VineCopulas: an open-source Python package for vine copula modelling

Judith N. Claassen [1], **Elco E. Koks** [1], **Marleen C. de Ruiter** [1], **Philip J. Ward** [1,2], **and Wiebke S. Jäger** [1]

1 Institute for Environmental Studies, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands **2** Deltares, Delft, The Netherlands

Summary

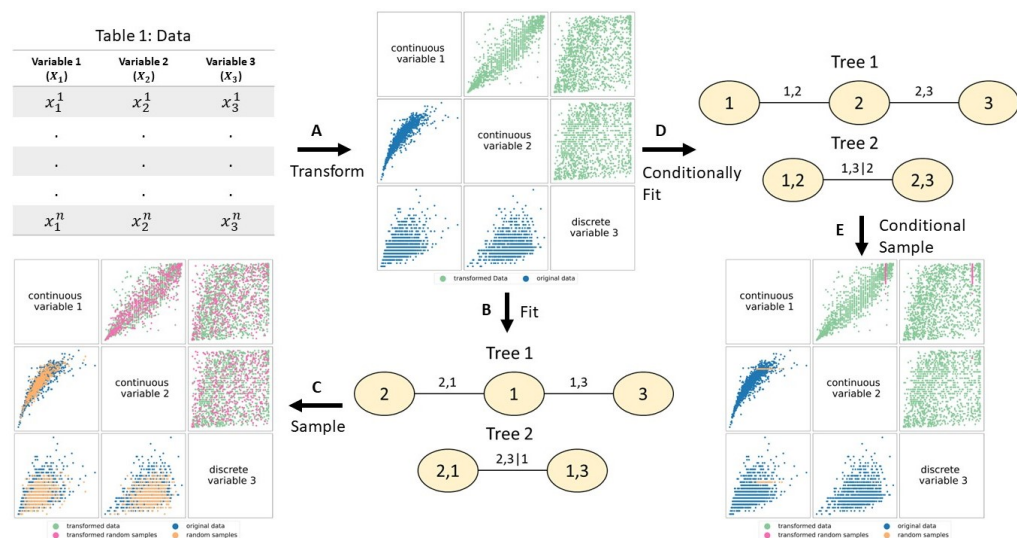A copula method can be used to describe the dependency structure between several random variables. Copula methods are used widely in various research fields across different disciplines, ranging from finance to the bio-geophysical sciences (Dißmann et al., 2013; Klein et al., 2020; Mitskopoulos et al., 2022). While some other multivariate distributions, for instance a multivariate normal distribution, allow for a highly symmetric dependency structure with the same univariate and multivariate marginal distributions, copulas can model the joint distribution of multiple random variables separately from their marginal distribution (Czado & Nagler, 2021; Sklar, 1959).

Once a copula distribution has been modelled, they allow for random samples of the data to be generated, as well as conditional samples. For example, if a copula has been fit between people's height and weight, this copula can create random correlated samples of both variables as well as conditional samples, e.g., samples of weight given a specific height.

Although copulas are an excellent tool to model dependencies in bivariate data, data with two variables, there are only a limited number of copulas capable of modelling larger multivariate datasets, for example, the Gaussian and Student-t copula. However, when modelling the dependencies between a large number of different variables, a more flexible multivariate modelling tool may be required that does not assume a single copula to capture all the individual dependencies. To this end, vine copulas have been proposed as a method to construct a multivariate model with the use of bivariate copulas as building blocks (Aas et al., 2009; Bedford & Cooke, 2001, 2002; Joe, 1997).

In the previous example related to height and weight, a vine copula could be used to also model age in relation to height and weight. Like bivariate copulas, vine copulas allow the user to generate random and conditional samples (Cooke et al., 2015). However, to draw conditional samples from a vine copula for a specific variable, the vine copula has to be structured in such a way that the order in which the samples are generated draws the variable of interest last, i.e. the sample is conditioned on the preceding samples of other variables. For example, if one wants to generate a conditional sample of height, the samples of age and weight have to be provided first. Additionally, while it is more common to use copulas for continuous data, such as weight and height, methods have been developed to also allow for discrete data, such as age, to be modelled (Mitskopoulos et al., 2022).

