Density ratios

Evaluating and improving the utility of synthetic data

Thom Benjamin Volker
t.b.volker@uu.nl

Imagine you have access to all the data in the world

What a privacy disaster would that be…

If real data is no option,

maybe synthetic data is!

If the synthetic data is good enough, it is almost as useful as real data

If the synthetic data is as useful as the real data, it is good enough

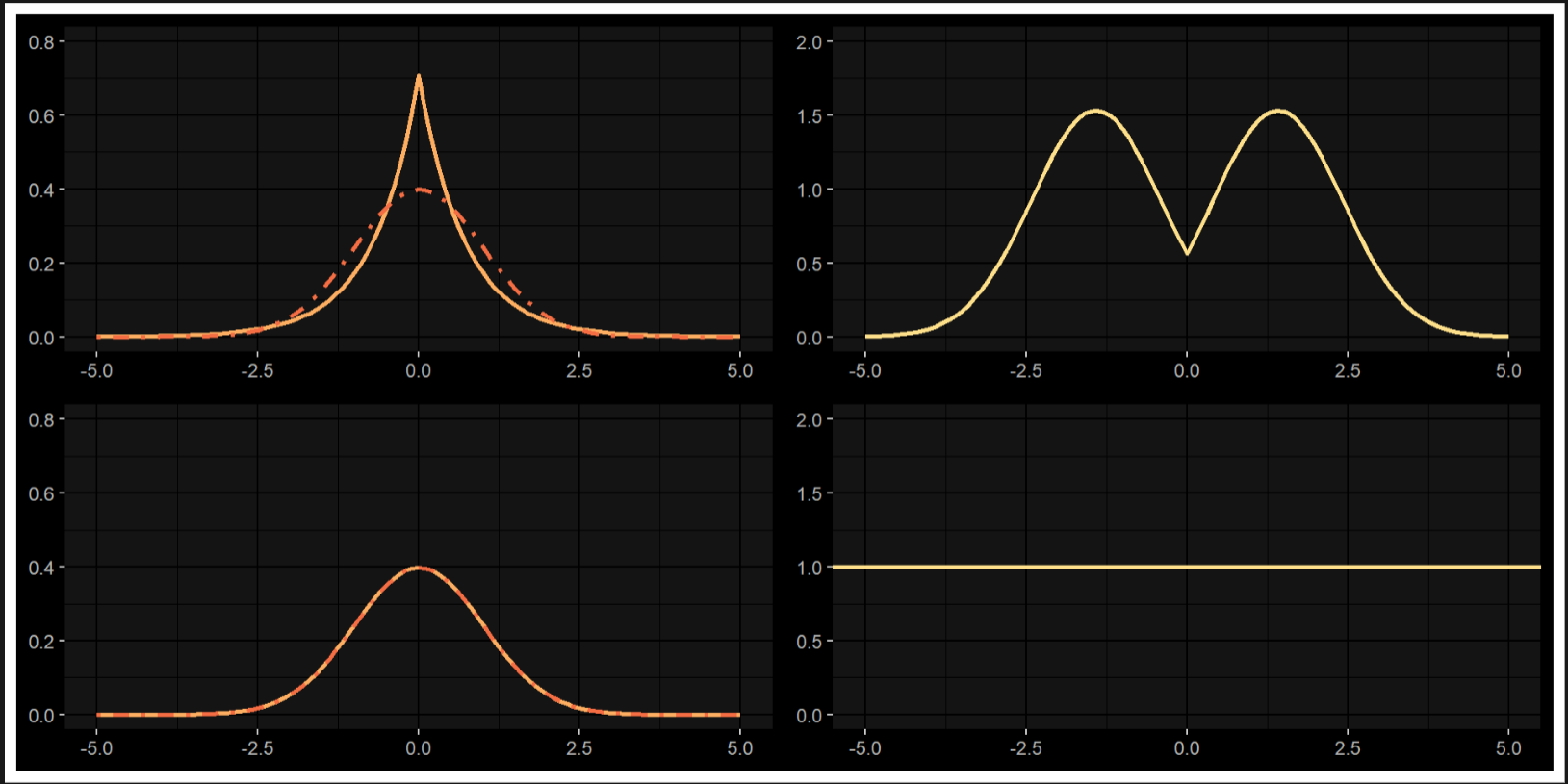
How can we tell whether the synthetic data is good enough?

