Copulas and Simplifying Assumption

Monica Li

June 22, 2024

Contents

1 Introduction

This report will focus on Copulas. It is meant as an introduction to the theory of Copulas. It should guide the reader from fairly basic principles to a somewhat complex model in Vine Copulas. There is also some commentary on the modeling and relevant to the former class of models, and the assumptions inherent. For some prerequisite mathematical definitions and theorems, please refer to Appendix [8].

2 Copulas

2.1 Copulas Definition

Let $[0,1]^d$ be the d-dimensional hypercube. A copula C is the distribution function on the hypercube with uniformly distributed marginals.

- A d-dimensional copula C is a multivariate distribution function on the d-dimensional hypercube $[0,1]^d$ with uniformly distributed marginals.

- The corresponding copula density for an absolutely continuous copula, denoted by c, can be obtained by partial differentiation, i.e., $c(\mathbf{u}) := \frac{\partial d}{\partial u_1,...,u_d} C(\mathbf{u})$ for all $\mathbf{u}$ in $[0,1]^d$.

We also often think of the copula density, which is the standard multivariate density function, i.e., $c(\mathbf{u}) := \frac{\partial d}{\partial u_1,...,u_d} C(\mathbf{u})$
When the marginals of variables are standardized to uniform distribution, 'copulas' characterizing the dependence between random variables. Separates the dependence between the components from the marginal distributions [1].

2.2 Sklar's Theorem

Theorem 1. *Sklar's Theorem*
Let X be a d-dimensional random vector with joint distribution function F and marginal distribution functions F_i, $i = 1, ..., d$, then the joint distribution function can be expressed as

$$F(x_1, ..., x_d) = C(F_1(x_1), ..., F_d(x_d))$$

with associated density or probability mass function

$$f(x_1, ..., x_d) = c(F_1(x_1), ..., F_a(x_d)) f_1(x_1) ... f_a(x_d)$$

for some d-dimensional copula C with copula density c. For absolutely continuous distributions, the copula C is unique.
The inverse also holds: the copula corresponding to a multivariate distribution function F with marginal distribution functions F_i for $i = 1, ..., d$ can be expressed as

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

and its copula density or probability mass function is determined by

$$c(u_1, ..., u_d) = \frac{f(F_1^{-1}(u_1), ..., F_d^{-1}(u_d))}{f_1(F_1^{-1}(u_1)) ... f_d(F_d^{-1}(u_d))}$$

3 Vines

The following two definitions are the same, and each offer their own conceptual advantages. The key difference lies in the proximity condition. Additionally, the two definitions use slightly different notations and terminologies.

3.1 Regular (R-) vine tree sequence - Czado Definition [1]

A set of trees $\{T_1, ..., T_{d-1}\}$ forms a regular vine tree sequence $V(T_1, ..., T_{d-1})$ on d elements if:

- Connected: Each tree $T_j = (N_j, E_j)$ is connected, meaning that there exists a path of nodes $a, n_1, ..., n_k, b \subset N_j$ between every pair of nodes a and b within a tree T_j.
- Initial Tree: The first tree T_1 is a tree with nodes numbered from 1 to d, and it has a set of edges denoted as E_1.
- Subsequent Trees: For $j \geq 2$, these subsequent trees T_j have as node set N_j defined to be the set of edges of the previous tree E_{j-1}. The set of edges is denoted E_j.
- Proximity: For $j = 2, ..., d-1$ and for every edge $\{a, b\}$ in the edge set E_j, it is required that the cardinality of the intersection of nodes in a and b is one. Reminder: $a, b \in E_j$ are sets so $a \cap b$ is a set and an edge in $E_{j-2} = N_{j-1}$.

Simplified Czado Definition:

- Nodes of $T(n)$ = Edges of $T(n-1)$
- You may connect Nodes in $T(n)$ which share one Node in $T(n-1)$

3.2 Regular vine - Joe Definition [2]

Let V be a regular vine on d elements, where $\epsilon(V) = \epsilon_1 \cup ... \cup \epsilon_{d-1}$ denotes the set of edges of V. Then, the vine V satisfies the conditions below:

- V consists of $d-1$ trees, denoted as $\{T_1, ..., T_{d-1}\}$.
- T_1 is a connected tree with nodes $N_1 = \{1, ..., d\}$ and edges ϵ_1.
- For $l = 2, ..., d-1$, T_l is a tree with nodes $N_l = \epsilon_{l-1}$ (the edges in a tree become the nodes in the next tree).
- Proximity: For $l = 2, ..., d-1$, n_1, n_2 denote the set of all elements in the nodes that forming the edge ϵ_l. For any edge ϵ_l, the symmetric difference of n_1 and n_2, denoted by $\#(n_1 \triangle n_2)$, equals 2 (nodes joined in an edge differ by two elements).

