Delft University of Technology
Faculty of EEMCS

Copula-Based Regression for Discrete Data

Conditional quantiles and variable selection under discretized covariates

by
Thomas Molendijk

to obtain the degree of Master of Science in Applied Mathematics
at the Delft University of Technology,
to be defended publicly on Friday August 29th, 2025, at 10:00 AM.

Student number:	5363357			
Thesis committee:	Prof.	Ö. Şahin,	TU Delft,	Daily Supervisor
	Prof.	G.F. Nane,	TU Delft,	Responsible Supervisor
	Prof.	M. Vittorietti,	TU Delft,	External Examiner

TU Delft

Contents

Acknowledgments

First things first, I want to thank my supervisor Özge Şahin for guiding me throughout this thesis. She gave me the freedom to turn this thesis into whatever I wanted it to be, and gently steered me back on course when I got lost in the weeds of this vine copula stuff. This thesis gave me a taste of what it is like to actually do research, and I learned research can taste a lot like failure. Her endless positivity made me feel like I wasn't failing at doing research even though, realistically, sometimes I definitely was.

I also want to thank my committee members Tina Nane, who introduced me this thesis topic, and Martina Vittorietti, who had to battle the examination committee to be here.

Before I started this masters, I could not have imagined I would be surrounded by so many great people. Here's to everyone who has made my time studying Applied Mathematics a fun one; there are too many of you to name, but you know who you are. That said, I want to give special thanks to the people who have been studying with me on the 4^{th} floor of EWI for the duration of my thesis, without whom I surely would have a better thesis, and a lot less fun writing it. Going to EWI everyday knowing I'm going to have a good time was invaluable, thank you. The Dutch & Italians, thanks for your company (and your assignments) in that hectic first year. You guys know how to make life memorable, balancing the Yin to the Dutch Yang. A special thanks to Ivan for pointing out all my mistakes when proofreading the draft of this thesis, and for always distracting me when I didn't feel like working.

How can I properly thank the people who gave me everything? To my whole family, to those who are here and to those who couldn't be: my mom, my dad, my sister and my grandparents, thank you for everything. I just don't know how else to say it.

Abstract

This thesis takes a step towards developing a vine-based regression method tailored to mixed-type (continuous & discrete) data. We visualize the difference between continuous conditional quantile functions of bivariate copulas and their discretized (discretized by binning the continuous variable into bins of equal probability) conditional quantile functions. This showed how discretization into a small number of bins loses dependence, especially in the tails. We show a different binning procedure can improve the preservation of dependence after discretization, motivating the development of methodologies which compute an 'optimal' (optimal in the sense of balancing tail dependence with dependence around the median) binning procedure given some data. For application in computing an 'optimal' binning procedure, we make an attempt at analytically quantifying the difference between discretized conditional quantiles and continuous conditional quantiles of bivariate copulas, but the analytical derivation is not possible for most settings (with exceptions for bivariate Clayton and bivariate Frank copulas, and the smallest conditioning value of discretized covariate).

From the aforementioned conclusion that discretization loses dependence, we infer a variable selection measure tailored to mixed-type data should be biased against discretized covariates, and that this bias should be monotone decreasing with the number of bins of the discretized covariate. The bias for/against discretized covariates of various variable selection measures is investigated, in a scenario with a single covariate and a scenario with two covariates. With a single covariate, Pearson's/polyserial correlation and Kendall's tau/tau-b were found unsuitable as variable selection measures for mixed-type data, due to a lack of bias against discretized covariates. Conditional log-likelihood and check-loss at the quantile level 0.05 seem nearly identically biased in the bivariate setting, although this should be said with the caveat that all simulation scenarios are homoskedastic. Finally, we show in a three-dimensional setting that correlation between covariates does not seem to affect the predictive performance when both covariates are continuous, but correlation between covariates has significant negative effects on the predictive performance when one of the covariates is discretized. This difference between the effect of discretization in two dimensions and three dimensions should be kept in mind when developing variable selection procedures for mixed-type data.

1 | Introduction

A statistician makes predictions by extracting information from data collected in the past. In theory (and only in theory), a perfect understanding of a perfect dataset should give one the ability to predict the future. In an effort to get one step closer to the ability to predict the future, this thesis investigates how dependence modeling methods using copulas should be adapted to account for discrete variables in the data.

Understanding and modeling dependencies between variables is a central challenge in modern statistics and data science. In many real-world applications, the data consists of a mix of continuous and discrete variables: for instance, in health data, one might encounter blood pressure measurements alongside binary indicators for disease presence; in finance, stock returns co-exist with credit ratings. Methods to model dependencies tend to be developed under the assumption of a dataset containing only continuous variables. There is a necessity to investigate whether these methods remain effective when applied to mixed-type (continuous & discrete) data.

Accurate modeling of dependencies requires a large amount of data. Data-collection is often expensive and time-consuming, making the sharing of data between organizations necessary to build effective models for prediction. Sharing data between organizations is not always legal, however. Take for instance healthcare data from hospitals in the Netherlands. Every hospital has some small amount of data on, say, bowel cancer patients' post-surgery recovery. The amount of data each hospital has is not enough to accurately model dependencies with, so they want to pool all their data together to create a sufficiently large dataset. However, such sharing of personal patient data with other organizations is heavily regulated under privacy laws. If it were possible to modify (by way of binning continuous data into discrete groups) patient data such that (1) patient-privacy is protected, and (2) dependencies in the data are largely unaffected, hospitals would be able to share privacy-protected datasets with each other, which would allow them to model their patients' post-surgery recovery more accurately. Building a privacy-preserving dataset which retains the most amount of information possible requires an understanding of how currently used dependence modeling methods handle discrete data, and of how discretization of covariates changes dependencies.

In short, a deeper understanding of how dependence modeling methods should be adapted to account for discrete variables in the data opens the door to more accurate prediction in a wide range of applications where discrete variables play a role (from healthcare to finance, and everything in between) and brings us one step closer to building informative, privacy-preserving datasets.

2 | Preliminaries

We begin by providing a brief introduction to topics relevant to understanding the contents of this thesis. In particular, this section will cover dependence measures, copulas, vine copulas and variable selection measures for the construction of vine copulas with the aim of conditional quantile estimation. A thorough introduction to the topics covered can be found in the book by Czado [1].

2.1 Dependence measures

Quantifying the strength of the dependence relationship between a pair of random variables is a tricky task. This is evident from the fact that there does not exist a single measure which quantifies completely the strength of dependence between a pair of random variables. Nevertheless, the literature is full of measures which quantify certain types of dependence. This section summarizes the dependence measures used in this thesis.

A *dependence measure* is a function of two variables which measures the degree to which two variables are dependent. In this thesis, we use the terms dependence measure and *correlation measure* interchangeably.

Pearson's correlation measure is a measure of linear correlation between two variables [2]. It is given by

$$\rho_{X,Y} = \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)\mathrm{Var}(Y)}.$$

Essentially, it is a measure of the covariance between random variables X and Y, normalized by their variance to take values in the interval $[-1,1]$. Since the covariance captures only linear dependence, Pearson's correlation also captures only linear dependence.

In the case where one of the variables is ordinal (discrete, with a clear ordering on its values), we can use *polyserial* correlation [3]. Polyserial correlation assumes the ordinal variable is produced from an underlying, normally distributed latent variable. To illustrate, say we observe a discrete (all discrete variables will be assumed to be ordinal in this thesis) variable X, which can take b values $\{x_1, x_2, \ldots, x_b\}$, and a continuous variable Y. When estimating the polyserial correlation between Y and X, we assume X is derived from an underlying normally distributed variable X^* based on some set of cut points, $c_1, \ldots, c_{b-1}$, i.e.

$$X = \begin{cases} x_1, & X^* < c_1 \\ x_2, & X^* \in [c_1, c_2) \\ \vdots & \\ x_{d-1}, & X^* \in [c_{b-2}, c_{b-1}) \\ x_d, & X^* \geq c_{b-1}. \end{cases}$$

When we assume the distribution of Y and the distribution of X^* are standard normal and that the bivariate distribution (Y, X^*) is Gaussian, we can estimate (e.g. through maximum-likelihood estimation) the parameters of the distribution of X^*, the cutpoints $c_1, \ldots, c_{b-1}$, and

finally the polyserial correlation ρ between Y and X as the Pearson's correlation between Y and the latent X^*.

When both X and Y are ordinal, we can similarly estimate their correlation by assuming both variables are derived from latent normal variables. This is called the *polychoric* correlation of X and Y [3].

Correlation measures which are independent of marginal distributions are so-called *rank-based* correlation measures. The rank of an observation refers to the rank of that observation after sorting all observations from low to high. Kendall's τ is a rank-based measure of monotonic dependence [4], which counts the number of discordant/concordant pairs of observations out of all possible pairs. A pair of observations $(x_1, y_1), (x_2, y_2)$ is concordant if $(x_1 > x_2 \wedge y_1 > y_2) \vee (x_1 < x_2 \wedge y_1 < y_2)$ is true. Conversely, the pair is discordant if $(x_1 > x_2 \wedge y_1 < y_2) \vee (x_1 < x_2 \wedge y_1 > y_2)$ is true. The sample-estimate of Kendall's τ is given by

$$\hat{\tau} = \frac{(\text{number of concordant pairs}) - (\text{number of discordant pairs})}{\text{number of pairs}}.$$

From this definition it is clear that Kendall's tau is always in the interval $[-1, 1]$. The definition of concordant/discordant pairs uses a strict inequality between observations, which means Kendall's tau can not take values -1 and 1 when there are ties in the data, such as when one of the two variables is discrete. To solve this, for discrete-continuous or discrete-discrete variable pairs the alternative Kendall's τ_b is most commonly used. Kendall's τ_b uses a different denominator, this denominator counts the maximum number of possible concordant/discordant pairs in the presence of ties so that under perfect dependence it may take values -1 and 1.

Tail-dependence coefficients quantify the probability of joint extreme events. The standard measures of lower/upper tail dependence between two random variables X and Y with marginal distributions F_X and F_Y are, respectively [5]:

$$\lambda_L = \lim_{u \downarrow 0} \mathbb{P}\left(Y > F_Y^{-1}(u) \mid X > F_X^{-1}(u)\right) \tag{2.1}$$

$$\lambda_U = \lim_{u \uparrow 1} \mathbb{P}\left(Y < F_Y^{-1}(u) \mid X < F_X^{-1}(u)\right), \tag{2.2}$$

2.2 Copulas

Copulas allow us to model a joint distribution/density in a way which separates the modeling of the marginal distributions from the modeling of the dependence between variables. Sklar's theorem provides us with the connection between the joint distribution, the marginal distributions of its components, and a copula. We define copulas as follows.

Definition 2.1 (Copula)**.** *A d-dimensional copula C is a distribution function $C : [0,1]^d \to [0,1]$ with uniform marginal distributions $U_1, U_2, \ldots, U_d \sim Unif[0,1]$, given by $C(u_1, u_2, \ldots, u_d) = P(U_1 \leq u_1, U_2 \leq u_2, \ldots, U_d \leq u_d)$. The copula density $c(u_1, \ldots, u_d)$ is then obtained by differentiating with respect to each component, $c(u_1, \ldots, u_d) = \frac{\partial^d}{\partial u_1 \ldots \partial u_d} C(u_1, \ldots, u_d)$.*

Following this definition we give Sklar's theorem.

Theorem 2.1 (Sklar's theorem [6])**.** *Let F be a d-dimensional cumulative distribution function with marginals $F_1, F_2, \ldots, F_d$. Then there exists a copula $C : [0,1]^d \to [0,1]$ such that for all $\mathbf{x} = (x_1, x_2, \ldots, x_d) \in \mathbb{R}^d$,*

$$F(x_1, x_2, \ldots, x_d) = C\big(F_1(x_1), F_2(x_2), \ldots, F_d(x_d)\big),$$

with associated density or probability mass function

$$f(x_1, x_2 \ldots, x_d) = c\big(F_1(x_1), F_2(x_2), \ldots, F_d(x_d)\big) f_1(x_1) \ldots f_d(x_d).$$

Given F, C is uniquely determined on the Cartesian product of ranges $F_1 \times F_2 \times \cdots \times F_d$. This implies C is unique if the marginals F_i are all continuous.

The converse statement is also true. Let $C : [0,1]^d \to [0,1]$ be a copula and $F_1, F_2, \ldots, F_d$ univariate marginal distributions, then the function $C\big(F_1(x_1), F_2(x_2), \ldots, F_d(x_d)\big)$ defines a d-dimensional joint distribution.

The result of Sklar's theorem allows data to be modeled in a two-step process. First, one models the marginal distribution of each variable. Next, data is transformed to the so-called *copula scale* by applying the corresponding marginal distribution of each variable to the data of that variable. This transformation is also called the *probability integral transform*. The probability integral transform is best explained by an example. Suppose we have observations $x_1, x_2, \ldots, x_n$ of some random variable X. We transform these observations to the copula scale by applying its marginal distribution $\mathbb{P}(X \leq x_1), \mathbb{P}(X \leq x_2), \ldots, \mathbb{P}(X \leq x_n)$. In essence, what the probability integral transform does is it removes the effect of marginal distributions by mapping all variables to a Unif$[0,1]$ distribution. The proof that this transformation yields uniformly distributed random variables is relatively simple, so we will provide it here. For any $q \in (0,1)$ and any continuous distribution function F_X,

$$\mathbb{P}(F_X(X) \leq q) = \mathbb{P}(X \leq F_X^{-1}(q)) = F_X(F_X^{-1}(q)) = q,$$

therefore $F_X(X) \sim \text{Unif}[0,1]$.

Using the copula scale data, we can model the dependence between variables in a space which is independent of their marginal distributions. This modeling of the dependence is done by finding the appropriate copula family and parameter which connects the probability integral transform of some set of variables, as in Sklar's theorem. Copula-based dependence modeling has the advantage of allowing for the modeling of *tail dependence* (co-movement between variables in their extremes, defined in Equation 2.1), non-linear dependence, as well as asymmetric dependence (different dependence between variables in their lower/upper tails).

2.2.1 Marginal distribution estimation

In order to obtain copula data, we require an estimate of the marginal distribution of each variable. Errors in the estimated marginal distribution propagate to errors in the estimated copula, because it gives us perturbed copula data. A 2007 paper by Kim et al. [7] aimed to compare the parametric inference for margins (IFM) method of marginal & copula estimation (which uses a parametric estimate of the marginals) with the (semi-parametric) pseudo-ML method (which uses a scaled empirical distribution function to estimate the marginals, i.e. a non-parametric estimate). A simulation study showed IFM is not robust to misspecification of the margins, and the non-parametric marginal estimation of the pseudo-ML method outperforms parametric estimation in most practical situations. It has become standard practice to use non-parametric estimation of marginal distributions, unless the parametric form of the marginal distribution is known.

2.2.2 Copula families

A copula family is a set of copulas $\{C_\theta : \theta \in \Theta\}$ indexed by a parameter θ in some parameter space Θ, where each C_θ defines a specific copula/dependence structure. In this thesis, we will work with copula families which have a one-to-one relationship between their parameter θ and their associated dependence strength given by the value of Kendall's τ, so rather than specifying the parameter value we will often specify Kendall's τ instead.

We separate the copula families discussed in this thesis into two categories, based on how that copula family is constructed. The first construction method gives us the class of *elliptical copulas*. These can be constructed by taking an 'inverse' of the equation in Sklar's theorem and

applying an elliptical joint distribution. The second class of copulas is the class of *Archimedean copulas*, produced by a generator function.

Definition 2.2 (Elliptical copula)**.** *Let F denote the cumulative distribution function (CDF) of a multivariate elliptical distribution, and let $F_1, \ldots, F_d$ be arbitrary continuous marginal distributions. The elliptical copula associated with these is obtained by the inverse of Sklar's theorem:*

$$C(u_1, \ldots, u_d) = F\left(F_1^{-1}(u_1), \ldots, F_d^{-1}(u_d)\right), \quad u_1, \ldots, u_d \in [0, 1]. \tag{2.3}$$

Two examples of elliptical copulas are the Gaussian- and t-copulas.

Example 2.1 (Bivariate Gaussian copula)**.** *Let Φ denote the CDF of the standard normal distribution, and let Φ_2 denote the CDF of a bivariate normal distribution with mean-vector $\mathbf{0}$, unit variance and correlation ρ. Equation 2.3 gives the bivariate Gaussian copula with correlation ρ as*

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

Its conditional distribution and conditional quantile function are

$$C_{2|1}(v \mid u) = \Phi\left(\frac{\Phi^{-1}(v) - \rho\,\Phi^{-1}(u)}{\sqrt{1-\rho^2}}\right), \quad u, v \in (0,1) \text{ and}$$
$$C_{2|1}^{-1}(\alpha \mid u) = \Phi\left(\Phi^{-1}(\alpha)\sqrt{1-\rho^2} + \rho\,\Phi^{-1}(u)\right), \quad u, \alpha \in (0,1).$$

Its upper and lower tail coefficients are

$$\lambda_U = \lambda_L = 0.$$

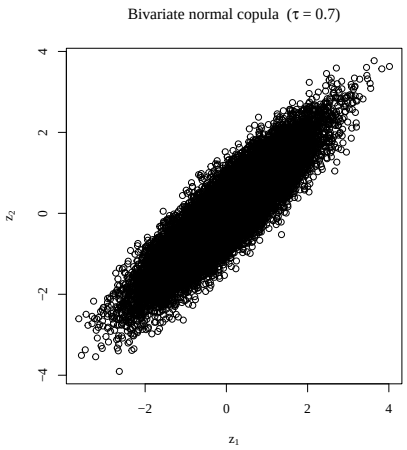


Figure 2.1: Each point represents one out of $n = 20000$ simulated observations of a bivariate Gaussian copula with parameter $\rho = 0.8$, or equivalently with Kendall's $\tau = 0.7$. The marginal distributions were chosen standard normal.

Example 2.2 (Bivariate Student's t-copula)**.** *Let t_ν denote the CDF of the univariate Student's t-distribution with parameter ν, and let $t_{2,\nu}$ denote the bivariate Student's t-distribution with mean-vector $\mathbf{0}$, unit variance and correlation ρ. Equation 2.3 gives the bivariate t-copula as*

$$C(u_1, u_2; \rho) = t_2(t_\nu^{-1}(u_1), t_\nu^{-1}(u_2); \nu, \rho), \quad u_1, u_2 \in (0, 1).$$

Its conditional distribution function and conditional quantile function are

$$C_{2|1}(v \mid u) = t_{\nu+1}\left(\frac{t_\nu^{-1}(v) - \rho\, t_\nu^{-1}(u)}{\sqrt{(1-\rho^2)\big(1 + (t_\nu^{-1}(u))^2/\nu\big)}}\right), \quad u, v \in (0,1) \text{ and}$$

$$C_{2|1}^{-1}(\alpha \mid u) = t_\nu^{-1}\left(\sqrt{\frac{\nu + (t_\nu^{-1}(u))^2}{\nu+1}}\,(t_{\nu+1}^{-1}(\alpha)) + \rho\, t_\nu^{-1}(u)\right), \quad u, \alpha \in (0,1).$$

Its upper and lower tail coefficients are

$$\lambda_U = \lambda_L = 2t_{\nu+1}\left(-\sqrt{v+1}\sqrt{\frac{1-\rho}{1+\rho}}\right).$$

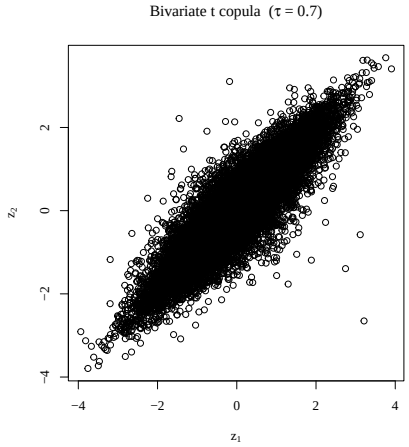


Figure 2.2: Each point represents one out of $n = 20000$ simulated observations of a bivariate Student's t-copula with parameter $\rho = 0.8$ and $\nu = 5$. The marginal distributions were chosen standard normal.

Definition 2.3 ([8] Archimedean copula)**.** *Let $\varphi : [0,1] \times \Theta \to [0,\infty)$ be a continuous, strictly decreasing, and convex function with $\varphi(1) = 0$, parametrized by some parameter $\theta \in \Theta$. Then the Archimedean copula with generator function φ and parameter θ is*

$$C_\varphi(u_1, \ldots, u_d; \theta) = \varphi^{-1}(\varphi(u_1;\theta) + \cdots + \varphi(u_d;\theta);\theta), \quad u_1, \ldots, u_d \in [0,1],$$

where φ^{-1} denotes the pseudo-inverse of φ given by

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

The class of Archimedean copulas is large, and includes the Clayton, Gumbel, Frank and Joe copulas. Archimedean copulas have the distinct advantage over elliptical copulas that they are able to model asymmetric dependence, i.e. different dependence between variables in the lower/upper tails.

Example 2.3 (Bivariate Clayton copula)**.** *The generator function of the Clayton copula is given by $\varphi(x;\delta) = \frac{1}{\delta}(x^{-\delta} - 1)$, for $\delta \in [0,\infty)$. The CDF of the bivariate Clayton copula is*

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

Its conditional distribution and conditional quantile function are

$$C_{2|1}(v \mid u) = \left(u^{-\delta} + v^{-\delta} - 1\right)^{-1/\delta - 1} v^{-\delta - 1}, \quad u, v \in (0,1) \text{ and}$$
$$C_{2|1}^{-1}(\alpha \mid u) = \left(\left(\alpha^{-\frac{\delta}{1+\delta}} - 1\right) u^{-\delta} + 1\right)^{-\frac{1}{\delta}}, \quad u, \alpha \in (0,1).$$

Its upper and lower tail coefficients are

$$\lambda_U = 0, \quad \lambda_L = 2^{\frac{-1}{\delta}}.$$

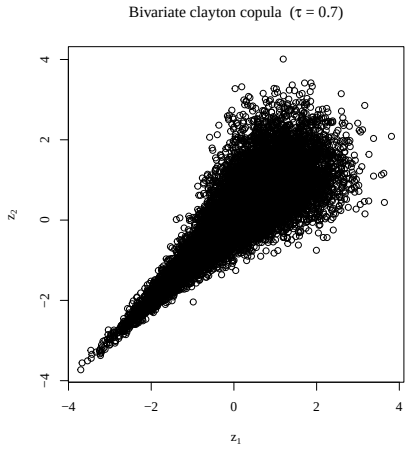


Figure 2.3: Each point represents one out of $n = 20000$ simulated observations of a bivariate Clayton copula with parameter $\delta = 4.67$, or equivalently with Kendall's $\tau = 0.7$. The marginal distributions were chosen standard normal.

Example 2.4 (Bivariate Gumbel copula)**.** *The generator function of the Clayton copula is given by* $\varphi(x; \delta) = -(\log(x))^{\delta}$*, for* $\delta \in [1, \infty)$*. The CDF of the bivariate Gumbel copula is*

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

Its conditional distribution is

$$C_{2|1}(v \mid u) = \frac{1}{u}(-\log u)^{\delta - 1}\left((-\log u)^{\delta} + (-\log v)^{\delta}\right)^{\frac{1-\delta}{\delta}} C_G(u, v), \quad u, v \in (0,1).$$

The Gumbel copula has no analytical expression for its conditional quantile function, it needs to be computed numerically by root-finding

$$C_{2|1}(v \mid u) - \alpha = 0, \quad u, \alpha \in (0,1)$$

Its upper and lower tail coefficients are

$$\lambda_U = 2 - 2^{\frac{1}{\delta}}, \quad \lambda_L = 0.$$

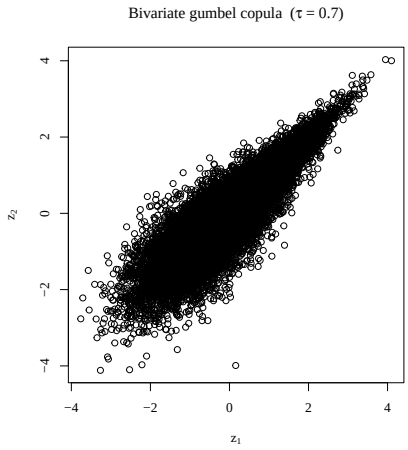


Figure 2.4: Each point represents one out of $n = 20000$ simulated observations of a bivariate Gumbel copula with parameter $\delta = 3.33$, or equivalently with Kendall's $\tau = 0.7$. The marginal distributions were chosen standard normal.

Example 2.5 (Bivariate Frank copula)**.** *The generator function of the Frank copula is given by* $\varphi(x;\delta) = -\log\left(\frac{e^{-\delta x}-1}{e^{-\delta}-1}\right)$*. The CDF of the bivariate Frank copula is*

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

Its conditional distribution and conditional quantile function are

$$C_{2|1}(v \mid u) = -\frac{1}{\delta}\log\left(1 - \frac{(1-e^{-\delta v})(1-e^{-\delta u})}{1-e^{-\delta}}\right), \quad u, v \in (0,1) \text{ and}$$
$$C^{-1}_{2|1}(\alpha \mid u) = -\frac{1}{\delta}\log\Big(1 - (1-e^{-\delta u})(1-e^{-\delta})e^{-\delta\alpha}\Big), \quad u, \alpha \in (0,1).$$

Its upper and lower tail coefficients are

$$\lambda_U = 0, \quad \lambda_L = 0.$$

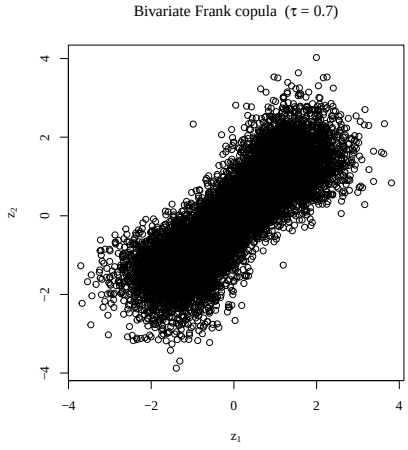


Figure 2.5: Each point represents one out of $n = 20000$ simulated observations of a bivariate Frank copula with parameter $\delta = 11.42$, or equivalently with Kendall's $\tau = 0.7$. The marginal distributions were chosen standard normal.

Example 2.6 (Bivariate Joe copula)**.** *The generator function of the Joe copula is given by* $\varphi(x;\theta) = -\log\left(1-(1-x)^{\theta}\right)$*. The CDF of the bivariate Joe copula is*

$$C_J(u_1,u_2;\theta) = 1 - \left[(1-u_1)^{\theta} + (1-u_2)^{\theta} - (1-u_1)^{\theta}(1-u_2)^{\theta}\right]^{\frac{1}{\theta}}, \quad u_1,u_2 \in [0,1].$$

Its conditional distribution is

$$C_{2|1}(v \mid u) = (1-C_J(u,v))^{1-\theta}\,(1-u)^{\theta-1}\left(1-(1-v)^{\theta}\right), \quad u,v \in (0,1) \text{ and}$$

The Joe copula has no analytical expression for its conditional quantile function, it needs to be computed numerically by root-finding

$$C_{2|1}(v \mid u) - \alpha = 0, \quad u,\alpha \in (0,1).$$

Its upper and lower tail coefficients are

$$\lambda_U = 2 - 2^{\frac{1}{\delta}}, \quad \lambda_L = 0.$$

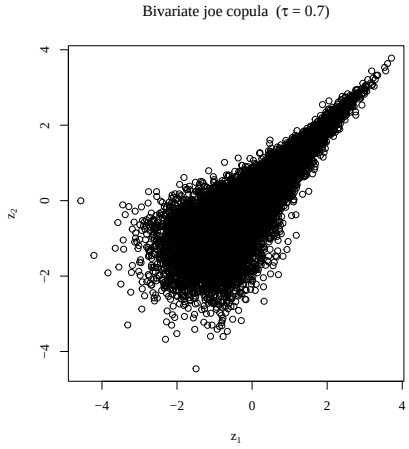


Figure 2.6: Each point represents one out of $n = 20000$ simulated observations of a bivariate Joe copula with parameter $\theta = 5.46$, or equivalently with Kendall's $\tau = 0.7$. The marginal distributions were chosen standard normal.

An overview of the CDF ($C(u,v)$), conditional distribution function ($C_{2|1}(v \mid u)$) and continuous quantile function ($C^{-1}_{2|1}(\alpha \mid u)$) of the Gaussian, Student t, Clayton, Gumbel, Frank and Joe copulas can also be found in Appendix 7.1.

2.2.3 Modeling discrete variables with copulas

For any set of variables Sklar's theorem guarantees uniqueness of the copula $C : [0,1]^d \to [0,1]$ on the Cartesian product of the ranges of the marginals (see Theorem 2.1). In the case of continuous marginal distributions, the range of the marginal is always the entirety of the interval $[0,1]$, so the Cartesian product of ranges is identical to the domain of the copula, thus implying complete uniqueness of the copula. The range of the marginal distribution of a discrete variable is not the entirety of the interval $[0,1]$, but some subset of disconnected values, e.g. $\{0, 0.5, 1\}$. One can see now that in a scenario with at least one discrete variable in the joint distribution, the copula is no longer unique on the copula domain $[0,1]^d$. Moreover, copula-based modeling with discrete variables is no longer margin-free since the marginal distribution of the discrete variable determines the range on which the copula is unique. Modeling discrete variables with copulas comes with a set of challenges, mainly due to this non-uniqueness of the copula. A detailed explanation of the challenges of modeling discrete data with copulas can be found in [9].

2.3 Vine copulas

When modeling multivariate ($d > 2$) data, the known closed-form multivariate copulas have a glaring weakness: they are inflexible. Elliptical multivariate copulas are unable to model asymmetric (different upper/lower tail) dependence between variables, and Archimedean copulas model the same dependence structure between every pair of variables. Flexibility of multivariate copula-based modeling can be improved through so-called *pair-copula constructions*. Pair-copula constructions are decompositions of a joint density into bivariate copulas. By modeling a joint density with many bivariate copulas instead of with a single multivariate copula, we regain flexibility in our models because we are free to choose a different dependence structure for each bivariate copula.

Before defining vines, we present a decomposition into bivariate copulas (i.e. a pair-copula construction) of a joint density which motivated the development of vines. This decomposition is originally due to Joe [10]. In our explanation, we will follow an approach similar to Aas et al. [11].

2.3.1 Decomposing a joint density into bivariate copulas

In this subsection we will build up to the general form of a decomposition of any multivariate joint density into the product of bivariate copulas and marginal densities. This decomposition will motivate the introduction of *vine copulas*, the framework used in vine-based regression.