Density ratios for utility[1]

$$r(x) = \frac{p(\boldsymbol{X}_{\text{syn}})}{p(\boldsymbol{X}_{\text{obs}})}$$

1. See *Masashi, Suzuki & Kanamori (2012). Density ratio estimation in machine learning.*

Density ratios for utility evaluation



Density ratios in practice

1. Estimate the density ratio directly and non-parametrically

- Implemented in R-package `densityratio`

2. Calculate a discrepancy measure for the synthetic data

- Kullback-Leibler divergence; Pearson divergence

3. Compare discrepancy measures for different data sets
4. Optionally: Test the null hypothesis $p(\boldsymbol{X}_{\text{syn}}) = p(\boldsymbol{X}_{\text{obs}})$

Density ratios for synthetic data (multivariate examples)

U.S. Current Population Survey (n = 5000)[1]

- Four continuous variables (*age, income, social security payments, household taxes*)
- Four categorical variables (*sex, race, marital status, educational attainment*)

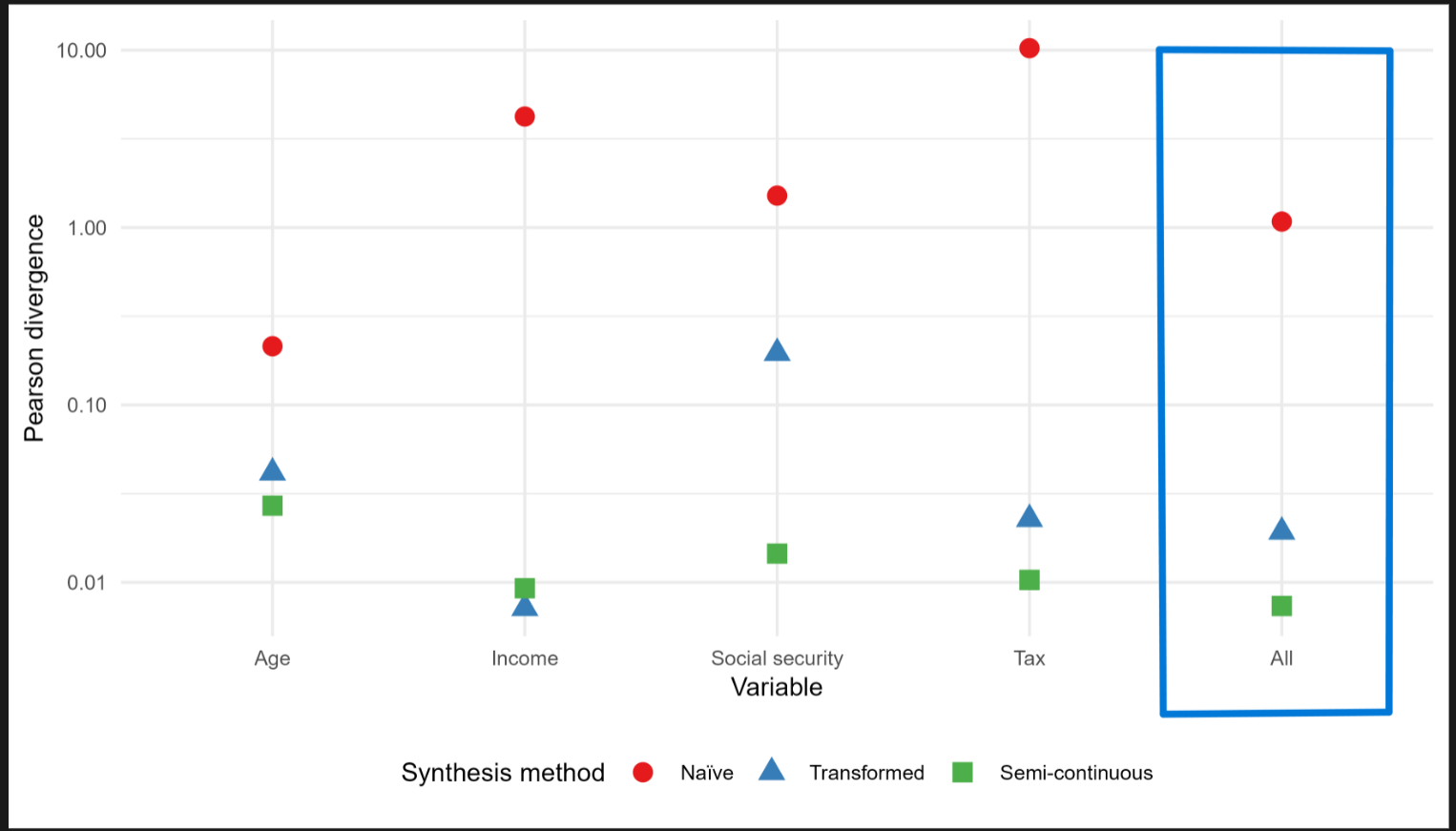
Synthetic data models

(Multinomial) logistic regression for categorical variables

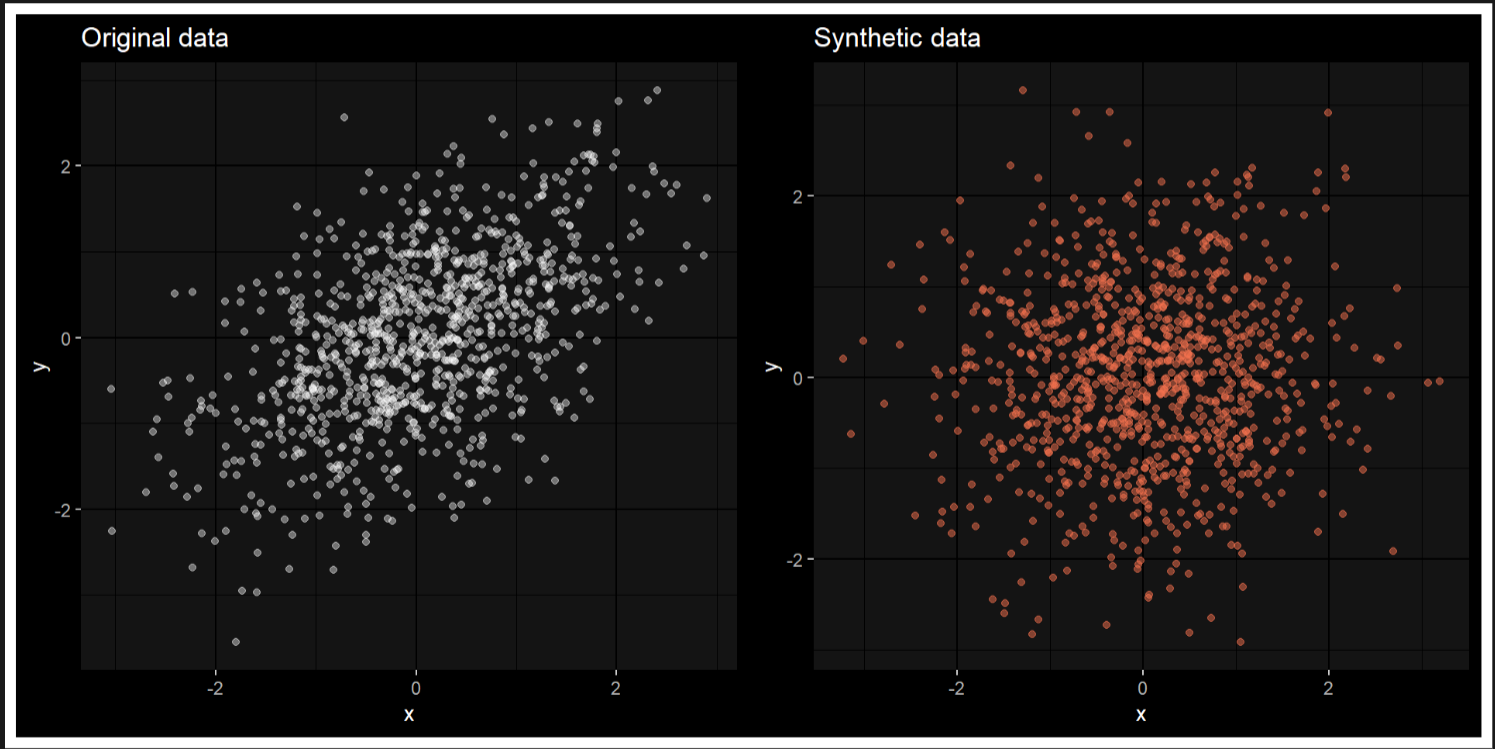
1. Linear regression
2. Linear regression with transformations (cubic root)
3. Linear regression with transformations and semi-continuous modelling

