Refinitiv 2021a). However, the ESG scores from 2017 to January 2022 can get modified. Moreover, the companies are no longer allowed to update or provide ESG information regarding 2016 or previous years of 2016, but such changes in ESG information are feasible for the ESG data from 2017 to January 2022.

Indeed, Company F in Example 5.1 has not yet provided any ESG information for the E pillar subscore of 2017; therefore, its E pillar subscore is zero. It also implies that the ESG category scores of 2017 building the E pillar subscore (Resource Use, Emissions, and Environmental Innovation) are zero. Company F can disclose its missing ESG information regarding the E pillar's categories until June 2022, given that ESG data for the last five years can be updated a posteriori (Refinitiv 2021a).

5.3 Data description and preprocessing

Using the non-definitive ESG data, for which companies can still disclose new ESG information, making their ESG data changed, our data consists of yearly ESG scores and E, S, and G pillar subscores composed of the ten ESG categories Resource Use, Emissions, Environmental Innovation, Workforce, Human Rights, Community, Product Responsibility, Management, Shareholders, and Corporate Social Responsibility scores of the constituents of the S&P 500 (extracted on February 4, 2021) and the constituents of the EuroStoxx 600 (extracted on March 28, 2022) over the period 2017 to 2019, i.e., the top market capitalization companies

in the USA and Europe, respectively. While we use only the former in Section 5.5, Section 5.6 analyzes both samples.

We compute the companies' daily log returns using their daily price data from January 2, 2017, to December 30, 2019. Since 17 companies in the S&P 500 and 109 companies in the EuroStoxx 600 do not report either ESG data or price data in 2017-2018 in the database, we excluded them from our analysis, working with 483 companies in the S&P 500 and 491 companies in the EuroStoxx 600. To have as many companies as possible in the sample, we argue that investors use the latest score available in the market to make their risk assessment. Hence, the ESG data of the companies, which do not have any values in 2019, is imputed by their ESG data in 2018, assuming their score has not yet been updated, and investors would still consider these scores in their decision-making. In total, we imputed the ESG data of 66 companies in the S&P 500 and 19 companies in the EuroStoxx 600 in 2019. The dependence estimated using Pearson correlation (Kendall's τ) between the ESG scores in 2017 and 2018 is equal to 0.94 (0.79), between the scores in 2018 and 2019 to 0.93 (0.78), while between the ESG scores in 2017 and 2019 to 0.89 (0.71) in the S&P 500. Similar results apply to the EuroStoxx 600.

We estimate the companies' annual 95% VaR as the empirical quantile and annual volatility as the market risk measure using the daily logarithmic returns for the risk measures. Then we calculate the companies' *vvrisk*, which is the multiplication of the VaR and annual volatility.

Both samples include companies from ten different Thomson Reuters Business Classifications Economic Sectors (Refinitiv 2021b): Basic Materials (23 and 50 companies), Consumer Cyclicals (77 and 79 companies), Consumer Non-Cyclicals (39 and 41 companies), Energy (24 and 17 companies), Financials (60 and 88 companies), Healthcare (56 and 33 companies), Industrials (65 and 84 companies), Real Estate (28 and 26 companies), Technology (82 and 45 companies), and Utilities (29 and 28 companies) in the S&P 500 and EuroStoxx 600, respectively. Even though the data provider determines the ESG, pillar, and category scores for 47 industry groups within the S&P 500 and 52 industry groups within the EuroStoxx 600, we work with ten economic sectors to have a larger sample size within each sector.

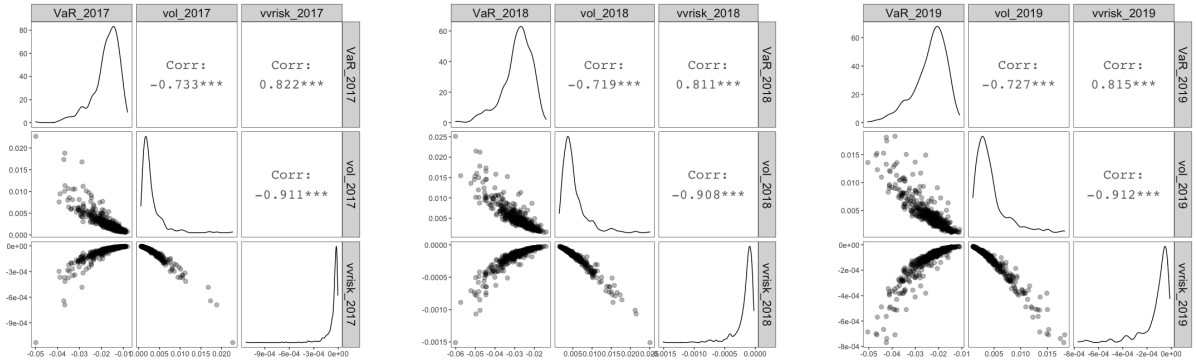
5.4 Exploratory data analysis

Risk measures

While exploring our risk data, Figure 5.2 reports the pairwise scatter plots with the empirical Kendall's τ of the VaR, volatility, and *vvrisk* for the S&P 500. We remark that less negative VaR tends to be associated with smaller volatility levels, and the *vvrisk* is highly negatively/positively dependent on the volatility/VaR.

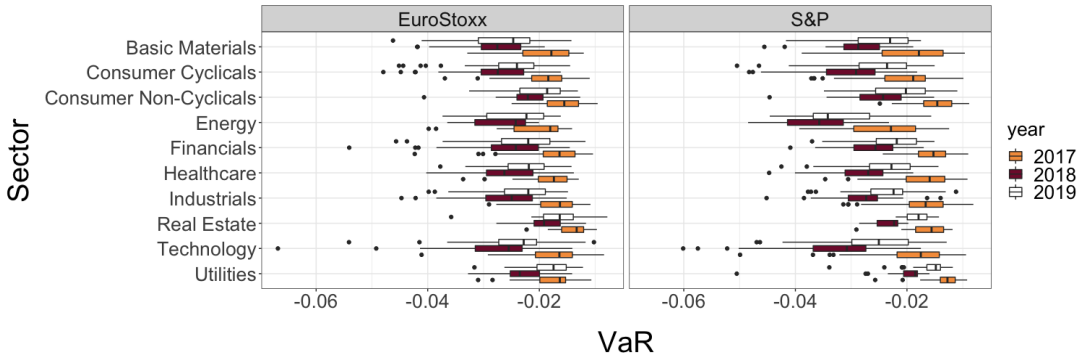
Moreover, we can observe the variability in the VaR of the companies in each of the ten sectors in the S&P 500 and EuroStoxx 600 across three years in Figure 5.3. While Financials and Real Estate have the smallest variation in 2017 and in 2018-2019 in the S&P 500,

Figure 5.2: Lower triangular panels: pairwise scatter plots of volatility, VaR, and *vvrisk* in a year in S&P 500, upper diagonal panels: the empirical Kendall's τ, diagonal: empirical marginal densities.



respectively, the variation of Real Estate in 2017-2018 and Utilities in 2019 is smaller than others in the EuroStoxx 600. Furthermore, the variation of Energy is larger than others across three years in the S&P 500. Additionally, Basic Materials, Consumer Cyclicals, and Consumer Non-Cyclicals in the S&P 500 have larger variations than those in the EuroStoxx 600 across three years, whereas Financials is the sole sector whose variation across three years is larger in the EuroStoxx 600 than the S&P 500.

Figure 5.3: 95% VaR across ten sectors and three years in the EuroStoxx 600 and S&P 500.



Missing ESG data

The number of ESG categories with not yet disclosed (missing) ESG information varies from zero to six, with more details given in the supplementary in Section 5.10. Our empirical findings provide that the non-financial disclosures are higher in the EuroStoxx 600 companies than in the S&P 500 companies. Since there are mandatory ESG disclosure regulations in

European Union but not in the USA, such observations might be expected (International Platform on Sustainable Finance 2021).

Table 5.1 shows that the percentage of the companies with not yet disclosed ESG information regarding at least one of the ten ESG categories ranges from a minimum of 15% in Consumer Non-Cyclicals in 2019 to a maximum of 71% in Healthcare in 2017 with an average of 47% across ten sectors and three years in the S&P 500. Its range is from 11% in Utilities in 2019 to 77% in Healthcare in 2017 with an average of 38% across ten sectors and three years in the EuroStoxx 600. Additionally, it is always higher in the S&P 500 than the EuroStoxx 600 per year in all sectors but Consumer Non-Cyclicals and Real Estate. Thus, missing ESG information changes with sectors and geographical regions. More than half of the companies in Consumer Cyclicals, Energy, and Healthcare in the S&P 500 and Real Estate in the EuroStoxx 600 have at least one of the ten ESG categories with not yet released ESG information each year. Moreover, we observe that the percentage of such companies tends to decrease in time. Such a result might imply that the companies disclose more information as it becomes available, which could provide new insights into their ESG performance and risk characteristics. Furthermore, since the ESG scores have had a strong impact on the company's value (Fatemi et al. 2018), one can expect the companies to publish more ESG information in the future.

Table 5.1: Percentage of the companies with missing ESG information at least in one of the ten ESG categories across ten sectors and three years in S&P500 and EuroStoxx600 combined.

Sector (S&P - EuroStoxx)	2017	2018	2019
Basic Materials	35% - 36%	30% - 24%	30% - 20%
Consumer Cycl.	56% - 43%	55% - 35%	53% - 30%
Consumer N-Cycl.	23% - 37%	18% - 34%	15% - 22%
Energy	67% - 35%	67% - 35%	58% - 29%
Financials	62% - 50%	53% - 36%	47% - 26%
Healthcare	71% - 70%	64% - 58%	59% - 45%
Industrials	48% - 46%	48% - 37%	45% - 30%
Real Estate	68% - 77%	54% - 65%	46% - 54%
Technology	48% - 42%	44% - 36%	34% - 33%
Utilities	45% - 32%	41% - 18%	34% - 11%

Figure 5.4 shows that companies with lower ESG scores tend to have more ESG categories with not yet published ESG information. It could be due to the lack of infrastructure allowing them to collect and then release such information. However, it would still imply that a company with a lower ESG score could have a large potential to upgrade its ESG score when not yet recorded information is disclosed.

ESG scores and pillar subscores

Figure 5.5 gives the boxplots of ESG scores and pillar subscores across ten sectors and three years in the EuroStoxx 600 and S&P 500, respectively. We see variability in all scores for all sectors.

The median ESG score of Consumer Non-Cyclicals is higher than that of others in the S&P 500. Since Consumer Non-Cyclicals has less missing information in the ESG categories than the others in Table 5.1, such a case can be expected. Similar results hold for Energy in the EuroStoxx 600.

In addition, the range of E-pillar subscores is higher than that of their other scores. It could be due to the fact that environmental issues may be hard to compute for some sectors like Financials. Some companies in Consumer Cyclicals, Financials, Healthcare, Industrials, Real Estate, and Technology have not yet published any *Environmental* information in 2018 in the S&P 500 as indicated by their zero *Environmental* pillar subscores.

Figure 5.4: Scatter plot of the companies' ESG scores and their number of ESG categories with undisclosed ESG information in Consumer Cyclicals in S&P 500 in 2017, where a triangle denotes the median ESG score of the respective row.

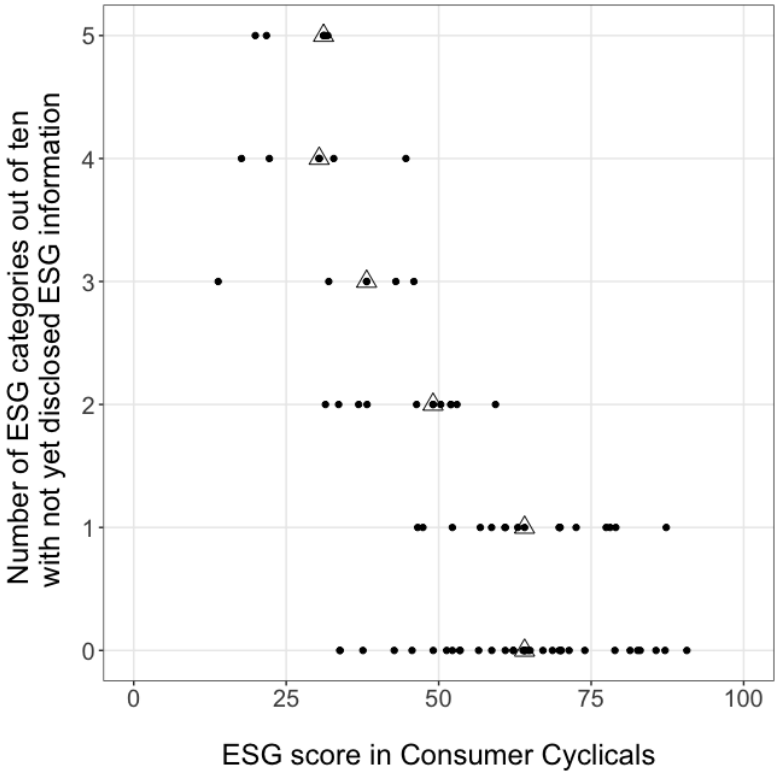
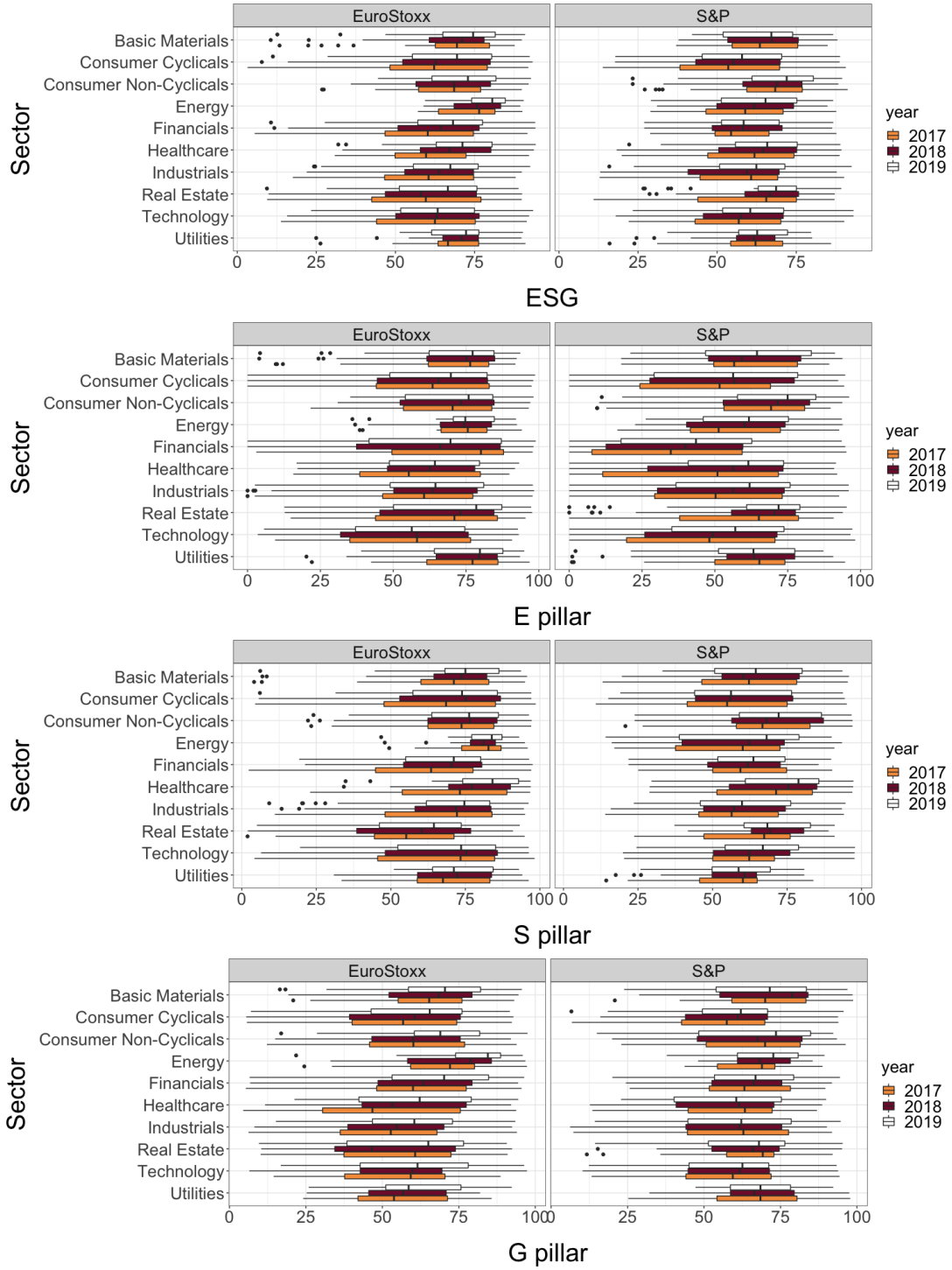


Figure 5.5: ESG and pillar subscores across ten sectors and three years in the EuroStoxx 600 and S&P 500.



5.5 The pitfalls of ESG scoring methodology

In this section, first, we discuss the need for non-definitive ESG scores from the provider's point of view. Then we study the dependence measured by Kendall's τ between our sample of non-definitive ESG scores and risk. Note that the dependence between the ESG scores and annual 95% VaR in 2017 for Basic Materials increased from 0.130 to 0.170, where we use the ESG scores from February 2021 in the former and September 2021 in the latter. Likewise, the Industrials' ESG scores by September 2021 have the dependence value of 0.124, increasing from 0.118. Therefore, such changes might be the first indication that the provider can be manipulating the ESG data thanks to its ESG scoring methodology.

We explain how the ESG scores of the five most recent years can be updated and the implications of the resulting non-definitive ESG scores using simple optimization and data mining approaches. In particular, we analyze the effect of the pillar subscore weights/category score weights to compute ESG scores/pillar subscores. Later, we discuss the impact of the initial disclosure of ESG information and an update of the published ESG information.

Non-definitive ESG scores from a provider perspective

The provider might aim to give the most updated ESG information to its customers. Thus, it is reasonable to *re-write* its recent ESG data. For example, suddenly, there could be big scandals about how a company approaches labor rights, and the company could not be regarded as responsible as previously assessed regarding its employee treatment. Therefore, the provider would need a modification in its ESG data, especially for its *Governance* pillar subscore, since the *Governance* pillar subscore takes into account employee treatment. Additionally, the world has been changing rapidly, and what is now unimportant might be the most important thing in the next five years. Currently, the *Environmental* pillar has had small importance in the aggregation to compute most Technology companies' ESG scores (Refinitiv 2021a). The provider's justification might be that *Environmental* information is not well related to the Technology sector's sustainability since big Technology companies do not disclose much *Environmental* information. However, it has been a hot debate how much *Environmental* damage a Technology company makes to develop its software and hardware (Bender et al. 2021; Strubell et al. 2019) and if the Technology companies correctly report their emissions (Klaaßen and Stoll 2021). Therefore, the provider might need to add much more disclosed *Environmental* information in the Technology's ESG data, even back in time, and modify the importance of the Technology's *Environmental* pillar in the future.

Non-definitive ESG scores and risk dependence

We consider Kendall's τ as a dependence measure between ESG scores and VaR in the same year for all sectors since our sample has nonlinear dependence as shown in Figure 5.6. The

number companies is 23 in Basic Materials, 77 in Consumer Cyclicals, 39 in Consumer Non-Cyclicals, 24 in Energy, 60 in Financials, 56 in Healthcare, 65 in Industrials, 28 in Real Estate, 82 in Technology, and 29 in Utilities.

Figure 5.6: The scatter plots of companies' ESG scores and their annual 95% VaR values in for Consumer Cyclicals and Industrials, where gray curve demonstrates a local polynomial regression fit, denoting nonlinear dependence between two variables.

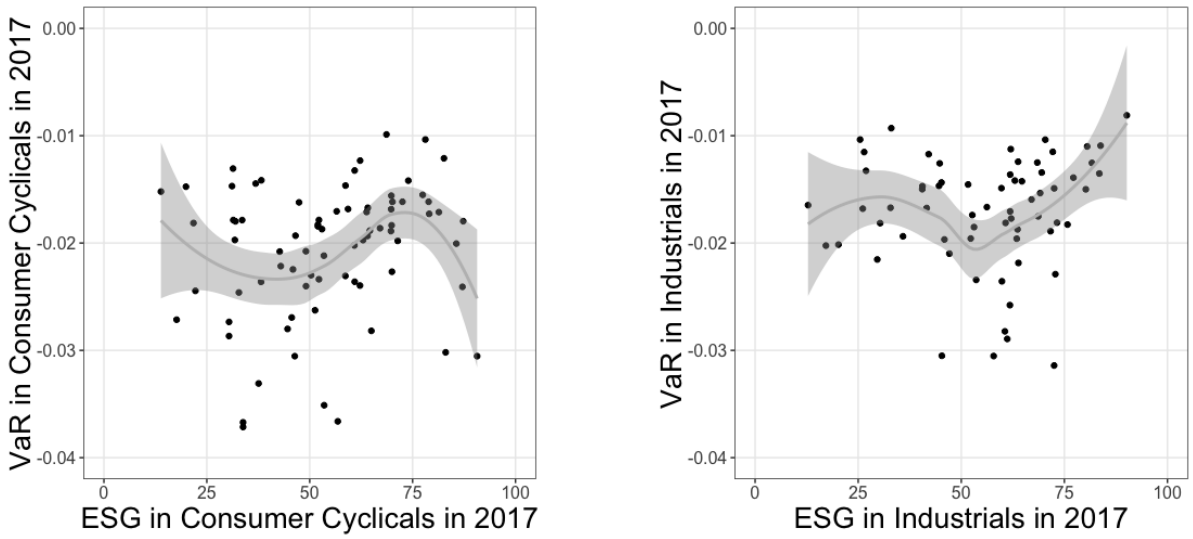


Table 5.2 shows that five sectors exhibit significant dependence for 2017, and it varies across sectors. The hypothesis testing is for $H_0{:}\tau = 0$ versus $H_A{:}\tau > 0$ (Hollander et al. 2013). *, **, and *** denote statistical significance at 10%, 5%, and 1% levels, respectively. Energy and Healthcare have a significant dependence at 10% and 5% levels, respectively, in all years. Significance exists for some years for Consumer Cyclicals, Industrials, and Technology.

Re-weighting scheme for pillar subscores

A reason for non-definitive ESG scores is that the weights of pillar subscores to calculate them can be updated. We will calculate the exemplary weights in this part. Refinitiv's current pillar subscore weights are determined for industry groups,[1] and Refinitiv can modify them annually.[2] The modifications are based on the information provided by large- and middle-cap companies, making them open to manipulations (Page 12 of Refinitiv (2021a)).

[1] To our knowledge, pillar subscore weights cannot be reported explicitly.

[2] As a reply to our inquiry on May 10, 2021, Refinitiv states, "Since the ESG scoring methodology which was implemented on April 6, 2020, we have not made any changes to the pillar subscore weights. They may change in the future to best adapt to potential changes that may occur within the industries and their ESG performance and impact."

Table 5.2: Kendall's τ between ESG/re-weighted ESG scores defined in Equation (5.1) (left/right columns under a year) and annual 95% VaR in 2017-2019 across ten sectors.

Sector	2017		2018		2019	
B. Materials	0.130	0.399 **	0.099	0.320 **	-0.075	0.130
Cons. Cycl.	0.161 **	0.175 **	0.144 **	0.182 **	0.036	0.101 *
Cons. N-Cycl.	-0.047	0.015	0.001	0.074	0.115	0.136
Energy	0.196 *	0.254 **	0.203 *	0.377 **	0.229 *	0.341 **
Financials	0.020	-0.003	-0.160	-0.096	-0.102	-0.005
Healthcare	0.309 ***	0.296 ***	0.278 ***	0.271 ***	0.161 **	0.161 **
Industrials	0.118 *	0.140 **	-0.071	0.051	-0.103	0.084
Real Estate	-0.138	-0.185	-0.312	-0.212	-0.310	-0.069
Technology	0.154 **	0.172 **	0.028	0.069	0.051	0.189 **
Utilities	-0.079	0.044	-0.030	0.059	-0.099	0.079

As expected, a modification in pillar subscore weights might result in very different ESG scores for the same company. For example, suppose that a company's *Environmental*, *Social*, and *Governance* pillar subscores are 30, 70, and 40, respectively. If the weights are 0.200, 0.500, and 0.300 for the *Environmental*, *Social*, and *Governance* pillar subscores, the company has an ESG score of 53. However, the company has an ESG score of 41 using the weights 0.500, 0.200, and 0.300 for the E, S, and G pillar subscores, respectively.

A data-mining approach can be easily applied to *re-write* ESG scores, which might find optimal weights of the pillar subscores to calculate the new, so-called re-weighted ESG scores. The optimality would also have a relationship between a risk measure and the re-weighted ESG scores. Since the pillar subscore weights are reported by three digits[1], there are 501501 possible weighting schemes for pillar subscores as discussed in the supplementary in Section 5.10. Therefore, we examine if and which re-weighting scheme out of 501501 can strengthen the relationship between the re-weighted ESG scores and VaR, solving optimization problems like proposed in Equation (5.2) (Powell 2015).

First, we define re-weighted ESG scores for each company i_p in sector p and year t via the new weights of the pillar subscores (w_p^E, w_p^S, w_p^G):

$$ReweightedESG_{i_p,p}^t = E_{i_p,p}^t \cdot w_p^E + S_{i_p,p}^t \cdot w_p^S + G_{i_p,p}^t \cdot w_p^G, \tag{5.1}$$

where $p \in \{Basic\ Materials, \ldots, Utilities\}$, $t \in \{2017, 2018, 2019\}$, $i_p = 1, \ldots, n_p$, and n_p is the number of companies in sector p.

