OUTLINE: Density ratio estimation as a technique for assessing the utility of synthetic data

Thom Benjamin Volker
Utrecht University, Utrecht, The Netherlands
E-mail: t.b.volker@uu.nl
Erik-Jan van Kesteren
Utrecht University, Utrecht, The Netherlands
E-mail: e.vankesteren@uu.nl

Summary. Abstract goes here

Keywords: keywords

1. Introduction

What is synthetic data?

What is high-utility synthetic data?

- Specific utility: more detailed, but often you do not have sufficient knowledge about the analyses that will be performed with the data.
- General utility: provide an intuition of the general quality of the synthetic data, and ideally cover the specific utility measures.

Current ways to assess the utility?

- pMSE - logistic, regression, CART models (Snoke, Raab, Nowok, Dibben & Slavkovic, 2018; General and specific utility measures for synthetic data AND Woo, Reiter, Oganian & Karr, 2009; Global measures of data utlity for microdata masked for disclosure limitation)
- Kullback-Leiber divergence (Karr, Kohnen, Oganian, Reiter & Sanil, 2006; A framework for evaluating the utility of data altered to protect confidentiality).

- According to multiple authors, both specific and general utility measures have important drawbacks (see Drechsler Utility PSD; cites others). Narrow measures potentially focus on analyses that are not relevant for the end user, and do not generalize to the analyses that are relevant. Global utility measures are generally too broad, and important deviations in the synthetic data might be missed. Moreover, the measures are typically hard to interpret.

- See Drechsler for a paragraph on fit for purpose measures, that lie between general and specific utility measures (i.e., plausibility checks such as non-negativity; goodness of fit measures as χ^2 for cross-tabulations; Kolmogorov-Smirnov).

- Drechsler also illustrates that the standardized $pMSE$ has substantial flaws, as the results are highly dependent on the model used to estimate the propensity scores, and unable to detect important differences in the utility for most of the model specifications. Hence, it is claimed that a thorough assessment of utility is required.

Our proposal

- density ratio methods

– Why?

- Implementation in R-package (SynUtility)

2. density estimation

Options: - Prediction models: logistic regression, SVM, CART? - Multivariate density estimation - density ratio?

Dimension reduction and visualization?

3. Methodology

TO DO

4. Simulations

TO DO

5. Real data example

TO DO

6. Results

TO DO

7. Discussion and conclusion

TO DO