A DENSITY RATIO FRAMEWORK FOR EVALUATING THE UTILITY OF SYNTHETIC DATA

A PREPRINT

Thom Benjamin Volker
Methodology and Statistics | Methodology
Utrecht University | Statistics Netherlands
Utrecht, 3584CH
t.b.volker@uu.nl

Peter-Paul de Wolf
Methodology
Statistics Netherlands
The Hague, 2490HA
pp.dewolf@cbs.nl

Erik-Jan van Kesteren
Methodology and Statistics
Utrecht University
Utrecht, 3584CH
e.vankesteren1@uu.nl

April 26, 2024

ABSTRACT

TODO TODO

Keywords synthetic data, utility, density ratio, privacy, disclosure limitation • synthetic data, utility, density ratio, privacy, disclosure limitation

1 Introduction

Openly accessible research data accelerates scientific progress tremendously. Open data allows third-party researchers to answer research questions with already collected data, freeing up resources that would otherwise be devoted to

data collection (Ramachandran, Bugbee, and Murphy 2021). Sharing data in combination with code allows others to validate research findings and build upon the work (Obels et al. 2020; Crosas et al. 2015). Students can benefit from open data, as it fosters education with realistic data (Atenas, Havemann, and Priego 2015), as well as the general public, through stimulating citizen science projects (Newman et al. 2012). However, making data openly available is often (rightfully) hampered by official legislation, like the General Data Protection Regulation (GDPR; European Parliament and Council of the European Union 2016), and general privacy concerns. In the worst case, sharing data may cause harm to individuals or organizations, which may withhold these entities from participating in future research. These privacy constraints have been named among the biggest hurdles in the advancement of computational social science (Lazer et al. 2009), and among top reasons for companies to not share their data with researchers (Future of Privacy Forum 2017).

Multiple approaches exist to balance the benefits of open data with potential privacy risks. Traditionally, data providers employed a suite of different disclosure limitation techniques before sharing the data, such as top-coding, record-swapping or adding noise (e.g., Hundepool et al. 2012; Willenborg and De Waal 2001). More recently, synthetic data has gained traction as a means to disseminate private data (Abowd, Stinson, and Benedetto 2006; Hawala 2008; Drechsler 2012; van de Wiel et al. 2023; Obermeyer et al. 2019; Zettler et al. 2021), although the conceptual framework traces back to the previous century (Little 1993; Rubin 1993). Simply put, the idea of synthetic data is to replace some, or all, of the observed values in a data set by synthetic values that are generated from some model (e.g., Drechsler 2011). If only some values are replaced, disclosure risks can be reduced because the sensitive or identifying values do not correspond to their true values anymore. If all values are replaced, there is also no one-to-one mapping between the original and the synthetic data, further reducing the disclosure risk. However, an increase in privacy typically comes at the cost of a decrease in utility. As more of the data is altered, the quality of the released data becomes more sensitive to the suitability of the generative model. Regardless of the approach to disclosure limitation, if the technique used to generate or alter the data does not align with the intricacies of the problem at hand, the utility of the released data will be further reduced than necessary.

Given that all disclosure limitation techniques reduce the utility of the data, the challenge that arises is how to determine whether the released data still has acceptable utility. Alternatively, one might consider different disclosure limitation techniques that all satisfy the defined privacy restrictions, and employ the one which yields data with the highest utility. That is, given a set of methods that all meet the privacy restrictions, one may aim to maximize the utility of the released data. Both strategies require a reliable and encompassing measure of data utility that allows to evaluate the quality of the released data, and that allows to compare different disclosure limitation techniques and/or synthesis models in terms of utility. Moreover, adequate utility measures often guide the synthesis process, by providing detailed feedback on important discrepancies between the original and synthetic data. Lastly, good utility measures help the data user in determining what the synthetic data can and cannot be used for.

In the synthetic data field, three classes of utility measures have been distinguished (see Drechsler and Haensch 2023 for a thorough review): fit-for-purpose measures, analysis-specific utility measures and global utility measures. Fit-

for-purpose measures are often the first step in assessing the quality of the synthetic data. These typically involve comparing the univariate distributions of the observed and synthetic data (for example using visualization techniques or goodness-of-fit measures). Although these measures provide an initial impression of the quality of the synthesis models used, this picture is by definition limited, because only one or two variables are assessed at the same time. Hence, complex relationships between variables will always be out of scope. Such relationships may be captured by analysis-specific utility measures, which quantify whether analyses on synthetic data provide results that are comparable to results from the same analysis performed on the observed data. These measures can, for example, evaluate how similar the coefficients of a regression model are (e.g., using the confidence interval overlap; Karr et al. 2006), or whether prediction models trained on the synthetic and observed data perform comparably in terms of evaluation metrics. However, analysis-specific utility generally does not carry over: 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 (see also Drechsler 2022).

Global utility measures may overcome the shortcomings of the previous approaches, as they evaluate the discrepancy between the entire multivariate distribution of the observed and synthetic data. As such, global utility measures yields the most promising class of utility measures, because if the observed and synthetic data have similar (multivariate) distributions, all potential analyses should yield similar results. Global utility can be evaluated using some divergence measure, such as the Kullback-Leibler divergence (Karr et al. 2006), or by evaluating whether the observed and synthetic data are distinguishable using a classification model (a technique called $pMSE$; Woo et al. 2009; Snoke et al. 2018). However, a common critique of global utility measures is that they tend to be too general (Drechsler 2022). That is, analyses on a synthetic data set that is overall quite similar to the observed data (i.e., has high global utility), may still yield results that are far from the results obtained from the real data. Also, commonly used methods for estimating the $pMSE$, as logistic regression and classification and regression trees, tend to become less reliable as the dimensionality of the data increases, and are vulnerable to model misspecification (Drechsler 2022). Lastly, the output of global utility measures can be hard to interpret, and say little about the regions in which the synthetic data do not resemble the true data accurately enough.

To overcome the issues related to traditional global utility measures, we propose to use the density ratio estimation framework (Sugiyama, Suzuki, and Kanamori 2012a) as a way of evaluating utility. Intuitively, if two data sets have similar multivariate distributions, the density ratio should be close to one over the range of the data. If the distributions of the observed and synthetic data are very different, the density ratio should be far from one at those regions where the distributions differ. As density estimation is known to be a difficult problem, the density ratio estimation framework provides techniques to directly estimate the density ratio, rather than the two separate densities, in a non-parametric way (Sugiyama, Suzuki, and Kanamori 2012a). These non-parametric estimation techniques come with automatic model specification, which mitigates the issue of model specification. This functionality is implemented in the `R`-package `densityratio` (Volker 2023). Importantly, the density ratio is estimated over the entire range of the data,

which provides measures of utility at every (possible) point in the data space. This point-specific quantification of utility turns out to be a useful side-product, as it allows to reweigh analyses on synthetic data when further improving the utility directly is not possible.

