

Title: Tuning DataSynthesizer
Authors: Jonathan P. Latner
Revision date: August 16, 2023

1 DataSynthesizer hyperparameters

https://github.com/DataResponsibly/DataSynthesizer/blob/master/notebooks/DataSynthesizer_correlated_attribute_mode.ipynb

- Total 63 combinations of tuning parameters (336 synthetic data sets)
- k (3: 0, 1, 2). The maximum number of parents in Bayesian network, i.e., the maximum number of incoming edges. Default is 0.
- epsilon (7: 0, 0.1, 1, 5, 10, 20, 30). It roughly means that removing a row in the input dataset will not change the probability of getting the same output more than a multiplicative difference of $\exp(\epsilon)$. Increase epsilon value to reduce the injected noises. Set epsilon=0 to turn off differential privacy. Default is 0.1.
- copies (3: 1, 5, 10): Number of synthetic data sets

2 Utility parameters

Following previous work, the utility of the synthetic and sample data was assessed using multiple measures. The confidence interval overlap (CIO) and ratio of counts/estimates (ROC) were calculated. This was to provide a more complete picture of the utility, rather than relying upon just one measure. The propensity score mean squared error (pMSE) was not used as, whilst it is suitable for analysing the synthetic data it is not suited to the analysis of sample data as it is structurally tied to the original data (since the sample data is a subset of the original data).

ROE is calculated by taking the ratio of the synthetic and original data estimates, where the smaller of these two estimates is divided by the larger one. Thus, given two corresponding estimates (e.g. totals, proportions), where y_{orig}^1 is the estimate orig from the original data and y_{synth}^1 is the corresponding estimate from the synthetic synth data, the ROE is calculated as:

$$ROE = \frac{\min(y_{orig}^1, y_{synth}^1)}{\max(y_{orig}^1, y_{synth}^1)} \quad (1)$$

If $y_{orig}^1 = y_{synth}^1$, then $ROE = 1$. The ROE will be calculated over bivariate and univariate data, and takes a value between 0 and 1. For each categorical variable the ratio of estimates are averaged across categories to give an overall ratio of estimates.

To calculate the CIO (using 95% confidence intervals), the coefficients from regression models built on the original and synthetic datasets are used. The CIO, proposed by Karr et al. [17],

is defined as:

$$CIO = \frac{1}{2} \left\{ \frac{\min(\mu_o, \mu_s) - \max(l_o, l_s)}{\mu_o - l_o} + \frac{\min(\mu_o, \mu_s) - \max(l_o, l_s)}{\mu_s - l_s} \right\} \quad (2)$$

where μ_o, l_o, l_s denote the respective upper and lower bounds of the confidence intervals for the original and synthetic data. This can be summarised by the average across all regression coefficients, with a higher CIO indicating greater utility (maximum value is 1 and a negative value indicating no overlap). Here, categorical variables are modified so that within each variable the maximum value within each variable is 1 and all other values are coded as 0. We use linear probability models to compare across variables. Each variable in the data set is a dependent variable and all other variables are independent variables, respectively.