We combine the joint density expression in Sklar's theorem (2.1) with Bayes theorem to obtain an expression for the conditional density as a function of marginals and the copula density. In two dimensions, we obtain

$$f(x_1 \mid x_2) = c_{12}(F_1(x_1), F_2(x_2); \theta) f_1(x_1). \tag{2.4}$$

We can expand this to a conditioning vector $\mathbf{v}$ of arbitrary dimensions,

$$f(x \mid \mathbf{v}) = c_{xv_j|\mathbf{v_{-j}}}(F(x \mid \mathbf{v_{-j}}), F(v_j \mid \mathbf{v_{-j}}); \theta) f(x \mid \mathbf{v_{-j}}) \tag{2.5}$$

where v_j is an arbitrary element of $\mathbf{v}$ and $\mathbf{v_{-j}}$ denotes the vector $\mathbf{v}$ excluding the j-th component. For ease of writing we will drop the arguments and the parameter, and denote $c_{xv_j|\mathbf{v_{-j}}}(F(x \mid \mathbf{v_{-j}}), F(v_j \mid \mathbf{v_{-j}}); \theta)$ simply as $c_{xv_j|\mathbf{v_{-j}}}$. Throughout this thesis we will make the following assumption, namely that the parameter $\theta_{jk|D}$ of any conditional copula $C_{jk|D}(F_{j|D}, F_{k|D} \mid \theta_{jk|D})$ depends on the conditioning sets only through which variables are in the conditioning sets, but does not depend on the specific values of the variables in the conditioning set D. In other words, the copula parameter $\theta_{jk|D}$ is independent of the values of the variables in the conditioning set D. This is called the *simplifying* assumption. It allows for a more efficient parametrization and fitting of copulas, at the risk of misspecification.

Note that Equation 2.5 of the conditional density is the product of a bivariate copula, and distribution/density functions conditioned on a smaller set than the left-hand side. Therefore we can apply this identity repeatedly to factorize any conditional density as a product of bivariate copulas and marginal densities. In each step of this iteration, we have a choice regarding which variable we take out of the conditioned set, with each choice yielding a different decomposition/-parametrization. We provide an example below.

Example 2.7. *Let us decompose a conditional density $f(x_1 \mid x_2, x_3)$ into the product of bivariate copulas and marginal densities in the manner of the iterative procedure shown above. What is more, we will give two distinct decompositions of $f(x_1 \mid x_2, x_3)$ to illustrate how a different choice in each iteration leads to a different decomposition/parametrization. In every iteration, we apply Equation 2.5 to the conditional density in the expression. This gives us a choice of which variable*

to take out of the conditioning set. If we first take x_2 out of the conditioning set, we obtain

$$\begin{aligned} f(x_1 \mid x_2, x_3) &= c_{12|3}(F(x_1 \mid x_3), F(x_2 \mid x_3)) f(x_1 \mid x_3) \\ &= c_{12|3}(F(x_1 \mid x_3), F(x_2 \mid x_3)) c_{13}(F_1(x_1), F_3(x_3)) f_1(x_1). \end{aligned}$$

If we first take x_3 out of the conditioning set, we obtain

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

The arguments of the conditional copulas in the decomposition are conditional distributions of the form $F(x \mid \mathbf{v})$. We can compute these by repeatedly applying the following identity, proven by Joe [12]. For every j,

$$F(x \mid \mathbf{v}) = \frac{\partial C_{xv_j|\mathbf{v_{-j}}}(F(x \mid \mathbf{v_{-j}}), F(v_j \mid \mathbf{v_{-j}}))}{\partial F(v_j \mid \mathbf{v_{-j}})} \tag{2.6}$$

As before, notice how the conditional distribution can be expressed as a function of a bivariate copula and conditioning distributions with smaller conditioning sets. For ease of notation, we will denote these copula derivatives as h-functions

$$h_{x|\mathbf{v}}(F_{x|v_{-\mathbf{j}}}(x \mid \mathbf{v_{-j}}) \mid F_{v_j|\mathbf{v_{-j}}}(v_j \mid \mathbf{v_{-j}})) \coloneqq \frac{\partial C_{xv_j|\mathbf{v_{-j}}}(F(x \mid \mathbf{v_{-j}}), F(v_j \mid \mathbf{v_{-j}}))}{\partial F(v_j \mid \mathbf{v_{-j}})}. \tag{2.7}$$

The arguments of the h-function are conditional distributions conditioned on a smaller set than the left-hand side. Therefore, we can calculate any conditional distribution by recursively applying Equation 2.6, yielding a representation of conditional distributions as nested h-functions.

We return to the modeling of a joint density in d dimensions, $f(x_1, \ldots, x_d)$, and rewrite the joint density as a product of conditional densities.

$$f(x_1, x_2, \ldots, x_d) = f(x_d \mid x_{d-1}, \ldots, x_1) f(x_{d-1}, \ldots, x_1) = \cdots = \left(\prod_{i=2}^{d} f(x_i \mid x_{i-1}, \ldots, x_1) \right) \cdot f(x_1)$$

Remark. In this decomposition we used a sequential ordering of variables, but this is an arbitrary choice made for convenience.

When we repeatedly apply Equation 2.5 and Equation 2.6 to this decomposition, we find the joint density can be written as the product of numerous bivariate copula and the marginal distributions, a pair-copula construction. As stated before, in each application of Equation 2.5 and Equation 2.6 we have the choice of which variable to take out of the conditioning set. One can see that the number of possible pair-copula constructions grows very quickly as a function of the dimension of the joint density. To be exact, the number of possible pair-copula constructions of this kind on d variables is $\frac{d!}{2} \cdot 2^{\binom{d-2}{2}}$ [13].

2.3.2 Introduction of vine copulas

To organize the large number of possible pair-copula constructions of the joint density laid out in the previous section, Bedford & Cooke introduce *regular vines* [14].

Definition 2.4. *A regular vine (also called R-vine) on d variables is a set of nested trees $T_1 = (E_1, N_1), \ldots, T_{d-1} = (E_{d-1}, N_{d-1})$ abiding by three conditions. For $i = 1, \ldots, d-1$:*

1. *T_1 has nodes $N_1 = \{1, \ldots, d\}$ and edges E_1.*
2. *Tree T_i has nodes $N_i = E_{i-1}$.*

3. Two edges in tree T_i are joined in tree T_{i+1} if they share a common node in tree T_i.

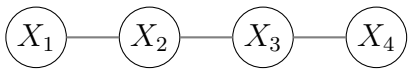
An R-vine is nothing more than a set of trees following some specific conditions, which can be used to graphically represent the previously laid out decomposition of a joint density into bivariate copulas. Because each of these possible joint density decompositions into bivariate copulas and marginal densities can be represented as a unique R-vine, these decompositions are called *vine copulas*, and we may refer to its graphical representation as an R-vine or to its representation as a decomposition interchangeably. Each tree in the R-vine represents all copulas of a certain level of conditioning. The first tree represents all unconditional copulas, the second tree represents all conditional copulas conditioned on one variable, etc. Within every tree, each edge represents a bivariate copula in the decomposition. More precisely, an edge between two nodes with index sets A and B corresponds to the copula $c_{(A\cup B\setminus A\cap B)|(A\cap B)}$. For example, an edge between nodes labeled X_{12}, X_{23} (index sets $\{1,2\}$ and $\{2,3\}$, respectively) corresponds to the copula $c_{13|2}$.

To illustrate how R-vines represent a decomposition of the joint density, we give two examples from two special classes of R-vines. The first is the class of D-vines, in which all nodes in every tree have at most two adjacent nodes. The second is the class of C-vines, in which in every tree, there is a node which is adjacent to all other nodes in that tree. These two classes of R-vines are far from exhaustive, but they are the most common classes used in the literature.

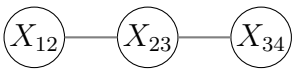
Example 2.8 (D-vine)**.** *Consider a density in 4 dimensions, $f(x_1, x_2, x_3, x_4)$. As shown earlier, we can decompose this density as*

$$f(x_1, x_2, x_3, x_4) = \big(\prod_{i=1}^{4} f(x_i)\big) c_{12}c_{23}c_{34}c_{13|2}c_{24|3}c_{14|23},$$

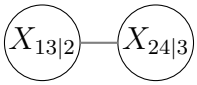
where for ease of writing we drop the inputs of the copula densities. We represent this decomposition graphically in the following D-vine.



(a) Tree T_1 in the D-vine.



(b) Tree T_2 in the D-vine.



(c) Tree T_3 in the D-vine.

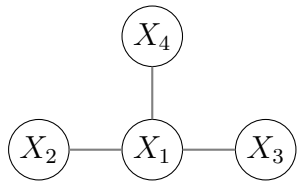
Figure 2.7: Figures of the trees in the D-vine.

We repeat that in an R-vine representation of a decomposition, it is not the nodes which represent copulas, but the edges. Each edge represents a single copula, and the specific copula it represents is determined by the indexes of its nodes. An edge between two nodes with indices A and B corresponds to the copula $c_{(A\cup B\setminus A\cap B)|(A\cap B)}$. The edges in tree T_1 represent copulas c_{12}, c_{23}, c_{34}. The edges in tree T_2 represent copulas $c_{13|2}, c_{24|3}$. The edge in tree T_3 represents the copula $c_{14|23}$. The decomposition of a D-vine is completely determined by the order of the variables in its first tree.

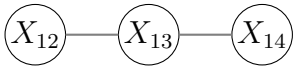
Example 2.9 (C-vine)**.** *Consider a density in 4 dimensions, $f(x_1, x_2, x_3, x_4)$. As shown earlier, we can decompose this density as*

$$f(x_1, x_2, x_3, x_4) = \big(\prod_{i=1}^{4} f(x_i)\big) c_{12}c_{13}c_{14}c_{23|1}c_{34|1}c_{24|13}.$$

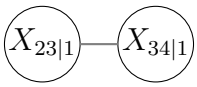
We represent this decomposition graphically in the following C-vine.



(a) Tree T_1 in the C-vine.



(b) Tree T_2 in the C-vine.



(c) Tree T_3 in the C-vine.

Figure 2.8: Figures of the trees in the C-vine.

The edges in tree T_1 represent copulas c_{12}, c_{13}, c_{14}. The edges in tree T_2 represent copulas $c_{23|1}, c_{34|1}$. The edge in tree T_3 represents the copula $c_{24|13}$.

In summary, vine copulas are highly flexible models for a joint density and consist of three components: (1) a vine structure (the lay-out of each tree) to determine the decomposition, (2) a set of copula families to fit on all copulas, and (3) a set of parameters of the copulas.

2.3.3 Extending vine copulas to discrete variables

An assumption we have made so far in the decompositions is that all variables are continuous. This need not be the case, and we will explain how vine copulas can be used to represent decompositions of multivariate probability mass functions of discrete variables, as shown by Panagiotelis et al. [15]. The extension to joint distributions of mixed-type (continuous & discrete) variables follows naturally.

In the case where $X_1, X_2, \ldots, X_d$ are discrete, the decomposition of the joint probability mass function (pmf) into conditional pmf's is identical to the continuous case.

$$\mathbb{P}(x_1, x_2, \ldots, x_d) = \left(\prod_{i=2}^{d} \mathbb{P}(x_i \mid x_{i-1}, \ldots, x_1) \right) \cdot \mathbb{P}(x_1).$$

Remark. Like in the continuous case, in this decomposition we used a sequential ordering of variables but this is an arbitrary choice made for convenience.

We can therefore extend vine copulas to discrete variables by providing alternative formulations of Equation 2.5 and Equation 2.6 for the discrete case. Let us show a discrete formulation of Equation 2.5.

$$\begin{aligned} \mathbb{P}(x \mid \mathbf{v}) &= \frac{\mathbb{P}(x, v_j \mid \mathbf{v}_{-\mathbf{j}})}{\mathbb{P}(v_j \mid \mathbf{v}_{-\mathbf{j}})} \\ &= \frac{\sum_{i_1=0}^{1} \sum_{i_2=0}^{1} (-1)^{i_1+i_2} C_{X,V_j|\mathbf{V}_{-\mathbf{j}}}(F_{X|\mathbf{V}_{-\mathbf{j}}}(x - i_1 \mid \mathbf{v}_{-\mathbf{j}}), F_{V_j|\mathbf{V}_{-\mathbf{j}}}(v_j - i_2 \mid \mathbf{v}_{-\mathbf{j}}))}{\mathbb{P}(v_j \mid \mathbf{v}_{-\mathbf{j}})} \end{aligned} \tag{2.8}$$

This gives us an expression for a conditional probability mass which requires computing the differences of some copulas with a conditioning set one smaller than the conditioning set we started with. This equation can therefore be applied recursively to obtain an expression for any conditional probability mass as differences of bivariate copulas.

The discrete alternative for Equation 2.6 is required since the arguments in the copula of Equation 2.8 again take the form of a conditional distribution.

$$F_{X|\mathbf{V}}(x \mid \mathbf{v}) = \frac{\mathbb{P}(X \leq x, V_j = v_j \mid \mathbf{v_{-j}})}{\mathbb{P}(v_j \mid \mathbf{v_{-j}})}$$
$$= \frac{C_{X,V_j|\mathbf{V_{-j}}}(F_{X|\mathbf{V_{-j}}}(x \mid \mathbf{v_{-j}}), F_{V_j|\mathbf{V_{-j}}}(v_j \mid \mathbf{v_{-j}})) - C_{X,V_j|\mathbf{V_{-j}}}(F_{X|\mathbf{V_{-j}}}(x \mid \mathbf{v_{-j}}), F_{V_j|\mathbf{V_{-j}}}(v_j - 1 \mid \mathbf{v_{-j}}))}{\mathbb{P}(v_j \mid \mathbf{v_{-j}})}. \tag{2.9}$$

One can now see how vine copulas can also be used to represent decompositions for joint probability mass functions of discrete variables, and that this naturally allows us to represent decompositions of joint densities of mixed-type variables as well. We give an example of a vine copula decomposition of a density with both continuous and discrete variables below.

Example 2.10. *We will provide a vine copula decomposition of a density of three variables X_1, X_2, X_3 (continuous, continuous, discrete, respectively). The vine copula will be a D-vine with order $X_1 - X_2 - X_3$ in the first tree. In the case of continuous variables, this would give us*

$$f_{123}(x_1, x_2, x_3) = f_1(x_1) f_2(x_2) \mathbb{P}(X_3 = x_3) c_{12} c_{23} c_{13|2}.$$

Because X_3 is discrete, the copula densities $c_{23}, c_{13|2}$ are not well-defined. A copula density is the derivative of a copula distribution w.r.t. both of its arguments (see Sklar's Theorem 2.1). Its arguments, which are marginal (conditional) distributions, are not continuous when X_3 is discrete. Instead, we look at what these copula densities represent, and find a way to compute them without taking the derivative. From Sklar's Theorem, the copula density equals $c_{23}(F_2(x_2), F_3(x_3)) = \frac{f_{23}(x_2,x_3)}{f_2(x_2)f_3(x_3)}$. The right-hand side is well-defined even when X_3 is discrete, and we can compute it as

$$\begin{aligned}
\tilde{c}_{23}(F_2(x_2), F_3(x_3)) &= \frac{f_{23}(x_2, x_3)}{f_2(x_2) f_3(x_3)} \\
&= \frac{1}{f_2(x_2) f_3(x_3)} \cdot \frac{\partial}{\partial x_2} \mathbb{P}(X_2 \leq x_2, X_3 = x_3), \quad \textit{(now use Equation 2.9 to obtain)} \\
&= \frac{1}{f_2(x_2) f_3(x_3)} \cdot \frac{\partial}{\partial x_2} \left[C_{23}(F_2(x_2), F_3(x_3)) - (C_{23}(F_2(x_2), F_3(x_3 - 1)) \right].
\end{aligned} \tag{2.10}$$

We can compute the version of $c_{13|2}$ when X_3 is discrete in a similar way,

$$\begin{aligned}
\tilde{c}_{13|2}(F_2(x_2), F_3(x_3)) &= \frac{f_{13|2}(x_1, x_3 \mid x_2)}{f_{1|2}(x_1 \mid x_2) f_{3|2}(x_3 \mid x_2)}, \quad \textit{where} \\
f_{1|2}(x_1 \mid x_2) &= \frac{f_{12}(x_1, x_2)}{f_2(x_2)} = c_{12}(F_1(x_1), F_2(x_2)) f_1(x_1), \\
f_{3|2}(x_3 \mid x_2) &= \frac{f_{23}(x_2, x_3)}{f_2(x_2)} \\
&= \frac{1}{f_2(x_2)} \cdot \frac{\partial}{\partial x_2} \left[C_{23}(F_2(x_2), F_3(x_3)) - (C_{23}(F_2(x_2), F_3(x_3 - 1)) \right] \quad \textit{(using Equation 2.9)}, \\
f_{13|2}(x_1, x_3 \mid x_2) &= \mathbb{P}(X_1 = x_1, X_3 = x_3 \mid X_2 = x_2) \quad \textit{(now use the denominator of Equation 2.8)} \\
&= \frac{\partial}{\partial x_1} \left[\mathbb{P}(X_1 \leq x_1, X_3 \leq x_3 \mid x_2) - \mathbb{P}(X_1 \leq x_1, X_3 \leq x_3 - 1 \mid x_2) \right] \\
&= \frac{\partial}{\partial x_1} \left[C_{13|2}(F_{1|2}(x_1 \mid x_2), F_{3|2}(x_3 \mid x_2)) - C_{13|2}(F_{1|2}(x_1 \mid x_2), F_{3|2}(x_3 - 1 \mid x_2)) \right].
\end{aligned}$$

The decomposition for X_3 discrete becomes

$$f_{123}(x_1, x_2, x_3) = f_1(x_1) f_2(x_2) \mathbb{P}(X_3 = x_3) c_{12} \tilde{c}_{23} \tilde{c}_{13|2}.$$

2.3.4 Estimation of vine copulas

Now that we have an understanding of what vine copula models are, we outline the procedure of estimating/fitting a vine copula model to a dataset. After modeling the marginal distribution of each variable to obtain each variable's (unconditional) copula data, fitting a vine copula is a three-step process: first (1) select a catalog of copula families which the bivariate copula building blocks in the vine will be fit to, then (2) select a vine structure (which determines the density decomposition, and thus which copulas will need to be estimated), and finally (3) choose for each copula in the vine which copula family in the catalogue fits best and estimate its parameter. We will explain each of these steps, in order of increasing complexity.

Step one is relatively simple, unless under severe computational constraints, it is usually best to keep the catalog of possible copula families as large as possible.

The details of the third step, copula selection and parameter estimation, require some explanation. We have seen that vine copulas consist of layers of trees, where each tree in the vine represents a degree of conditioning (the first tree models all unconditional dependence, the second tree models all conditional dependence conditioned on a single variable, the third models all conditional dependence conditioned on two variables, etc.). Parameter estimation of vine copulas is done sequentially, tree-by-tree, starting with the tree which models unconditional dependencies. This is necessary, because in order to estimate the parameter(s) of a conditional copula $c_{AB|\mathbf{C}}$ between two variables A, B conditioned on a vector of variables $\mathbf{C}$, we require access to the copula data of that copula. From Equation 2.5, we see that the copula data is obtained by applying conditional distributions $F_{A|\mathbf{C}}, F_{B|\mathbf{C}}$ to the data, and from Equation 2.6 we see that this requires estimation of conditional copulas conditioned on a set with cardinality one smaller than the cardinality of $\mathbf{C}$. Thus, parameter estimation of vine copulas can be done tree-by-tree, but estimation must start from the tree which models unconditional dependence. This has an important consequence; estimation errors are propagated throughout the vine copula estimation process. Minimizing this propagation of parameter estimation errors is the basis of many vine structure selection methods.

The second step in vine-copula estimation, vine structure selection, is a much more complex problem. With everything covered so far, one can see that although all possible vine-copula decompositions of a joint density model the same density, not all vine-copula decompositions are made equal. Specifically, (1) the copulas induced by different vine structures vary in the severity in which they violate the simplifying assumption and how well the copula parameters can be estimated, (2) since estimation errors are propagated in the sequential estimation of the vine, one might want to model the most important dependencies first and lastly, (3) although vine copulas model joint densities, in this thesis we are interested in the conditional density of a response given a set of covariates, i.e. 'distributional regression' or 'conditional distribution estimation'. This conditional density can always be obtained through numerical integration, which can be computationally expensive. In decompositions where the response variable is not in the conditioning set of any copula, the conditional density can be obtained without numerical integration, by simply multiplying all copulas with the response variable in the conditioned set with the marginal distribution of the response variable (see [16]). In the R-vine, this restriction on the conditioning sets is equivalent to having the node with the response variable as a leaf node in every tree. Intuitively, when the response variable is a leaf node in every tree, we can cut from each tree the node containing the response and obtain an R-vine for the covariates alone. Then Bayes' theorem allows us to evaluate the conditional density of the response given covariates without numerical integration. The following example illustrates this.

Example 2.11. *If we consider the vine shown in Figure 2.7 and imagine X_1 to be the response variable, we can obtain the density $f(x_1, x_2, x_3, x_4)$ from the vine copula as a whole and we can obtain the density for the covariates $f(x_2, x_3, x_4)$ by 'cutting off' the node containing X_1 from every tree, thereby obtaining a vine copula for the covariates X_2, X_3, X_4. Bayes' theorem*

then allows us to obtain the conditional density of the response given covariates without need for numerical integration.

On the other hand, if we imagine X_2 to be the response variable, we can not cut off every node containing X_2 from each tree in the vine because this would not result in a complete vine structure of the covariates X_1, X_3, X_4. Therefore in the scenario where X_2 is imagined as the response variable, the conditional density of the response given covariates would need to be obtained by numerical integration of the following expression

$$f(x_2 \mid x_1, x_3, x_4) = \frac{f(x_1, x_2, x_3, x_4)}{\int f(x_1, x_2^*, x_3, x_4) dx_2^*}.$$

In conclusion, the problem of finding a (near-)optimal vine structure lies in the balancing of all the aforementioned forces. We review current literature on vine structure selection in Section 3.

2.3.5 Simulating from vine copulas

The ability to simulate data from a vine copula is essential because it allows for the testing and validation of models. To simulate data from a vine copula, we use the inverse Rosenblatt transform. The inverse Rosenblatt transform maps a vector $\mathbf{U} = (U_1, \ldots, U_d)$ of independent uniform random variables to a vector $\mathbf{X}$ with some distribution $F(F_1^{-1}(U_1), \ldots, F_d^{-1}(U_d))$ for arbitrary marginals and joint distribution [17]. It is given by

$$\begin{aligned}
X_1 &= F_1^{-1}(U_1) \\
X_2 &= F_{2|1}^{-1}(U_2 \mid X_1) = F_2^{-1}(h_{2|1}^{-1}(U_2 \mid X_1)) \quad \text{(see Equation 2.7 for the definition of } h\text{-functions)} \\
&\vdots \\
X_d &= F_{d|d-1,\ldots,1}^{-1}(U_d \mid X_{d-1}, X_1) = F_d^{-1}(h_{d|d-1,\ldots,1}^{-1}(U_d \mid X_{d-1}, \ldots, X_1)).
\end{aligned}$$

The second equality in each line is the inverse of Equation 2.6. To illustrate how we can calculate these (inverse) conditional distributions with h-functions recursively as claimed, let us show the recursion for some arbitrary distribution function $F_{1|23}$. By repeatedly applying Equation 2.6, we obtain the expression

$$\begin{aligned}
F_{1|23}(X_1 \mid X_2, X_3) &= h_{1|23}(F_{1|3}(X_1 \mid X_3) \mid F_{2|3}(X_2 \mid X_3)) \\
&= h_{1|23}(h_{1|3}(F_1(X_1) \mid F_3(X_3)) \mid h_{2|3}(F_2(X_2) \mid F_3(X_3))).
\end{aligned}$$

The inverse then becomes computable using h-functions as well and is equal to

$$F_{1|23}^{-1}(U_1 \mid X_2, X_3) = F_1^{-1}(h_{1|3}^{-1}(h_{1|23}^{-1}(U_1 \mid F_3(X_3)) \mid h_{2|3}(F_2(X_2) \mid F_3(X_3)))).$$

In the case where at least one of the variables is discrete, we need to handle the simulation a bit more carefully. To simulate from a vine copula with discrete variables is a more tricky procedure than the continuous case shown above because the conditional distributions are no longer nested h-functions, they also contain differences of copula distribution functions (as can be seen from Equation 2.9). We solve this problem by simulating copula data from a vine copula with only continuous variables with the above procedure, and interpret that copula data as the latent variables which generate the discrete variables. This interpretation makes the procedure for simulating from vine copulas of continuous variables the same as for vine copulas of discrete (or mixed-type) variables, but it also makes it harder to justify choosing a discrete marginal distribution which can take infinitely many values (since the copula data will be computed at finite precision, limiting the actual values the discrete variable can take). This is resolved by limiting our choice of discrete marginal distributions to those taking a finite number of values.

2.4 Variable selection measures

After fitting a vine copula model to some data, we obtain an estimate of the conditional distribution of the response variable given the covariates. To understand the relation between this estimate and the true conditional distribution, we compare the estimated conditional distribution to observations of the response variable with a *performance measure* which quantifies their similarity.

As will be seen in the literature review (Section 3), these performance measures are often also used as *variable selection measures* during the vine structure selection step of the vine copula estimation process. Many vine structure selection methods rely on comparing performance measures of different possible vine structures in order to choose the most favourable structure for conditional distribution estimation. In this thesis, we focus on variable selection rather than performance quantification, so from now on we will refer to these measures as variable selection measures.

The conditional log-likelihood (CLL) is a natural first choice for a variable selection measure. Although many readers will be very familiar with the CLL, we repeat the definition of the for the sake of completeness. Let $f_{Y|X_1,\dots,X_d}$ denote the conditional density of the response variable Y given the covariates $X_1,\dots,X_d$. The conditional log-likelihood over some data $\mathcal{D} = \{(Y_1, X_{1,1},\dots,X_{d,1}),\dots,(Y_n, X_{1,n},\dots,X_{d,n})\}$ is given by

$$\mathcal{L}(f_{Y|X_1,\dots,X_d} \mid \mathcal{D}) \coloneqq \sum_{i=1}^{n} \log(f_{Y|X_1,\dots,X_d}(Y_i \mid X_{1,i},\dots,X_{d,i})).$$

The CLL is a good measure of overall model performance, but in many areas of application (e.g. healthcare/financial risk modeling) accurate modeling of the tails of the distribution in order to obtain, say, an accurate estimate of the 99-th quantile of the conditional distribution is more important than overall fit. To assess model fit on a specific α-quantile, we use the *check-loss* function, given by

$$\rho_\alpha(u) \coloneqq \begin{cases} \alpha|u|, & u \geq 0 \\ (1-\alpha)|u|, & u < 0 \end{cases} = u(\alpha - 1_{\{u<0\}})$$

Substituting u for the estimation error $Y - \tilde{Y}_\alpha$ (where $\tilde{Y}_\alpha$ is the estimate of the α-th quantile of Y, note this is a constant estimate, i.e. without covariates), Koenker starts his 1978 paper by stating that the minimizer over all data points of the check-loss function over $\tilde{Y}_\alpha$ is equal to the α-th sample quantile of Y [18] (stating also that he was not the first to show this). One can think of minimizing the check-loss function as aiming to balance positive and negative estimation errors, in the case where the weights for positive and negative errors are α and $1-\alpha$, respectively. Note that over all data points, a check-loss value of zero is unattainable when Y is not a constant. The minimum value of the sum of $\rho_\alpha(Y - \tilde{Y}_\alpha)$ over all data points is generally unknown, but check-loss values can nevertheless be used to compare two models on a specific quantile.

3 | Literature review

In this section, we review current literature on vine-based regression methods. We cover some of the main topics in current research which are relevant to the contents of this thesis: vine structure selection, improving computational complexity of vine-based regression, and vine-based regression for discrete variables.

Vine structure selection

Vine structure selection is an active research area. The number of R-vine structures is vast, and heuristics are necessary to search for the structure which provides the best fitting model.

Dissmann [19] introduced a greedy algorithm for R-vine structure selection. The algorithm chooses each tree structure sequentially based on a maximum spanning tree (MST) maximizing the sum of absolute empirical Kendall's τ among all copulas in the tree. This is based on the idea of fitting the strongest dependencies first. This greedy structure is justified by the fact that the lowest trees have the greatest influence on the overall model fit, and since estimation errors are propagated one wants to estimate the most important copulas early in the estimation procedure. In Dissmann's paper, Kendall's τ was chosen as a metric for strength of dependence, but other metrics for the strength of dependence have been tested as well. Czado et al. used the Akaike Information Criterion (AIC) for each pair-copula as a metric for strength of dependence [20]. This is more expensive in higher dimensions than Kendall's τ, since it requires maximizing a likelihood for each pair-copula family. A notable flaw in these metrics for strength of dependence is that they do not take the effect of the simplifying assumption into account. Two algorithms which combine Kendall's τ with tests for severity of violation of the simplifying assumption are given and tested in [21]. The authors show that their algorithm often provides a better model fit, at the expense of computational cost. They suggest to use Dissmann's algorithm, check whether the number of pair-copulas which violate the simplifying assumption is greater than expected, and repeat vine structure selection using their more computationally intensive method if this is the case. Following this, Brechmann & Joe show Dissmann's method can be improved by additionally searching in a neighbourhood of each MST [22].

In a novel approach, Chang et al. adapt a Monte Carlo Tree Search (MCTS) algorithm for vine structure learning [23]. This method seems to find a well-fitting structure, outperforming both Dissmann and Brechmann & Joe in low- and high-dimensional settings, although the difference in performance becomes smaller as the set truncation level grows larger (as the truncation level grows larger, the vines eventually become regular, untruncated vines). The computational complexity of the algorithm is significantly greater than the previously mentioned alternative algorithms.

All methods discussed so far construct the vine sequentially, starting from the first tree. In 2011, Kurowicka published a method which works in the opposite direction, starting from the last tree [24]. The aim of their algorithm was to construct an R-vine with a maximum number of independence copulas in the last trees, i.e. a maximally truncated R-vine. The motivation for such an algorithm was the same as for Dissmann's algorithm: to force the strongest dependencies to be modeled in the first trees of the vine. In simulation however, it was observed that constructing an R-vine in the non-sequential direction (i.e the tree which models conditional

dependencies with largest conditioning sets first) is not ideal as the choice of a small partial correlation in tree T_j may severely constrain the choices available in tree T_{j-1}.