Then Equation (5.2a) determines new pillar subscore weights so that re-weighted ESG scores show the highest positive dependence with the VaR within each sector in all years, where it holds $\boldsymbol{ReweightedESG}_p^t = (ReweightedESG_{1,p}^t, \ldots, ReweightedESG_{n_p,p}^t)^\top$ and $\boldsymbol{VaR}_p^t = (VaR_{1,p}^t, \ldots, VaR_{n_p,p}^t)^\top$. As shown by Diemont et al. (2016), it is natural to assume that companies with higher (re-weighted) ESG scores should have stronger VaR

relationships. Following the ESG scoring methodology of Refinitiv (Refinitiv 2021a), Equation (5.2b) ensures that re-weighted ESG scores are between zero and 100; Equation (5.2c) assumes each pillar subscore has a non-negative impact on the re-weighted ESG scores. Our simple scheme aggregates the dependence over the three recent years. Alternative approaches could consider the re-weighting in a single year or impose a positive lower bound on weights (e.g., $w_p^E, w_p^S, w_p^G \geq 0.2$) to avoid putting too much concentration on a single pillar subscore as shown in Table 5.4. To have a larger sample size within the S&P 500, we investigate the re-weighting on a broader level, i.e., ten economic sectors rather than 47 industry groups.[3]

$$\max_{w_p^E, w_p^S, w_p^G} \quad \sum_{t=2017}^{2019} \hat{\tau}\Big(\boldsymbol{ReweightedESG}_p^t,\ \boldsymbol{VaR}_p^t\Big) \tag{5.2a}$$

$$\text{subject to} \quad w_p^E + w_p^S + w_p^G = 1, \quad \forall p, \tag{5.2b}$$

$$w_p^E, w_p^S, w_p^G \geq 0, \quad \forall p. \tag{5.2c}$$

To minimize the sensitivity of the numerical optimization concerning initial parameter values, we run the models using 20 different starting values. We choose the optimal weights corresponding to the highest dependence value of 20 runs within each sector. We do not observe multiple optimal solutions.

Table 5.2 presents the dependence measured by Kendall's τ between re-weighted ESG scores and VaR. We observe increases in the dependence using re-weighted ESG scores, but five minimal reductions result from the fact that Refinitiv determines the weights for industry groups within a sector, having more parameters and flexibility to strengthen the dependence. Table 5.2 shows that the re-weighted ESG scores of Consumer Cyclicals, Energy, and Healthcare show significance at least at 10%, 5%, and 5% levels in all years, respectively. Significance exists in some years for Basic Materials, Industrials, and Technology using re-weighted ESG scores. However, Healthcare's significance level is also at 5% using ESG scores for all years.

Table 5.3 reports the new pillar subscore weights as a solution of Equation (5.2). The *Environmental* pillar in Utilities and Basic Materials is the sole driver of strengthening the risk performance using re-weighted ESG scores. Determining such a high *Environmental* pillar subscore weight might be unreasonable for the provider, but having its weight as high as possible while ensuring that all pillar subscore weights are larger than 0.200, like in Table 5.4 still provides strong risk dependence for Basic Materials in 2017 as seen in the left three columns of the same table. Likewise, Energy's *Environmental* pillar plays a vital role in having significant risk dependence using re-weighted ESG scores, but the provider currently gives the highest importance to its *Social* pillar subscore[1].

The *Governance* pillar is the main factor in reducing risks in Consumer Cyclicals, Financials, Industrials, and Real Estate using re-weighted ESG scores. However, none of the industry groups in Consumer Cyclicals currently has the highest weight assigned to their *Governance*

[3]The variance in current pillar weights within sectors is low, except for Consumer Cyclicals, Healthcare, and Technology.

pillar[1]. Moreover, using re-weighted ESG scores, stronger *Social* activities are associated with lower tail risk in Technology.

Table 5.3: New pillar subscore weights resulting in re-weighted ESG scores in Table 5.2 across ten sectors as a result of solving the optimization problem in Equation (5.2).

Sector	E	S	G
Basic Materials	1.000	0.000	0.000
Consumer Cycl.	0.085	0.174	0.741
Consumer N-Cycl.	0.639	0.346	0.015
Energy	0.904	0.000	0.096
Financials	0.007	0.099	0.894
Healthcare	0.133	0.481	0.386
Industrials	0.000	0.000	1.000
Real Estate	0.000	0.024	0.976
Technology	0.065	0.935	0.000
Utilities	1.000	0.000	0.000

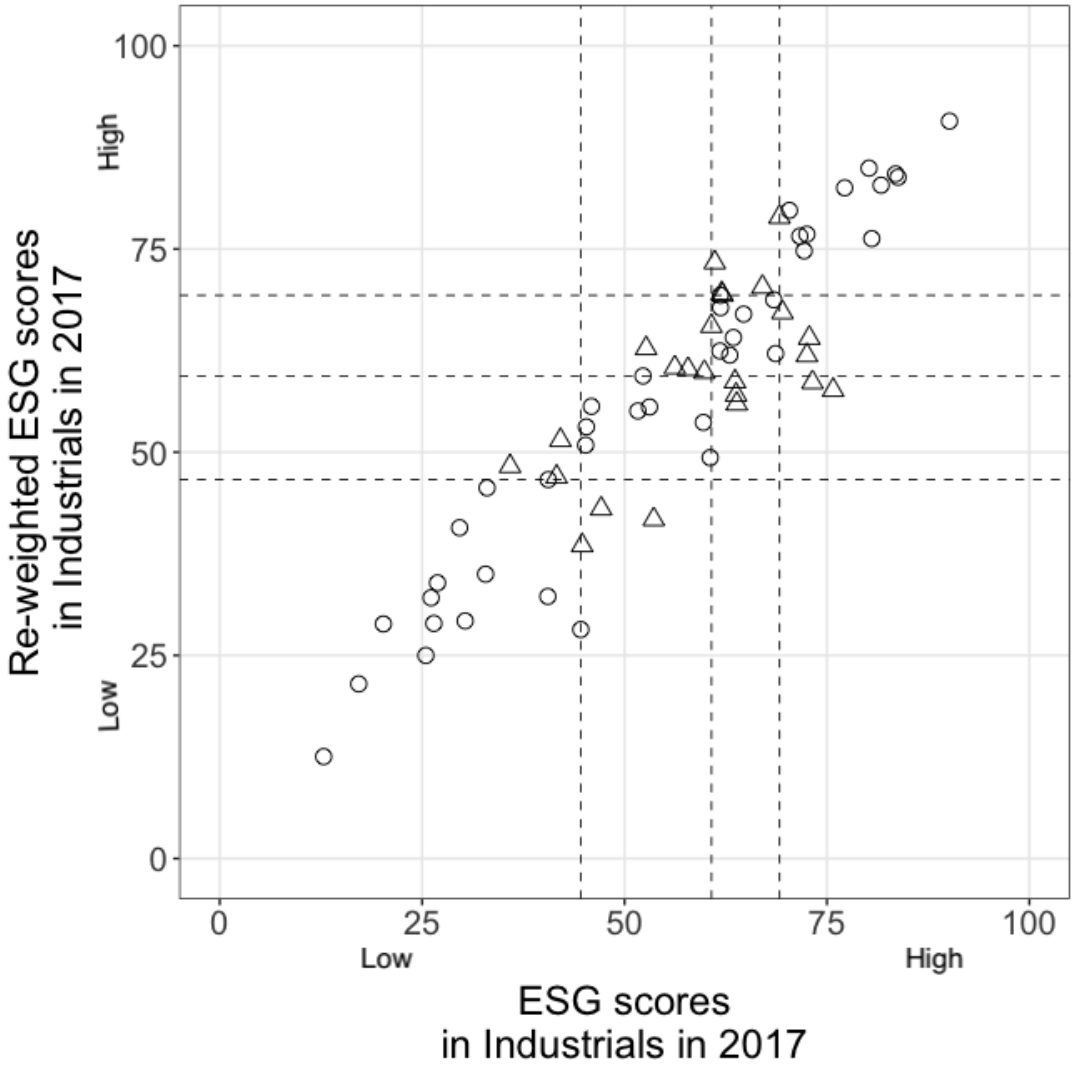
Table 5.4: Kendall's τ between re-weighted ESG scores and 95% VaR, and new pillar subscore weights in 2017-2019 across ten sectors after imposing a positive lower bound (0.2) on weights.

Sector	2017		2018		2019		E	S	G
B. Materials	0.209	*	0.154		-0.028		0.598	0.202	0.200
Consumer Cycl.	0.192	***	0.184	***	0.055		0.200	0.247	0.553
Consumer N-Cycl.	-0.009		0.036		0.142		0.559	0.230	0.211
Energy	0.239	*	0.341	**	0.297	**	0.567	0.200	0.233
Financials	0.010		-0.155		-0.075		0.200	0.201	0.599
Healthcare	0.304	***	0.262	***	0.145	*	0.203	0.423	0.374
Industrials	0.130	*	-0.032		0.000		0.200	0.201	0.599
Real Estate	-0.164		-0.296		-0.180		0.200	0.205	0.595
Technology	0.149	**	0.041		0.119	*	0.202	0.596	0.201
Utilities	-0.044		0.034		-0.054		0.600	0.200	0.200

Figure 5.7 illustrates the companies' ESG and re-weighted ESG scores for 2017 in Industrials, where we use the re-weighting scheme given in Table 5.4. Even though the dependence (Kendall's τ) between ESG and re-weighted ESG scores is around 0.72, more than 35% of the companies out of 65 are in different quartiles regarding them. Thus, also some top and bottom quartile companies concerning their ESG scores, regarded as the good and bad ESG responsible companies, respectively, change with a modification in the pillar subscore weights. Nonetheless, we observe that the companies with the highest and lowest ESG score have the same rank after the re-weighting. Therefore, the re-weighting scheme considered is not

enough to change their ESG rank since they have the highest and lowest pillar subscores for two pillars among their peers.

Figure 5.7: Scatter plot of the Industrials' ESG and re-weighted ESG scores for 2017 obtained using the re-weighting scheme in Table 5.4, where a triangle denotes the companies whose score quartiles change, and a grid line represents a quartile of the respective axis.



As a result, the ESG scores of the sectors, except Healthcare, can be *re-written* by re-weighting the pillar subscores so that better ESG scores are associated with fewer risks, in line with the idea of Berg et al. (2021). Therefore, using non-definitive ESG scores in the research might lead to different conclusions based on their data extraction date, like a significant/insignificant (re-weighted/non-definitive ESG scores) risk relationship in Basic

Materials for 2017. However, such a simple re-weighting scheme still does not provide a significant VaR dependence for all sectors. Therefore, further research could use more complex models to strengthen risk dependence.

Re-weighting scheme for category scores

A re-weighting scheme can also be applied for category scores to compute (new, re-weighted) pillar subscores. Thus, pillar subscores might be subject to modifications (Appendix E in Refinitiv (2021a)). Then since pillar subscores compute ESG scores as discussed in Section 5.5, ESG scores can also change.

We can argue that a more flexible re-weighting scheme with more parameters, such as one for category scores, leads to new ESG scores whose risk dependence is at least as given in Table 5.2. Therefore, one can expect that a re-weighting scheme for category scores alters the significance of the (new) ESG scores and risk dependence, as well as the quartiles of the companies regarding their (new) ESG scores.

The initial disclosure of ESG information

Besides changing pillar subscore and category score weights, another pitfall that can alter ESG scores and research findings is the initial disclosure of ESG information. If companies have not yet released the ESG indicators' information (value) under a category, the respective category score is encoded as zero. However, the companies with zero scores are not considered in the ranking scheme to find the category score of their peers (Page 9 of Refinitiv (2021a)). As soon as the companies with zero scores publish the ESG indicators' values, they are included in the ranking computation, changing the peers' category scores even though the peers did not change any ESG indicator values. Such disclosure methodology results in the *re-written* ESG scores in that industry group without any announcements by the provider.

In the following hypothetical Example 5.2, we illustrate such an effect of the initial disclosure of ESG information on ESG scores.

Example 5.2 *(An effect of the initial disclosure of ESG information on ESG scores).*
Suppose that there are five companies with a generic name in the industry group Chemicals, and their fictitious category score Emission is given in Table 5.5. Assume that the companies might disclose a carbon dioxide emission (COe) value as the ESG indicator of Emission. We see in Table 5.5 that the company X publishes its initial COe value since a number replaces N/A, and the others do not update it. However, all companies' COe and Emission scores later increase by the company X's disclosure of the COe value. For instance, the COe score of the company B changes from 0.125 to 0.300. Thus, its Emission score of 12.50 is replaced by 30.00. Given that higher scores represent more responsible companies, the company X's disclosure makes the impression that all other companies are more responsible than before.

Table 5.5: Changes in the category score Emission before and after the disclosure of the company X, assuming lower COe value is better.

Company	COe value (before)	COe score (before)	Emission (before)	COe value (after)	COe score (after)	Emission (after)
X	N/A	0.00	0.00	50.4	$\frac{1/2}{5} = 0.100$	10.00
Y	20.5	$\frac{1/2}{4} = 0.125$	12.50	20.5	$\frac{1+1/2}{5} = 0.300$	30.00
Z	14.5	$\frac{1+1/2}{4} = 0.375$	37.50	14.5	$\frac{2+1/2}{5} = 0.500$	50.00
Q	10.5	$\frac{2+1/2}{4} = 0.625$	62.50	10.5	$\frac{3+1/2}{5} = 0.700$	70.00
U	5.5	$\frac{3+1/2}{4} = 0.875$	87.50	5.5	$\frac{4+1/2}{5} = 0.900$	90.00

Moreover, since the Environmental pillar subscore is a weighted sum of the category scores, such disclosure causes ongoing and unannounced modifications in the companies' Environmental pillar subscore as shown in Table 5.6. Remarkably, companies Q and U have the highest Environmental pillar subscore before the company X's disclosure. Nonetheless, after the company X's disclosure, the company Q is regarded as the best company in terms of the Environmental pillar subscore due to updates in the EM scores. Despite the change in the scores, the Environmental information of the companies Q and U is the same before and after the disclosure of the company X. A similar result applies to comparing the Environmental pillar subscore of the companies Y and Z.

Table 5.6: Modifications in the E pillar before and after the disclosure of the company X in Table 5.5, assuming that the weights to calculate the Environmental (E) pillar subscore are 0.400, 0.400, and 0.200 for the Emission (EM), Resource Use (RU), and Environmental Innovation (EI) category scores, respectively.

	EM score (before)	RU score (before)	EI score (before)	E pillar (before)	EM score (after)	RU score (after)	EI score (after)	E pillar (after)
X	0.00	0.00	0.00	0.00	10.00	0.00	0.00	4.00
Y	12.50	37.50	25.00	25.00	30.00	37.50	25.00	32.00
Z	37.50	12.50	25.00	25.00	50.00	12.50	25.00	30.00
Q	62.50	87.50	75.00	75.00	70.00	87.50	75.00	78.00
U	87.50	62.50	75.00	75.00	90.00	62.50	75.00	76.00

Finally, we observe that all companies' ESG scores are re-written as given in Table 5.7 due to the modifications in the companies' Environmental pillar subscore before and after the disclosure of the company X. Thus, a company's initial disclosure of ESG information changes its peers' respective category scores, respective pillar subscores, and ESG scores.

We remark that the change in the scores due to the initial disclosure of ESG information might decrease as the number of companies in an industry group or the number of indicators to

Table 5.7: Modifications in the ESG scores before and after the disclosure of the company X in Table 5.5, assuming that the weights to calculate the ESG score are 0.200, 0.300, and 0.500 for the E, S, and G pillar subscores, respectively.

Company	E pillar (before)	S pillar (before)	G pillar (before)	ESG (before)	E pillar (after)	S pillar (after)	G pillar (after)	ESG (after)
X	0.00	30.00	10.00	14.00	4.00	30.00	10.00	14.80
Y	25.00	10.00	30.00	23.00	32.00	10.00	30.00	24.40
Z	25.00	90.00	70.00	67.00	30.00	90.00	70.00	68.00
Q	75.00	50.00	90.00	75.00	78.00	50.00	90.00	75.60
U	75.00	70.00	50.00	61.00	76.00	70.00	50.00	61.20

compute a category score increases. However, companies can still disclose around 50 numerical ESG indicators for the last five years, making unannounced score changes accumulated over time. We remark that the average number of numerical ESG indicators is based on the key performance indicators document provided by Refinitiv on request.

In our data sample, five companies out of 65 in Industrials have not yet reported any *Environmental* information for 2018 and can report them until June 2023. For instance, they have not published their carbon dioxide emission and water usage indicators for 2018 to compute their *Environmental* pillar subscore. However, when such *Environmental* information is released, they will get a non-zero *Environmental* pillar subscore for 2018 based on their relative performance among their peers.

With this motivation, consider the following additional hypothetical example in 2023: these five companies publish their *Environmental* information for 2018, thereby getting a non-zero *Environmental* pillar subscore for 2018. Then their ESG score for 2018 increases since an ESG score is a weighted sum of pillar subscores. Accordingly, they get better ESG score ranks for 2018 among their peers in the data extraction in 2023 than in 2021. Assume their ESG score rank for 2018 improves around 25-35 points in 2023 compared to 2021, e.g., the company on the 61st rank in 2021 is on the 35th rank in 2023. As a result of this assumption, the remaining companies' ranks change accordingly, e.g., the 56th rank in 2021 obtains the 61st rank in 2023. However, we assume that the ranks within the remaining companies do not differ. For instance, while the 56th rank in 2021 obtains the 61st rank in 2023, the 55th rank in 2021 obtains the 60th rank in 2023.[4] Under these assumptions, the dependence (Kendall's τ) between ESG scores and annual 95% VaR for Industrials for 2018 changes from -0.071 in 2021 to 0.002 in 2023. If the companies disclose more and better information to improve their rank around 55-60, the associated dependence between ESG scores and VaR for Industrials for 2018 equals 0.078 in 2023, even stronger, as given by the simple optimization problem in Equation (5.3),

[4]Our motivation comes from a company in Industrials, whose rank for 2018 in February 2021 decreased value by 26 points among its peers in September 2021 thanks to the initial ESG information disclosure.

where $\boldsymbol{ESG}_{Ind}^{2018} = \{ESG_{Ind}^{i,2018} : i = 1, \ldots, 65\}$ and $\boldsymbol{VaR}_{Ind}^{2018} = \{VaR_{Ind}^{i,2018} : i = 1, \ldots, 65\}$ denote the 65 companies' ESG scores for 2018 in Industrials (Ind) and their associated annual 95% VaR, respectively. The parameters corresponding to the new ESG scores of the five companies after the ESG information disclosure are denoted by $ESG_{Ind}^{1,2018}, \ldots, ESG_{Ind}^{5,2018}$.

$$\max_{ESG_{Ind}^{1,2018},\ldots,ESG_{Ind}^{5,2018}} \quad \hat{\tau}\Big(\boldsymbol{ESG}_{Ind}^{2018}, \quad \boldsymbol{VaR}_{Ind}^{2018}\Big) \tag{5.3a}$$

$$\text{subject to} \quad ESG_{Ind}^{1,2018}, \ldots, ESG_{Ind}^{5,2018} \leq 100, \tag{5.3b}$$

$$ESG_{Ind}^{1,2018}, \ldots, ESG_{Ind}^{5,2018} \geq 0. \tag{5.3c}$$

The significance of providing transparent corporate social responsibility information on companies' financial benefits has been recently shown by Du and Yu (2021). Given the growing importance of ESG scores on companies' value (e.g., Fatemi et al. (2018) and Wong et al. (2021)), it is reasonable to expect companies to publish more ESG information over time. Hence, such information potentially alters non-definitive ESG scores, and the unannounced ESG score modification is a fact that could affect research findings.

Update of the published ESG information

Another factor that can lead to modifications in the ESG scores is an update of the published ESG information, including its deletion. For instance, in Example 5.2, the company Y's carbon dioxide emission value can be updated if such previous information appears incorrect. If the new value affects the company Y's rank among its peers to find its carbon dioxide emission score, its and its peers' ESG scores are *re-written* as discussed in Example 5.2.

Consider another hypothetical example from our data sample: if almost 6% of the companies out of 65 in Industrials update their ESG scores and ranks for 2018, assuming that the others' ranks within themselves are preserved, the dependence (Kendall's τ) between ESG scores and VaR for Industrials for 2018 increases from -0.071 to a significant value of 0.119 using the hypothesis testing of $H_0 : \tau = 0$ versus $H_A : \tau > 0$ at 10% level.

For more discussion, we refer to Berg et al. (2021).

5.6 The quantification of missing data and its impact

In this section, first, we quantify the companies' potential to disclose more ESG information, including missing ones, in time and analyze the impact of the disclosure on the ESG scores and the risk of the portfolios. Accordingly, we formulate a *Missing (M) pillar* subscore, which explicitly captures the amount of not yet reported ESG information regarding the ten ESG categories. Later, we define *Environmental, Social, Governance, and Missing (ESGM)* scores and propose an optimization approach for their computation, linking them to companies' riskiness. The optimization scheme aims to harvest the potential to strengthen the risk

relationship in disclosing missing ESG information (EBA 2020). Our analyses are based on the constituents of the S&P 500 and EuroStoxx 600 explored in Section 5.4. In this part, we base our four pillar subscore weights estimation on training (in-sample) data to avoid overfitting the data. Finally, we compare the relationship between risk, ESGM scores and ESG scores using test (out-of-sample) data.

Missing pillar (M pillar) subscore

Since a zero ESG category score denotes that the company has not yet reported any information regarding it, like in Example 5.1, we define a new pillar accounting for zero values, i.e., missing information, in the ten ESG category scores in a given business class: the *Missing (M) pillar* subscore. A business class can be an industry group or an economic sector based on the business classification of Thomson Reuters (Refinitiv 2021a). For instance, the assets from the S&P 500 in Section 5.3 belongs to 61 industry groups from ten economic sectors.

Definition 5.1: The M pillar subscore

For company p in a given business class a and year t, first, we find the total number of zero values in its ten ESG categories, i.e., $x^a_{zero,p,t}$. The set $S^a_{zero,t}$ contains these values for all a, p, t. Denoting the total number of companies with same value as p in $S^a_{zero,t}$ (including the company p itself) by $e^a_{p,t}$, and the total number of companies with a higher value than p in $S^a_{zero,t}$ by $l^a_{p,t}$, the M pillar subscore of company p in a given business class a and year t is defined as

$$x^a_{M,p,t} = 100 \cdot \frac{l^a_{p,t} + \frac{e^a_{p,t}}{2}}{n_a}, \quad p = 1, \ldots, n_a, \quad t = 1, \ldots, T, \quad a = 1, \ldots, A, \tag{5.4}$$

where n_a is the total number of companies in the given business class a, T is the total number of years, and A is the total number of given business classes. The detailed calculations and notations are given in the supplementary in Section 5.10.

From the definition, when a company has more zero values in its ESG categories than all other companies in its business class, its M pillar subscore will be the highest. This is because it has a higher extent of not yet published ESG information reflected in a high M pillar subscore. Moreover, the M pillar subscore is continuous and between zero and 100, with a mean value of 50 (proof in the supplementary in Section 5.10). Such a formulation makes it robust to outliers and comparable with the three ESG pillar subscores. Its formulation is similar to our data provider's ESG category score methodology (Refinitiv 2021a).

Since our data provider has ten ESG category scores, the highest total number of zero values a company can have in its ESG categories in our empirical analysis is ten. Accordingly, Example 5.3 shows the M pillar subscore calculation steps using ten ESG category scores, where the business class is an economic sector.

Example 5.3 *(The M pillar subscore calculation). Let* $\boldsymbol{x}^a_{p,t} = (x^a_{RU,p,t}, \ldots, x^a_{CS,p,t})^\top$ *contain ten ESG category scores of company* p *in business class* a *in year* t*. Suppose there are four companies* ($n_a = 4$) *in the economic sector Consumer Cyclicals* ($a = 2$)*, and their fictitious ESG category scores in 2017* ($t = 2017$) *are given as follows:*
$\boldsymbol{x}^2_{1,2017} = (99.3, 50.1, 12.3, 52.2, 0.00, 67.9, 0.00, 11.2, 20.4, 0.00)^\top$,
$\boldsymbol{x}^2_{2,2017} = (63.5, 70.1, 52.3, 84.3, 10.2, 77.9, 88.9, 55.2, 80.4, 86.3)^\top$,
$\boldsymbol{x}^2_{3,2017} = (36.3, 0.00, 12.3, 23.2, 0.00, 17.9, 0.00, 21.2, 50.5, 58.3)^\top$,
$\boldsymbol{x}^2_{4,2017} = (85.2, 0.00, 12.3, 12.2, 0.00, 54.3, 52.5, 81.2, 75.6, 24.3)^\top$.
For company p *with* $p = 1, \ldots, 4$*, we determine the total number of zero values in its ESG categories and have* $S^2_{zero,2017} = \{3, 0, 3, 2\}$*. Consider the fourth company: it holds* $x^2_{zero,4,2017} = 2$*,* $e^2_{4,2017} = 1$*, and* $l^2_{4,2017} = 1$*. Accordingly, we calculate its M pillar subscore as given in Equation (5.4):* $x^2_{M,4,2017} = 100 \cdot \frac{1+1/2}{4} = 37.5$*. For the second company without any zero values in its ESG categories, the M pillar subscore is given by* $x^2_{M,2,2017} = 100 \cdot \frac{0+1/2}{4} = 12.5$*. Since we also calculate* $x^2_{M,1,2017} = 75$ *and* $x^2_{M,3,2017} = 75$*, the M pillar subscore is between zero and 100; the average M pillar subscore of all companies is 50, as postulated.*

Environmental, Social, Governance, and Missing (ESGM) scores

Next, we incorporate the M pillar into the three ESG pillars and builds new scores, the *ESGM: Environmental, Social, Governance, and Missing* scores.

Definition 5.2: The ESGM score

The ESGM score of company p in a given business class a and year t is defined as a weighted sum:

$$x^a_{ESGM,p,t} = x^a_{E,p,t} \cdot w^a_E + x^a_{S,p,t} \cdot w^a_S + x^a_{G,p,t} \cdot w^a_G + x^a_{M,p,t} \cdot w^a_M, \quad \forall a, p, t. \tag{5.5}$$

We denote the ESGM of companies in sector a and year t by $\boldsymbol{x}^a_{ESGM,t} = (x^a_{ESGM,1,t}, \ldots, x^a_{ESGM,n_a,t})^\top$.

The ESGM scores have four weighted pillar subscores, and the unknown pillar subscore weight varies according to its business class. The next task is to estimate the pillar subscore weights. Since even regulatory authorities, including the European Banking Authority, have acknowledged the role of ESG scores in quantifying the company's riskiness and have identified a need to incorporate ESG risks into overall business strategies and risk management frameworks (EBA 2020), we propose the following optimization scheme in Equation (5.6) to estimate them, connecting the companies' ESGM scores with their risk performance, where $\boldsymbol{x}^a_{risk,t} = (x^a_{risk,1,t}, \ldots, x^a_{risk,n_a,t})^\top$ denotes the risk values of companies in sector a and year t as specified in Equation (5.7).

