IEEE Access
Multidisciplinary ⁝ Rapid Review ⁝ Open Access Journal

Received 1 July 2024, accepted 22 July 2024, date of publication 29 July 2024, date of current version 6 August 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3434582

RESEARCH ARTICLE

Synthetic Tabular Data Validation: A Divergence-Based Approach

PATRICIA A. APELLÁNIZ, ANA JIMÉNEZ, BORJA ARROYO GALENDE, JUAN PARRAS, AND SANTIAGO ZAZO

Information Processing and Telecommunications Center, ETS Ingenieros de Telecomunicación, Universidad Politécnica de Madrid, 28040 Madrid, Spain

Corresponding author: Patricia A. Apellániz (patricia.alonsod@upm.es)

This work was supported by the Genomics and Personalized Medicine for all though Artificial Intelligence in Haematological Diseases (GenoMed4All) and Synthetic generation of hematological data over federated computing frameworks (SYNTHEMA) Projects from European Union's Horizon 2020 Research and Innovation Program under Grant 101017549 and Grant 101095530.

ABSTRACT The ever-increasing use of generative models in various fields where tabular data is used highlights the need for robust and standardized validation metrics to assess the similarity between real and synthetic data. Current methods lack a unified framework and rely on diverse and often inconclusive statistical measures. Divergences, which quantify discrepancies between data distributions, offer a promising avenue for validation. However, traditional approaches calculate divergences independently for each feature due to the complexity of joint distribution modeling. This paper addresses this challenge by proposing a novel approach that uses divergence estimation to overcome the limitations of marginal comparisons. Our core contribution lies in applying a divergence estimator to build a validation metric considering the joint distribution of real and synthetic data. We leverage a probabilistic classifier to approximate the density ratio between datasets, allowing the capture of complex relations. We specifically calculate two divergences: the well-known Kullback-Leibler (KL) divergence and the Jensen-Shannon (JS) divergence. KL divergence offers an established use in the field, while JS divergence is symmetric and bounded, providing a reliable metric. The efficacy of this approach is demonstrated through a series of experiments with varying distribution complexities. The initial phase involves comparing estimated divergences with analytical solutions for simple distributions, setting a benchmark for accuracy. Finally, we validate our method on a real-world dataset and its corresponding synthetic counterpart, showcasing its effectiveness in practical applications. This research offers a significant contribution with applicability beyond tabular data and the potential to improve synthetic data validation in various fields.

INDEX TERMS Divergence, Kullback-Leibler, Jensen-Shannon, synthetic data generation, tabular data, validation, density ratio.

I. INTRODUCTION

Generative models (GMs) are emerging as a disruptive force within artificial intelligence. By learning the underlying patterns and distributions of real-world data, these models can create novel synthetic data replicating the characteristics of the original domain. These synthetic data offer significant advantages for various research and development applications. In particular, synthetic data can be used for data augmentation, addressing limitations in data collection efforts. In addition, it can ensure privacy protection by replacing sensitive real-world information. Furthermore, it facilitates the exploration of rare or extreme scenarios (edge cases) that might be difficult or expensive to encounter in real-world settings. As a result, the synthetic data generated by these models is accelerating innovation in various fields, including healthcare [1], finance [2], and the development of autonomous vehicles [3], to mention some.

Following the impressive advances in GMs for data generation, a critical aspect emerges: ensuring the quality and effectiveness of synthetic data. Replicating the source data's statistical properties, accuracy, consistency, and domain-specific relevance is essential. This challenge becomes significantly pronounced as GMs gain traction within engineering design communities. The need for robust and standardized validation metrics becomes paramount.

The associate editor coordinating the review of this manuscript and approving it for publication was Sawyer Duane Campbell.

The current literature highlights a significant gap in this area. Although there are standardized evaluation metrics for the generation of synthetic images [4], [5], and the generation of text [6], [7], measuring the quality of synthetic tabular data presents unique challenges. Unlike image data, qualitative evaluation through visual inspection is not feasible. Additionally, relying solely on expert insight can be highly inefficient. The existing landscape for tabular data validation often focuses primarily on the efficacy or utility in machine learning tasks [8], [9]. However, a consistent approach to assessing the similarity between synthetic and real tabular data remains sparse. Studies employ a diverse set of metrics, including pairwise correlation difference [10], support coverage [11], likelihood fitness [12], alongside various statistic values [13], [14]. Reviews such as [15] attempt to categorize these techniques, but a genuinely standardized approach or a single metric that captures the full spectrum of statistical information from a distribution remains a challenge.