3 Graphs

Figure 1: Frequency of counts

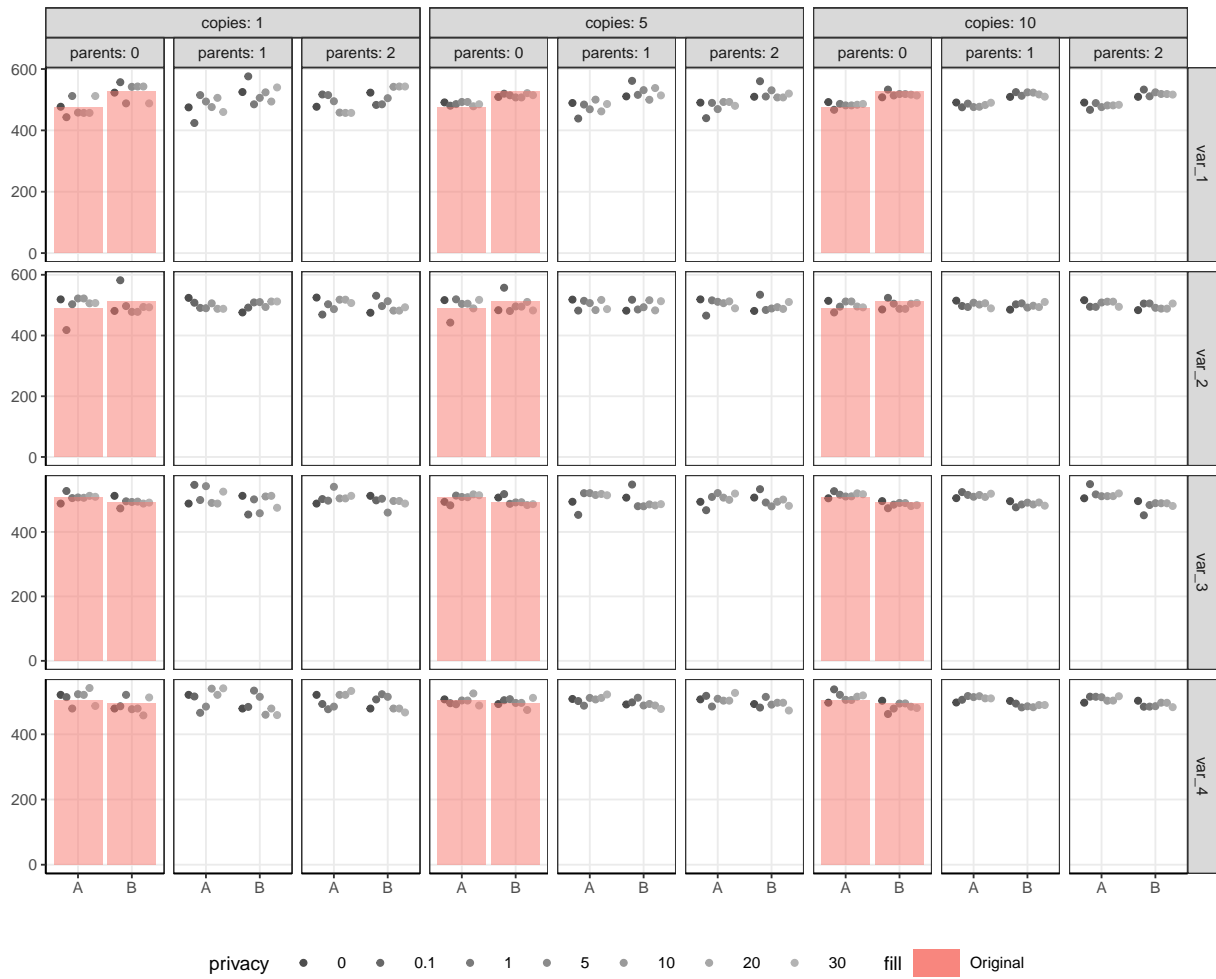


Figure 2: Ratio of estimates (ROE) - univariate

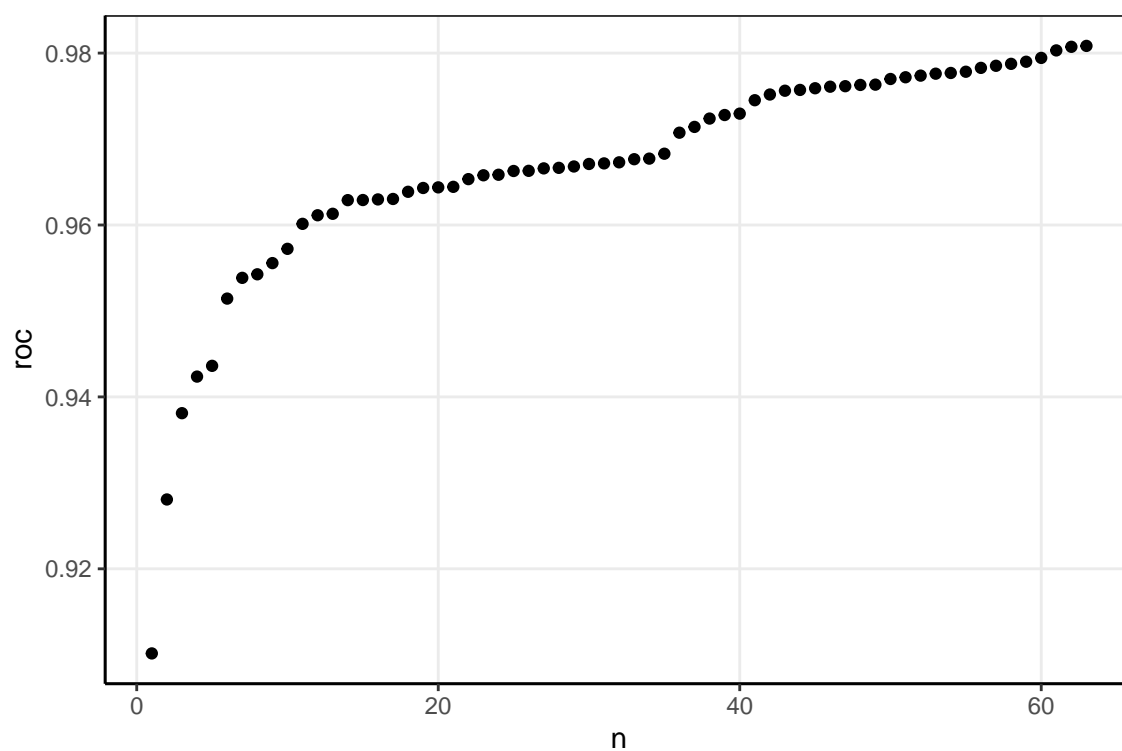