Refinitiv allows investors to build custom ESG scores by assigning customized pillar weights (Refinitiv 2022). Therefore, our proposed optimization scheme that links the scores and

riskiness can be applied for such a custom aggregation by investors. From the angle of corporate investments' net present value estimation process, Kudratova et al. (2020) also presented an optimization model for quantitative sustainability measurements.

$$
\begin{aligned}
(\hat{w}_E^a, \hat{w}_S^a, \hat{w}_G^a, \hat{w}_M^a) = \underset{w_E^a, w_S^a, w_G^a, w_M^a}{\arg\max} \quad & \sum_{t=t_1}^{t_2} \hat{\tau}_{risk}\Big(\boldsymbol{x}_{ESGM,t}^a, \quad \boldsymbol{x}_{risk,t}^a\Big) && \text{(5.6a)}\\
\text{subject to} \quad & w_E^a + w_S^a + w_G^a + w_M^a = 1, \quad \forall a, && \text{(5.6b)}\\
& w_E^a, w_S^a, w_G^a \geq 0.100, \quad \forall a, && \text{(5.6c)}\\
& w_M^a \geq 0, \quad \forall a, && \text{(5.6d)}\\
& w_E^a, w_S^a, w_G^a \geq w_M^a, \quad \forall a. && \text{(5.6e)}
\end{aligned}
$$

We analyze the ESGM scores' influence on their risk performance by focusing on their dependence in Equation (5.6a). We choose Kendall's tau (τ) as our dependence measure like in Section 5.5. Likewise, we can specify a generic risk function in Equation (5.7), where volatility values of companies in sector a and year t is denoted by $\boldsymbol{x}_{vol,t}^a = (x_{vol,1,t}^a, \ldots, x_{vol,n_a,t}^a)^\top$, and *vvrisk* values of companies in sector a and year t is given by $\boldsymbol{x}_{vv,t}^a = (x_{vv,1,t}^a, \ldots, x_{vv,n_a,t}^a)^\top$. More complex objective functions aiming to find optimal ESG portfolios for investors, such as proposed in Ahmed et al. (2021) and Pedersen et al. (2021), are subject to future research. Nevertheless, we propose a flexible framework that can take only VaR and volatility or their joint interaction with the artificially introduced risk measure, i.e., *vvrisk*. High ESGM scores should be linked to the strong VaR (e.g., Diemont et al. (2016)) and *vvrisk*, as well as the low volatility (e.g., Kumar et al. (2016)). Moreover, the pillar subscore weights can be estimated using data from the period $[t_1, t_2]$. For instance, we will be using the two recent years of the ESG data for an in-sample estimation, i.e., $t_1 = 2017, t_2 = 2018$, while $t_3 = 2019$ will be used for an out-of-sample evaluation.

$$
\hat{\tau}_{risk}\Big(\boldsymbol{x}_{ESGM,t}^a, \quad \boldsymbol{x}_{risk,t}^a\Big) = \begin{cases} \hat{\tau}\Big(\boldsymbol{x}_{ESGM,t}^a, \quad \boldsymbol{x}_{vv,t}^a\Big), & \text{if } risk = vvrisk, \\ \hat{\tau}\Big(\boldsymbol{x}_{ESGM,t}^a, \quad \boldsymbol{x}_{VaR,t}^a\Big), & \text{if } risk = VaR, \\ -\hat{\tau}\Big(\boldsymbol{x}_{ESGM,t}^a, \quad \boldsymbol{x}_{vol,t}^a\Big), & \text{if } risk = vol. \end{cases} \tag{5.7}
$$

The constraint in Equation (5.6b) ensures that the ESGM scores are between zero and 100, similar to the ESG scores. To exclude unrealistic scenarios, Equation (5.6c) ensures that each pillar subscore except the M pillar subscore has a positive lower bound, which is motivated by the lowest weight ever given to one of the E, S, and G pillar subscores in one of the industry groups by our data provider. To account for cases with no impacts of disclosing new ESG information on the risk performance, we set a lower bound of zero for the M pillar subscore weight in Equation (5.6d).

In Equation (5.6e), we assume that the E, S, and G pillar subscore weights are larger than or equal to the M pillar subscore weight. It considers the relative importance of already disclosed ESG information in the E, S, and G pillar subscores compared to the potential disclosure represented by the M pillar subscore. Disclosing reduces the companies' weighted M pillar subscore due to a decrease in the number of zero values entering the computation of the M pillar subscore. By assigning higher weights to the E, S, and G pillar subscores, our scheme encourages companies to disclose new ESG information, which usually positively impacts both their ESGM and ESG scores. Additionally, such an optimization scheme allows us to see for which business classes not yet disclosed ESG information might play a role and for which business classes a re-weighting scheme for the E, S, and G pillar subscores matters to strengthen the risk dependence.

Overall, in Equation (5.6), the constraints are linear, and the objective function is nonlinear in terms of the parameters with unknown derivatives. Thus, such a scheme can be solved by a derivative-free optimization algorithm dealing with linear constraints. Larson et al. (2019) provide a recent review of derivative-free optimization methods.

When our ESGM pillar subscore weights are estimated, we compute a company's ESGM score as follows in Example 5.4.

Example 5.4 *(The ESGM score calculation).*
Suppose the estimated pillar weights of the companies in Consumer Cyclicals ($a = 2$) are given by $w_E^2 = 0.258$, $w_S^2 = 0.122$, $w_G^2 = 0.498$, $w_M^2 = 0.122$. Then a company's ESGM score ($p = 1$) in Consumer Cyclicals with the following pillar subscores in 2017, $x^a_{E,1,2017} = 40.0$, $x^a_{S,1,2017} = 60.0$, $x^a_{G,1,2017} = 20.0$, $x^a_{M,1,2017} = 50.0$, is calculated as follows:
$x^2_{ESGM,1,2017} = 0.258 \cdot 40.0 + 0.122 \cdot 60.0 + 0.498 \cdot 20.0 + 0.122 \cdot 50.0 = 33.70.$

Empirical results: M pillar subscores

Since the companies in the data sets introduced in Section 5.3 contain the ESG categories with not yet reported ESG information as shown in Table 5.1, we account for their information disclosure, assigning an M pillar subscore as in Equation (5.4) in each of the ten sectors.

After computing the companies' M pillar subscores across ten sectors and three years in the S&P 500 and EuroStoxx 600, we observe that the mean M pillar subscore within each sector and year in both data sets is 50, as constructed. The M pillar subscore shows variation among ten sectors, and the standard deviation of the M pillar subscore changes from 18.31 for Consumer Non-Cyclicals to 28.16 for Real Estate in 2017 in the S&P 500. Likewise, the lowest and highest M pillar standard deviations are 15.75 for Utilities in 2019 and 28.28 for Real Estate in 2017 in the EuroStoxx 600. Consumer Non-Cyclicals and Utilities have the lowest percentage of the companies with undisclosed ESG information regarding an ESG category in Table 5.1 for the S&P 500 and EuroStoxx 600, respectively, and their M pillar subscore has the lowest standard deviation.

Table 5.8 reports the empirical Kendall's τ between the ESG scores, E, S, G, and M pillar subscores in Consumer Cyclicals in 2017 in the S&P 500. We see that the E, S, and G pillar subscores have positive medium-sized dependence on the ESG scores, while the M pillar subscore negatively depends on the ESG scores and the other pillars, as detailed now: when more ESG information is available for a company, the number of ESG categories with undisclosed ESG information decreases. Accordingly, its ESG category scores increase, assuming nothing changes in the other available information. Then since an ESG pillar subscore aggregates the underlying ESG category scores, the respective E, S, and G pillar subscores increase, increasing its ESG score. However, since the number of zero values used for the computation of the M pillar subscore decreases, its M pillar subscore goes down. Our findings are characteristically similar when considering other years, sectors, and EuroStoxx 600.

Table 5.8: Empirical Kendall's τ matrix of the ESG scores, E, S, G, and M pillar subscores in Consumer Cyclicals in the S&P 500 in 2017.

	ESG	E	S	G	M
ESG	1.00	0.66	0.76	0.46	-0.52
E		1.00	0.54	0.25	-0.57
S			1.00	0.31	-0.44
G				1.00	-0.20
M					1.00

Empirical results: in sample analysis

After computing the companies' E, S, G, and M pillar subscores, now, we focus on estimating the E, S, G, and M pillar subscore weights using the in sample data from 2017 and 2018 across ten sectors, linking the resulting ESGM scores to their risk measures as formulated in Equation (5.6). Precisely, we find $(\hat{w}_E^a, \hat{w}_S^a, \hat{w}_G^a, \hat{w}_M^a)$ for all sectors a in both data sets and aim to analyze for which sectors there is a risk strengthening effect using the M pillar, i.e., potential disclosure of ESG information.

We use the derivative-free optimization solver, LINCOA (Linearly Constrained Optimization Algorithm). LINCOA solves linearly constrained optimization problems without using derivatives of the objective function and uses a trust region method (Powell 2015). As Powell (2015) mentioned, we transform the linear equality in Equation (5.6b) into two inequalities. After running sensitivity analyses, the initial and final trust-region radii are set to 0.2 and 0.0005, respectively. The maximum number of function evaluations allowed is 10000. As the numerical optimization problems are sensitive to initial parameter values, we use ten different starting values and choose the optimal weights in correspondence with the best objective function value of ten runs. We do not observe multiple optimal solutions. All results are available upon request.

Tables 5.9 and 5.10 report the sectors' estimated four pillar subscore weights, the dependence of ESG, ESGM scores and risk for the S&P 500 and EuroStoxx 600, respectively. According to Tables 5.9 and 5.10, as postulated in Equation (5.6), the pillar subscore weights sum up to one; the E, S, and G pillar subscore weights are at least 0.100; the M pillar subscore is non-negative, and the E, S, and G pillar subscore weights are larger than or equal to the M pillar subscore weight in both data sets. We also see that ESGM scores are built on the M pillar (non-zero M pillar weight) for Consumer Cyclicals, Energy, Industrials, Technology, and Utilities in the S&P 500 and Consumer Cyclicals, Energy, Industrials, Financials, Healthcare, and Real Estate in the EuroStoxx 600. The sectors above the horizontal line have the positive M pillar subscore weight in Tables 5.9 and 5.10. Even though we observe the M pillar effect on the risk dependence in Consumer Cyclicals, Energy, and Industrials in the S&P 500 and EuroStoxx 600, the estimated pillar weights differ in both data sets. For instance, the M pillar weight for Consumer Cyclicals is 0.084 in the S&P 500 and 0.189 in the EuroStoxx 600. Moreover, Technology has the M pillar weight of 0.245 in the S&P 500 and of zero in the EuroStoxx 600. Therefore, the impact of the potential disclosure of ESG information, i.e., missing ESG information, on the VaR dependence changes by sectors and geographical regions.

Remarkably, the dependence of ESG scores and risk also depends on sectors and geographical regions as shown in Tables 5.9 and 5.10. While the dependence on Industrials is significant at the 10% level with the value of 0.118 in 2017 in the S&P 500, it is 0.060 without being significant at the same level in the EuroStoxx 600.

Table 5.9: New E, S, G, M pillar subscore weights (left) and the Kendall's τ between ESG, ESGM scores, and annual 95% VaR in 2017-2018 across sectors (right) in the S&P 500. The sectors above the horizontal line have the positive M pillar subscore weight.

Sector (S&P)	E	S	G	M	ESG and VaR 2017	ESG and VaR 2018	ESGM and VaR 2017	ESGM and VaR 2018
C. Cycl.	0.259	0.195	0.357	0.189	0.161 **	0.144 **	0.216 ***	0.215 ***
Energy	0.650	0.100	0.172	0.078	0.196 *	0.203 *	0.261 **	0.355 ***
Industrials	0.102	0.103	0.751	0.044	0.118 *	-0.071	0.137 *	0.019
Tech.	0.245	0.245	0.265	0.245	0.154 **	0.028	0.189 ***	0.094
Utilities	0.323	0.323	0.177	0.177	-0.079	-0.030	0.015	0.177 *
B. Materials	0.800	0.100	0.100	0.000	0.130	0.099	0.296 **	0.194
C. N-Cycl.	0.486	0.410	0.104	0.000	-0.047	0.001	0.015	0.053
Financials	0.100	0.100	0.800	0.000	0.020	-0.160	0.009	-0.121
Healthcare	0.730	0.100	0.169	0.000	0.309 ***	0.278 ***	0.297 ***	0.282 ***
Real Estate	0.100	0.800	0.100	0.000	-0.138	-0.312	-0.032	-0.286

Table 5.10: New E, S, G, M pillar subscore weights (left) and the Kendall's τ between ESG, ESGM scores, and annual 95% VaR in 2017-2018 across sectors (right) in the EuroStoxx 600. The sectors above the horizontal line have the positive M pillar subscore weight.

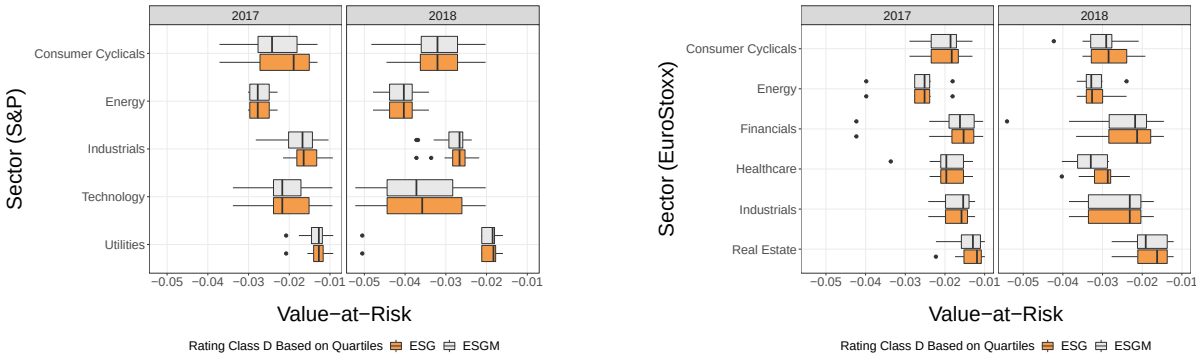
Sector (EuroStoxx)	E	S	G	M	ESG and VaR 2017	ESG and VaR 2018	ESGM and VaR 2017	ESGM and VaR 2018
C. Cycl.	0.240	0.541	0.135	0.084	0.144 **	0.157 **	0.187 ***	0.194 ***
Energy	0.238	0.485	0.175	0.102	0.471 ***	0.515 ***	0.559 ***	0.574 ***
Financials	0.100	0.100	0.700	0.100	-0.132	-0.088	-0.059	0.001
Healthcare	0.236	0.562	0.101	0.101	0.205 **	0.432 ***	0.258 **	0.466 ***
Industrials	0.100	0.714	0.100	0.086	0.060	0.029	0.094	0.070
Real Estate	0.700	0.100	0.100	0.100	-0.040	-0.151	-0.015	-0.114
B. Materials	0.100	0.618	0.282	0.000	-0.096	-0.171	-0.056	-0.099
C. N-Cycl.	0.683	0.217	0.100	0.000	0.054	0.085	0.129	0.115
Tech.	0.790	0.100	0.110	0.000	0.089	0.168 *	0.117	0.230 **
Utilities	0.233	0.542	0.225	0.000	0.127	0.296 **	0.206 *	0.333 ***

A company is assigned to a rating class (i.e., A, B, C, or D) based on its ESG score using thresholds or quartiles (Refinitiv 2021a). Thus, we group the companies with the highest to lowest ESG and ESGM scores in the first/second/third/fourth quartile as ESG and ESGM rating class A/B/C/D in each of the sectors where we observe the M pillar effect, respectively. Class D contains the companies with the lowest scores and might be excluded from ESG portfolios. For both data sets, we witness that low ESGM scores (class D) are associated with higher or equal median risks than low ESG scores (class D), except for Industrials in 2017 in the EuroStoxx 600, as demonstrated in Figure 5.8. Nonetheless, the median risk of ESG and ESGM scores in the EuroStoxx 600 is closer than that of those in the S&P500.

Additionally, since the VaR variation is low in some sectors, such as Utilities in the S&P 500, dividing the companies into classes with different VaR characteristics is hard. Thus, we can argue that the companies which have not yet released ESG information as much as their peers do, thereby having lower ESG scores than them in some sectors, might result in risk underestimation in the ESG portfolios using negative screening. Instead, the ESGM scores quantify better the companies that can be excluded, e.g., ESGM class D, and provide stronger risk performances for such portfolios than the ESG scores as seen for Consumer Cyclicals in the S&P 500 in 2017 and Healthcare in the EuroStoxx 600 in 2018.

Comparing the ESG and ESGM rating classes in Consumer Cyclicals in the S&P 500 in 2017 presents that ESGM scores move three/one companies from the ESG class D to the ESGM class C/B in 2017. Likewise, one company in the ESGM class B and three companies in the ESGM class C belong to the ESG class D in Healthcare in the EuroStoxx 600 in 2018. Such results reveal that the companies with low ESG scores might not necessarily provide the worst risk performances. Rather, their ESG scores could be low due to not yet disclosed ESG information the data provider does not explicitly point out. Still, these companies might publish more ESG information in the future, increasing their ESG scores, as modeled by their

Figure 5.8: Empirical 95% VaR of rating class D for ESG and ESGM in the S&P 500 (left) and EuroStoxx 600 (right) in 2017, 2018.



ESGM scores. Therefore, the ESGM scores can work not only to include missing information but also to allocate the companies to more appropriate risk classes. Similar results hold for the remaining sectors in both data sets with a non-zero M pillar weight (available upon request).

Empirical results: out-of-sample analysis

ESGM scores provide stronger risk dependence and identify more reliably the companies for exclusion strategies in ESG portfolios than the ESG scores in the previous part. Now we use the estimated E, S, G, and M pillar subscore weights for another year to calculate ESGM scores, where the ESG data is still non-definitive, and companies can publish their ESG information in time. Accordingly, we perform the out-of-sample analyses using the data sets for 2019.

First, we calculate the companies' predicted ESGM scores for 2019 in each sector and data set:

$$x^a_{ESGM,p,2019} = x^a_{E,p,2019} \cdot \hat{w}^a_E + x^a_{S,p,2019} \cdot \hat{w}^a_S + x^a_{G,p,2019} \cdot \hat{w}^a_G + x^a_{M,p,2019} \cdot \hat{w}^a_M, \forall a, p,$$

where $\hat{w}^a_E$, $\hat{w}^a_S$, $\hat{w}^a_G$, and $\hat{w}^a_M$ are the estimated weights using the training data of 2017 and 2018 in Tables 5.9 and 5.10.

Second, we analyze the dependence between the ESG scores, ESGM scores and VaR in 2019 in Table 5.11, where the sectors above the horizontal line have the positive M pillar subscore weight in Tables 5.9 and 5.10. As seen, the out-of-sample analysis also confirms the higher risk dependence for the ESGM scores than the ESG scores in all cases but Energy in the EuroStoxx 600. However, the significance of the risk dependence is better using the ESGM scores than the ESG scores in Energy in the S&P 500. Likewise, the ESGM scores provide better risk dependence than the ESG scores in all sectors, except Consumer Non-Cyclicals and Healthcare in the S&P 500. Nonetheless, the M pillar weight is zero for both sectors, i.e., we do not find the impact of not yet disclosed ESG information on the risk dependence.

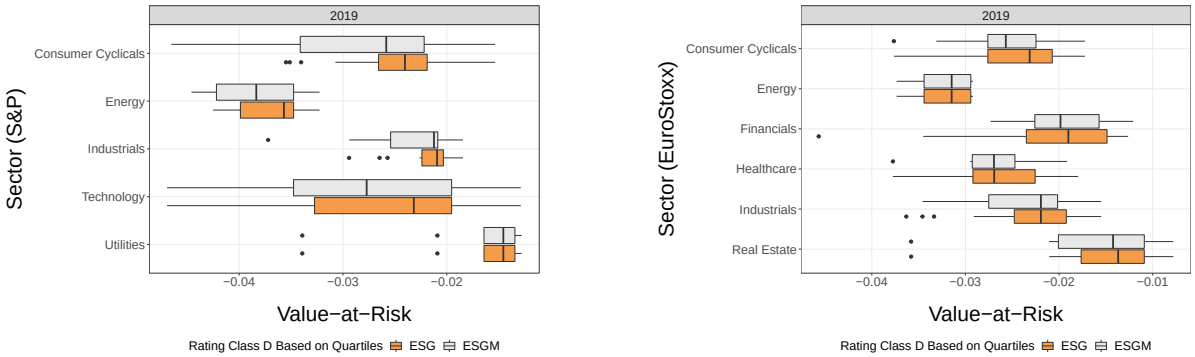
Table 5.11: Kendall's τ between ESG, ESGM scores, and 95% VaR in 2019 across sectors (out-of-sample) in the S&P 500 (left) and EuroStoxx 600 (right). The sectors above the horizontal line have the positive M pillar subscore weight in Tables 5.9 and 5.10, respectively.

Sector (S&P500)	ESG	ESGM
C. Cycl.	0.036	0.096
Energy	0.229 *	0.326 **
Industrials	-0.103	0.039
Tech.	0.051	0.068
Utilities	-0.099	-0.010
B. Materials	-0.075	0.043
C. N-Cycl.	0.115	0.115
Financials	-0.102	-0.031
Healthcare	0.161 **	0.088
Real Estate	-0.310	-0.259

Sector (EuroStoxx)	ESG	ESGM
C. Cycl.	-0.019	0.015
Energy	0.529 ***	0.529 ***
Financials	-0.123	-0.031
Healthcare	0.375 ***	0.386 ***
Industrials	0.030	0.076
Real Estate	-0.182	-0.120
B. Materials	-0.064	-0.040
C. N-Cycl.	0.168 *	0.200 **
Tech.	0.073	0.101
Utilities	0.270 **	0.286 **

Finally, Figure 5.9 suggests that the ESGM class D presents higher median risks than the ESG class D in all but Utilities in the S&P 500. Similarly, the median risk ESGM scores provided for class D is higher than that of ESG scores provided for Consumer Cyclicals, Financials, and Real Estate in the EuroStoxx 600. However, the median risk of the ESG and ESGM scores for class D seems to be closer in the EuroStoxx 600 than the S&P 500. Hence, the ESG portfolios using negative screening could benefit from the ESGM scores in terms of the risk performance, supporting our findings using the in sample data, even though there are some differences across sectors and regions.

Figure 5.9: Empirical 95% VaR of rating class D for ESG and ESGM in the S&P 500 (left) and EuroStoxx 600 (right) in 2019.



5.7 Application: vine copula mixture models

In this section, we present a case study of vine copula mixture models (VCMM) explained in Chapter 3. Since COVID-19 greatly impacted the Energy sector, our analysis focuses on 24 Energy companies in the S&P 500 and uses their daily log returns from January 2, 2017, to December 30, 2020, including the COVID-19 period. In total, we have 997 observations.

We divide the companies into four ESG rating classes $I_i \in \{A, B, C, D\}$ using their 2017 ESG scores' quartiles $i = 1, \ldots, 24$, as we did in Section 5.6. The companies in the class A is regarded as the most ESG responsible companies, while the class B follows it. However, the classes C and D can be regarded as the ESG irresponsible companies. Then we calculate the four ESG indices $ESG_{t,j}$ as a linearly weighted combination of the associated companies' daily log returns $y_{t,i}$ and market capitalizations weights M_i for $j \in \{A, B, C, D\}$, $t = 1, \ldots, 997$, and $i = 1, \ldots, 24$ as follows:

$$ESG_{t,j} = \frac{\sum\limits_{\substack{g \in [1,24] \\ I_g \in j}} y_{t,g} \cdot M_g}{\sum\limits_{\substack{g \in [1,24] \\ I_g \in j}} M_g} \quad \text{for} \quad t = 1, \ldots, 997, \quad j \in \{A, B, C, D\}.$$

For instance, the ESG index A at day t, $ESG_{t,A}$, is based on the daily log returns and market capitalization weights of the six companies whose ESG scores in 2017 are higher than the others. As a result, we have time series data of four ESG indices with 997 observations.

The main interest is to estimate the dependence structures among ESG responsible and irresponsible companies, represented by the corresponding ESG indices, in 2017-2020, where the ESG scores are non-definitive. Thus, we will run the VCMM with a different number of components and initial partition. Moreover, we work with C-vines to see which ESG index is the main driver of the dependence (if it is). Alternatively, one may work with D-vines, where the order of the nodes from left to right corresponds to the ESG indices A, B, C, and D. Even though such an application might be a natural order among the four different ESG indices, we work with the non-definitive ESG scores whose order may change in the future.

Marginal models

Since the daily log returns of the four ESG indices have the stochastic volatility shown up in the (partial) autocorrelation functions in Figure 5.10, we apply the generalized autoregressive conditional heteroskedastic model GARCH(1,1) with the innovation distribution standardized and being symmetric Student's t_ν with mean zero, variance one, and $\nu > 2$ to remove the serial dependence, getting the GARCH-filtered data. The (partial) autocorrelation functions of standardized residuals in Figure 5.10 show the removal of the serial dependence and GARCH

effects. Here we employ a quasi-maximum-likelihood estimator (QMLE) for GARCH parameters. Then we estimate the vine copula mixture model parameters based on the residuals obtained from the QMLE of GARCH parameters.

Let R_t $(t = 1, \ldots, n)$ be the daily log returns of an ESG index so that for each index j, we denote the returns as $R_{t,j}$ at time t. Then the univariate GARCH(1,1) model is

$$R_{t,j} = \mu_j + \sigma_{t,j} Z_{t,j}, \quad \sigma^2_{t,j} = \omega_j + \alpha_j R^2_{t-1,j} + \beta_j \sigma^2_{t-1,j}, \quad j \in \{A, B, C, D\}, t = 1, \ldots, 997,$$

where $\omega_j > 0$, $\alpha_j > 0$, $\beta_j > 0$ for each j, and $Z_{t,j}$ are independent and identically distributed innovations over t. Fitting univariate GARCH(1,1) model to each of the four daily log returns gives the estimated parameters in Table 5.12.

Table 5.12: GARCH(1,1) parameters with standardized Student's t_ν distributed innovations for ESG indices' (absolute) daily log returns in 2017-2020.

Index	μ_j	ω_j	α_j	β_j	ν_j
A	-8.96 x 10^{-6}	6.94 x 10^{-7}	0.113	0.885	5.477
B	1.21 x 10^{-4}	2.63 x 10^{-7}	0.097	0.894	5.456
C	7.55 x 10^{-5}	9.88 x 10^{-8}	0.124	0.871	6.437
D	4.93 x 10^{-5}	2.41 x 10^{-7}	0.076	0.905	4.959

Now we have independent observations across time, which can be combined by considering a VCMM. Rather than converting the GARCH-filtered data into the copula data, we let the VCMM do this by selecting univariate parametric margins for each component. Therefore, we run our analysis based on GARCH models whose innovations follow a vine copula mixture model. Before our study, Lee and Lee (2011) developed a forecasting algorithm for VaR based on GARCH-type models whose innovations follow a Gaussian mixture model.

