Title: Tuning CTGAN (SDV package)
Authors: Jonathan P. Latner
Revision date: August 16, 2023

1 CTGAN hyperparameters

https://sdv.dev/SDV/user_guides/single_table/ctgan.html#advanced-usage

- Total 108 combinations of tuning parameters (576 synthetic data sets)
- **epochs** (3: 300, 600, 900): the number of iterations that the model will perform to optimize its parameters. Default is 300.
- **batch_size** (2: 500, 1000): as well as the number of samples used in each step. Default is 500.
- **log_frequency** (2: True, False): Whether to use log frequency of categorical levels in conditional sampling. It defaults to True. This argument affects how the model processes the frequencies of the categorical values that are used to condition the rest of the values. In some cases, changing it to False could lead to better performance.
- **discriminator_steps** (3: 1, 5, 10): Number of discriminator updates to do for each generator update. From the WGAN paper: https://arxiv.org/abs/1701.07875. WGAN paper default is 5. Default is 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^1_{orig} is the estimate orig from the original data and y^1_{synth} is the corresponding estimate from the synthetic synth data, the ROE is calculated as:

$$ROE = \frac{min(y^1_{orig}, y^1_{synth})}{max(y^1_{orig}, y^1_{synth})} \tag{1}$$

If $y^1_{orig} = y^1_{synth}$, 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:

