School of Informatics

Informatics Project Proposal
Copula and Marginal Generative Flows

[redacted]

Abstract

The proposed research project explores the use of Flow-based generative models for bivariate and multivariate copula estimation. Bivariate copula estimation is achieved using Copula and Marginal Generative Flows (CM Flows), as proposed by Wiese et al. [1]. For multivariate estimation, CM flows are combined with R-vines estimation techniques, as in Dissmann et al. [2], yielding a new estimation technique for multivariate copulas. Estimations are evaluated on simulated data from known copula families.

Date: Monday 27th April, 2020

[redacted]

1 Introduction

Modeling and estimation of multivariate distributions is an important task in numerous fields, such as finance, medicine and physics [3, 4, 5]. One approach to multivariate modeling is to choose an existing distributional family of multivariate distributions, such as the Gaussian distribution, and fit it to the data. This can be rather restrictive, since both the marginals and the dependence structure are determined by the chosen distributional family. Copulas offer a remedy to this problem: they are a mathematical construct to fully describe the dependence between random variables. By Sklar's Theorem [6], any joint distribution can be separated into a copula distribution and a marginal distribution. Conversely, copula and marginal can be used to model any multivariate distribution, allowing for marginals and dependence structure to be modeled separately.

While the marginal and joint distributions are directly observable, the copula remains the hidden dependence structure that connects joint and marginal distribution [7]. Both parametric and semi-parametric estimation techniques have been proposed [8, 9, 10]. One downside of these methods is again their restrictiveness. To estimate copulas, they take assumptions about the underlying distributional family, which can lead to false certainty in predictions.

One consequential example of this occurred in what led up to the financial crisis of 2008: a copula function approach by Li [11] had become the standard tool to model the joint distribution of the time until default of financial products given only measurements of the marginal distributions derived from market information, such as risky bond prices or asset swap spreads. This allowed complex risks to be modeled from available data. The formula made, however, a critical assumption: that the dependence between the random variables follows a gaussian distribution. This is especially problematic in the tails, where the gaussian copula predicts a low dependence. This low dependence led to an under-estimation of the time until default in the case of unlikely events, and thus a false sense of security when joining financial products such as bonds.

One possible solution to this is the use of non-parametric methods. These methods do not place an assumption on the form of the copula or the marginals, and thus offer greater generality. Two common methods of non-parametric estimation are kernel-based methods and neural networks. For copula estimation, kernel-based methods have been well explored [7, 12], while literature on copula estimation using neural networks remains sparse.

Recently, Wiese et al. [1] proposed Copula and Marginal Generative Flows (CM Flows), a method that uses Flow-based generative models for the estimation of copula and marginal from a bivariate joint distribution. Generative models try to generate examples that come from the same distribution as the distribution of the data they are trained on. So, given samples from a joint distribution p(X, Y) or marginal distributions p(X) and p(Y), they generate examples that follow these distributions. This makes them suitable for tasks, where the underlying distribution is unknown and of interest. There are different kinds of generative models, such as Generative Adversarial Networks (GANs), Variational autoencoders (VAEs) and Flow-based Generative Models (also called generative flows). Unlike GANs and VAEs, generative flows explicitly model the input distribution. They are more straightforward to train, since the loss function is simply the negative log-likelihood [13].

In the proposed research project, I will explore the use of CM flows for estimating copulas. I will first create a Pytorch implementation of CM flows, which I will evaluate using known copula families. In addition to the evaluation performed by Wiese et al. [1], I will evaluate the predictions of the marginals explicitly, and evaluate the model without assuming tail beliefs.

This will give further insight on the behaviour of CM flows in the tails, where only few data points are sampled.

Then, I will extend CM flows on multivariate joint distributions: In their paper, Wiese et al. [1] only introduce bivariate CM flows, and hint to the use of vine copulas for multivariate CM flows. R-vine copulas are a graphical tool to generate a multivariate copula from connected bivariate copulas. Finding a suitable R-vine structure is a difficult task, since for n dimensions, there exist $\frac{n!}{2} \times 2^{\binom{n-2}{2}}$ different R-vines [14]. This is why most approaches use a heuristic to find an R-vine structure. Dissmann et al. [2] use a sequential approach for this task: they first identifying the first tree, its pair-copula families and estimating their parameters, then the second, and so on. In the proposed research, the selection of R-vine structure is distinct from the selection of the copula family and estimation of copula parameters. Exploiting this, I will replace the selection of copula family and estimation of copula parameters by the CM flows implementation. This will result in a new estimator for multivariate copulas.