In the remainder of the article, we introduce the density ratio framework and the associated estimation techniques, and connect the framework to traditional utility measures as the $pMSE$ and the Kullback-Leibler divergence. We then present simulations to demonstrate the performance of density ratio estimation in stylized settings and compare it to traditional utility measures. Subsequently, we apply density ratio estimation in a case study where we evaluate the utility of multiple synthetic versions of the U.S. Current Population Survey. We conclude with a discussion of the results, highlight the strengths and weaknesses of the density ratio framework, and provide recommendations for future research.

2 Background

Over the years, many methods have been introduced to generate synthetic data, all with the aim of providing a suitable balance between privacy and utility. These methods can be relatively simple, such as a sequence of generalized linear models (e.g., Reiter 2004), or as complex as deep learning models with thousands of parameters (e.g., Xu et al. 2019), with many options in between. However, the complexity of the generation process is not necessarily a good indicator of the quality of the synthetic data. That is, relatively simple methods could still capture the most important aspects of the data that complex methods fail to capture (and vice versa). Hence, data providers typically do not know which synthesis method will provide the highest utility a priori, and might compare multiple synthesis strategies to determine which data set will be released. Good global utility measures can help in this process, by allowing to quantify the quality of the candidate synthetic data sets.[1] Moreover, such global utility measures may guide the synthesis process itself, if they provide sufficiently specific information about the degree of misfit of the synthetic data. In the upcoming section, we provide an overview of existing global utility measures, and introduce the density ratio framework as an encompassing approach to evaluating global utility.

2.1 Existing global utility measures

Global utility measures typically attempt to quantify the distributional similarity between the observed and synthetic data samples. The intuition is that if two data sets have similar distributions, the data sets can be used for the same purposes, and analyses on the two data sets should yield similar results. A common way to formalize distributional similarity is through the Kullback-Leibler (KL) divergence, as proposed in Karr et al. (2006). The KL-divergence measures the relative entropy from the probability distribution of the observed data $p_{\text{obs}}(\mathbf{x})$ to the probability distribution of the synthetic data $p_{\text{syn}}(\mathbf{x})$ (with $\mathbf{x} \in \mathbb{R}^{n \times d}$), and is defined as

$$D_{\text{KL}}(p_{\text{syn}}||p_{\text{obs}}) = \int p_{\text{syn}}(\mathbf{x}) \log \frac{p_{\text{syn}}(\mathbf{x})}{p_{\text{obs}}(\mathbf{x})} \mathrm{d}\mathbf{x}. \tag{1}$$

[1] We focus on global utility measures, because in many situations the data provider does not know which analysis will be performed with the synthetic data. If the data provider knows for which purposes the data will be used, analysis-specific utility measures may be more informative.

Karr et al. (2006) argue that the KL-divergence can be constructed from density estimators $\hat{p}_{\text{obs}}$ and $\hat{p}_{\text{syn}}$, after which the integral can be approximated using numerical quadrature. Alternatively, if the observed and synthetic data are both (multivariate) normally distributed, the KL-divergence can be computed analytically. Both implementations might be unfeasible: the combination of density estimation with numerical quadrature is cumbersome in high dimensional settings, and assuming multivariate normality might be too restrictive. Yet, we show in later sections that the density ratio estimation framework gives rives to an alternative way of estimating the KL-divergence that overcomes these challenges.

An alternative way to evaluate whether two distributions are statistically indistinguishable is by evaluating whether a classification model can tell samples from the two distributions apart (see Kim et al. 2021, who formalize the connection between classification accuracy and two-sample testing). In the context of synthetic data, this implies that if a classification model can distinguish observed from synthetic samples, the distributional similarity is low, and so is the global utility. If a classifier cannot distinguish between the observed and synthetic data, one would conclude that the global utility is high. The propensity score mean-squared error ($pMSE$), introduced by Woo et al. (2009) and further developed in Snoke et al. (2018), formalizes this intuition. Let I_i denote an indicator variable that equals 1 if observation i ($i \in 1, \dots, N$, $N = n_{\text{obs}} + n_{\text{syn}}$) belongs to the synthetic data, and 0 otherwise. We then train a classifier that outputs the predicted probability of observation i being a synthetic record $\hat{s}_i$ based on the observation's scores on the variables (this can be the set of all variables, but also a subset). From these, we can calculate the utility statistic

$$pMSE = \frac{1}{N} \sum_{i=1}^{N} \left(\hat{s}_i - \frac{n_{\text{syn}}}{N} \right)^2, \tag{2}$$

which ought to be smaller when the synthetic data is more like the observed data. Crucially, the $pMSE$ depends on the classification model used and increases in the flexibility of the classification model, making it prone to overfitting and hard to interpret. To combat these issues, Snoke et al. (2018) suggests to compare the $pMSE$-value with its expectation under the null distribution. Provided that the classification model is a logistic regression model with k parameters (including the intercept), Snoke et al. (2018) show that the expected $pMSE$ is given by

$$\mathbb{E}[pMSE] = \left(\frac{k-1}{N} \right) \left(\frac{n_{\text{obs}}}{N} \right)^2 \left(\frac{n_{\text{syn}}}{N} \right).$$

For other classification models, the expectation can be approximated through a resampling procedure. Accordingly, the $pMSE$-ratio is given by

$$pMSE\text{-ratio} = \frac{pMSE}{\mathbb{E}[pMSE]},$$

with values smaller than 10 deemed acceptable (Gillian M. Raab, Nowok, and Dibben 2021; although the authors remark that the smaller the better). Apart from the $pMSE$, several other measures exist that can be constructed from the estimated propensity scores, such as the percentage of records correctly predicted (Gillian M. Raab, Nowok, and Dibben 2021) or the Kolmogorov-Smirnov statistic (Bowen, Liu, and Su 2021), both which are strongly correlated with the $pMSE$ (Gillian M. Raab, Nowok, and Dibben 2021).

Due to its intuitive nature, multiple studies advice the use of the $pMSE$ as a promising technique to evaluate the quality of synthetic data (e.g., Gillian M. Raab, Nowok, and Dibben 2017; Gillian M. Raab, Nowok, and Dibben 2021; Hu and Bowen 2024). Yet, it is not free of criticism. The usefulness of the $pMSE$ hinges on choosing a model that can capture the important intricacies of the observed data. Drechsler (2022) illustrated that the utility score is highly dependent on the model used to estimate the propensity scores, and that clear improvements in synthesis models are not necessarily picked up in the $pMSE$. Moreover, the $pMSE$ is prone to overfitting and users may find it difficult to select an appropriate model for estimating the propensity scores. Hence, current methods to evaluate global utility are either difficult to estimate, especially in high-dimensional situations, or depend strongly on model specification. Density ratio estimation might alleviate these issues. In what follows, we explain how the density ratio can be estimated, and how it can be used to evaluate the utility of synthetic data.