Simplified Joe Definition:

- Nodes of T(n) = Edges of T(n-2)
- Connecting Nodes in T(n) which have symmetric difference=2

3.3 An example of Vine

Vines from both Czado's and Joe's point of view

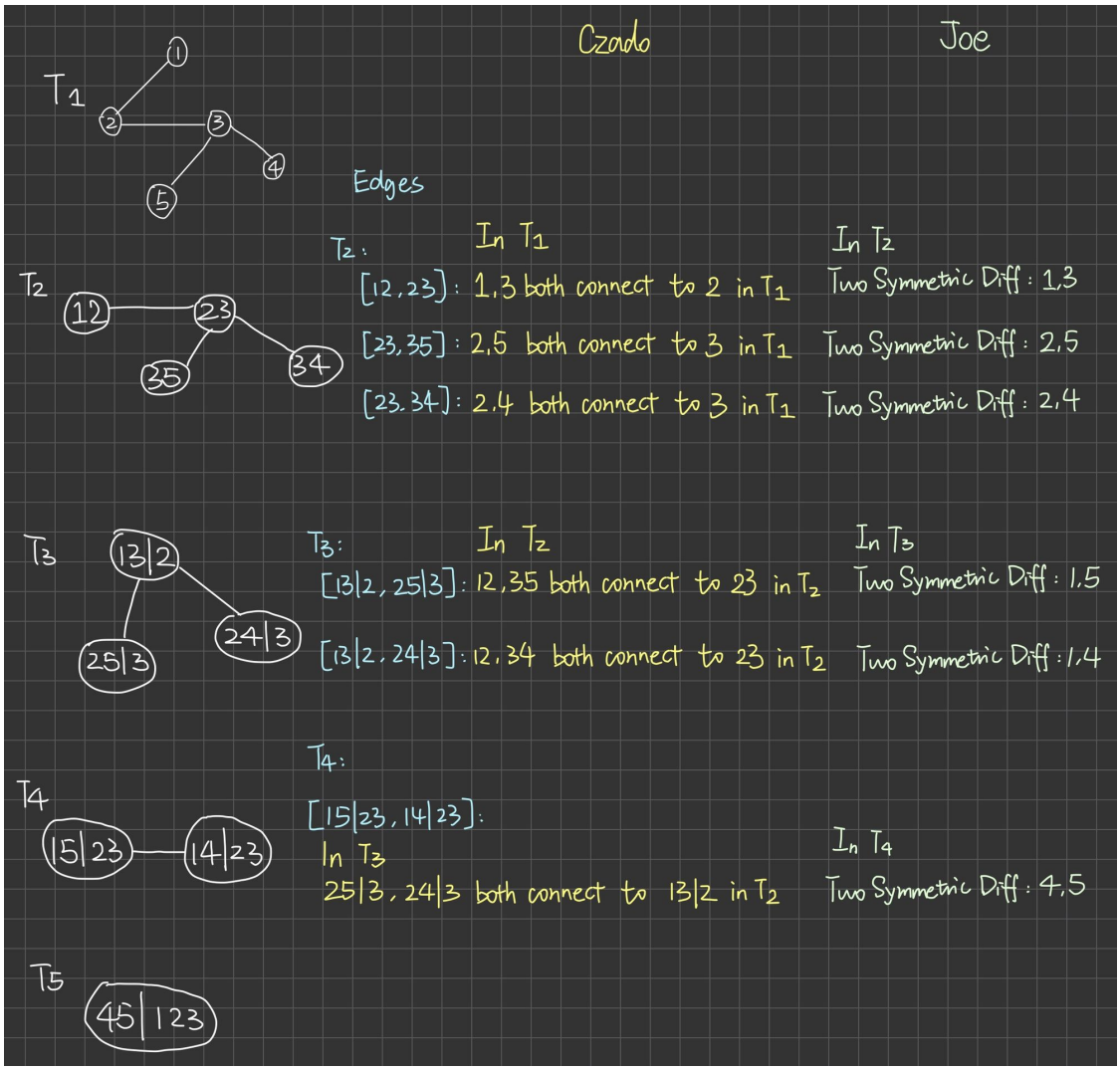

Figure 1: Vine started with a Tree 1 of five Nodes

4 Conditional Distributions

In statistics, when modeling joint distributions, we sometimes use conditioning and/or mixing to create models. This approach allows us to break down complex distributions into simpler, more manageable components by focusing on the relationships between variables. It is useful to consider these methods when thinking about how Vines work, as Vines provide a convenient way to express joint distributions of two or more random variables and their associated conditional distributions. First, we will give some definitions and explanations of key concepts in conditional distributions.

Definition 1. (**Chain Rule for Conditional Probability**)
for any events $A_1, ..., A_n$,

$$P\left[\cap_{i=1}^{n} A_i\right] = Pr[A_1] \times Pr[A_2|A_1] \times Pr[A_3|A_1 \cap A_2] \times \ldots \times Pr[A_n|\prod_{i=1}^{n-1} A_i]$$

Definition 2. (**Chain Rule for Conditional Density**)
Similarly, for any $x_1, ..., x_n$,

$$f_{x_1,...,x_n}(x_1, ..., x_n) = f_{x_1}(x_1) \times f_{x_2|x_1}(x_2|x_1) \times f_{x_3|x_1,x_2}(x_3|x_1, x_2) \times \ldots \times f_{x_n|x_1,...,x_{n-1}}(x_n|x_1, ..., x_{n-1})$$

There are infinitely many forms of expressing $f_{x_1,...,x_n}(x_1, ..., x_n)$, for example:

$$f_{x_1,...,x_n}(x_1, ..., x_n) = f_{x_1,...,x_{n-1}|x_n}(x_1, ..., x_{n-1}|x_n) \times f_{x_n}(x_n)$$

5 Mixture