The research questions are:

1. Are CM flows a suitable method for bivariate copula estimation?
2. Can CM flows improve multivariate copula estimation using R-vines?

The research objectives are:

1. Implement and evaluate CM flows on bivariate distributions
2. Create a multivariate copula estimator using CM flows and R-vines

Beneficiaries The proposed research will benefit the academic field on non-parametric copula estimation in three ways: first, it will further explore the suitability of generative flows for copula estimation. Second, it will provide a method of incorporating CM Flows in R-vine estimation. Third, I am planning to make my python implementation of CM Flows publicly available on Github, so applied researchers and practitioners can easily implement CM flows to estimate copulas.

Feasibility CM flows combine two generative flow models: Real-NVPs and DDSFs. Both of these have publicly available Pytorch implementations. These can be combined to create CM flows, without the need to write the models from scratch. The implementation of CM flows for R-vines poses a more difficult problem, since it could take considerable time to incorporate CM flows into the R-vines estimation technique.

This research proposal is structured as follows: section 2 gives a brief overview over the relevant literature. Section 3 provides the mathematical background needed. Section 4 describes the methodology, evaluation techniques, and expected outcomes. Section 5 outlines the research plan, milestones and deliverables.

2 Literature Review

Copulas can be estimated via maximum-likelihood estimation by placing assumptions on both the form of the copula and the marginal. This parametric estimation has been proposed by

Oakes [8], for example. Semi-parametric estimation methods place assumptions only on the form of the copula, as seen in Genest et al. [9] and Chen and Fan [10]. Kernel-based nonparametric estimation methods have been proposed by Chen and Huang [7] and Scaillet and Fermanian [12]. Elidan [15] combine Bayesian Networks with copulas for multi-dimensional modeling.

Flow-based generative models use normalizing flows for density estimation. Notable normalizing flows include Real-valued Non-Volume Preserving (RealNVP) transformations [16], Non-linear Independent Components Estimation (NICE) [16] and Deep Dense Sigmoidal Flows (DDSFs) [17]. Wiese et al. [1] use a DDSF for estimation of the marginal distribution, and a RealNVP for copula estimation.

Since constructing high-dimensional copulas is difficult, several methods attempt to combine bivariate copulas to achieve higher dimensions. There exist hierarchical tree compositions [18], mixture of trees [19] and recursive construction of conditional bivariate copulas [20, 21, 22]. A sequential bayesian method of R-vines estimation have been suggested by Gruber et al. [23]. An approach using Bayesian Networks has been proposed by [15].

Dissmann et al. [2] use a sequential approach for finding an appropriate R-vine tree structure: they first identify the first tree, its pair-copula families and then estimate their parameters. Then, they estimate the second tree and so on. For tree selection they use a maximum spanning algorithm. Pair-copulas are chosen using the Akaike information criterion. Parameter estimation follows Aas et al. [20].

3 Background

This section provides the definitions for the most important concept used in the proposed research. Namely, they are the definition of a copula and Sklar's theorem, which forms the basis for connection the joint and marginal distributions. For normalizing flows, the change of variable formula is needed to calculate the loss function. For the extension on R-vines, the definition of R-vines is given.

A **copula** is the cumulative distribution function (CDF) of a random vector defined on the hypercube $[0,1]^d$ with uniform marginals over [0,1]: $C(u_1, ..., u_d) = P(U_1 \leq u_1, ..., U_d \leq u_d)$ for d dimensions, where $U_i \sim U_{[0,1]}$ and $C(u1, ..., ud) : [0,1]^d \rightarrow [0,1]$ [24]. The copula probability density function (PDF) is defined as $c(u_1, ..., u_d) = \partial^d C(u_1, ..., u_d)/\partial u_1 ... \partial u_d$.