Figure 3: Parametric estimates of tuning parameters on ROE

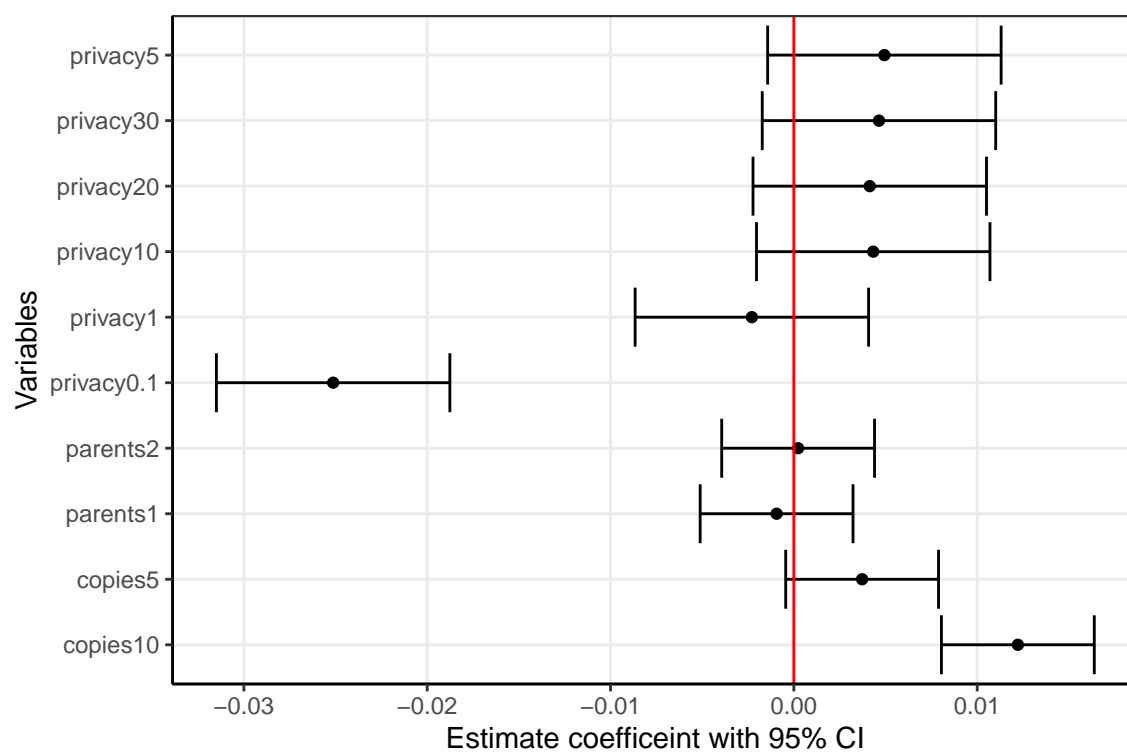


Figure 4: Parametric estimates of tuning parameters on ROE by copies (m)

