Federal State Autonomous Educational Institution of Higher Education

ITMO University

Methods and models for multivariate data analysis

Report on learning practice # 3

Sampling of multivariate random variables

Performed by:

Nemat Allah Aloush

(J4134c)

Saint-Petersburg

2021

Contents

1.Substantiation of chosen subsample

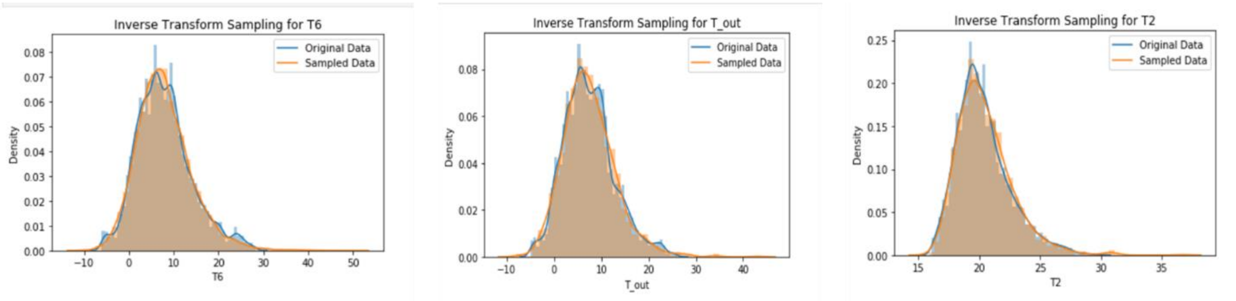
For this lab work, the following variables were chosen to investigate them:

- Overall – 10 variables: 'T1', 'T4', 'T5', 'RH_3','RH_7','RH_9', 'RH_out','T6','T2','T_out'.
- 3– target variables: 'T6','T2','T_out'.
- The rest - predictors: 'T1', 'T4', 'T5', 'RH_3','RH_7','RH_9', 'RH_out'.

2.Sampling of chosen target variables using univariate parametric distributions with 2 different sampling methods.

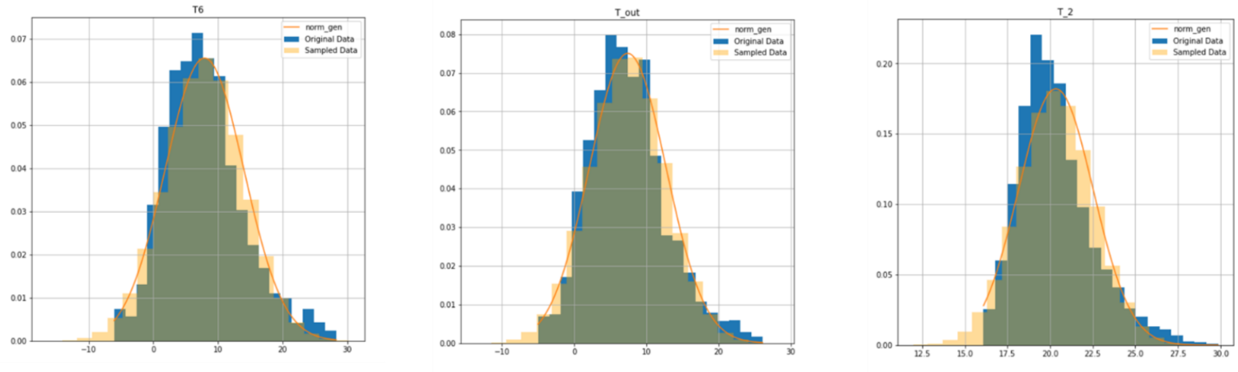
1. Inverse transform sampling

In the following figure, we can see in each plot the original data and the sampled data using Inverse transform sampling, for each target variable. We can see that the sampled data is quite similar to the original data.



