Correlation Pattern Recognizer (Copatrec): Automating Nonlinear Model Selection from Data

Abstract. Regression relations, particularly those belonging to nonlinear families, are frequently utilized in statistical and simulation-based studies. However, despite the fact that many human-related systems are complex and exhibit nonlinear patterns, these methods have often been overlooked or are reduced to linear forms or discrete values due to issues related to the identification of functions and associated parameters from empirical data.

A currently developed package, the *Correlation Pattern Recognizer* (Copatrec), aims at simplifying the process of deriving accurate function models from input data. Copatrec achieves this through the identification of representative mathematical equations using Machine Learning techniques to perform parameter estimation, and automated evaluation, as well as visualization features to assist in a reliable and transparent (function) model selection process. Since, for instance, machine learning methods are highly dependent on the quality of the data and their distributions, built-in outlier detection functions are included in the package. The combination of the preprocessing, auto-model detection, and output analytical objects with the prediction (based on parameterizable techniques with comprehensively reported metrics) and visualization capabilities make Copatrec a tool that can be used for the purpose of data exploration, model fitting in statistical analysis, and for data-driven agent-based models (or simulation models more broadly) as a means to establish functions for the use as input parameters from empirical datasets. At the same time, the tool can be used for the analysis of simulation output (e.g., to assess the relationship to empirical or theoretical functions statistically).

This package is designed using Python (versions 3.6 and above), integrates functionality from packages such as SciPy, Scikit-learn, and Matplotlib, and is openly available via pip (https://pypi.org/project/copatrec/). Ongoing developments include a web application that makes it possible to use the application in your browser directly, hence making the analysis accessible to individuals that do not have any Python experience.

Keywords: Statistical Analysis · Nonlinear Regression · Model Selection · data-driven simulation models · simulation parameterization