This paper aims to address this gap by proposing a divergence estimator. Divergences offer a comprehensive approach to the validation of similarity. These measures quantify the discrepancy between two probability distributions, effectively capturing the differences between real and synthetic data. Unlike traditional methods that often focus on individual attributes [16], [17], divergences can consider the joint distribution of all attributes, providing a more holistic view of the data. However, it is essential to acknowledge the challenge associated with joint distributions. Modeling the intricate relations between all features in a dataset can be computationally expensive and intractable. Although some existing papers explore divergences or distances for validation, they often apply them to single attributes because of this limitation [11], [18], [19]. Crucially, even if individual features appear similar, the joint distribution, which captures their relations, may differ significantly. Estimating divergences between marginal distributions neglects the correlation between features. For example, in survival analysis databases, the risk associated with a disease is often closely related to the patient's age [20]. Although these two variables might be generated to mimic the original distributions, this method might fail to preserve their intrinsic, possibly nonlinear, correlation. In practice, disrupting original correlations can result in synthetic data that, while statistically similar in univariate terms, diverges significantly in multivariate contexts. This discrepancy is critical because it can affect the utility and reliability of synthetic data in applications where feature interdependencies play a crucial role, such as predictive modeling in healthcare. To address this issue, correlation matrices are sometimes employed to assess relations between features [21]. However, these matrices primarily reveal linear relations, provide complex representations usually interpreted graphically, and are challenging to analyze quantitatively. This underscores the importance of considering the joint distribution for an accurate assessment. Our proposed divergence estimator addresses this challenge by efficiently capturing these complex joint distributions. This approach enables the computation of divergences between real and synthetic data, allowing for a more comprehensive and robust evaluation of the quality of synthetic data. The advantages of using divergences for validation remain the same:

- Sensitivity to Complex Distributions: Divergences can capture the full complexity of real-world data distributions, including nonlinear relations and rare events, which traditional methods might miss.
- Robustness to Noise: Divergences are less sensitive to noise and outliers, making them more reliable for practical applications.
- Interpretability: Divergences can be easily interpreted and visualized, providing valuable insights into the discrepancies between real and synthetic data. They condense statistical similarity between distributions into a single scalar, facilitating quantitative interpretation of the findings.

Therefore, we introduce a novel methodology to validate real and synthetic data similarity:

- **Divergence-based Similarity Estimation:** Our approach leverages divergence estimation to quantify the discrepancy between the joint probability distributions of real and synthetic tabular data. This approach goes beyond traditional validation metrics, often focusing on marginal statistics.
- **Probabilistic Discriminator Network for Density Ratio Estimation:** To facilitate divergence calculations, we introduce an innovative approximation technique based on a probabilistic discriminator network [22]. This network plays a crucial role in estimating the density ratio between real and synthetic distributions, which forms the basis for divergence computations.
- **KL Divergence and JS Divergence for Validation:** We acknowledge the widespread use of Kullback-Leibler (KL) divergence in information theory (often referred to as relative entropy) for divergence calculations. However, we recognize its limitations: KL divergence is not an actual distance in terms of mathematics [23] due to its asymmetry and the potential for infinite values in specific scenarios. To address this, we propose the inclusion of the Jensen-Shannon (JS) divergence, which offers the desirable properties of being bounded and symmetric, thus defining a proper distance metric. This bounded metric is crucial to assess the quality of synthetic data generation.
- **Rigorous Evaluation Framework:** We introduce a comprehensive evaluation framework to assess the efficacy of our proposed method. We initially apply our method to simple distributions where analytical solutions for divergences exist. This serves as a benchmark to validate the accuracy of our estimated divergences. Following the initial validation, we progress to more complex scenarios and finally apply our method to real-world datasets and their corresponding

synthetically generated counterparts. This allows us to assess the effectiveness of our approach in practical applications.
- **Broad Applicability:** Due to the adaptability of our divergence-based validation method, it can be applied to various types of data beyond tabular formats. This has the potential to significantly impact numerous fields, such as healthcare, finance, and manufacturing. Adopting our approach can substantially improve the accuracy and reliability of synthetic data validation in various domains.

