
Copula application manual

Generating samples for high dimensional data

HOANG Caroline * 1

1. Introduction

We detail about our application contents to build a model, analyse it and generating samples from a dataset using copula for users. It contents 5 tabs:

- Loading data and pseudo observation transformation
- Data visualisation
- Building model
- Model evaluation
- Generating samples

2. Loading data and pseudo observation transformation

This tab contains loading the dataset by importing/dragging file. Then followed by a pseudo-observation transformation that will set uniformly the dataset from 0 to 1 values. If the dataset contains timeseries, it will be converted and divided into year, month, day, and seconds and depending the amount of time range we have we drop some of those.

Before starting

Upload your data

Drag and Drop or Select Files

Submit files

Figure 1. Upload your data

3. Data visualization

This tab only contains data visualization by pairs using pseudo-observation data. It is used to see the (tail) dependances between 2 variables.

4. Building model

This tab contains the process of building the model. As we mostly want to generate samples from dependent variables, we can manually select the columns we want to build the

Bivariate plot

Scatter plot with marginal histogram

Select first column

A51-3B01.WZ01

Select second column

PSC1-1P01.JZ03

Scatter plot

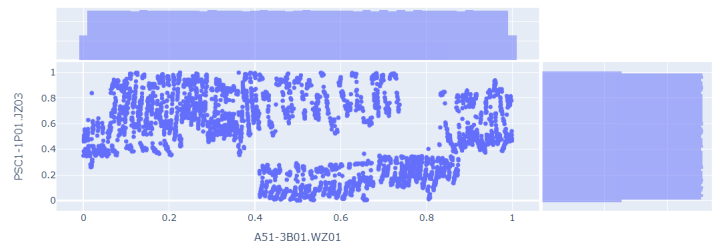


Figure 2. Bivariate plot for observed data in copula space

copula model. Building a copula model is exponential to the dimension of the data. It is based on pairs of variables and its dependency relation. Therefore, some pairs may not be correlated. In order to figure out their correlation we use either the mutual information or the Kendall's τ metric. We display their absolute value for a given metric each possible pair. For the mutual information criteria, we have to choose the parameters "bins" that is for the histogram plot that will be used to compute the mutual information. This plot allows us to choose a threshold value that will set to 0 the MI or Kendall's τ value during our model building process. Therefore, setting to independent those pairs. Once building the model, we can download it and visualized the structure with different level of graphs. The graph is interactive, we can click on the node to see the corresponding conditioning variables involved and also click on the edge to have the weight of it which is the Kendall's τ of the incoming node of the next tree We can select

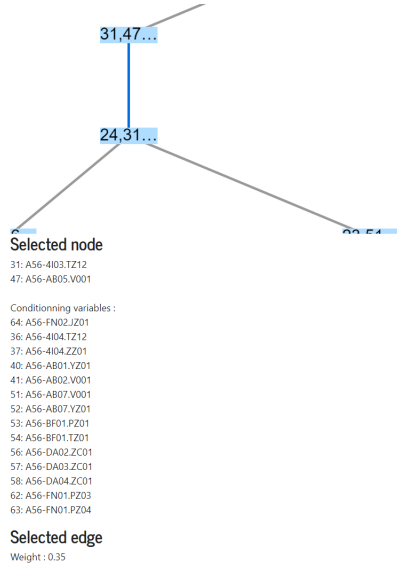


Figure 7. Result values while interacting with a node and an edge

Model evaluation

We want to assess whether it is reliable and adequate for future use, i.e., evaluate it in absolute terms.

Model evaluation therefore proceeds by comparing characteristics of the observed data, which was used for model specification, with simulated observations from the specified R-vine model

Upload observation and model	Bivariate plot	Empirical copula distribution	General QQ-plot	Mean of copula	Mutual Information	Mahalanobis distance & KS-test
<div>Upload your model</div> <div>Drag and Drop or Select model file</div>						
<div>Upload your pseudo observation data</div> <div>Drag and Drop or Select your data file</div>						
<div>Submit files and simulate samples</div>						

Figure 8. Interface model evaluation section

Data mean distribution We need alternative quantities. The most commonly used one is given by the mean of the copula data over its d components $S_i^K = \frac{1}{d} \sum_{r=1}^d u_{ir}^K$, and $S_i = \frac{1}{d} \sum_{r=1}^d u_{ir}$ for all $i \in [1, n]$ where n is the number of samples/observation. The appropriateness of model(K) can then be assessed by comparing histograms and empirical quantiles based on set of $\{S_i^K, i = 1, \dots, n\}$ and $\{S_i, i = 1, \dots, n\}$. We can extend the formula by adding personalize weight for each dimension. This approaches may not be in the copula space meaningful for independence pairs which means that we need to be careful about the variables chosen since independent induces an uniform distribution.

