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.02863

Software

- Review
- Repository
- Archive

Editor: Olivia Guest

Reviewers:

- @khinsen
- @mnarayan

Submitted: 22 October 2020
Published: 24 September 2021

Summary and Statement of Need

Synthetic data – artificially generated data that mimic the original (observed) data by preserving relationships between variables (Nowok et al., 2016) – may be useful in several areas such as healthcare, finance, data science, and machine learning (Dahmen & Cook, 2019; Kamthe et al., 2021; Nowok et al., 2016; Patki et al., 2016). As such, copula-based data generation models – probabilistic models that allow for the statistical properties of observed data to be modelled in terms of individual behavior and (inter-)dependencies (Joe, 2014) – have shown potential in several applications such as finance, data science, and meteorology (Kamthe et al., 2021; Li et al., 2020; Meyer, Nagler, et al., 2021; Patki et al., 2016). Although copula-based data generation tools have been developed for tabular data – e.g. the Synthetic Data Vault project using Gaussian copulas and generative adversarial networks (Patki et al., 2016; Xu & Veeramachaneni, 2018), or the Synthetic Data Generation via Gaussian Copula (Li et al., 2020) – in computational sciences such as weather and climate, data often consist of large, labelled multidimensional datasets with complex dependencies.

Here we introduce Synthia, an open-source multidimensional synthetic data generator written in Python for xarray's (Hoyer & Hamman, 2017) labelled arrays and datasets with support for parametric and vine copulas models and functional principal component analysis (fPCA) – an extension of principal component analysis where data consist of functions instead of vectors (Ramsay & Silverman, 2005) – to allow for a wide range of data and dependent structures to be modelled. For efficiency, algorithms are implemented in NumPy (Harris et al., 2020) and SciPy (SciPy 1.0 Contributors et al., 2020) for Gaussian (parametric) copula and fPCA classes and rely on the C++ library vinecopulib (Nagler & Vatter, 2020b) through pyvinecopulib's (Nagler & Vatter, 2020a) bindings for fast computation of vines.

Recent applications of Synthia include the generation of dependent (Meyer, Nagler, et al., 2021) and independent (Meyer, Hogan, et al., 2021) samples for improving the prediction of machine learning emulators in weather and climate. In this release we include examples and tutorials for univariate and multivariate synthetic data generation using copula and fPCA methods and look forward to enabling the generation of synthetic data in various scientific communities and for several applications.

Acknowledgments

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

References

Dahmen, J., & Cook, D. (2019). SynSys: A Synthetic Data Generation System for Healthcare Applications. *Sensors*, *19*(5), 1181. https://doi.org/10.3390/s19051181

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., … Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, *585*(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2

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* (1st Edition). Chapman and Hall/CRC. https://doi.org/10.1201/b17116

Kamthe, S., Assefa, S., & Deisenroth, M. (2021). Copula Flows for Synthetic Data Generation. *arXiv:2101.00598 [cs, Stat]*. http://arxiv.org/abs/2101.00598

Li, Z., Zhao, Y., & Fu, J. (2020). SynC: A Copula based Framework for Generating Synthetic Data from Aggregated Sources. *2020 International Conference on Data Mining Workshops (ICDMW)*, 571–578. https://doi.org/10.1109/ICDMW51313.2020.00082

Meyer, D., Hogan, R. J., Dueben, P. D., & Mason, S. L. (2021). Machine learning emulation of 3D cloud radiative effects. *Journal of Advances in Modeling Earth Systems*. https://doi.org/10.1029/2021MS002550

Meyer, D., Nagler, T., & Hogan, R. J. (2021). Copula-based synthetic data augmentation for machine-learning emulators. *Geoscientific Model Development*, *14*(8), 5205–5215. https://doi.org/10.5194/gmd-14-5205-2021

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

Nowok, B., Raab, G. M., & Dibben, C. (2016). Synthpop: Bespoke Creation of Synthetic Data in R. *Journal of Statistical Software*, *74*(11). https://doi.org/10.18637/jss.v074.i11

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

Ramsay, J. O., & Silverman, B. W. (2005). *Functional data analysis* (2nd ed). Springer. ISBN: 978-0-387-40080-8

SciPy 1.0 Contributors, 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., … van Mulbregt, P. (2020). SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods*, *17*(3), 261–272. https://doi.org/10.1038/s41592-019-0686-2

Xu, L., & Veeramachaneni, K. (2018). Synthesizing Tabular Data using Generative Adversarial Networks. *arXiv:1811.11264 [cs, Stat]*. http://arxiv.org/abs/1811.11264