In summary, our contribution diverges from the traditional emphasis on divergence estimation, which is already a well-established and robust field of study. Instead, we focus on applying divergence estimation to validate synthetic data generation. As highlighted in our literature review, a notable opportunity exists to establish a standardized validation metric in this area. We propose that estimating divergences between real and synthetic data can serve as a definitive measure of the accuracy of the data generation process. By leveraging divergence metrics, we offer a quantitative measure that is both interpretable and grounded in solid theoretical principles. This approach introduces a robust metric that enhances the validation process for synthetic data generation, ensuring that synthetic data not only resemble the original in terms of individual feature distributions but also preserve the intricate interdependencies between features. These interdependencies are crucial for the data's authenticity and applicability in real-world scenarios. This advancement represents a significant step forward in standardizing validation practices and improving the reliability of synthetic datasets across various fields.

The structure of the paper is carefully designed to facilitate a comprehensive understanding of our proposal. It begins with Section I, an introductory section that sets the stage by highlighting the importance of robust validation metrics in the use of synthetic data across various domains. This is followed by Section II, a detailed background section that delves into the theoretical and analytical aspects of divergence measures, specifically the KL and JS divergences. Furthermore, this section discusses the density ratio estimation based on probabilistic classification, as outlined in reference [22], which is crucial for the subsequent divergence calculation. The methodology section (Section III) then defines our approach using generative models to obtain this density ratio estimation, detailing the process analytically, visually, and programmatically. The fourth section describes the experimental setup and discusses the validation results from four different experiments, each increasing in complexity, to demonstrate the efficacy of our approach. The paper concludes in Section V with a summary of our findings, highlighting key conclusions and potential future research lines. An appendix is also included, which presents additional results from our experiments, providing deeper insights into our validation process.

II. BACKGROUND

A. DIVERGENCE DEFINITION

Let us consider two probability distributions, $p(x)$ and $q(x)$, for a random variable x. In the context of probability distributions, divergence quantifies the dissimilarity between $p(x)$ and $q(x)$. In essence, it measures how different these distributions are regarding the probabilities they assign to each possible value of x. Divergences play a crucial role in various fields, including Machine Learning (ML), as they provide a way to compare and analyze the behavior of different probability distributions.

The Kullback-Leibler divergence is a common measure in information theory to quantify the discrepancy between $p(x)$ and $q(x)$. Denoted $D_{\mathbb{KL}}\left(p(x)\|q(x)\right)$, the KL divergence captures the asymmetry of this difference by measuring the expected penalty incurred by assuming $q(x)$ to be an approximation of the true distribution $p(x)$. Due to its extensive applicability in various fields, the KL divergence remains a dominant metric for comparing probability distributions. The KL divergence between $p(x)$ and $q(x)$ is given by

$$D_{\mathbb{KL}}\left(p(x)\|q(x)\right) = \int p(x)\log\frac{p(x)}{q(x)}dx. \tag{1}$$

Unfortunately, this expression is usually intractable. Thus, we must approximate it using Monte Carlo (MC) simulation, assuming we can draw L samples from $p(x)$:

$$\int p(x)\log\frac{p(x)}{q(x)}dx \approx \frac{1}{L}\sum_{i=1}^{L}\log\frac{p\left(x_i\right)}{q\left(x_i\right)}, \quad x_i \sim p(x). \tag{2}$$

To create a symmetric and bounded measure of divergence between $p(x)$ and $q(x)$, the Jensen-Shannon divergence is based on the KL divergence. The JS divergence denoted $D_{\mathbb{JS}}\left(p(x)\|q\left(x\right)\right)$, is a true distance defined as the KL average between both distributions and a common reference:

$$\begin{aligned} &D_{\mathbb{JS}}\left(p(x)\|q(x)\right) \\ &\quad = \frac{1}{2}D_{\mathbb{KL}}\left(p(x)\|m(x)\right) + \frac{1}{2}D_{\mathbb{KL}}\left(q(x)\|m(x)\right). \end{aligned} \tag{3}$$

The common reference distribution, $m(x)$, is typically chosen as the midpoint between the two original distributions, such as the average: $m(x) = \left(p(x) + q(x)\right)/2$. Notably, the JS divergence is inherently bounded. When employing a base-2 logarithm, the maximum divergence value is 1.