For the sake of computational complexity, we desire a vine structure for which the conditional density of the response variable is readily obtained without numerical integration. This is equivalent to restricting ourselves to vine structures in which the response variable is a leaf node in every tree. However, the algorithms mentioned thus far do not guarantee the conditional density is available without numerical integration. In [25], the authors suggest a way of getting around this problem, by using any of the aforementioned methods to find an optimal R-vine for the set of covariates, and then link the response variable to the node which yields the conditional response density while maximizing correlations with the response variable, thereby producing an R-vine which satisfies our constraint.

In a regression problem, one might consider this approach of modeling the covariates first and later adding the response variable an unnatural approach. The alternative, building a vine structure around the response variable in such a way that the conditional density is available, was first explored by Kraus & Czado in 2017, who proposed a method for D-vine regression which built a D-vine with the response variable as a leaf node, sequentially adding the covariate which maximized the conditional likelihood [26]. Note that this automatically performs feature selection as well. The method has three major limitations, (1) the vine structure is restricted to D-vines, (2) greedily maximizing the conditional likelihood does not guarantee a globally optimal likelihood and is computationally intensive, and (3) all pair-copulas considered are parametric.

Zhu et al. extended this method to R-vines [27] by introducing a heuristic which chose the R-vine in a way which maximizes the sum of partial correlations of the copulas in the R-vine structure. For univariate conditional distribution estimation, the improvement of the extension to R-vines was shown to be limited. An important distinction between the D-vine based method and the R-vine based extension is that the latter can be used for estimating multivariate conditional distributions as well.

Tepegjozova presented a modified version of Kraus & Czado's 2017 algorithm which (partially) addressed each of the three major limitations mentioned above [28]. They expanded the algorithm to include both C-vines and D-vines, made the algorithm less greedy by looking two steps ahead in the construction of the vines, and estimated all copulas nonparametrically. Moreover, their method allows for controlling the computational cost of building the vine independently of the number of explanatory variables, although the number of computations is generally high. In simulation, it was shown that Tepegjozova's algorithm outperforms Kraus & Czado's in most scenarios, and the author suggested their algorithm could be improved by combining both parametric and nonparametric estimation of copulas in an efficient way.

Computational complexity

Among the methods for vine structure selection presented in the previous section, some methods automatically include variable selection. These methods therefore provide a start-to-finish algorithm for vine-based regression. Each of these methods considered thus far has computational complexity of $\mathcal{O}(d^3)$ [29], where d is the number of variables in the data. However, certain applications require faster conditional distribution estimation; for instance, in many domains of application (genomics, healthcare, finance) the data may be very high-dimensional ($d > 100$), and existing vine-based regression methods become incomputable. For this reason, decreasing the computational complexity of vine-based regression algorithms is one of the major problems in vine-based regression research.

A recent 2024 paper by Şahin & Czado provided a modification to Kraus & Czado's algorithm which improved computational complexity to $\mathcal{O}(d^2)$ [30]. The result was obtained by speeding up the variable selection process. As explained in the previous section, Kraus & Czado's algorithm chooses which variable to add to the D-vine by selecting the candidate variable which yields the highest conditional likelihood of the response variable given the explanatory variables in the

D-vine and the candidate variable. To compute this conditional likelihood, all bivariate copulas induced by appending the candidate variable to the D-vine need to be estimated (this number of copulas is equal to the number of explanatory variables already in the D-vine). The method of Şahin & Czado requires less bivariate copula estimation in the variable selection process because the variable to add is chosen through maximizing the conditional likelihood of the copula of the *residuals* of the response variable (the difference in the value of the response variable and the median as estimated by the D-vine constructed so far) and the candidate variable. To compute this conditional likelihood, only a single bivariate copula needs to be estimated. The algorithms by Kraus & Czado and Şahin & Czado were compared in a simulation study, which concluded Şahin & Czado's algorithm is not only computationally faster, but outperformed in eight out of nine settings simulated.

Vine-based regression with discrete variables

As explained in Section 2.3.3, Panagiotelis et al. introduced vine copulas for modeling multivariate discrete data [15] and later presented a model selection method for discrete vine copulas which modified Dissmann's algorithm to select maximum spanning trees based on a modified AIC, as opposed to Kendall's τ [31].

Nagler proved that under mild restrictions on the type of noise being added, one can estimate a joint distribution of mixed-type data with an estimation of the joint distribution of continuous data (obtained by adding continuous noise to the discrete variables, also called jittering) [32]. The latter estimation should be done nonparametrically since the convoluted discrete data can usually not be appropriately captured by parametric forms. A comparison of nonparametric copula estimators did not find any estimator to be uniformly better than all others, so the choice should be made with context (i.e. strength of dependence, tail dependence) in mind [33].

Schallhorn et al. adapted the D-vine regression of Kraus & Czado ([26]) to allow for regression on mixed-type data [34]. Schallhorn followed the scheme of Kraus & Czado, but used the aforementioned continuous convolutions by Nagler and the nonparametric copula estimator proposed in [35]. In simulations this method was shown to provide fast and accurate estimation compared to other well-known quantile regression methods (linear/boosted additive/nonparametric quantile regression). Jittering has some drawbacks, however. First, the results are dependent on the choice of jittering distribution, therefore results may need to be averaged over many runs with different jittering distributions to converge to an average. Second, the jittered data does not necessarily have the same dependence structure as the original discrete data. Especially for discrete variables with few categories, jittering can lead to biased correlation estimates [36]. Thus, there is still a need for vine-based regression methods for mixed-type data which do not rely on jittering.

4 | (Discretized) conditional quantile functions of (bivariate) copulas

This chapter investigates how the conditional quantile function of a bivariate copula changes when the marginal distribution of the covariate is discrete rather than continuous, and how this affects vine-based regression methods. First, we visualize and compare (empirically) continuous conditional quantiles and discrete conditional quantiles. After that, we attempt to quantify the difference between continuous conditional quantiles and discrete conditional quantiles analytically.

We will investigate the conditional quantile of a continuous response variable Y given a covariate X, whose joint distribution is governed by a single-parameter copula (Gaussian/Clayton/Gumbel/Frank/Joe copula). In the continuous case, the variables Y and X are assumed to have standard normal marginal distributions, so as to more easily see conditional quantile curves in the tails. In the discretized case, Y remains standard normally distributed but the discretized X is generated by binning continuous X into b bins of equal probability. Throughout this thesis, we will often generate discretized variables by binning a latent variable into bins of equal probability, so for ease of reference we give a formal definition.

Definition 4.1 (Discretized variable)**.** *For a continuous random variable X, we generate a new, discrete variable 'discretized X' with b values by separating X into b bins of equal probability. Let $0 = c_0 < c_1 < \cdots < c_b = 1$ denote the quantile-boundaries. Discretized X takes value k when $X \in B_k \coloneqq (F_X^{-1}(c_{k-1}), F_X^{-1}(c_k)]$. We impose that all bins are equally probable, so by uniformity of the copula data this means the quantile-boundaries are evenly spaced out in the interval $[0, 1]$. For example, for a discretization into 4 bins, the set of quantile-boundaries would be $\{0, 0.25, 0.5, 0.75, 1\}$.*

$$X_{discretized} \coloneqq \begin{cases} 1, & X \in B_1 \\ 2, & X \in B_2 \\ \vdots & \\ b, & X \in B_b. \end{cases}$$

Before investigating the effect of discretization of the covariate on conditional quantiles, we briefly explain how the continuous and discretized conditional quantile function are obtained through copulas. When the covariate is continuous, we use that the continuous conditional α-quantile of Y given X is

$$F_{Y|X}^{-1}(\alpha \mid X = x) = F_Y^{-1}\left(C_{Y|X}^{-1}(\alpha \mid F_X(x))\right), \quad \alpha \in (0, 1),$$

which follows directly from $F_{Y|X}(Y \leq y \mid X = x) = C_{Y|X}(F_Y(y) \mid F_X(x))$, where the latter is the following

$$
\begin{aligned}
C_{Y|X}(u_Y \mid u_X) &= \mathbb{P}(F_Y(Y) \leq u_Y \mid F_X(X) = u_X) && \text{(where } F_Y(Y), F_X(X) \sim \text{Unif}[0,1]\text{)} \\
&= \frac{\mathbb{P}(F_Y(Y) \leq u_Y, F_X(X) = u_X)}{\mathbb{P}(F_X(X) = u_X)} \\
&= \frac{\frac{\partial}{\partial u_X}\mathbb{P}(F_Y(Y) \leq u_Y, F_X(X) \leq u_X)}{\frac{\partial}{\partial u_X}\mathbb{P}(F_X(X) \leq u_X)} && \text{(the denominator equals 1; } F_X(X) \sim \text{Unif}[0,1]\text{)} \\
&= \frac{\partial}{\partial u_X}\mathbb{P}(F_Y(Y) \leq u_Y, F_X(X) \leq u_X) \\
&= \frac{\partial}{\partial u_X} C_{Y,X}(u_Y, u_X). && (4.1)
\end{aligned}
$$

The continuous conditional α-quantile ($F^{-1}_{Y|X}(\alpha \mid X = x)$) can then be found by solving $F_{Y|X}(Y \leq y \mid X = x) - \alpha = 0$ for y. For some copula families (Gaussian, Frank, Clayton), the continuous conditional quantile function has a closed-form expression. For those families for which a closed-form expression exists, the continuous conditional quantile function can be found in the table in Appendix 7.1.

Next, we formalize the notion of a 'discretized' conditional α-quantile.

Definition 4.2 (Discretized conditional quantile)**.** *Partition the support of a continuous covariate X into b bins defined by the quantile-boundaries*

$$B_k := \big(F_X^{-1}(c_{k-1}),\, F_X^{-1}(c_k)\big], \qquad k = 1, \ldots, b,$$

where $0 = c_0 < c_1 < \cdots < c_b = 1$. For any fixed $\alpha \in (0,1)$ and bin B_k, define the discretized conditional α-quantile as

$$F^{-1}_{Y|X}(\alpha \mid X \in B_k) := \inf\big\{\, y : F_{Y|X}(Y \leq y \mid X \in B_k) \geq \alpha \big\}.$$

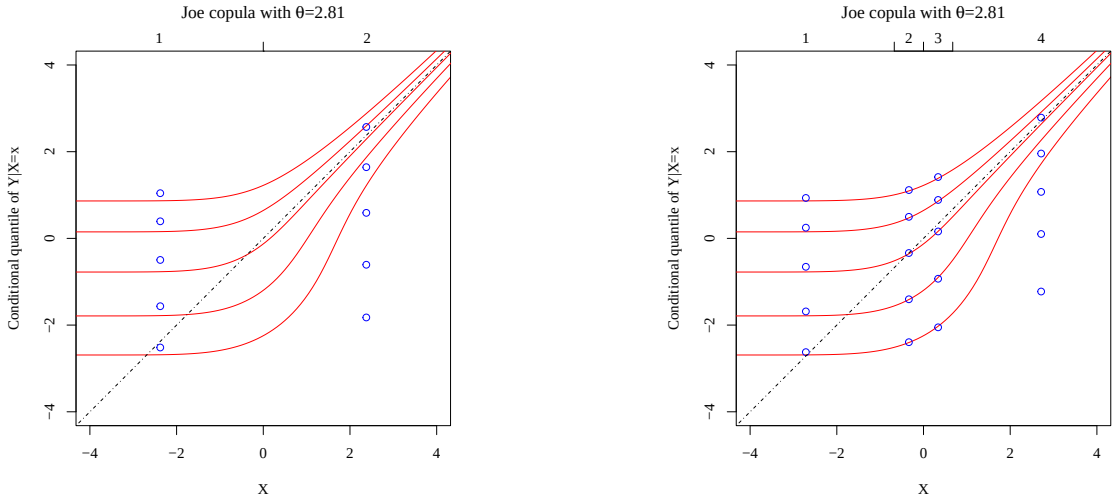
In the case where X is discretized, we need a different procedure for obtaining the conditional distribution function since it is no longer possible to take derivatives of the copula distribution function with respect to $F_X(x)$ (it is no longer continuous w.r.t. $F_X(x)$). We will provide a formulation for the conditional distribution function of Y conditioned on discretized X, then the discretized conditional quantile function is naturally its inverse. Let $k \in \{1, 2, \ldots, b\}$, and let c_{k-1}, c_k be the quantile-boundaries of the k-th bin. Then the conditional distribution function is

$$
\begin{aligned}
F_{Y|X}\left(Y \leq y \mid X \in B_k\right) &= F_{Y|X}\left(Y \leq y \mid X \in \left(F_X^{-1}(c_{k-1}), F_X^{-1}(c_k)\right]\right) \\
&= \frac{F_{YX}\left(Y \leq y, X \leq F_X^{-1}(c_k)\right) - F_{YX}\left(Y \leq y, X \leq F_X^{-1}(c_{k-1})\right)}{\mathbb{P}\left(X \leq F_X^{-1}(c_k)\right) - \mathbb{P}\left(X \leq F_X^{-1}(c_{k-1})\right)} \\
&= \frac{C_{Y,X}\left(F_Y(y), c_k\right) - C_{Y,X}(F_Y(y), c_{k-1})}{c_k - c_{k-1}}. && (4.2)
\end{aligned}
$$

Just like for the continuous conditional quantile function, the continuous discretized α-quantile can then be found by solving $F_{Y|X}(Y \leq y \mid X \in B_k) - \alpha = 0$ for y.

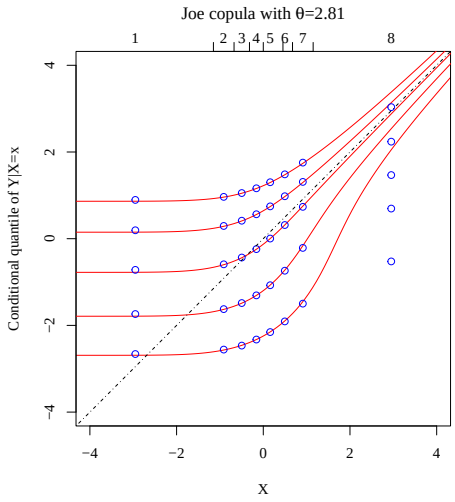
Now that we have formulations for the conditional quantile functions in both the continuous and discrete case, we can continue our investigation of the impact of discretization on conditional quantile functions. In their 2015 paper, Bernard & Czado present plots of the continuous conditional quantile function for normal marginals of Y and X, and various copula families [37]. We extend those plots to include the discretized conditional quantile function. In Figure 4.1 we plot the conditional quantiles for continuous X and discretized X of the Joe copula. The plots for Frank/Clayton/Gaussian/Gumbel copulas are found in Appendix 7.2.1. We chose to

present the Joe copula here as it illustrates the effect of discretization on the conditional quantile function when the copula does not model tail dependence (lower tail, asymptotically constant), and is strongly tail dependent (upper tail, asymptotically linear) with different number of bins for discretized X.

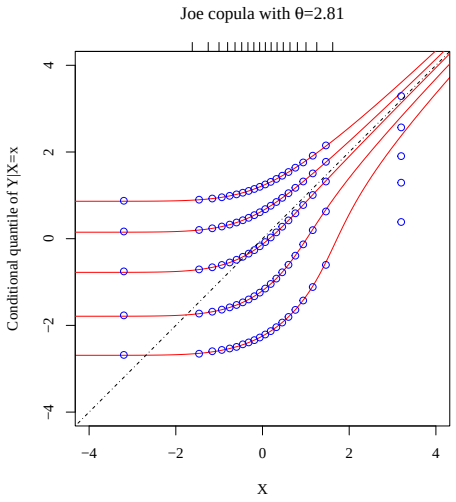


(a) $b = 2$

(b) $b = 4$



(c) $b = 8$



(d) $b = 20$

Figure 4.1: Plots of the conditional quantile function of the Joe copula with $\theta = 2.81$ (Kendall's $\tau = 0.49$), for quantiles $\alpha \in \{0.01, 0.1, 0.5, 0.9, 0.99\}$. On the lower x-axis, the conditioning value of X in the continuous case is shown, and the corresponding continuous conditional quantile functions are red lines. On the top x-axis, the conditioning value of X in the discretized case, with the corresponding discretized conditional quantile values given as blue circle in the middle of the associated bin. To illustrate, the eighth bin in plot (c) corresponds to the range of values $(1.15, \infty)$ on the lower x-axis. The values of the conditional quantiles were obtained numerically through root-finding.

Inspecting the figures for every copula (see Appendix 7.2.1) gives insight into how much information on the conditional quantiles can be lost through discretization of the covariate. As shown in Table 4.1, when $b = 2$, the conditional 0.5-quantiles in each bin for discretized X are nearly identical for all copulas.

	Gaussian ($\rho = 0.7$)	Clayton ($\delta = 1.95$)	Gumbel ($\delta = 1.97$)	Frank ($\delta = 5.63$)	Joe ($\theta = 2.81$)
bin 1 ($b = 2$)	-0.54	-0.59	-0.52	-0.56	-0.50
bin 2 ($b = 2$)	0.54	0.50	0.56	0.56	0.59
bin 1 ($b = 4$)	-0.88	-1.06	-0.79	-0.86	-0.66
bin 2 ($b = 4$)	-0.23	-0.16	-0.27	-0.28	-0.34
bin 3 ($b = 4$)	0.23	0.33	0.19	0.28	0.16
bin 4 ($b = 4$)	0.88	0.67	0.98	0.86	1.07

Table 4.1: Discretized conditional 0.5-quantile values of Gaussian ($\rho = 0.7$), Clayton ($\delta = 1.95$), Gumbel ($\delta = 1.97$), Frank ($\delta = 5.63$) and Joe ($\theta = 2.81$) copulas, for number of bins $b = 2$ and $b = 4$. The parameter of each copula was chosen such that each copula has Kendall's $\tau = 0.49$. The marginal distribution of Y is standard normal. The values of the discretized conditional quantiles were obtained numerically through root-finding.

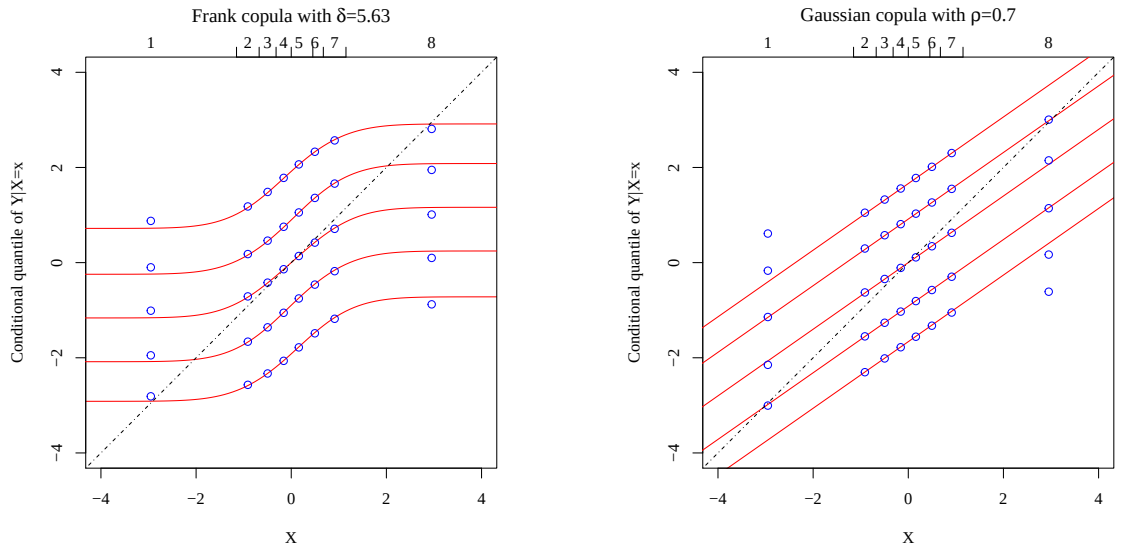
The difference between discretized conditional quantile values in the same bin for different copulas is very small, illustrating how discretization of a continuous covariate into a small ($b \leq 4$) number of bins of equal probability loses a lot of the characteristics of the dependency of the response with the continuous covariate. For a closer look at the robustness of this conclusion to different marginal distributions of Y, we give the same table for Y uniform. Conditional quantile plots with Y uniform can be found in Appendix 7.2.2.

	Gaussian ($\rho = 0.7$)	Clayton ($\delta = 1.95$)	Gumbel ($\delta = 1.97$)	Frank ($\delta = 5.63$)	Joe ($\theta = 2.81$)
bin 1 ($b = 2$)	0.29	0.28	0.30	0.29	0.31
bin 2 ($b = 2$)	0.71	0.69	0.71	0.71	0.72
bin 1 ($b = 4$)	0.19	0.14	0.21	0.20	0.26
bin 2 ($b = 4$)	0.41	0.44	0.39	0.39	0.37
bin 3 ($b = 4$)	0.59	0.63	0.58	0.61	0.56
bin 4 ($b = 4$)	0.81	0.75	0.84	0.80	0.86

Table 4.2: Discretized conditional 0.5-quantile values of Gaussian ($\rho = 0.7$), Clayton ($\delta = 1.95$), Gumbel ($\delta = 1.97$), Frank ($\delta = 5.63$) and Joe ($\theta = 2.81$) copulas, for number of bins $b = 2$ and $b = 4$. The parameter of each copula was chosen such that each copula has Kendall's $\tau = 0.49$. The marginal distribution of Y is uniform on $[0, 1]$. The values of the discretized conditional quantiles were obtained numerically through root-finding.

For Y uniform, we find again that discretization into a small ($b \leq 4$) number of bins of equal probability makes different copula families nearly indistinguishable.

As b increases, discretized conditional quantile values for different copulas become easier to distinguish, but even then certain copulas give remarkably similar discretized conditional quantiles, such as the Frank and Gaussian copulas (see Figure 4.2 below), despite them having different asymptotic conditional quantile behaviours (asymptotically constant and weakly linear, respectively) [25].



(a) $b = 8$ **(b)** $b = 8$

Figure 4.2: Plots of the conditional quantile function of the Frank and Gaussian copula, both with parameter such that Kendall's $\tau = 0.49$, for quantiles $\alpha \in \{0.01, 0.1, 0.5, 0.9, 0.99\}$. On the lower x-axis, the conditioning value of X in the continuous case is shown, and the corresponding continuous conditional quantile functions are red lines. On the top x-axis, we have the conditioning value of X in the discretized case, with the corresponding discretized conditional quantile values given as blue circle in the middle of the associated bin. The values of the conditional quantiles were obtained numerically through root-finding.

In short, discretization of a covariate into a small number of bins of equal probability causes the loss of significant dependence characteristics, to the point where different copulas may become indistinguishable.

In the conditional quantile plots, as the number of bins increases the discretized conditional quantiles become more accurate around the median of the covariate relatively quickly, but even for a large number of bins ($b = 20$) tail dependence is not captured well. This is problematic, considering the fact that vine-based regression is often favoured over other quantile regression methods specifically for its ability to flexibly capture tail dependence. For scenarios where the discretization can be chosen by the researcher, such as for building privacy-preserving datasets in healthcare, the degree to which tail dependence is lost through discretization may be improved through a different approach to binning. One could make the bins around the median conditioning value of X larger, accepting that this introduces some more error when conditioning on a value close to the median of X, but obtaining more accurate discretized conditional quantile values on the tails in return.

For a proof-of-concept, we recreate Table 4.1 with a different approach to binning, making the bins around the median wider. For instance, for $b = 4$, bins of equal probability would have quantile-boundaries $\{0, 0.25, 0.5, 0.75, 1\}$. We could discretize X in a way which preserves more of the dependence characteristics between Y and X in the tails of X, by choosing quantile-boundaries $\{0, 0.1, 0.5, 0.9, 1\}$. By preserving more dependence characteristics in the tails, different copula families should become more easily distinguishable based on discretized conditional quantiles, see the table below.

	Gaussian ($\rho = 0.7$)	Clayton ($\delta = 1.95$)	Gumbel ($\delta = 1.97$)	Frank ($\delta = 5.63$)	Joe ($\theta = 2.81$)
bin 1 ($b = 4$)	-1.22	-1.57	-1.04	-1.04	-0.74
bin 2 ($b = 4$)	-0.39	-0.39	-0.40	-0.45	-0.44
bin 3 ($b = 4$)	0.39	0.44	0.38	0.45	0.40
bin 4 ($b = 4$)	1.22	0.76	1.47	1.04	1.58

Table 4.3: Discretized conditional 0.5-quantile values of Gaussian ($\rho = 0.7$), Clayton ($\delta = 1.95$), Gumbel ($\delta = 1.97$), Frank ($\delta = 5.63$) and Joe ($\theta = 2.81$) copulas, for number of bins $b = 4$. The quantile boundaries of the bins were chosen $\{0, 0.1, 0.5, 0.9, 1\}$. The parameter of each copula was chosen such that each copula has Kendall's $\tau = 0.49$. The marginal distribution of Y is standard normal. The values of the discretized conditional quantiles were obtained numerically through root-finding.

With this tail-favoured binning procedure, the discretized conditional quantiles of the studied copula families are much easier to distinguish than the discretized conditional quantiles under the equal probability binning procedure. For instance, note how in Table 4.3 the Gaussian and Frank copulas are now easy to distinguish by the discretized conditional quantiles in their lowest and highest bins, whereas before they were nearly identical (see Table 4.1).

We have seen that although dependence around the median of the variable which generates the discretized covariate is captured well even for a moderate number of bins, tail dependence is not well-captured unless the binning procedure is tailored specifically to capturing tail-dependence. For application of discretization in building privacy-preserving datasets, we would want to be able to compute an 'optimal choice of binning' (i.e. quantile boundaries) which retains the most important dependence characteristics. Although we will not go so far as to solve this problem, we aim to take a small step towards solving this 'optimal choice of binning' by quantifying the difference between some continuous conditional α-quantile and its corresponding discretized conditional α-quantile,

$$F_{Y|X}^{-1}\left(\alpha \mid X \in \left(F_X^{-1}(c_{k-1}), F_X^{-1}(c_k)\right]\right) - F_{Y|X}^{-1}(\alpha \mid X = x), \quad \left(x \in (F_X^{-1}(c_{k-1}), F_X^{-1}(c_k)\right).$$

This quantification may be simplified if we can express the discretized conditional α-quantile as the continuous conditional α-quantile conditioned on some continuous conditioning value. The following lemma shows the existence of such a conditioning value.

Lemma 4.1. *Let Y, X be two random variables with continuous marginal distributions F_Y, F_X, governed by a copula C with continuously differentiable CDF $C_{Y,X}$. Let X be discretized into b bins where the k-th bin has quantile-boundaries c_{k-1}, c_k, i.e.*

$$B_k := \left(F_X^{-1}(c_{k-1}),\ F_X^{-1}(c_k)\right], \qquad k = 1, \ldots, b.$$

The discretized conditional α-quantile in the k-th bin $F_{Y|X}^{-1}(\alpha \mid X \in B_k)$ is equal to the conditional α-quantile $F_{Y|X}^{-1}(\alpha \mid X = x^*)$ for some $x^* \in B_k$. Moreover, this x^* is unique.

Proof. As shown in Equation 4.1, the conditional distribution for Y given X in terms of their copula $C_{Y,X}$ is defined as

$$C_{Y|X}\left(F_Y(y) \mid F_X(x)\right) = \frac{\partial}{\partial F_X(x)} C_{Y,X}(F_Y(y), F_X(x)).$$

By the assumption of continuous differentiability, $C_{Y|X}\left(F_Y(y) \mid F_X(x)\right)$ is continuous.

The discretized conditional α-quantile $F^{-1}_{Y|X}(\alpha \mid X \in B_k)$ is equal to y such that

$$\begin{aligned}
& F_{Y|X}(y \mid X \in B_k) = \alpha \\
\iff & \frac{C_{Y,X}(F_Y(y), c_k) - C_{Y,X}(F_Y(y), c_{k-1})}{c_k - c_{k-1}} = \alpha \quad \text{(using Equation 4.2)} \\
\iff & \int_{c_{k-1}}^{c_k} \frac{\partial}{\partial u_X} C_{Y,X}(F_Y(y), u_X) \mathrm{d}u_X = \alpha(c_k - c_{k-1}) \\
\iff & \int_{c_{k-1}}^{c_k} C_{Y|X}(F_Y(y) \mid u_X) \mathrm{d}u_X = \alpha(c_k - c_{k-1}).
\end{aligned}$$

By continuity of the integrand and the mean-value theorem for integrals, for any y there exists some

$$u_X^* \in (c_{k-1}, c_k] \Rightarrow x^* = F_X^{-1}(u_X^*) \ \in \ B_k,$$

such that

$$\int_{c_{k-1}}^{c_k} C_{Y|X}(F_Y(y) \mid u_X) \mathrm{d}u_X = C_{Y|X}(F_Y(y) \mid u_X^*)(c_k - c_{k-1}) = \alpha(c_k - c_{k-1}),$$

which implies

$$F^{-1}_{Y|X}(\alpha \mid X \in B_k) = F^{-1}_{Y|X}(\alpha \mid X = x^*).$$

Thus the discretized conditional α-quantile on the bin B_k coincides with the continuous conditional α-quantile at some intermediate covariate value x^*. This shows existence, and by continuity and monotonicity of the continuous conditional quantile function also uniqueness of x^*. □

The previous lemma shows existence of a conditioning value x^* such that the discretized conditional quantile in some bin is equal to the continuous conditional quantile function at x^*. This x^* which allow for the discretized conditional quantile to be 'pinned' onto the continuous conditional quantile theoretically exist for every copula family and every bin, but an analytical expression is not always available. From now on, we will define x^* as a function of copula family, bin, and quantile.

Definition 4.3. *Let Y and X be continuous random variables with marginal distribution functions F_Y, F_X, and assume their dependence is modeled by some copula family C with continuously differentiable CDF $C_{Y,X}$. Let X be discretized into b bins where the k-th bin has quantile-boundaries c_{k-1}, c_k, i.e.*

$$B_k = \big(F_X^{-1}(c_{k-1}),\, F_X^{-1}(c_k)\big], \qquad k = 1, \ldots, b.$$

For each $\alpha \in (0,1)$ and bin index k, define $x_C^*(\alpha, k)$ to be the value in B_k satisfying

$$F^{-1}_{Y|X}\big(\alpha \mid X \in B_k\big) = F^{-1}_{Y|X}\big(\alpha \mid X = x_C^*(\alpha, k)\big).$$

Analytical expressions for $x_C^*(\alpha, k)$ may be useful in proving asymptotic behaviour of discretized conditional quantile functions, or for finding analytical solutions to the 'optimal binning' problem for privacy-preserving discretization. The following lemma shows that in the case of bivariate Clayton and Frank dependence structures, we can find an analytical expression for $x_C(\alpha, 1)$. However, such expressions are in general not available. To the best of our knowledge, such x^* are only analytically available for bin 1 of the bivariate Clayton and Frank copulas because to obtain an expression for x^*, an analytical expression for both the discretized conditional quantile and the continuous conditional quantile need to be available.

