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). Through these developments, the quality of synthetic data improved significantly, and while the notion of using fake data for research purposes was originally regarded as laughable, it is nowadays increasingly being used in practice.

The main challenge when generating synthetic data is to adequately balance the privacy leakage with the utility (i.e., quality) of the synthetic data. On the upper limit of this privacy-utility trade-off, the synthesis model is so good (or, rather, bad), that the real data is exactly reproduced, resulting in the same privacy loss as when disseminating the real data. In statistical terms, the synthesis model is overparameterized to such an extent that there are no degrees of freedom left, and there is thus no randomness involved in the generation of the synthetic values. On the lower limit of the trade-off, synthetic values are generated without borrowing any information from the real data. For example, we could place the value 0 or a random draw from a standard normal distribution for every record and every variable, such that the synthetic data contains only noise. A synthesis model usually sits somewhere between these two extremes, and contains some information from the real data, which implies that the synthetic data resembles the real data to some extent, yielding more than zero utility, but also some disclosure risk. The data provider typically wants to know what privacy loss is incurred by disseminating the synthetic data, while the user wants to know whether any analysis can be reliably performed. At the same time, knowledge about the utility can help the data provider to improve the quality of the synthesis model. Hence, the synthetic data provider is left with the complicated task of qualifying where on this continuum the synthetic data is located.

Based on this privacy-utility trade-off, it seems obvious that any attempt of data synthesis results in the loss of information. Accordingly, the utility of the synthetic data will always be lower than the utility of the real data. The questions that naturally arise are how much information is sacrificed, and how much the synthetic data deviates from the observed data. In the synthetic data literature, three classes of utility measures have been distinguished (for a thorough review of these measures, see Drechsler and Haensch 2023): fit-for-purpose measures, analysis-specific utility measures and global utility measures. Fit-for-purpose measures are typically the first step in assessing the quality of the synthetic data. They typically involve comparing the univariate distributions of the observed and synthetic data (for example using visualization techniques or goodness-of-fit measures). Although very useful to get an initial impression of the quality of the data synthesis models used, this picture is by definition limited, because only one or two variables variables are assessed at the same time. Hence, complex relationships between variables will always be out of scope. Global utility measures build up on the fit-for-purpose measures, by comparing the distribution of the synthetic data with the distribution of the observed data in a more general way. This can be done using some distance measure [e.g., the Kullback-Leibler divergence; see Karr et al. (2006)], but also by estimating how well a prediction model can distinguish between the observed and synthetic data, and using the predicted probabilities [propensity scores; Rosenbaum and Rubin (1983)] as a measure of discrepancy [e.g., the propensity score mean squared error, $pMSE$; Woo et al. (2009); Snoke et al. (2018)]. While global utility measures paint a rather complete picture, that is, over the entire range of the data, they tend to be too general. That is, global utility measures can be so

broad that important discrepancies between the real and synthetic data can be missed, and an a synthetic data set with high global utility might still yield analyses with results that are far from the results from real data analyses (see Drechsler 2022). Lastly, the analysis-specific utility measures quantify to what extent analyses performed on the synthetic data align with the same analyses on the observed data. These measures can evaluate to what degree the coefficients of a regression model are similar [e.g., using the confidence interval overlap; Karr et al. (2006)], but also to what extent prediction models trained on the synthetic and observed data are perform similarly in terms of evaluation metrics. However, also these measures have important shortcomings. Analysis-specific utility generally does not carry over, even not to closely related analyses: high specific utility for one analysis does not at all imply high utility for another analysis. Since data providers typically do not know which analyses will be performed with the synthetic data, it is impossible to provide analysis-specific utility measures for all potentially relevant analyses (for a more thorough discussion, see Drechsler 2022).

From the other perspective on the privacy-utility trade-off, that is, the privacy-side, several promising advances have been made with respect to building formal privacy guarantees into the data generation mechanism through differential privacy (CITE DWORK; CITE DP-SYNTHESIS METHOD). In addition to these privacy-by-design mechanisms, some measures exist to quantify privacy loss of synthetic data after generation. However, the practical applicability of these measures depends on whether the data is fully or partially synthetic, and especially in case of the former, the practical applicability of these measures is often limited (for an extensive discussion of these issues, see Drechsler and Haensch 2023). More research on measures to evaluate disclosure risks in synthetic is thus certainly needed, but the current paper focuses on utility measures for synthetic data.