2.2 Density ratio estimation

The density ratio estimation framework 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 to 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 (i.e., sample selection bias; Kanamori, Hido, and Sugiyama 2009), and two-sample homogeneity tests (Sugiyama et al. 2011). The general idea of density ratio estimation is depicted in Figure 1, and boils down to comparing two distributions by modelling the density ratio $r(\mathbf{x})$ between the probability distributions of the numerator samples, taken from the synthetic data distribution, $p_{\text{syn}}(\mathbf{x})$, and the denominator samples, taken from the observed data distribution, $p_{\text{obs}}(\mathbf{x})$, such that

$$r(\mathbf{x}) = \frac{p_{\text{syn}}(\mathbf{x})}{p_{\text{obs}}(\mathbf{x})}. \tag{3}$$

This specification has the intuitive interpretation that in locations where the density ratio takes large values, too many synthetic observations will be generated in that region and at the locations where the density ratio is small, there will be too few synthetic observations, both relative to the observed data. An intuitive approach to estimating $r(\mathbf{x})$ from samples of $p_{\text{obs}}(\mathbf{x})$ and $p_{\text{syn}}(\mathbf{x})$ would be to estimate the observed and synthetic data density separately, for example using kernel density estimation (e.g., see Scott 1992 for an overview), and subsequently compute the ratio of these estimated densities. However, density estimation is one of the hardest tasks in statistical learning, unavoidably leading to estimation errors for both densities, especially in high dimensions. When subsequently taking the ratio of the estimated densities, estimation errors tend to be magnified. Direct density ratio estimation avoids this issue by specifying and estimating a model directly for the ratio without first estimating the separate densities. Extensive simulations on a wide variety of tasks showed that this approach typically outperforms density ratio estimation through naive kernel density estimation, especially when the dimensionality of the data increases (e.g., Kanamori, Suzuki, and Sugiyama 2012b; Hido et al. 2008; Kanamori, Hido, and Sugiyama 2009).

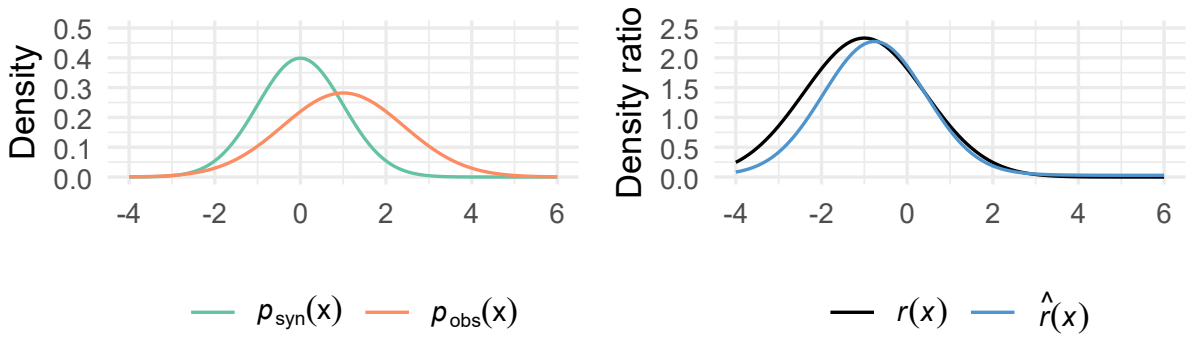

Figure 1: Example of the true and estimated density ratio of two normal distributions with different means and variances (i.e., $p_{\text{syn}}(\mathbf{x}) = N(0, 1)$ and $p_{\text{obs}}(\mathbf{x}) = N(1, 2)$). The function $r(\mathbf{x}) = p_{\text{syn}}(\mathbf{x})/p_{\text{obs}}(\mathbf{x})$ denotes the true density ratio, the function $\hat{r}(\mathbf{x})$ denotes an estimate of the density ratio based on $n_{\text{syn}} = n_{\text{obs}} = 200$ samples from each distribution obtained with unconstrained Least-Squares Importance Fitting (uLSIF). Note that the density ratio is itself not a proper density.

2.2.1 Estimating the density ratio

Over the past years, several methods for direct density ratio estimation have been developed. A large class of these methods attempt to directly minimize the error between the true density ratio $r(\mathbf{x})$ and its estimate $\hat{r}(\mathbf{x})$. Following this approach, we define a loss function $\mathcal{L}(r(\mathbf{x}), \hat{r}(\mathbf{x}))$ that measures the discrepancy between the true and estimated density ratio. To give an example, consider the following loss based on the squared error

$$\begin{aligned}\mathcal{L}_S(r(\mathbf{x}), \hat{r}(\mathbf{x})) &= \frac{1}{2}\int (\hat{r}(\mathbf{x}) - r(\mathbf{x}))^2 p_{\text{obs}}(\mathbf{x})\mathrm{d}\mathbf{x} \\ &= \frac{1}{2}\int \hat{r}(\mathbf{x})^2 p_{\text{obs}}(\mathbf{x})\mathrm{d}\mathbf{x} - \int \hat{r}(\mathbf{x}) r(\mathbf{x}) p_{\text{obs}}(\mathbf{x})\mathrm{d}\mathbf{x} + \frac{1}{2}\int r(\mathbf{x})^2 p_{\text{obs}}(\mathbf{x})\mathrm{d}\mathbf{x}.\end{aligned} \tag{4}$$

The second term can be rewritten, because the denominator in $r(\mathbf{x})$ cancels with $p_{\text{obs}}(\mathbf{x})$, while the third term is a constant with respect to the parameters in the density ratio model and can thus be ignored. Hence, we are left with the following loss function to minimize

$$\mathcal{L}'_S(r(\mathbf{x}), \hat{r}(\mathbf{x})) = \frac{1}{2}\int \hat{r}(\mathbf{x})^2 p_{\text{obs}}(\mathbf{x})\mathrm{d}\mathbf{x} - \int \hat{r}(\mathbf{x}) p_{\text{syn}}(\mathbf{x})\mathrm{d}\mathbf{x}, \tag{5}$$

which is termed least squares importance fitting (LSIF) by Kanamori, Hido, and Sugiyama (2009).

This approach can be straightforwardly generalized to a wide class of loss functions that fall under the family of Bregman divergences (see Sugiyama, Suzuki, and Kanamori 2012b; Mohamed and Lakshminarayanan 2017). Then, a general class of losses is encompassed by the expression

$$\mathcal{L}_f(r(\mathbf{x}), \hat{r}(\mathbf{x})) = \int \Big(f(r(\mathbf{x})) - f(\hat{r}(\mathbf{x})) - f'(\hat{r}(\mathbf{x}))(r(\mathbf{x}) - \hat{r}(\mathbf{x})) \Big) p_{\text{obs}}(\mathbf{x})\mathrm{d}\mathbf{x}, \tag{6}$$

where f is a differentiable and strictly convex function with derivative f'. Then, ignoring the terms independent of $\hat{r}(\mathbf{x})$ and noting that $r(\mathbf{x})p_{\text{obs}}(\mathbf{x}) = p_{\text{syn}}(\mathbf{x})$, we obtain the following objective

$$\mathcal{L}'_f(r(\mathbf{x}), \hat{r}(\mathbf{x})) = \int \Big(f'(\hat{r}(\mathbf{x}))\hat{r}(\mathbf{x}) - f(\hat{r}(\mathbf{x})) \Big) p_{\text{obs}}(\mathbf{x})\mathrm{d}\mathbf{x} - \int f'(\hat{r}(\mathbf{x}))p_{\text{syn}}(\mathbf{x})\mathrm{d}\mathbf{x}. \tag{7}$$

Minimizing this loss for different functions f yields different estimators for the density ratio, that focus on different regions of the density ratio.[2] Specifically, some estimators place more emphasis on accurately modelling the regions in which the true density ratio is large, but less emphasis on accurately estimating the smaller density ratios, and vice versa (see Sugiyama, Suzuki, and Kanamori 2012b; Menon and Ong 2016). Importantly, Sugiyama, Suzuki, and Kanamori (2012b) show that the Bregman divergence minimization approach to density ratio estimation is equivalent to estimating the f-divergence (Ali and Silvey 1966) between the synthetic and observed data distributions. Hence, when estimating the density ratio, one implicitly estimates the divergence between the synthetic and observed data distributions.