For d random variables $X_1, ..., X_d$ with multivariate distribution F. S is a non-empty subset of $1, ..., d$, which represents the conditioning set of variables. T is a subset of the complement of S, denoted as S^c, with at least two elements, which will serve as the set of conditioned variables. Denoting $M = S \cup T$, we express:

$$F_M(x_M) = \int_{(-\infty, x_S]} F_{T|S}(x_T|y_S) dF_S(y_S);$$

the conditional distribution $F_{T|S}(\cdot|x^S)$ exists almost everywhere on a set $X \subset \mathbb{R}^{|S|}$ with $P(X_S \in X) = 1$ [2].

proof.

$$F_M(x_M) = F_{T|S} = \int_{-\infty}^{x_s} \frac{\frac{\partial^{|S|} F_{T\cup S}}{\prod_{k\in S} \partial y_k}}{f_s(y_s)} dF_s(y_s)[5] = \int_{-\infty}^{x_s} \frac{\frac{\partial^{|S|} F_{T\cup S}}{\prod_{k\in S} \partial y_k}}{f_s(y_s)} f_s(y_s) dy_s = \int_{-\infty}^{x_s} \frac{\partial^{|S|} F_{T\cup S}}{\prod_{k\in S} \partial y_k} dy_s$$

$$= \int_{-\infty}^{x_s} \frac{\partial^{|S|} F_{T\cup S}}{\prod_{k\in S} \partial y_k} d(\prod_{k\in S} \partial y_k) = F_{T\cup S}[3] = F_M$$

□

6 Simplifying assumption

The theorem 4.7 from Czado(2019)[1] illustrates the fundamental principle of vine copulas, where the joint density function of a multivariate distribution can be factorized into the product of marginal densities and copulas. The copulas capture the dependencies between pairs of variables, conditioned on subsets of other variables.
The theorem states:

Theorem 2. *Every joint density $f_{1,...,d}$ can be decomposed as*

$$f_{1,...,d}(x_1, ..., x_d) = \prod_{j=1}^{d-1} \left(\prod_{i=1}^{d-j} c_{i,(i+j);(i+1),...,(i+j-1)} \right) \times \prod_{k=1}^{d} f_k(x_k) \tag{1}$$

Making a simplifying assumption eases the modeling with vine copulas, which assumes that the copulas for the conditional distributions do not depend on the specific values of the conditioning variables (i+1),...,(i+j-1). The following is a more formalized definition.