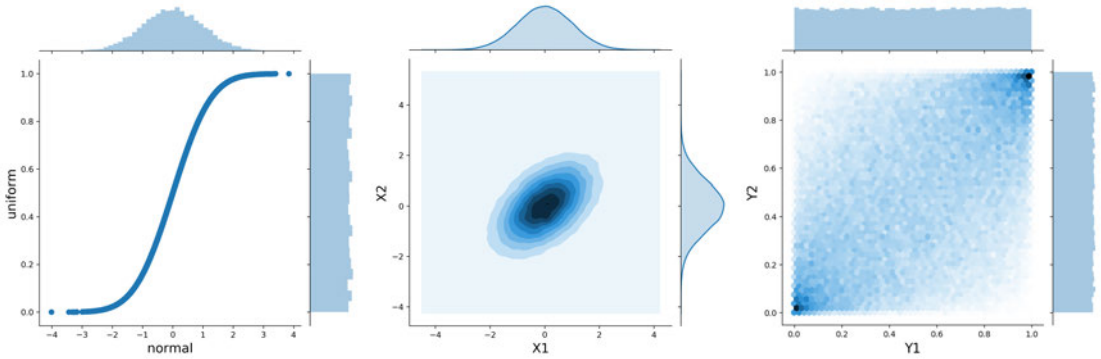
Figure 1 illustrates the copula of a multivariate normal joint distribution. The probability integral transform can be used to transform the marginals of a multivariate joint distribution to become uniform: for any distribution X with CDF F_X, $F_X(X)$ has uniform distribution. Sklar's theorem [24] formalized how copulas connect the marginal and joint distributions:

Sklar's Theorem: Let $X = (X_1, ..., X_d)$ be a d-dimensional random vector with CDF F_X that has the marginals $F_1, ..., F_d$. Then there exists a copula C such that $\forall x \in Rd : F_X(x_1, ..., x_d) = C(F_1(x_1), ..., F_d(x_d)), x_i \in R$. When the marginals are continuous, the copula is unique. Note, that $U_i = F_i(X_i)$ is uniformly distributed, and $x_i = F_i^{-1}(u_i)$ if $u_i \sim U$ and F_i being the marginal of x_i. For $u_i \in [0,1]$:

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

Following Safaai et al. [25], copula density can then be written as

$$c(u_1, ..., u_d) = \frac{f_X(F_1^{-1}(u_1), ..., F_d^{-1}(u_d))}{\prod_{i=1}^{d} f_i(F_i^{-1}(u_i))} \tag{2}$$



(a) Probability Integral Transform (b) Joint Distribution (c) Copula

Figure 1: Copula Illustration for a multivariate normal joint distribution

and thus, the joint PDF f_X can be separated into the copula c and the product of marginal densities. This allows for an interpretation of the copula as the part of the joint density function that is independent from the marginals and captures the dependencies between the variables [25].

Normalizing Flows Both DDSFs and Real NVPs are normalizing flows. They are used to transform one random variable into another and are denoted as $f_\Theta : X \to Y$. They consist of invertible transformations with a tractable Jacobian [17]. Being invertible, the change of variables formula can be applied on the densities $p_Y(y)$ and $p_X(x)$:

$$p_Y(y) = \left| \frac{\partial f(x)}{\partial x} \right|^{-1} p_X(x) \tag{3}$$

The determinant of f's Jacobian is needed on the RHS to adjust the expanding or contracting of regions of X by f [17].

R-Vines A regular (R-) vine tree sequence is a set of trees $\mathcal{V} = (T_1, ..., T_{d-1})$ if:

1. Each tree $T_j = (N_j, E_j)$ is connected.
2. T_1 is a tree with node set $N_1 = 1, ..., d$ and edge set E_1.
3. For $j \geq 2$, T_j is a tree with node set $N_j = E_{j-1}$ and edge set E_j.
4. For $j = 2, ..., d-1$ and $a, b \in E_j$ it must hold that $|a \cap b| = 1$ [26].

Thus, a regular vine has a nested structure: the edges in the first tree and the nodes in the second tree, the edges in the second tree are the nodes in the first tree, and so on.

The regular vine distribution consists of the R-vine tree sequence $\mathcal{V}$ on d elements, the marginal distributions $F = (F_1, ..., F_d)$ of the random variable $X_i, i = 1, ..., d$, the bivariate copulas $B = C_l | e \in E_i, i = 1, ..., d-1$ and the relationship between the R-vine tree sequence $\mathcal{V}$ and the set B of bivariate copulas [26].

4 Methodology

CM flows The model follows Wiese et al. [1]: they propose CM flows composed of a copula flow $h_\eta : [0,1]^2 \to [0,1]^2$ and a marginal flow $m_\theta : [0,1]^2 \to \mathbb{R}$. The marginal flow approximates the inverse marginal CDFs $F_{X_1}^{-1}$ and $F_{X_2}^{-1}$, the copula flow approximates the generating function of the copula $(C_1, C_2) = (F_{X_1}(X_1), F_{X_2}(X_2))$. The CM flow is given by $g_{\theta,\eta}(u) = m_\theta \circ h_\eta(u)$ for $u \in [0,1]^2$ [1].