After defining a loss function, we need a model for the density ratio function $\hat{r}(\mathbf{x})$. There are many possibilities to specify this model, but a common choice is to use a linear model for the density ratio (e.g., Huang et al. 2006; Kanamori, Hido, and Sugiyama 2009; Izbicki, Lee, and Schafer 2014; Gruber et al. 2024). That is, we define the density ratio model as

$$\hat{r}(\mathbf{x}) = \varphi(\mathbf{x})\theta, \tag{8}$$

where $\varphi(\mathbf{x})$ is a basis function vector, that transforms the data from a p-dimensional to a b-dimensional space, and θ is a b-dimensional parameter vector. Although the model is linear in its parameters, the density ratio itself is typically a non-linear function of the data due to the basis functions. The parameter vector θ is estimated to minimize the discrepancy with the true density ratio using the loss function $\mathcal{L}(r(\mathbf{x}), \hat{r}(\mathbf{x}))$ in Equation 7.

Also for the basis functions, several specifications are possible, ranging from an identity function (Qin 1998) to normalizing flows (Choi, Liao, and Ermon 2021) or neural networks (Tiao 2018; Uehara et al. 2016). However, a commonly used basis function in the density ratio literature is the Gaussian kernel function (e.g., Huang et al. 2006; Sugiyama et al. 2007; Kanamori, Hido, and Sugiyama 2009; Liu et al. 2013; Gruber et al. 2024), which conveniently allows to model the non-linear density ratio using models that are linear in their parameters. The Gaussian kernel function is defined as

$$\varphi(\mathbf{x}) = \mathcal{K}(\mathbf{x}_i, c_j) = \exp\left(-\frac{\|\mathbf{x} - c_j\|^2}{2\sigma^2} \right),$$

where c_j ($j \in 1, \ldots, J$) denotes the centers of the Gaussian kernel functions, and σ controls the kernel width (i.e., it defines over which distance differences between the observations and the centers are considered relevant; for more information on kernel functions, see, e.g., Murphy 2022). The appropriate width of the kernel can be determined using cross-validation. The centers are typically sampled from the data (in our case, the observed and synthetic data can both be used, but it is also possible to solely use the synthetic samples).

[2] It is easy to see that using $f(x) = \frac{1}{2}(x-1)^2$ turns the Bregman divergence (Equation 7) into the squared error (Equation 5).

After defining the model and the loss function, the density ratio can be easily estimated. The `densityratio` package in R (Volker 2023) provides an implementation of commonly used loss functions with a Gaussian kernel basis function. The package comes with an easy-to-use interface, automatic cross-validation of hyperparameters and builds on `C++` for fast computation. However, for increased flexibility with respect to loss functions and basis functions, one could use classic optimization techniques as gradient descent or (quasi-)Newton methods, for example as implemented in the `optim` function in the statistical software `R` (R Core Team 2023). Both approaches are briefly illustrated in Appendix A (TODO).

2.2.2 Evaluating data utility with the density ratio

After estimating the density ratio, the information provided by the density ratio can be summarized in a divergence measure, using the fact that estimating the density ratio is equivalent to estimating the f-divergence between the synthetic and observed data distributions (Sugiyama, Suzuki, and Kanamori 2012b). Accordingly, we can use the estimated density ratio function to estimate an f-divergence between the observed and synthetic data, and use this divergence measure as a utility measure for the synthetic data. Although this divergence statistic is difficult to interpret in an absolute sense, it can be used as a relative measure of quality of the synthetic data. That is, for different synthetic data sets, we can calculate the divergence to the observed data, and compare these values to determine which synthetic data set is most similar to the observed data. In a more formal way, the f-divergence gives rise to a test statistic that can be used to test the null hypothesis that the synthetic data is generated from the same distribution as the observed data. The corresponding p-value can be calculated by comparing the observed test statistic to the null distribution of test statistics obtained using random permutations of the observed and synthetic data (see, for example, Sugiyama et al. 2011; Wornowizki and Fried 2016; Kanamori, Suzuki, and Sugiyama 2012a). Lastly, the density ratio function itself can be plotted against individual variables or pairs of variables to discover whether particular variables are particularly poorly captured in the synthetic data. Each of these strategies can be readily executed using the `densityratio` package in `R` (Volker 2023), and will be further illustrated in the upcoming sections.

Ik twijfel nog een beetje over de volgende zaken, en voordat ik alles ga zitten uitwerken hoor ik graag eerst jullie mening:

1. Ik heb nu het deel over regularisatie eruit gelaten, omdat ik vond dat het wat afleid van de main ideas, maar kan deze ook gewoon toevoegen aan Equation 7. Het kan ook casual opgemerkt worden wanneer het daadwerkelijk gebruikt wordt. Wat vinden jullie?

2. Dimension reduction: ik was niet zo zeker hoe ik dit er direct moest bouwen, want effectief is het niet veel anders dan de huidige aanpak, alleen met een andere basis function (namelijk $\phi(x) = \mathcal{K}(x_i^T U, c_j^T U)$), waarbij U ook parameters heeft die geschat worden. Anyway, ik kan aan het eind een sectie maken met dat DRE nog steeds lastig kan zijn in hoge dimensies, en dat daar oplossingen voor zijn (bijvoorbeeld een LFDA-stap voor ulsif, of LHSS, of the spectral density ratio estimation methode (laatste twee zitten allebei in het package)). Het punt is vooral dat ik niet wil dat het teveel afleidt van het main punt, namelijk density ratio

estimation, maar qua content zou het misschien een eigen sectie verdienen, ook omdat het niet direct enorm aansluit bij de twee secties hierboven.

3. Denken jullie dat dit een goede plek is om wat dieper in te gaan op mogelijke voordelen van density ratio estimation? Ik dacht dat na sectie 2.2.2 een voor de hand liggende plek zou zijn om hier wat dieper op in te gaan, en dan specifiek de volgende punten te benoemen: utility quantifications at every point in multivariate space, easy and flexible model specification through automatic model selection with cross-validation, potential for importance weighting, wellicht dan ook direct het punt van dimension reduction hierbij pakken.

