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/draging file. Then followed by a pseudo-observation transformation that will set uniformally 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.

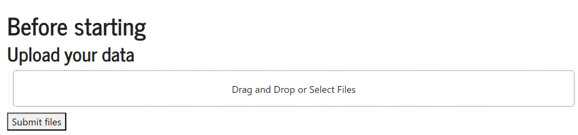

Figure 1. Upload your data

3. Data visualization

This tab only contains data visualization by pairs using pseudo-obervation 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 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 independaent 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 conditionning variables involved and also click on the edge to have the weight of it which is the Kendall's τ of the incomming node of the next tree We can select

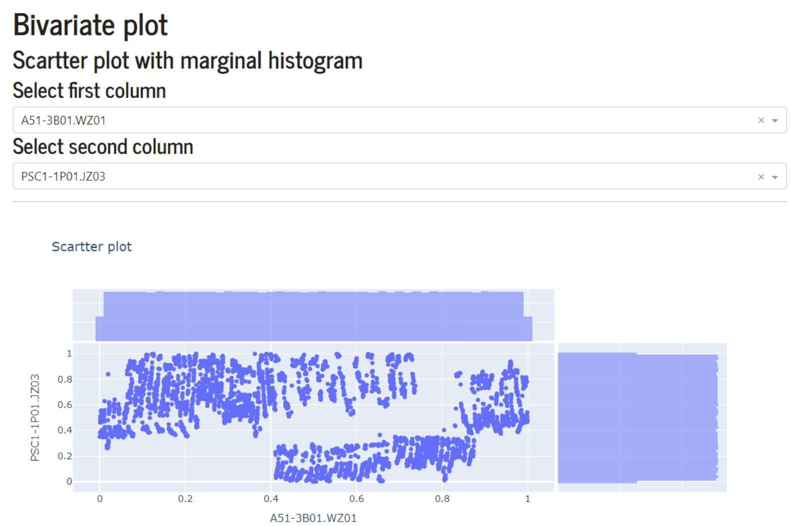

Figure 2. Bivariate plot for observed data in copula space

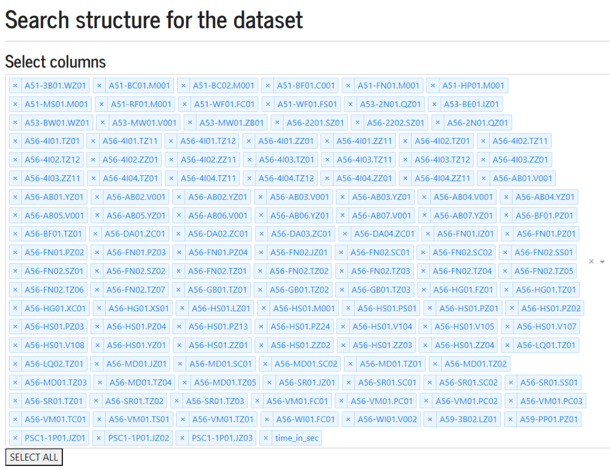

Figure 3. Bivariate plot for observed data in copula space

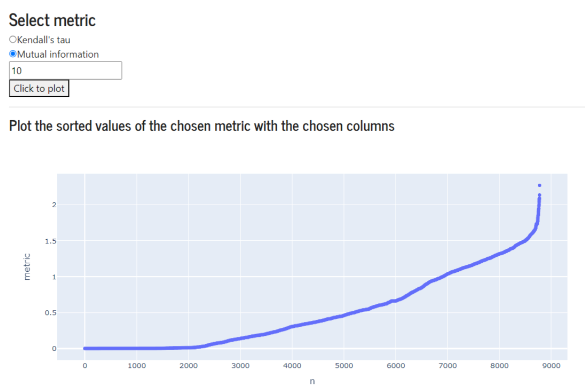

Figure 4. Mutual information by pairs with "bins" = 10

5. Model evaluation

We introduce many evaluations of the model including mostly bivariate comparison with the observed data.

QQ plot QQ-plot is used to compare the distribution of the observed and theorical distribution. This plot allows to compare two empirical cdf's, in our case the empirical cdf's of the observed and simulated data, where the observed data comes from the unknown true distribution and we want to investigate if the distribution induced by copula model is close to this true distribution. We compute :

$$w_i^K = \frac{1}{n} \sum_{j=1}^{n} \mathbf{1}_{u_{jr}^K \leq u_{is}^K, u_{js}^K \leq u_{ir}^K}$$

and

$$w_i = \frac{1}{n} \sum_{j=1}^{n} \mathbf{1}_{u_{jr} \leq u_{is}, u_{js} \leq u_{ir}}$$