Definition 3. (**Simplifying assumption**) Let F be a multivariate Gaussian distribution of Z, the copulas for the conditional distributions do not depend on the values v_S of the conditioning variables, i.e.

$$C_{T;S}(\cdot) = C_{T;S}(\cdot; v_S)$$

depends only on the variables in the set $M = S \cup T$ (or the correlation matrix for Z_M).

A discrete example:
In credit risk assessment, the assumption is made that the copulas, representing the interdependence between a borrower's credit score (S) and income level (T) given various conditions, remain consistent irrespective of the specific values of conditioning variables, like loan amounts or terms.

7 Application in Paper[3]

The paper introduces the modeling of multivariate dependencies using PCCs, particularly focusing on trivariate cases. The simplifying assumption states that the dependency captured by the copula is invariant to the values of the conditioning variables. The paper suggests that this assumption might not hold in reality, which could lead to misleading inferences about the dependence structure among variables in real-life cases. Therefore, the paper introduces a nonparametric smoothing methodology to relax this assumption. It describes the methodology, demonstrates its performance through simulations, and applies it to real data. This discussion will focus on the implications for trivariate PCCs when the conditional independence assumption does not hold.

7.1 Trivariate PCCs

First, it is worth mention the **Trivariate PCCs** introduced in the paper.
Let X_1, X_2, X_3 be random variables with joint distribution function F and continuous margins F_1, F_2, F_3, respectively. Sklar's Representation Theorem [1] states that, for all $x_1, x_2, x_3 \in \mathbb{R}$,

$$F(x_1, x_2, x_3) = C\{F_1(x_1), F_2(x_2), F_3(x_3)\},$$

where C is a copula [2.1], i.e., a distribution function with margins that are uniform on $(0, 1)$. If F is absolutely continuous, its density can be written in terms of the density c of C as

$$f(x_1, x_2, x_3) = c\{F_1(x_1), F_2(x_2), F_3(x_3)\} \cdot f_k(x_k),$$

where, for each $k \in \{1, 2, 3\}$, f_k is the density of F_k. For more variables, it can be expressed as density[1].

A PCC is based on the fact that f can be decomposed as

$$f(x_1, x_2, x_3) = f_3(x_3) \cdot f_{2|3}(x_2|x_3) \cdot f_{1|23}(x_1|x_2, x_3). \tag{1}$$

Note that this factorization is unique up to relabeling. For any index set $A \subset \{1, 2, 3\}$ and $k \in A$, let $A - k = A \backslash \{k\}$. Using Sklar's Representation Theorem, one can then write, for arbitrary $j \notin A$,

$$f_{j|A} = c_{jk|A-k}(F_{j|A-k}, F_{k|A-k}) \cdot f_{j|A-k}. \tag{2}$$

proof of (2).

$$\frac{\partial}{\partial j \partial k} C_{jk|A-k}(F_{j|A-k}, F_{k|A-k}) = \frac{\partial}{\partial j \partial k} F_{jk|A-k}[1]$$

$$c_{jk|A-k}(F_{j|A-k}, F_{k|A-k}) \cdot f_{j|A-k} \cdot f_{k|A-k} = f_{jk|A-k}$$

$$c_{jk|A-k}(F_{j|A-k}, F_{k|A-k}) \cdot f_{j|A-k} = \frac{f_{jk|A-k}}{f_{k|A-k}} = \frac{\frac{f_{jk,(A-k)}}{f_{A-k}}}{\frac{f_{k,(A-k)}}{f_{A-k}}} = \frac{\frac{f_{j,A}}{f_{A-k}}}{\frac{f_A}{f_{A-k}}} = \frac{f_{j,A}}{f_A} = f_{j|A}$$

□

Repeated applications of relation (2) in (1) make it possible to express f as

$$f(x_1, x_2, x_3) = f_1(x_1)f_2(x_2)f_3(x_3) \cdot c_{12}\{F_1(x_1), F_2(x_2)\} \cdot c_{23}\{F_2(x_2), F_3(x_3)\}$$
$$\cdot c_{13|2}\{F_{1|2}(x_1|x_2), F_{3|2}(x_3|x_2); x_2\}, \tag{3}$$

which reduces to

$$c(u_1, u_2, u_3) = c_{12}(u_1, u_2)c_{23}(u_2, u_3)c_{13|2}(u_1|2, u_3|2; u_2)$$

if the margins of F are uniform.

