Assessing the utility of synthetic data: A density ratio perspective

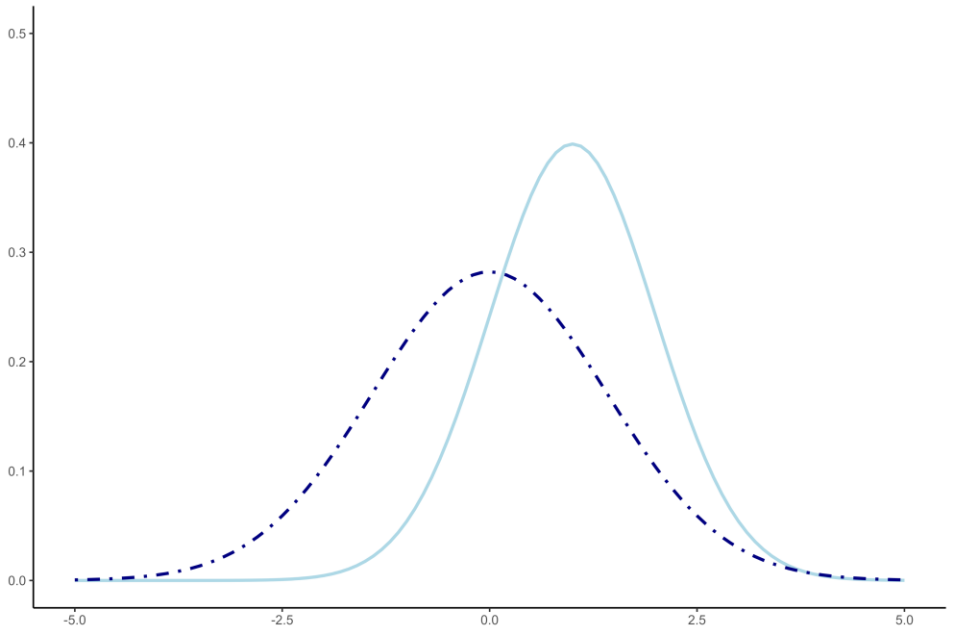
Thom Benjamin Volker, Peter-Paul de Wolf & Erik-Jan van Kesteren

UNECE Expert Meeting on Statistical
Data Confidentiality

Synthetic data utility

Intuitively

- How different are the synthetic data from the observed data?
- Is the synthetic data (almost) as useful as the observed data?



Practically

- Can we tell the observed and synthetic data apart?
- Can we obtain inferences from the synthetic data that are similar to inferences from the observed data?

Utility is hard to measure

The utility of synthetic data depends on what it's used for

How can we know what the synthetic data will be used for?

We can't...

We need good measures for general utility (distributional similarity)

- If the observed and synthetic data have indistinguishable distributions, they should allow for similar inferences.

Existing general utility measures

pMSE

- Practical, easy to use
- Not straightforward to specify the propensity score model
- Increasingly difficult to use when the dimensionality of the data increases relative to the sample size