4. Ik wil graag nog ergens iets zeggen over de relaties tussen density ratio estimation en de andere methodes (KL-divergence, pMSE), maar ik vraag me een beetje af of dat niet beter in de discussie kan.

5. Is dit echt de goede plaats om iets te zeggen over categorische variabelen? Deze kunnen namelijk gewoon geïncludeerd worden, maar met een Gaussian kernel wordt dat misschien een beetje awkward. Desalniettemin kan het prima, en kunnen we er ook voor kiezen om in de discussie te benoemen dat we een oplossing hebben gekozen, maar dat er ook andere opties zijn (bijv. andere kernels).

3 Numerical illustrations: Simulation studies

To evaluate the performance of density ratio estimation and compare it to existing methods, we conduct a series of simulation studies. We consider a variety of scenarios, including univariate and multivariate models, varying correlations, number of variables, and sample sizes. To perform density ratio estimation, we use unconstrained Least-Squares Importance Fitting (uLSIF; Kanamori, Hido, and Sugiyama 2009) as implemented in the `densityratio` package in `R` (Volker 2023), because it fast and has been shown to perform well in a variety of settings [ADD CITATIONS]. We compare the performance of uLSIF to the performance of the p-MSE as implemented in the `R`-package `synthpop`.

3.1 Univariate simulations

For the sake of illustrational clarity, we first evaluate the performance of density ratio estimation in a univariate setting where its performance can be evaluated using visualizations. In the simulations, we generate data according to four data-generating mechanisms, (1) a Laplace distribution, (2) a log-normal distribution, (3) a location-scale t-distribution and (4) a normal distribution. Subsequently, we approximate the true data generating mechanism using a Gaussian model with the same mean and variance as the original data. This setting is similar to situations that are commonly encountered in the synthetic data field, in the sense that the true data generating mechanism is unknown and needs to be approximated using a simpler approximating model. The exact specifications of the data generating mechanisms are as follows:

1. Laplace distribution with location parameter $\mu = 1$ and scale parameter $b = 1$.
2. Log-normal distribution with log-mean parameter $\mu_{\text{log}} = \log\{1/\sqrt{3}\}$ and log-standard deviation parameter $\sigma_{\text{log}} = \sqrt{\log 3}$.

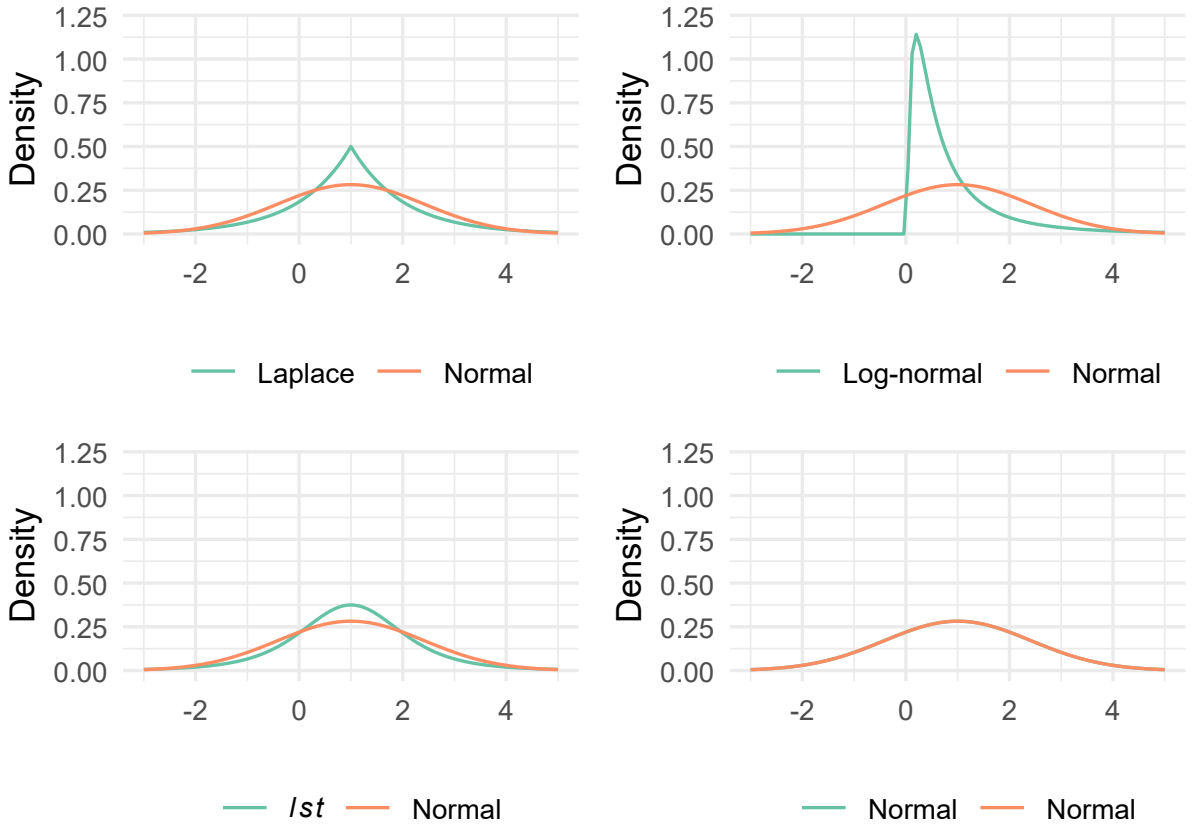

Figure 2: True and synthetic data densities for the four simulations with Laplace, Log-normal, location-scale t- and Normal densities. All data-generating mechanisms have the same mean $\mu = 1$ and variance $\sigma^2 = 2$. Note that the true and synthetic data density in the bottom right panel are completely overlapping.

3. Location-scale t-distribution with location parameter $\mu = 1$, scale parameter $\tau^2 = 1$ and degrees of freedom $\nu = 4$.
4. Normal distribution with mean $\mu = 1$ and variance $\sigma^2 = 2$.

These data generating mechanisms are chosen such that they all have the same population mean $\mu = 1$ and variance $\sigma^2 = 2$. The approximating model is a Gaussian distribution with mean $\mu = 1$ and variance $\sigma^2 = 2$. This distribution has the same mean and variance, but differs in higher-order moments, except in the fourth scenario. In the last simulation, we model the data generating mechanism correctly to obtain some understanding into how density ratio estimation behaves under a well-specified synthesis model. In all scenarios, we generate 500 data sets with $n_{\text{obs}} = 250$ observations from the true data generating mechanism, and $n_{\text{syn}} = 250$ synthetic observations from the approximating Gaussian model. The density ratio model is estimated using uLSIF with a Gaussian kernel and an L_2-regularization parameter. Both the kernel bandwidth and the regularization parameter are selected using cross-validation, using the default settings as implemented in the `densityratio` package.

TODO: Add interpretation of output; add significance tests and compare with $pMSE$ and kolmogorov-smirnov.