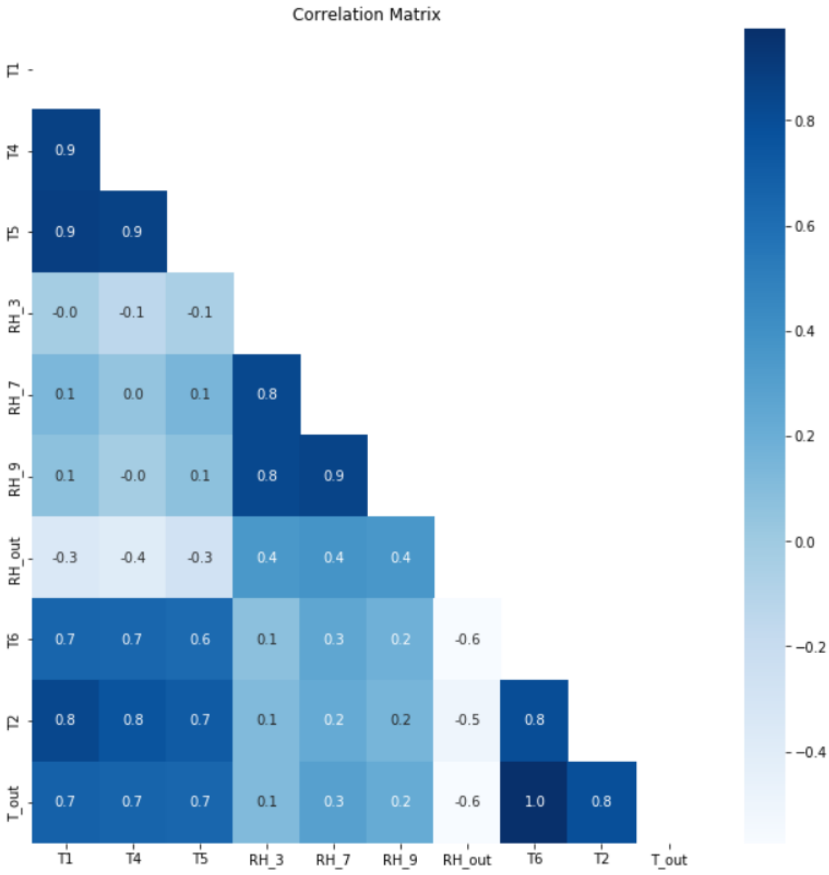
2. Accept-Reject Sampling

In the following figure, we can see in each plot the original data and the sampled data using Accept Reject Sampling, for each target variable. We can see that the sampled data is quite similar to the original data.



3. Estimation of relations between predictors and chosen target variables.

In the following figure, we can find the correlation matrix for our data to check the correlation between the chosen targets and the predictors.

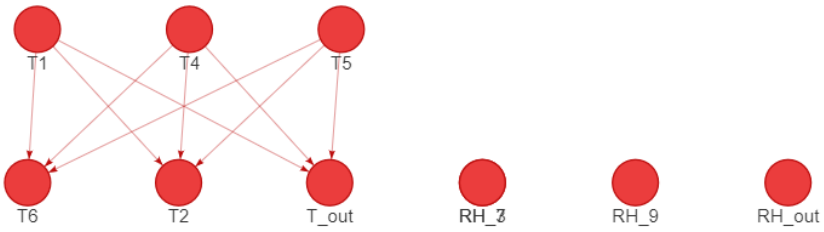


From the correlation matrix, we can notice that for each of our targets ('T6', 'T2', 'T_out') it seems that the most correlated predictors with them are ('T1', 'T5', 'T4').