Lemma 4.2. *Let c_1 denote the quantile-boundary for the first bin. For continuous marginal distributions F_Y, F_X, and a copula C in the bivariate Clayton or Frank copula families, we can find the unique conditioning value $x^*_C(\alpha, 1) \in (0, c_1]$ such that*

$$F^{-1}_{Y|X}(\alpha \mid X \in \left(F^{-1}_X(0), F^{-1}_X(c_1)\right]) = F^{-1}_{Y|X}(\alpha \mid X = x^*_C(\alpha, 1)),$$

i.e. we can anchor the discretized conditional quantile function in the first bin on the continuous conditional quantile function. These x^* are as follows.

Clayton: $x^*(\alpha, 1) = F^{-1}_X\left(\left(\frac{-1+\alpha^{-\delta}}{-1+\alpha^{\frac{-\delta}{\delta+1}}}\right)^{\frac{-1}{\delta}} c_1\right)$

Frank: $x^*(\alpha, 1) = F^{-1}_X\left(\frac{-1}{\delta}\log\left(\frac{1}{1-\alpha}\left(\frac{\alpha(e^{-\delta c_1}-1)}{e^{-\delta\alpha c_1}-1} - \alpha\right)\right)\right)$

Proof. The derivation of x^* follows the same three steps, no matter the copula family. The derivations of the x^* corresponding to each copula family will be in the order presented above.

First, assume a **Clayton** dependence structure with parameter δ.

Step 1: Find a closed-form expression for the discretized conditional quantile in bin 1

From our definition of a discretized conditional α-quantile (4.2):

$$\begin{aligned}
& F_{Y|X}(\alpha \mid X \in (0, c_1]) = \alpha, \quad \text{using Equation 2.9:} \\
\iff & \frac{C_{Y,X}(F_Y(y), c_1) - C_{Y,X}(F_Y(y), 0)}{c_1} = \alpha \\
\iff & C_{Y,X}(F_Y(y), c_1) = \alpha c_1, \quad \text{under Clayton dependence:} \\
\iff & \left(F_Y(y)^{-\delta} + c_1^{-\delta} - 1\right)^{\frac{-1}{\delta}} = \alpha c_1 \\
\iff & F_Y(y)^{-\delta} + c_1^{-\delta} - 1 = (\alpha c_1)^{-\delta} \\
\iff & F_Y(y) = \left((\alpha c_1)^{-\delta} - c_1^{-\delta} + 1\right)^{\frac{-1}{\delta}} \\
\iff & y = F^{-1}_Y\left(\left((\alpha c_1)^{-\delta} - c_1^{-\delta} + 1\right)^{\frac{-1}{\delta}}\right)
\end{aligned}$$

Step 2: Find a closed-form expression for the conditional quantile

We use the definition of the conditional quantile, as given in Equation 4.1. $F^{-1}_{Y|X}(\alpha \mid X = x)$ is equal to y such that:

$$\begin{aligned}
& C_{Y|X}(F_Y(y) \mid F_X(x)) = \alpha, \quad \text{under Clayton dependence:} \\
\iff & \left(F_X(x)^{-\delta} + F_Y(y)^{-\delta} - 1\right)^{\frac{-1}{\delta}-1} F_X(x)^{-\delta-1} = \alpha \\
\iff & y = F^{-1}_Y\left(\left(1 - F_X(x)^{-\delta} + (\alpha F_X(x)^{\delta+1})^{\frac{-\delta}{\delta+1}}\right)^{\frac{-1}{\delta}}\right)
\end{aligned}$$

Step 3: Solve the resulting equality

Now we can find a conditioning value x^* as desired by setting the closed-form expressions for the discretized conditional quantile and the conditional quantile equal to each other. This gives us:

$$\begin{aligned}
& \left((\alpha c_1)^{-\delta} - c_1^{-\delta} + 1\right)^{\frac{-1}{\delta}} = \left(1 - F_X(x^*)^{-\delta} + (\alpha F_X(x^*)^{\delta+1})^{\frac{-\delta}{\delta+1}}\right)^{\frac{-1}{\delta}} \\
\iff & (-1 + \alpha^{\frac{-\delta}{\delta+1}}) F_X(x^*)^{-\delta} = (-1 + \alpha^{-\delta}) c_1^{-\delta} \\
\iff & F_X(x^*)^{-\delta} = \left(\frac{-1+\alpha^{-\delta}}{-1+\alpha^{\frac{-\delta}{\delta+1}}}\right) c_1^{-\delta} \\
\iff & x^* = F^{-1}_X\left(\left(\frac{-1+\alpha^{-\delta}}{-1+\alpha^{\frac{-\delta}{\delta+1}}}\right)^{\frac{-1}{\delta}} c_1\right)
\end{aligned}$$

Assume a **Frank** dependence structure with parameter δ.
Step 1: Find a closed-form expression for the discretized conditional quantile in bin 1
We know $F^{-1}_{Y|X}\left(\alpha \mid X \in \left(F_X^{-1}(c_{k-1}), F_X^{-1}(c_k)\right]\right)$ is y such that

$$\begin{aligned}
& C_{Y,X}(F_Y(y), c_1) = \alpha c_1, \quad \text{under Frank dependence:} \\
\iff & \frac{-1}{\delta} \log\left(1 + \frac{(e^{-\delta c_1} - 1)(e^{-\delta F_Y(y)} - 1)}{e^{-\delta} - 1}\right) = \alpha c_1 \\
\iff & \frac{(e^{-\delta c_1} - 1)(e^{-\delta F_Y(y)} - 1)}{e^{-\delta} - 1} = e^{-\delta \alpha c_1} - 1 \\
\iff & e^{-\delta F_Y(y)} - 1 = \frac{(e^{-\delta} - 1)(e^{-\delta \alpha c_1} - 1)}{e^{-\delta c_1} - 1} \\
\iff & y = F_Y^{-1}\left(\frac{-1}{\delta} \log\left(\frac{(e^{-\delta} - 1)(e^{-\delta \alpha c_1} - 1)}{e^{-\delta c_1} - 1} + 1\right)\right)
\end{aligned}$$

Step 2: Find a closed-form expression for the conditional quantile
We use the definition of the conditional quantile, as given in Equation 4.1. $F^{-1}_{Y|X}(\alpha \mid X = x)$ is equal to y such that:

$$\begin{aligned}
& C_{Y|X}(F_Y(y) \mid F_X(x)) = \alpha, \quad \text{under Frank dependence:} \\
\iff & \frac{e^{-\delta F_X(x)}(1 - e^{-\delta F_Y(y)})}{-e^{-\delta} - e^{-\delta(F_X(x) + F_Y(y))} + e^{-\delta F_X(x)} + e^{-\delta F_Y(y)}} = \alpha \\
\iff & y = F_Y^{-1}\left(\frac{-1}{\delta} \log\left(1 - \frac{\alpha(1 - e^{-\delta})}{e^{-\delta F_X(x)} + \alpha(1 - e^{-\delta F_X(x)})}\right)\right) \\
\iff & y = F_Y^{-1}\left(\frac{-1}{\delta} \log\left(1 - \frac{\alpha(1 - e^{-\delta})}{(1 - \alpha)e^{-\delta F_X(x)} + \alpha}\right)\right)
\end{aligned}$$

Step 3: Solve the resulting equality
Now we can find a conditioning value x^* as desired by setting the closed-form expressions for the discretized conditional quantile and the conditional quantile equal to each other. This gives us:

$$\begin{aligned}
& 1 - \frac{\alpha(1 - e^{-\delta})}{e^{-\delta F_X(x^*)} + \alpha(1 - e^{-\delta F_X(x^*)})} = \frac{(e^{-\delta} - 1)(e^{-\delta \alpha c_1} - 1)}{e^{-\delta c_1} - 1} + 1 \\
\iff & e^{-\delta F_X(x^*)} + \alpha(1 - e^{-\delta F_X(x^*)}) = \frac{-\alpha(1 - e^{-\delta})(e^{-\delta c_1} - 1)}{(e^{-\delta} - 1)(e^{-\delta \alpha c_1} - 1)} \\
\iff & (1 - \alpha)e^{-\delta F_X(x^*)} + \alpha = \frac{\alpha(e^{-\delta c_1} - 1)}{e^{-\delta \alpha c_1} - 1} \\
\iff & x^* = F_X^{-1}\left(\frac{-1}{\delta} \log\left(\frac{1}{1 - \alpha}\left(\frac{\alpha(e^{-\delta c_1} - 1)}{e^{-\delta \alpha c_1} - 1} - \alpha\right)\right)\right)
\end{aligned}$$

□

This section investigated the amount of information lost in discretization of a covariate through examining the difference between discretized conditional quantiles and continuous conditional quantiles. We demonstrated that the right choice of 'binning' (i.e. the quantile boundaries of the discretized covariate) has the potential to retain more information on the dependence structure with the response than discretization into bins of equal probability, although an analytical solution to the 'optimal binning' problem seems out of reach.

It is clear a variable selection measure tailored to mixed-type data should be biased against discrete covariates. Moreover, this bias should be monotone decreasing with the number of bins of the discretized covariate. In the next section, we look at how commonly used variable selection measures are biased for/against discretized covariates.

5 | Variable selection measures under discretization

This section investigates how common variable selection measures (Pearson's/polyserial correlation, Kendall's tau/tau-b, conditional log-likelihood, check-loss) behave under discretization. Specifically, we compute the value of these variable selection measures with a continuous covariate and compare this to the value of the variable selection measure when computed with a discretized version of that covariate. Understanding how variable selection measures behave under discretization is an initial step towards designing a variable selection method tailored to mixed-type data with vines.

5.1 Variable selection measures under discretization, in 2D

First, we investigate the behaviour of variable selection measures in the simplest possible scenario. We will simulate a response variable Y and a single covariate X, both continuous with standard normal marginal distribution. The discretized X is generated from X with the procedure outlined in Definition 4.1. The dependence between Y and X will be governed by a Gaussian/Clayton/Gumbel/Frank/Joe copula, whose parameters will be changed so as to see the variable selection measures' biases for/against discretized variables over the range of dependence strength. The copula-parameter is determined by Kendall's τ. The range of Kendall's τ we iterate over is $\tau \in [0.20, 0.95]$, because in the lower end of the range variable selection measures barely differentiate between continuous and discretized variables, and in the upper end of the range variable selection measures such as CLL increases exponentially, which makes it harder to see what happens in the rest of the range of τ.

5.1.1 Pearson's/polyserial correlation

Figure 5.1 below plots estimated Pearson's correlation between Y and X, and estimated polyserial correlation between Y and discretized X, for a variety of copula families and a range of copula parameters.

The main takeaway from these figures is that for some copula families (Gaussian, Gumbel, Clayton & Joe for $\tau \leq 0.5$), polyserial correlation between Y and discretized X is very close to Pearson's correlation between Y and X, and for other copula families (Frank, Clayton & Joe for $\tau > 0.5$), polyserial correlation between Y and discretized X is greater than Pearson's correlation between Y and X. This means that when Pearson's/polyserial correlation is used as a variable selection tool for vine-based regression, a discretized variable generated from a latent variable with some Pearson's correlation with the response will be measured as equally good/better to add to the vine when compared to the generating variable itself.

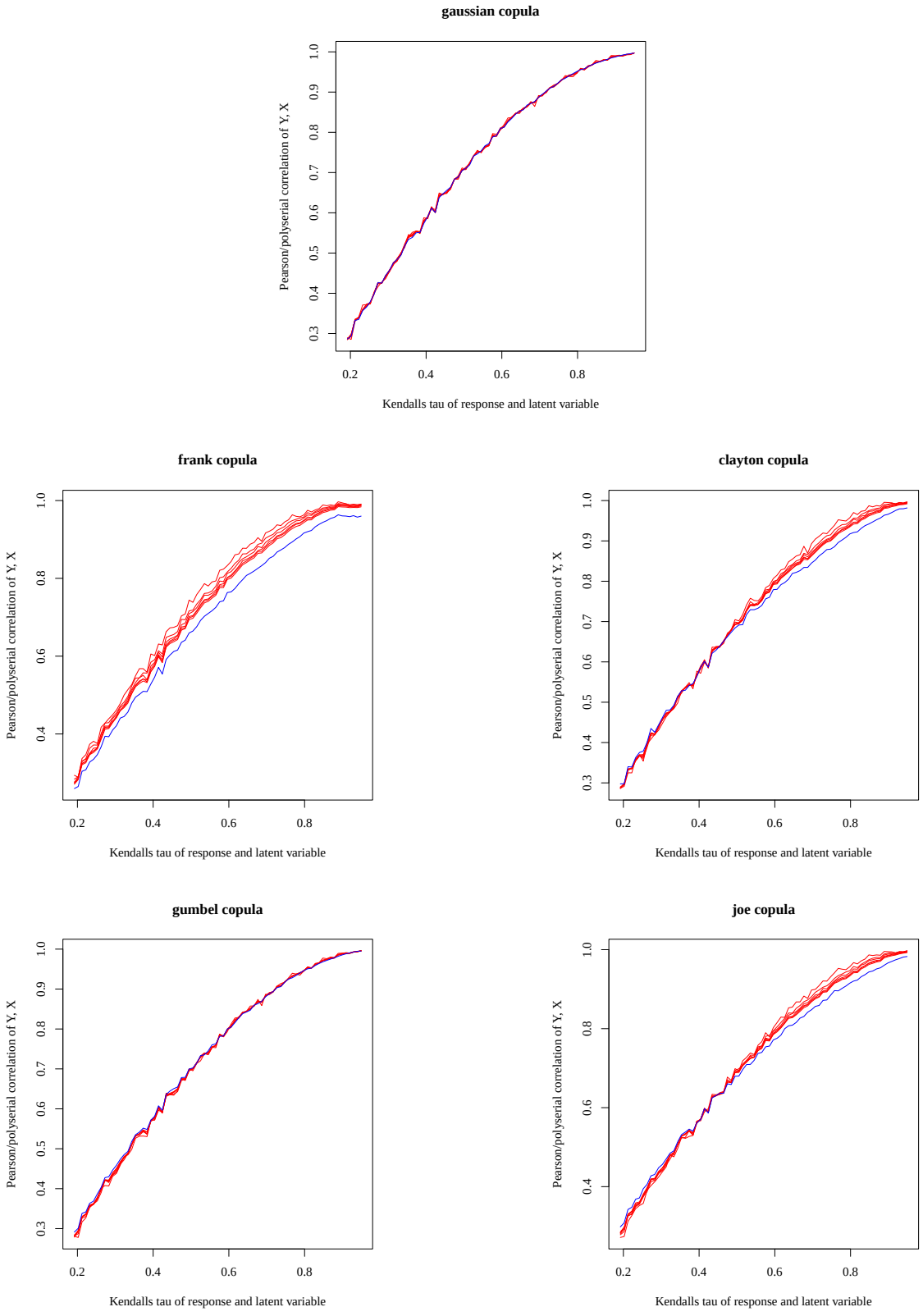


Figure 5.1: Both Y and X have $\mathcal{N}(0,1)$ marginal distributions. Dependence structure of Y and X is given in the subtitle of each plot. Presented on the x-axis is Kendall's τ of the response and continuous covariate X, which determines the parameter of the copula between Y and X. Pearson's/polyserial correlation was computed for $\tau = 0.20, 0.21, \ldots, 0.95$, but curves were drawn in a continuous manner by linear interpolation between points. The blue line represents Pearson correlation of Y and X. The red lines represent polyserial correlation between Y and discretized X. The red lines (although they are not easily distinguishable) correspond to $b = 2, 3, 4, 6, 8, 20$. The number of observations over which Pearson's/polyserial correlation were computed is $n = 10000$.

Pearson's and polyserial correlation are not margin-free, that is, their values are affected by the marginal distributions of Y and X. For a short look at the robustness of the conclusion that Pearson's/polyserial correlation is unsuitable as a variable selection measure for mixed-type variables, we repeat the simulation from Figure 5.1, this time with uniform marginal distributions for Y and X.

From Figure 5.2 below, it seems the issues with overestimation get worse when the assumption of normality from polyserial correlation fails. For uniform marginals we see again that polyserial correlation gives a greater estimate of correlation with the response than Pearson's correlation, confirming our conclusion that Pearson's/polyserial correlation is not an ideal variable selection measure for mixed-type data because it is biased in favour of discrete variables, rather than against.

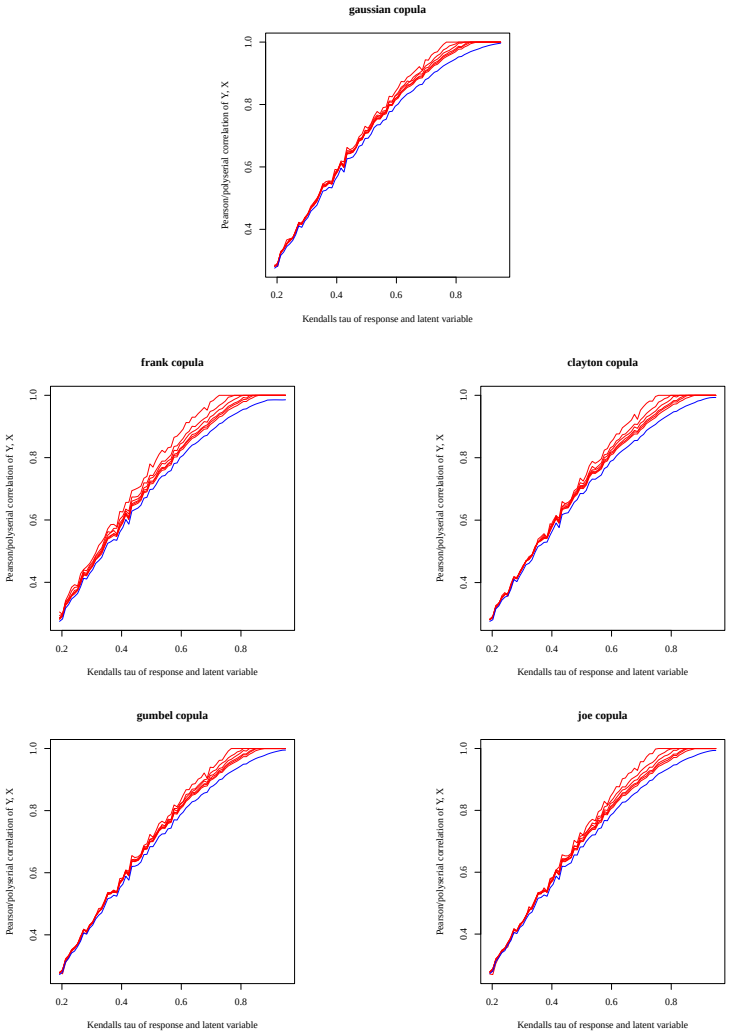


Figure 5.2: Both Y and X have Unif$[0, 1]$ marginal distributions. Dependence structure of Y and X is given in the subtitle of each plot. Presented on the x-axis is Kendall's τ of the response and continuous covariate X, which determines the parameter of the copula between Y and X. Pearson's/polyserial correlation was computed for $\tau = 0.20, 0.21, \ldots, 0.95$, but curves were drawn in a continuous manner by linear interpolation between points. The blue line represents Pearson correlation of Y and X. The red lines represent polyserial correlation between Y and discretized X. The red lines (although they are not easily distinguishable) correspond to $b = 2, 3, 4, 6, 8, 20$. The number of observations over which Pearson's/polyserial correlation were computed is $n = 10000$.

5.1.2 Kendall's tau/tau-b

Figure 5.3 below plots estimated Kendall's tau between Y and X, and estimated Kendall's tau-b between Y and discretized X, for a variety of copula families and a range of copula parameters.

As expected, estimated Kendall's tau and Kendall's tau-b seem (visually, at least) unaffected by the copula family governing Y and X. This is as expected because Kendall's tau characterizes the strength of dependence of a copula, but it is not dependent on the specific dependence characteristics (i.e. shape of distribution function) of that copula. An important drawback of Kendall's tau/tau-b as a variable selection measure is that a discretized variable is not necessarily scored as less attractive than its generating variable; Kendall's tau-b of Y and discretized X may be greater than Kendall's tau of Y and X. In fact for discretized variables with a large number of bins ($b = 6, 8, 20$) Kendall's tau-b of Y and the discretized variable is consistently greater than estimated Kendall's tau of Y and the continuous covariate.

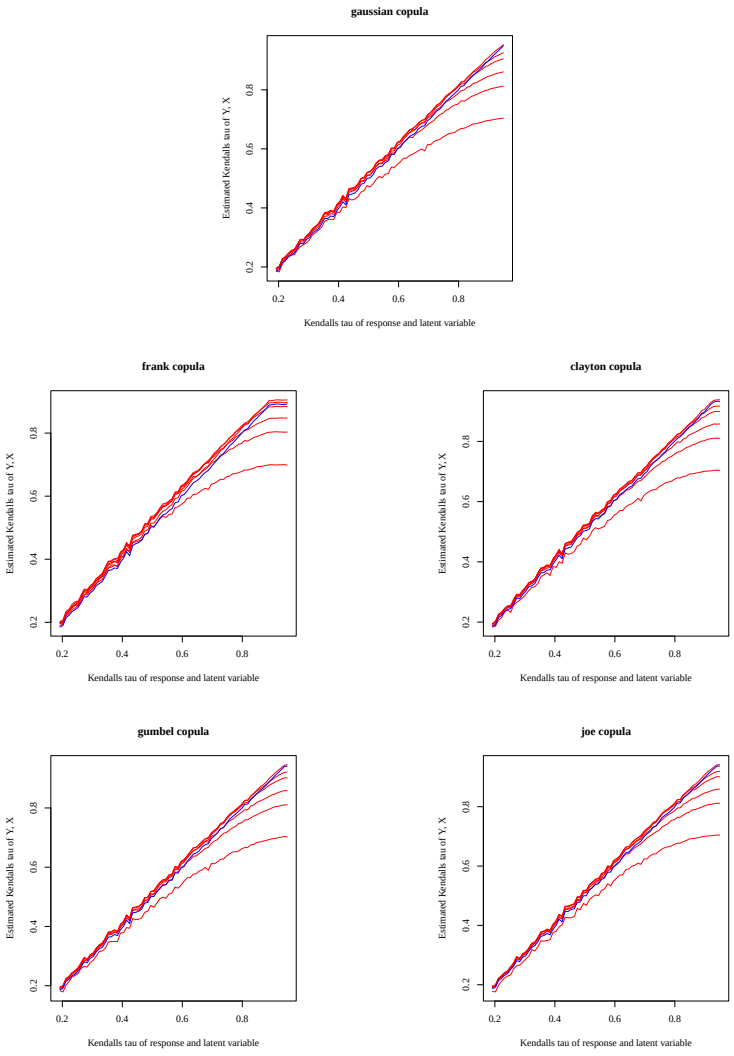


Figure 5.3: Dependence structure of Y and X is given in the subtitle of each plot. Presented on the x-axis is Kendall's τ of the response and continuous covariate X, which determines the parameter of the copula between Y and X. Estimated Kendall's tau and tau-b were computed for $\tau = 0.20, 0.21, \ldots, 0.95$, but curves were drawn in a continuous manner by linear interpolation between points. The blue line represents estimated Kendall's tau of Y and X. The red lines represent Kendall's tau-b between Y and discretized X. The red lines correspond, from lowest to highest, to $b = 2, 3, 4, 6, 8, 20$. The number of observations over which Kendall's tau/tau-b were computed is $n = 10000$.

5.1.3 Conditional log-likelihood

Figure 5.4 below plots the conditional log-likelihood (CLL) of Y given X, and conditional log-likelihood of Y given discretized X for a variety of copula families and a range of copula parameters.

Note in these figures that the CLL of Y given discretized X is monotone in the number of bins b of the discretized covariate (as opposed to previously seen Pearson's/polyserial correlation and Kendall's tau/tau-b), but not linear. For example, for any fixed dependence strength τ the increase in CLL when going from $b = 2$ to $b = 6$ is roughly the same as the increase from $b = 6$ to the CLL with the continuous covariate. This is likely due to an effect observed with the discretized conditional quantiles in Section 4, where we observed *"...as the number of bins increases the discretized conditional quantiles become more accurate around the median of the covariate relatively quickly, but even for a large number of bins ($b = 20$) tail dependence is not captured well"*.

The CLL is predominantly affected by predictive performance around the conditional mean of the response. For a variable selection measure which focuses more on predictive performance on the tail of the response, such as check-loss at $\alpha = 0.05$, the spacing between bins may look quite different, with larger differences between check-loss at $\alpha = 0.05$ using continuous covariates and check-loss at $\alpha = 0.05$ using discretized covariates.

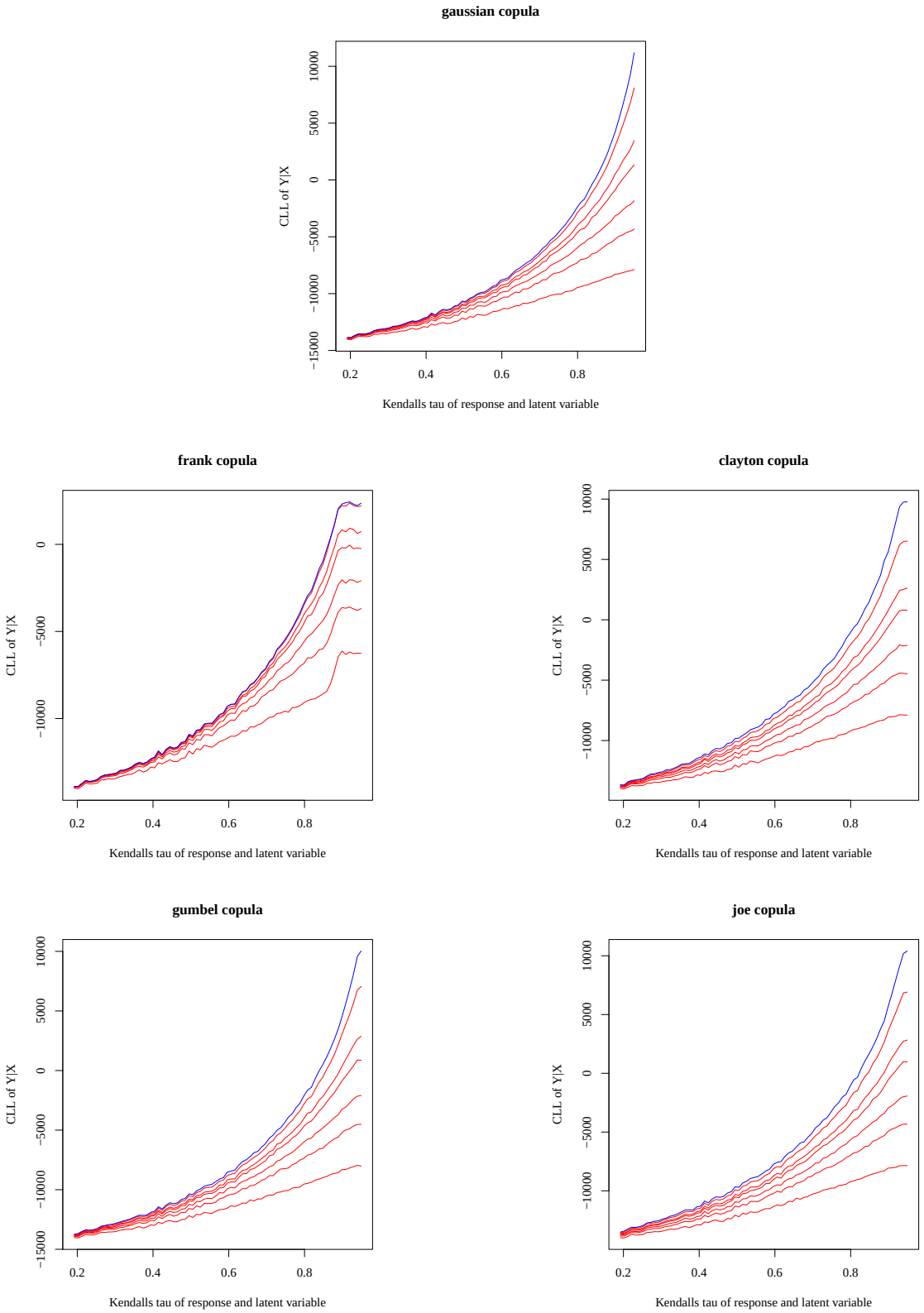


Figure 5.4: Both Y and X have $\mathcal{N}(0,1)$ marginal distributions. Dependence structure of Y and X is given in the subtitle of each plot. Presented on the x-axis is Kendall's τ of the response and continuous covariate X, which determines the parameter of the copula between Y and X. Conditional log-likelihood of Y given X and Y given discretized X were computed for $\tau = 0.20, 0.21, \ldots, 0.95$, but curves were drawn in a continuous manner by linear interpolation between points. The blue line represents CLL of Y and X. The red lines represent CLL between Y and discretized X. The red lines correspond, from lowest to highest, to $b = 2, 3, 4, 6, 8, 20$. The number of observations over which CLL was computed is $n = 10000$.

To quantify the bias of CLL as a variable selection measure against discretized covariates, we want to quantify the difference in CLL between continuous and its discretized covariate as a function of dependence strength between response and covariate (Kendall's τ) and number of bins of the discretized covariate (b). In lieu of an analytical approach to quantifying this difference in CLL between some continuous variable and its discretized variable, we fit a function to the CLL for each setting (continuous covariate, discretized covariate with $b \in \{2, 3, 4, 6, 8, 20\}$) by way of minimizing mean-squared error. These fitted functions provide an estimate of the CLL of Y given a covariate based on characteristics of the marginal of the covariate (whether it is continuous and otherwise the number of bins of the marginal) and Kendall's tau of Y and the variable which generates the covariate. This allows us to compare the estimated CLL of response given a covariate between two covariates with different type (continuous/discrete), discretization procedure (number of bins), and Kendall's τ with Y. A more detailed explanation of the fitting procedure can be found in Appendix 7.3.1. For an illustration, Figure 5.5 below plots the function fitted on the Gaussian subplot of Figure 5.4.