In future applications, we will allow the estimation of the GARCH parameters with that of the VCMM parameters simultaneously so that GARCH-filtered data is obtained by considering the dependence structure of the variables. In this case, we obtain a maximum log-likelihood estimator (MLE). However, since the calculation of the MLE in GARCH models is complicated, we apply a simple step-by-step estimation procedure in this section. We also remark that the standardized residuals of GARCH models have zero mean and unit variance. Therefore, one might need to impose some constraints on the VCMM estimator or modify it afterward.

VCMM results

We fit the VCMM from one to three components using three initial partitions: Gaussian mixture models (GMM), k-means, and model-based hierarchical clustering (hcVVV) (Scrucca et al. 2016). As we aim to identify the driver index of the dependence in different periods (if it exists), we only work with C-vines.

Figure 5.10: Autocorrelation (ACF) and partial autocorrelation (PACF) function plots with the blue confidence interval bands: ESG B index's daily log returns in 2017-2020 (top) and its GARCH-filtered standardized residuals (bottom).

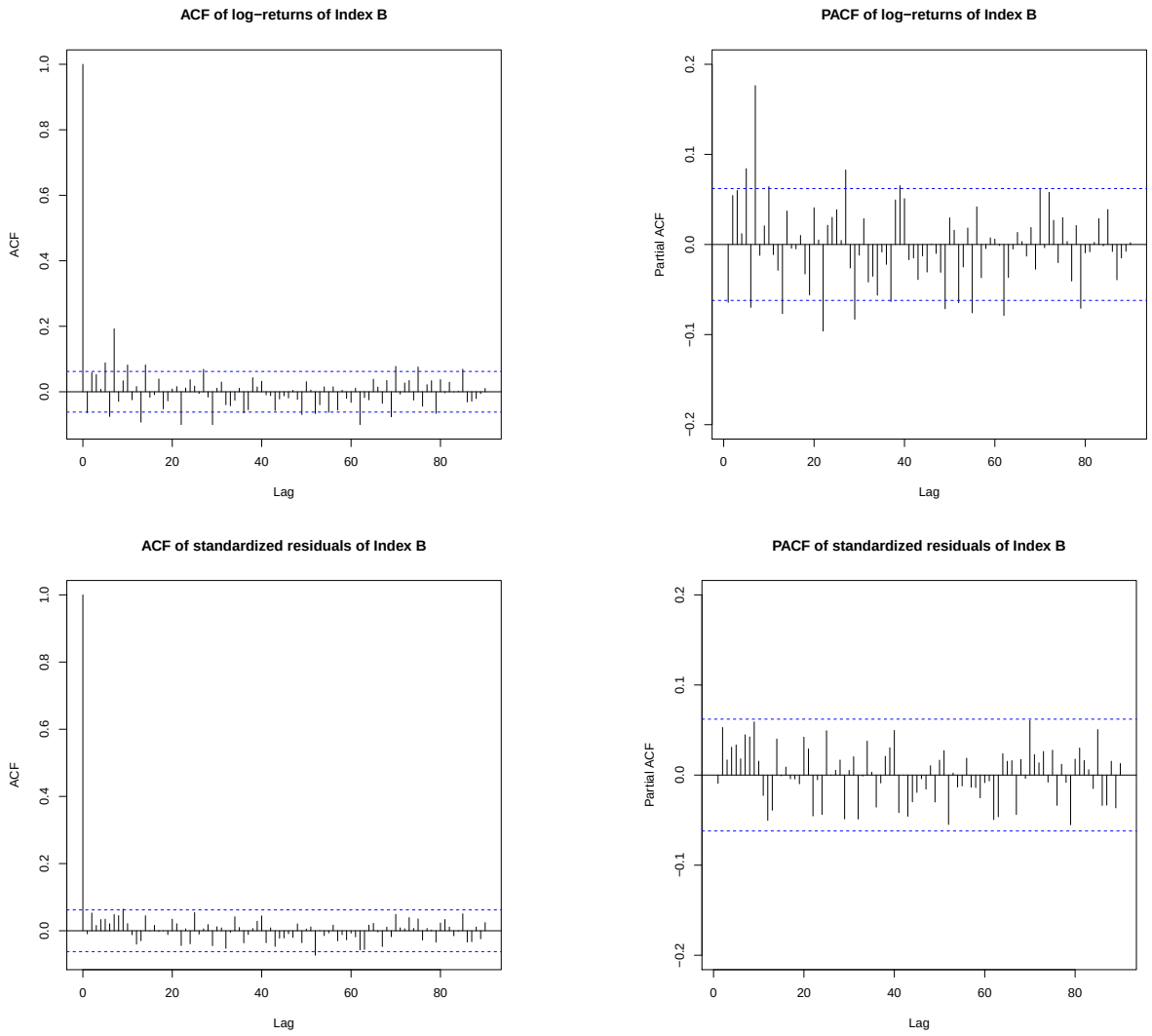


Table 5.13 gives the VCMM results. We see that the BIC of the single-component C-vine copula model is better than the others; however, the two-component C-vine copula mixture model initialized by GMM has the best AIC among them. Moreover, the log-likelihood of the three-component C-vine copula mixture model initialized by k-means is higher than that of others. As discussed in Chapter 3, the BIC is not a good measure to compare vine copula mixture models, and tests must be developed to decide if a mixture model is needed for data.

In addition, the GARCH-filtered data is assumed to have zero mean and unit variance,

and the single-component C-vine copula model fits the univariate margins which satisfy it. Likewise, the assumption is satisfied for the marginal distributions fitted by the two-component C-vine copula mixture model initialized by GMM and by the three-component C-vine copula mixture model initialized by k-means. Full results are available on request.

Table 5.13: ESG index's GARCH-filtered (residuals) daily log returns in 2017-2020: vine tree structure of each component, initial partition of the VCMM, the number of components, log-likelihood, AIC and BIC values. The best value for each column is highlighted.

Vine structure	Initial partition	#components	loglik.	#par	BIC	AIC
C-vine	-	1	-3333.48	23	6825.77	6712.96
C-vine	kmeans	2	-3326.32	54	7025.50	6760.64
C-vine	GMM	2	-3326.07	30	6859.29	6712.15
C-vine	hcVVV	2	-3334.57	54	7042.00	6777.14
C-vine	kmeans	3	-3293.00	67	7048.62	6720.00
C-vine	GMM	3	-3310.67	50	6966.58	6721.35
C-vine	hcVVV	3	-3338.73	68	7146.99	6813.47

Financial interpretation

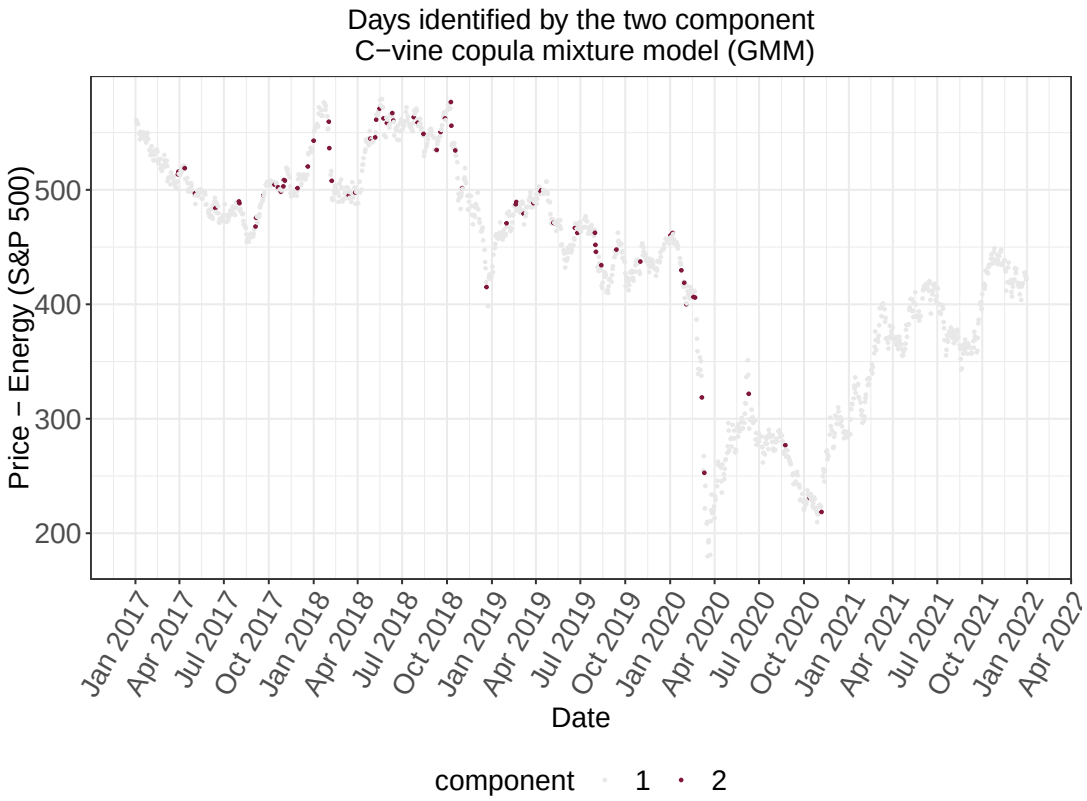
To compare the models for financial interpretation, we focus on the (unconditional) dependence structure of each component given by the model and the observations belonging to the component.

Table 5.14 gives the estimated C-vine copula models' first tree level. For the single-component C-vine copula model, we see that the root node is the ESG index B and has a symmetric lower and upper tail dependence with others. Therefore, when very good/bad news comes to the ESG index B, they also affect the other indices. The multivariate Gaussian distribution assumption would not be adequate for risk management in this setup. Further, an explanation of the selection of index B as the root node could be that investor preferences are oriented toward the ESG responsible companies. Moreover, given that such companies' market capitalization is higher than those in the indices C and D, another reason could be the size of the companies. However, why the ESG index B is more dependent on others than the index A needs further investigation.

Table 5.14 shows that the two-component C-vine copula mixture model chooses the index B/C as the root node in the first/second component. However, the number of observations (days) assigned to the second component is 74 out of 997. Such days can be the beginning of the pre-crisis, crisis, and post-crisis periods. To investigate them, we focus on the price data of the Energy sector in the S&P 500 in 2017-2021 in Figure 5.11. We observe that the second component includes the dates which can be regarded as the end of the crisis periods, such as 21 December 2018 and 06 November 2020. The latter might be the end

of the COVID-19 crisis for the Energy sector. Likewise, 30 January 2020 is included in the second component, around when the COVID-19 crisis may have impacted the Energy sector negatively. Nevertheless, other dates included in the second component, such as 7 May 2018 and 19 February 2020, need further financial interpretation and investigation.

Figure 5.11: The S&P 500's Energy sector price in 2017-2020: time series plot with the days colored by the two-component C-vine copula mixture model initialized by the GMM.



We see in Table 5.14 that the dependence structure and strength among the ESG indices differ by their components identified by the three-component C-vine copula mixture model. While the ESG index B is the driver of the dependence in the second and third components, the first component identifies index C as the root node. Figure 5.12 illustrates that the first component usually corresponds to the days with negative log returns. On the other hand, the second and third components can be regarded as positive and stable days, respectively. On the negative days, i.e., when bad news arrives, the less ESG responsible companies tend to drive dependence, which is consistent with the findings of Czado et al. (2022), who found that the ESG index C has the strongest dependence with the S&P500 in times of crisis. What has also been seen is that the dependence strength among the indices B and C is 0.70 with the single vine copula model, but it changes from 0.39 to 0.43 based on the positive, negative,

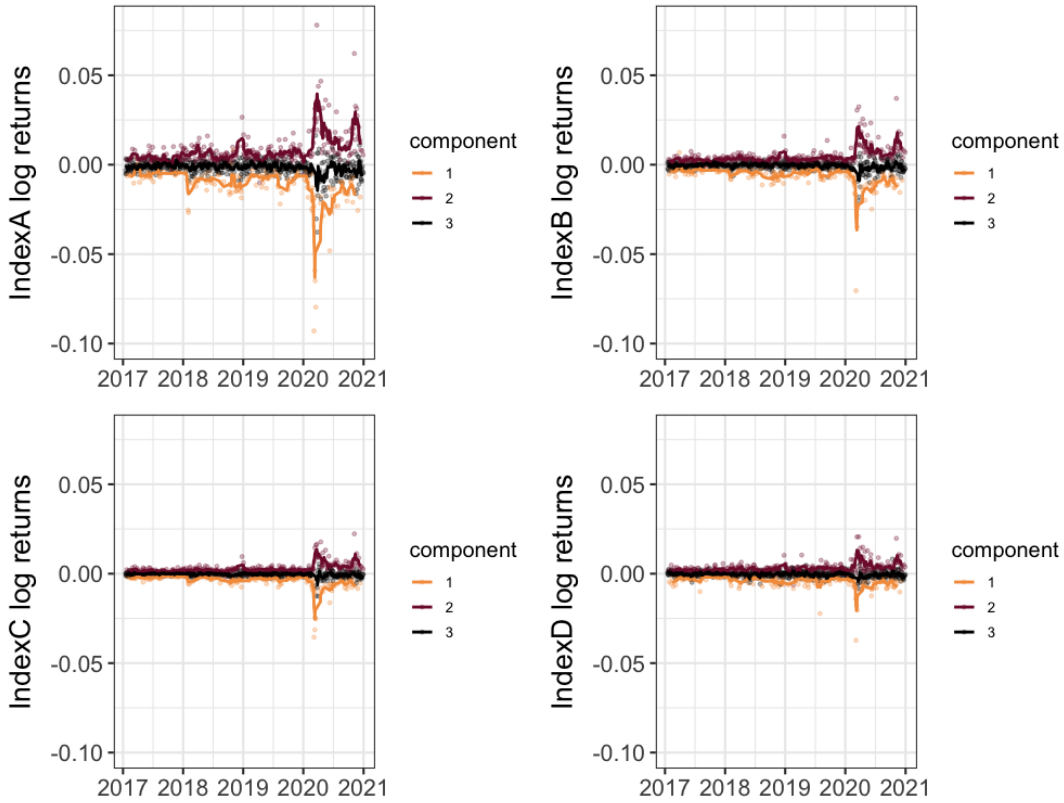
or stable days with the three-component C-vine copula mixture model. In addition, while a symmetric tail dependence exists between the indices B and C with the single-component C-vine copula model, the three-component C-vine copula mixture model identifies an asymmetric tail dependence between them in positive days. As expected, the tail dependence does not show up in the stable days as represented by the bivariate Frank copula in the third component.

Table 5.14: ESG index's GARCH-filtered (residuals) daily log returns in 2017-2020: The first tree level of the selected vine copula models with their dependence estimated by Kendall's $\hat{\tau}$, the estimated upper and lower tail dependence coefficients, $\hat{\lambda}^{upper}$ and $\hat{\lambda}^{lower}$, defined in Equations (2.1) and (2.1), respectively, the estimated mixture weights of the corresponding components $\hat{\pi}_j$ defined in Equation (3.14) and the number of observations assigned to each component n_j.

Model	Component (j)	Edge	Copula family	$\hat{\tau}$	$\hat{\lambda}^{lower}$	$\hat{\lambda}^{upper}$	$\hat{\pi}_j$	n_j
Single-component C-vine	1	B,D	t	0.61	0.46	0.46	1.00	997
		B,C	t	0.70	0.49	0.49		
		B,A	t	0.65	0.57	0.57		
Two-component C-vine (GMM)	1	B,D	Gaussian	0.63	0.00	0.00	0.89	923
		B,C	Gaussian	0.71	0.00	0.00		
		B,A	Gaussian	0.68	0.00	0.00		
	2	C,D	Gumbel	0.51	0.00	0.60	0.11	74
		C,B	Gaussian	0.67	0.00	0.00		
		C,A	BB1	0.57	0.61	0.45		
Three-component C-vine (k-means)	1	C,D	Clayton	0.24	0.33	0.00	0.19	174
		C,B	Clayton	0.42	0.62	0.00		
		C,A	Survival Gumbel	0.39	0.48	0.00		
	2	B,D	Joe	0.33	0.00	0.55	0.31	301
		B,C	Survival BB1	0.43	0.19	0.55		
		B,A	Joe	0.35	0.00	0.58		
	3	B,D	Frank	0.30	0.00	0.00	0.50	522
		B,C	Frank	0.39	0.00	0.00		
		B,A	Survival BB8	0.30	0.00	0.00		

All in all, even though the single-component C-vine copula model and two-component, three-component C-vine copula mixture models provide some financial interpretation, we can compare the candidate models for portfolio VaR and conditional tail expectation in financial risk management. Research is high on the agenda to deal with such cases. We also note that the ESG scores used to build ESG indices are based on non-definitive ESG scores. Hence, they may not be good representatives of companies' ESG responsibility levels.

Figure 5.12: ESG indices' daily log returns in 2017-2020: time series plot with the observations' component given by the three-component C-vine copula mixture model initialized by k-means. Each line represents the moving average of the last five days.



5.8 Discussion

Our research is in alignment with the recent discussions and has implications for researchers, companies, legislators, investors and asset managers, and data providers. The Sustainable Finance Roadmap, released in February 2022 by European Securities and Markets Authority (ESMA), also lists the main challenges which need action in analyzing ESG-related risks. They report that "Overall, data gaps, low quality and a lack of transparency may lead to misrepresentation and to a misallocation/mispricing of investments" and call for actions to assess the issue of the data quality affecting the ESG data users (Page 19 of ESMA (2022)).

For researchers

It is reasonable to expect researchers to use non-definitive ESG scores (e.g., Dicuonzo et al. (2022), Ding et al. (2021), Valbuena-Hernandez and Mandojana (2022), and Zanin (2021))

since they include the five most recent years. Hence, researchers should report their data extraction date, including its week, month, and year, to understand the research inconsistencies better since Refinitiv updates its ESG scores weekly.

Moreover, even though the concept of sustainability reporting and performance measures has been discussed for many years by scholars (Milne and Gray 2013; Searcy 2012), not only providers but also researchers can investigate how to assign unified and justifiable pillar/category weights. Even though the recent efforts include the incorporation of financial statements items with a machine learning approach to represent ESG scores (D'Amato et al. 2021a,b), there is room to develop more stable ESG scoring methodologies. EBA has recently set out the pillar subscore disclosure guidelines to improve companies' ESG information consistency and comparability (EBA 2022).

The fact that the missing information's potential impact on the risk differs for sectors and geographical regions suggests investigating the determinants of the occurrence and distribution of the companies' missing ESG information among ESG categories across regions and sectors. Such an analysis could also provide insights into which and how ESG information should be disclosed to measure the companies' ESG performance and responsible investing accurately. Recently, ESMA also stated regarding the ESG data that "These data needs are currently not fulfilled by the data disclosed by companies. The data gaps can neither be fully bridged by third-party ESG data or by rating providers whose methodologies, limitations and assumptions need to become more transparent." (Page 19 of ESMA (2022)). Lagasio and Cucari (2019) already identified board independence, board size, and women's directorship as the empowering factors of non-financial disclosures. However, as they pointed out, there is a gap in the studies investigating the determinants of the non-financial information released by companies across sectors.

Furthermore, for researchers, in addition to showing companies' potential to disclose missing ESG information, the M pillar can be used as a proxy of the current ESG disclosure quality adapted for sectoral peculiarities like other pillars. Future studies can perform a regression analysis, where the dependent variable is a financial performance measure, and independent variables are E, S, G, and M pillar subscores and company specifics like their market capitalization. This would allow for analyzing the impact of the current disclosure quality on financial measures. However, we remark that such an analysis does not provide scores comparable to ESG scores; thus, they might be difficult to be used by investors in the same manner. Alternative approaches for a regression analysis can encode the current disclosure quality as a binary variable based on the industry median by counting the number of ESG categories disclosed as proposed in Santamaria et al. (2021).

For companies

We show that the amount and time of the ESG information shared by the companies greatly affect how the ESG scores are constructed. Moreover, our analyses imply that companies could be excluded from investment portfolios, not necessarily because of their actual ESG

performance but possibly due to their lack of and low speed of the ESG disclosure mechanism. Therefore, a focus should be on providing complete disclosure material, allowing investors to manage ESG-related risks better and make sustainable investments for the world.

For legislators

We also show that EuroStoxx 600 companies have less missing information in the ESG categories than the S&P 500 companies. In the European Union (EU), sustainable finance legislations have been fine-tuned. For instance, the EU's Directive 2014/95/EU sets out that relevant, large, and public interest EU companies must disclose ESG information annually.[5] On the other hand, there have not yet existed any mandatory sustainability-related disclosures in the United States (International Platform on Sustainable Finance 2021). However, the Federal Reserve joined the Network for Greening the Financial System, a global network of central banks, and might take action about the applicable commitments towards a sustainable economy.[6] The mandatory context might have a positive impact on the credibility of the non-financial disclosures for the companies (Mazzotta et al. 2020).

For investors and asset managers

Our research also has several implications for investors and asset managers. For example, Revelli (2017) argues that ESG information is important in mainstream analysts' and fund managers' investment strategies. But, on the other hand, the identification of socially responsible and irresponsible companies to be included in socially responsible funds (SRF) has been an issue (Gangi and Varrone 2018). As shown in Section 5.5, if the selection of socially responsible companies regarding the ESG scores is based on the recent ESG scores, the time of the selection may significantly impact the selected companies and their risk dependence. It implies that the social responsibility level of the same SRF based on the ESG scores might differ in time. Therefore, it is important to follow the ESG scores and their methodology changes to ensure the stable selection of the companies. It is also clear that the ESG scores show the "relative goodness" of the companies based on the available ESG information of their sector peers. With the amount of missing information in the recent ESG scores reported and discussed in Section 5.6, investors may even consider constructing their own scores. Refinitiv also suggests such construction for investors (Refinitiv 2022).

Moreover, our results show the importance of knowing and understanding what is behind the ESG scores for investors. The investors can exclude the companies based on their low ESG scores in a given sector from their portfolios; however, the companies still disclose their ESG information in time, increasing their ESG scores.

[5] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32014L0095

[6] https://www.federalreserve.gov/newsevents/pressreleases/bcreg20201215a.htm

Additionally, we show that the dependence among daily log returns of company groups formed by the companies' ESG performances changes based on the characteristics of the day, e.g., positive or negative returns, and the VaR estimation of portfolios should be investigated regarding these.

For data providers

We conjecture that such results in Section 5.5 can hold using ESG scores from other providers. For instance, Morgan Stanley Capital International (MSCI) states in their ESG scoring methodology, "Companies are monitored on a systematic and ongoing basis, including daily monitoring of controversies and governance events. New information is reflected in reports on a weekly basis and significant score changes trigger analyst review." (Page 13 of MSCI (2022)). Another provider, Sustainalytics, shares, "Scores are updated on an annual basis... Clients are given advance notice of upcoming structural changes, like the addition of new data points, that can be implemented once a year." (Pages 7 and 14 of Sustainalytics (2022)). In addition, Sustainalytics, also encodes the missing raw scores based on company disclosures as zero.[7] However, the efforts to make the ESG scoring construction methodology transparent by data providers would support data users.

5.9 Conclusion

As the divergence of ESG scores from different ESG providers has been a vital discussion point, a suggestion to deal with the issue has been to use ESG scores from various providers. Nevertheless, we show that the ESG scores might not even be stable within the same provider due to allowing ESG scores to change for the five most recent years. Such ESG scores are called non-definitive.

Even though we work with simple optimization and data-mining schemes, we show that a re-weighting scheme for pillar subscores and category scores might dramatically alter ESG scores. If a weighting scheme changes, we expect it to be announced in the methodology document by Refinitiv. However, we show that the initial disclosure or an update of ESG information leads to changes in ESG scores, and Refinitiv does not currently state such modifications. Moreover, these reasons for modifications in ESG scores cause the companies' ranks regarding their ESG scores to change. Thus, some companies are regarded as more sustainable than before, even though they are not. Further, since researchers use the ranks to classify companies as high/low ESG companies (e.g., Demers et al. (2021) and Engelhardt et al. (2021)), a cause of the mixed literature results regarding the link between ESG scores and financial performance could be such score instabilities and the studies' different data extraction dates. The results are consistent with Bae et al. (2021), using non-definitive

[7] https://www.sustainalytics.com/docs/default-source/meis/kone_oyj_riskratingsreport_18032021.pdf?sfvrsn=577d22c8_0

ESG scores, who believe that their inconsistency with previous studies can be due to the data extraction timing.

Moreover, we show that even focusing on the S&P 500 companies, which are very much under scrutiny, the link between ESG scores and risk is not clear-cut. Additionally, such a link depends on sectors and the date and amount of ESG information disclosure.

We propose a new pillar subscore, the so-called Missing (M) pillar subscore, to explicitly consider the companies' potential of disclosing missing ESG information. By doing so, we introduce a new Environmental, Social, Governance, and Missing (ESGM) score. In addition, our study formulates an optimization scheme to link the companies' ESGM scores and riskiness. Such a scheme encourages companies to disclose more ESG information and evaluate the impact of such additional information on financial risk. Furthermore, it can be used by investors to build customized scores.

We evaluate the risk performance of the proposed ESGM scores and ESG scores using in-sample and out-of-sample data. The in-sample data analysis allows us to estimate the ESGM scores' pillar weights, which are used to compute the ESGM scores. The out-of-sample data analysis tests the power of the ESGM scores regarding their risk dependence. Using the S&P 500 and EuroStoxx 600 companies' non-definitive ESG data provided by Refinitiv, we show that the ESGM scores provide stronger risk dependence than the ESG scores in some sectors using both the in-sample and out-of-sample data. We argue that a potential disclosure of missing ESG information impacts the risk in these sectors. This approach potentially supports the investment decisions as an ESG exclusion strategy depends on the ESG rating class, which is affected by missing information. Furthermore, incorporating the potential of possible disclosure allows to include companies in the portfolio that would otherwise be excluded as too much data is missing at the time of the investment decision. Nonetheless, the dependence of risk and ESG/ESGM scores and the impact and amount of the missing ESG information change with sectors and geographical regions. We discuss our research's further implications for researchers, companies, legislators, investors and asset managers, and data providers.

Our study is limited by the small number of companies within each sector. Future studies should consider applying our re-weighting, M pillar subscore, and ESGM scores formulations to data from a larger number of companies. The second limitation of our research is that it does not account for common risk factors. Nevertheless, research is high on the agenda to deal with such issues. For future studies, researchers can consider a multivariate analysis controlling for common factors to investigate the impact of the (non-definitive) ESG score changes on risk. Moreover, our data-mining schemes might be applied to the link between the (non-definitive) ESG scores and financial performance measures. Such an application would reveal how much such a relationship might change in time. Likewise, the performance of the socially responsible funds based on the ESG scores might be compared to the current scores versus the "data-mined" scores. Finally, necessary modifications for the M pillar subscore and exploring optimization schemes with different objective functions can be considered in the future to evaluate the ESGM scores also from other ESG data providers.