The univariate marginal flow is defined as a DDSF $f : \mathbb{R} \times \Theta \to \mathbb{R}$, the bivariate marginal flow is simply defined as

$$m_\theta(u) = \left[m_{\theta_1}^{(1)}(u_1), m_{\theta_2}^{(2)}(u_2)\right]^T \tag{4}$$

with $u \in [0,1]^d$ [1].

The bivariate copula flow is constructed using a Real NVP $\tilde{h} : \mathbb{R} \times H \to \mathbb{R}^2$ and the function $\Psi : \mathbb{R} \to [0,1]$ [1]:

$$\begin{aligned} h : [0,1]^2 \times H &\to [0,1]^2 \\ (u, \eta) &\mapsto \Psi \circ \tilde{h}_\eta \circ \Psi^{-1}(u) \end{aligned} \tag{5}$$

The function Ψ projects the input from the unit square onto the real numbers, and Ψ^{-1} projects the output of the RealNVP back onto the unit square. Wiese et al. [1] use the sigmoid function as Ψ. Alternatively, one could use the inverse CDF of the normal distribution.

Tail Belief Wiese et al. [1] use tail beliefs to model exact tails of the marginals. The univariate marginal flow $m(u, \theta)$ follows the tail belief in the tails and becomes a scaled version of the DDSF f outside the tails. The parameter θ can be optimized by minimizing the NLL only outside of the tails. In the proposed research, I will first implement the model without tail beliefs, to see how CM flows behave in the tails. Evaluating the model with tail beliefs is an option.

Evaluation CM Flows While the CM flows are trained only on samples from the joint distribution, I evaluate them using known copula densities. The copula flow's objective it to achieve a uniform distribution in the marginals of the copula, $U \sim U([0,1]^2)$, and minimize the difference between (C_1, C_2) and $(\tilde{C_1}, \tilde{C_2})) = (h_{\eta,1}(U), h_{\eta,2}(U))$. The latter is computed using the Jensen-Shannon divergence (JSD). The uniformity of the marginals of the copula is assessed with two metrics:

$$T(i,n) = \frac{1}{n} \sum_{k=1,\dots,n} \left| \log \mathbb{P}(\tilde{C_i} \in A_k + \log n \right| \tag{6}$$

which is being approximated using Monte-Carlo, and

$$M(i,n) = \max_{k=1,\dots,n} \left| \log \mathbb{P}(\tilde{C_i} \in A_k + \log n \right| \tag{7}$$

with $A_k = [(k-1)/n, k/n], k = 1, ..., n$.

Additionally, I will evaluate the estimation of the marginals, using JSD for the predicted marginals and the true marginal.

Extension on R-vines Dissmann et al. [2] deploy a sequential, heuristic algorithms to choose R-vines. Their tasks can be split into:

1. selecting an R-vine structure
2. choose a bivariate copula family for each variable pair
3. estimate parameters for each copula

For step 1, they use a maximum spanning algorithm to select the tree that maximizes the sum of absolute empirical Kendall's taus. To estimate the bivariate copulas, they select a copula from a known copula family and estimate its parameters. In my model, Steps 2 and 3 are replaced by CM flows, for non-parametric copula estimation.

Evaluation R-vines Dissmann et al. [2] use a simulation study to evaluate their algorithm. In the proposed research, I will evaluate the performance of the R-vine CM flows using simulated data of multivariate distributions using known copula families. Dissmann et al. [2] compare the predicted model to the true model using the general tau-difference and the lower and upper tau-difference. The general tau-difference is the mean absolute difference between pairwise empirical Kendall's taus of the simulated data from the true and selected models. The lower and upper tau-difference is the mean absolute difference between pairwise empirical lower and upper exceedance Kendall's taus. Another evaluation method would be to measure KL-divergence of the predicted copula and marginal to the true copula and marginal.

Data Evaluation on simulated data allows for an exact comparison between the true underlying copula and the predicted copula. Hence, both models are evaluated on simulated data. Wiese et al. [1] evaluate CM flows for the Clayton, Gumbel and Frank copula. Evaluation on other copula families is possible. For the R-vine extension, higher dimensional data can be simulated using combinations of different copulas, from the same or other different families.