Similarly to KL, the JS divergence can also be estimated using MC simulation:

$$\begin{aligned} &D_{\mathbb{JS}}\left(p(x)\|q(x)\right) \\ &\approx \frac{1}{2L}\sum_{i=1}^{L}\log\frac{p\left(x_i\right)}{m\left(x_i\right)} + \frac{1}{2L}\sum_{i=1}^{L}\log\frac{q\left(\tilde{x}_i\right)}{m\left(\tilde{x}_i\right)} \\ &= \frac{1}{2L}\sum_{i=1}^{L}\log\left(\frac{2p\left(x_i\right)}{p\left(x_i\right)+q\left(x_i\right)}\right) \\ &\qquad + \frac{1}{2L}\sum_{i=1}^{L}\log\left(\frac{2q\left(\tilde{x}_i\right)}{p\left(\tilde{x}_i\right)+q\left(\tilde{x}_i\right)}\right), \end{aligned} \tag{4}$$

assuming that we can generate L samples from both distributions

$$\begin{cases} x_i \sim p(x) \\ \tilde{x}_i \sim q(x) \end{cases}. \tag{5}$$

B. ESTIMATING DENSITY RATIO USING PROBABILISTIC CLASSIFICATION

Density ratio estimation focuses on estimating the ratio $r^*(x)$ of two probability densities, $p(x)$ and $q(x)$, based solely on samples drawn from these distributions. This field has seen significant theoretical development, particularly regarding complex ratio estimators and their convergence properties [24]. The ratio is therefore defined as

$$r^*(x) = \frac{p(x)}{q(x)}, \tag{6}$$

where $*$ refers to the exact value. This approach avoids directly estimating individual densities since any errors in the denominator (i.e., $q(x)$) are dramatically amplified during the ratio calculation. The classical probabilistic classification approach remains popular among the various methods of estimating the density ratio due to its relative simplicity. This subsection reviews the method described by [22] for estimating the KL divergence. We further demonstrate its application to the JS divergence estimation as well.

Estimating (1) and (3) is based on determining the unknown density ratio r^*. This section establishes a connection between the density ratio of two distributions, $p(x)$ and $q(x)$, and a probabilistic classifier that optimally distinguishes samples drawn from these distributions. Consider a scenario in which we possess samples from both distributions, $p(x)$ and $q(x)$, with each sample labeled according to its origin. Using a classifier to estimate the class-membership probabilities for each sample, we can derive an estimator for the density ratio. Let $X_p = x^{(1)}, x^{(2)}, \ldots, x^{(M_p)}$ and $X_q = \tilde{x}^{(1)}, \tilde{x}^{(2)}, \ldots, \tilde{x}^{(M_q)}$ represent sets of real and synthetic data samples, respectively, where M_p and M_q denote the corresponding sample sizes.

We proceed by forming a combined dataset, denoted as $D = (x_n, y_n)_{n=1}^{N}$, where $M_T = M_p + M_q$ represents the total number of samples. The label y_n associated with each sample x_n indicates its source distribution: $y_n = 1$ for samples from $p(x)$ and $y_n = 0$ for samples from $q(x)$. Consequently, we have $p(x) = \mathcal{P}(x \mid y = 1)$ and $q(x) = \mathcal{P}(x \mid y = 0)$, where $\mathcal{P}$ denotes the probability.

By applying Bayes' theorem, we can express the density ratio as:

$$\begin{aligned} r^*(x) = \frac{p(x)}{q(x)} &= \frac{\mathcal{P}(x \mid y = 1)}{\mathcal{P}(x \mid y = 0)} \\ &= \left(\frac{\mathcal{P}(y = 1 \mid x)\mathcal{P}(x)}{\mathcal{P}(y = 1)}\right)\left(\frac{\mathcal{P}(y = 0 \mid x)\mathcal{P}(x)}{\mathcal{P}(y = 0)}\right)^{-1} \\ &= \frac{\mathcal{P}(y = 0)}{\mathcal{P}(y = 1)}\frac{\mathcal{P}(y = 1 \mid x)}{\mathcal{P}(y = 0 \mid x)}. \end{aligned} \tag{7}$$

Equal prior probabilities are assumed for the source distributions, specifically $\mathcal{P}(y = 0) = \mathcal{P}(y = 1)$. This assumption is strategically made to prevent the estimation process from being skewed towards any particular class. By setting the prior probabilities to be equal, the ratio of marginal probabilities effectively cancels out, leaving:

$$r^*(x) = \frac{\mathcal{P}(y = 1 \mid x)}{\mathcal{P}(y = 0 \mid x)}. \tag{8}$$

This expression reveals that r^* can be estimated solely based on the posterior probability, $\mathcal{P}(y = 1 \mid x)$, which represents the probability that a sample originates from $p(x)$ given its features.