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

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

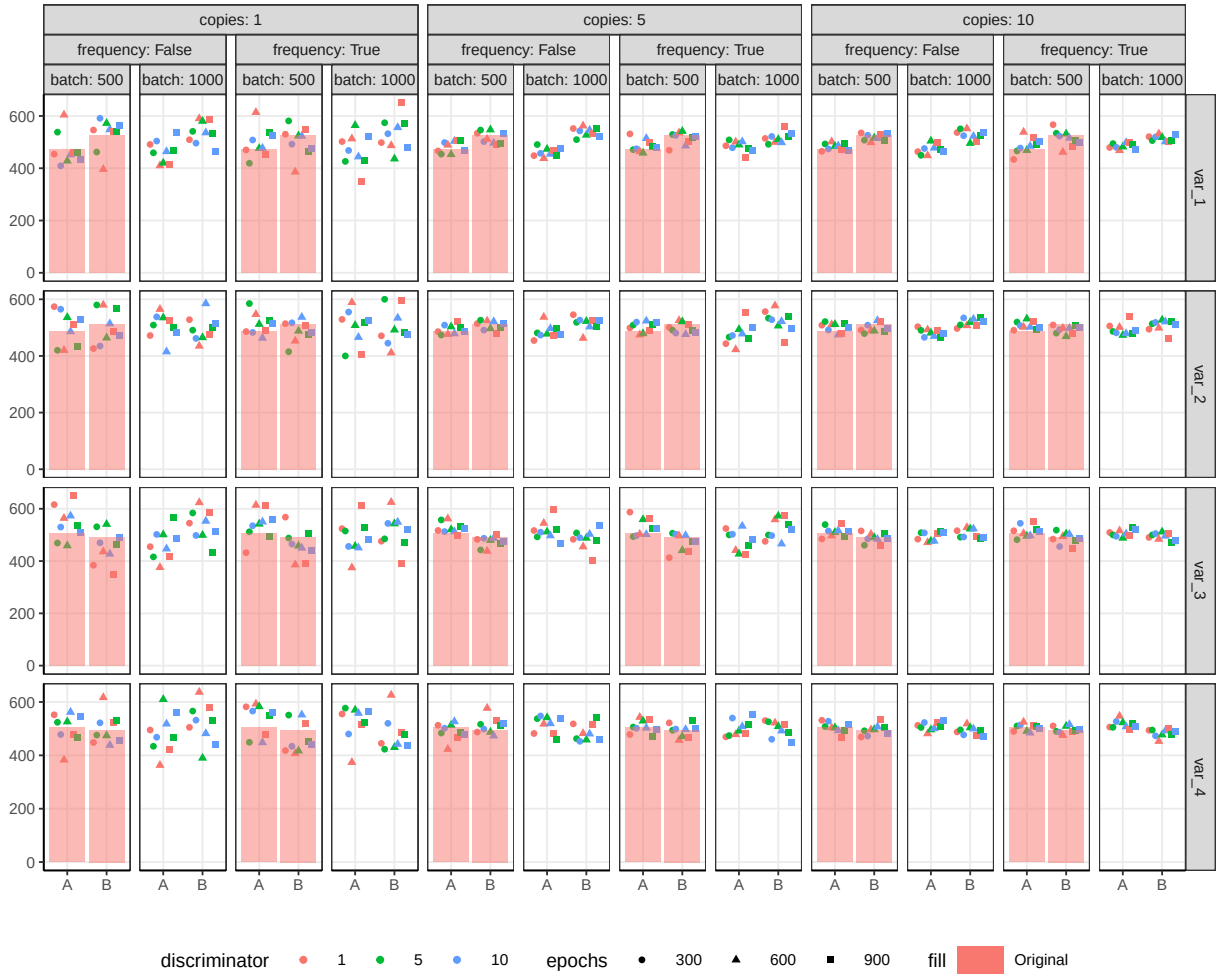

Figure 2: Ratio of estimates (ROE) - univariate