Specifically, we illustrate a group of methods under the umbrella of density ratio estimation (for a thorough evaluation of work done is this area, see Sugiyama, Suzuki, and Kanamori 2012a) for assessing synthetic data utility that incorporates fit-for-purpose, global and analysis-specific utility measures. In short, the framework of density ratio estimation attempts to compare the multivariate distributions of two data sets (e.g., two different samples or groups) by directly estimation the ratio of their densities. The idea is that if two data sets are drawn from the same data-generating mechanism, their densities should be similar, and the ratio of their densities should thus be close to one at all possible points in the multivariate space. This approach easily extends from univariate to bivariate and multivariate densities. As such, it bridges the gap between fit-for-purpose and global utility measures. Moreover, a density ratio can also be estimated for the posterior distributions of parameters, and thus also incorporates specific utility measures. The only requirement of density ratio estimation is that there are samples (e.g., observations, samples from a posterior) from the observed and synthetic data (or parameter) distributions, or that these distributions can be approximated, for example with a (multivariate) normal distribution. It is not required to assume some parametric distribution, but in cases where no samples are available, it can help to simplify the process of estimation the density ratio. Note that density ratio estimation deliberately does not entail estimating the densities of the observed and synthetic data separately, and taking their ratio. Estimating the probability distribution of a data set is one of the hardest challenges in statistics, unavoidably resulting in estimation errors. When performing this task for two data sets, and subsequently taking the ratio, may magnify the estimation errors, resulting in a poorer estimate of the density ratio then necessary when directly estimating the density ratio.

In what follows, we intuitively describe the density ratio estimation framework, reviewing some of the work done in this area, while attempting to avoid technicalities where possible, while referring to the underlying mathematical foundations. We will briefly relate the method to existing utility measures, and describe how existing utility measures either fall under the umbrella of density ratio estimation, or are related to those. Subsequently, we will illustrate how the method can be used in practice by providing multiple examples. Lastly, we will discuss shortcomings of the method and relate these to avenues for future work.

Density ratio estimation

The framework of density ratio estimation was originally developed in the machine learning community for the comparison of two probability distributions (for an overview, see Sugiyama, Suzuki, and Kanamori 2012a). The framework has been shown to be applicable for prediction (Sugiyama et al. 2010; Sugiyama 2010), outlier detection (Hido et al. 2008), change-point detection in time-series (Liu et al. 2013), importance weighting under domain adaptation [or, in statistical terms, sample selection bias; Kanamori, Hido, and Sugiyama (2009)], and, importantly, two-sample homogeneity tests (Sugiyama et al. 2011). Regardless of the exact goal, the idea is to compare two distributions by estimating the density ratio $r(\boldsymbol{x})$ between the probability distributions of the numerator samples, which we take to be the synthetic data samples,

$p_{syn}(\boldsymbol{x}_{syn})$, and the denominator samples, which we take to be the observed data samples, $p_{obs}(\boldsymbol{x}_{obs})$. A naive approach would be to estimate the observed and synthetic data density separately, for example using kernel density estimation [REF], and subsequently compute the ratio from these estimated densities. However, density estimation is one of the hardest tasks in statistical learning, and is prone to estimation errors that would affect both estimated densities. When subsequently taking the ratio of the estimated densities, the estimation errors might be magnified, and the validity of the resulting density ratio might be lower than strictly necessary. In fact, extensive simulations on a wide variety of tasks showed that directly estimating the density ratio typically outperforms naive kernel density estimation, especially when the dimensionality of the data increases (e.g., Kanamori, Suzuki, and Sugiyama 2012; Hido et al. 2008; Kanamori, Hido, and Sugiyama 2009).

Over the past years, several methods for direct density ratio estimation have been developed, that typically employ some divergence measure to estimate the distance from the density ratio model to the true density ratio function. To keep the discussion general, we focus here on a class of methods that employ the *Bregman* divergence to quantify this distance (for a thorough review and technical discussion, see Sugiyama, Suzuki, and Kanamori 2012b). In some sense, this framework bears some resemblance to model selection using information criteria (e.g., Akaike's information criterion), where the distance from a statistical model to the true data-generating mechanism is estimated (e.g., by relying on an estimate of the Kullback-Leibler divergence). In fact, the Kullback-Leibler divergence is a special case of the more general class of Bregman divergences, and one of the density ratio estimation techniques (the Kullback-Leibler importance estimation procedure; Sugiyama et al. 2008) relies on the Kullback-Leibler divergence to estimate the density ratio function.

Illustration

```
# normal versus normal
#
# normal versus laplace
#
# normal versus lognormal
#
# normal versus t
```

```
# multivariate normal versus multivariate normal
#
# multivariate normal versus something else
```

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.

———. 2022. "Challenges in Measuring Utility for Fully Synthetic Data." In *Privacy in Statistical Databases*, edited by Josep Domingo-Ferrer and Maryline Laurent, 220–33. Cham: Springer International Publishing. https://doi.org/https://doi.org/10.1007/978-3-031-13945-1_16.