Expected Outcomes The implementation of CM Flows is expected to yield similar results to Wiese et al. [1]. I expect the evaluation of the marginals to yield positive results. Since I am not assuming tails, I expect the prediction in the tails to be less accurate, since only few data points are available for training. The outcomes of the R-vine extension are less predictable. In the desired case, using CM Flows for R-vine estimation could result in a non-parametric model that is able to estimate R-vines well. But problems could arise: since more than one CM Flow has to be trained per R-vine, training time for the whole model could become high.

Overall, I expect the proposed research to yield further insights into CM flows, especially when not assuming tail beliefs. In addition, the proposed research could result in a new non-parametric estimation method for R-vines.

Limitations The proposed research is limited to evaluation on simulated data. A model that performs well on simulated data might not do so on empirical data, due to errors in data collection or biased sampling. However, since the model is non-parametric, it can be assumed to generalize better to empirical data than parametric models. Other limitations of the proposed research are given by personal time constraints: the baseline CM flows model will not be changed, for example by trying different types of generative flows. For the R-vine extension, there are many different types of R-vine structure selection techniques, some of which might perform better with CM flows.

Risk Assessment There are a few risks involved in the project: first, the CM flows implementation could take more time than expected. Since it is the first step, this would delay the other works. Second, the CM flows might not work as well as expected on new examples. Especially the analysis of the marginals could prove worse results than expected. If CM flows are not promising, there is no point in implementing them for R-vine estimation. One way out would be to further analyze why the CM flows don't perform well.

Ethics Since I am only using computer-generated data, there are no apparent ethical risk in regards to privacy or data misuse. The application of copula estimation is very broad, it is difficult to predict how advances in this field will be used later.

5 Research Plan, Milestones and Deliverables

The research project is split into two stages: the implementation of CM Flows, and the extension to R-vines. Figure 2 illustrates the proposed timeline: after completing most of the background reading in week 1, I will focus on implementing and evaluating the CM flows, allowing for two week for implementation, and one week for implementation. Then, I will focus on the R-vine estimation, allowing for a few days of further background reading, and then spending three weeks on R-vine implementation. One week is allocated for R-vine evaluation. The last month of the dissertation period will be spend on write-up. Table 1 shows the milestones of the project: they are the implementation of the CM flows, the implementation of R-vine estimation, and the write-up of the dissertation. The deliverables are the drafts for the evaluation of the CM flows and the R-vine extension, as well as the dissertation, as shown in 2. Deliverable 1 depends on the completion of milestone 1, and deliverable 2 depends on the completion of milestone 2.

The tools used for this project will be python, git and latex. Python, and specifically the Pytorch framework will be used for the implementations. Git will be used as version control. Latex will be used to continuously record my progress and for the delivery of the evaluation drafts and the final dissertation.

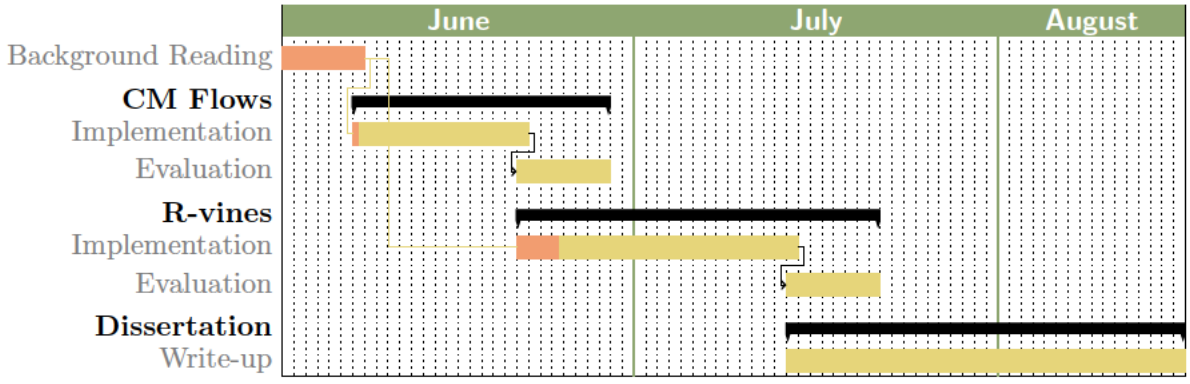


Figure 2: Gantt Chart of the activities defined for this project.

Milestone	Week	Description
M_1	3	Implement CM Flows
M_2	8	Implement R-vine extension
M_3	11	Dissertation Write-up

Table 1: Milestones defined in this project.