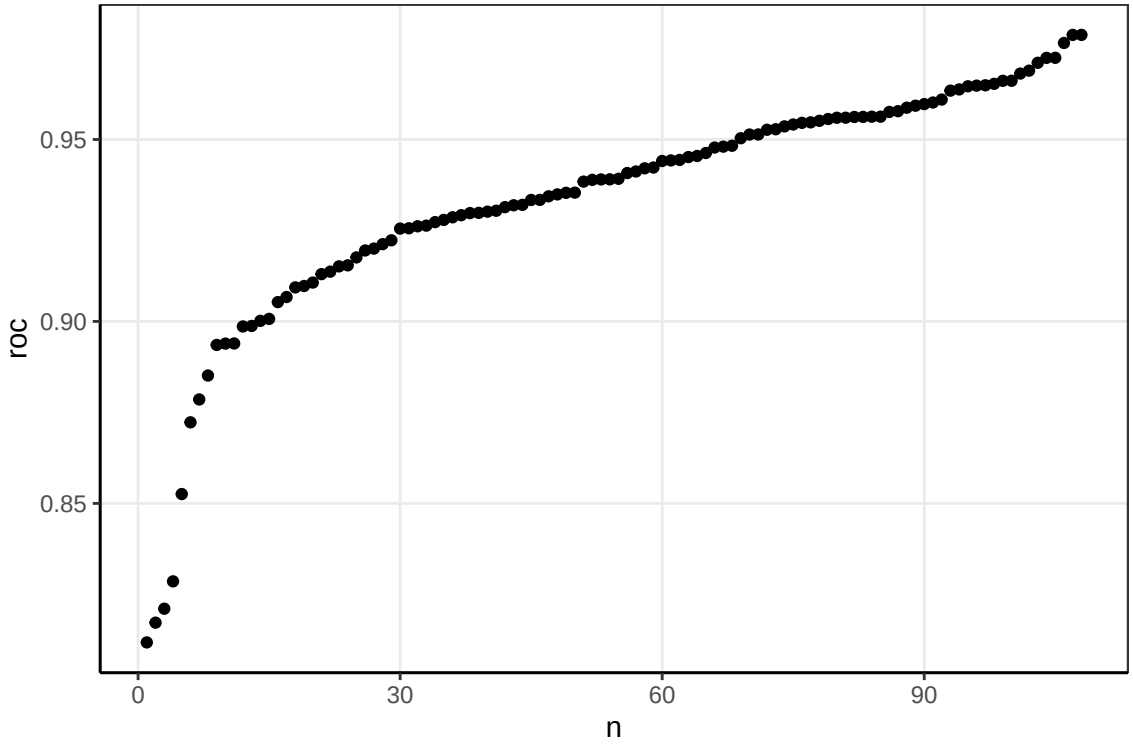

Figure 3: Parametric estimates of tuning parameters on ROE

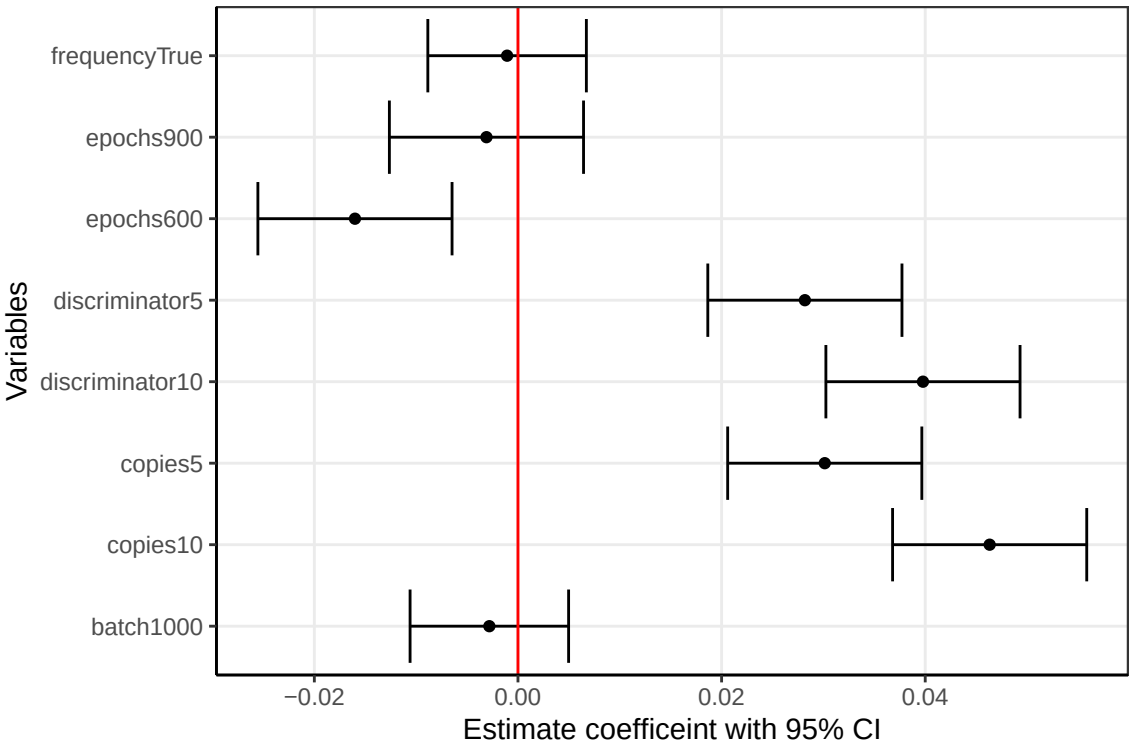

Figure 4: Parametric estimates of tuning parameters on ROE by copies (m)