5.10 Supplementary materials

Missing information in ESG categories

Table 5.15: Percentage of the companies whose ESG category score is missing (zero) classified by year, sector, and category in the S&P 500 and EuroStoxx 600. Only the categories for which at least a company in a given year does not contain any information are reported.

$t = 2017$ S&P - EuroSt.	Resource Use	Emissions	Environmental Innovation	Human Rights	Product Responsibility	CSR Strategy
B. Materials	0% - 0%	4% - 2%	26% - 18%	9% - 18%	4% - 2%	4% - 2%
C. Cycl.	16% - 4%	13% - 5%	43% - 29%	27% - 28%	0% - 10%	34% - 6%
C. N-Cycl.	0% - 0%	3% - 0%	15% - 20%	8% - 24%	0% - 0%	8% - 2%
Energy	4% - 0%	0% - 0%	62% - 29%	38% - 12%	8% - 0%	4% - 0%
Financials	18% - 9%	18% - 3%	42% - 28%	52% - 35%	0% - 11%	30% - 3%
Healthcare	14% - 0%	20% - 0%	66% - 55%	29% - 36%	0% - 6%	29% - 3%
Industrials	14% - 4%	12% - 2%	32% - 21%	28% - 33%	3% - 6%	25% - 8%
Real Estate	18% - 4%	14% - 4%	25% - 23%	61% - 69%	4% - 35%	29% - 15%
Technology	12% - 2%	21% - 2%	29% - 27%	18% - 22%	0% - 9%	40% - 7%
Utilities	3% - 0%	3% - 0%	14% - 4%	34% - 32%	0% - 0%	7% - 0%
$t = 2018$ S&P - EuroSt.	Resource Use	Emissions	Environmental Innovation	Human Rights	Product Responsibility	CSR Strategy
B. Materials	0% - 0%	4% - 0%	26% - 18%	4% - 8%	0% - 0%	4% - 2%
C. Cycl.	16% - 4%	10% - 3%	44% - 29%	21% - 15%	0% - 4%	30% - 3%
C. N-Cycl.	0% - 0%	0% - 0%	15% - 17%	3% - 20%	0% - 0%	5% - 0%
Energy	0% - 0%	0% - 0%	62% - 29%	25% - 12%	4% - 0%	4% - 0%
Financials	13% - 7%	15% - 2%	42% - 19%	42% - 22%	0% - 6%	22% - 3%
Healthcare	7% - 0%	12% - 0%	62% - 48%	21% - 18%	0% - 0%	25% - 0%
Industrials	12% - 4%	9% - 2%	34% - 20%	18% - 21%	2% - 6%	28% - 5%
Real Estate	11% - 4%	11% - 0%	25% - 19%	43% - 58%	0% - 31%	11% - 0%
Technology	9% - 2%	20% - 0%	29% - 29%	17% - 16%	0% - 9%	34% - 4%
Utilities	0% - 0%	3% - 0%	14% - 11%	31% - 11%	0% - 0%	3% - 0%
$t = 2019$ S&P -EuroSt.	Resource Use	Emissions	Environmental Innovation	Human Rights	Product Responsibility	CSR Strategy
B. Materials	0% - 0%	4% - 0%	26% - 18%	4% - 2%	0% - 0%	0% - 2%
C. Cycl.	14% - 4%	8% - 1%	44% - 27%	16% - 10%	0% - 1%	26% - 0%
C. N-Cycl.	0% - 0%	0% - 0%	13% - 17%	0% - 5%	0% - 0%	5% - 0%
Energy	0% - 0%	0% - 0%	54% - 24%	21% - 6%	4% - 0%	0% - 0%
Financials	8% - 5%	8% - 1%	35% - 16%	32% - 15%	0% - 1%	10% - 2%
Healthcare	4% - 0%	4% - 0%	59% - 45%	11% - 3%	0% - 0%	14% - 0%
Industrials	8% - 1%	5% - 1%	31% - 19%	14% - 11%	2% - 4%	23% - 2%
Real Estate	11% - 4%	4% - 0%	18% - 15%	43% - 42%	0% - 19%	7% - 0%
Technology	7% - 2%	15% - 0%	26% - 27%	12% - 11%	0% - 7%	20% - 2%
Utilities	0% - 0%	3% - 0%	10% - 7%	24% - 4%	0% - 0%	3% - 0%

Notation for missing values

Table 5.16: Mathematical indices, data, and their notation used in the chapter.

Index&Data	Notation
Sector (Business class)	$a = 1, \ldots, 10$
Total number of companies	n
Company z	$z = 1, \ldots, n$
Total number of companies in sector a	n_a
Company p in sector a	$p = 1, \ldots, n_a$
Year	$t = 2017, \ldots, 2019$
VaR of company p in sector a and year t	$x^a_{VaR,p,t}$
VaR values of companies in sector a and year t	$\boldsymbol{x}^a_{VaR,t} = (x^a_{VaR,1,t}, \ldots, x^a_{VaR,n_a,t})^\top$
Volatility of company p in sector a and year t	$x^a_{vol,p,t}$
Volatility values of companies in sector a and year t	$\boldsymbol{x}^a_{vol,t} = (x^a_{vol,1,t}, \ldots, x^a_{vol,n_a,t})^\top$
vvrisk of company p in sector a and year t	$x^a_{vv,p,t}$
vvrisk values of companies in sector a and year t	$\boldsymbol{x}^a_{vv,t} = (x^a_{vv,1,t}, \ldots, x^a_{vv,n_a,t})^\top$
ESG of company p in sector a and year t	$x^a_{ESG,p,t}$
ESG of companies in sector a and year t	$\boldsymbol{x}^a_{ESG,t} = (x^a_{ESG,1,t}, \ldots, x^a_{ESG,n_a,t})^\top$
ESGM of company p in sector a and year t	$x^a_{ESGM,p,t}$
ESGM of companies in sector a and year t	$\boldsymbol{x}^a_{ESGM,t} = (x^a_{ESGM,1,t}, \ldots, x^a_{ESGM,n_a,t})^\top$
E pillar of company p in sector a and year t	$x^a_{E,p,t}$
E pillar of companies in sector a and year t	$\boldsymbol{x}^a_{E,t} = (x^a_{E,1,t}, \ldots, x^a_{E,n_a,t})^\top$
S pillar of company p in sector a and year t	$x^a_{S,p,t}$
S pillar of companies in sector a and year t	$\boldsymbol{x}^a_{S,t} = (x^a_{S,1,t}, \ldots, x^a_{S,n_a,t})^\top$
G pillar of company p in sector a and year t	$x^a_{G,p,t}$
G pillar of companies in sector a and year t	$\boldsymbol{x}^a_{G,t} = (x^a_{G,1,t}, \ldots, x^a_{G,n_a,t})^\top$
M pillar of company p in sector a and year t	$x^a_{M,p,t}$
M pillar of companies in sector a and year t	$\boldsymbol{x}^a_{M,t} = (x^a_{M,1,t}, \ldots, x^a_{M,n_a,t})^\top$
Resource use of company p in sector a and year t	$x^a_{RU,p,t}$
Emissions of company p in sector a and year t	$x^a_{EM,p,t}$
Environmental innovation of company p in sector a and year t	$x^a_{EI,p,t}$
Workforce of company p in sector a and year t	$x^a_{WF,p,t}$
Human rights of company p in sector a and year t	$x^a_{HR,p,t}$
Community of company p in sector a and year t	$x^a_{CO,p,t}$
Product Responsibility of company p in sector a and year t	$x^a_{PR,p,t}$
Management of company p in sector a and year t	$x^a_{MG,p,t}$
Shareholders of company p in sector a and year t	$x^a_{SH,p,t}$
CSR strategy of company p in sector a and year t	$x^a_{CS,p,t}$
Total number of zero values in ESG categories of company p in sector a and year t	$x^a_{zero,p,t}$
Set of ESG category indices	$S_{CAT} = \{RU, EM, EI, WF, HR, CO, PR, MG, SH, CS\}$
Set of total number of zero values in ESG categories of companies in sector a and year t	$S^a_{zero,t} = \{x^a_{zero,1,t}, \ldots, x^a_{zero,n_a,t}\}$

Total number of zero values in ESG categories of company p in sector a and year t ($x^a_{zero,p,t}$):

$$x^a_{zero,p,t} = \sum_{\substack{j' \in S_{CAT} \\ j' : x^a_{j',p,t} = 0}} 1, \quad \forall a, p, t. \tag{5.8}$$

Total number of companies that company p has the same total number of zero values in ESG categories in sector a and year t ($e^a_{p,t}$):

$$e^a_{p,t} = \sum_{\substack{j' \in [1, n_a] \\ j' : x^a_{zero,j',t} = x^a_{zero,p,t}}} 1, \quad \forall a, p, t. \tag{5.9}$$

Total number of companies that company p has a higher total number of zero values in ESG categories in sector a and year t ($l^a_{p,t}$):

$$l^a_{p,t} = \sum_{\substack{j' \in [1, n_a] \\ j' : x^a_{zero,j',t} < x^a_{zero,p,t}}} 1, \quad \forall a, p, t. \tag{5.10}$$

Number of re-weighting schemes for pillar subscores

Assume that the maximum increment a pillar subscore weight can have in $(0, 1)$ is k such that $mod(\frac{1}{k}, 1) = 0$. We know that a pillar subscore weight is in $[0, 1]$. Then a pillar subscore weight can have $1/k + 1$ different values. Moreover, there are three pillar subscore weights, and they have to sum up to 1.

Without loss of generality, suppose that the E pillar subscore weight is 0. Then the S pillar subscore weight can have $1/k + 1$ different values so that the G pillar subscore weight is 1 - (the S pillar subscore weight). As a result, there are $1/k + 1$ different weighting schemes. Now, assume that the E pillar subscore weight increases by a value of k. Then the S pillar subscore weight can have $1/k$ different values so that the G pillar subscore weight is 1 - (the E pillar subscore weight + the S subscore score weight). Hence, there are $1/k$ different weighting schemes. In a similar fashion, it can be shown that the number of re-weighting schemes for pillar subscores is given by

$$(1/k + 1) + (1/k) + \ldots + 1 = \frac{(1/k + 1) \cdot (1/k + 2)}{2}.$$

Since it holds $k = 0.001$ in Refinitiv's methodology[1], there are 501501 different re-weighting schemes for pillar subscores.

Proof of the M pillar subscore bounds and average

Let $S = \{x_1, \ldots, x_N\}$ be a set with N elements such that $x_1 \leq \ldots \leq x_N$. For the dth element, x_d, assume the number of elements whose value is smaller/larger than x_d in S are denoted by n_s^d/n_l^d. Also, n_e^d corresponds to the number of elements x_d has the same value in S (including itself). It holds that $n_s^d + n_l^d + n_e^d = N$ for $d = 1, \ldots N$. Then its M pillar subscore is given by:

$$x_M^d = 100 \cdot \frac{n_s^d + \frac{n_e^d}{2}}{N} \quad \text{for} \quad d = 1, \ldots, N.$$

Lower bound on M pillar subscore: Since it holds $n_s^d \geq 0, n_e^d \geq 0, N \geq 0$, we have $x_M^d \geq 0$ for $\forall_d$. Additionally,

$$\lim_{N \to \infty} 100 \cdot \frac{0 + 1/2}{N} = 0.$$

Upper bound on M pillar subscore: Since it holds $x_M^d = 100 \cdot \frac{n_s^d + \frac{n_e^d}{2}}{N} \leq 100 \cdot \frac{n_s^d + \frac{n_e^d}{2}}{n_s^d + n_e^d} \leq 100 \cdot \frac{n_s^d + n_e^d}{n_s^d + n_e^d}$, we have $x_M^d \leq 100, \forall_d$. Additionally,

$$\lim_{N \to \infty} 100 \cdot \frac{(N-1) + 1/2}{N} = 100.$$

Average value of M pillar subscore: Denoting the average value of the M pillar subscore for the elements in S by $\bar{x}_M$, we can write:

$$\bar{x}_M = 100 \cdot \frac{(n_s^1 + \ldots + n_s^N) + \frac{(n_e^1 + \ldots + n_e^N)}{2}}{N \cdot N}.$$

In the first scenario, assume $x_1 < \ldots < x_N$. Then it holds

$$\bar{x}_M = 100 \cdot \frac{(0 + \ldots + N - 1) + \frac{(1 + \ldots + 1)}{2}}{N \cdot N} = 100 \cdot \frac{\frac{N \cdot (N-1)}{2} + \frac{N}{2}}{N \cdot N} = 50.$$

In the second scenario, assume $x_1 = \ldots = x_j < x_{j+1} < \ldots < x_N$. Then we have

$$\begin{aligned} \bar{x}_M &= 100 \cdot \frac{(0 + \ldots + 0 + J + (J+1) + \ldots + N - 1) + \frac{(J + \ldots + J + 1 + \ldots + 1)}{2}}{N \cdot N} \\ &= 100 \cdot \frac{\frac{(N-1+J) \cdot (N-J)}{2} + \frac{J \cdot J + (N-J)}{2}}{N \cdot N} = 50. \end{aligned}$$

The second scenario can be easily adopted for the equal elements, which exist more than once in the set, and it can be proven that the average M pillar subscore is 50.

Chapter 6

Genomic prediction

Main materials in this chapter are based on Sahin and Czado (2022a), but we extend real data analysis in Section 6.4 and add other discussion points in Section 6.5.

6.1 Motivation

Genomic prediction (GP) aims at predicting a breeding value using genotypic measurements. Then an unobserved trait can be predicted using its genotype information like single-nucleotide polymorphism (SNP). With rapid developments in genomic technologies, researchers have high-dimensional SNP data sets available. However, it poses some challenges in prediction modeling, such as a small number of observations and a large number of explanatory variables, skewness in variables, irrelevant and redundant variables, interactions among variables, and nonconstant error variance.

To solve the drawbacks regarding the data dimensionality in GP, statistical or machine learning based approaches have been applied (Li et al. 2018). Recently, quantile regression approaches, which model the conditional distribution of the response, have been utilized to deal with the skewness and outliers in the data (Montesinos-López et al. 2019; Pérez-Rodríguez et al. 2020). Still, the question has been *how to model conditional quantiles flexibly while handling data dimensionality in GP*.

It is also important for researchers to *identify the SNPs relevant* for predicting breeding values to design future genotype studies. For instance, since the human population has been growing, the stability of food supplies has gained much more importance. Hence, recent plant breeding efforts have been given to the genetic improvement of crops. The key task is to predict a breeding trait well so crops can be generated based on future requirements. Hölker et al. (2019) provided agronomic measurements and more than 500 thousand SNPs to make European flint maize landraces available for such a task.

In this chapter, we apply our high-dimensional sparse vine copula regression methods introduced in Chapter 4 for genomic prediction of maize traits, proposing data preprocessing

and feature extraction steps on the data given by Hölker et al. (2019). Such steps can be further applied and improved in future studies. To the best of our knowledge, there has not yet been any study performing genomic prediction using vine copula models. Further, the real data analysis supports the advantage of vine copula based prediction methods over linear and forest-based models in the presence of nonlinear dependence and dependent features.

The remainder of the chapter is organized as follows. Section 6.2 describes the data and data preprocessing steps. Section 6.3 explains the feature extraction steps. The results using sparse vine copula regression methods on the data are given in Section 6.4. Section 6.5 discusses our findings, and Section 6.6 concludes the chapter.

6.2 Data description and preprocessing

We describe a real-data application on the doubled-haploid (DH) lines from European flint maize landraces that motivates our sparse vine copula regression methods' usage given in Chapter 4. Hölker et al. (2019) evaluated 899 DH lines whose data contains genotypic measurements with the SNP array technology and phenotypic measurements of agronomic traits across environments. We can regard a DH line as an observation, a SNP as an explanatory variable, and a trait as a response in regression.

We are interested in the relationship between a DH line's genotype encoded by its SNPs and its phenotypic outcome described by its traits, i.e., the genomic prediction of maize traits. More specifically, we would like to find relevant SNPs for a trait in a multivariate prediction model using our vine copula regression methods, performing the variable selection.

There are three landraces in the data, and we focus on the Kemater Landmais Gelb (KE) landrace, which has the largest number of observations (471 out of 899). There are 501,124 explanatory variables, SNPs, which contain only zero and two values, e.g., zero corresponds to the genotype TT, and two denotes the genotype CC.

In addition, we predict four responses of agronomic traits separately: early plant height measured by centimetres at the fourth and sixth stages (PH_V4/V6), female flowering time (FF), and male flowering time (MF) measured by days. They are quantitative and continuous as shown in Figure 6.2. Table 6.1 gives the descriptive summary statistics of the four traits, and Figure 6.1 shows the traits' histogram. We see in Table 6.1 that PH_V4 trait has a minimum value of 19.39 and a maximum value of 56.28, indicating a relatively wide range of values. Moreover, PH_V6 has a larger range, with a minimum of 33.95 and a maximum of 107.32. Its mean is slightly lower than its median, indicating a slightly left-skewed distribution as also seen in Figure 6.1. Further, PH_V6 stands out with the highest standard deviation, showing greater dispersion. FF has the smallest range among the four traits, and its mean and median are close, showing a roughly symmetrical distribution. Moreover, MF has a similar range to the FF.

Plant breeders need to increase early plant development and adopt maize genotypes in future studies. Also, plant breeders may avoid decreasing or increasing female and male

Table 6.1: Summary statistics of the four traits.

Trait	Minimum	First quartile	Median	Mean	Third quartile	Maximum	Standard deviation
PH_V4	19.39	36.91	40.42	39.93	43.46	56.28	5.76
PH_V6	33.95	72.16	78.49	77.32	83.69	107.32	10.02
FF	69.40	76.94	79.71	79.68	82.38	95.37	4.23
MF	66.15	73.80	76.30	76.56	79.19	97.23	4.14

flowering times during adoption. Thus, the traits' prediction from the genotypic measurements is crucial.

To compare the performance of regression methods, we partition our data sets randomly into training (67%) and test (33%) sets. As a result, the training and test sets contain 314 and 157 observations, respectively.

Moreover, we remove the duplicate explanatory variables, retaining only one. Next, we remove the explanatory variables with common values among the observations. We use the threshold of 5%, also known as minor allele frequency (Rédei 2008), to identify such variables. For instance, assume an explanatory variable in our training set contains 300 zero values and 14 two values. Then such a variable does not differ among the observations and might not be expected to have predictive power on a response. As a result, the number of explanatory variables in the training and test sets decreases from 501124 to 44789, i.e., we retain around 9% of the initial explanatory variables.

Hence, the number of observations (DH lines) in the training sets is 314, whereas 147 for their test sets. The number of explanatory variables (SNPs) is 44789 ($p = 1, \ldots, 44789$), and there are four univariate responses (traits) ($k = 1, \ldots, 4$).

6.3 Feature extraction

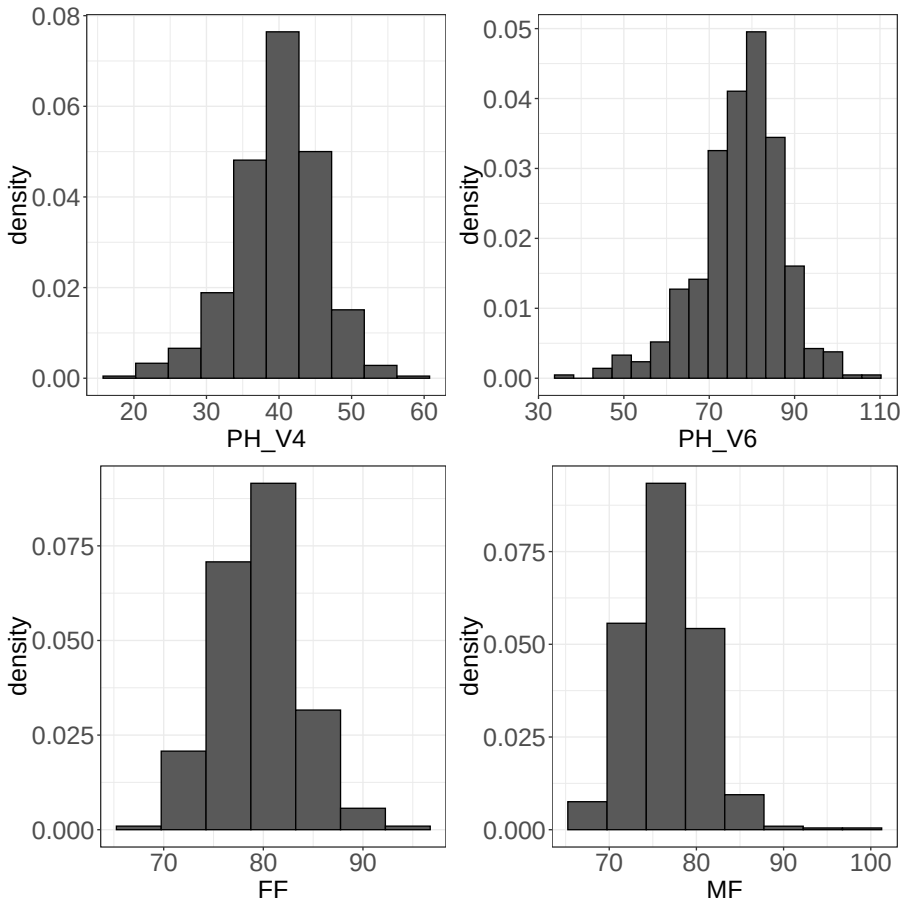
Since our explanatory variables are binary, and there can be associated latent variables with a prediction power on the response, we focus on estimating these latent variables and using them as extracted features (new explanatory variables) in a regression model. Our approach is to group the explanatory variables and estimate their weights in their groups so that such weights are used to estimate the latent variables representing each group.

Let $\boldsymbol{y}_k$ and $\boldsymbol{SNP}_p$ denote the response vector for trait k and explanatory variable vector p, respectively.

1. Fit a linear regression between a response and an explanatory variable, SNP:

$$\boldsymbol{y}_k = \hat{\beta}_0^{\,p,k} + \hat{\beta}_1^{\,p,k} \cdot \boldsymbol{SNP}_p, \quad for \quad k = 1, \ldots, 4, \quad p = 1, \ldots, 44789.$$

Figure 6.1: Histogram of the four traits.



2. Perform a two-tailed Wald test for $H_0 : \beta_1^{p,k} = 0$ versus $H_1 : \beta_1^{p,k} \neq 0$ and determine the associated P-values $P^{p,k}$ for $k = 1, \ldots, 4, \quad p = 1, \ldots, 44789$.

3. Screen the explanatory variables whose P-value from the second step are smaller than 0.10 and have the screened set S_k:

   $S_k = \{\boldsymbol{SNP_p} : P^{p,k} < 0.10 | p = 1, \ldots, 44789\} \quad \text{for} \quad k = 1, \ldots, 4.$

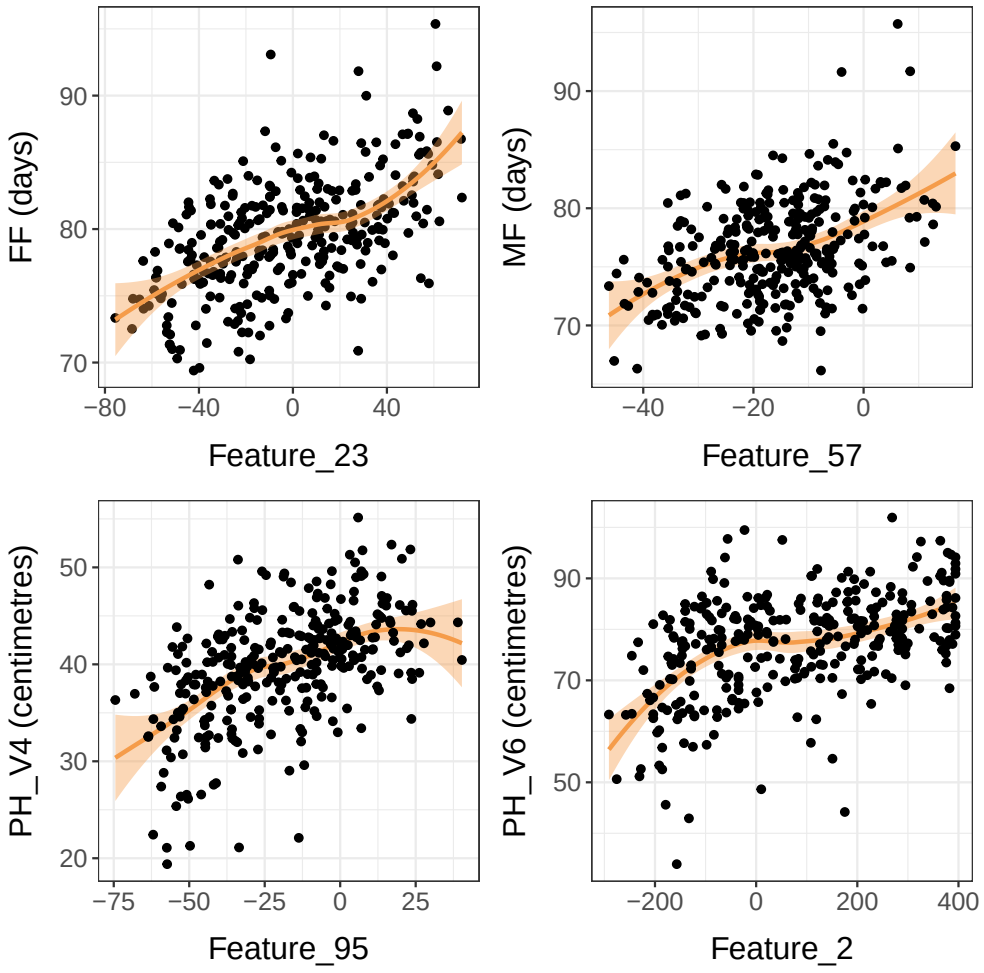
4. Order the set of the explanatory variables S_k based on their P-value non-decreasingly: $O_k = \{\boldsymbol{SNP_{w_1}}, \ldots, \boldsymbol{SNP_{w_t}}\}$, where $P^{w_1,k} \leq \ldots \leq P^{w_t,k}$ for $O_k = S_k$, $k = 1, \ldots, 4$.