The univariate conditional distributions featuring in (3) are given by $F_{j|k}(x_j|x_k) = h_{jk}\{F_j(x_j), F_k(x_k)\}$, where, for all $u, v \in (0, 1)$,

$$h_{jk}(u, v) = \frac{\partial C_{jk}(u, v)}{\partial v}.$$

7.2 Investigating the Simplifying Assumption

Definition 4. (**Kendall's tau (τ)**)
The Kendall τ coefficient, denoted as τ, is defined as the difference between the proportion of concordant[7] pairs and the proportion of discordant[7] pairs among all possible pairs of observations. Mathematically, it is expressed as:

$$\tau = \frac{\text{number of concordant pairs} - \text{number of discordant pairs}}{\text{total number of pairs}}$$

Assume for simplicity that $(X_1, X_2, X_3) = (U_1, U_2, U_3)$ is a random vector with standard uniform margins. Further suppose that

1. C_{12} is a Clayton copula with parameter $\theta_{12} = 1.2$;
2. C_{23} is a Gumbel–Hougaard copula with parameter $\theta_{23} = 3$;
3. given $U_2 = u_2$, $C_{13|2}$ is a Frank copula with parameter $\theta_{13|2}(u_2) = \gamma(4u_2 - 2)^3$,

where $\gamma \in \{0, 1\}$.
When $\gamma = 0$, the Frank Copula with parameter 0 is an independence copula. The variables U_1 and U_3 are conditionally independent given U_2, and hence the simplifying assumption is satisfied. When $\gamma = 1$, however, the conditional copula $C_{13|2}$ depends on the value of U_2, and the resulting model is not a simplified PCC.

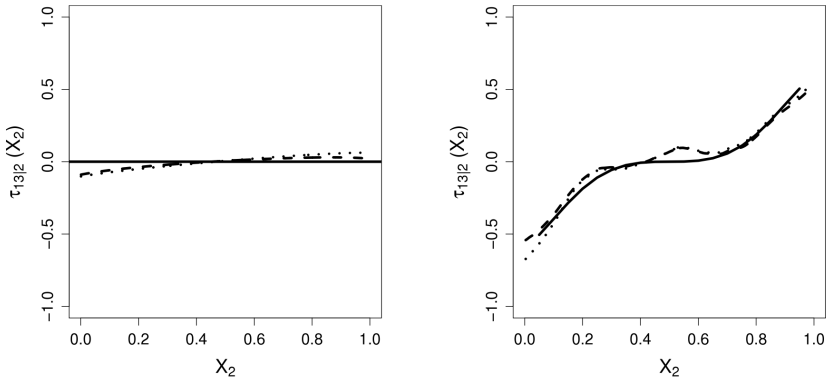

Figure 2: Plots of $\tau(X_1, X_3|X_2 = x_2)$[4] as a function of x_2 assuming a Frank copula for $C_{13|2}$, as derived from $\hat{\theta}_{13|2}$ (dashed) and $\tilde{\theta}_{13|2}$ (dotted) for the data when $\gamma = 0$ (left) and $\gamma = 1$ (right). In both graphs, the true function is shown as a solid curve.

The figure above reveals several key findings. Firstly, regardless of whether the margins are known or estimated, the estimates are similar to the true curve, indicating robustness in the estimation process. Most importantly, what the paper aims to show is in the left panel where the parameter $\gamma = 0$, the curves demonstrate a flat trend near $\tau = 0$, consistent with the underlying assumption, suggesting that $C_{13|2}$ is not functionally dependent on the conditioning variable X_2. Conversely, the right panel, representing $\gamma = 1$, exhibits a nonlinear pattern, implying potential dependency between X_2 and $C_{13|2}$, challenging the simplifying assumption.

This technique provides a methodology to examine the data for potential violations of the simplifying assumption. The plots illustrate the differences between scenarios with and without a violation of this assumption. These results indicate that the assumption should not be made blindly. Therefore, the proposed technique in the following section of the paper is designed for validation before applying the assumption.

