UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Expert meeting on Statistical Data Confidentiality

26–28 September 2023, Wiesbaden

ASSESSING THE UTILITY OF SYNTHETIC DATA: A DENSITY RATIO PERSPECTIVE

Thom Benjamin Volker (Utrecht University, the Netherlands; Statistics Netherlands, the Netherlands)
Erik-Jan van Kesteren (Utrecht University, the Netherlands)
t.b.volker@uu.nl, e.vankesteren1@uu.nl

Abstract

High quality synthetic data can be a solution to overcome disclosure risks that arise when disseminating research data to the public. However, for inferential purposes, the usefulness of the synthetic data largely depends on whether the synthetic data model approximates the distribution of the observed data close enough. Often, multiple candidate models are considered and improvements are made iteratively. Evaluating the utility of the synthetic data after the model building steps is a crucial but complicated endeavor, and although many methods exist, their results may tell an incomplete story. We propose to evaluate the utility of synthetic data using density ratio estimation techniques, which allows to embed both general and specific utility measures into a common framework. Using techniques from the density ratio estimation field, we show how an interpretable global utility measure can be obtained from the ratio of the observed and synthetic data densities that can even be used to express the utility of individual (synthetic) data points. Applying these techniques on (an approximation to the) posterior distributions of parameters gives rise to analysis-specific utility measures. Using empirical examples, we show that framing utility from a density ratio perspective improves on existing global utility measures. Lastly, we implemented the procedure along with existing utility measures in an R-package.

Introduction

In recent years, the academic interest in synthetic data has exploded. Synthetic data is increasingly being used as a solution to overcome privacy and confidentiality issues that are inherently linked to the dissemination of research data. National statistical institutes and other government agencies have started to disseminate synthetic data to the public while restricting access to the original data to protect sensitive information (e.g., Abowd, Stinson, and Benedetto 2006; Hawala 2008; Drechsler 2012). At the same time, researchers start to share a synthetic version of their research data to comply with open science standards (e.g., Wiel et al. 2023; Obermeyer et al. 2019; Zettler et al. 2021). Rather than sharing the original research data, a synthetic surrogate is shared to facilitate reviewing the data processing and analysis pipeline. Additionally, synthetic data is increasingly being used for training machine learning models (Nikolenko 2021). On a lower level, synthetic data can be used in model testing pipelines (before access to the real data is provided), for data exploration (misschien YOUTH citeren all), and for educational purposes.

In short, the general idea of synthetic data is to substitute values from the collected data with synthetic values that are generated from a model. In this way, it is possible to generate an entirely new synthetic data set (commonly referred to as the *fully* synthetic data approach; Rubin 1993), but one could also replace only those values that would yield a high risk of disclosure when released (called *partially* synthetic data; Little 1993). Both approaches essentially attempt to build a model that incorporates as much of the information in the real data as possible. The models used to generate synthetic data were originally closely related to methods used for multiple imputation of missing data, such as fully conditional specification (Volker and Vink 2021) or sequential regression (Nowok, Raab, and Dibben 2016). Recently, significant improvements in generative modelling in combination with work on formal privacy guarantees sparked the development of a great deal of novel methods in the computer science community (e.g., Patki, Wedge, and Veeramachaneni 2016; Xu et al. 2019).

References

Abowd, John M., Martha Stinson, and Gary Benedetto. 2006. "Final Report to the Social Security Administration on the SIPP/SSA/IRS Public Use File Project." Longitudinal Employer-Household Dynamics Program, U.S. Bureau of the Census, Washington, DC. https://ecommons.cornell.edu/bitstream/handle/1813/43929/SSAfinal.pdf?sequence=3&isAllowed=y.

Drechsler, Jörg. 2012. "New Data Dissemination Approaches in Old Europe – Synthetic Datasets for a German Establishment Survey." *Journal of Applied Statistics* 39 (2): 243–65. https://doi.org/10.1080/02664763.2011.584523.

Hawala, Sam. 2008. *Producing Partially Synthetic Data to Avoid Disclosure*. http://www.asasrms.org/Proceedings/y2008/Files/301018.pdf.

Little, Roderick J. A. 1993. "Statistical Analysis of Masked Data." *Journal of Official Statistics* 9 (2): 407–7.

Nikolenko, Sergey I. 2021. *Synthetic Data for Deep Learning*. Springer International Publishing. https://doi.org/10.1007/978-3-030-75178-4.

Nowok, Beata, Gillian M. Raab, and Chris Dibben. 2016. "**Synthpop**: Bespoke Creation of Synthetic Data in *R*." *Journal of Statistical Software* 74 (11). https://doi.org/10.18637/jss.v074.i11.

Obermeyer, Ziad, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. "Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations." *Science* 366 (6464): 447–53. https://doi.org/10.1126/science.aax2342.

Patki, Neha, Roy Wedge, and Kalyan Veeramachaneni. 2016. "The Synthetic Data Vault." *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, October. https://doi.org/10.1109/dsaa.2016.49.

Rubin, Donald B. 1993. "Statistical Disclosure Limitation." *Journal of Official Statistics* 9 (2): 461–68.

Volker, Thom Benjamin, and Gerko Vink. 2021. "Anonymiced Shareable Data: Using Mice to Create and Analyze Multiply Imputed Synthetic Datasets." *Psych* 3 (4): 703–16. https://doi.org/10.3390/psych3040045.

Wiel, Mark A. van de, Gwenaël G. R. Leday, Jeroen Hoogland, Martijn W. Heymans, Erik W. van Zwet, and Ailko H. Zwinderman. 2023. "Think Before You Shrink: Alternatives to Default Shrinkage Methods Can Improve Prediction Accuracy, Calibration and Coverage." https://doi.org/10.48550/ARXIV.2301.09890.

Xu, Lei, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. 2019. "Modeling Tabular Data Using Conditional GAN." In *Advances in Neural Information Processing Systems*, edited by H. Wallach, H. Larochelle, A. Beygelzimer, F. dAlché-Buc, E. Fox, and R. Garnett. Vol. 32. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2019/file/254ed7d2de3b23ab10936522dd547b78-Paper.pdf.

Zettler, Ingo, Christoph Schild, Lau Lilleholt, Lara Kroencke, Till Utesch, Morten Moshagen, Robert Böhm, Mitja D. Back, and Katharina Geukes. 2021. "The Role of Personality in COVID-19-Related Perceptions, Evaluations, and Behaviors: Findings Across Five Samples, Nine Traits, and 17 Criteria." *Social Psychological and Personality Science* 13 (1): 299–310. https://doi.org/10.1177/19485506211001680.