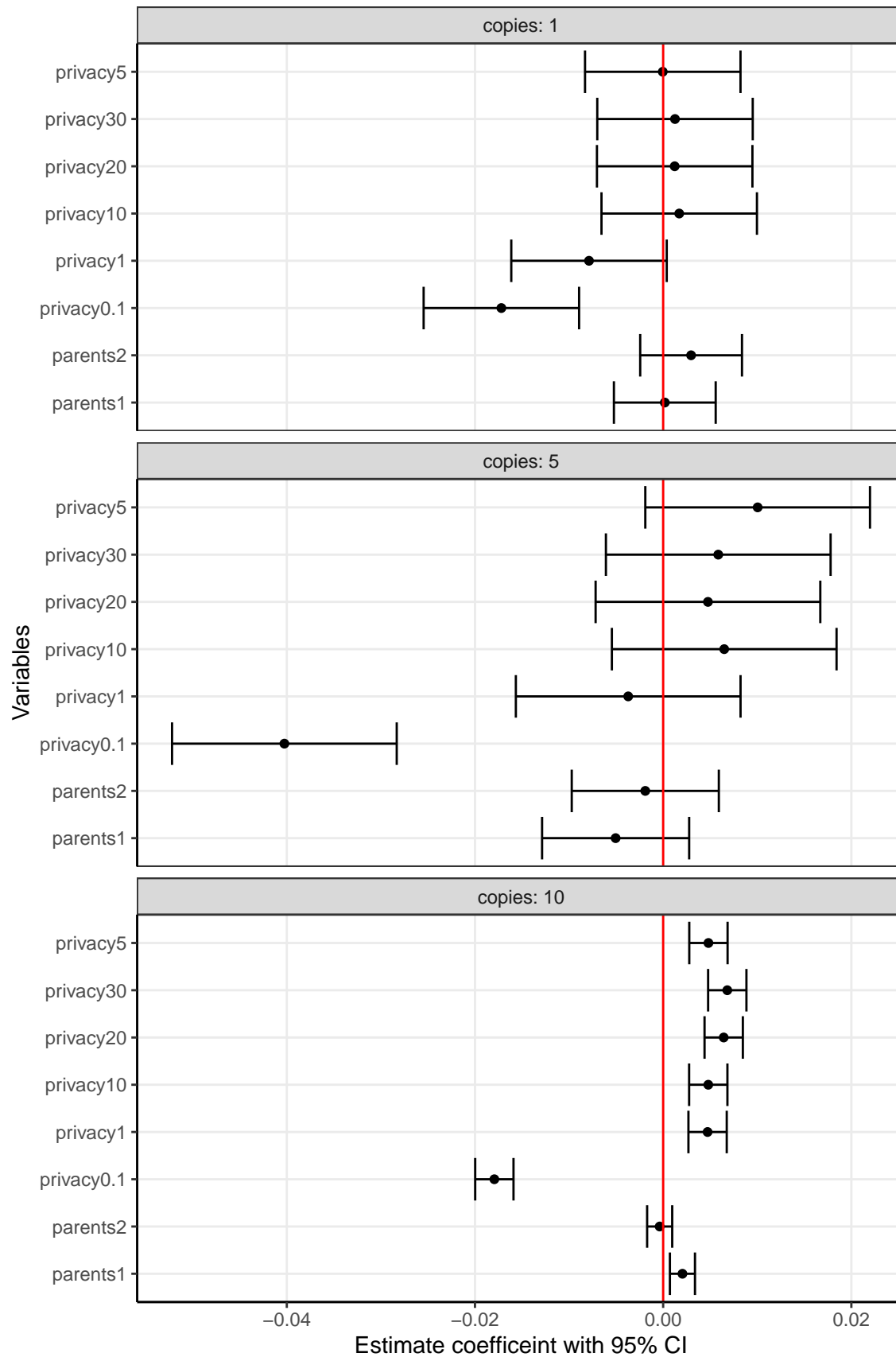


Figure 5: Ratio of estimates - confidence interval overlap (CIO)