8 Appendix

Definition 5. (**Conditional Distribution of a multivariate**)
Let $(X_1, \ldots, X_d) \sim F$, where $F \in \mathcal{F}(F_1, \ldots, F_d)$. If $X_1, \ldots, X_d$ are all discrete, then conditional distributions of the form $P(X_j \leq x_j, j \in S_1 \,|\, X_k = x_k, k \in S_2)$ are defined from conditional probability applied to events. If $X_1, \ldots, X_d$ are all continuous random variables, and $F_1, \ldots, F_d$ are absolutely continuous with respective densities $f_1, \ldots, f_d$, then the **Conditional CDFs** are defined via limits [2].
If $S_2 = \{k\}$, then

$$F_{S_1 \,|\, k}(x_{S_1} \,|\, x_k) := \lim_{\varepsilon \to 0^+} \frac{P(X_j \leq x_j, j \in S_1, x_k \leq X_k < x_k + \varepsilon)}{P(x_k \leq X_k < x_k + \varepsilon)} = \frac{\frac{\partial F_{S_1 \cup \{k\}}}{\partial x_k}}{f_k(x_k)}$$

proof.

$$\lim_{\varepsilon \to 0^+} \frac{P(X_j \leq x_j, j \in S_1, x_k \leq X_k < x_k + \varepsilon)}{P(x_k \leq X_k < x_k + \varepsilon)} = \frac{\frac{\partial F_{S_1 \cup \{k\}}}{\partial x_k}}{\frac{\partial F_{\{k\}}}{\partial x_k}} = \frac{\frac{\partial F_{S_1 \cup \{k\}}}{\partial x_k}}{f_k(x_k)}$$

□

If the cardinality of S2 is greater than or equal to 2, then the definition of the conditional CDF is:

$$F_{S_1 \,|\, S_2}(x_{S_1} \,|\, x_{S_2}) := \lim_{\varepsilon \to 0} \frac{P(X_j \leq x_j,\ j \in S_1; x_k \leq X_k < x_k + \epsilon,\ k \in S_2)}{P(x_k \leq X_k < x_k + \varepsilon, k \in S_2)} = \frac{\frac{\partial^{|S_2|} F_{S_1 \cup S_2}}{\prod_{k \in S} \partial x_k}}{f_{S_2}(x_{S_2})},$$

provided F_{S_2} is absolutely continuous.

Definition 6. (**Conditional PDF**)

$$f_{S_1 \,|\, k}(x_{S_1} \,|\, x_k) = \frac{\partial F_{S_1 \,|\, k}(x_{S_1} \,|\, x_k)}{\partial x_{S_1}}$$

Given the joint CDF $F_{S_1 \cup S_2}$ and marginal PDF f_{S_2}, the **conditional PDF** $f_{S_1|S_2}(x_{S_1}|x_{S_2})$ is defined as follows:

$$f_{S_1|S_2}(x_{S_1}|x_{S_2}) = \frac{\frac{\partial^{|S_2|} F_{S_1 \cup S_2}}{\prod_{k \in S} \partial x_k}}{f_{S_2}(x_{S_2})}$$

This formula holds true only if the cardinality of S_2 is greater than or equal to 2, and F_{S_2} is absolutely continuous.

Definition 7. (**Concordant** and **Discordant**)
Let $(x_1, y_1), \ldots, (x_n, y_n)$ be a set of observations of the joint random variables X and Y, such that all the values of x_i and y_i are unique (ties are neglected for simplicity). Any pair of observations (x_i, y_i) and (x_j, y_j), where $i < j$, are said to be **Concordant** if the sort order of (x_i, x_j) and (y_i, y_j) agrees: that is, if either both $x_i > x_j$ and $y_i > y_j$ holds or both $x_i < x_j$ and $y_i < y_j$; otherwise they are said to be **Discordant**.

Theorem 3. ***(Fundamental Theorem of Calculus, Part I)***
Let f be continuous on the closed interval $[a, b]$, and let F be the function defined, for all x in $[a, b]$, by

$$F(x) = \int_a^x f(t)\, dt.$$

Then F is continuous on $[a, b]$ and differentiable on (a, b), and $F'(x) = f(x)$ for all x in (a, b).

9 Acknowledgments

I would like to thank my mentor, Yanees Dobberstein, for his invaluable guidance throughout the semester. Our weekly discussions, along with his insightful input, have significantly enhanced my understanding and the quality of this work.

References

1. Czado C. Analyzing dependent data with vine copulas. Lecture Notes in Statistics, Springer 2019;222.
2. Joe H. Dependence modeling with copulas. CRC press, 2014.
3. Acar EF, Genest C, and Nešlehová J. Beyond simplified pair-copula constructions. Journal of Multivariate Analysis 2012;110:74–90.