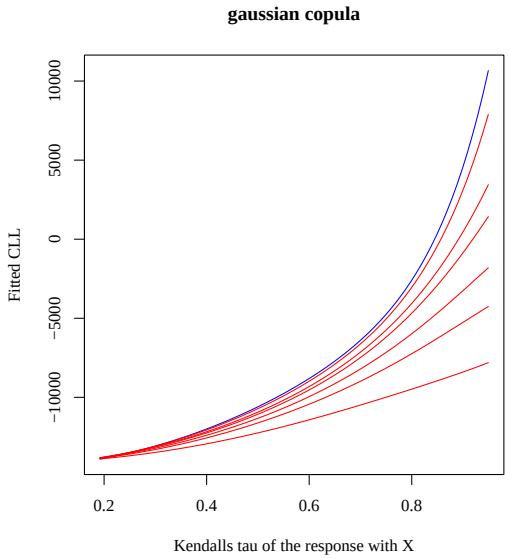


Figure 5.5: Both Y and X have $\mathcal{N}(0, 1)$ marginal distributions. Dependence structure of Y and X is Gaussian. Presented on the x-axis is Kendall's τ of the response Y and continuous covariate X. The blue line represents the function fitted on CLL of Y given X of Figure 5.4. The red lines represent the function fitted on CLL of Y given discretized X of Figure 5.4. The red lines correspond, from top to bottom, to $b = 2, 3, 4, 6, 8, 20$.

As said earlier, we can use these functions fitted on the CLL for different types of marginal distributions (continuous/discretized) to compare any pair of variables based on their 'estimated' (estimation by way of evaluating the fitted function) CLL under a given copula. Figure 5.6 below plots a conversion of Kendall's tau of the variable which generates the discretized covariate to the Kendall's tau of the continuous covariate with equal 'estimated' CLL. For an example of how to interpret Figure 5.6, from the line for $b = 2$ in the subplot corresponding to Gaussian dependence we read that the estimated CLL for a discretized variable with $b = 2$ bins generated from a variable which has Kendall's tau with the response equal to 0.6, is equal to the estimated CLL for continuous covariate with Kendall's tau with the response equal to 0.4. In other words, when using CLL as a variable selection measure, a continuous covariate with Kendall's tau with the response equal to 0.4 is measured as equally informative for Y as a discretized covariate with $b = 2$ whose latent variable has Kendall's tau with the response equal to 0.6. These figures allow for a rough comparison in CLL between any two variables, without needing to fit copulas

connecting the variables to the response to compute the CLL exactly. This gives insight into the biases of CLL as a variable selection measure in the bivariate setting.

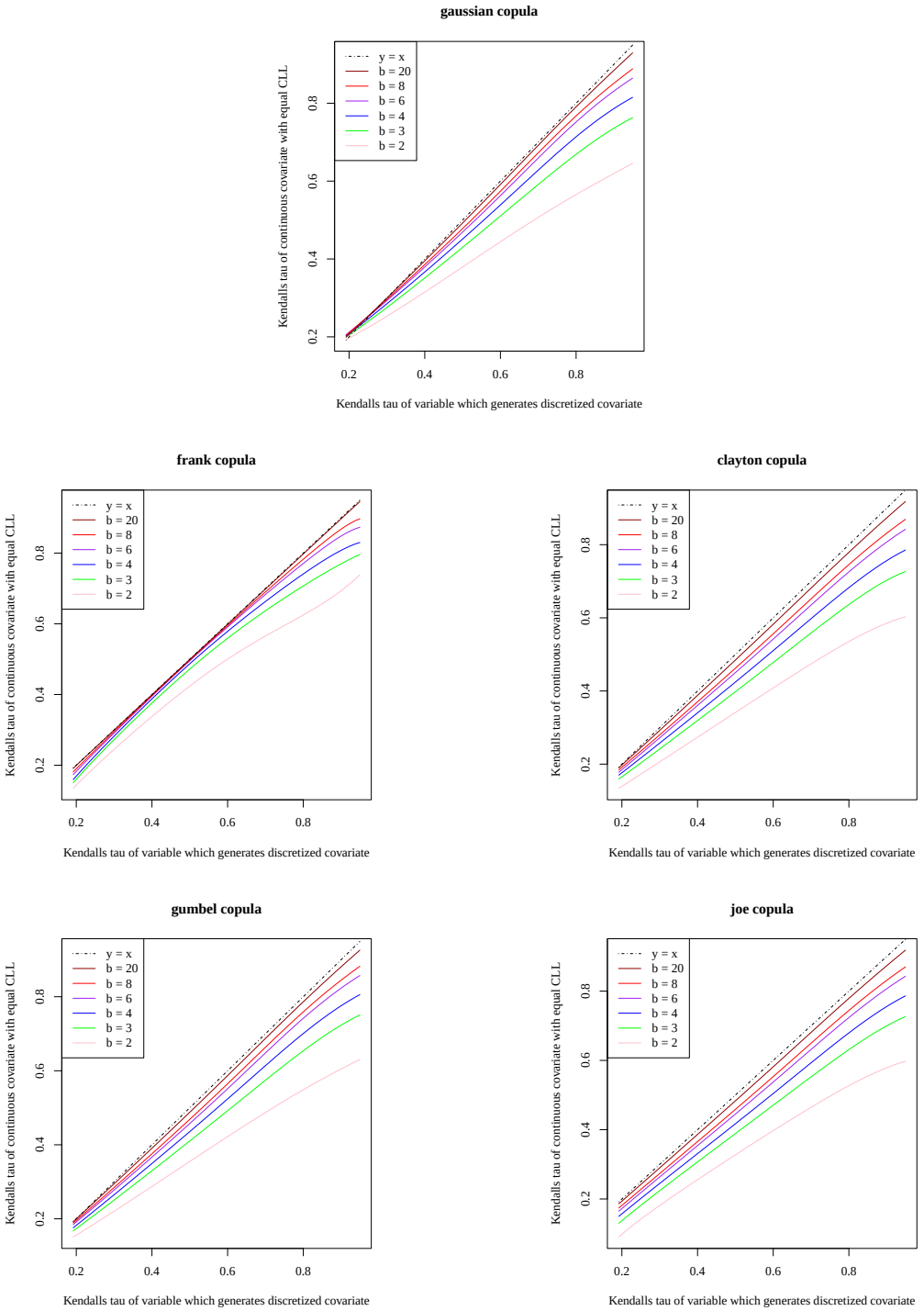
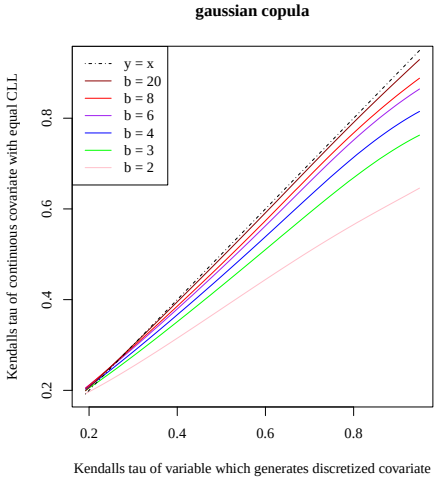


Figure 5.6: Both Y and X have $\mathcal{N}(0,1)$ marginal distributions. Dependence structure of Y and X is given in the subtitle of each plot. Presented on the x-axis is Kendall's τ of the response and the variable which generates the discretized X. By way of evaluating the function fitted on the empirical CLL-values of Figure 5.4, we estimate the CLL corresponding to that discretized covariate. On the y-axis is Kendall's τ of the response and continuous X with the same estimated CLL.

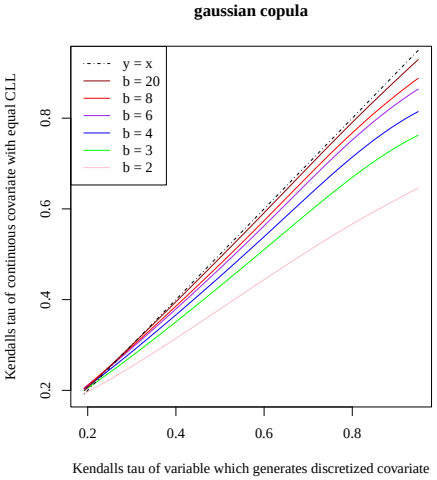
Remark. The fitted curves on the Frank subplot of Figure 5.6 and Figure 7.10 deviate more from the empirical values compared to other families. While fitting an alternative functional form might give a better visual fit, we chose to use a consistent fitting approach across all subplots to ensure comparability. Since the discrepancy does not affect the overall interpretation or conclusions, no additional adjustment was applied.

As mentioned earlier, these plots reinforce the idea that CLL punishes discretization mostly for a low ($b \leq 4$) number of bins, but a discretized variable with a moderate to high number of bins ($b \geq 6$) is measured as only slightly less informative for prediction of Y than its continuous latent variable.

We reproduce Figure 5.6 for uniform marginals of Y and X, to understand whether the marginal distributions have an effect on the bias of CLL against discretized covariates when used as a variable selection measure. The conversion plots for all copula families can be found in Figure 7.10 in the Appendix. The plot below compares the Gaussian subplot of Figure 5.6 (conversion plots of CLL with normal marginals) to the Gaussian subplot of Figure 7.10 (conversion plots of CLL with uniform marginals).



(a) Both Y and X have $\mathcal{N}(0,1)$ marginal distributions. Dependence structure of Y and X is Gaussian. Presented on the x-axis is Kendall's τ of the response and the variable which generates the discretized X. By way of evaluating the function fitted on the empirical CLL-values of Figure 5.4, we estimate the CLL corresponding to that discretized covariate. On the y-axis is Kendall's τ of the response and continuous X with the same estimated CLL.



(b) Both Y and X have Unif$[0,1]$ marginal distributions. Dependence structure of Y and X is Gaussian. Presented on the x-axis is Kendall's τ of the response and the variable which generates the discretized X. By way of evaluating the function fitted on the empirical CLL-values of Figure 5.4, we estimate the CLL corresponding to that discretized covariate. On the y-axis is Kendall's τ of the response and continuous X with the same estimated CLL.

Figure 5.7

Comparing Figure 5.6 with Figure 7.10, the shape of the plots are visually identical. This implies that the bias of variables using CLL as a variable selection measure does not depend significantly on the marginal distributions of the response or continuous covariate.

5.1.4 Check-loss at $\alpha = 0.05$

From the previous section regarding CLL, we found the CLL of a discretized covariate improves very quickly as the number of bins b of the discretized covariate increases from 2 to 3/4/6. We

hypothesized that a variable selection measure which favours predictive performance on the tails would see a smaller effect of increasing b from low to a moderate number of bins, because we have seen in Section 4 that tail dependence remains poorly captured as the number of bins increases. We test this hypothesis by repeating the simulation, this time with check-loss at $\alpha = 0.05$ as the variable selection measure.

For an accurate comparison between the biases of CLL and check-loss at $\alpha = 0.05$ as variable selection measures, we create a conversion plot similar to Figure 5.6, this time for check-loss at $\alpha = 0.05$. A more detailed explanation of how these conversion plots for check-loss at $\alpha = 0.05$ were created can be found in Appendix 7.3.3. We present the conversion plots in Figure 5.8 below.

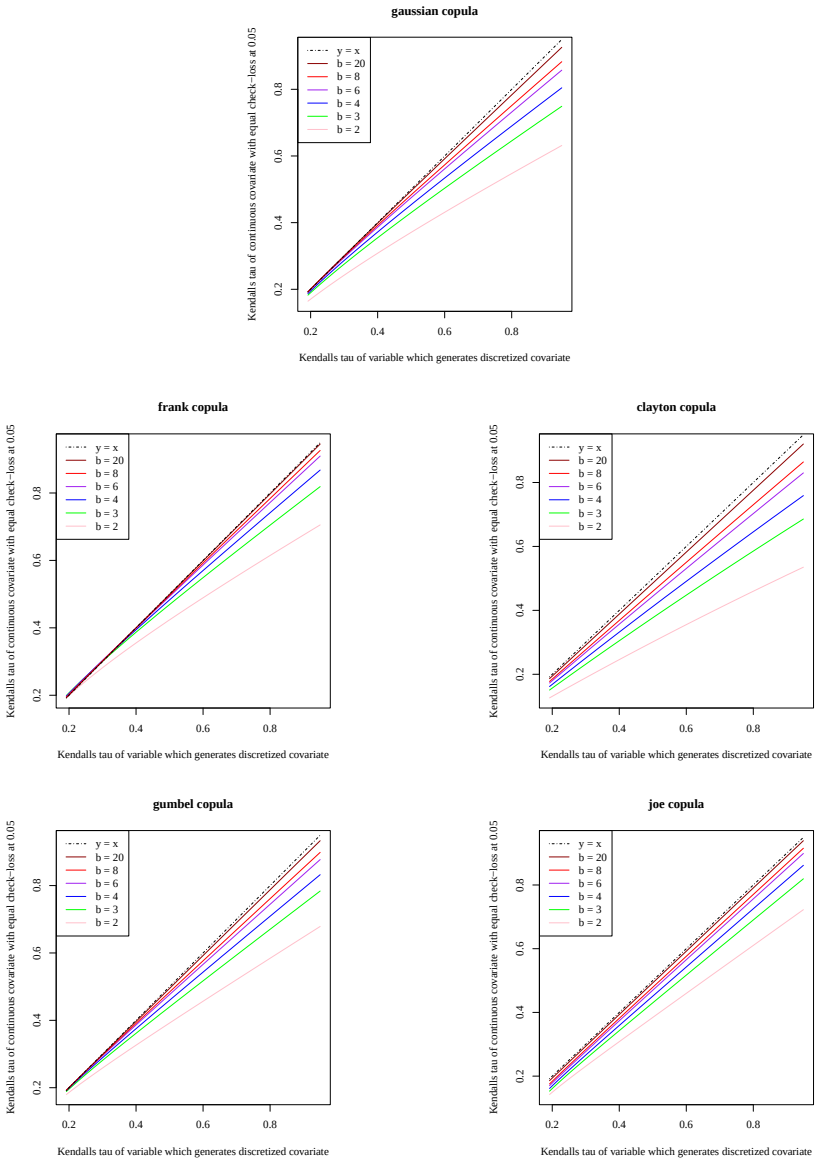
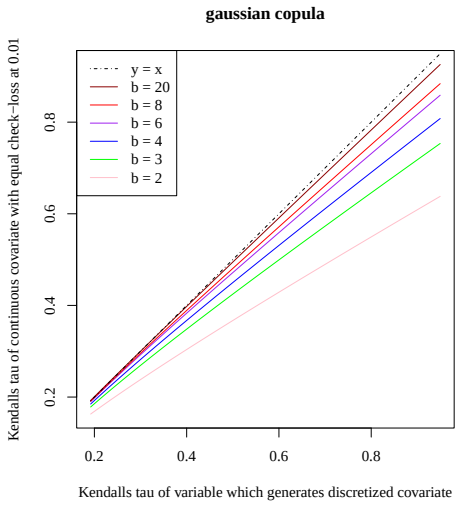


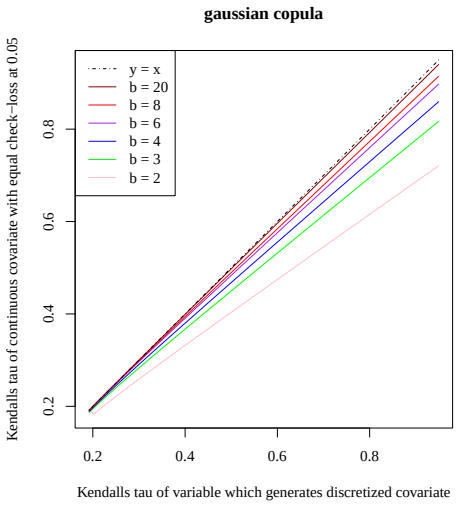
Figure 5.8: Both Y and X have $\mathcal{N}(0, 1)$ marginal distributions. Dependence structure of Y and X is given in the subtitle of each plot. Presented on the x-axis is Kendall's τ of the response and the variable which generates the discretized X. By way of evaluating the function fitted on the empirical check-loss values of Figure 7.11, we estimate the check-loss ($\alpha = 0.05$) corresponding to that discretized covariate. On the y-axis is Kendall's τ of the response and continuous X with the same estimated check-loss ($\alpha = 0.05$).

Despite our hypothesis that a variable selection measure which favours covariates which capture tail dependence would see a greater bias against discretization than CLL, this is not what the conversion plot of check-loss at $\alpha = 0.05$ tells us. In fact, the conversion plots of check-loss at $\alpha = 0.05$ and CLL are nearly identical. One possible explanation for this is that the 0.05-quantile is not far enough in the tails to see a clear difference in the behaviours of CLL and check-loss.

To see if the behaviour of check-loss as a variable selection measure (i.e. the shape of the conversion plots of Figure 5.8) depends significantly on the chosen marginal distributions of Y and X and the α-quantile at which check-loss is computed, we plot Figure 5.8, this time with (1) check-loss at $\alpha = 0.01$ (see Figure 7.13 in Appendix), a more extreme tail quantile, and (2) Y and X uniform (see Figure 7.14 in Appendix). For illustration, we plot the Gaussian sub-plots of the aforementioned figures below, the full figures can be found in the Appendix.



(a) Both Y and X have $\mathcal{N}(0,1)$ marginal distributions. Dependence structure of Y and X is Gaussian. Presented on the x-axis is Kendall's τ of the response and the variable which generates the discretized X. By way of evaluating the function fitted on the empirical check-loss ($\alpha = 0.01$) values we estimate the check-loss ($\alpha = 0.01$) corresponding to that discretized covariate. On the y-axis is Kendall's τ of the response and continuous X with the same estimated check-loss ($\alpha = 0.01$).



(b) Both Y and X have Unif$[0,1]$ marginal distributions. Dependence structure of Y and X is Gaussian. Presented on the x-axis is Kendall's τ of the response and the variable which generates the discretized X. By way of evaluating the function fitted on the empirical check-loss ($\alpha = 0.05$) values, we estimate the check-loss ($\alpha = 0.05$) corresponding to that discretized covariate. On the y-axis is Kendall's τ of the response and continuous X with the same estimated check-loss ($\alpha = 0.05$).

Comparing both Gaussian sub-plots, we see the behaviour (i.e. the bias against discretized covariates) of check-loss as variable selection measure is somewhat dependent on the marginals of Y and X, but largely independent of the tail-quantile at which check-loss is evaluated. The fact that under uniform marginals check-loss is less biased against discretization is possibly due to the boundedness of the uniform distribution. Because the uniform distribution is bounded, even a bad prediction (for instance, using a discretized covariate) stays quite close to the correct prediction (using the continuous covariate), which is not the case for normal marginal distributions.

5.2 Variable selection measures under discretization, in 3D

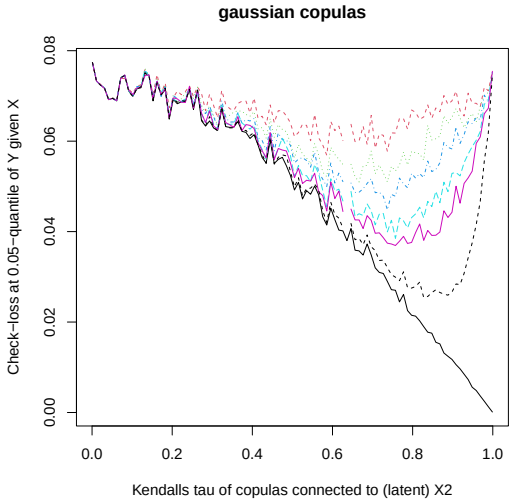
In the previous section, we quantified the impact of discretization on variable selection measures in a setting with a single covariate and saw that in the bivariate setting Pearson/polyserial correlation and Kendall's τ/τ_b are not biased enough against discretized covariates to be used in vine-based regression for mixed-type data, and that CLL and check-loss ($\alpha = 0.05$) were closely aligned in the simulated bivariate scenario. This section investigates CLL and check-loss at $\alpha = 0.05$ in the 3-dimensional scenario.

In this section, we aim to achieve a similar result in a setting with a single response Y and two covariates, X_1 and X_2, where one of the covariates is discretized (again in the manner laid out in Definition 4.1). Observations will be generated from a D-vine with order $Y - X_1 - X_2$. In a setting with two covariates, we naturally have a choice of which of the two covariates to discretize. Discretization of X_1 may not yield equal CLL/check-loss values as the discretization of X_2, as the roles these covariates play in the vine are not identical. The most notable difference is that due to X_2 being a leaf node in the vine, X_2 is never conditioned on, whereas X_1 is conditioned on in the copula $c_{YX_2|X_1}$. Discretization may have a different impact on a variable in the conditioned set than a variable which is never conditioned on. As we will see, the multivariate setting is much more complicated than the univariate setting.

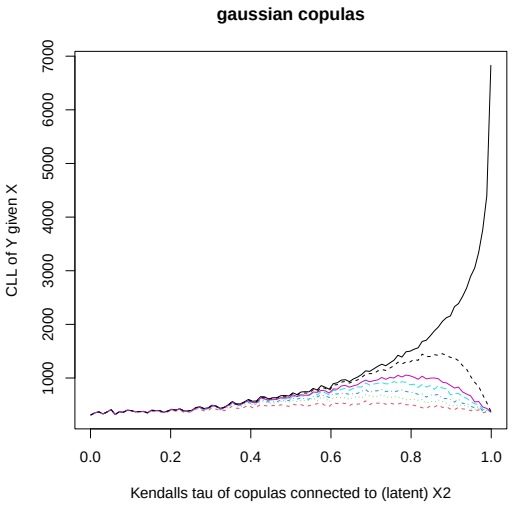
In the bivariate analysis in Section 5.1, we computed the effect of discretization on CLL and check-loss for many strengths of dependence, from $\tau = 0.20$ to $\tau = 0.95$. There, we saw the effect of discretization is dependent on the strength of dependence of the discretized variable with the response. In the 3-dimensional setting with X_2 as the discretized variable, there are two copulas which affect the strength of dependence between Y and either covariate. Thus, there are two copulas whose dependence strength we want to adjust. When X_2 is discretized, we adjust the dependence strength of the copula $c_{X_1X_2}$ and the conditional copula $c_{YX_2|X_1}$. Due to time-constraints, we only consider the D-vine with X_2 discretized.

Vine with X_2 discretized

We will do a simulation study to compute CLL and check-loss at $\alpha = 0.05$ of vines with different parameters and number of bins b with which discretized X_2 is generated. We sample $n = 1000$ observations from the D-vine with order $Y - X_1 - X_2$. The marginal distributions of the continuous variables, Y and X_1, are standard normal. Discretized X_2 is generated through binning its latent variable X_2 into bins of equal probability, as illustrated in Definition 4.1.



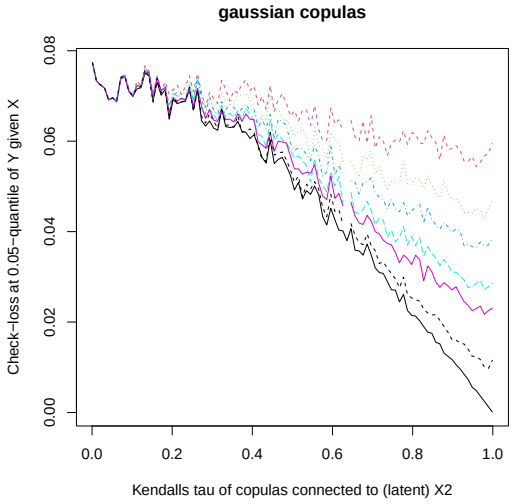
(a) On the y-axis, check-loss at $\alpha = 0.05$ of Y given X_1 and (discretized or continuous) X_2 over $n = 1000$ observations. On the x-axis, Kendall's τ of copulas $c_{X_1X_2}$ and $c_{YX_2|X_1}$. Kendall's τ of copula c_{YX_1} is fixed at $\tau = 0.5$. From lowest to highest line in the graph, we see check-loss at 0.05-quantile for continuous X_2, then check-loss at 0.05 for discretized X_2 with bins $b = 20, 8, 6, 4, 3, 2$, respectively.



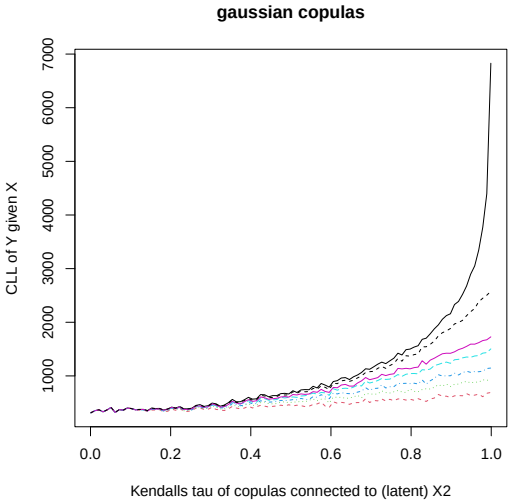
(b) On the y-axis, CLL of Y given X_1 and (discretized or continuous) X_2 over $n = 1000$ observations. On the x-axis, Kendall's τ of copulas $c_{X_1X_2}$ and $c_{YX_2|X_1}$. Kendall's τ of copula c_{YX_1} is fixed at $\tau = 0.5$. From highest to lowest line in the graph, you see CLL for continuous X_2, then CLL for discretized X_2 with bins $b = 20, 8, 6, 4, 3, 2$, respectively.

Figure 5.10

Observe the strange effect discretization has on the CLL and check-loss in this scenario. One would expect them to be decreasing as τ goes to 1 (as they did in the bivariate analysis), but for discretized X_2 predictive performance gets worse after reaching an optimum around $\tau \approx 0.8$. One might find a hypothetical reason for this worsening predictive performance when one considers that on the x-axis, two different copula parameters are changed; the parameter of $c_{X_1X_2}$ and the parameter of $c_{YX_2|X_1}$. It is known that correlation between covariates is undesirable for predictive performance in regression problems. So, perhaps the negative effect of increasing the dependence between covariates X_1 and X_2 takes over from the positive effect of increasing the dependence between $Y, X_2 \mid X_1$ around $\tau \approx 0.8$. This hypothesis would explain why predictive performance using discretized X_2 gets worse after reaching an optimum at $\tau < 1$, but it does not explain why predictive performance using the continuous version of X_2 steadily improves as τ goes up. Still, to test this hypothesis let us make a similar plot as Figure 5.10, this time keeping the dependence strength of the copula c_{X_1,X_2} constant at $\tau = 0.7$ (deliberately chosen smaller than the value where CLL and check-loss move back in the unfavourable direction), and varying only the dependence strength of the conditional copula $c_{Y,X_2|X_1}$.



(a) Generated in a near-identical setting as Figure 5.10, only with Kendall's τ of copula $c_{X_1X_2}$ fixed at $\tau = 0.7$. On the y-axis, check-loss at $\alpha = 0.05$ of Y given X_1 and (discretized or continuous) X_2. On the x-axis, Kendall's τ of copula $c_{YX_2|X_1}$. Kendall's τ of copulas c_{YX_1} and $c_{X_1X_2}$ are fixed at $\tau = 0.5$ and $\tau = 0.7$, respectively. From lowest to highest line in the graph, you see check-loss at 0.05-quantile for continuous X_2, then check-loss at $\alpha = 0.05$ for discretized X_2 with bins $b = 20, 8, 6, 4, 3, 2$, respectively.

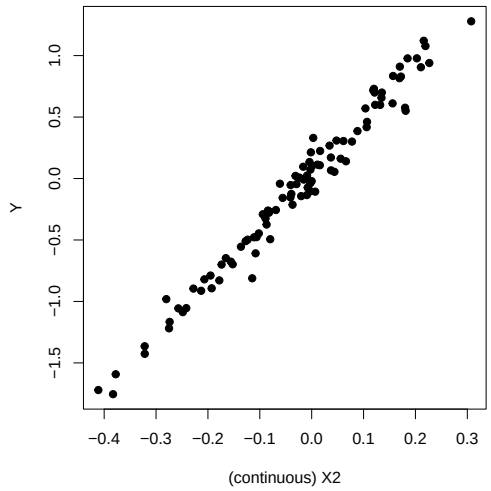


(b) Generated in a near-identical setting as Figure 5.10, only with Kendall's τ of copula $c_{X_1X_2}$ fixed at $\tau = 0.7$. On the y-axis, CLL of Y given X_1 and (discretized or continuous) X_2. On the x-axis, Kendall's τ of copula $c_{YX_2|X_1}$. Kendall's τ of copulas c_{YX_1} and $c_{X_1X_2}$ are fixed at $\tau = 0.5$ and $\tau = 0.7$, respectively. From highest to lowest line in the graph, you see CLL for continuous X_2, then CLL for discretized X_2 with bins $b = 20, 8, 6, 4, 3, 2$, respectively.

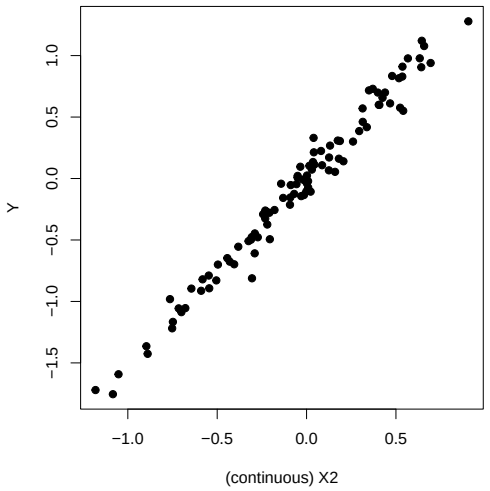
Figure 5.11

This looks more like what one would expect, with CLL and check-loss using discretized X_2 decreasing steadily as τ becomes larger. This leads us to conclude that in this 3-dimensional setting with X_2 discretized, very high correlation between X_1 and X_2 has a very strong negative effect on predictive performance. Now, let us explain why it is only the CLL and check-loss for discretized X_2 which suffer so much from high correlation between X_1 and X_2, whereas CLL and check-loss using continuous X_2 do not.

To explain this, let us make a first observation; changing the parameter of $c_{X_1X_2}$ does not change the Y- or X_1-values of the observations generated by the D-vine. In other words, under the simulation settings of Figure 5.10 and Figure 5.11, where the parameter of $c_{X_1X_2}$ is the only difference between the two settings, the columns corresponding to Y and X_1 are identical in both datasets. Thus, we may gauge visually the predictive power X_2 has after conditioning on some value of X_1, by plotting for both simulation settings all values (X_2, Y) which correspond to a value for X_1. Then we can see how much predictive power X_2 has after removing the effect of X_1.