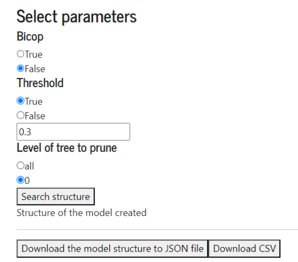

Figure 5. Select parameters : only structure (bicop = false), with threshold = 0.3 applied for the first level only

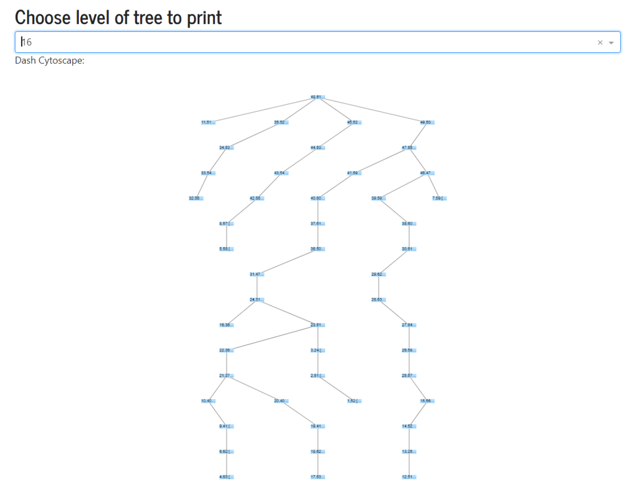

Figure 6. Tree visualization for the level 19

for $i \in [|1, n|]$ and $r, s \in [|1, d|]$ where n is the number of obersation and (r, s) a pair of variable. w_i^K corresponding to the truncated one so the simulated one

Empirical copula estimation and tail dependence If one is particularly interested in an accurate modeling of the joint tail behavior of variables, it might be interesting to consider the empirical copula distribution functions in the tails,

$$C_n(u_1, ..., u_d) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{U_{i1} \leq u1, ..., U_{id} \leq ud}$$

the pseudo-samples $U_1, ..., U_n$. These can be considered as samples from the underlying copula C. In this case it's interesting to look at bivariate one So we compute $C_n(\alpha, \alpha)$ (lower tail) and $C_n(1-\alpha, 1-\alpha)$ (upper tail) for $\alpha \in [0, 0.1]$. It allows us to observe the tail dependances, if the simulated one respects the behavior from observed one that we can see by computing the bivariate plot.

Bivariate scatter plot This section is for plotting in copula space the pair of variables. It tells the form of dependence between them, especially the tail.

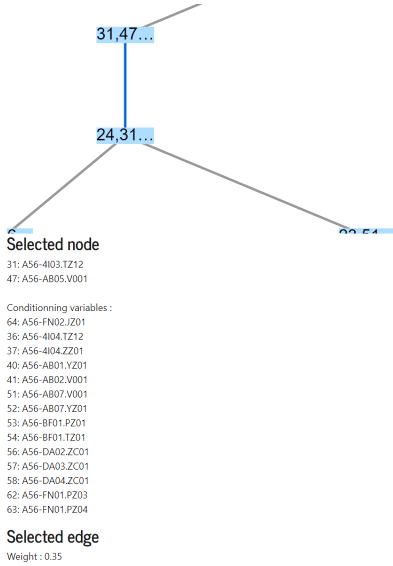

Figure 7. Result values while interacting with a node and an edge

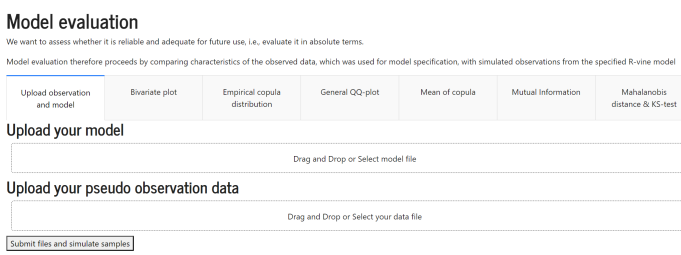

Figure 8. Interface model evaluation section

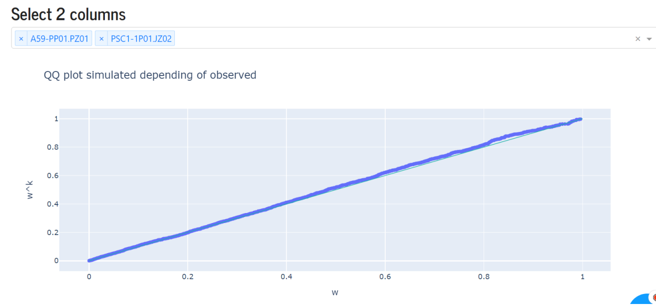

Figure 9. QQ-plot using 2 variables
We can see that the distribution induced from simulated data of this pair variable fits the observed data.