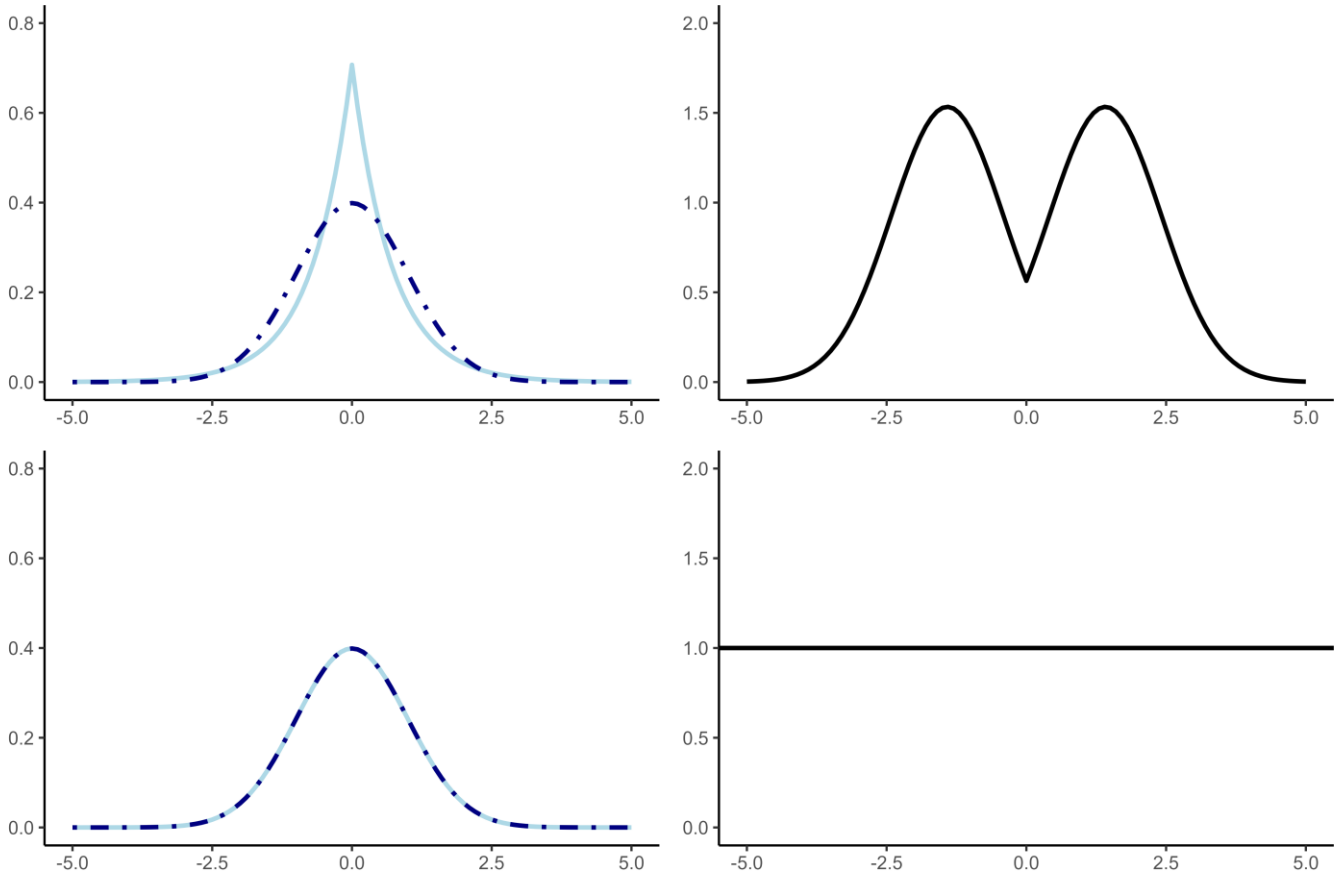
***Kullback-Leibler* divergence**

- Theoretically very elegant
- Hard to estimate in practice

Utility as a density ratio

$$r(x) = \frac{p_{syn}(x)}{p_{obs}(x)}$$

Evaluating utility: density ratios



Example: Density ratios for utility

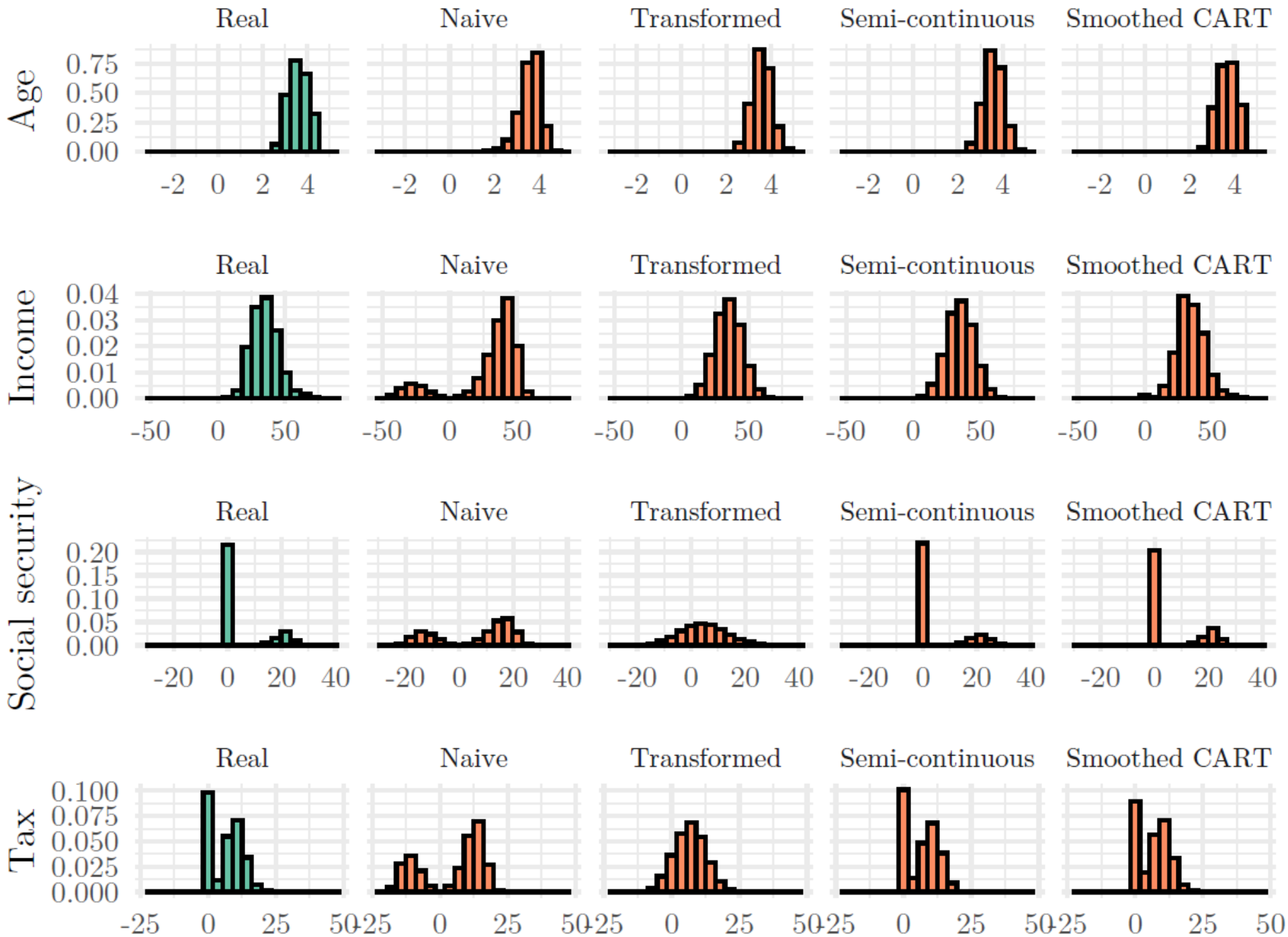
Data – U.S. Current Population Survey (*n* = 5000)*

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

Synthesis strategies

- Linear models; transformations; semi-continuous + transformations; smoothed CART
- Logistic / multinomial regression

* Our gratitude goes out to Jörg Drechsler for his willingness to share the data and synthesis script


Real
Naive
Transformed
Semi-continuous
Smoothed CART
Age
0.75
0.50
0.25
0.00
-2 0 2 4
Income
0.04
0.03
0.02
0.01
0.00
-50 0 50
Social security
0.20
0.15
0.10
0.05
0.00
-20 0 20 40
Tax
0.100
0.075
0.050
0.025
0.000
-25 0 25 50

Assessing utility

1. Estimate the density ratio using a kernel model
   - Unconstrained least-squares importance fitting (uLSIF; Kanamori et al., 2009)
2. Compute a discrepancy measure for each synthetic dataset