To simplify even more, the logit function, denoted by σ^{-1}, offers a convenient way to transform the posterior probability into the ratio of the original probabilities. The logit function is defined as the inverse of the logistic function σ:

$$\sigma^{-1}(z) = \ln\left(\frac{z}{1 - z}\right), \tag{9}$$

where z represents any probability. The logit function has the property that the log odds of two probabilities equal the log of their ratio. In our case, applying the logit function to both the numerator and denominator of the expression for (8) results in the log ratio of the original probabilities:

$$\begin{aligned} \ln&\left(\frac{\mathcal{P}(y = 1 \mid x)}{\mathcal{P}(y = 0 \mid x)}\right) \\ &= \sigma^{-1}\left(\mathcal{P}(y = 1 \mid x)\right) - \sigma^{-1}\left(\mathcal{P}(y = 0 \mid x)\right) \end{aligned} \tag{10}$$

Since the prior probabilities are assumed to be equal, the second term cancels out, leaving us with:

$$\ln\left(r^*(x)\right) = \sigma^{-1}(\mathcal{P}(y = 1 \mid x)) \tag{11}$$

Taking the exponent of both sides recovers the original expression for $r^*(x)$ but expressed in terms of the odds of the posterior probability:

$$r^*(x) = \exp\left[\sigma^{-1}(\mathcal{P}(y = 1 \mid x))\right]. \tag{12}$$

Therefore, by applying the logit function, we can simplify the calculation of the density ratio. The transformation converts the posterior probability into a log-odds representation, ultimately leading to the desired log ratio. This establishes a crucial link between the density ratio and the probabilistic classifier. By training a classifier to effectively distinguish between samples from $p(x)$ and $q(x)$, we can obtain estimates of posterior probabilities and subsequently derive an estimate of the density ratio using the logit transformation.

III. METHODOLOGY

This research uses a neural network (NN) classifier, denoted D_θ with parameters θ, to approximate the posterior probability $P(y = 1 \mid x)$. This network acts as a discriminator to classify input samples originating from the $p(x)$ or $q(x)$ distributions. The network architecture is designed to capture the discriminative features that distinguish these distributions effectively. In this study, we aim to understand how input data variations influence the ratio estimation between distributions. To isolate the impact of input data, we deliberately

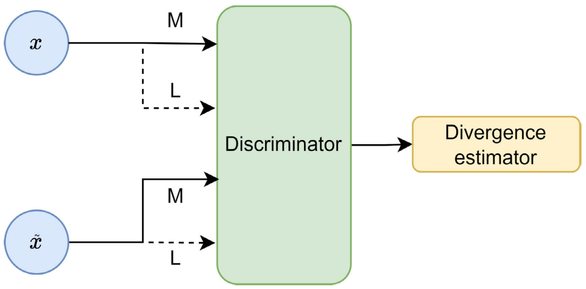


FIGURE 1. Architecture of the neural network-based divergence estimator to assess the dissimilarity between samples from two datasets. The discriminator takes two sets of samples as input: *M* samples from each set to train and *L* samples from each to infer the divergence.

chose not to fine-tune the model across different experiments, maintaining a consistent architecture throughout. This approach allows us to examine the robustness of the estimator across various data types without the confounding effects of model optimization. The model parameters were selected to balance interpretability and provide meaningful insights, ensuring that our results illustrate the estimator's practical capabilities rather than achieving optimal performance in each scenario. This strategy emphasizes the versatility and applicability of the estimator in real-world contexts, offering valuable insights into the challenges of synthetic data validation.

Specifically, D_θ employs a three-layer architecture with a decreasing number of neurons per layer: 256 in the first hidden layer, followed by 64 and 32 neurons in the subsequent hidden layers, respectively. The Leaky Rectified Linear Unit activation function is used throughout the hidden layers to introduce nonlinearity into the model. In addition, dropout, batch normalization, and early stopping techniques are incorporated to prevent overfitting during the training process.