5. Estimate the latent variables, i.e., create the features $\boldsymbol{feature}_{k,d_k}^G$ by using a grouping size G of explanatory variables in O_k and using their coefficients from the first step:

   $\boldsymbol{feature}_{k,d_k}^G = \hat{\beta}_1^{w_{d_k},k} \cdot \boldsymbol{SNP_{w_{d_k}}} + \ldots + \hat{\beta}_1^{w_{d_k+G-1},k} \cdot \boldsymbol{SNP_{w_{d_k+G-1}}} \quad \text{for} \quad G \in \{100, 200\}, \quad n_{k_G} = \left\lceil \frac{|O_k|}{G} \right\rceil, \quad d_k = 1, \ldots, n_{k_G}, \quad k = 1, \ldots, 4.$

At the end of the third step above, we obtain the screened SNP sets with the length of 17363, 9163, 19727, and 18284 for FF, MF, PH_V4, and PH_V6, respectively. As a result, we have 174 (87) features for FF, 92 (46) features for MF, 198 (99) features for PH_V4, and 183 (93) features for PH_V6 for grouping size $G = 100$ ($G = 200$). Figure 6.2 shows a scatter plot of a selected continuous feature and a trait, where relationships between a trait and different features are observable.

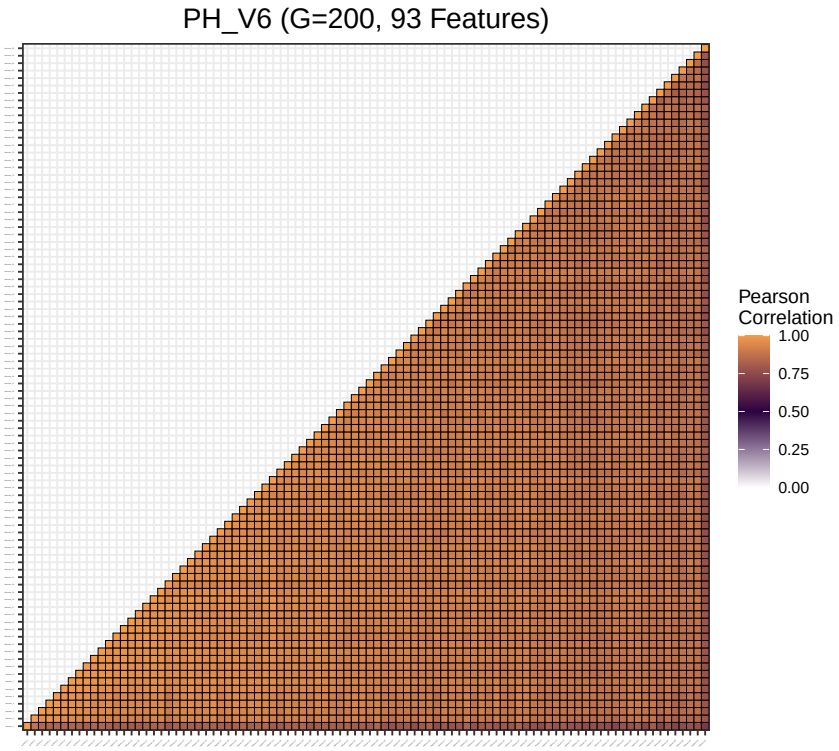
Figure 6.2: Scatter plots of an extracted feature, combining 100 SNPs in a feature, and a trait. Feature_2 corresponds to the combination of the SNPs whose p-value from the OLS of the associated trait is higher than 100 SNPs but lower than others. Similar correspondence applies to other features. Orange curves demonstrate a local polynomial regression fit.



The correlation matrix of features using a grouping size $G = 200$ in Figure 6.3 shows that many correlated features exist for PH_V6. We remark that we do not observe any negative correlation between the four traits' features as shown in Figure 6.3 in the supplementary in

Section 6.7. Further, the minimum correlation between features is 0.58 (0.67) for FF, 0.11 (0.30) for MF, 0.45 (0.77) for PH_V4, and 0.55 (0.69) for PH_V6, whereas its maximum is 0.98 (0.97) for FF, 0.91 (0.93) for MF, 0.97 (0.99) for PH_V4, and 0.98 (0.99) for PH_V6.

Figure 6.3: Features' Pearson correlation map using a grouping size of $G = 200$ for PH_V6, where orange, purple, and white denote the strongest positive, medium, and zero dependence, respectively.



6.4 Application: sparse vine copula regression

We have our data $D_k^G = (\boldsymbol{y}_k, \boldsymbol{feature}_{k,1}^G, \ldots, \boldsymbol{feature}_{k,n_{k_G}}^G)$ for each response $k = 1, \ldots, 4$ and $G \in \{100, 200\}$. To identify if a feature is relevant, redundant, or irrelevant, we first conduct a bivariate analysis by fitting a vine copula regression on each feature and trait, i.e., D-vines with two nodes: response and one feature. If a feature is relevant or redundant given the others, our methods add it to the model; otherwise, it is not selected as explained in Section 4.5. For instance, we conduct 174 (87) bivariate analyses for FF using a grouping size of $G = 100$ (200). As a result, all features for the four responses are classified as relevant or redundant using a grouping size of $G = 100$ and $G = 200$, where the relevant features are

listed in Table 6.6 in the supplementary in Section 6.7.

Next, we explain how to find the most relevant features, thereby redundant ones given them: we apply our methods to eight different data sets' training sets. Moreover, we compare them with linear quantile regression penalized with a LASSO function ($LQRLasso$) and quantile regression forests (QRF), which are described in Section 4.7. For $LQRLasso$, we provide more details on its fit in the supplementary in Section 6.7. We evaluate the models' predictive performance on test sets using the pinball loss defined in Equation (4.7) in Section 4.7 at the quantile levels $0.05, 0.50$, and 0.95.

Table 6.2 shows that vine copula based methods perform worse than $LQRLasso$ and QRF for MF. We observe that dependencies among MF and its selected features by $vineregRes$ are more linear than those among other traits since it fits mostly the Gaussian copula in the first tree of the D-vine for MF as seen in Table 6.4 in the supplementary in Section 6.7. We remark that linear quantile regression by $LQRLasso$ may perform well if it can avoid crossing quantile curves, but there is no guarantee that the 95% quantile curve exceeds the 90% quantile curve everywhere as shown in Figure 6.5 in the supplementary in Section 6.7. Further, whenever $LQRLasso$ is more accurate than $vineregRes$ for PH_V4, it includes more features, giving a trade-off between model sparsity and accuracy. Even though QRF provides the lowest pinball loss at all quantiles for FF for $G = 200$, $vineregRes$ has better performance than it for $G = 100$, except the upper quantile. Moreover, the method $vineregRes$ is the most sparse and accurate model at all quantiles for PH_V6 using $G = 200$. It chooses two features for $G = 200$, identifying more than 95% of the features as redundant. It has the best accuracy for PH_V6 for four cases out of six, with three quantiles evaluated for two G values.

We provide the selected feature indices in Tables 6.6, 6.7, and 6.8 in the supplementary in Section 6.7. Given them, the other features are redundant for a given trait and a grouping of G using $vineregRes$. For instance, given the first and 88th features for PH_V6 using a grouping size of $G = 200$, the remaining 91 features are redundant using $vineregRes$. Since the features of PH_V6 are highly correlated but are not needed in a model, in parallel to the simulation study results in Section 4.7, the reason for our methods' better accuracy than QRF may be many correlated but redundant features for PH_V6. In addition, Table 6.7 in the supplementary in Section 6.7 shows that the first feature is selected by all methods considered for the prediction of PH_V6. Moreover, Figures 6.6 and 6.7 in the supplementary in Section 6.7 show that QRF and LQR tend to do worse at the extremes than $vineregRes$.

Our SNP screening and feature extraction steps are similar to Qian et al. (2020). However, they fit a simple linear regression on the first feature, which includes linearly and marginally the most important SNPs. Thus, we compare our models' performance for the pinball loss at levels $0.05, 0.50$, and 0.95 by fitting a vine copula regression only on the first feature for PH_V4 and PH_V6. However, any variable selection is not allowed. The bivariate copula family selection between the response and the first feature is conducted as explained in Section 4.3. We call such a bivariate copula based regression with the first feature $bicopreg$. Then $bicopreg$ using a grouping size of $G = 200$ for PH_V6 has a pinball loss of 0.97, 3.18, and 0.94 at

Table 6.2: Comparison of the methods' performance on the test set for the pinball loss (PL_α) and on the training set for the number of selected features (No. Ftr.), where (a,b,c) under the $LQRLasso$ column corresponds to the quantile levels (0.05,0.50,0.95). The best performance on the test set for each quantile level α, trait, and G is highlighted.

		$vregRes$	$vregParCor$	$LQRLasso$	QRF	$vregRes$	$vregParCor$	$LQRLasso$	QRF
Trait	Measure	$G = 100$				$G = 200$			
FF	$PL_{0.05}$	0.35	0.49	0.40	0.38	0.39	0.39	0.39	0.37
	$PL_{0.50}$	1.43	1.51	1.50	1.48	1.48	1.56	1.47	1.45
	$PL_{0.95}$	0.47	0.47	0.41	0.38	0.41	0.43	0.39	0.39
	No. Ftr.	11	22	(8,41,4)	174	4	14	(8,29,5)	87
MF	$PL_{0.05}$	0.35	0.36	0.34	0.33	0.35	0.36	0.32	0.34
	$PL_{0.50}$	1.41	1.42	1.39	1.36	1.39	1.40	1.36	1.37
	$PL_{0.95}$	0.45	0.47	0.41	0.39	0.44	0.45	0.40	0.39
	No. Ftr.	12	16	(7,45,8)	92	8	13	(5,15,12)	46
PH_V4	$PL_{0.05}$	0.51	0.51	0.55	0.55	0.51	0.55	0.56	0.55
	$PL_{0.50}$	1.93	1.87	1.92	1.94	1.96	1.99	1.92	1.94
	$PL_{0.95}$	0.56	0.58	0.55	0.60	0.57	0.57	0.55	0.62
	No. Ftr.	6	11	(9,15,8)	198	3	11	(7,17,4)	99
PH_V6	$PL_{0.05}$	1.01	1.01	0.98	1.00	0.96	0.98	1.05	1.00
	$PL_{0.50}$	3.09	3.10	3.04	3.27	3.06	3.47	3.14	3.31
	$PL_{0.95}$	0.91	0.89	0.92	0.97	0.90	1.05	0.94	1.04
	No. Ftr.	4	12	(8,49,5)	183	2	12	(6,29,6)	93

$0.05, 0.50$, and 0.95, respectively. Thus, $vineregRes$, which selects the first and 88th features as relevant for the prediction of PH_V6, performs better than $bicopreg$ as seen in Table 6.9 in the supplementary in Section 6.7. However, having more features in $vineregParCor$, including the first one, worsens the performance compared to $bicopreg$ using a grouping size of $G = 200$. On the other hand, $vineregParCor$ has better accuracy than $bicopreg$ using a grouping size of $G = 100$ for PH_V6 as shown in Table 6.9 in the supplementary in Section 6.7. Likewise, the mean of selected feature indices using a grouping size of $G = 100$ for PH_V4 is 102.33 for $vineregRes$ and 106.18 for $vineregParCor$. Further, both methods do not select the first feature as relevant but redundant for PH_V4. Nevertheless, their prediction power is stronger than $bicopreg$, except at the level 0.50 using $G = 200$. Hence, linearly and marginally most important SNP groups might not be considered the most relevant for prediction when allowing nonlinear dependencies as in our methods.

6.5 Discussion

High-dimensional sparse vine copula regression is a significant tool for efficiently allowing non-linear relationships between explanatory variables and response and selecting relevant variables. In genomic prediction, genotypic measurements like SNPs are often very high-dimensional, which might be reduced by considering some SNP groups and their interactions. Also, many groups may be irrelevant for prediction. Our methods can handle such situations and predict

responses at different quantile levels. Their performance might be improved with bivariate copula families allowing for more asymmetries, e.g., more than two parameters.

For our application, consider the following question: Which SNPs impact the low and high quantiles of the trait PH_V6? Our method $vineregRes$ identifies two SNP groups (features) that consist of 400 SNPs in total. In the first feature, the corresponding SNPs' p-values out of the linear regression with the trait PH_V6 change from 10^{-12} to 10^{-7}, whereas its range is $[0.087, 0.090]$ in the 88th feature. Thus, the marginal impacts of the selected SNPs differ. Moreover, given these 400 SNPs, others are redundant to predict the trait PH_V6. In addition, 107 SNPs in the first feature are located in the first chromosome, while the highest number of SNPs per chromosome in the 88th feature is 33 for the ninth chromosome. Thus, plant breeders can assess the selected SNP groups' impact on the trait's various quantile levels and identify the associated SNPs using our methods. Moreover, chosen SNP groups can be compared to other genome-wide association studies, helping plant breeders to decide on future genotype adoption. In our application, comparing the identified SNPs with those in Mayer et al. (2020) is high on the agenda.

Feature extraction is a vital step that may impact our methods' genomic prediction power. For instance, the choice of SNPs' weights estimating their latent variable is open to future research. Also, even though it offers a trade-off between a computational burden and prediction power, one can apply cross-validation for the choice of the SNP's group size G. In addition, some SNPs might affect the trait, not marginally only in the presence of certain other SNPs. Alternatively, one may remove the P-value screening of the SNPs at the 10% level described in Section 6.3 and consider all possible extracted features. Likewise, some SNPs might influence the trait marginally, but not when certain other SNPs are in the model. For such cases, some post-processing steps for feature extraction might be applied.

In addition, other feature extraction techniques using SNP information include the principal component analysis of the genotype data (Crossa et al. 2017). The specified number of the most important components are regarded as features and evaluated in a prediction model. However, the approach does not consider the interaction between the extracted features and the response. Thus, it might not predict the response well. Therefore, different feature extraction methods' prediction power can be compared using different models.

Another feature extraction approach considers finding the features which are not correlated to each other but impact the response since most genomic prediction models suffer from the correlation among features (Cuevas et al. 2014). However, vine copula based prediction approaches deal well with them and provide a good alternative to the existing methods.

Finally, our feature extraction steps given in Section 6.3 can be extended to chromosomes. For instance, there are ten maize chromosomes. Therefore, considering the SNPs' chromosome information would allow having ten different SNP sets corresponding to each chromosome in Step 4 of Section 6.3. Accordingly, one can create the features based on the sets, allowing more interpretation of the selected features regarding the chromosome impact on the trait.

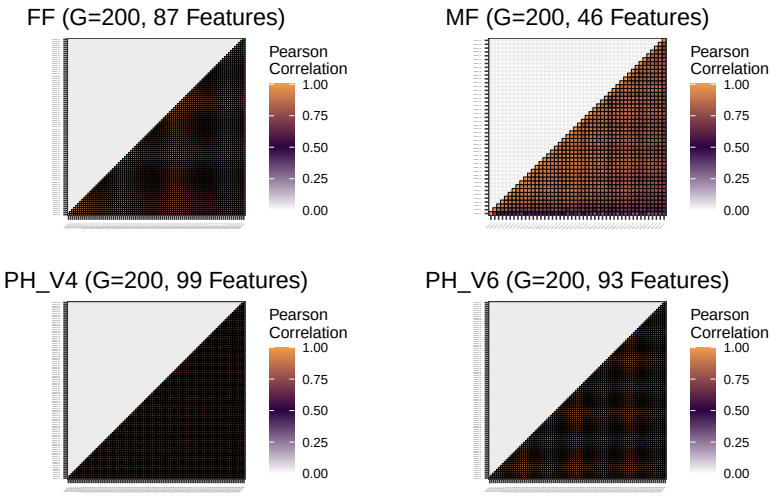
6.6 Conclusion

We apply our high-dimensional sparse vine copula based regression approaches to the genomic prediction of maize traits, providing guidelines on real-data preprocessing and feature extraction. In addition, we discuss future research directions regarding the further application of vine copulas in genomic prediction.

6.7 Supplementary materials

Pearson correlation map of features

Figure 6.4: Features' Pearson correlation map using a grouping size of $G = 200$ for a trait, where orange, purple, and white denote the strongest positive, medium, and zero dependence, respectively.



Estimated parameters by $LQRLasso$

Table 6.3: Estimated the LASSO penalty parameters λ used in Section 6.4.

Trait	Quantile	$G = 100$	$G = 200$	Trait	Quantile	$G = 100$	$G = 200$
	0.05	0.017	0.007		0.05	0.020	0.018
FF	0.50	0.014	0.010	PH_V4	0.50	0.027	0.014
	0.95	0.024	0.009		0.95	0.013	0.009
	0.05	0.029	0.007		0.05	0.015	0.010
MF	0.50	0.009	0.011	PH_V6	0.50	0.011	0.009
	0.95	0.032	0.029		0.95	0.013	0.013

Fitted D-vines' first tree level by $vineregRes$

Table 6.4: Pair-copulas with the estimated Kendall's τ in the fitted D-vine's first tree for MF (left) and FF (right). The numbers under the edge denote the selected feature indices.

G	Edge	Family	τ
100	(MF, 7)	Gaussian	0.41
	(7, 80)	Gaussian	0.45
	(80, 82)	Gaussian	0.45
	(82, 2)	BB7	0.33
	(2, 77)	BB7	0.30
	(77, 71)	Gumbel	0.48
	(71, 66)	BB1	0.46
	(66, 70)	Gaussian	0.46
	(70, 63)	BB1	0.46
	(63, 33)	BB8	0.42
	(33, 69)	Gaussian	0.51
	(69, 57)	Gaussian	0.51
200	(MF, 4)	Gaussian	0.41
	(4, 35)	Gaussian	0.53
	(35, 1)	BB7	0.30
	(1, 29)	BB7	0.31
	(29, 46)	Gaussian	0.59
	(46, 32)	BB7	0.49
	(32, 9)	Gumbel	0.51
	(9, 40)	Gaussian	0.54

G	Edge	Family	τ
100	(FF, 34)	Gaussian	0.38
	(34, 164)	Frank	0.51
	(164, 158)	Frank	0.46
	(158, 137)	BB8	0.46
	(137, 18)	BB8	0.58
	(18, 20)	Gaussian	0.73
	(20, 144)	Frank	0.49
	(144, 23)	Frank	0.50
	(23, 155)	BB8	0.55
	(155, 168)	Frank	0.54
	(168, 51)	Frank	0.48
200	(FF, 19)	Gaussian	0.38
	(19, 68)	BB8	0.58
	(68, 69)	BB8	0.60
	(69, 9)	Gaussian	0.59

Table 6.5: Pair-copulas with the estimated Kendall's τ in the fitted D-vine's first tree for PH_V4 (left) and PH_V6 (right). The numbers under the edge denote the selected features.

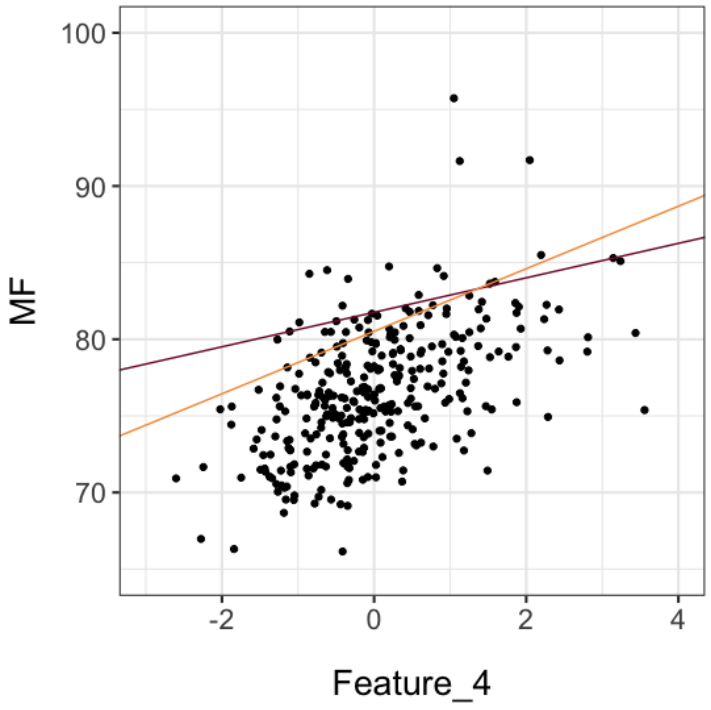
G	Edge	Family	τ
100	(PH_V4, 44)	Gaussian	0.39
	(44, 150)	Frank	0.57
	(150, 162)	BB8	0.47
	(162, 160)	BB8	0.45
	(160, 3)	BB8	0.57
	(3, 95)	Frank	0.61
200	(PH_V4, 2)	Gaussian	0.38
	(2, 69)	Frank	0.66
	(69, 49)	Frank	0.71

G	Edge	Family	τ
100	(PH_V6, 1)	Gumbel	0.34
	(1, 175)	BB8	0.40
	(175, 110)	BB8	0.53
	(110, 169)	BB8	0.58
200	(PH_V6, 1)	BB7	0.34
	(1, 88)	Frank	0.53

Quantile crossing of $LQRLasso$

$LQRLasso$ performs variable selection and linear quantile regression at a specified quantile level at the same time. However, it might result in the crossing of quantile curves. For an illustration, we choose the first and fourth features obtained using a grouping size of $G = 100$ as explanatory variables for predicting the trait MF at levels 90% and 95% and let $LQRLasso$ run for both levels. Then only the fourth feature is chosen for the prediction, thereby estimating its coefficient. However, as seen in the following plot, the coefficient estimated by $LQRLasso$ leads to the quantile crossing.

Figure 6.5: Fitted linear quantile regression curves at 90% (orange) and 95% (red) levels for MF versus the selected Feature_4 by $LQRLasso$, where candidate features to select from are the first and fourth features using a grouping size of $G = 100$.



Selected feature indices

We remark that the features are based on the grouping G. For instance, the first feature using a grouping size of $G = 100$ (200) includes 100 (200) SNPs whose p-value in a simple linear regression for a trait is smaller than others. Thus, selected features using a grouping size of $G = 100$ and $G = 200$ cannot be compared. Moreover, QRF uses all features to make predictions.

Table 6.6: The relevant feature indices by $vineregRes$ and $vineregParCor$ for each grouping G and trait. The same feature indices selected by $vineregRes$ and $vineregParCor$ for each grouping G and trait are highlighted.

	$vineregRes$	$vineregParCor$	$vineregRes$	$vineregParCor$
Trait	$G = 100$		$G = 200$	
FF	34, 164, **158**,**137**, 18, 20 144, **23**, **155**, 168, **51**	37, **158**, **137**, 68, 166, 163, 140, **155**, 101, 104, 126, **23**, 47, 80, 41, 42, 52, 55, 49, **51**, 136, 36	**19**, 68, 69, **9**	**19**, 79, 24, **9**, 48, 84, 70, 83, 82, 66, 63, 50, 67, 45
MF	**7, 80, 82, 2**, 77, 71, 66, 70, **63**, 33, 69, **57**	**7, 80, 2**, 45, 27, 24, **57, 82**, 16, 52, 15, 76, 62, **63**, 19, 81	**4, 35, 1, 29**, 46, 32, 9, **40**	**4, 35, 1, 29, 40**, 3, 38, 8, 14, 23, 20, 15, 21
PH_V4	**44**, 150, 162, 160, 3, **95**	**44**, 5, 179, 153, 93, **95** 92, 182, 51, 86, 88	**2**, 69, 49	**2**, 77, 99, 26, 59, 73, 4, 72, 20, 29, 83
PH_V6	**1**, 175, 110, 169	**1**, 26, 2, 119, 3, 116, 86, 163, 72, 101, 130, 98	**1, 88**	13, **1**, 82, 49, 66, 56, 53, 30, 46, 9, 68, **88**

Table 6.7: The selected feature indices by $LQRLasso$ for each G, trait, and quantile level α. The intercept is included in all.

Trait	α	$G = 100$	$G = 200$
FF	0.05	76, 81, 83, 140, 157, 18, 163	9,28,40,41,42,69,79
	0.50	3,4,23,37,46,49,52,54,61,63,68, 71,78,80,85,88,99,100,101,104,105,107,108, 126,130,132,134,138,142,144,147,151,155, 158,160,163,164,166,168,171	2,8,9,19,23,24,36,39,40, 44,47,48,50,51,60,61,63,70, 72,74,76,78,79,80,82,83,84,86
	0.95	17, 37, 88	9,19,49,61
MF	0.05	2, 14, 24, 61, 63, 70	1,3,10,16,19,31,32,35,36,38,39
	0.50	1,4,6,13,15,18,19,20,23,24,27,28, 30,33,36,37,38,39,42,45,46,47,49,52, 53,54,56,57,58,59,60,62,63,64, 69,71,75,77,78,80,81,84,85,89	1,2,3,4,7,15,23,25,29,30, 35,38,40,41
	0.95	7, 8, 9, 63, 69, 70, 71	4,5,35,36
PH_V4	0.05	19, 95, 109, 125, 139, 171, 174, 189	35,44,48,79,80,94
	0.50	2,5,44,90,130,139,141,146, 153,158,179,181,196,197	1,2,22,30,44,59,60,65,68,73, 74,77,83,86,90,98
	0.95	51, 90, 95, 130, 139, 185, 191	48,65,86
PH_V6	0.05	1, 76, 93, 106, 127, 130, 175	1,38,43,53,55
	0.50	1,2,10,15,22,25,29,32,39,43,45,53, 59,65,70,72,73,77,86,88,91,95,96, 97,98,99,101,106,110,116,123,125, 128,130,131,132,138,142,156,159, 160,163,165,169,177,178,179,180	1,3,6,13,15,25,30,33,35,38, 48,49,53,54,55,58,59,60,62, 66,69,74,80,83,86,88,89,90
	0.95	1, 86, 135, 138	1,58,63,84,89

Table 6.8: The same selected feature indices by $vineregRes$, $vineregParCor$, and $LQRLasso$ for each G and trait and quantile level α. (-) denotes the empty set.

Trait	$G = 100$	$G = 200$
FF	-	9
MF	63	35
PH_V4	-	-
PH_V6	1	1

Comparison of the vine copula based regression methods

Table 6.9: Comparison of the vine copula based regression methods' performance on the test set of the real data in Section 6.2. The best performance on the test set for each quantile level, trait, and the grouping G is highlighted.

		$vineregRes$	$vineregParCor$	$bicopreg$	$vineregRes$	$vineregParCor$	$bicopreg$
Trait	Measure	$G = 100$			$G = 200$		
PH_V4	$PL_{0.05}$	0.51	0.51	0.53	0.51	0.51	0.52
	$PL_{0.50}$	1.93	1.87	1.92	1.96	1.99	1.93
	$PL_{0.95}$	0.56	0.58	0.58	0.57	0.57	0.58
PH_V6	$PL_{0.05}$	1.01	1.01	1.01	0.96	0.98	0.97
	$PL_{0.50}$	3.09	3.10	3.18	3.06	3.47	3. 18
	$PL_{0.95}$	0.91	0.89	0.93	0.90	1.05	0.94

Predictions by different methods

Figure 6.6: Scatter plots of the trait PH_V6 measured by centimetres and the median predictions for PH_V6 by $LQRLasso$, QRF, and $vineregRes$ in the training set.