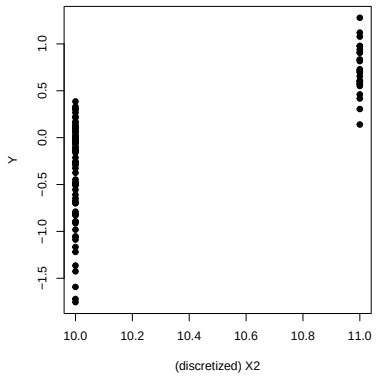
(a) Values (X_2, Y) of all observations whose corresponding X_1-value is in $(-0.025, 0]$. Observations were generated with a D-vine with Kendall's tau of copulas c_{YX_1}, $c_{X_1X_2}$, $c_{YX_2|X_1}$ equal to 0.5, 0.9 and 0.9, respectively. This is equivalent to the simulation setting of Figure 5.10 at $\tau = 0.9$.



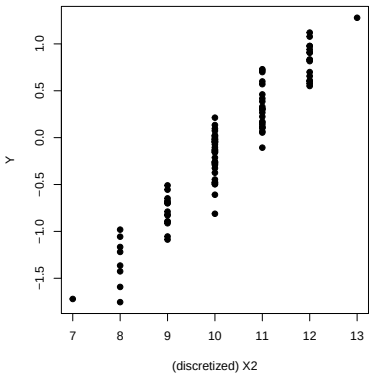
(b) Values (X_2, Y) of all observations whose corresponding X_1-value is in $(-0.025, 0]$. Observations were generated with a D-vine with Kendall's tau of copulas c_{YX_1}, $c_{X_1X_2}$, $c_{YX_2|X_1}$ equal to 0.5, 0.7 and 0.9, respectively. This is equivalent to the simulation setting of Figure 5.11 at $\tau = 0.9$.

Figure 5.12

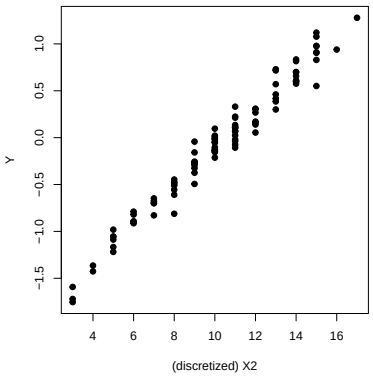
Higher correlation between X_1 and X_2 does not affect the predictive power of X_2 after adjusting for X_1. Note one big difference between Figure 5.12a and Figure 5.12b is the width of the x-axis, i.e. the width of the range of values X_2 attains in observations with corresponding $X_1 \in (0.025, 0]$. Now let us look at plots produced in the same conditions, only with discretized X_2, discretized into $b = 20$ bins, because the effect of worsening CLL and check-loss as τ goes to 1 in Figure 5.10 is most pronounced for $b = 20$.



(a) Values (X_2, Y) of all observations whose corresponding X_1-value is in $(-0.025, 0]$. Observations were generated with a D-vine with Kendall's tau of copulas c_{YX_1}, $c_{X_1X_2}$, $c_{YX_2|X_1}$ equal to 0.5, 0.98 and 0.98, respectively. This is equivalent to the simulation setting of Figure 5.10 at $\tau = 0.98$. Discretized X_2 is obtained by quantile binning with $b = 20$ bins of equal quantile-width.



(b) Values (X_2, Y) of all observations whose corresponding X_1-value is in $(-0.025, 0]$. Observations were generated with a D-vine with Kendall's tau of copulas c_{YX_1}, $c_{X_1X_2}$, $c_{YX_2|X_1}$ equal to 0.5, 0.9 and 0.9, respectively. This is equivalent to the simulation setting of Figure 5.10 at $\tau = 0.9$. Discretized X_2 is obtained by quantile binning with $b = 20$ bins of equal quantile-width.



(c) Values (X_2, Y) of all observations whose corresponding X_1-value is in $(-0.025, 0]$. Observations were generated with a D-vine with Kendall's tau of copulas c_{YX_1}, $c_{X_1X_2}$, $c_{YX_2|X_1}$ equal to 0.5, 0.7 and 0.9, respectively. This is equivalent to the simulation setting of Figure 5.11 at $\tau = 0.9$. Discretized X_2 is obtained by quantile binning with $b = 20$ bins of equal quantile-width.

Figure 5.13

Clearly, the higher correlation between X_1 and X_2 has a significant impact on the predictive power discretized X_2 has over Y after adjusting for X_1. This is because the higher correlation between X_1 and X_2 narrows the range of values X_2 takes for any given X_1 value. Thus, after discretization, the number of bins X_2 falls in for a given X_1 value decreases as the correlation of X_1 and X_2 increases, causing discretized X_2 to contribute less and less new information on Y additional to the information on Y already given by X_1. The slice of X_1 values used in Figure 5.12 and Figure 5.13 is equal to its 0.49^{th} and 0.50^{th} quantiles, which is very close to the quantile-boundary between the 10^{th} and 11^{th} bins. Because this slice is right on the boundary between two bins, discretized X_2 takes two values in Figure 5.13a. If the slice of X_1 values lies closer to the middle of a quantile bin, then discretized X_2 might only attain a single value, making it completely uninformative for Y after adjusting for X_1. This is why in Figure 5.10, the CLL and check-loss at $\tau = 0.99$ using discretized X_2 go back to their respective values at $\tau = 0.01$, because when correlation between X_1 and X_2 is nearly perfect, discretized X_2 attains only a single value for any small slice of X_1 values, making it completely uninformative for Y after adjusting for X_1, despite correlation of its generating variable (X_2) with Y given X_1 being extremely high. Finally, careful inspection of Figure 5.10 reveals that the value of τ at which the CLL and check-loss of discretized X_2 start to creep back up is lower for smaller number of bins. This is also explainable; when b is smaller, less correlation between X_1 and X_2 is needed before discretized X_2 becomes uninformative because the interior of quantile bins is larger when b is smaller.

In summary, in a D-vine with order $Y - X_1 - X_2$, correlation between covariates does not seem to affect predictive performance when both covariates are continuous, but when X_2 is discretized, correlation with X_1 has a very significant impact on predictive performance. As correlation between X_1 and X_2 goes to $\tau = 1$, discretized X_2 contributes less and less new information on Y additional to the information on Y already given by X_1, no matter how high the correlation between $Y, X_2 \mid X_1$ is. One question left unanswered in this section is whether correlation between a variable which generates a discretized covariate and any other continuous covariate has this effect on the predictive power of a discretized covariate, or whether it is only correlation with covariates which are in the conditioning set of the copula which connect the discretized covariate with the response which has this effect.

6 | Discussion & further research

This thesis investigated the effect of discretization of covariates on conditional quantile functions of bivariate copulas and variable selection measures, in an effort to take a step towards building privacy-preserving datasets using discretization of covariates and towards vine-based regression methods tailored to mixed-type (continuous & discrete) data.

In Section 4, we investigated the impact of discretization on different (bivariate) copula families, marginal distributions and binning procedures. We visualized discretized conditional quantiles over continuous conditional quantiles, giving some sense of how much information is lost in discretization. Initially, discretization was done in bins of equal probability. These discretized conditional quantiles lost so much dependence characteristics that some copula families became nearly indistinguishable, regardless of the marginal distribution of the response. The amount of information a discretized covariate provides about the dependence of the response and its generating variable may be improved significantly with a different approach to binning. In an effort to understand how one can choose the 'optimal' approach to discretizing a covariate (i.e. the optimal choice of quantile-boundaries), we took a step towards analytical quantification of the error in conditional quantiles caused by discretization (see Lemma 4.2), although a fully analytical solution to this problem seems unlikely due to the impossibility of analytically solving equations with certain copula families' (conditional) distribution functions. Solutions to the optimal-binning problem will require a numerical approach and/or heuristics.

The study of discretized conditional quantile functions showed us a variable selection measure tailored to mixed-type data should be biased against discrete variables due to the significant loss of information caused by discretization. Moreover, we desire such a variable selection measure to be monotone increasing in the number of bins of the discretized covariate. Section 5 investigated how common variable selection measures (Pearson's/polyserial correlation, Kendall's tau/tau-b, conditional log-likelihood, check-loss at a tail quantile) are biased for/against discretized variables. Pearson's/polyserial correlation and Kendall's tau/tau-b were found unsuitable for mixed-type data because they were not biased against discretized covariates and/or non-monotone. In the bivariate setting, conditional log-likelihood and check-loss at a tail quantile were found to have nearly identical biases against a discretized covariate, but note the caveat that we only investigated homoskedastic simulation scenarios. Before drawing conclusions about the exchangeability of CLL and check-loss at $\alpha = 0.05$ as variable selection measures in the bivariate setting, the biases against discretized covariates of conditional log-likelihood and check-loss at $\alpha = 0.05$ should be compared in a heteroskedastic scenario as well. Finally, the bias of conditional log-likelihood against discretized covariates seemed unaffected by the choice of marginals. On the other hand, the bias of check-loss $\alpha = 0.5$ was somewhat affected by choice of marginals, but seemed unaffected by the choice of α-quantile at which check-loss was evaluated when tested with $\alpha = 0.01$.

In a 3-dimensional setting with two covariates, we saw discretization has a different effect than in the 2-dimensional setting. Specifically, correlation between the covariates may have a strong negative effect on predictive performance when a covariate is discretized, to the point of making the discretized covariate completely uninformative for Y, as if it were (conditionally) independent from the response. This has important implications for the computational complexity of vine-

based regression tailored to mixed-type data. Namely, the method by Şahin & Czado ([30]) which reduced computational complexity of vine-based regression to $\mathcal{O}(d^2)$ by considering variable selection measures with the residual rather than with the vine as a whole, does not take into account the effect of correlation on discretized covariates, thereby overestimating the additional information discretized covariates would provide to the vine.

We present some avenues for further research. Some of which are related to the content of this thesis, others were ideas which could not be investigated fully due to time constraints.

6.1 D-vine regression for mixed data through latent variable estimation

Under the assumption that any discrete variable is generated from a continuous latent variable, the value of the discrete variable gives information about the range its latent variable is in. Rather than using variable selection measures tailored to mixed-type data, we might work around the problem of discrete variables by constructing a second dataset, where the discrete variables are replaced by their estimated latent data. Standard variable selection measures for continuous variables can then be used. Moreover, estimating latent data may be useful after variable selection as well. Copulas could be fit on the second dataset containing estimated latent data and predictions made using estimated latent data, completely bypassing the need for developing methods tailored to mixed-type data.

A direction for further research could be to investigate whether estimating the value of the latent variable through its polyserial/polychoric correlations with other covariates improves the predictive performance of the model, when compared to using the discrete variable. Estimating the value of the latent variable has two parts; (1) computing the distribution of the latent variable conditional on the value of its discrete variable and the other covariates, and (2) sampling from this distribution effectively.

For an illustration of how latent data may be estimated, we show how we obtain the distribution of the latent variable conditioned on covariate values. Let $X_1, X_2, \ldots, X_k$ denote continuous covariates. Let X_d, X_d^* denote a discrete covariate and its latent variable, respectively. We work under the assumption that all marginals and dependence is Gaussian. In our data, we have access to the value of X_d only. This tells us the range its latent variable X_d^* is in. We will condition on covariate values and the event $X_d = k$, letting c_k, c_{k-1} be the right- and left-boundaries of the k-th bin, respectively. We find

$$\begin{aligned}\mathbb{P}(X_d^* \leq x \mid X_1, X_2, \ldots, X_k, X_d = k) &= \mathbb{P}(X_d^* \leq x \mid X_1, X_2, \ldots, X_k, X_d^* \in (c_k, c_{k-1}]) \\ &= \frac{\mathbb{P}(X_d^* \leq x \mid X_1, X_2, \ldots, X_k) - \mathbb{P}(X_d^* \leq c_{k-1} \mid X_1, X_2, \ldots, X_k)}{\mathbb{P}(X_d^* \leq c_k \mid X_1, X_2, \ldots, X_k) - \mathbb{P}(X_d^* \leq c_{k-1} \mid X_1, X_2, \ldots, X_k)}.\end{aligned}$$

Each term is known, since by our assumption that all marginals and dependence structures are Gaussian we can use the well-known result of conditioning a multivariate normal, which states that when

$$\mathbf{X} = \begin{bmatrix} X_1 \\ \mathbf{X_{-1}} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_1 \\ \boldsymbol{\mu}_{-1} \end{bmatrix}, \begin{bmatrix} \sigma_{11} & \boldsymbol{\sigma}_{12} \\ \boldsymbol{\sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix} \right),$$

the conditional distribution of X_1 given $\mathbf{X_{-1}}$ is

$$F_{X_1|\mathbf{X_{-1}}}(x) = \Phi\left(\frac{x - \left(\mu_1 + \boldsymbol{\sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(X_{-1} - \boldsymbol{\mu}_{-1})\right)}{\sqrt{\sigma_{11} - \boldsymbol{\sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\sigma}_{21}}} \right).$$

To illustrate how the density of a latent variable which generates a binary discrete variable changes when conditioning on other covariate values, we set the covariance of all covariates to

0.5 and condition on the covariate values $(X_1 = -1, X_2, = -1, X_3 = -1, X_4 = -1, X_d^* \in (-\infty, 0])$ to present the following figure.

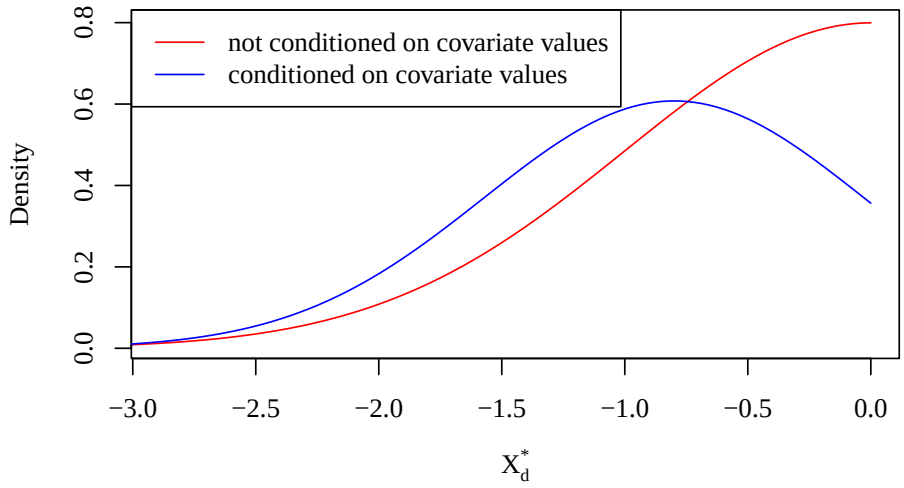


Figure 6.1: The red line shows the density $f_{X_d^*}(\cdot \mid X_d^* \in (-\infty, 0])$, the blue line shows the density $f_{X_d^*}(\cdot \mid X_1 = -1, X_2 = -1, X_3 = -1, X_4 = -1, X_d^* \in (-\infty, 0])$. The variables each have standard normal marginals, and covariance between any pair of variables is set to 0.5.

As can be seen from the figure, conditioning on covariate values may be informative for the distribution of the latent variable.

The next step to investigate is the sampling strategy. The sampling strategy should effectively explore the distribution of the latent variable computed previously, while balancing computational limits. Remember, each datapoint in the original dataset (D) produces its own distribution of the latent variable through its unique set of covariate values. To effectively explore each such distribution, we will sample from it multiple (n_{approx}) times. These n_{approx} samples will be put into a new dataset, $NewD$, which will thus contain n_{approx} times as many datapoints as D. To find all copula families and parameters of the D-vine, weighted MLE may be used on the new dataset $NewD$. Computing the weight for each datapoint will be part of the sampling strategy.

Possible sampling strategies include:

1. Sampling n_{approx} quantiles from a Unif$[0, 1]$ distribution, taking each sample as the inverse of the conditional latent variable distribution at that quantile and the weight of that sample as the density of the conditional latent variable evaluated at that sample.

2. Sampling n_{approx} quantiles such that they are equidistant and in the range $(0, 1)$, taking each sample as the inverse of the conditional latent variable distribution at that quantile and the weight of that sample as the density of the conditional latent variable evaluated at that sample.

3. Sampling n_{approx} quantiles such that they are equidistant and in the range $(0, 1)$, taking each sample as the inverse of the conditional latent variable distribution at that quantile q and the weight of that sample as the density of the conditional latent variable evaluated at that sample multiplied by a factor $1 + \lambda(0.5 - q)^2$, so that points on the tail are weighted more.

6.2 Quantification of error in conditional quantiles caused by discretization

In Section 4, we considered a way of quantifying the error in conditional quantiles caused by discretization by way of expressing this error, the difference between a continuous conditional quantile and the discretized conditional quantile, as a difference between two continuous conditional quantile functions at different conditioning values. This lead to error quantification for a very small subset of copula families and bins. Better conditional quantile error quantification/estimation would aid in designing an optimal binning strategy, to be used in building privacy-preserving datasets. Considering the difficulty of analytical solutions for this problem, we suggest efforts should be put towards estimation of discretization error in conditional quantile functions, rather than solving exactly.

6.3 Estimation of Kendall's tau of latent variable with response

The conversion plots of Figure 5.8 and Figure 5.6 plot on the x-axis the value of Kendall's tau between the response and the latent variable of a discretized covariate. From Figure 5.3, we saw Kendall's tau-b is not necessarily a good estimator for Kendall's tau between latent variable and the response. In order to use the conversion plots more effectively, a better estimate of Kendall's tau between latent variable and response may be investigated, in the situation when one has access only to data of the discretized covariate and the response.

6.4 Variable selection for vine-based regression for mixed-type data

The larger problem this thesis aims to address is vine-based regression for mixed-type data. The difference in impact of discretization in a two-dimensional or multi-dimensional setting should be taken into account when developing variable selection procedures for vine-based regression tailored to mixed-type data. We consider the main problem to be the loss of predictive power of discretized covariates when they are highly correlated with a continuous covariate. Further research could investigate when correlation between covariates of different type cause the loss of predictive power, and try to quantify this loss. One could start with the question left unanswered in this thesis, namely whether correlation between a variable which generates a discretized covariate and any other continuous covariate has a negative effect on the predictive power of a discretized covariate, or whether it is only correlation with covariates which are in the conditioning set of the copula which connect the discretized covariate with the response which has this effect.

7 | Appendix

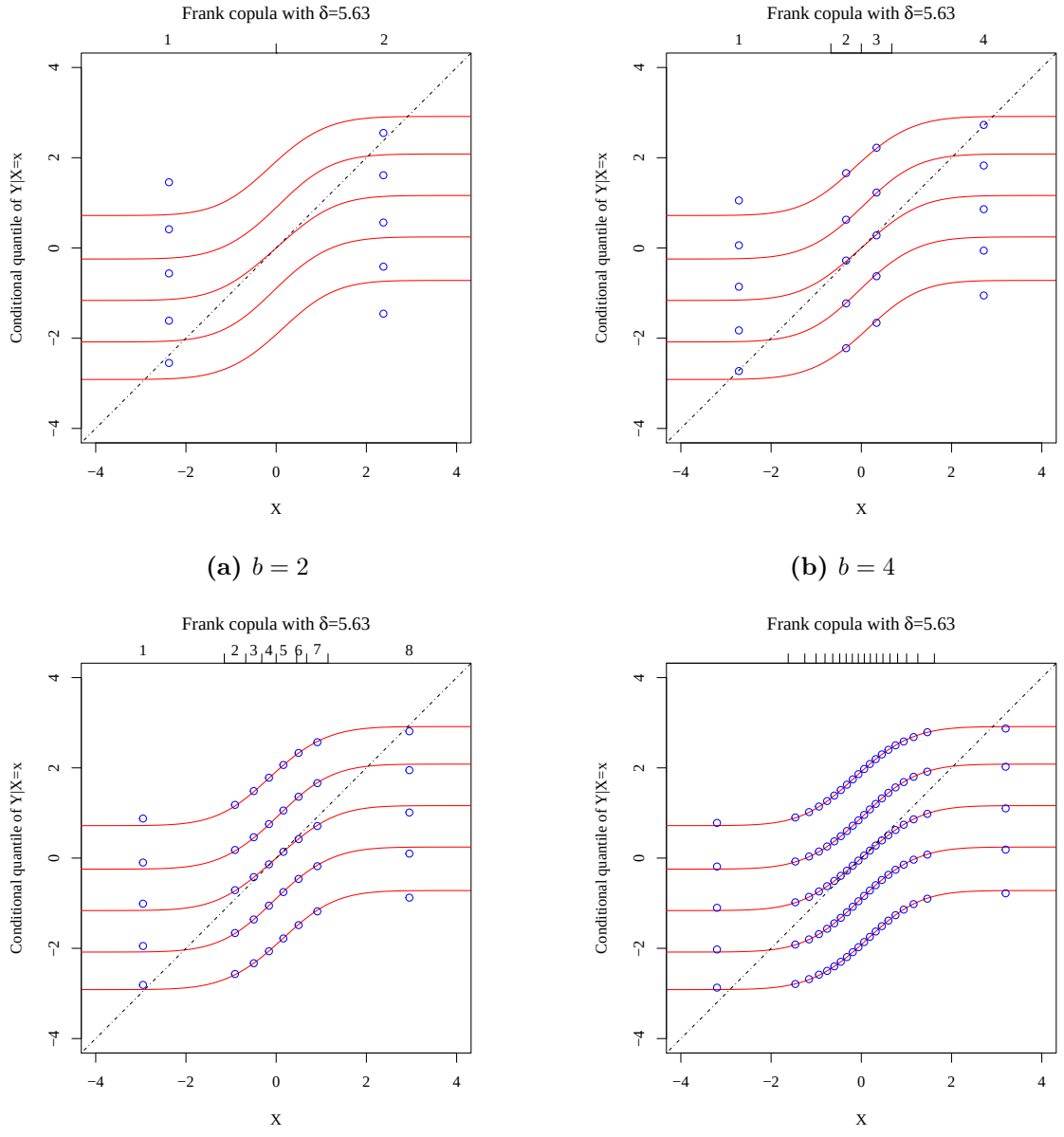
7.1 List of copula functions

Table 7.1: Bivariate copulas: expressions for CDF, conditional distribution, conditional quantile function.

Copula	Functions
Gaussian copula, $\rho \in [-1, 1]$	$C(u,v) = \Phi_\rho\big(\Phi^{-1}(u), \Phi^{-1}(v)\big)$,
Φ_ρ is the CDF of a standard bivariate normal with correlation ρ	$C_{2\|1}(v \mid u) = \Phi\left(\frac{\Phi^{-1}(v) - \rho\,\Phi^{-1}(u)}{\sqrt{1-\rho^2}}\right)$,
	$C_{2\|1}^{-1}(\alpha \mid u) = \Phi\Big(\Phi^{-1}(\alpha)\sqrt{1-\rho^2} + \rho\,\Phi^{-1}(u)\Big)$.
Student t **copula**, $\rho \in [-1, 1]$, $\nu \in \mathbb{N} \setminus \{0\}$	$C(u,v) = t_{\rho,\nu}\big(t_\nu^{-1}(u), t_\nu^{-1}(v)\big)$,
t_ν is the CDF of the Student t with ν degrees of freedom	$C_{2\|1}(v \mid u) = t_{\nu+1}\left(\frac{t_\nu^{-1}(v) - \rho\, t_\nu^{-1}(u)}{\sqrt{(1-\rho^2)\left(1+(t_\nu^{-1}(u))^2/\nu\right)}}\right)$,
$t_{\rho,\nu}$ is the CDF of the bivariate Student t (correlation ρ, degrees of freedom ν)	$C_{2\|1}^{-1}(\alpha \mid u) = t_\nu^{-1}\left(\sqrt{\frac{\nu+(t_\nu^{-1}(u))^2}{\nu+1}}\,(t_{\nu+1}^{-1}(\alpha)) + \rho\, t_\nu^{-1}(u)\right)$.
Frank copula, $\delta \neq 0$	$C(u,v) = -\frac{1}{\delta}\log\Big(1 + \frac{(e^{-\delta u}-1)(e^{-\delta v}-1)}{e^{-\delta}-1}\Big)$,
	$C_{2\|1}(v \mid u) = \frac{e^{-\delta u}(1-e^{-\delta v})}{-e^{-\delta} - e^{-\delta(u+v)} + e^{-\delta u} + e^{-\delta v}}$,
	$C_{2\|1}^{-1}(\alpha \mid u) = -\frac{1}{\delta}\log\left(1 - \frac{\alpha(1-e^{-\delta})}{e^{-\delta u} + \alpha(1-e^{-\delta u})}\right)$.
Clayton copula, $\delta > 0$	$C(u,v) = (u^{-\delta} + v^{-\delta} - 1)^{-1/\delta}$,
	$C_{2\|1}(v \mid u) = \left(u^{-\delta} + v^{-\delta} - 1\right)^{-1/\delta - 1}\, v^{-\delta-1}$,
	$C_{2\|1}^{-1}(\alpha \mid u) = \left(\left(\alpha^{-\frac{\delta}{1+\delta}} - 1\right) u^{-\delta} + 1\right)^{-\frac{1}{\delta}}$.
Gumbel copula, $\delta \geq 1$	$C_G(u,v) = \exp\Big\{-\big[(-\log u)^\delta + (-\log v)^\delta\big]^{1/\delta}\Big\}$
	$C_{2\|1}(v \mid u) = \frac{1}{u}(-\log u)^{\delta-1}\left((-\log u)^\delta + (-\log v)^\delta\right)^{\frac{1-\delta}{\delta}} \mathrm{C}_G(u,v)$
Joe copula, $\theta \geq 1$	$C_J(u,v) = 1 - \big[(1-u_1)^\theta + (1-u_2)^\theta - (1-u_1)^\theta(1-u_2)^\theta\big]^{\frac{1}{\theta}}$
	$C_{2\|1}(v \mid u) = \left(1 - \mathrm{C}_J(u,v)\right)^{1-\theta}(1-u)^{\theta-1}\left(1-(1-v)^\theta\right)$

7.2 Appendix to Section 4

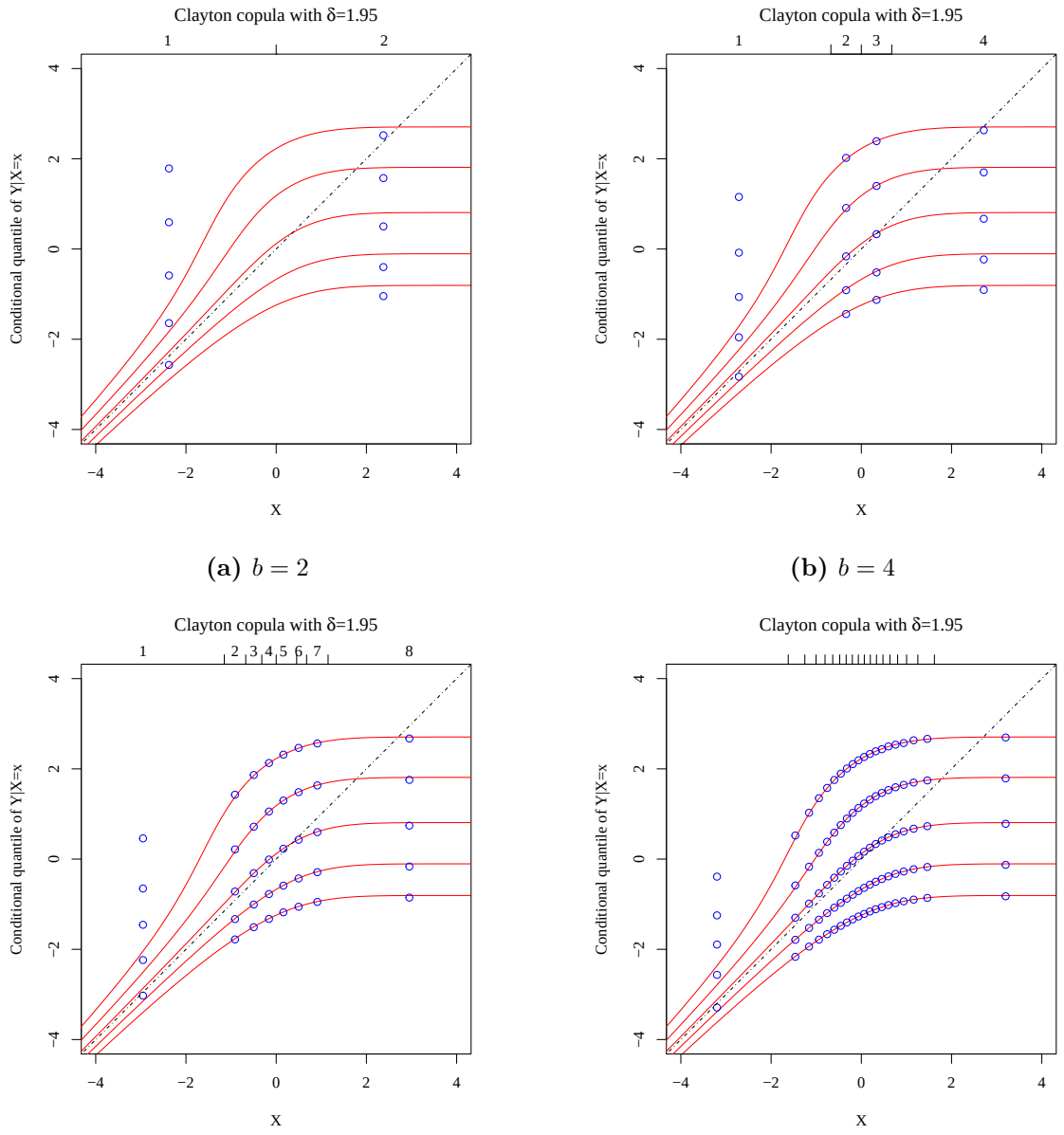
7.2.1 Conditional quantile plots, Y standard normal



(a) $b = 2$ **(b)** $b = 4$

(c) $b = 8$ **(d)** $b = 20$

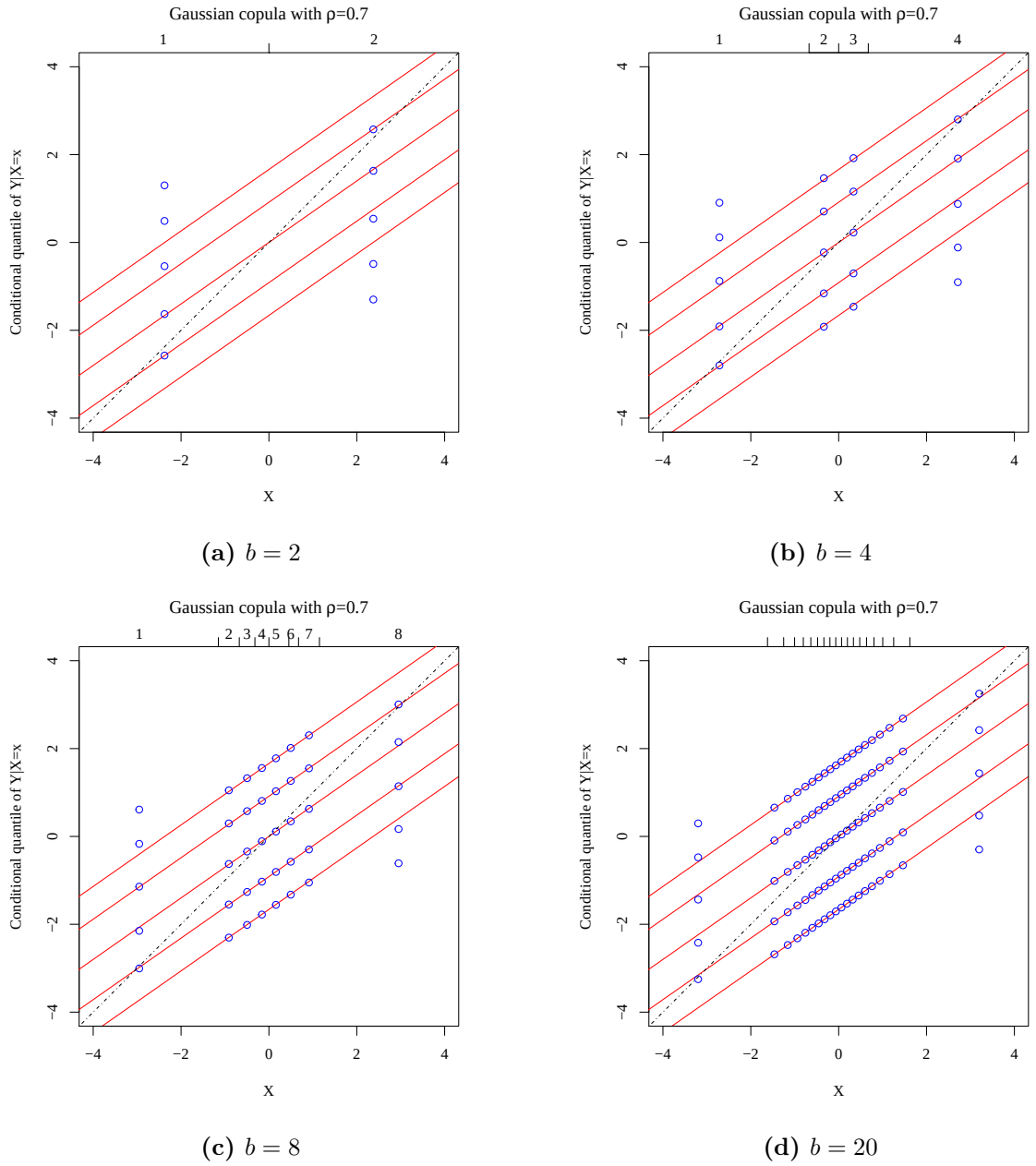
Figure 7.1: Plots of the conditional quantile function of the Frank copula with $\delta = 5.63$ (corresponding $\tau = 0.49$), for quantile $\tau \in \{0.01, 0.1, 0.5, 0.9, 0.99\}$. On the lower x-axis, the conditioning value of X in the continuous case, the corresponding conditional quantile functions in red lines. On the top x-axis, the conditional value of X in the discretized case, with the corresponding conditional quantile values given as blue circle in the middle of the associated bin. The marginal distribution of Y is standard normal. The values of the conditional quantiles were obtained numerically through root-finding.