The estimator for r^* can be constructed as a function of the classifier's output:

$$\begin{aligned} r_\theta(x) &= \exp\left[\sigma^{-1}\left(D_\theta(x)\right)\right] \\ &\approx \exp\left[\sigma^{-1}(\mathcal{P}(y=1 \mid x))\right] = r^*(x). \end{aligned} \tag{13}$$

The optimal class probability estimator is learned by minimizing a suitable loss function, such as the binary cross-entropy loss:

$$\begin{aligned} &\mathcal{L}(\theta) \\ &= -\mathbb{E}_{p(x)}\big[\log \mathcal{D}_\theta(x)\big] - \mathbb{E}_{q(x)}\big[\log(1-\mathcal{D}_\theta(x))\big]. \end{aligned} \tag{14}$$

Once trained, the class probability estimator can construct MC estimates of the KL and JS divergences. The KL divergence between $p(x)$ and $q(x)$ can be estimated as

$$\begin{aligned} D_{\mathbb{KL}}\big(p(x)\|q(x)\big) &= \mathbb{E}_{p(x)}\big[\log r^*(x)\big] \\ &\approx \frac{1}{L}\sum_{i=1}^{L}\log r^*(x_i) \approx \frac{1}{L}\sum_{i=1}^{L}\log r_\theta(x_i) \\ &= \frac{1}{L}\sum_{i=1}^{L}\sigma^{-1}\big(\mathcal{D}_\theta\,(x_i)\big), \end{aligned} \tag{15}$$

where $x_i \sim p(x)$, L is the number of samples used for the estimate.

Similarly, the JS divergence can be estimated as:

$$\begin{aligned} &D_{\mathbb{JS}}\big(p(x)\|q(x)\big) \\ &\approx \frac{1}{2L}\sum_{i=1}^{L}\log\big(2\mathcal{D}_\theta(x_i)\big) \\ &\quad + \frac{1}{2L}\sum_{i=1}^{L}\log\big(2-2\mathcal{D}_\theta(\tilde{x}_i)\big) \\ &= \frac{1}{2L}\sum_{i=1}^{L}\log\big(\mathcal{D}_\theta(x_i)\big) \\ &\quad + \frac{1}{2L}\sum_{i=1}^{L}\log\big(1-\mathcal{D}_\theta(\tilde{x}_i)\big), \end{aligned} \tag{16}$$

where $\tilde{x}_i \sim q(x)$ denotes the number of samples used from the second distribution during divergence estimation.

Interestingly, the discriminator loss function, $\mathcal{L}(\theta)$, converges to a lower bound of the JS divergence up to a constant, $-2 \cdot D_{\mathbb{JS}}\big(p(x)\|q(x)\big) + \log 4$ [22]. This relation can be established by analyzing the upper bound of the loss function:

$$\begin{aligned} &\sup_\theta \mathbb{E}_{p(x)}\big[\log D_\theta(x)\big] + \mathbb{E}_{q(x)}\big[\log(1-D_\theta(x))\big] \\ &= \mathbb{E}_{p(x)}[\log \mathcal{P}(y=1 \mid x)] + \mathbb{E}_{q(x)}[\log \mathcal{P}(y=0 \mid x)] \\ &= \mathbb{E}_{p(x)}\Big[\log\frac{p(x)}{p(x)+q(x)}\Big] + \mathbb{E}_{q(x)}\Big[\log\frac{q(x)}{p(x)+q(x)}\Big] \\ &= \mathbb{E}_{p(x)}\Big[\log\frac{1}{2}\frac{p(x)}{m(x)}\Big] + \mathbb{E}_{q(x)}\Big[\log\frac{1}{2}\frac{q(x)}{m(x)}\Big] \\ &= \mathbb{E}_{p(x)}\Big[\log\frac{p(x)}{m(x)}\Big] + \mathbb{E}_{q(x)}\Big[\log\frac{q(x)}{m(x)}\Big] - 2\log 2 \\ &= 2 \cdot D_{\mathbb{JS}}\big(p(x)\|q(x)\big) - \log 4. \end{aligned} \tag{17}$$

Then, we can establish:

$$\begin{aligned} &2 \cdot D_{\mathbb{JS}}\big(p(x)\|q(x)\big) - \log 4 \\ &\geq \sup_\theta \mathbb{E}_{p(x)}\big[\log D_\theta(x)\big] + \mathbb{E}_{q(x)}\big[\log(1-D_\theta(x))\big]. \end{aligned} \tag{18}$$