Figure 10. Empitical copula estimation on tails

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 $\{S_i^K, i = 1, ..., n\}$ and $\{S_i, i = 1, ..., n\}$ We can extand 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 unirform 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 for interpretation and count the number of variables where the hypothesis that those distributions (obeserved 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.

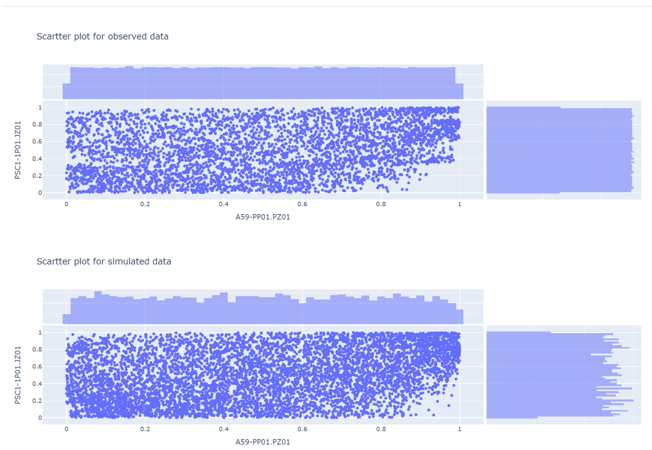

Figure 11. Bivariate scatter plot

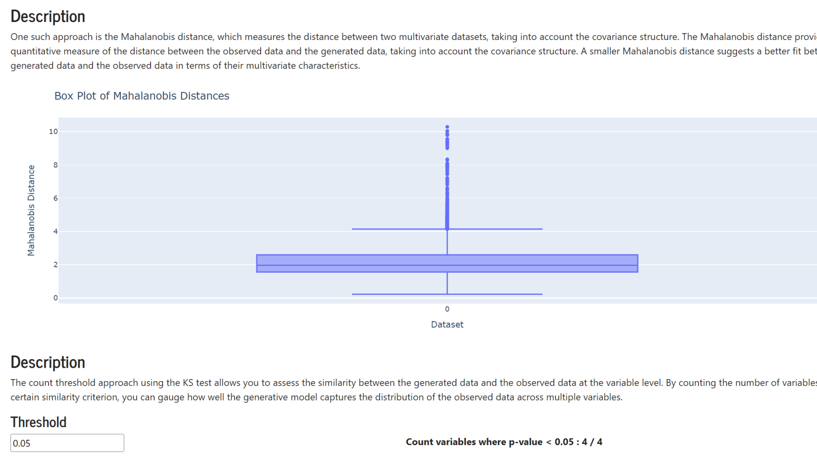

Figure 13. Mahalanobis distance and KS-test

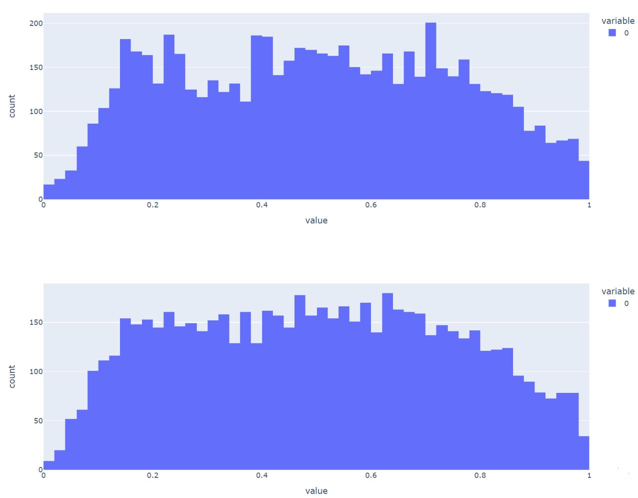

Figure 12. Histrogram of the mean of each instances

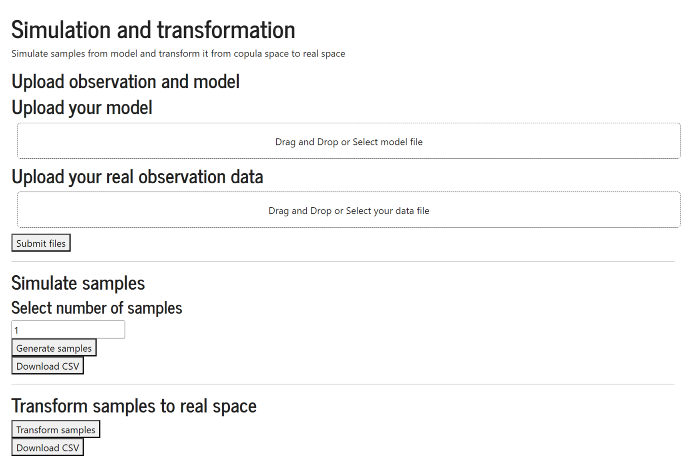

Figure 14. Interface to generate samples and transform it to real space