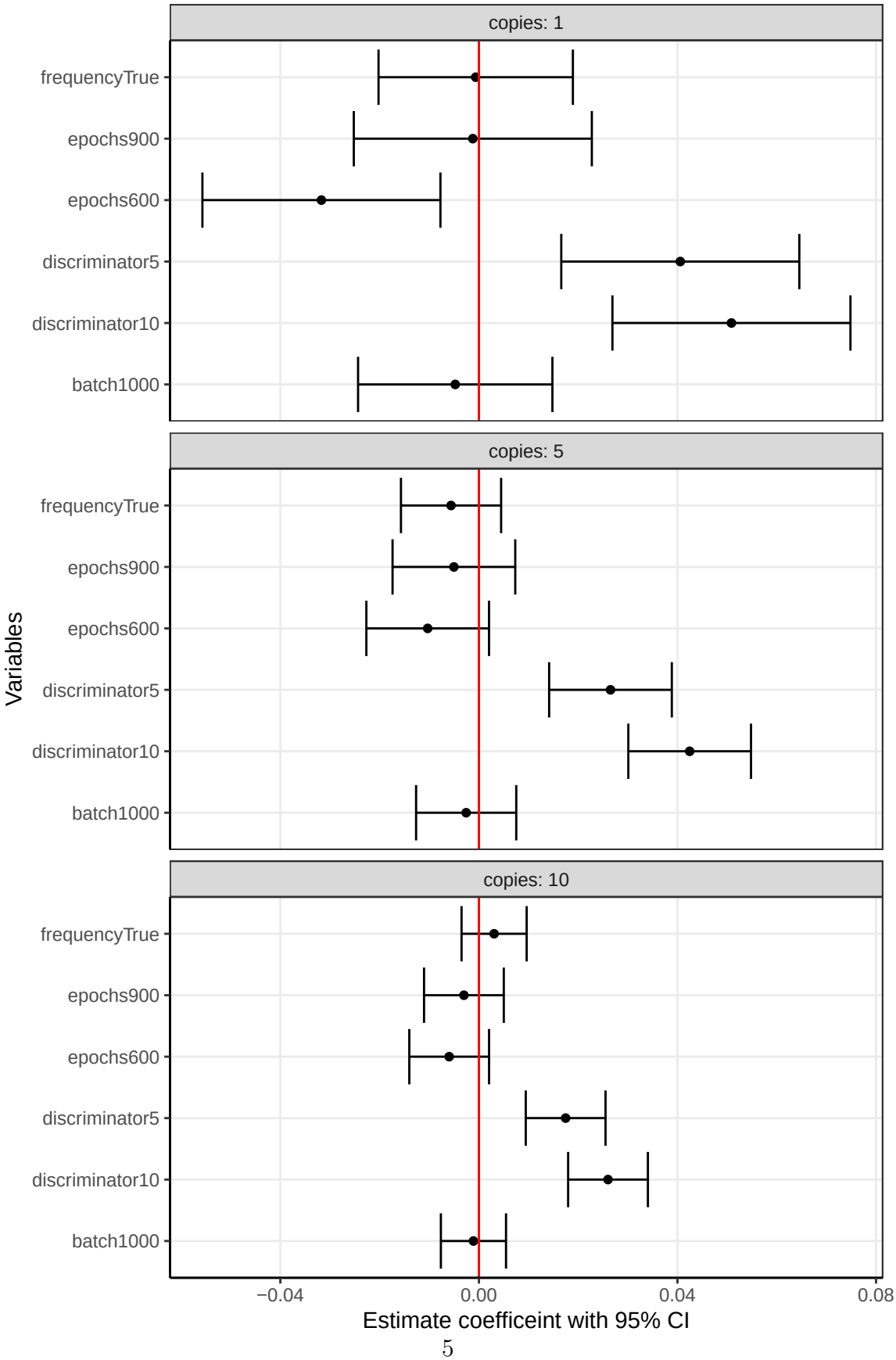

Figure 5: Ratio of estimates - confidence interval overlap (CIO)