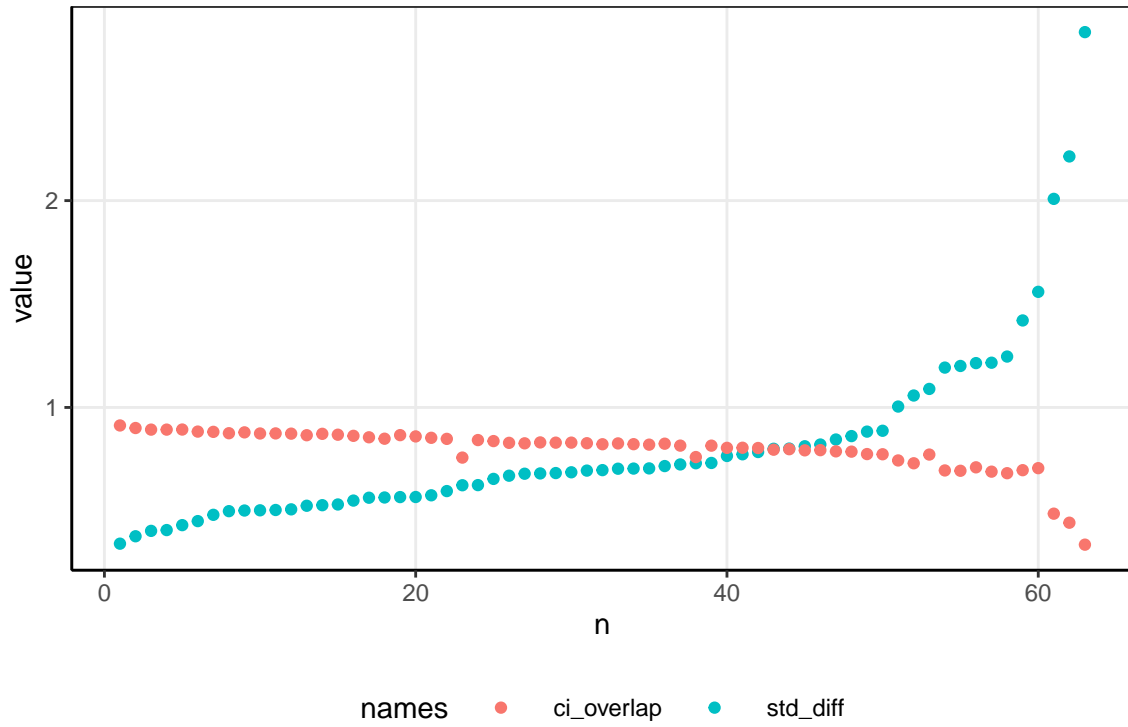


Figure 6: Parametric estimates of tuning parameters on CIO

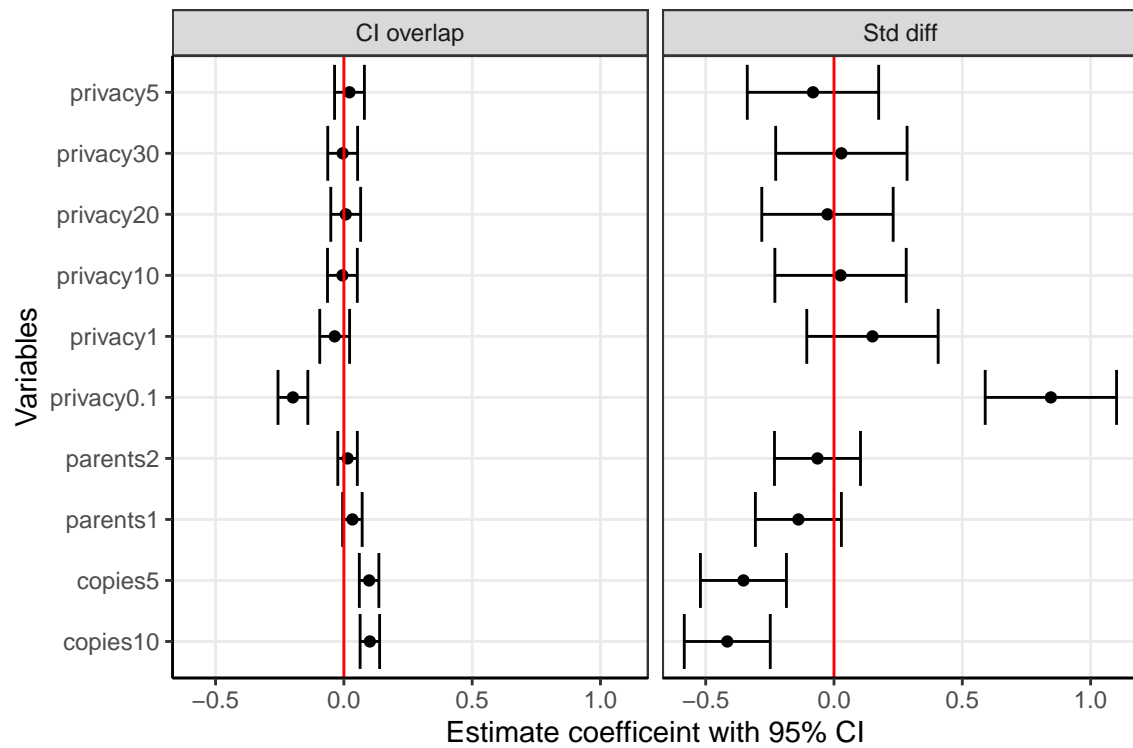


Figure 7: Parametric estimates of tuning parameters on CIO by copies (m)