Following this analysis, we can arrive at the desired inequality. However, we can manipulate the inequality by negating both sides for a more convenient form.

$$\begin{aligned} &-2 \cdot D_{\mathbb{JS}}\big(p(x)\|q(x)\big) + \log 4 \\ &\leq -\sup_\theta \mathbb{E}_{p(x)}\big[\log D_\theta(x)\big] + \mathbb{E}_{q(x)}\big[\log(1-D_\theta(x))\big] \\ &= \inf_\theta -\mathbb{E}_{p(x)}\big[\log D_\theta(x)\big] - \mathbb{E}_{q(x)}\big[\log(1-D_\theta(x))\big] \\ &= \inf_\theta \mathcal{L}(\theta). \end{aligned} \tag{19}$$

The architecture of the divergence estimator proposed between two distributions is depicted in Fig. 1. The discriminator network receives two sets of samples: M samples from the first distribution $p(x)$ labeled class 1 and M samples from the second distribution $q(x)$ labeled class 0. During training, the discriminator aims to distinguish between these two sets. Subsequently, L samples from each distribution

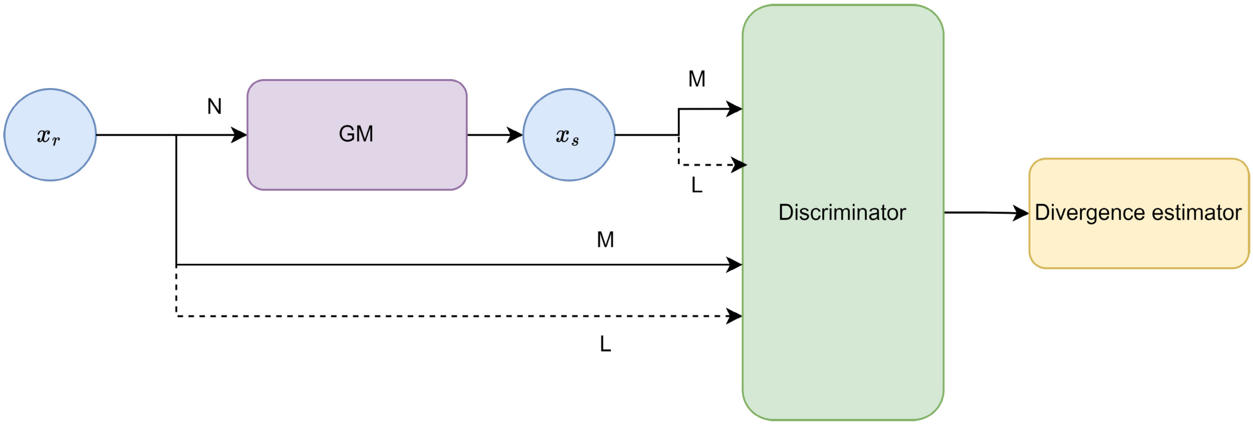


FIGURE 2. Generative model for divergence estimation between real and synthetic data. The generative model learns an approximation of N samples from real data x_r, denoted x_S. Subsequently, M samples are drawn from each distribution to train the divergence estimation discriminator. Finally, L samples from each distribution estimate the divergence between real and synthetic data.

TABLE 1. Experiments configuration summary.

Experiment	Compared Distributions	Generative Model	N	M	L
Experiment 1	Multivariate Gaussian Distributions	-	-	[20, 200, 2000]	[20, 200, 2000]
Experiment 2	Gaussian Mixture Distributions	-	-	[20, 200, 2000]	[20, 200, 2000]
Experiment 3	Gaussian Mixture and Synthetic Data	Gaussian Mixture Model	[10, 20, 30, ... , 150]	2,000	2,000
Experiment 4	Real and Synthetic Data	CTGAN [12], VAE [25]	10,000	7,500	1,000

Experiments carried out to assess the effectiveness of the proposed approach. Experiments 1 and 2 investigate the impact of distribution complexity by varying the number of M and L. Experiments 3 and 4 focus on the application of generative processes. A generative model is introduced to create synthetic data based on N samples, enabling a comparison between real and synthetic scenarios. Different values of N, M, and L could have been used; however, we believe that the chosen values are sufficiently illustrative of our aims. Moreover, we sought a trade-off between simplicity in interpretation and the ability to produce exemplary results.

are used to estimate the divergence between the underlying probability distributions.