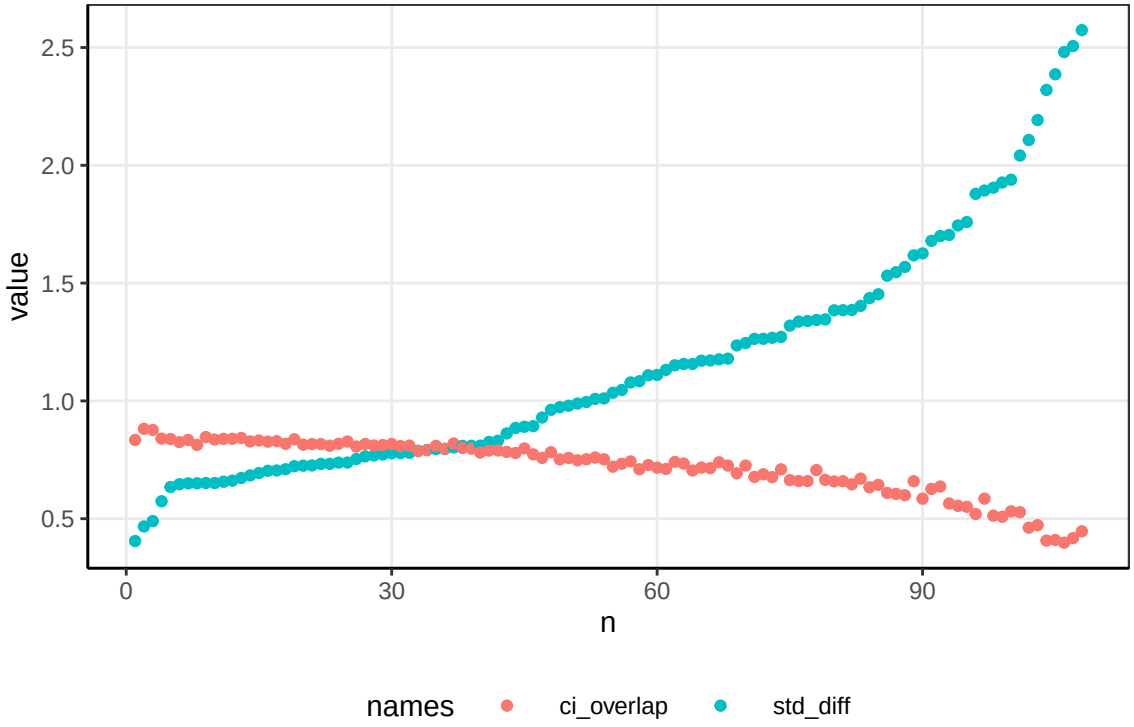

Figure 6: Parametric estimates of tuning parameters on CIO

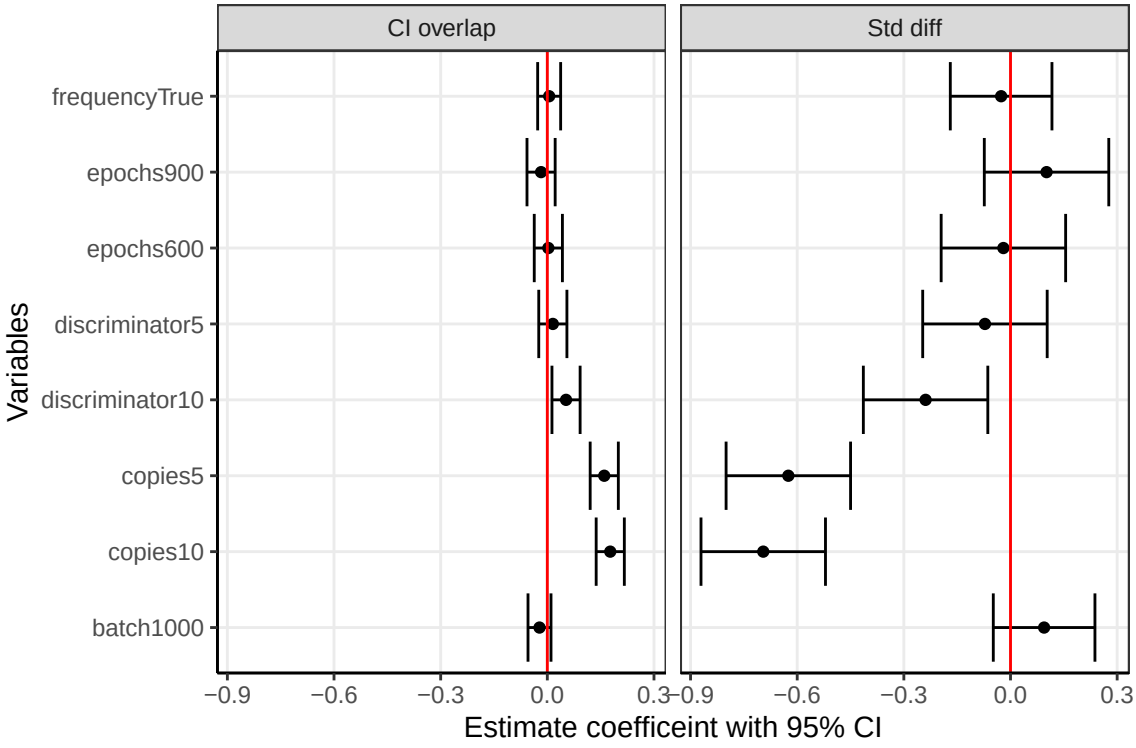

Figure 7: Parametric estimates of tuning parameters on CIO by copies (m)