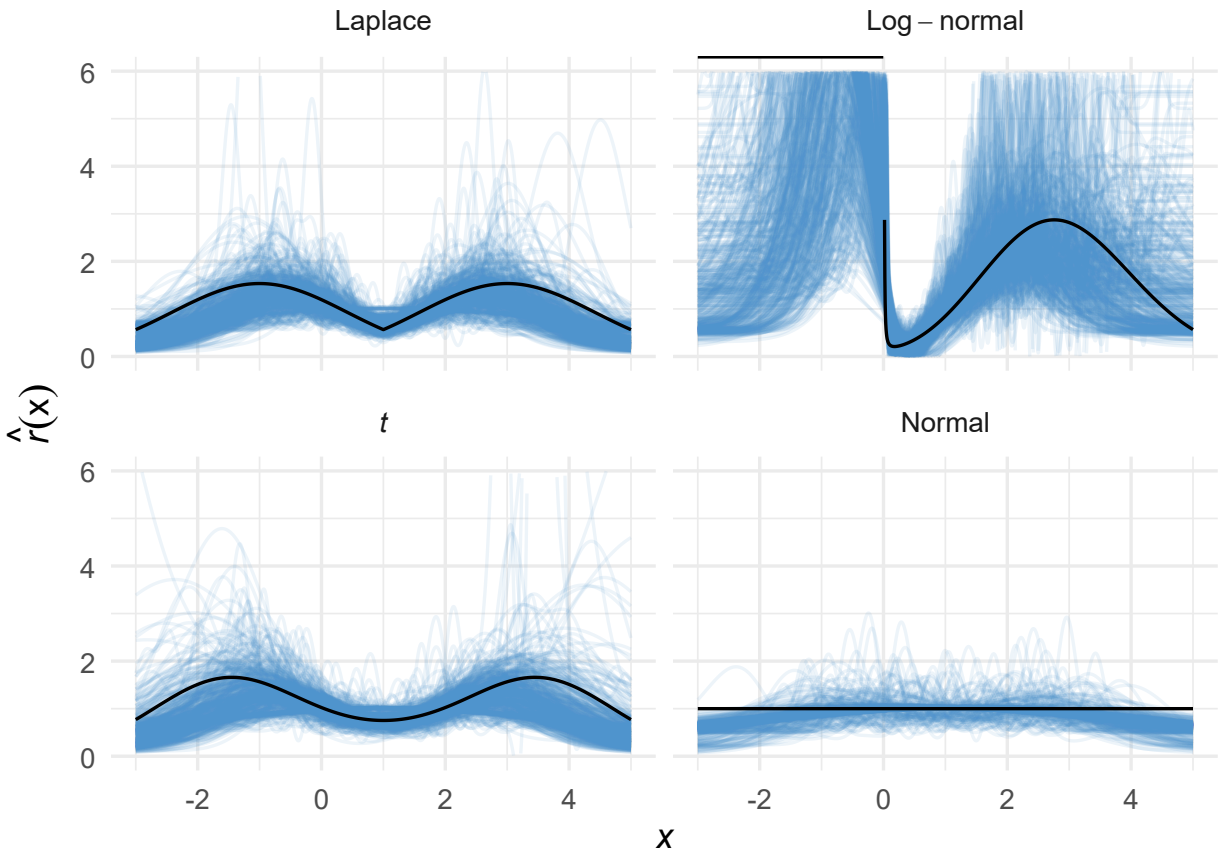

Figure 3: Estimated density ratios by unconstrained least-squares importance fitting in four univariate examples: A Laplace distribution, a log-normal distribution, a location-scale t-distribution and a normal distribution, all approximated by a normal distribution with the same mean and variance as the sample from the true distribution.

3.2 Multivariate simulations

TODO: Adjust section title

Currently, my plan is as follows, specify a (two/four) true data generating mechanism: A multivariate normal distribution with a given correlation structure, 7 and 17 variables each for two different sample size (say n = 500 and n = 2000). Append the multivariate normal distribution with three variables with a different distribution that depend on the other variables through some non-linear function. Have three synthetic data generating mechanisms, a simple multivariate normal distribution with the same means and variances but zero covariances, a model that uses a multivariate normal distribution with correlation structure taken from the real data but misses the non-linear effects (and marginal distributions of these variables), and a model that is equivalent to the true model. Then, evaluate the quality of the synthetic data sets using density ratio estimation and the $pMSE$, and see which of the two ranks the models correctly most of the time.