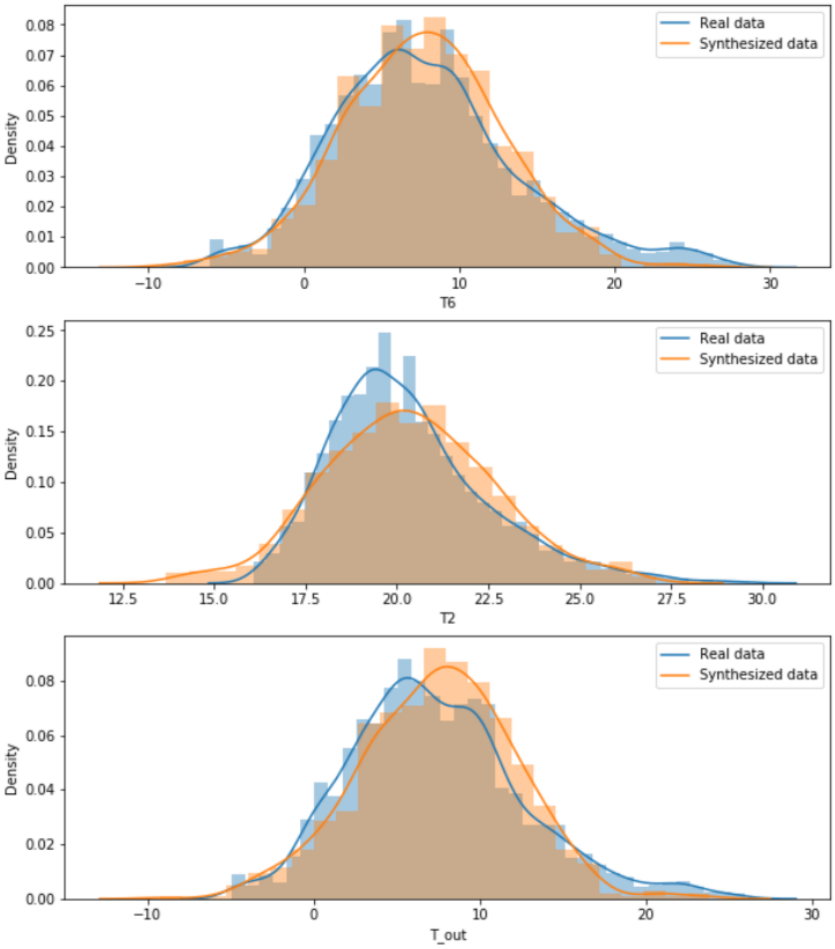
4.Bayesian networks

1. Manual Bayesian network

for this model I choose its structure depending on the correlation matrix. As already mentioned that for each of our targets ('T6', 'T2', 'T_out') it seems that the most correlated predictors with them are ('T1', 'T5', 'T4'). Thus, the chosen structure for this model is as following:



In the following figure we can find in each plot (for each target variable: T6, T2, T_out) the real data and the synthesized data.

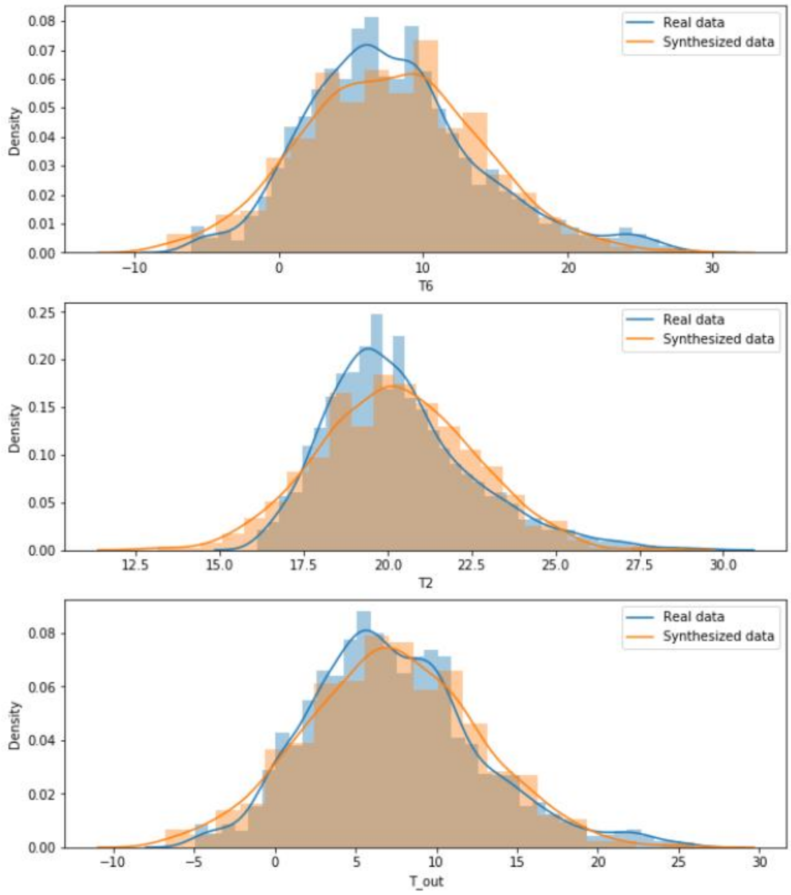


2. Structural learning model: Hill-Climbing with K2 score function

The structure for this model is as follows:

```
[['T1', 'T4'], ['T1', 'T2'], ['T1', 'RH_3'], ['T4', 'T_out'], ['T4', 'T2'],
['T4', 'RH_out'], ['T4', 'RH_3'], ['T5', 'T1'], ['T5', 'T4'], ['T5', 'RH_3'],
['T5', 'T_out'], ['T5', 'RH_out'], ['T5', 'T2'], ['RH_3', 'T2'], ['RH_3',
'T_out'], ['RH_3', 'RH_out'], ['RH_7', 'RH_out'], ['RH_7', 'RH_3'], ['RH_7',
'T_out'], ['RH_7', 'T4'], ['RH_7', 'T5'], ['RH_7', 'T1'], ['RH_7', 'T2'],
['RH_9', 'RH_7'], ['RH_9', 'RH_3'], ['RH_9', 'T5'], ['RH_9', 'T1'], ['RH_9',
'T2'], ['RH_9', 'T4'], ['RH_9', 'T_out'], ['RH_9', 'T6'], ['T2', 'T_out'], ['T2',
'RH_out'], ['T2', 'T6'], ['T_out', 'T6'], ['T_out', 'RH_out']].
```

In the following figure we can find in each plot (for each target variable: T6, T2, T_out) the real data and the synthesized data gained through Hill-Climbing with K2 score function.

