The Journal of Open Source Software

Synthia: multidimensional synthetic data generation in Python

David Meyer[1, 2] **and Thomas Nagler**[3]

1 Department of Meteorology, University of Reading, Reading, UK **2** Department of Civil and Environmental Engineering, Imperial College London, London, UK **3** Mathematical Institute, Leiden University, Leiden, The Netherlands

DOI: 10.21105/joss.0XXXX

Software

- Review
- Repository
- Archive

Editor: Editor Name

Submitted: 01 January XXXX
Published: 01 January XXXX

Summary

Synthetic data are artificially generated data, not obtainable by direct measurements (McGraw-Hill, 2003). To serve a similar purpose to real data, they need to preserve the statistical properties in terms of their individual behavior and (inter-)dependences (Meyer et al., 2021). Copula and functional Principle Component Analysis (fPCA) are statistical models that allow these properties to be simulated (Joe, 2014). As such, copula generated data have shown potential for improving the generalization of machine learning (ML) emulators (Meyer et al., 2021) or for anonymizing real-data datasets (Patki et al., 2016). Although several synthetic data generation software exist (Patki et al., 2016; Xu & Veeramachaneni, 2018), to our knowledge, none offer a simple interface for working with multidimensional labelled datasets using copula and fPCA models.

Synthia is an open source Python package to model univariate and multivariate data, parameterize data using empirical and parametric methods, and manipulate marginal distributions. It is designed to enable scientists and practitioners to handle labelled multivariate data typical of computational sciences. Synthia supports three methods of multivariate data generation through fPCA, parametric (Gaussian) copula, and vine copula models for continuous (all), discrete (vine), and categorical (vine) variables. It has a simple and succinct API to natively handle xarray's (Hoyer & Hamman, 2017) labelled arrays and datasets. It uses a pure Python implementation for fPCA and Gaussian copula, and relies on the fast and well tested C++ library vinecopulib (Nagler & Vatter, 2020b) through pyvinecopulib's (Nagler & Vatter, 2020a) bindings for fast and efficient computation of vines.

Synthia has already been used to generate augmented datasets in Meyer et al. (2021) for improving the predictions of a ML emulator. With the release of Synthia, we look forward to enabling the generation of synthetic data from various scientific communities and experts alike.

Acknowledgments

We thank Maik Riechert for his comments and contributions to the project.

References

Hoyer, S., & Hamman, J. (2017). Xarray: N-D labeled arrays and datasets in Python. *Journal of Open Research Software*, *5*(1). https://doi.org/10.5334/jors.148

Joe, H. (2014). *Dependence Modeling with Copulas* (Zeroth). Chapman and Hall/CRC. https://doi.org/10.1201/b17116

McGraw-Hill. (2003). *McGraw-Hill Dictionary of Scientific and Technical Terms* (6th ed.). McGraw-Hill Education.

Meyer, D., Nagler, T., & Hogan, R. J. (2021). *Copula-based synthetic data generation for machine learning emulators in weather and climate: Application to a simple radiation model*. https://doi.org/10.5194/gmd-2020-427

Nagler, T., & Vatter, T. (2020a). *Pyvinecopulib*. Zenodo. https://doi.org/10.5281/zenodo.4288292

Nagler, T., & Vatter, T. (2020b). *Vinecopulib*. Zenodo. https://doi.org/10.5281/zenodo.4287554

Patki, N., Wedge, R., & Veeramachaneni, K. (2016). The Synthetic Data Vault. *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, 399–410. https://doi.org/10.1109/DSAA.2016.49

Xu, L., & Veeramachaneni, K. (2018). *Synthesizing tabular data using generative adversarial networks*. http://arxiv.org/abs/1811.11264