`VineCopulas` is a Python package that is able to fit and simulate both bivariate and vine copulas. This package allows for both discrete as well as continuous input data, and can draw conditional samples for any variables of interest with the use of different vine structures (see Figure 1).

Figure 1: A schematic representation of VineCopulas functionalities, where the lettering refers to the different arrows (Python functions). A) Samples from Table 1 - data, consisting of both continuous and discrete variables (plotted in blue) are transformed into pseudo-observations using their marginal distributions (shown in green). B) A vine copula is fit to the transformed data. Here, the first tree has nodes containing the variables and edges denoting the bivariate dependencies. The edges in the second tree denote the dependency between all variables. C) Using the fitted vine copula, random samples are generated. D) As not every vine copula structure is suitable to generate conditional samples of every variable, due to its inherent sampling order, a vine copula can also be fit conditionally. Here, a vine copula is fit conditionally for variable 1. E) The conditionally fit vine copula is used to draw conditional samples of variable 1 given specific values of variables 2 and 3.

Statement of need

The programming language R is widely known as the most advanced statistical programming language and hence has many well-developed packages for copulas, such as `copula` (Hofert et al., 2023), `VineCopula` (Nagler et al., 2023), and `CDVineCopulaConditional` (Bevacqua, 2017). However, with the open source programming language Python gaining more popularity for statistical programming, there is an increasing interest in Python-based copula packages. Therefore, we have developed the package `VineCopulas`, a pure Python implementation for (vine) copulas.

`VineCopulas` integrates many of the standard copula package features, including fitting, Probability Density Function (PDFs), and random sample generation for bivariate and vine copulas. This package can also fit the best marginal distributions of the individual variables based on the univariate distributions available in the statistical Python package `SciPy` (Virtanen et al., 2020). Furthermore, the `VineCopulas` can compute cumulative distribution functions (CDFs) of bivariate copulas. In addition, the package also enables the user to generate conditional samples, fit vine structures to facilitate specific conditional probabilities and fit as well as simulate discrete data, all of which are unique to have in a single package.

While there are two well-used python copula packages, `copulas` (DataCebo, n.d.), and `pyvinecopulib` (Nagler & Vatter, 2023), neither of these packages includes the above-mentioned unique features. Furthermore, `copulas` is mostly suitable for bivariate copulas, and has limited vine copula capabilities, while `pyvinecopulib` is a C++ library with a Python interface, meaning that it is not fully Python-based, and therefore less adaptable for a Python user. Therefore, `VineCopulas` is targeted towards data analysts, researchers and modellers in various fields, who are python users or require functionality specifically for discrete data and conditional sampling.

VineCopulas is currently being used in a study on multi-hazards to model the dependencies between different natural hazard intensities. For this study, the ability to generate conditional samples is required to evaluate possible magnitudes of one natural hazard given multiple others e.g., levels of extreme precipitation given specific extreme wind speeds and relative humidity. The capability to also simulate discrete data may be useful for hazards with intensity measures of a discrete nature, such as the Volcanic Explosivity Index (VEI). Applications of this type, are growing in the field of compound and multi-hazard risk research (Bevacqua et al., 2017; Eilander et al., 2023).VineCopulas will allow python users to continue this research at a higher dimensionality, showing the clear need for this package.

Acknowledgements

References

Aas, K., Czado, C., Frigessi, A., & Bakken, H. (2009). Pair-copula constructions of multiple dependence. *Insurance: Mathematics and Economics*, *44*, 182–198. https://doi.org/10.1016/j.insmatheco.2007.02.001

Bedford, T. J., & Cooke, R. (2001). *Monte carlo simulation of vine dependent random variables for applications in uncertainty analysis*.