Statistical test We will use the KS-test on single variables in the copula to assess the similarity between the generated data and the observed data at the variable level. As it's only for single variable, we can choose a threshold for p-value

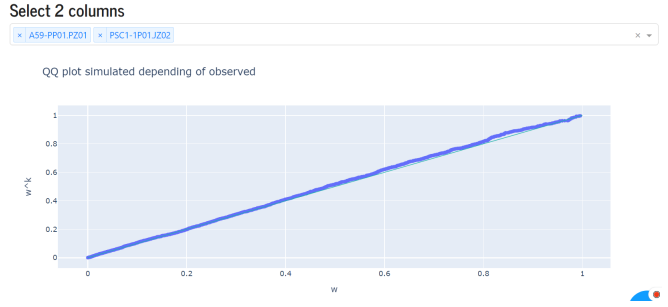


Figure 9. QQ-plot using 2 variables
We can see that the distribution induced from simulated data of this pair variable fits the observed data.



Figure 10. Empirical copula estimation on tails

for interpretation and count the number of variables where the hypothesis that those distributions (observed and simulated) are similar. We will use as well a metric called the Mahalanobis distance, which measures the distance between two multivariate datasets, taking into account the covariance structure. The smaller the value is, the better it is as it means that the 2 datasets are similar as the distance is closer.

6. Generating samples

This section is to generate samples from a given model and a given observed dataset used to transform the samples generated into the marginal space. The generated samples can be downloaded.

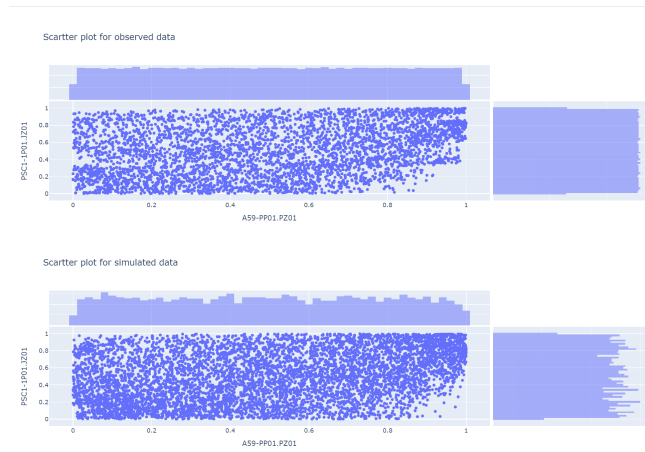
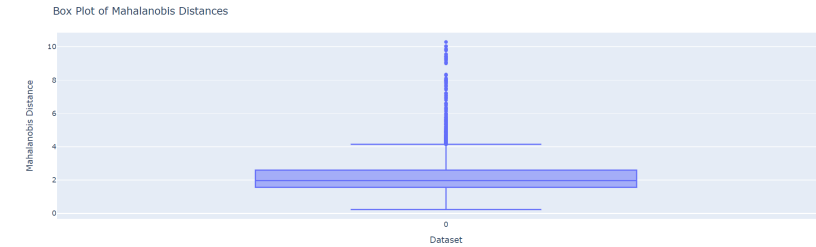


Figure 11. Bivariate scatter plot

Description

One such approach is the Mahalanobis distance, which measures the distance between two multivariate datasets, taking into account the covariance structure. The Mahalanobis distance provides a quantitative measure of the distance between the observed data and the generated data, taking into account the covariance structure. A smaller Mahalanobis distance suggests a better fit between the generated data and the observed data in terms of their multivariate characteristics.



Description

The count threshold approach using the KS test allows you to assess the similarity between the generated data and the observed data at the variable level. By counting the number of variables that meet a certain similarity criterion, you can gauge how well the generative model captures the distribution of the observed data across multiple variables.

Threshold

0.05

Count variables where p-value < 0.05 : 4 / 4

Figure 13. Mahalanobis distance and KS-test

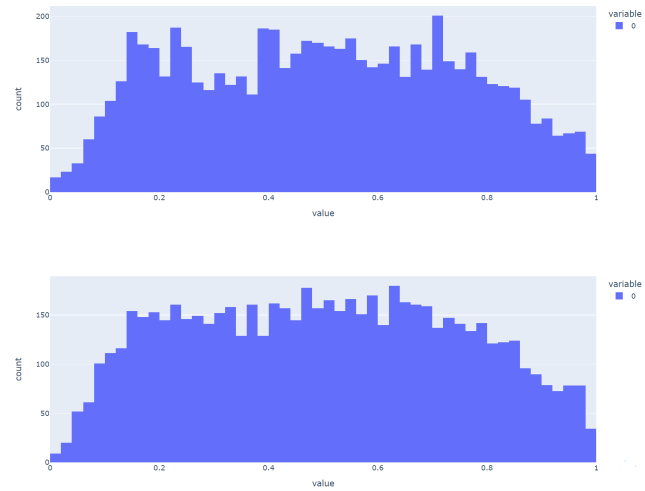


Figure 12. Histogram of the mean of each instances

Simulation and transformation

Simulate samples from model and transform it from copula space to real space

Upload observation and model

Upload your model

Drag and Drop or Select model file

Upload your real observation data

Drag and Drop or Select your data file

Submit files

Simulate samples

Select number of samples

1

Generate samples

Download CSV

Transform samples to real space

Transform samples

Download CSV

Figure 14. Interface to generate samples and transform it to real space