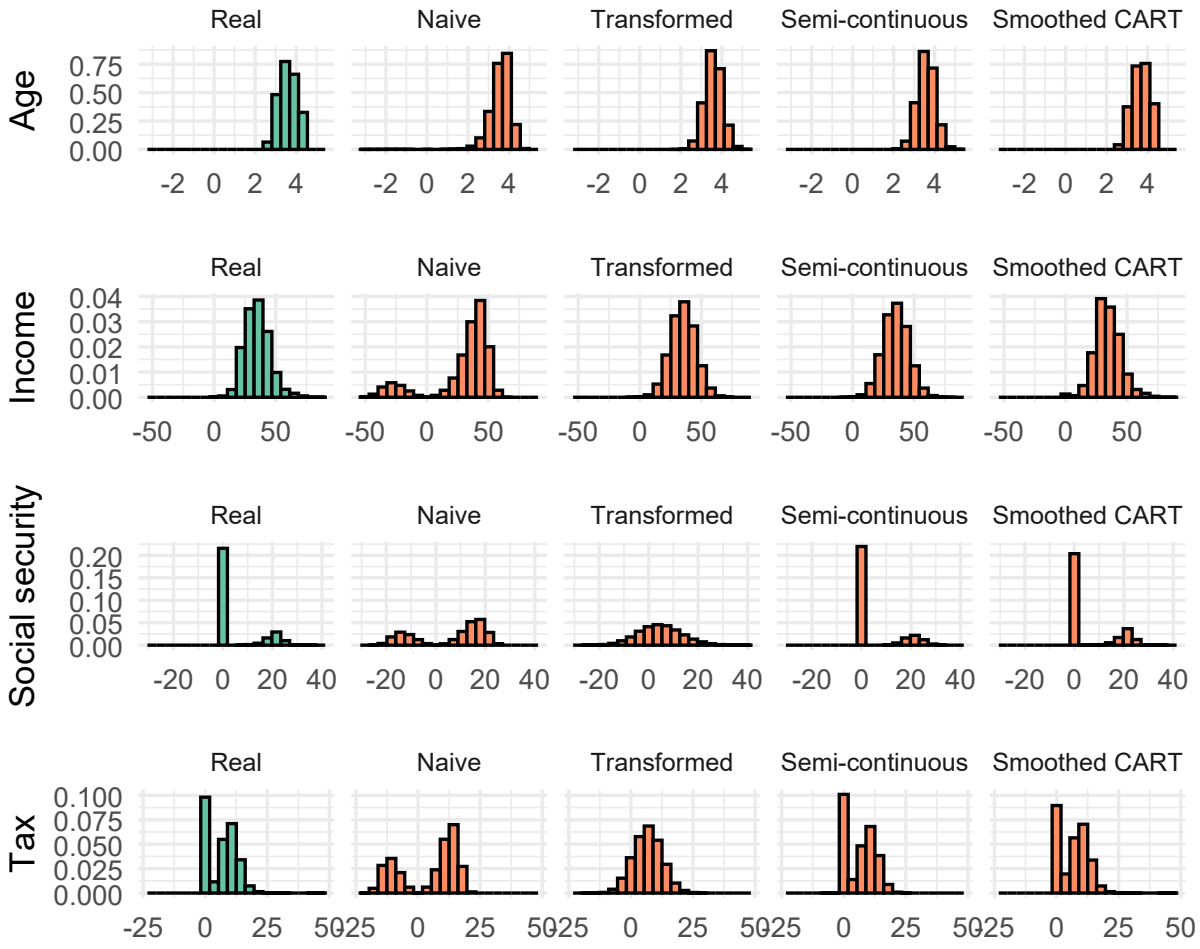

Figure 4: Real and synthetic data distributions for the variables age, household income (income), household property taxes (tax) and social security benefits (social security) on a cubic root scale.

4 Application: Synthetic data generation for the U.S. Current Population Survey

ADD TEXT

TODO: Expand example

High-dimensional density ratio estimation

Use density ratio values for weighted analysis

Incorporate point on individual data utility

- High dimensional example
- Expansion of discussion points: empirical example with weighted analyses individual data utility

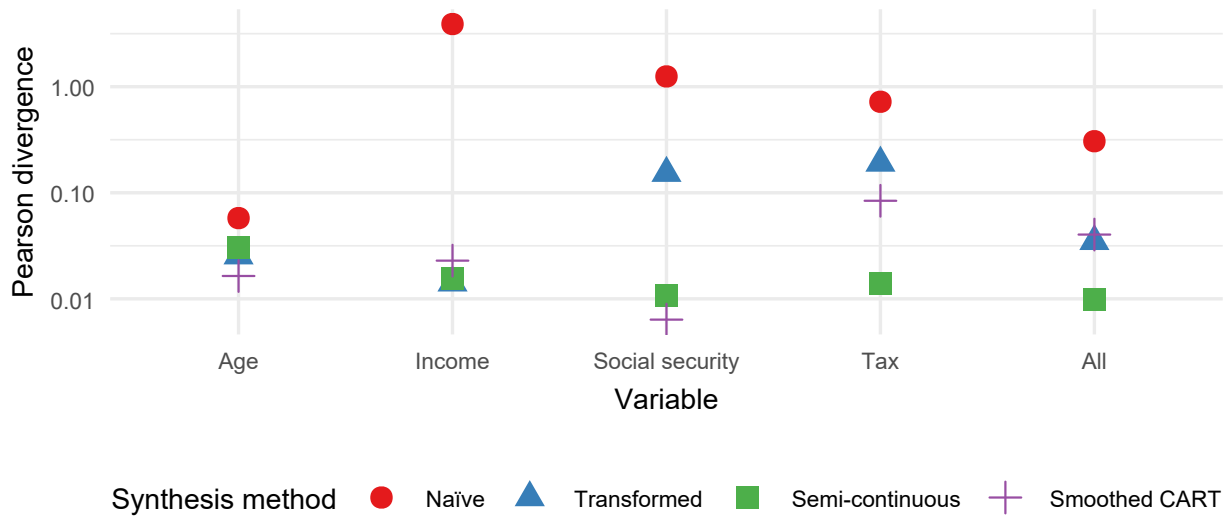

Figure 5: Pearson divergence estimates after different synthesis strategies for the separate variables and the synthetic data as a whole.

5 Discussion and conclusion

Adapt from previous version

- Privacy remark?
- When does it not work?

Density ratios can be used directly to train generative models, see: Mohamed & Lakshminarayanan (2016) and Uehara, Sata, Suzuki, Nakayama & Matsuo (2016)

Density ratio estimation is equivalent to class probability estimation in classification problems, only differing in the loss function that is employed (assuming the same model class is used for density ratio estimation and classification). See Menon & Ong, 2016

Logistic regression achieves the minimum asymptotic variance for correctly specified models (Qin, Biometrika 1998), but is not reliable for misspecified models (Kanamori, Suzuki, Sugiyama, IECE, 2010)

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.

Ali, S. M., and S. D. Silvey. 1966. "A General Class of Coefficients of Divergence of One Distribution from Another." *Journal of the Royal Statistical Society. Series B (Methodological)* 28 (1): 131–42. https://doi.org/10.1111/j.2517-6161.1966.tb00626.x.

Atenas, Javiera, Leo Havemann, and Ernesto Priego. 2015. "Open Data as Open Educational Resources: Towards Transversal Skills and Global Citizenship." *Open Praxis* 7 (4): 377–89. https://www.learntechlib.org/p/161986.

Bowen, Claire McKay, Fang Liu, and Bingyue Su. 2021. "Differentially Private Data Release via Statistical Election to Partition Sequentially." *METRON* 79 (1): 1–31. https://doi.org/10.1007/s40300-021-00201-0.

Choi, Kristy, Madeline Liao, and Stefano Ermon. 2021. "Featurized Density Ratio Estimation." In *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, edited by Cassio de Campos and Marloes H. Maathuis, 161:172–82. Proceedings of Machine Learning Research. PMLR. https://proceedings.mlr.press/v161/choi21a.html.

Crosas, Mercè, Gary King, James Honaker, and Latanya Sweeney. 2015. "Automating Open Science for Big Data." *The ANNALS of the American Academy of Political and Social Science* 659 (1): 260–73. https://doi.org/10.1177/0002716215570847.

Drechsler, Jörg. 2011. *Synthetic Datasets for Statistical Disclosure Control: Theory and Implementation*. New York: Springer Science & Business Media. https://doi.org/10.1007/978-1-4614-0326-5.

———. 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/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.

European Parliament, and Council of the European Union. 2016. "Regulation (EU) 2016/679 of the European Parliament and of the Council. Of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation)." OJ L 119, 4.5.2016, p. 1–88. May 4, 2016. https://eur-lex.europa.eu/eli/reg/2016/679/oj.

Future of Privacy Forum. 2017. "Understanding Corporate Data Sharing Decisions: Practices, Challenges, and Opportunities for Sharing Corporate Data with Researchers."

Gruber, Lukas, Markus Holzleitner, Johannes Lehner, Sepp Hochreiter, and Werner Zellinger. 2024. "Overcoming Saturation in Density Ratio Estimation by Iterated Regularization." https://doi.org/10.48550/arXiv.2402.13891.

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.

Hu, Jingchen, and Claire McKay Bowen. 2024. "Advancing Microdata Privacy Protection: A Review of Synthetic Data Methods." *WIREs Computational Statistics* 16 (1): e1636. https://doi.org/https://doi.org/10.1002/wics.1636.