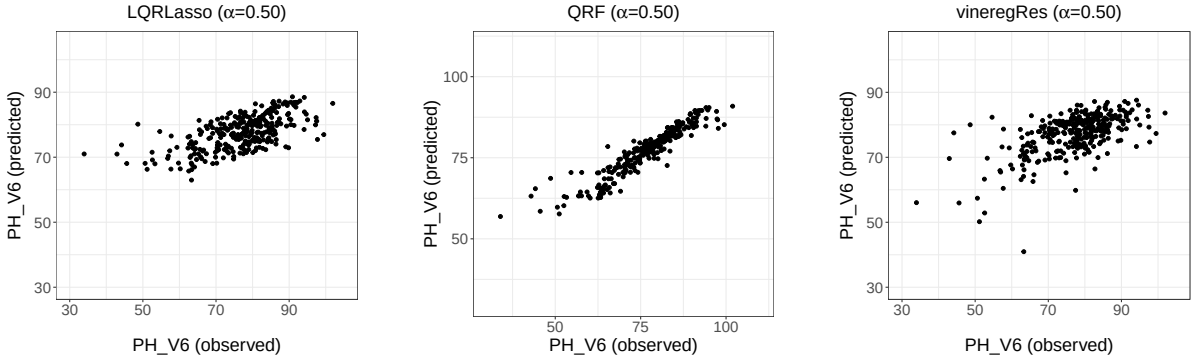
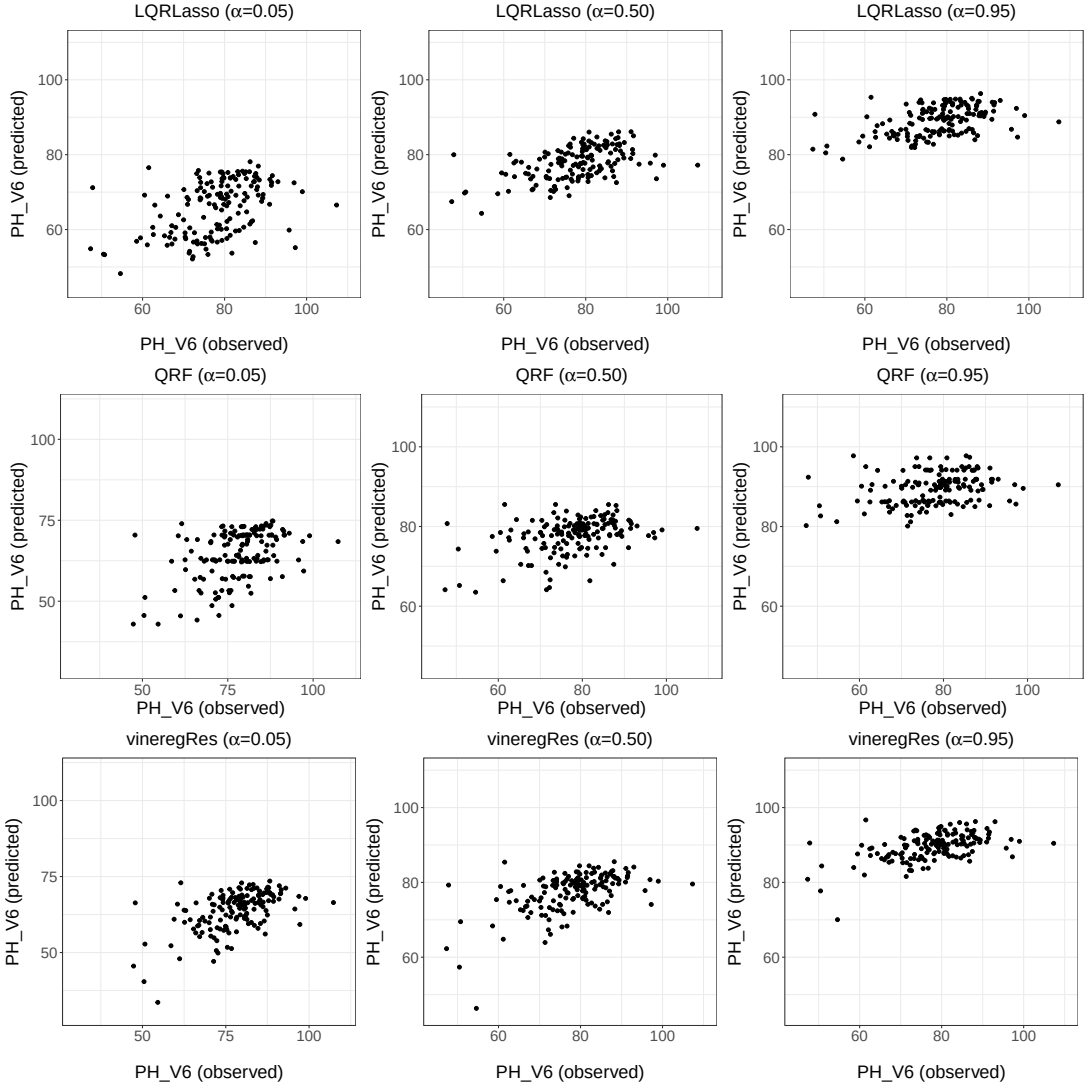


Figure 6.7: Scatter plots of the trait PH_V6 measured by centimetres and the predictions at different quantile levels α for PH_V6 by $LQRLasso$, QRF, and $vineregRes$ in the test set.



Chapter 7

Conclusion

In this thesis, we demonstrate the potential of using vine copulas for statistical learning and the importance of their consideration in data analysis, focusing on financial and genomic applications. Through developing new algorithms and methods, we show that vine copula based statistical learning models are flexible and interpretable. Further, our comprehensive comparison with the existing literature demonstrates the usefulness of vine copulas for clustering and prediction and highlights the importance of proper modeling of nonlinear dependence. We discuss further research directions and provide guidelines for their starting points.

As data volume, including observations and features, increases with the increasing storage power, data observations are likely to be grouped. To find them, Chapter 3 describes a vine copula mixture model (VCMM) framework and shows the advantages of the corresponding model-based clustering algorithm compared to other statistical distributions, especially in the presence of heavy tail and asymmetric tail dependence between a pair of features.

Another crucial data inference method is the prediction of an outcome of interest not only on the mean but also on quantiles. Vine copulas can express the conditional distribution of a response given features analytically under some assumptions and be used for (quantile) prediction. However, given that many features are not needed in a predictive model, we define feature types for quantile regression and introduce efficient feature selection methods for vine copula based regression in Chapter 4. We show their advantage over machine learning and linear models when many features are correlated but are not needed.

Chapter 5 works with one of the crucial data types in finance: Environmental, Social, and Governance (ESG) scores. We present a case study of VCMM using them and show that the driver of the dependence among companies has neither the best nor the worst scores and changes with stock price movements. Going beyond, we analyze data collection problems behind the scores using data-mining schemes and propose approaches to deal with missing data, considering companies' sectoral peculiarities.

In Chapter 6, we show how selecting features in vine copula based regression can be useful for genomic prediction using real data on maize genotypes and phenotypes. We also propose genomic feature extraction techniques that consider the relationship between the phenotype

and the genotype information.

As an outlook, we can incorporate machine learning and copula based models into each other to boost their capabilities. For example, consider the prediction task with many dependent features. A machine learning model, such as random forests, might not make accurate predictions. However, a vine copula based regression with feature selection might be used to eliminate some features to get input data and then make predictions with them for random forests. Alternatively, suppose the dependence among important features needs to be modeled explicitly in very high dimensional data. Since vine copulas might suffer from data dimensionality more than random forests, the latter can be fitted. Vine copulas can then model the most important features for interpretation and actionable insights, such as the likelihood or probability of extreme value/events.

An extension of the statistical learning approaches we proposed in the thesis is to combine them as a cluster-weighted model (Gershenfeld 1997). It assumes that a finite mixture model expresses the joint density of the response and features, and a mixture component density is composed of the conditional density of the response given features and feature densities. It allows modeling non-monotonicity in observations using a vine copula based regression model.

An application of our statistical learning methods is to use them for imputing missing values. For instance, after excluding missing observations, a feature can be predicted using other features as predictors in our vine copula based regression. Then the fit can be used to predict the excluded missing observations. Alternatively, an expectation-minimization type algorithm may be developed for missing value imputation similar to Ding and Song (2016). Moreover, when observations are clustered in some features, fitting a VCMM to real-valued observations and then predicting the missing value of an observation based on the mixture model likelihood can be applied. Yan et al. (2015) give an application with Gaussian mixture models. Using them, an interesting real-life application is to impute missing ESG scores.

Statistical learning algorithms based on dependence modeling tools have further application areas. An important area to apply our models is disease progression and treatment. Diseases have been predicted from genomes or medical features using inference methods (Eskidere et al. 2012). However, exploring disease progression where the dependence among some genome regions or features fires the disease in a vine copula based regression model is a novel idea to investigate. Likewise, the disease treatment's success can be predicted by modeling the dependence among features, such as heart rate and blood pressure.

Object recognition has many real-life applications in computer vision that aid decision-making processes. For example, objects have features like color and texture or are characterized by their image. Dependence among features and pixels can be captured by VCMM for image segmentation, improving object recognition. For instance, an application based on Dirichlet Gaussian mixture models is discussed by Nguyen and Wu (2011).

While new algorithms make data analyses more appropriate and improved, the availability of big data has made statistical software tools much more sophisticated. For example, the `R` software provides many packages for fitting statistical learning algorithms (Koenker

2022; Nagler 2022; Sahin 2021). However, `Python` software is particularly well suited for deploying large-scale algorithms. There has been considerable effort to create open-source `Python` libraries for vine copula models (Nagler and Vatter 2022b), and extending them to the statistical learning algorithms discussed here would contribute to many fields and experts.

Overall, we contribute to the growing literature on vine copulas and statistical learning, providing new insights into using vine copula based statistical learning models for data analysis. We hope our work will inspire further research and develop more advanced methods to model high-dimensional complex data sets.

Bibliography

Aas, K., Czado, C., Frigessi, A., and Bakken, H. (2009). "Pair-copula constructions of multiple dependence". In: *Insurance: Mathematics and Economics* 44.2, pp. 182–198. DOI: 10.1016/j.insmatheco.2007.02.001.

Abhayawansa, S. and Tyagi, S. (2021). "Sustainable investing: the black box of environmental, social, and governance (ESG) ratings". In: *The Journal of Wealth Management* 24.1, pp. 49–54. DOI: 10.3905/jwm.2021.1.130.

Ahmed, M. F., Gao, Y., and Satchell, S. (2021). "Modeling demand for ESG". In: *The European Journal of Finance*, pp. 1–15. DOI: 10.1080/1351847X.2021.1924216.

Akaike, H. (1998). "Information theory and an extension of the maximum likelihood principle". In: *Selected papers of Hirotugu Akaike*. New York, NY: Springer, pp. 199–213. DOI: 10.1007/978-1-4612-1694-0_15.

Alessandrini, F. and Jondeau, E. (2020). "ESG investing: from sin stocks to smart beta". In: *The Journal of Portfolio Management* 46.3, pp. 75–94. DOI: 10.3905/jpm.2020.46.3.075.

Alessandrini, F., Baptista Balula, D., and Jondeau, E. (2021). "ESG screening in the fixed-income universe". In: *Available at SSRN 3966312*. DOI: 10.2139/ssrn.3966312.

Amel-Zadeh, A. and Serafeim, G. (2018). "Why and how investors use ESG information: evidence from a global survey". In: *Financial Analysts Journal* 74.3, pp. 87–103. DOI: 10.2469/faj.v74.n3.2.

Andrews, J. L. and McNicholas, P. D. (2011). "Mixtures of modified t-factor analyzers for model-based clustering, classification, and discriminant analysis". In: *Journal of Statistical Planning and Inference* 141.4, pp. 1479–1486. DOI: 10.1016/j.jspi.2010.10.014.

Aslan, A., Poppe, L., and Posch, P. (2021). "Are sustainable companies more likely to default? Evidence from the dynamics between credit and ESG ratings". In: *Sustainability* 13.15. DOI: 10.3390/su13158568.

Ayton, J., Krasnikova, N., and Malki, I. (2022). "Corporate social performance and financial risk: further empirical evidence using higher frequency data". In: *International Review of Financial Analysis*, p. 102030. DOI: 10.1016/j.irfa.2022.102030.

Azzalini, A. and Capitanio, A. (2003). "Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t-distribution". In: *Journal of the Royal Statistical*

Society: Series B (Statistical Methodology) 65.2, pp. 367–389. DOI: 10.1111/1467-9868.00391.

Bae, K.-H., El Ghoul, S., Gong, Z. J., and Guedhami, O. (2021). "Does CSR matter in times of crisis? Evidence from the COVID-19 pandemic". In: *Journal of Corporate Finance* 67, p. 101876. DOI: 10.1016/j.jcorpfin.2020.101876.

Bannier, C. E., Bofinger, Y., and Rock, B. (2021). "Corporate social responsibility and credit risk". In: *Finance Research Letters*. DOI: 10.1016/j.frl.2021.102052.

Barko, T., Cremers, M., and Renneboog, L. (2021). "Shareholder engagement on environmental, social, and governance Performance". In: *Journal of Business Ethics*. DOI: 10.1007/s10551-021-04850-z.

Bax, K., Sahin, Ö., Czado, C., and Paterlini, S. (2023). "ESG, risk, and (tail) dependence". In: *International Review of Financial Analysis*, p. 102513. DOI: 10.1016/j.irfa.2023.102513.

Bedford, T. and Cooke, R. M. (2001). "Probability density decomposition for conditionally dependent random variables modeled by vines". In: *Annals of Mathematics and Artificial Intelligence* 32, 245–268. DOI: 10.1023/A:1016725902970.

Bedford, T. and Cooke, R. M. (2002). "Vines - A new graphical model for dependent random variables". In: *Annals of Statistics* 30.4, pp. 1031–1068. DOI: 10.1214/aos/1031689016.

Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. (2021). "On the dangers of stochastic parrots: can language models be too big?" In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. New York, NY, USA: Association for Computing Machinery. DOI: 10.1145/3442188.3445922.

Berg, F., Fabisik, K., and Sautner, Z. (2021). "Is history repeating itself? The (un)predictable past of ESG ratings". In: *Available at SSRN 3722087*. DOI: 110.2139/ssrn.3722087.

Berg, F., Koelbel, J. F., and Rigobon, R. (2022). "Aggregate confusion: the divergence of ESG ratings". In: *Review of Finance* 26.6, pp. 1315–1344. DOI: 10.1093/rof/rfac033.

Billio, M., Costola, M., Hristova, I., Latino, C., and Pelizzon, L. (2021). "Inside the ESG ratings: (dis)agreement and performance". In: *Corporate Social Responsibility and Environmental Management* 28.5, pp. 1426–1445. DOI: 10.1002/csr.2177.

Bloomberg (2021). *ESG assets may hit 53 trillion dollars by 2025, a third of global AUM*. https://www.bloomberg.com/professional/blog/esg-assets-may-hit-53-trillion-by-2025-a-third-of-global-aum/. visited on 2021-11-22.

Bouveyron, C. and Brunet-Saumard, C. (2014). "Model-based clustering of high-dimensional data: a review". In: *Computational Statistics and Data Analysis* 71, pp. 52–78. DOI: 10.1016/j.csda.2012.12.008.

Brechmann, E. (2010). "Truncated and simplified regular vines and their applications". MSc Thesis. Technical University of Munich.

Browne, R. P. and McNicholas, P. D. (2015). "A mixture of generalized hyperbolic distributions". In: *Canadian Journal of Statistics* 43.2, pp. 176–198. DOI: 10.1002/cjs.11246.

Cabral, C. R. B., Lachos, V. H., and Prates, M. O. (2012). "Multivariate mixture modeling using skew-normal independent distributions". In: *Comput. Stat. Data Anal.* 56.1, 126–142. DOI: 10.1016/j.csda.2011.06.026.

Cannon, A. J. (2011). "Quantile regression neural networks: implementation in R and application to precipitation downscaling". In: *Computers & geosciences* 37.9, pp. 1277–1284. DOI: 10.1016/j.cageo.2010.07.005.

Cannon, A. J. (2018). "Non-crossing nonlinear regression quantiles by monotone composite quantile regression neural network, with application to rainfall extremes". In: *Stochastic environmental research and risk assessment* 32, pp. 3207–3225. DOI: 10.1007/s00477-018-1573-6.

Celeux, G. and Govaert, G. (1995). "Gaussian parsimonious clustering models". In: *Pattern Recognition* 28.5, pp. 781–793. DOI: 10.1016/0031-3203(94)00125-6.

Chang, B. and Joe, H. (2019). "Prediction based on conditional distributions of vine copulas". In: *Computational Statistics & Data Analysis* 139, pp. 45–63. DOI: 10.1016/j.csda.2019.04.015.

Conn, D., Ngun, T., Li, G., and Ramirez, C. M. (2019). "Fuzzy forests: extending random forest feature selection for correlated, high-dimensional data". In: *Journal of Statistical Software* 91.9, 1–25. DOI: 10.18637/jss.v091.i09.

Crossa, J., Pérez-Rodríguez, P., Cuevas, J., Montesinos-López, O., Jarquín, D., De Los Campos, G., Burgueño, J., González-Camacho, J. M., Pérez-Elizalde, S., Beyene, Y., et al. (2017). "Genomic selection in plant breeding: methods, models, and perspectives". In: *Trends in plant science* 22.11, pp. 961–975. DOI: 10.1016/j.tplants.2017.08.011.

Cuevas, J., Pérez-Elizalde, S., Soberanis, V., Pérez-Rodríguez, P., Gianola, D., and Crossa, J. (2014). "Bayesian genomic-enabled prediction as an inverse problem". In: *G3 Genes—Genomes—Genetics* 4.10, pp. 1991–2001. DOI: 10.1534/g3.114.013094.

Cuvelier, E. and Noirhomme-Fraiture, M. (2005). "Clayton copula and mixture decomposition". In: *ASMDA 2005*, pp. 699–708.

Czado, C. (2019). "Analyzing dependent data with vine copulas: a practical guide with R". In: *Lecture Notes in Statistics*. Cham: Springer. DOI: 10.1007/978-3-030-13785-4_1.

Czado, C., Bax, K., Sahin, Ö., Nagler, T., Min, A., and Paterlini, S. (2022). "Vine copula based dependence modeling in sustainable finance". In: *The Journal of Finance and Data Science* 8, pp. 309–330. DOI: 10.1016/j.jfds.2022.11.003.

Dang, U. J., Browne, R. P., and McNicholas, P. D. (2015). "Mixtures of multivariate power exponential distributions". In: *Biometrics* 71.4, pp. 1081–1089. DOI: 10.1111/biom.12351.

Demers, E., Hendrikse, J., Joos, P., and Lev, B. (2021). "ESG did not immunize stocks during the COVID-19 crisis, but investments in intangible assets did". In: *Journal of Business Finance & Accounting* 48.3-4, pp. 433–462. DOI: 10.1111/jbfa.12523.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). "Maximum likelihood from incomplete data via the EM algorithm". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 39.1, pp. 1–38. DOI: 10.1111/j.2517-6161.1977.tb01600.x.

Dheeru, D. and Karra Taniskidou, E. (2017). *UCI machine learning repository*. URL: http://archive.ics.uci.edu/ml.

Díaz, V., Ibrushi, D., and Zhao, J. (2021). "Reconsidering systematic factors during the Covid-19 pandemic–the rising importance of ESG". In: *Finance Research Letters* 38, p. 101870. DOI: 10.1016/j.frl.2020.101870.

Dicuonzo, G., Donofrio, F., Iannuzzi, A. P., and Dell'Atti, V. (2022). "The integration of sustainability in corporate governance systems: an innovative framework applied to the European systematically important banks". In: *International Journal of Disclosure and Governance*, pp. 1–15. DOI: 10.1057/s41310-021-00140-2.

Diday, E. and Vrac, M. (2005). "Mixture decomposition of distributions by copulas in the symbolic data analysis framework". In: *Discrete Applied Mathematics* 147.1, pp. 27–41. DOI: 10.1016/j.dam.2004.06.018.

Diemont, D., Moore, K., and Soppe, A. (2016). "The downside of being responsible: corporate social responsibility and tail risk". In: *Journal of Business Ethics* 137.2, pp. 213–229. DOI: 10.1007/s10551-015-2549-9.

Ding, W. and Song, P. X.-K. (2016). "EM algorithm in Gaussian copula with missing data". In: *Computational Statistics & Data Analysis* 101, pp. 1–11. DOI: 10.1016/j.csda.2016.01.008.

Ding, W., Levine, R., Lin, C., and Xie, W. (2021). "Corporate immunity to the COVID-19 pandemic". In: *Journal of Financial Economics* 141.2, pp. 802–830. DOI: 10.1016/j.jfineco.2021.03.005.

Dißmann, J., Brechmann, E. C., Czado, C., and Kurowicka, D. (2013). "Selecting and estimating regular vine copulae and application to financial returns". In: *Computational Statistics and Data Analysis* 59, pp. 52–69. DOI: 10.1016/j.csda.2012.08.010.

Du, S. and Yu, K. (2021). "Do corporate social responsibility reports convey value relevant information? Evidence from report readability and tone". In: *Journal of Business Ethics* 172.2, pp. 253–274. DOI: 10.1007/s10551-020-04496-3.

D'Amato, V., D'Ecclesia, R., and Levantesi, S. (2021a). "ESG score prediction through random forest algorithm". In: *Computational Management Science*, pp. 1–27. DOI: 10.1007/s10287-021-00419-3.

D'Amato, V., D'Ecclesia, R., and Levantesi, S. (2021b). "Fundamental ratios as predictors of ESG scores: a machine learning approach". In: *Decisions in Economics and Finance* 44.2, pp. 1087–1110. DOI: 10.1007/s10203-021-00364-5.

EBA (2020). *Management and supervision of ESG risks for credit institutions and investment firms*. visited on 2021-10-26. European Banking Authority. URL: https://www.eba.europa.eu/eba-launches-consultation-incorporate-esg-risks-governance-risk-management-and-supervision-credit.

EBA (2022). *Final draft implementing technical standards on prudential disclosures on ESG risks in accordance with Article 449a CRR*. visited on 2022-01-27. European Banking Authority. URL: https://www.eba.europa.eu/eba-publishes-binding-standards-pillar-3-disclosures-esg-risks.

Engelhardt, N., Ekkenga, J., and Posch, P. (2021). "ESG ratings and stock performance during the COVID-19 crisis". In: *Sustainability* 13.13. DOI: 10.3390/su13137133.

Eskidere, Ö., Ertaş, F., and Hanilçi, C. (2012). "A comparison of regression methods for remote tracking of Parkinson's disease progression". In: *Expert Systems with Applications* 39.5, pp. 5523–5528. DOI: 10.1016/j.eswa.2011.11.067.

ESMA (2022). *Sustainable finance roadmap 2022-2024*. https://www.esma.europa.eu/sites/default/files/library/esma30-379-1051_sustainable_finance_roadmap.pdf. visited on 2022-02-25.

Farah, T., Li, J., Li, Z., and Shamsuddin, A. (2021). "The non-linear effect of CSR on firms' systematic risk: international evidence". In: *Journal of International Financial Markets, Institutions and Money* 71, p. 101288. DOI: 10.1016/j.intfin.2021.101288.

Fatemi, A., Glaum, M., and Kaiser, S. (2018). "ESG performance and firm value: the moderating role of disclosure". In: *Global Finance Journal* 38, pp. 45–64. DOI: 10.1016/j.gfj.2017.03.001.

Fernández, C. and Steel, M. F. (1998). "On Bayesian modeling of fat tails and skewness". In: *Journal of the american statistical association* 93.441, pp. 359–371. DOI: 10.1080/01621459.1998.10474117.

Fraley, C. and Raftery, A. E. (1998). "How many clusters? Which clustering method? Answers via model-based cluster analysis". In: *The Computer Journal* 41.8, pp. 578–588. DOI: 10.1093/comjnl/41.8.578.

Franczak, B. C., Browne, R. P., and McNicholas, P. D. (2014). "Mixtures of shifted asymmetric laplace distributions". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36.6, pp. 1149–1157. DOI: 10.1109/TPAMI.2013.216.

Friede, G., Busch, T., and Bassen, A. (2015). "ESG and financial performance: aggregated evidence from more than 2000 empirical studies". In: *Journal of Sustainable Finance & Investment* 5.4, pp. 210–233. DOI: 10.1080/20430795.2015.1118917.

Gambacciani, M. and Paolella, M. S. (2017). "Robust normal mixtures for financial portfolio allocation". In: *Econometrics and Statistics* 3, pp. 91–111. DOI: 10.1016/j.ecosta.2017.02.003.

Gangi, F. and Varrone, N. (2018). "Screening activities by socially responsible funds: a matter of agency?" In: *Journal of Cleaner Production* 197, pp. 842–855. DOI: 10.1016/j.jclepro.2018.06.228.

Garel, A. and Petit-Romec, A. (2021). "Investor rewards to environmental responsibility: evidence from the COVID-19 crisis". In: *Journal of Corporate Finance* 68, p. 101948. DOI: 10.1016/j.jcorpfin.2021.101948.

Genest, C., Ghoudi, K., and Rivest, L.-P. (1995). "A semiparametric estimation procedure of dependence parameters in multivariate families of distributions". In: *Biometrika* 82.3, pp. 543–552. DOI: 10.1093/biomet/82.3.543.

Genuer, R., Poggi, J.-M., and Tuleau-Malot, C. (2010). "Variable selection using random forests". In: *Pattern recognition letters* 31.14, pp. 2225–2236. DOI: 10.1016/j.patrec.2010.03.014.

Gershenfeld, N. (1997). "Nonlinear inference and cluster-weighted modeling". In: *Annals of the New York Academy of Sciences* 808.1, pp. 18–24.

Gyönyörová, L., Stachoň, M., and Stašek, D. (2021). "ESG ratings: relevant information or misleading clue? Evidence from the S&P global 1200". In: *Journal of Sustainable Finance & Investment*, pp. 1–35. DOI: 10.1080/20430795.2021.1922062.

Hartigan, J. A. and Wong, M. A. (1979). "Algorithm AS 136: a k-means clustering algorithm". In: *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 28.1, pp. 100–108. DOI: 10.2307/2346830.

Hastie, T., Tibshirani, R., and Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Vol. 2. New York, NY: Springer. DOI: 10.1007/978-0-387-84858-7.