   - **Pearson divergence**: $\widehat{PE}\left(X_{syn}, X_{obs}\right) = \frac{1}{2n_{syn}} \sum_{i=1}^{n_{syn}} \hat{r}\left(x_{syn}^{(i)}\right) - \frac{1}{n_{obs}} \sum_{j=1}^{n_{obs}} \hat{r}\left(x_{obs}^{(j)}\right) + \frac{1}{2}$
3. Compare the test statistics between synthetic data sets

All implemented in the R-package **`densityratio`**

Assessing utility in R

```
library(densityratio)

# for every variable in every synthetic dataset, do:
dr <- ulsif(cps_synthetic$var, cps_real$var)
summary(dr_var)

# and for every entire synthetic dataset, do:
dr <- ulsif(cps_synthetic, cps_real)
summary(dr)
```


Pearson divergence
1.00
0.10
0.01
Age
Income
Social security
Tax
All
Variable
Synthesis method
Naïve
Transformed
Semi-continuous
Smoothed CART

The way forward...

Density ratio estimation provides:

- An intuitive framework for the evaluation of synthetic data;
- Cross-validation for automatic hyperparameter selection;
- Readily available extensions to high dimensional settings;
- Utility scores for individual data points

But requires research into

- How to deal with categorical data;
- How to make best use of side products (e.g., individual utility scores)