(a) $b = 2$ **(b)** $b = 4$

(c) $b = 8$ **(d)** $b = 20$

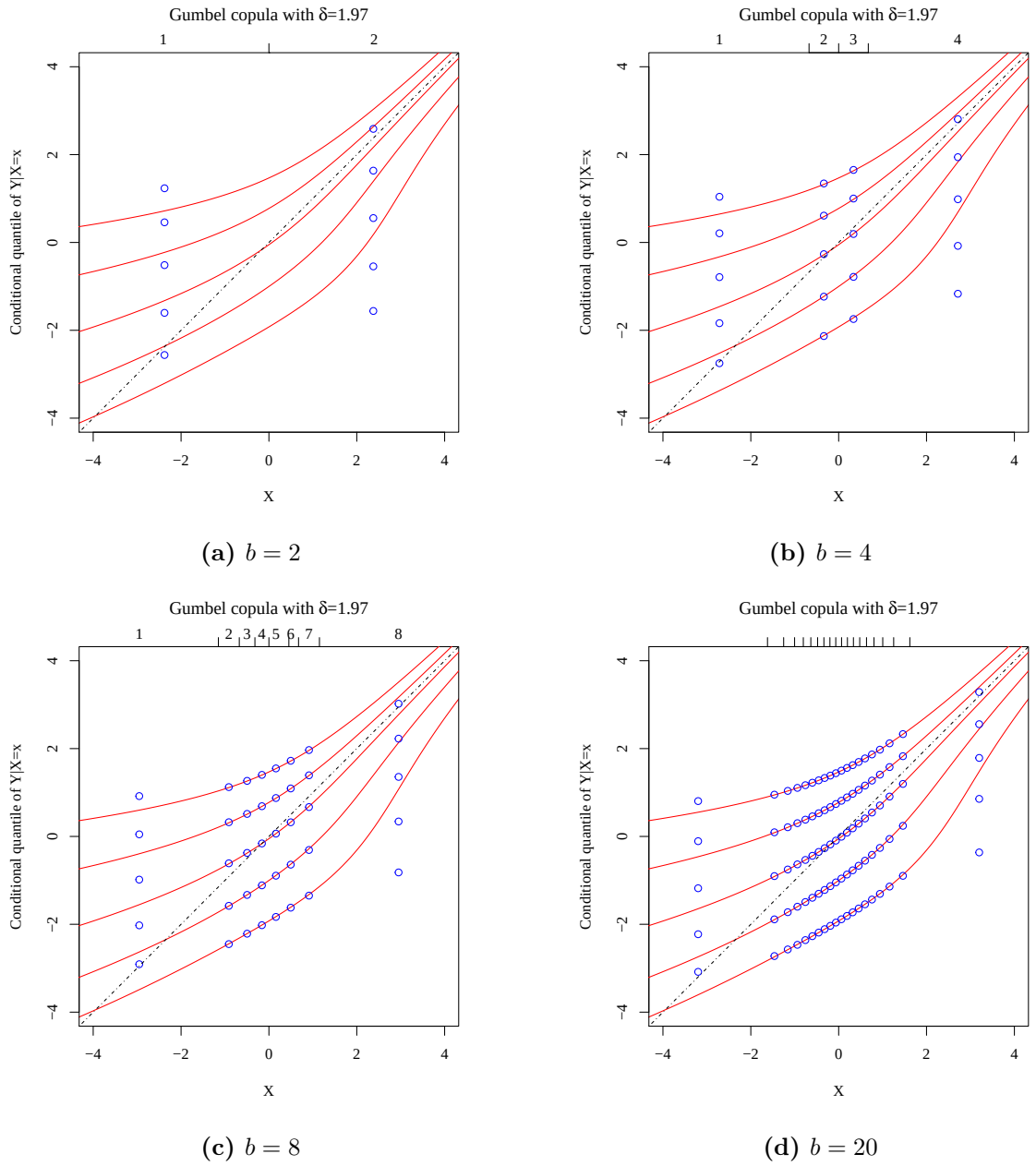
Figure 7.2: Plots of the conditional quantile function of the Clayton copula with $\delta = 1.95$ (corresponding $\tau = 0.49$), for quantile $\tau \in \{0.01, 0.1, 0.5, 0.9, 0.99\}$. On the lower x-axis, the conditioning value of X in the continuous case, the corresponding conditional quantile functions in red lines. On the top x-axis, the conditional value of X in the discretized case, with the corresponding conditional quantile values given as blue circle in the middle of the associated bin. The marginal distribution of Y is standard normal. The values of the conditional quantiles were obtained numerically through root-finding.



(a) $b = 2$ **(b)** $b = 4$

(c) $b = 8$ **(d)** $b = 20$

Figure 7.3: Plots of the conditional quantile function of the Gaussian copula with $\rho = 0.7$ (corresponding $\tau = 0.49$), for quantile $\tau \in \{0.01, 0.1, 0.5, 0.9, 0.99\}$. On the lower x-axis, the conditioning value of X in the continuous case, the corresponding conditional quantile functions in red lines. On the top x-axis, the conditional value of X in the discretized case, with the corresponding conditional quantile values given as blue circle in the middle of the associated bin. The marginal distribution of Y is standard normal. The values of the conditional quantiles were obtained numerically through root-finding.

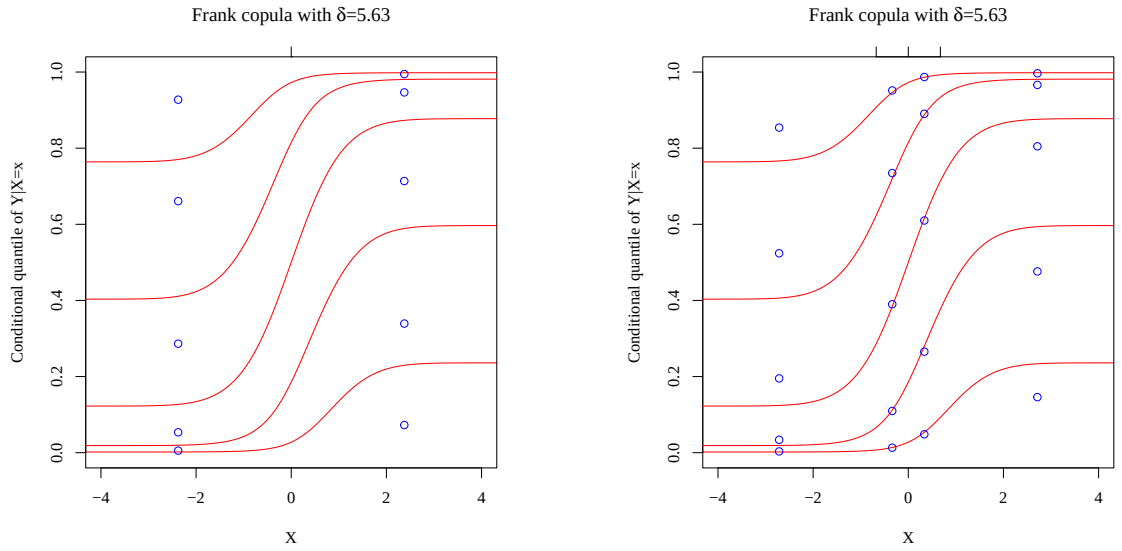


(a) $b = 2$ **(b)** $b = 4$

(c) $b = 8$ **(d)** $b = 20$

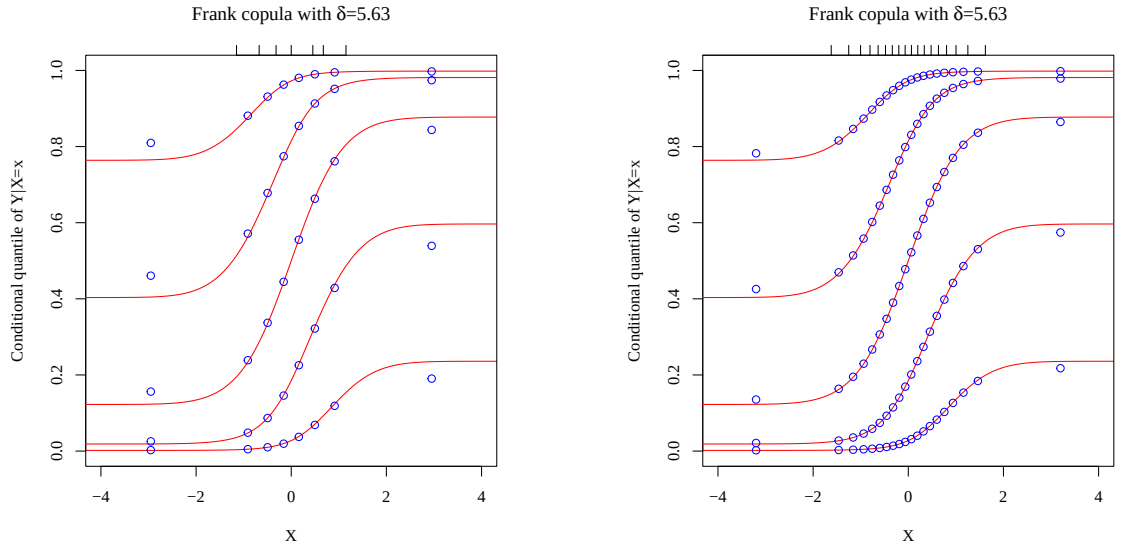
Figure 7.4: Plots of the conditional quantile function of the Gumbel copula with $\delta = 1.97$ (corresponding $\tau = 0.49$), for quantile $\tau \in \{0.01, 0.1, 0.5, 0.9, 0.99\}$. On the lower x-axis, the conditioning value of X in the continuous case, the corresponding conditional quantile functions in red lines. On the top x-axis, the conditional value of X in the discretized case, with the corresponding conditional quantile values given as blue circle in the middle of the associated bin. The marginal distribution of Y is standard normal. The values of the conditional quantiles were obtained numerically through root-finding.

7.2.2 Conditional quantile plots, Y uniform



(a) $b = 2$

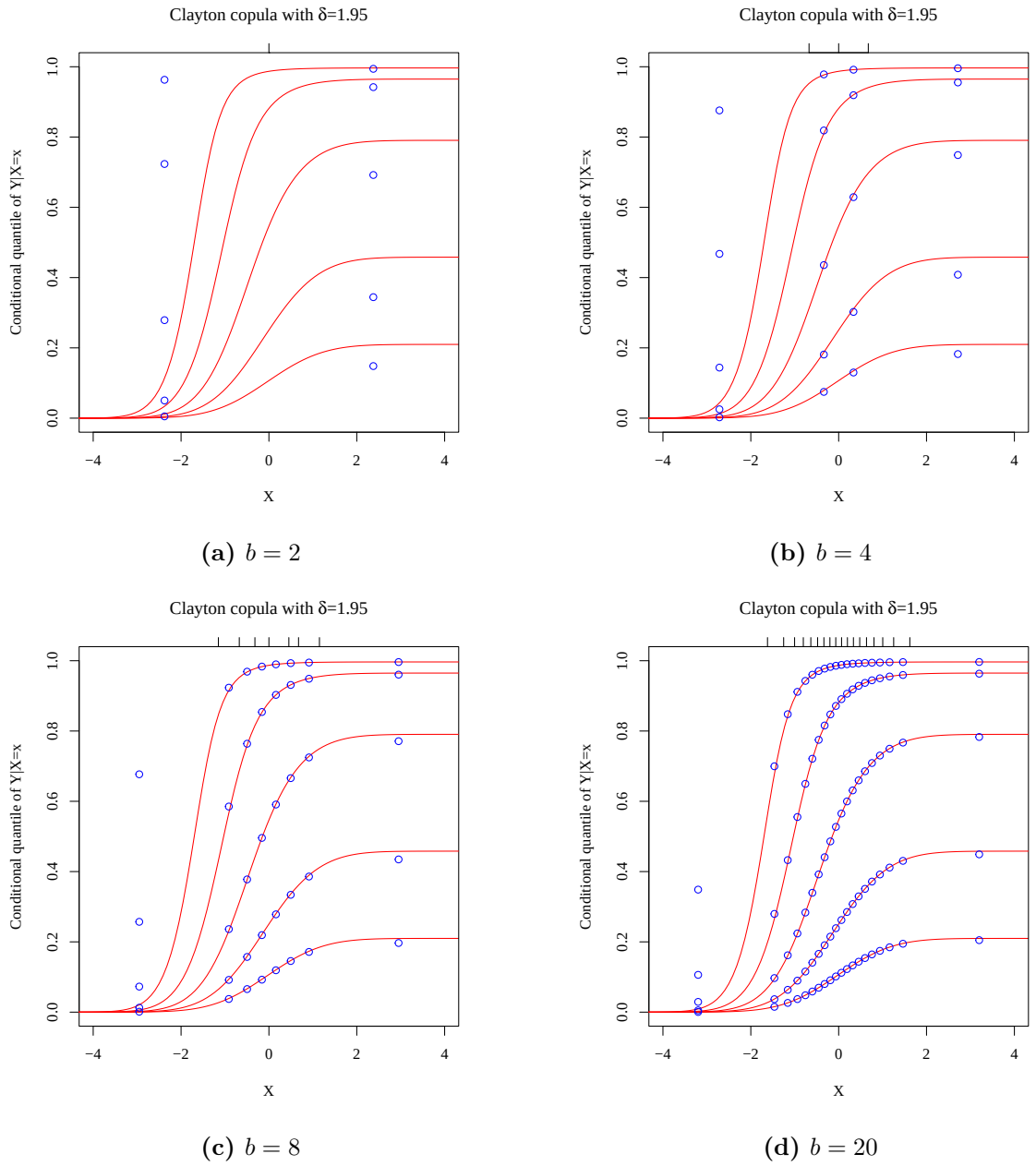
(b) $b = 4$



(c) $b = 8$

(d) $b = 20$

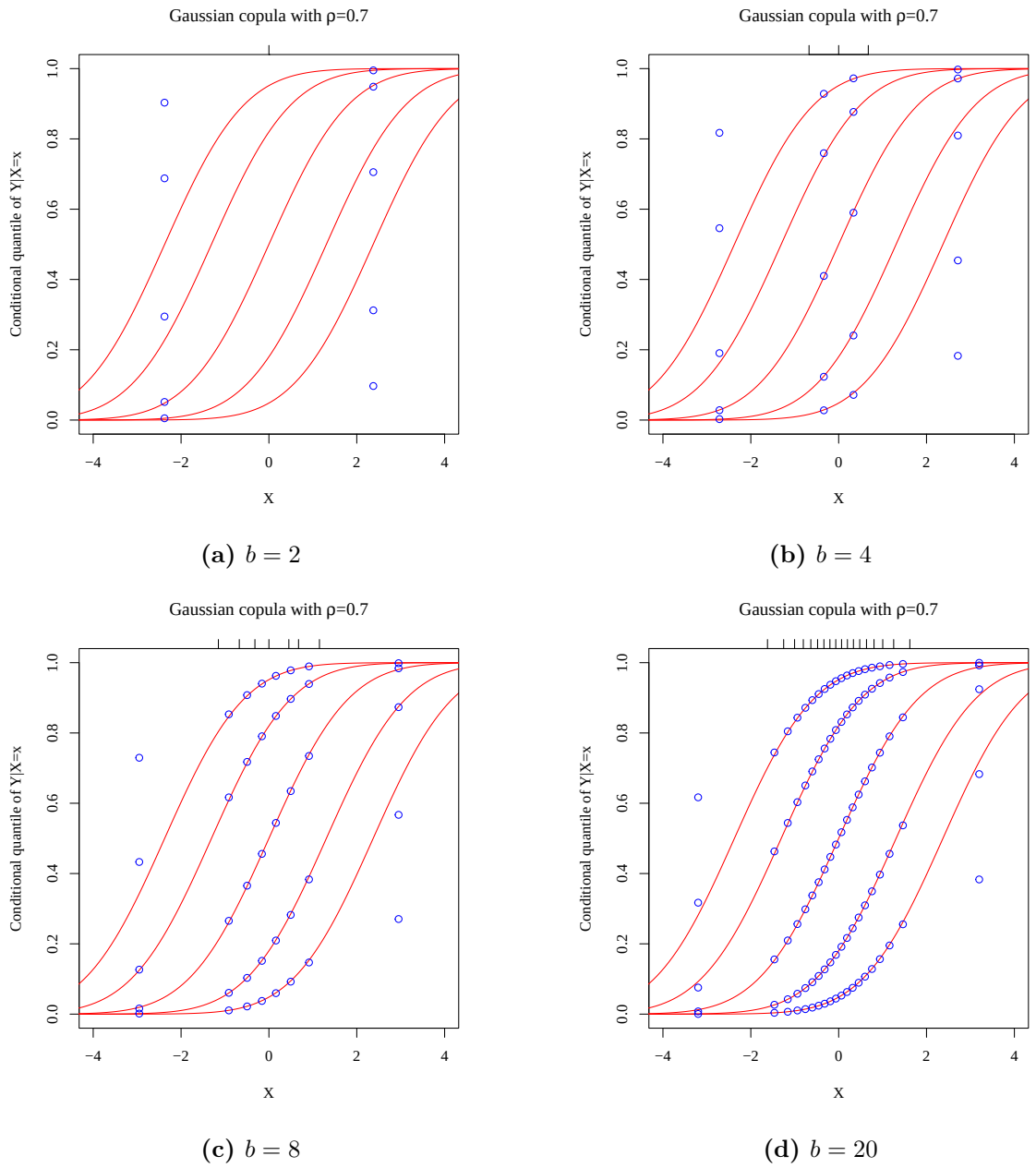
Figure 7.5: Plots of the conditional quantile function of the Frank copula with $\delta = 5.63$ (corresponding $\tau = 0.49$), for quantile $\tau \in \{0.01, 0.1, 0.5, 0.9, 0.99\}$. On the lower x-axis, the conditioning value of X in the continuous case, the corresponding conditional quantile functions in red lines. On the top x-axis, the conditional value of X in the discretized case, with the corresponding conditional quantile values given as blue circle in the middle of the associated bin. The marginal distribution o Y is uniform. The values of the conditional quantiles were obtained numerically through root-finding.



(a) $b = 2$ **(b)** $b = 4$

(c) $b = 8$ **(d)** $b = 20$

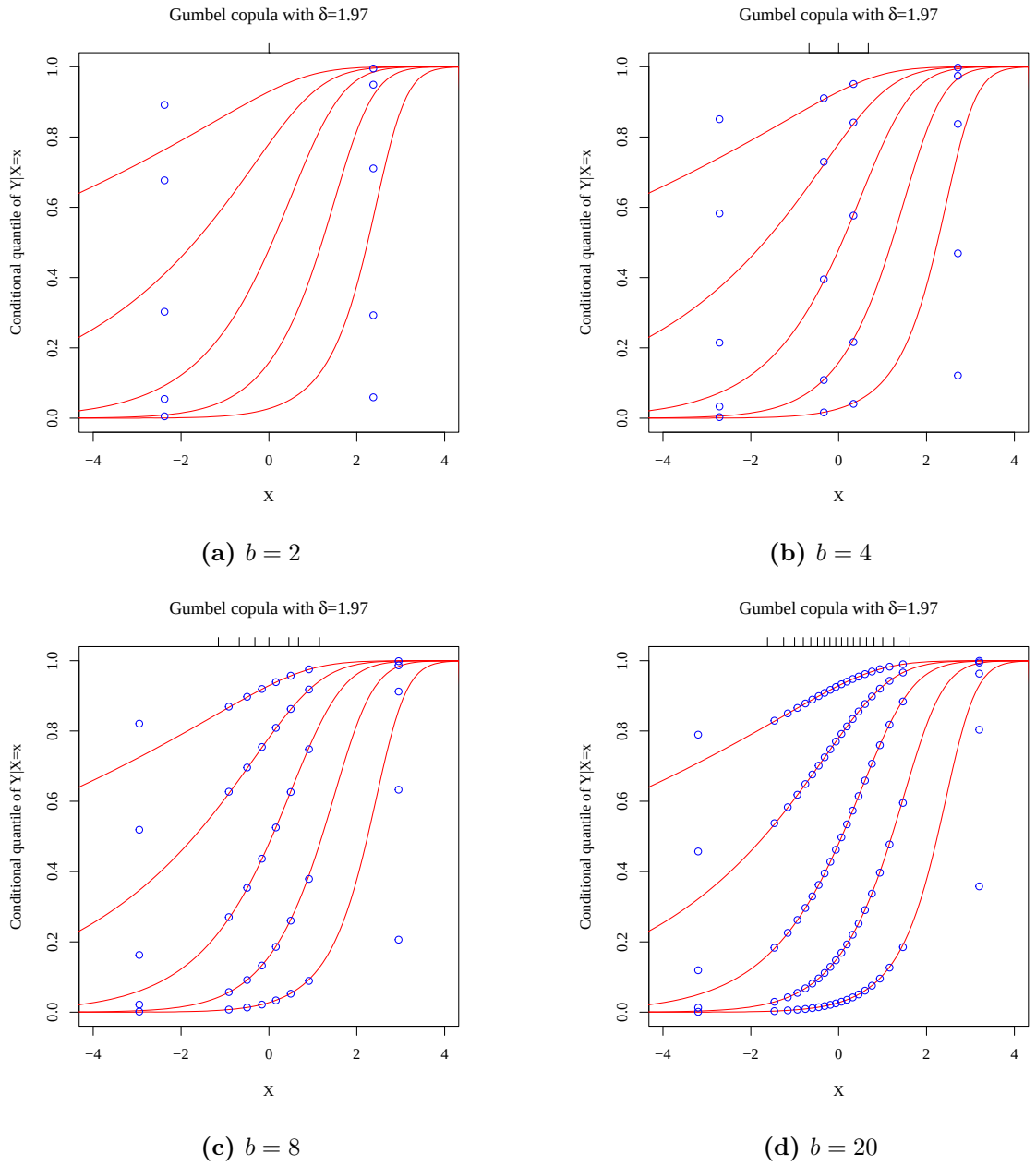
Figure 7.6: Plots of the conditional quantile function of the Clayton copula with $\delta = 1.95$ (corresponding $\tau = 0.49$), for quantile $\tau \in \{0.01, 0.1, 0.5, 0.9, 0.99\}$. On the lower x-axis, the conditioning value of X in the continuous case, the corresponding conditional quantile functions in red lines. On the top x-axis, the conditional value of X in the discretized case, with the corresponding conditional quantile values given as blue circle in the middle of the associated bin. The marginal distribution of Y is uniform. The values of the conditional quantiles were obtained numerically through root-finding.



(a) $b = 2$ **(b)** $b = 4$

(c) $b = 8$ **(d)** $b = 20$

Figure 7.7: Plots of the conditional quantile function of the Gaussian copula with $\rho = 0.7$ (corresponding $\tau = 0.49$), for quantile $\tau \in \{0.01, 0.1, 0.5, 0.9, 0.99\}$. On the lower x-axis, the conditioning value of X in the continuous case, the corresponding conditional quantile functions in red lines. On the top x-axis, the conditional value of X in the discretized case, with the corresponding conditional quantile values given as blue circle in the middle of the associated bin. The marginal distribution of Y is uniform. The values of the conditional quantiles were obtained numerically through root-finding.



(a) $b = 2$ **(b)** $b = 4$

(c) $b = 8$ **(d)** $b = 20$

Figure 7.8: Plots of the conditional quantile function of the Gumbel copula with $\delta = 1.97$ (corresponding $\tau = 0.49$), for quantile $\tau \in \{0.01, 0.1, 0.5, 0.9, 0.99\}$. On the lower x-axis, the conditioning value of X in the continuous case, the corresponding conditional quantile functions in red lines. On the top x-axis, the conditional value of X in the discretized case, with the corresponding conditional quantile values given as blue circle in the middle of the associated bin. The marginal distribution of Y is uniform. The values of the conditional quantiles were obtained numerically through root-finding.

7.3 Appendix to Section 5

7.3.1 Creating conversion plots from conditional log-likelihood plots of Figure 5.4, normal marginals

With some empirical testing, functions of the form $f(x) = ax^2 + bx^8 + cx^{10} + d$ were found to have lowest MSE when fit on the CLL plots of Figure 5.4.

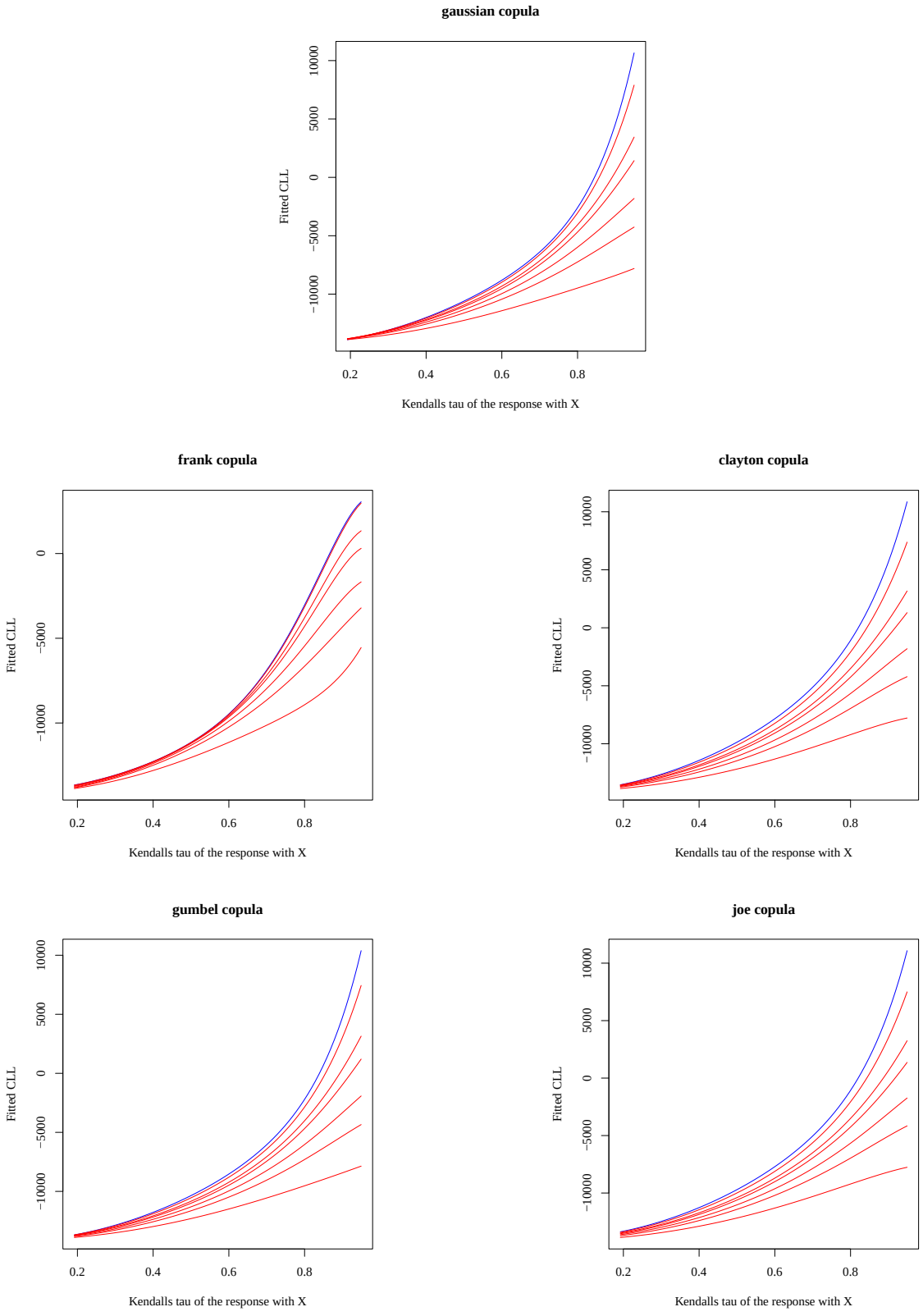


Figure 7.9: Both Y and X have $\mathcal{N}(0,1)$ marginal distributions. Dependence structure of Y and X is given in each subplot. Presented on the x-axis is Kendall's τ of the response Y and continuous covariate X. The blue line represents the function fitted on CLL of Y given X of Figure 5.4. The red lines represent the function fitted on CLL of Y given discretized X of Figure 5.4. The red lines correspond, from top to bottom, to $b = 2, 3, 4, 6, 8, 20$.