Drechsler, Jörg, and Anna-Carolina Haensch. 2023. "30 Years of Synthetic Data." https://doi.org/10.48550/ARXIV.2304.02107.

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

Hido, Shohei, Yuta Tsuboi, Hisashi Kashima, Masashi Sugiyama, and Takafumi Kanamori. 2008. "Inlier-Based Outlier Detection via Direct Density Ratio Estimation." In *2008 Eighth IEEE International Conference on Data Mining*,

edited by Fosca Giannotti, Dimitrios Gunopulos, Franco Turini, Carlo Zaniolo, Naren Ramakrishnan, and Xindong Wu, 223–32. https://doi.org/10.1109/ICDM.2008.49.

Kanamori, Takafumi, Shohei Hido, and Masashi Sugiyama. 2009. "A Least-Squares Approach to Direct Importance Estimation." *Journal of Machine Learning Research* 10 (48): 1391–1445. http://jmlr.org/papers/v10/kanamori09a.html.

Kanamori, Takafumi, Taiji Suzuki, and Masashi Sugiyama. 2012. "Statistical Analysis of Kernel-Based Least-Squares Density-Ratio Estimation." *Machine Learning* 86 (3): 335–67. https://doi.org/10.1007/s10994-011-5266-3.

Karr, Alan F., Christine N. Kohnen, Anna Oganian, Jerome P. Reiter, and Ashish P. Sanil. 2006. "A Framework for Evaluating the Utility of Data Altered to Protect Confidentiality." *The American Statistician* 60 (3): 224–32. https://doi.org/10.1198/000313006X124640.

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

Liu, Song, Makoto Yamada, Nigel Collier, and Masashi Sugiyama. 2013. "Change-Point Detection in Time-Series Data by Relative Density-Ratio Estimation." *Neural Networks* 43: 72–83. https://doi.org/https://doi.org/10.1016/j.neunet.2013.01.012.

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.

Rosenbaum, Paul R., and Donald B. Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70 (1): 41–55. https://doi.org/10.1093/biomet/70.1.41.

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

Snoke, Joshua, Gillian M. Raab, Beata Nowok, Chris Dibben, and Aleksandra Slavkovic. 2018. "General and Specific Utility Measures for Synthetic Data." *Journal of the Royal Statistical Society. Series A (Statistics in Society)* 181 (3): pp. 663–688. https://doi.org/https://doi.org/10.1111/rssa.12358.

Sugiyama, Masashi. 2010. "Superfast-Trainable Multi-Class Probabilistic Classifier by Least-Squares Posterior Fitting." *IEICE Transactions on Information and Systems* E93-D (10). https://doi.org/10.1587/transinf.E93.D.2690.

Sugiyama, Masashi, Taiji Suzuki, Yuta Itoh, Takafumi Kanamori, and Manabu Kimura. 2011. "Least-Squares Two-Sample Test." *Neural Networks* 24 (7): 735–51. https://doi.org/https://doi.org/10.1016/j.neunet.2011.04.003.

Sugiyama, Masashi, Taiji Suzuki, and Takafumi Kanamori. 2012a. *Density Ratio Estimation in Machine Learning*. Cambridge University Press. https://doi.org/10.1017/CBO9781139035613.

———. 2012b. "Density-Ratio Matching Under the Bregman Divergence: A Unified Framework of Density-Ratio Estimation." *Annals of the Institute of Statistical Mathematics* 64 (5): 1009–44. https://doi.org/10.1007/s10463-011-0343-8.

Sugiyama, Masashi, Taiji Suzuki, Shinichi Nakajima, Hisashi Kashima, Paul von Bünau, and Motoaki Kawanabe. 2008. "Direct Importance Estimation for Covariate Shift Adaptation." *Annals of the Institute of Statistical Mathematics* 60 (4): 699–746. https://doi.org/10.1007/s10463-008-0197-x.

Sugiyama, Masashi, Ichiro Takeuchi, Taiji Suzuki, Takafumi Kanamori, Hirotaka Hachiya, and Daisuke Okanohara. 2010. "Conditional Density Estimation via Least-Squares Density Ratio Estimation." In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, edited by Yee Whye Teh and Mike Titterington, 9:781–88. Proceedings of Machine Learning Research. Chia Laguna Resort, Sardinia, Italy: PMLR. https://proceedings.mlr.press/v9/sugiyama10a.html.

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.

Woo, Mi-Ja, Jerome P. Reiter, Anna Oganian, and Alan F. Karr. 2009. "Global Measures of Data Utility for Microdata Masked for Disclosure Limitation." *Journal of Privacy and Confidentiality* 1 (1). https://doi.org/10.29012/jpc.v1i1.568.

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.