Matrix-based vine copula building algorithms

Diploma Abstract

Author:
Dániel Pfeifer

Supervisor:
Edith Alice Kovács, PhD

Vine copulas were introduced by Bedford T. and R. M. Cooke in 2002 [1]. They are used to model joint probability distributions, allowing for separate modeling of the one-dimensional marginals and connections between variables.

$$f(\mathbf{x}) = \prod_{j=1}^{n} \underbrace{f_j(x_j)}_{\substack{\text{one-dimensional}\\ \text{marginals}}} \prod_{i=1}^{n-1} \prod_{(e_1,e_2|T)\in E_i} \underbrace{c_{e_1,e_2|T}}_{\substack{\text{conditional pair-copulas}\\ \text{in the vine structure}}} \underbrace{(F_{e_1|T}(x_{e_1}|\mathbf{x}_T), F_{e_2|T}(x_{e_2}|\mathbf{x}_T))}_{\substack{\text{conditional}\\ \text{c.d.f.'s}}}$$

The pair- and conditional pair-copula densities (denoted by $c_{e_1,e_2|T}$) can be read from the vine structure (see left of Figure 1). Vines can also be modeled using a special junction tree sequence, called a cherry tree sequence (see right of Figure 1).

Vine copulas as useful because they can efficiently model differently correlated variables, and also symmetric and non-symmetric multivariate connections between variables.

To store vines in a digital environment, Dissmann et al used vine matrices [2]. However these did not uniquely correspond to a vine. In this study, we corrected this, and gave two vine matrix building algorithms, and proved that under certain conditions they are equivalent.

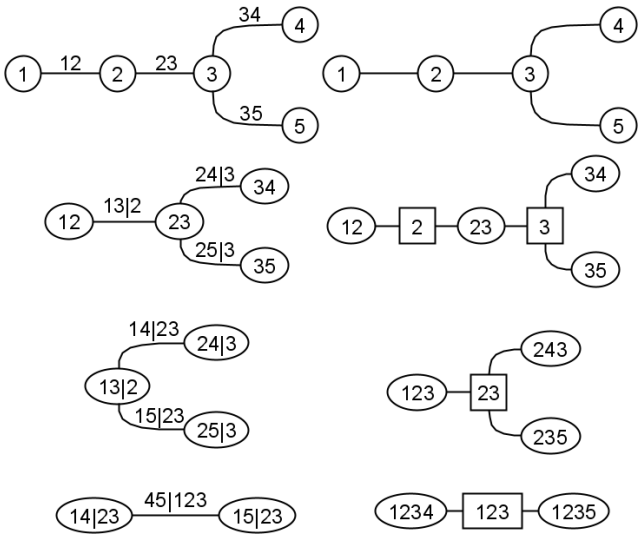

Figure 1: Example of a vine structure, regular and cherry tree representation

Vines can be truncated to obtain an approximation for the actual f joint p.d.f. This new, truncated approximation is efficient if most of the "information content" is captured in the first couple of trees. The most important new result of the study achieves this with a greedy vine building algorithm that takes a data frame as its input, and builds up a vine such that when truncating on any level, the resulting tree will be "locally optimal", and minimize the Kullback-Leibler divergence between f and its truncated vine approximation. (This is based on an important result by Edith Kovács and Tamás Szántai [3].)

The information content of the vine is defined by the so-called "copula entropy", introduced by [5] in 2011. They also proved that the negative of the copula entropy is a generalization of the Mutual Information in higher dimensions, and developed an R and Python package called `copent` that approximates this value for any given combination of columns in a data frame [4].

We have implemented the mentioned algorithms in Python, and ran it on two Kaggle datasets, a Red Wine Quality and a MAGIC Gamma Telescope data frame. It was tested on different data sizes, and the runtime was measured. The complexity of the algorithm is dependent on the the number of rows m with a runtime of $O(m^2)$, which makes it hard to apply it for large datasets, so only 100 and 1000 rows were tested. The greedy vine building algorithm successfully found a locally optimal vine for both datasets, from which the connections between variables can easily be read (see pages 34-35 of the study).

[1] T. Bedford and R.M. Cooke. "Vines—a new graphical model for dependent random variables". In: (2002).

[2] J. Dißmann et al. "Selecting and estimating regular vine copulae and application to financial returns". In: (2013).

[3] Edith Kovács and Tamás Szántai. "On the connection between cherry-tree copulas and truncated R-vine copulas". In: *Kybernetika* 53.3 (2017), pp. 437–460.

[4] Jian Ma. "copent: Estimating Copula Entropy and Transfer Entropy in R". In: *arXiv preprint arXiv:2005.14025* (2020).

[5] Jian Ma and Zengqi Sun. "Mutual information is copula entropy". In: *Tsinghua Science & Technology* 16.1 (2011), pp. 51–54.