Deliverable	Week	Description
D_1	4	Draft Evaluation of CM Flows
D_2	9	Draft Evaluation R-vine extension
D_3	11	Dissertation

Table 2: List of deliverables defined in this project.

References

[1] Magnus Wiese, Robert Knobloch, and Ralf Korn. Copula & marginal flows: Disentangling the marginal from its joint. *arXiv preprint arXiv:1907.03361*, 2019.

[2] Jeffrey Dissmann, Eike C Brechmann, Claudia Czado, and Dorota Kurowicka. Selecting and estimating regular vine copulae and application to financial returns. *Computational Statistics & Data Analysis*, 59:52–69, 2013.

[3] Umberto Cherubini, Elisa Luciano, and Walter Vecchiato. *Copula methods in finance*. John Wiley & Sons, 2004.

[4] Masanori Sato, Kiyotomo Ichiki, and Tsutomu T Takeuchi. Precise estimation of cosmological parameters using a more accurate likelihood function. *Physical Review Letters*, 105(25):251301, 2010.

[5] D Beaudoin and L Lakhal-Chaieb. Archimedean copula model selection under dependent truncation. *Statistics in medicine*, 27(22):4440–4454, 2008.

[6] Fabrizio Durante, Juan Fernandez-Sanchez, and Carlo Sempi. A topological proof of sklar's theorem. *Applied Mathematics Letters*, 26(9):945–948, 2013.

[7] Song Xi Chen and Tzee-Ming Huang. Nonparametric estimation of copula functions for dependence modelling. *Canadian Journal of Statistics*, 35(2):265–282, 2007.

[8] David Oakes. A model for association in bivariate survival data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 44(3):414–422, 1982.

[9] Christian Genest, Kilani Ghoudi, and L-P Rivest. A semiparametric estimation procedure of dependence parameters in multivariate families of distributions. *Biometrika*, 82(3):543–552, 1995.

[10] Xiaohong Chen and Yanqin Fan. Estimation and model selection of semiparametric copula-based multivariate dynamic models under copula misspecification. *Journal of econometrics*, 135(1-2): 125–154, 2006.

[11] David X Li. On default correlation: A copula function approach. *The Journal of Fixed Income*, 9 (4):43–54, 2000.

[12] Olivier Scaillet and Jean-David Fermanian. Nonparametric estimation of copulas for time series. *FAME Research paper*, (57), 2002.

[13] Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. In *Advances in Neural Information Processing Systems*, pages 10215–10224, 2018.

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

[15] Gal Elidan. Copula bayesian networks. In *Advances in neural information processing systems*, pages 559–567, 2010.

[16] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*, 2016.

[17] Chin-Wei Huang, David Krueger, Alexandre Lacoste, and Aaron Courville. Neural autoregressive flows. *arXiv preprint arXiv:1804.00779*, 2018.

[18] C Savu and M Trede. Hierarchical archimedean copulas: International conference on high frequency finance. *Konstanz, Germany*, page 27, 2006.

[19] Sergey Kirshner. Learning with tree-averaged densities and distributions. In *Advances in Neural Information Processing Systems*, pages 761–768, 2008.

[20] Kjersti Aas, Claudia Czado, Arnoldo Frigessi, and Henrik Bakken. Pair-copula constructions of multiple dependence. *Insurance: Mathematics and economics*, 44(2):182–198, 2009.

[21] Tim Bedford and Roger M Cooke. Vines: A new graphical model for dependent random variables. *Annals of Statistics*, pages 1031–1068, 2002.

[22] Dorota Kurowicka and Roger M Cooke. The vine copula method for representing high dimensional dependent distributions: application to continuous belief nets. In *Proceedings of the Winter Simulation Conference*, volume 1, pages 270–278. IEEE, 2002.

[23] Lutz Gruber, Claudia Czado, et al. Sequential bayesian model selection of regular vine copulas. *Bayesian Analysis*, 10(4):937–963, 2015.

[24] Roger B Nelsen. *An introduction to copulas*. Springer Science & Business Media, 2007.

[25] Houman Safaai, Arno Onken, Christopher D Harvey, and Stefano Panzeri. Information estimation using nonparametric copulas. *Physical Review E*, 98(5):053302, 2018.

[26] Claudia Czado. Analyzing dependent data with vine copulas. *Lecture Notes in Statistics, Springer*, 2019.