1. Thanks to Jörg Drechsler for sharing the data.

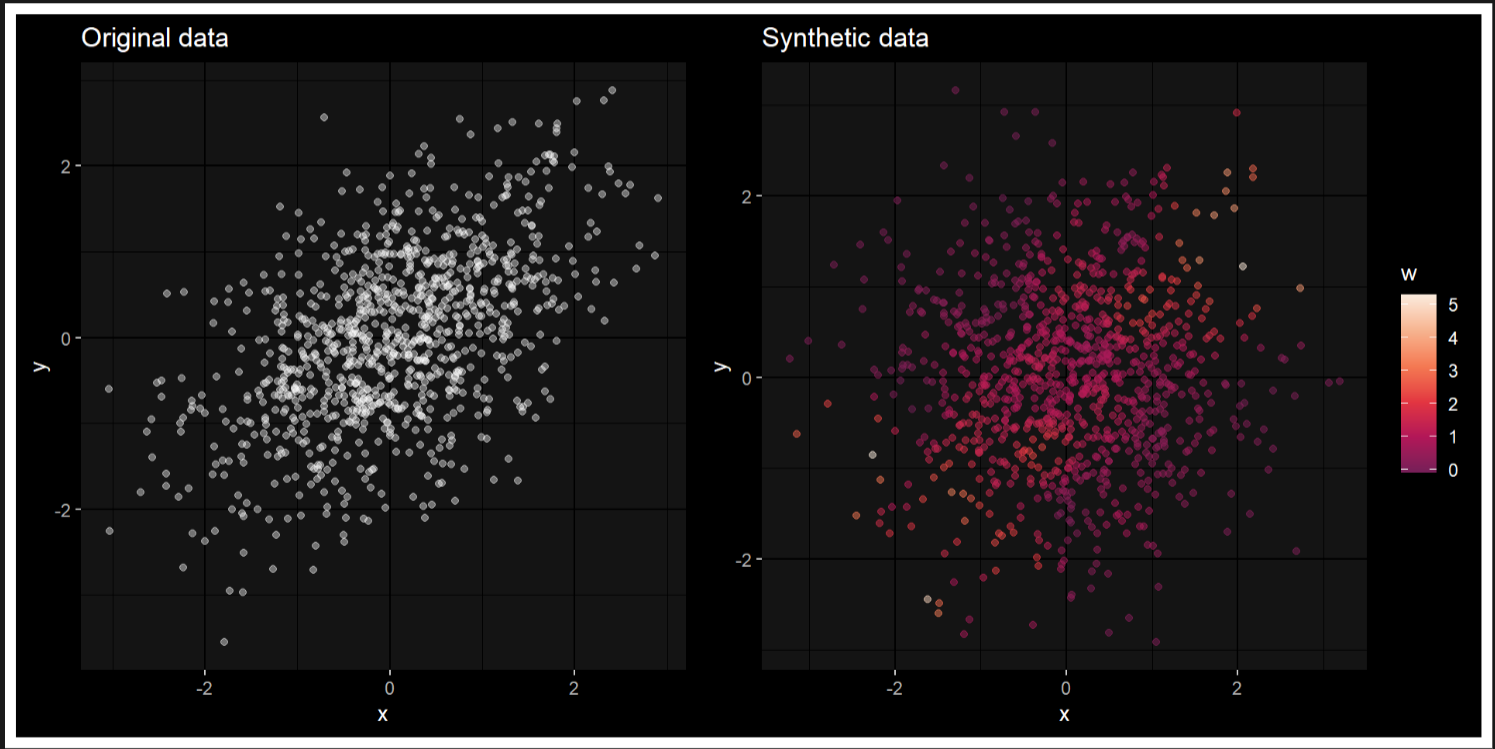
Utility of the synthetic data



Reweighting synthetic data: regression coefficients



Reweighting synthetic data: regression coefficients



Reweighting synthetic data: regression coefficients

	Observed	Synthetic	Reweighted
b0	-0.0179771	0.0203401	-0.0189604
b1	0.5152371	0.0385202	0.4779748

Other advantages of density ratios for utility

Use density ratios to discard synthetic outliers

High-dimensional extensions

- Find a $m < p$-dimensional subspace in which the synthetic and observed data are maximally different
- Estimate the density ratio in this subspace

Automatic cross-validation for hyperparameter selection

Thanks for your attention!

Even if it was simulated…

Questions?

t.b.volker@uu.nl