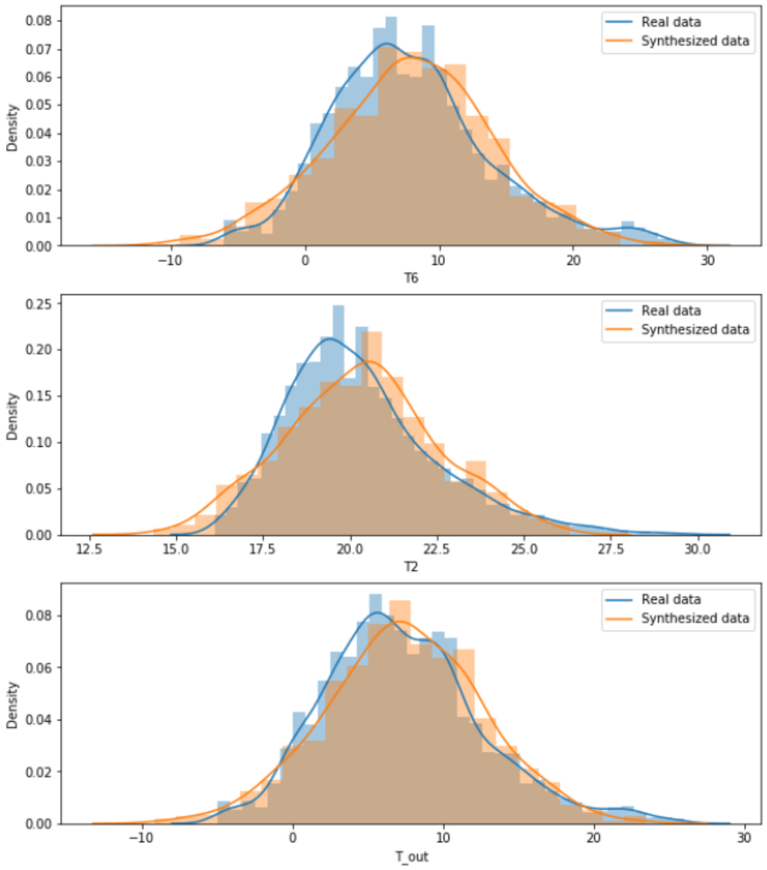


3. Structural learning model: Search strategy PC algorithms with MI score function

The structure for this model is as follows:

```
[['RH_7', 'T1'], ['T5', 'T1'], ['T5', 'RH_7'], ['T2', 'T1'], ['T2', 'RH_7'],
['T2', 'T5'], ['RH_out', 'T1'], ['RH_out', 'RH_7'], ['RH_out', 'T5'], ['RH_out',
'T2'], ['T6', 'T1'], ['T6', 'RH_7'], ['T6', 'T5'], ['T6', 'T2'], ['T6',
'RH_out'], ['T6', 'T4'], ['RH_9', 'T1'], ['RH_9', 'RH_7'], ['RH_9', 'T5'],
['RH_9', 'T2'], ['RH_9', 'RH_out'], ['RH_9', 'T4'], ['RH_9', 'T6'], ['T4', 'T1'],
['T4', 'RH_7'], ['T4', 'T5'], ['T4', 'T2'], ['T4', 'RH_out'], ['RH_3', 'T1'],
['RH_3', 'RH_7'], ['RH_3', 'T5'], ['RH_3', 'T2'], ['RH_3', 'RH_out'], ['RH_3',
'T4'], ['RH_3', 'T6'], ['RH_3', 'RH_9'], ['T_out', 'T1'], ['T_out', 'RH_7'],
['T_out', 'T5'], ['T_out', 'T2'], ['T_out', 'RH_out'], ['T_out', 'T4'], ['T_out',
'T6'], ['T_out', 'RH_9'], ['T_out', 'RH_3']].
```

In the following figure we can find in each plot (for each target variable: T6, T2, T_out) the real data and the synthesized data gained through PC algorithms with MI score function.



5.Quality analysis.

In the following table, we can see the accuracy for each target,

	Manual Bayesian network	Hill-Climbing with K2 score function	PC algorithms with MI score function
T6	1.182	1.321	1.321
T2	4.489	1.122	1.123
T_out	3.811	3.034	3.036

Through comparing the numerical results (accuracy) and the plotted graphs for the three targets and the three methods, it seems that Manual Bayesian network worked the best for T2 and T_out. While the other structural models yield similar results for the 3 targets.

6.Source code

Please find my code in the following GitHub link:

https://github.com/neematAllosh/MultiderivativeDataAnalysis/blob/master/lab_03.ipynb