Huang, Jiayuan, Alexander J. Smola, Arthur Gretton, Karsten M. Borgwardt, and Bernhard Schölkopf. 2006. "Correcting Sample Selection Bias by Unlabeled Data." In *Advances in Neural Information Processing Systems*, edited by B. Schölkopf, J. Platt, and T. Hoffman. Vol. 19. MIT Press. https://proceedings.neurips.cc/paper_files/paper/2006/file/a2186aa7c086b46ad4e8bf81e2a3a19b-Paper.pdf.

Hundepool, Anco, Josep Domingo-Ferrer, Luisa Franconi, Sarah Giessing, Eric Schulte Nordholt, Keith Spicer, and Peter-Paul De Wolf. 2012. *Statistical Disclosure Control*. John Wiley & Sons. https://doi.org/10.1002/9781118348239.

Izbicki, Rafael, Ann Lee, and Chad Schafer. 2014. "High-Dimensional Density Ratio Estimation with Extensions to Approximate Likelihood Computation." In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, edited by Samuel Kaski and Jukka Corander, 33:420–29. Proceedings of Machine Learning Research. Reykjavik, Iceland: PMLR. https://proceedings.mlr.press/v33/izbicki14.html.

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. 2012a. "f -Divergence Estimation and Two-Sample Homogeneity Test Under Semiparametric Density-Ratio Models." *IEEE Transactions on Information Theory* 58 (2): 708–20. https://doi.org/10.1109/TIT.2011.2163380.

———. 2012b. "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.

Kim, Ilmun, Aaditya Ramdas, Aarti Singh, and Larry Wasserman. 2021. "Classification accuracy as a proxy for two-sample testing." *The Annals of Statistics* 49 (1): 411–34. https://doi.org/10.1214/20-AOS1962.

Lazer, David, Alex Pentland, Lada Adamic, Sinan Aral, Albert-László Barabási, Devon Brewer, Nicholas Christakis, et al. 2009. "Computational Social Science." *Science* 323 (5915): 721–23. https://doi.org/10.1126/science.1167742.

Little, Roderick J. A. 1993. "Statistical Analysis of Masked Data." *Journal of Official Statistics* 9 (2): 407–7. https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/statistical-analysis-of-masked-data.pdf.

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/10.1016/j.neunet.2013.01.012.

Menon, Aditya, and Cheng Soon Ong. 2016. "Linking Losses for Density Ratio and Class-Probability Estimation." In *Proceedings of the 33rd International Conference on Machine Learning*, edited by Maria Florina Balcan and Kilian Q. Weinberger, 48:304–13. Proceedings of Machine Learning Research. New York, New York, USA: PMLR. https://proceedings.mlr.press/v48/menon16.html.

Mohamed, Shakir, and Balaji Lakshminarayanan. 2017. "Learning in Implicit Generative Models." https://doi.org/10.48550/arXiv.1610.03483.

Murphy, Kevin P. 2022. *Probabilistic Machine Learning: An Introduction*. MIT Press. probml.ai.

Newman, Greg, Andrea Wiggins, Alycia Crall, Eric Graham, Sarah Newman, and Kevin Crowston. 2012. "The Future of Citizen Science: Emerging Technologies and Shifting Paradigms." *Frontiers in Ecology and the Environment* 10 (6): 298–304. https://doi.org/10.1890/110294.

Obels, Pepijn, Daniël Lakens, Nicholas A. Coles, Jaroslav Gottfried, and Seth A. Green. 2020. "Analysis of Open Data and Computational Reproducibility in Registered Reports in Psychology." *Advances in Methods and Practices in Psychological Science* 3 (2): 229–37. https://doi.org/10.1177/2515245920918872.

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.

Qin, Jing. 1998. "Inferences for case-control and semiparametric two-sample density ratio models." *Biometrika* 85 (3): 619–30. https://doi.org/10.1093/biomet/85.3.619.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Raab, Gillian M., Beata Nowok, and Chris Dibben. 2017. "Guidelines for Producing Useful Synthetic Data." https://arxiv.org/abs/1712.04078.

Raab, Gillian M, Beata Nowok, and Chris Dibben. 2021. "Assessing, Visualizing and Improving the Utility of Synthetic Data." https://doi.org/10.48550/arXiv.2109.12717.

Ramachandran, Rahul, Kaylin Bugbee, and Kevin Murphy. 2021. "From Open Data to Open Science." *Earth and Space Science* 8 (5): e2020EA001562. https://doi.org/10.1029/2020EA001562.

Reiter, Jerome P. 2004. "Releasing Multiply Imputed, Synthetic Public use Microdata: An Illustration and Empirical Study." *Journal of the Royal Statistical Society Series A: Statistics in Society* 168 (1): 185–205. https://doi.org/10.1111/j.1467-985X.2004.00343.x.

Rubin, Donald B. 1993. "Statistical Disclosure Limitation." *Journal of Official Statistics* 9 (2): 461–68. https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/discussion-statistical-disclosure-limitation2.pdf.

Scott, David W. 1992. *Multivariate Density Estimation: Theory, Practice, and Visualization*. Wiley. https://doi.org/10.1002/9780470316849.

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/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, Shinichi Nakajima, Hisashi Kashima, Paul Buenau, and Motoaki Kawanabe. 2007. "Direct Importance Estimation with Model Selection and Its Application to Covariate Shift Adaptation." In *Advances in Neural*

Information Processing Systems, edited by J. Platt, D. Koller, Y. Singer, and S. Roweis. Vol. 20. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2007/file/be83ab3ecd0db773eb2dc1b0a17836a1-Paper.pdf.

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/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, 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.

Tiao, Louis C. 2018. "Density Ratio Estimation for KL Divergence Minimization Between Implicit Distributions." *Tiao.io*. https://tiao.io/post/density-ratio-estimation-for-kl-divergence-minimization-between-implicit-distributions/.

Uehara, Masatoshi, Issei Sato, Masahiro Suzuki, Kotaro Nakayama, and Yutaka Matsuo. 2016. "Generative Adversarial Nets from a Density Ratio Estimation Perspective." https://doi.org/10.48550/arXiv.1610.02920.

van de Wiel, Mark A., 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.

Volker, Thom Benjamin. 2023. "Densityratio: Direct Estimation of the Ratio of Densities of Two Groups of Observations." https://github.com/thomvolker/densityratio.

Willenborg, Leon, and Ton De Waal. 2001. *Elements of Statistical Disclosure Control*. Springer Science & Business Media. https://doi.org/10.1007/978-1-4613-0121-9.

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.

Wornowizki, Max, and Roland Fried. 2016. "Two-Sample Homogeneity Tests Based on Divergence Measures." *Computational Statistics* 31 (1): 291–313. https://doi.org/10.1007/s00180-015-0633-3.

Xu, Lei, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. 2019. *Modeling Tabular Data Using Conditional GAN*. 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.

A Density ratio estimation in `R`

TO DO

Small example of ulsif in `densityratio`

Do the same with `optim`

Do the same for a different loss function