Hennig, C. (2010). "Methods for merging Gaussian mixture components". In: *Advances in Data Analysis and Classification* 4, pp. 3–34. DOI: 10.1007/s11634-010-0058-3.

Hennig, C. (2022). "An empirical comparison and characterisation of nine popular clustering methods". In: *Advances in Data Analysis and Classification* 16.1, pp. 201–229. DOI: 10.1007/s11634-021-00478-z.

Hölker, A. C., Mayer, M., Presterl, T., Bolduan, T., Bauer, E., Ordas, B., Brauner, P. C., Ouzunova, M., Melchinger, A. E., and Schön, C.-C. (2019). "European maize landraces made accessible for plant breeding and genome-based studies". In: *Theoretical and Applied Genetics* 132.12, pp. 3333–3345.

Hollander, M., Wolfe, D. A., and Chicken, E. (2013). *Nonparametric statistical methods*. Vol. 751. John Wiley & Sons. DOI: 10.1002/9781119196037.

Hu, L. (2006). "Dependence patterns across financial markets: a mixed copula approach". In: *Applied Financial Economics* 16.10, pp. 717–729. DOI: 10.1080/09603100500426515.

International Platform on Sustainable Finance (2021). *International platform on sustainable finance*. https://ec.europa.eu/info/sites/default/files/business_economy_euro/banking_and_finance/documents/211104-ipsf-esg-disclosure-report_en.pdf. visited on 2022-04-20.

Jarjir, S. L., Nasreddine, A., and Desban, M. (2020). "Corporate social responsibility as a common risk factor". In: *Global Finance Journal*, p. 100577. DOI: 10.1016/j.gfj.2020.100577.

Joe, H. (1996). "Families of m-variate distributions with given margins and $m(m-1)/2$ bivariate dependence parameters". In: *Lecture Notes-Monograph Series* 28, pp. 120–141. DOI: 10.1214/lnms/1215452614.

Joe, H. (2014). *Dependence modeling with copulas*. New York: Chapman and Hall/CRC. DOI: 10.1201/b17116.

Joe, H. and Xu, J. J. (1996). "The estimation method of inference functions for margins for multivariate models". In: *Technical Report no. 166, Department of Statistics, University of British Columbia*, pp. 1–21. DOI: 10.14288/1.0225985.

Jones, M. C. and Faddy, M. J. (2003). "A skew extension of the t-distribution, with applications". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 65.1, pp. 159–174. DOI: 10.1111/1467-9868.00378.

Karlis, D. and Xekalaki, E. (2003). "Choosing initial values for the EM algorithm for finite mixtures". In: *Computational Statistics and Data Analysis* 41.3, pp. 577–590. DOI: 10.1016/S0167-9473(02)00177-9.

Kendall, M. G. (1938). "A new measure of rank correlation". In: *Biometrika* 30.1/2, pp. 81–93. DOI: 10.2307/2332226.

Kim, J. M., Kim, D., Liao, S. M., and Jung, Y. S. (2013). "Mixture of D-vine copulas for modeling dependence". In: *Computational Statistics and Data Analysis* 64, pp. 1–19. DOI: 10.1016/j.csda.2013.02.018.

Klaaßen, L. and Stoll, C. (2021). "Harmonizing corporate carbon footprints". In: *Nature communications* 12.1, pp. 1–13. DOI: 10.1038/s41467-021-26349-x.

Klugman, S. A., Panjer, H. H., and Willmot, G. E. (2019). *Loss models: from data to decisions*. John Wiley & Sons. DOI: 10.1002/9780470391341.

Koenker, R. (2011). "Additive models for quantile regression: model selection and confidence bandaids". In: *Brazilian Journal of Probability and Statistics* 25.3, pp. 239–262. DOI: 10.1214/10-BJPS131.

Koenker, R. (2022). *quantreg: quantile regression*. R package version 5.88.

Koenker, R. and Bassett Jr, G. (1978). "Regression quantiles". In: *Econometrica: journal of the Econometric Society*, pp. 33–50. DOI: 10.2307/1913643.

Kosmidis, I. and Karlis, D. (2016). "Model-based clustering using copulas with applications". In: *Statistics and Computing* 26, pp. 1079–1099. DOI: 10.1007/s11222-015-9590-5.

Kraus, D. and Czado, C. (2017). "D-vine copula based quantile regression". In: *Computational Statistics & Data Analysis* 110, pp. 1–18. DOI: 10.1016/j.csda.2016.12.009.

Krupskii, P. and Joe, H. (2013). "Factor copula models for multivariate data". In: *Journal of Multivariate Analysis* 120, pp. 85 –101. DOI: 10.1016/j.jmva.2013.05.001.

Krupskii, P. and Joe, H. (2015). "Tail-weighted measures of dependence". In: *Journal of Applied Statistics* 42.3, pp. 614–629. DOI: 10.1080/02664763.2014.980787.

Kudratova, S., Huang, X., Kudratov, K., and Qudratov, S. (2020). "Corporate sustainability and stakeholder value trade-offs in project selection through optimization modeling:

Application of investment banking". In: *Corporate Social Responsibility and Environmental Management* 27.2, pp. 815–824. DOI: 10.1002/csr.1846.

Kumar, N. C. A., Smith, C., Badis, L., Wang, N., Ambrosy, P., and Tavares, R. (2016). "ESG factors and risk-adjusted performance: a new quantitative model". In: *Journal of Sustainable Finance & Investment* 6.4, pp. 292–300. DOI: 10.1080/20430795.2016.1234909.

Lagasio, V. and Cucari, N. (2019). "Corporate governance and environmental social governance disclosure: a meta-analytical review". In: *Corporate Social Responsibility and Environmental Management* 26.4, pp. 701–711. DOI: 10.1002/csr.1716.

Larson, J., Menickelly, M., and Wild, S. M. (2019). "Derivative-free optimization methods". In: *Acta Numerica* 28, 287–404. DOI: 10.1017/S0962492919000060.

Lee, D., Joe, H., and Krupskii, P. (2018). "Tail-weighted dependence measures with limit being the tail dependence coefficient". In: *Journal of Nonparametric Statistics* 30.2, pp. 262–290. DOI: 10.1080/10485252.2017.1407414.

Lee, S. and Lee, T. (2011). "Value-at-risk forecasting based on Gaussian mixture ARMA–GARCH model". In: *Journal of Statistical Computation and Simulation* 81.9, pp. 1131–1144. DOI: 10.1080/00949651003752320.

Lee, S. and McLachlan, G. J. (2014). "Finite mixtures of multivariate skew t-distributions: some recent and new results". In: *Statistics and Computing* 24, pp. 181–202. DOI: 10.1007/s11222-012-9362-4.

Lemaréchal, C. (2005). "A view of line-searches". In: *Optimization and Optimal Control: Proceedings of a Conference Held at Oberwolfach, March 16–22, 1980*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 59–78. DOI: 10.1007/BFb0004506.

Lemonte, A. J. (2014). "The beta log-logistic distribution". In: *Brazilian Journal of Probability and Statistics* 28.3, pp. 313 –332. DOI: 10.1214/12-BJPS209.

Li, B., Zhang, N., Wang, Y.-G., George, A. W., Reverter, A., and Li, Y. (2018). "Genomic prediction of breeding values using a subset of SNPs identified by three machine learning methods". In: *Frontiers in genetics* 9, p. 237. DOI: 10.3389/fgene.2018.00237.

Li, K., Liu, X., Mai, F., and Zhang, T. (2021). "The role of corporate culture in bad times: evidence from the COVID-19 pandemic". In: *Journal of Financial and Quantitative Analysis* 56.7, pp. 2545–2583. DOI: 10.1017/S0022109021000326.

Liaw, A. and Wiener, M. (2002). "Classification and regression by randomForest". In: *R News* 2.3, pp. 18–22.

Lin, T. I., Lee, J. C., and Yen, S. Y. (2007). "Finite mixture modelling using the skew normal distribution". In: *Statistica Sinica* 17.3, pp. 909–927.

Lins, K. V., Servaes, H., and Tamayo, A. (2017). "Social capital, trust, and firm performance: the value of corporate social responsibility during the financial crisis". In: *the Journal of Finance* 72.4, pp. 1785–1824. DOI: 10.1111/jofi.12505.

Liu, C. and Rubin, D. B. (1994). "The ECME algorithm: a simple extension of EM and ECM with faster monotone convergence". In: *Biometrika* 81.4, pp. 633–648. DOI: 10.2307/2337067.

Loader, C. (2006). *Local regression and likelihood*. New York, NY: Springer. DOI: 10.1007/b98858.

Lööf, H., Sahamkhadam, M., and Stephan, A. (2021). "Is corporate social responsibility investing a free lunch? The relationship between ESG, tail risk, and upside potential of stocks before and during the COVID-19 crisis". In: *Finance Research Letters*. DOI: 10.1016/j.frl.2021.102499.

Mangasarian, O. L., Street, W. N., and Wolberg, W. H. (1995). "Breast cancer diagnosis and prognosis via linear programming". In: *Operations Research* 43.4, pp. 570–577. DOI: 10.1287/opre.43.4.570.

Maugis, C., Celeux, G., and Martin-Magniette, M. L. (2009). "Variable selection for clustering with gaussian mixture models". In: *Biometrics* 65.3, pp. 701–709. DOI: 10.1111/j.1541-0420.2008.01160.x.

Mayer, M., Hölker, A. C., González-Segovia, E., Bauer, E., Presterl, T., Ouzunova, M., Melchinger, A. E., and Schön, C.-C. (2020). "Discovery of beneficial haplotypes for complex traits in maize landraces". In: *Nature communications* 11.1, pp. 1–10. DOI: 10.1038/s41467-020-18683-3.

Mazzotta, R., Bronzetti, G., and Veltri, S. (2020). "Are mandatory non-financial disclosures credible? Evidence from Italian listed companies". In: *Corporate Social Responsibility and Environmental Management* 27.4, pp. 1900–1913. DOI: 10.1002/csr.1935.

McLachlan, G. J. and Peel, D. (2000). *Finite mixture models/Geoffrey McLachlan, David Peel*. Wiley New York; Chichester. ISBN: 0471006262. DOI: 10.1002/0471721182.

McNicholas, P. D. (2016). "Model-based clustering". In: *Journal of Classification* 33, pp. 331–373. DOI: 10.1007/s00357-016-9211-9.

Meinshausen, N. (2006). "Quantile regression forests". In: *Journal of Machine Learning Research* 7.35, pp. 983–999.

Meinshausen, N. (2022). *quantregForest: quantile regression forests*. R package version 1.3-7.

Meng, X.-L. and Rubin, D. B. (1993). "Maximum likelihood estimation via the ECM algorithm: a general framework". In: *Biometrika* 80.2, pp. 267–278. DOI: 10.2307/2337198.

Milne, M. J. and Gray, R. (2013). "W(h)ither ecology? The triple bottom line, the global reporting initiative, and corporate sustainability reporting". In: *Journal of Business Ethics* 118.1, pp. 13–29. DOI: 10.1007/s10551-012-1543-8.

Mohr, A., Schumacher, C., and Kiefner, V. (2022). "Female executives and multinationals' support of the UN's sustainable development goals". In: *Journal of World Business* 57.3, p. 101304. DOI: 10.1016/j.jwb.2021.101304.

Montesinos-López, A., Montesinos-López, O. A., Villa-Diharce, E. R., Gianola, D., and Crossa, J. (2019). "A robust Bayesian genome-based median regression model". In: *Theoretical and Applied Genetics* 132.5, pp. 1587–1606. DOI: 10.1007/s00122-019-03303-6.

Morales-Nápoles, O. (2010). "Counting vines". In: *Dependence Modeling: Vine Copula Handbook*. World Scientific, pp. 189–218. ISBN: 9789814299886. DOI: 10.1142/9789814299886_0009.

Morelli, G. and D'Ecclesia, R. (2021). "Responsible investments reduce market risks". In: *Decisions in Economics and Finance* 44.2, pp. 1211–1233. DOI: 10.1007/s10203-021-00351-w.

MSCI (2022). *MSCI ESG ratings methodology*. visited on 2022-01-23. URL: https://www.msci.com/documents/1296102/4769829/MSCI+ESG+Ratings+Methodology+-+Exec+Summary+Dec+2020.pdf/15e36bed-bba2-1038-6fa0-2cf52a0c04d6?t=1608110671584 (visited on 01/23/2022).

Murray, P. M., Browne, R. P., and McNicholas, P. D. (2017). "A mixture of SDB skew-t factor analyzers". In: *Econometrics and Statistics* 3, pp. 160–168. DOI: 10.1016/j.ecosta.2017.05.001.

Nagler, T., Bumann, C., and Czado, C. (2019). "Model selection in sparse high-dimensional vine copula models with an application to portfolio risk". In: *Journal of Multivariate Analysis* 172, pp. 180–192. DOI: 10.1016/j.jmva.2019.03.004.

Nagler, T. (2022). *vinereg: D-Vine quantile regression*. R package version 0.8.1.

Nagler, T. and Vatter, T. (2022a). *kde1d: univariate kernel density estimation*. R package version 1.0.4.

Nagler, T. and Vatter, T. (2022b). *pyvinecopulib*. Version v0.6.2. Zenodo.

Nagler, T., Schepsmeier, U., Stoeber, J., Brechmann, E. C., Graeler, B., and Erhardt, T. (2021). *VineCopula: statistical inference of vine copulas*. R package version 2.4.3.

Nelsen, R. B. (2007). *An introduction to copulas*. New York, NY: Springer. DOI: 10.1007/0-387-28678-0.

Nguyen, T. M. and Wu, Q. J. (2011). "Dirichlet Gaussian mixture model: application to image segmentation". In: *Image and Vision Computing* 29.12, pp. 818–828. DOI: 10.1016/j.imavis.2011.09.001.

Panagiotelis, A., Czado, C., and Joe, H. (2012). "Pair copula constructions for multivariate discrete data". In: *Journal of the American Statistical Association* 107.499, pp. 1063–1072. DOI: 10.1080/01621459.2012.682850.

Panagiotelis, A., Czado, C., Joe, H., and Stöber, J. (2017). "Model selection for discrete regular vine copulas". In: *Computational Statistics and Data Analysis* 106, pp. 138–152. DOI: 10.1016/j.csda.2016.09.007.

Parzen, E. (1962). "On estimation of a probability density function and mode". In: *The annals of mathematical statistics* 33.3, pp. 1065–1076. DOI: 10.1214/aoms/1177704472.

Pedersen, L. H., Fitzgibbons, S., and Pomorski, L. (2021). "Responsible investing: The ESG-efficient frontier". In: *Journal of Financial Economics* 142.2, pp. 572–597. DOI: 10.1016/j.jfineco.2020.11.001.

Peel, D. and McLachlan, G. J. (2000). "Robust mixture modelling using the t distribution". In: *Statistics and Computing* 10, pp. 339–348. DOI: 10.1023/A:1008981510081.

Pérez-Rodríguez, P., Montesinos-López, O. A., Montesinos-López, A., and Crossa, J. (2020). "Bayesian regularized quantile regression: a robust alternative for genome-based prediction of skewed data". In: *The Crop Journal* 8.5, pp. 713–722. DOI: 10.1016/j.cj.2020.04.009.

Pingel, R. (2014). "Some approximations of the logistic distribution with application to the covariance matrix of logistic regression". In: *Statistics & Probability Letters* 85, pp. 63–68. DOI: 10.1016/j.spl.2013.11.007.

Powell, M. J. D. (2015). "On fast trust region methods for quadratic models with linear constraints". In: *Mathematical Programming Computation* 7.3, pp. 237–267. DOI: 10.1007/s12532-015-0084-4.

Powell, M. J. (2003). "On trust region methods for unconstrained minimization without derivatives". In: *Mathematical programming* 97, pp. 605–623. DOI: 10.1007/s10107-003-0430-6.

Prates, M. O., Cabral, C. R. B., and Lachos, V. H. (2013). "mixsmsn: fitting finite mixture of scale mixture of skew-normal distributions". In: *Journal of Statistical Software* 54.12, pp. 1–20. DOI: 10.18637/jss.v054.i12.

PRI (2021). *A practical guide to ESG integration for equity investing*. visited on 2021-11-22. URL: https://www.unpri.org/download?ac=10.

Qian, J., Tanigawa, Y., Du, W., Aguirre, M., Chang, C., Tibshirani, R., Rivas, M. A., and Hastie, T. (2020). "A fast and scalable framework for large-scale and ultrahigh-dimensional sparse regression with application to the UK Biobank". In: *PLoS genetics* 16.10. DOI: 10.1371/journal.pgen.1009141.

Raftery, A. E. and Dean, N. (2006). "Variable selection for model-based clustering". In: *Journal of the American Statistical Association* 101.473, pp. 168–178. DOI: 10.1198/016214506000000113.

Rédei, G. P. (2008). *Encyclopedia of genetics, genomics, proteomics, and informatics*. Dordrecht: Springer. DOI: 10.1007/978-1-4020-6754-9.

Refinitiv (2021a). *Environmental, social and governance (ESG) scores from Refinitiv*. visited on 2021-10-26. URL: https://www.refinitiv.com/content/dam/marketing/en_us/documents/methodology/refinitiv-esg-scores-methodology.pdf (visited on 10/26/2021).

Refinitiv (2021b). *The Refinitiv business classification*. visited on 2021-10-26. URL: https://www.refinitiv.com/content/dam/marketing/en_us/documents/quick-reference-guides/trbc-business-classification-quick-guide.pdf (visited on 10/26/2021).

Refinitiv (2022). *Create custom ESG scoring*. visited on 2022-03-28. URL: https://www.refinitiv.com/content/dam/marketing/en_us/documents/fact-sheets/build-custom-esg-scores-using-refinitiv-esg-data-in-eikon.pdf.

Rehman, Z. u., Khan, A., and Rahman, A. (2020). "Corporate social responsibility's influence on firm risk and firm performance: the mediating role of firm reputation". In: *Corporate Social Responsibility and Environmental Management* 27.6, pp. 2991–3005. DOI: 10.1002/csr.2018.

Revelli, C. (2017). "Socially responsible investing (SRI): From mainstream to margin?" In: *Research in International Business and Finance* 39, pp. 711–717. DOI: 10.1016/j.ribaf.2015.11.003.

Roy, A. and Parui, S. K. (2014). "Pair-copula based mixture models and their application in clustering". In: *Pattern Recognition* 47.4, pp. 1689–1697. DOI: 10.1016/j.patcog.2013.10.004.

Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D. A., and Nolan, G. P. (2005). "Causal protein-signaling networks derived from multiparameter single-cell data". In: *Science* 308.5721, pp. 523–529. DOI: 10.1126/science.1105809.

Sahin, Ö. (2021). *vineclust*. Version 0.1.0. URL: https://github.com/oezgesahin/vineclust.

Sahin, Ö. and Czado, C. (2022a). "High-dimensional sparse vine copula regression with application to genomic prediction". In: *arXiv preprint arXiv:2208.12383*. DOI: 10.48550/arXiv.2208.12383.

Sahin, Ö. and Czado, C. (2022b). "Vine copula mixture models and clustering for non-Gaussian data". In: *Econometrics and Statistics* 22, pp. 136–158. DOI: 10.1016/j.ecosta.2021.08.011.

Sahin, Ö., Bax, K., Czado, C., and Paterlini, S. (2022). "Environmental, social, governance scores and the missing pillar - why does missing information matter?" In: *Corporate Social Responsibility and Environmental Management* 29.5, pp. 1782–1798. DOI: 10.1002/csr.2326.

Sahin, Ö., Bax, K., Paterlini, S., and Czado, C. (2023). "The pitfalls of (non-definitive) environmental, social, and governance scoring methodology". In: *Global Finance Journal* 56, p. 100780. DOI: 10.1016/j.gfj.2022.100780.

Santamaria, R., Paolone, F., Cucari, N., and Dezi, L. (2021). "Non-financial strategy disclosure and environmental, social and governance score: Insight from a configurational approach". In: *Business Strategy and the Environment* 30.4, pp. 1993–2007. DOI: 10.1002/bse.2728.

Schwarz, G. (1978). "Estimating the dimension of a model". In: *The Annals of Statistics* 6.2, pp. 461–464. DOI: 10.1214/aos/1176344136.

Scrucca, L. and Raftery, A. (2015). "Improved initialisation of model-based clustering using Gaussian hierarchical partitions". In: *Advances in Data Analysis and Classification* 9, pp. 447–460. DOI: 10.1007/s11634-015-0220-z.

Scrucca, L., Fop, M., Murphy, T. B., and Raftery, A. E. (2016). "Mclust 5: Clustering, classification and density estimation using Gaussian finite mixture models". In: *R Journal* 8.1, pp. 289–317. DOI: 10.32614/rj-2016-021.

Searcy, C. (2012). "Corporate sustainability performance measurement systems: a review and research agenda". In: *Journal of Business Ethics* 107.3, pp. 239–253. DOI: 10.1007/s10551-011-1038-z.

Sheather, S. J. and Jones, M. C. (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. DOI: 10.1111/j.2517-6161.1991.tb01857.x.

Sherwood, B. and Maidman, A. (2020). *rqPen: penalized quantile regression*. R package version 2.2.2.

Sklar, A (1959). "Fonctions de répartition à n dimensions et leurs marges". In: *Publications de L'Institut de Statistique de L'Université de Paris* 8, pp. 229–231.

Sorensen, D. C. (1981). "Trust-region methods for unconstrained minimization". In:

Speiser, J. L., Miller, M. E., Tooze, J., and Ip, E. (2019). "A comparison of random forest variable selection methods for classification prediction modeling". In: *Expert systems with applications* 134, pp. 93–101. DOI: 10.1016/j.eswa.2019.05.028.

Steinwart, I. and Christmann, A. (2011). "Estimating conditional quantiles with the help of the pinball loss". In: *Bernoulli* 17.1, pp. 211–225. DOI: 10.3150/10-BEJ267.

Stöber, J. and Schepsmeier, U. (2013). "Estimating standard errors in regular vine copula models". In: *Computational Statistics* 28, pp. 2679–2707. DOI: 10.1007/s00180-013-0423-8.

Stöber, J., Joe, H., and Czado, C. (2013). "Simplified pair copula constructions-Limitations and extensions". In: *Journal of Multivariate Analysis* 119, pp. 101–118. DOI: 10.1016/j.jmva.2013.04.014.

Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., and Zeileis, A. (2008). "Conditional variable importance for random forests". In: *BMC bioinformatics* 9, pp. 1–11. DOI: 10.1186/1471-2105-9-307.

Strubell, E., Ganesh, A., and McCallum, A. (2019). "Energy and policy considerations for deep learning in NLP". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy, pp. 3645–3650. DOI: 10.18653/v1/P19-1355.

Sun, M., Konstantelos, I., and Strbac, G. (2017). "C-Vine copula mixture model for clustering of residential electrical load pattern data". In: *IEEE Transactions on Power Systems* 32.3, pp. 2382–2393. DOI: 10.1109/TPWRS.2016.2614366.

Sustainalytics (2022). *ESG risk ratings methodology*. visited on 2022-01-23. URL: https://connect.sustainalytics.com/esg-risk-ratings-methodology.

Tepegjozova, M. (2019). "D-and C-vine quantile regression for large data sets". MSc Thesis. Technical University of Munich.

Tepegjozova, M., Zhou, J., Claeskens, G., and Czado, C. (2022). "Nonparametric C-and D-vine-based quantile regression". In: *Dependence Modeling* 10.1, pp. 1–21. DOI: 10.1515/demo-2022-0100.

Valbuena-Hernandez, J. P. and Mandojana, N. Ortiz-de (2022). "Encouraging corporate sustainability through effective strategic partnerships". In: *Corporate Social Responsibility and Environmental Management*. DOI: 10.1002/csr.2188.

Verheyden, T., Eccles, R. G., and Feiner, A. (2016). "ESG for all? The impact of ESG screening on return, risk, and diversification". In: *Journal of Applied Corporate Finance* 28.2, pp. 47–55. DOI: 10.1111/jacf.12174.

Vrac, M., Chédin, A., and Diday, E. (2005). "Clustering a global field of atmospheric profiles by mixture decomposition of copulas". In: *Journal of Atmospheric and Oceanic Technology* 22.10, pp. 1445–1459. DOI: 10.1175/JTECH1795.1.

Waerden, B. Van der (1953). "Ein neuer Test für das Problem der zwei Stichproben". In: *Mathematische Annalen* 126.1, pp. 93–107. DOI: 10.1007/BF01343153.

Wang, W.-L. and Lin, T.-I. (2015). "Robust model-based clustering via mixtures of skew-t distributions with missing information". In: *Advances in Data Analysis and Classification* 9.4, pp. 423–445. DOI: 10.1007/s11634-015-0221-y.

Weiß, G. N. and Scheffer, M. (2015). "Mixture pair-copula-constructions". In: *Journal of Banking & Finance* 54, pp. 175–191. DOI: 10.1016/j.jbankfin.2015.01.008.

Wong, W. C., Batten, J. A., Mohamed-Arshad, S. B., Nordin, S., Adzis, A. A., et al. (2021). "Does ESG certification add firm value?" In: *Finance Research Letters* 39, p. 101593. DOI: 10.1016/j.frl.2020.101593.

Wood, S. N. (2017). *Generalized additive models: an introduction with R*. Boca Raton: Chapman and Hall/CRC. DOI: 10.1201/9781315370279.

Wright, S. and Nocedal, J. (2006). *Numerical optimization*. New York, NY: Springer. ISBN: 978-0-387-30303-1. DOI: 10.1007/978-0-387-40065-5.

Yan, X., Xiong, W., Hu, L., Wang, F., and Zhao, K. (2015). "Missing value imputation based on gaussian mixture model for the internet of things". In: *Mathematical Problems in Engineering* 2015. DOI: 10.1155/2015/548605.

Zanin, L. (2021). "Estimating the effects of ESG scores on corporate credit ratings using multivariate ordinal logit regression". In: *Empirical Economics*, pp. 1–32. DOI: 10.1007/s00181-021-02121-4.

Zhang, Q. and Shi, X. (2017). "A mixture copula Bayesian network model for multimodal genomic data". In: *Cancer Informatics* 16. DOI: 10.1177/1176935117702389.

Zhu, K., Kurowicka, D., and Nane, G. F. (2021). "Simplified R-vine based forward regression". In: *Computational Statistics & Data Analysis* 155, p. 107091. DOI: 10.1016/j.csda.2020.107091.

Zhuang, H., Diao, L., and Yi, G. Y. (2021). "A Bayesian nonparametric mixture model for grouping dependence structures and selecting copula functions". In: *Econometrics and Statistics*. DOI: 10.1016/j.ecosta.2021.03.009.