With these fitted functions, the conversion plots can be made by evaluating the fitted function of a discretized covariate at some Kendall's τ_1, and finding the value of Kendall's τ_2 such that continuous covariate at τ_2 has the same CLL as the discretized covariate at τ_1. For instance, the fitted function of the Gaussian subplot shows a discretized covariate with $b = 2$ and Kendall's tau (between response and its latent variable) equal to 0.6 has the same CLL as a continuous covariate with Kendall's tau equal to 0.4. So in the conversion plot, the line corresponding to $b = 2$ will take value 0.4 on the y-axis at 0.6 on the x-axis.

7.3.2 Conversion plots for conditional log-likelihood, uniform marginals

Below are the conversion plots for the conditional log-likelihood, with Y and X uniform.

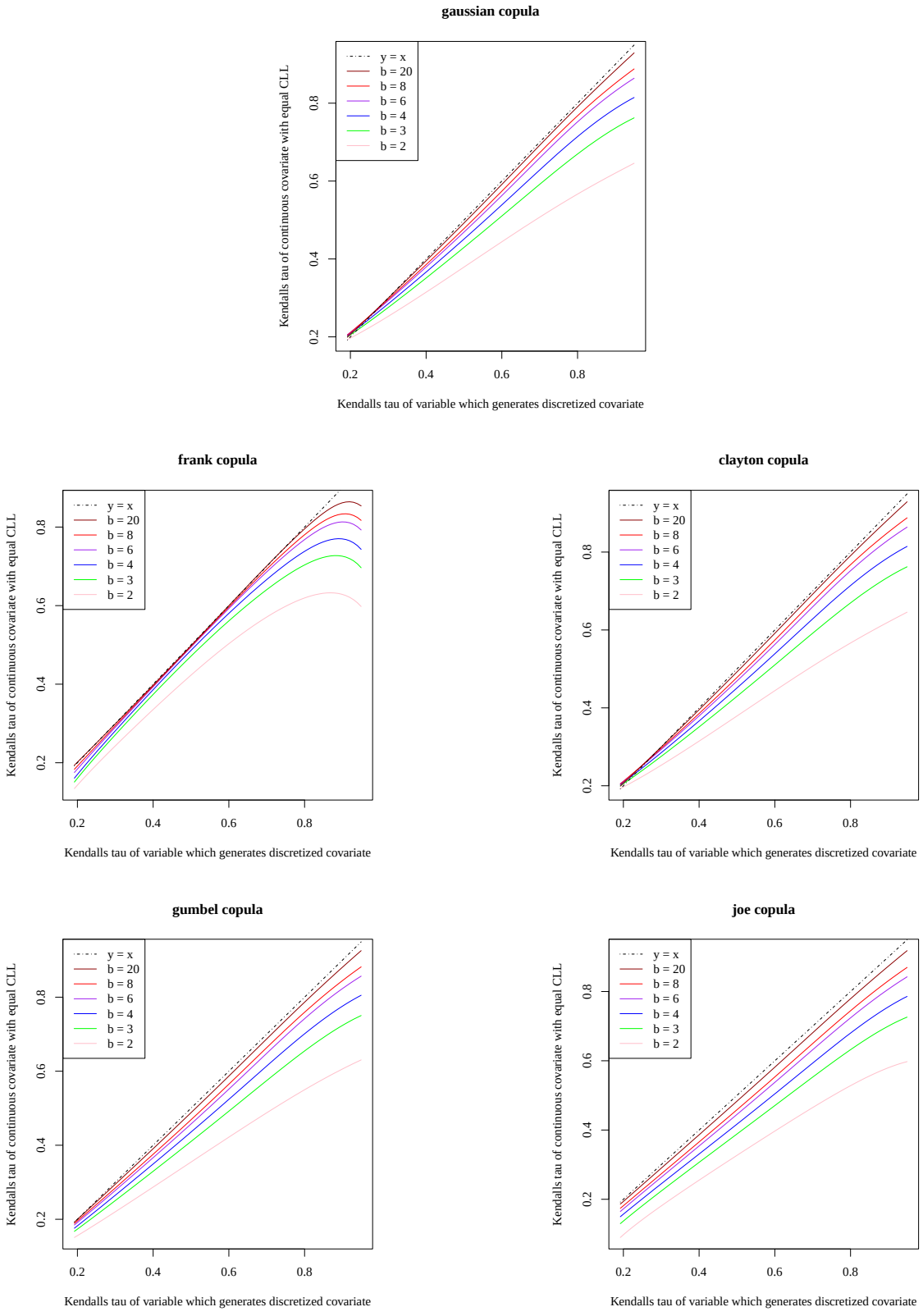


Figure 7.10: Both Y and X have Unif$[0, 1]$ marginal distributions. Dependence structure of Y and X is given in the subtitle of each plot. Presented on the x-axis is Kendall's τ of the response and continuous covariate X, which determines the parameter of the copula between Y and X. The blue line represents CLL of Y and X. The red lines represent CLL between Y and discretized X. The red lines correspond, from lowest to highest, to $b = 2, 3, 4, 6, 8, 20$.

7.3.3 Creating conversion plots from check-loss ($\alpha = 0.05$) plots of Figure 7.11

The simulation study gave the check-loss ($\alpha = 0.05$) plots posted below.

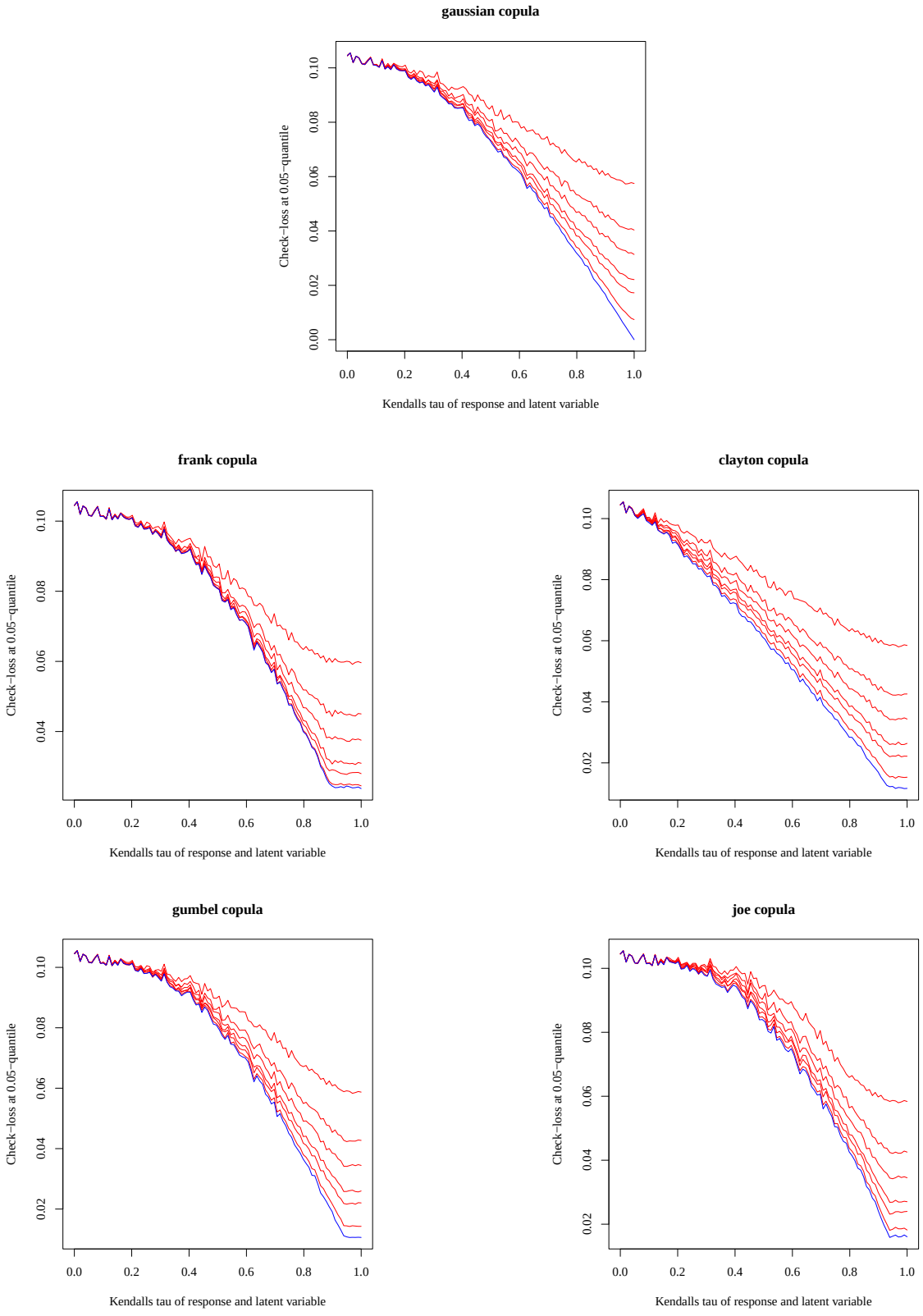


Figure 7.11: Both Y and X have $\mathcal{N}(0,1)$ marginal distributions. Dependence structure of Y and X is given in the subtitle of each plot. Presented on the x-axis is Kendall's τ of the response Y and continuous covariate X, which determines the parameter of the copula between Y and X. Check-loss was computed for $\tau = 0.01, 0.02, 0.03, \ldots, 0.99, 1$, but curves were drawn in a continuous manner by linear interpolation between points. The blue line represents check-loss at $\alpha = 0.05$ for prediction of Y given X. The red lines represent check-loss at $\alpha = 0.05$ for prediction of Y given discretized X. The red lines correspond, from top to bottom, to $b = 2, 3, 4, 6, 8, 20$. Number of observations $n = 10.000$.

With some empirical testing, functions of the form $f(x) = ax + bx^2 + c$ were found to have lowest MSE when fit on the check-loss ($\alpha = 0.05$) plots of Figure 7.11.

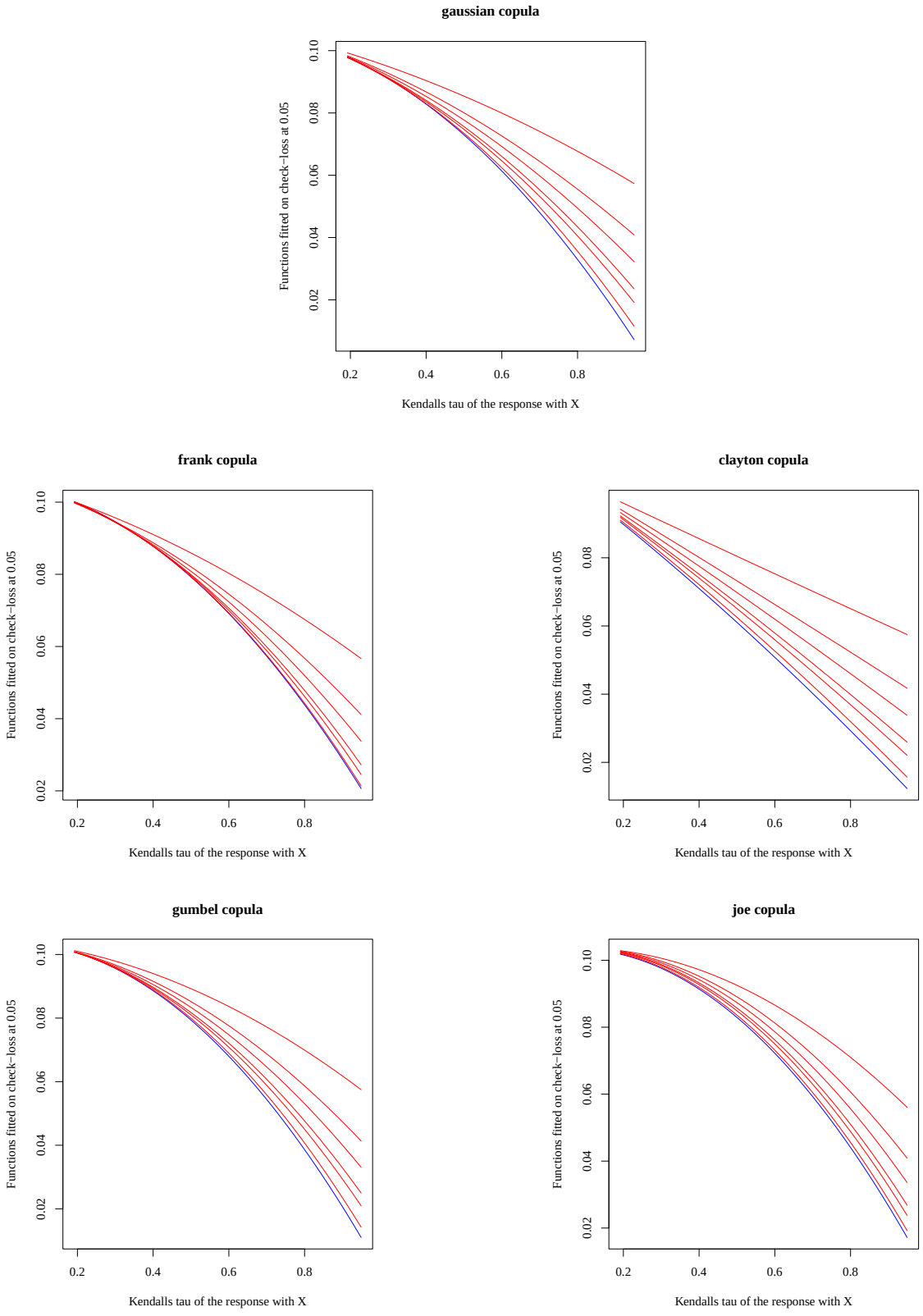


Figure 7.12: Both Y and X have $\mathcal{N}(0,1)$ marginal distributions. Dependence structure of Y and X is given in each subplot. Presented on the x-axis is Kendall's τ of the response Y and continuous covariate X. The blue line represents the function fitted on check-loss ($\alpha = 0.05$) of Y given X of Figure 7.11. The red lines represent the function fitted on CLL of Y given discretized X of Figure 7.11. The red lines correspond, from top to bottom, to $b = 2, 3, 4, 6, 8, 20$.

These fitted functions allow us to create the conversion plots of Figure 5.8, in the manner explained in the previous section.

7.3.4 Conversion plots for check-loss ($\alpha = 0.01$)

Posted below are the conversion plots for check-loss at a more extreme tail quantile, $\alpha = 0.01$. The conversion plots for $\alpha = 0.01$ are nearly identical to the conversion plots for check-loss at $\alpha = 0.05$, showing the amount of bias check-loss has against discretized covariates does not depend significantly on the tail-quantile where check-loss is evaluated. The main difference between the conversion plots for the different values of α is that for $\alpha = 0.01$, the subplot for the Joe copula had check-loss for discretized covariate which started at a higher value than check-loss for continuous covariate, so the conversion from the former to the latter does not exist on the left-hand side of the x-axis.

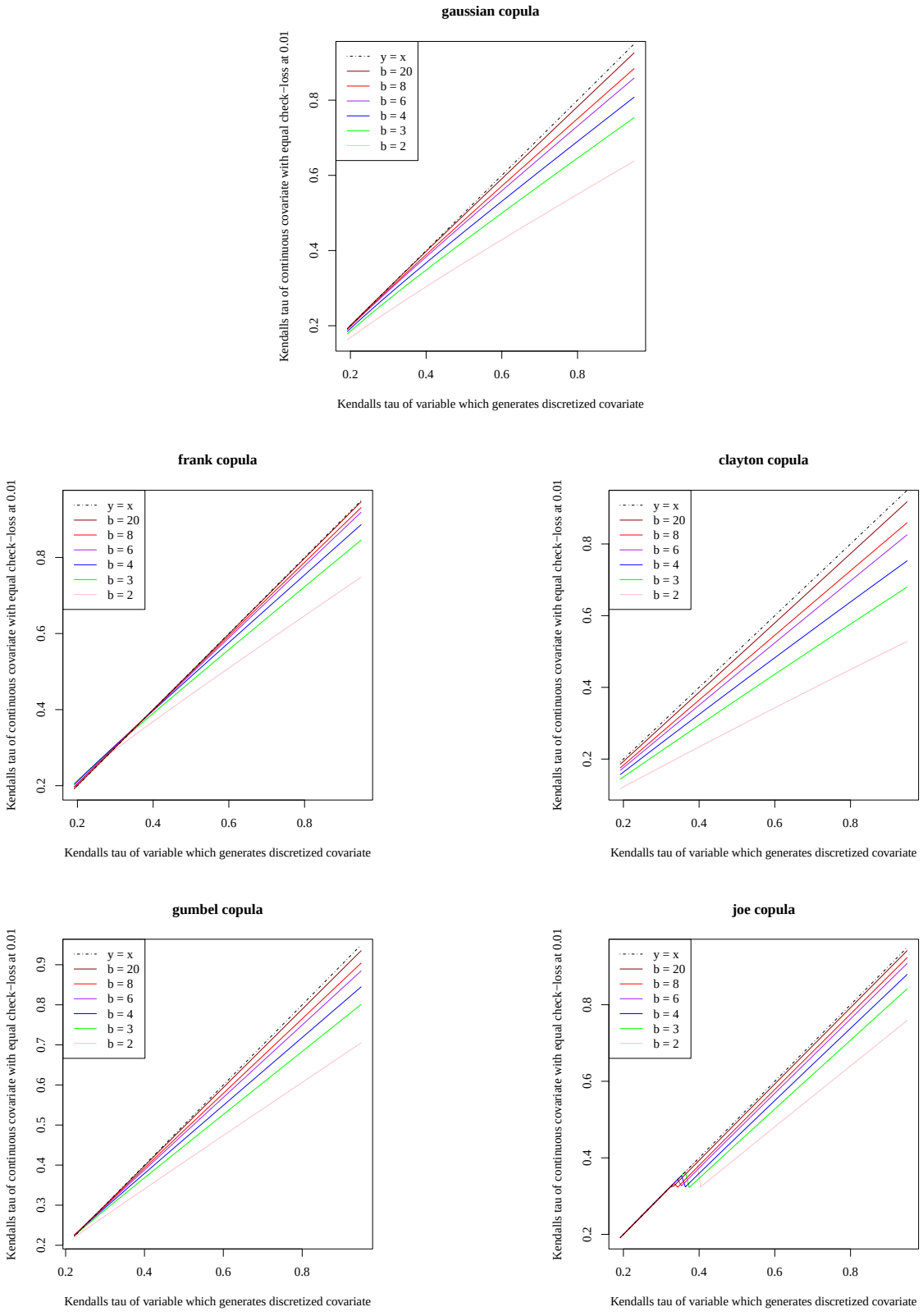


Figure 7.13: Both Y and X have $\mathcal{N}(0,1)$ marginal distributions. Dependence structure of Y and X is given in the subtitle of each plot. Presented on the x-axis is Kendall's τ of the response and continuous covariate X, which determines the parameter of the copula between Y and X. The blue line represents check-loss at $\alpha = 0.01$ of Y given X. The red lines represent check-loss at $\alpha = 0.01$ of Y given discretized X. The red lines correspond, from lowest to highest, to $b = 2, 3, 4, 6, 8, 20$.

7.3.5 Conversion plots for check-loss ($\alpha = 0.05$), uniform marginals

To observe whether the behaviour/bias of check-loss at $\alpha = 0.05$ changes when the marginal distributions of Y and X change, we plot the conversion plots below for uniform marginals.

Note the subplot for the Joe copula had fitted check-loss for discretized covariate which started at a higher value than fitted check-loss for continuous covariate, so the conversion from the former to the latter does not exist on the left-hand side of the x-axis.

The conversion plots with Y, X uniform below are hard to distinguish from the conversion plots with Y, X normal, supporting the idea that the amount of bias check-loss at $\alpha = 0.05$ has against discretized covariates is very consistent for different marginals.

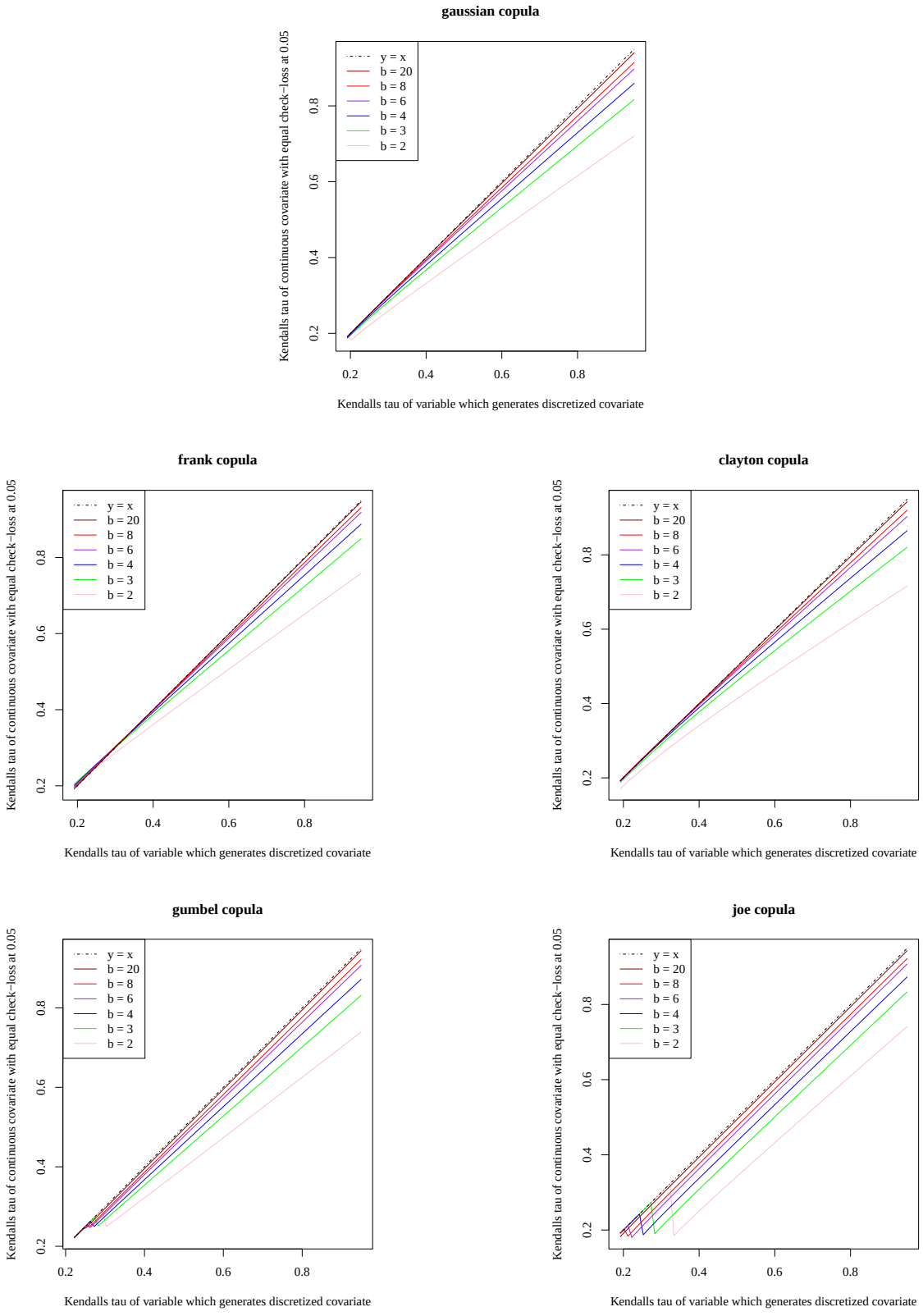


Figure 7.14: Both Y and X have Unif$[0, 1]$ marginal distributions. Dependence structure of Y and X is given in the subtitle of each plot. Presented on the x-axis is Kendall's τ of the response and continuous covariate X, which determines the parameter of the copula between Y and X. The blue line represents check-loss at $\alpha = 0.05$ of Y given X. The red lines represent check-loss at $\alpha = 0.05$ of Y given discretized X. The red lines correspond, from lowest to highest, to $b = 2, 3, 4, 6, 8, 20$.

References

[1] Claudia Czado. "Analyzing dependent data with vine copulas". In: *Lecture Notes in Statistics, Springer* 222 (2019).

[2] Israel Cohen et al. "Pearson correlation coefficient". In: *Noise reduction in speech processing* (2009), pp. 1–4.

[3] Fritz Drasgow. "Polychoric and polyserial correlations". In: *Encyclopedia of statistical sciences* (2004).

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

[5] Gabriel Frahm, Markus Junker, and Rafael Schmidt. "Estimating the tail-dependence coefficient: properties and pitfalls". In: *Insurance: mathematics and Economics* 37.1 (2005), pp. 80–100.

[6] M Sklar. "Fonctions de répartition à n dimensions et leurs marges". In: *Annales de l'ISUP*. Vol. 8. 3. 1959, pp. 229–231.

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

[8] Roger B Nelsen. *An introduction to copulas*. Springer, 2006.

[9] Christian Genest and Johanna Nešlehová. "A primer on copulas for count data". In: *ASTIN Bulletin: The Journal of the IAA* 37.2 (2007), pp. 475–515.

[10] Harry Joe. "Families of m-variate distributions with given margins and m (m-1)/2 bivariate dependence parameters". In: *Lecture notes-monograph series* (1996), pp. 120–141.

[11] Kjersti Aas et al. "Pair-copula constructions of multiple dependence". In: *Insurance: Mathematics and economics* 44.2 (2009), pp. 182–198.

[12] Harry Joe and James Jianmeng Xu. "The estimation method of inference functions for margins for multivariate models". In: (1996).

[13] Oswaldo Morales-Napoles. "Counting vines". In: *Dependence modeling: Vine copula handbook*. World Scientific, 2010, pp. 189–218.

[14] Tim Bedford and Roger M Cooke. "Vines–a new graphical model for dependent random variables". In: *The Annals of Statistics* 30.4 (2002), pp. 1031–1068.

[15] Anastasios Panagiotelis, Claudia Czado, and Harry Joe. "Pair copula constructions for multivariate discrete data". In: *Journal of the American Statistical Association* 107.499 (2012), pp. 1063–1072.

[16] Roger M Cooke, Dorota Kurowicka, and K Wilson. "Sampling, conditionalizing, counting, merging, searching regular vines". In: *Journal of Multivariate Analysis* 138 (2015), pp. 4–18.

[17] Murray Rosenblatt. "Remarks on a multivariate transformation". In: *The annals of mathematical statistics* 23.3 (1952), pp. 470–472.

[18] Roger Koenker and Kevin F Hallock. "Quantile regression". In: *Journal of economic perspectives* 15.4 (2001), pp. 143–156.

[19] Jeffrey Dissmann et al. "Selecting and estimating regular vine copulae and application to financial returns". In: *Computational Statistics & Data Analysis* 59 (2013), pp. 52–69.

[20] Claudia Czado, Stephan Jeske, and Mathias Hofmann. "Selection strategies for regular vine copulae". In: *Journal de la Société Française de Statistique* 154.1 (2013), pp. 174–191.

[21] Daniel Kraus and Claudia Czado. "Growing simplified vine copula trees: improving Dissmann's algorithm". In: *arXiv preprint arXiv:1703.05203* (2017).

[22] Eike C Brechmann and Harry Joe. "Truncation of vine copulas using fit indices". In: *Journal of Multivariate Analysis* 138 (2015), pp. 19–33.

[23] Bo Chang, Shenyi Pan, and Harry Joe. "Vine copula structure learning via Monte Carlo tree search". In: *The 22nd International Conference on Artificial Intelligence and Statistics*. PMLR. 2019, pp. 353–361.

[24] Dorota Kurowicka. "Optimal truncation of vines". In: *Dependence modeling: Vine copula handbook*. World Scientific, 2010, pp. 233–247.

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

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

[27] Kailun Zhu, Dorota Kurowicka, and Gabriela F Nane. "Simplified R-vine based forward regression". In: *Computational Statistics & Data Analysis* 155 (2021), p. 107091.

[28] Marija Tepegjozova et al. "Nonparametric C-and D-vine-based quantile regression". In: *Dependence Modeling* 10.1 (2022), pp. 1–21.

[29] Marija Tepegjozova. "D-and C-vine quantile regression for large data sets". In: (2019).

[30] Özge Sahin and Claudia Czado. "High-dimensional sparse vine copula regression with application to genomic prediction". In: *Biometrics* 80.1 (2024), ujad042.

[31] Anastasios Panagiotelis et al. "Model selection for discrete regular vine copulas". In: *Computational Statistics & Data Analysis* 106 (2017), pp. 138–152.

[32] Thomas Nagler. "A generic approach to nonparametric function estimation with mixed data". In: *Statistics & Probability Letters* 137 (2018), pp. 326–330.

[33] Thomas Nagler, Christian Schellhase, and Claudia Czado. "Nonparametric estimation of simplified vine copula models: comparison of methods". In: *Dependence Modeling* 5.1 (2017), pp. 99–120.

[34] Niklas Schallhorn et al. "D-vine quantile regression with discrete variables". In: *arXiv preprint arXiv:1705.08310* (2017).

[35] Gery Geenens, Arthur Charpentier, and Davy Paindaveine. "Probit transformation for nonparametric kernel estimation of the copula density". In: (2017).

[36] Aristidis K Nikoloulopoulos. "Efficient estimation of high-dimensional multivariate normal copula models with discrete spatial responses". In: *Stochastic environmental research and risk assessment* 30.2 (2016), pp. 493–505.

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

—-