Bedford, T. J., & Cooke, R. (2002). Vines - a new graphical model for dependent random variables. *Annals of Statistics*, *30*(4), 1031–1068. https://doi.org/10.1214/aos/1031689016

Bevacqua, E. (2017). *CDVineCopulaConditional: Sampling from conditional c- and d-vine copulas*. https://CRAN.R-project.org/package=CDVineCopulaConditional

Bevacqua, E., Maraun, D., Hobæk Haff, I., Widmann, M., & Vrac, M. (2017). Multivariate statistical modelling of compound events via pair-copula constructions: Analysis of floods in ravenna (italy). *Hydrology and Earth System Sciences*, *21*(6), 2701–2723. https://doi.org/10.5194/hess-21-2701-2017

Cooke, R. M., Kurowicka, D., & Wilson, K. (2015). Sampling, conditionalizing, counting, merging, searching regular vines. *Journal of Multivariate Analysis*, *138*, 4–18. https://doi.org/10.1016/j.jmva.2015.02.001

Czado, C., & Nagler, T. (2021). Vine Copula Based Modeling. *Annual Review of Statistics and Its Application*, *9*, 453–477. https://doi.org/10.1146/annurev-statistics-040220-101153

DataCebo. (n.d.). *Copulas: Create tabular synthetic data using copulas-based modeling*. PyPI. Retrieved March 7, 2024, from https://pypi.org/project/copulas/

Dißmann, J., Brechmann, E. C., Czado, C., & Kurowicka, D. (2013). Selecting and estimating regular vine copulae and application to financial returns. *Computational Statistics & Data Analysis*, *59*, 52–69. https://doi.org/10.1016/j.csda.2012.08.010

Eilander, D., Couasnon, A., Sperna Weiland, F. C., Ligtvoet, W., Bouwman, A., Winsemius, H. C., & Ward, P. J. (2023). Modeling compound flood risk and risk reduction using a globally applicable framework: A pilot in the sofala province of mozambique. *Natural Hazards and Earth System Sciences*, *23*(6), 2251–2272. https://doi.org/10.5194/nhess-23-2251-2023

Hofert, M., Kojadinovic, I., Maechler, M., & Yan, J. (2023). *Copula: Multivariate dependence

with copulas. https://CRAN.R-project.org/package=copula

Joe, H. (1997). *Multivariate models and multivariate dependence concepts*. Chapman; Hall/CRC. https://doi.org/10.1201/9780367803896

Klein, N., Kneib, T., Marra, G., & Radice, R. (2020). Bayesian mixed binary-continuous copula regression with an application to childhood undernutrition. *Elsevier eBooks*, 121–152. https://doi.org/10.1016/b978-0-12-815862-3.00011-1

Mitskopoulos, L., Amvrosiadis, T., & Onken, A. (2022). Mixed vine copula flows for flexible modeling of neural dependencies. *Frontiers in Neuroscience*, *16*. https://doi.org/10.3389/fnins.2022.910122

Nagler, T., Schepsmeier, U., Stoeber, J., Brechmann, E. C., Graeler, B., Erhardt, T., Almeida, C., Min, A., Czado, C., Hofmann, M., Killiches, M., Joe, H., & Vatter, T. (2023). *VineCopula: Statistical inference of vine copulas*. https://cran.r-project.org/web/packages/VineCopula/index.html

Nagler, T., & Vatter, T. (2023). *Pyvinecopulib*. Zenodo. https://doi.org/10.5281/ZENODO.10435751

Sklar, M. (1959). Fonctions de repartition an dimensions et leurs marges. *Publ. Inst. Statist. Univ. Paris*, *8*, 229–231.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., … SciPy 1.0 Contributors. (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, *17*, 261–272. https://doi.org/10.1038/s41592-019-0686-2