This study evaluates the dissimilarity between real and synthetic data generated by a Generative Model (GM). Minimizing the divergence between the real and synthetic data distributions is crucial in this context. Ideally, the divergence should approach zero, indicating that the generated data become virtually indistinguishable from the real data. This interchangeability allows the use of synthetic data in various applications where real data might be scarce or sensitive. From the discriminator's perspective, achieving minimal divergence implies that it cannot reliably differentiate between real and synthetic samples, indicating successful data generation. For GMs, the distributions of interest become:

- $p_R(x) = p(x)$, representing the real data distribution.
- $p_S(x) = q(x)$, representing the synthetic data distribution generated by the model.

Training data consist of N samples drawn from the real data distribution, denoted $x_r \sim p_R(x)$. These data are used to train the GM, as illustrated in Fig. 2. The GM learns to approximate the real data distribution as $p_S(x)$, allowing us to sample from this distribution and obtain synthetic data points $x_s \sim p_S(x)$. Following the architecture described in Fig. 1, these synthetic samples are then used along with the real data samples to train the discriminator and estimate the divergence between the real and synthetic data distributions.

IV. EXPERIMENTS

We aim to validate the effectiveness of our proposed divergence estimator for data generation. We achieve this through an exhaustive experimental evaluation in four different experiments, described in Table 1 designed to assess the estimator's performance under various conditions. The initial set of experiments focuses on controlled scenarios that involve comparisons of simple theoretical data distributions. Subsequently, the complexity is gradually increased by incorporating generative scenarios that reproduce real-world applications. This progression in complexity not only tests the estimator's ability to handle increasingly challenging conditions but also demonstrates its scalability and reliability. The increasing complexity is essential to ensure that our estimator remains accurate and efficient, even as the distributional assumptions become less defined and more reflective of the irregularities typically found in real data.

We comprehensively analyze the influence of sample size on the estimator's performance. This analysis covers the impact of the training and validation sets sizes for the discriminator network and the number of samples used during the generative process. To gain a deeper understanding, we evaluated different configurations for each parameter, as the table shows. For every experiment, all possible combinations of N, M, and L are executed five times with unique random seeds. This rigorous approach employing multiple random seeds mitigates the effects of inherent randomness within training

TABLE 2. Experiments configuration summary.

Experiment	Validation Technique
Experiment 1	Analytical, MC and Discriminator Estimations
Experiment 2	MC and Discriminator Estimations
Experiment 3	MC and Discriminator Estimations
Experiment 4	Discriminator Estimation

Validation procedure for each experiment depending on their level of complexity.

processes. Consequently, it provides a statistically robust evaluation and a more reliable representation of the observed trends. Ultimately, this comprehensive analysis facilitates identifying the configuration that yields the most accurate assessment of the divergence between data distributions.

The inherent variability in data nature and complexity across different experiments needs diverse divergence estimation methods. Table 2 details the specific techniques employed for each experiment, allowing for a comparative analysis whenever possible. The experiment's level of complexity guides the selection of the validation technique.

- **Analytical Divergence:** This method, calculated only for $D_{\mathbb{KL}}$ for simplicity, represents the true divergence value due to the specific distributions used in the experiment.
- **MC Estimated Divergence:** This is a widely used estimation approach, but it can be computationally expensive.
- **Discriminator Estimated Divergence:** This method uses our proposed discriminator network to learn the density ratio between the two distributions and estimate the divergence.

In the experiments conducted, the parameter values of the different models used were explicitly chosen to yield illustrative results to analyze the performance of the proposed estimator in validating synthetic data generation. We recognize that alternative parameters could be employed or fine-tuning could be undertaken to identify optimal settings; however, we intend to observe the effect of the input data on the models as such and to demonstrate their potential. This approach allows us to focus on understanding how variations in input data influence the estimator's ability to validate synthetic datasets, thereby showcasing the practical applicability and robustness of the estimator across different data scenarios.

For complete transparency and reproducibility, the data and code used in this study are publicly available on our repository (https://github.com/Patricia-A-Apellaniz/divergence_estimator).

A. EXPERIMENT 1

Table 3 presents the results obtained for $D_{\mathbb{KL}}$ and $D_{\mathbb{JS}}$ between random multivariate Gaussian distributions. A significant advantage of this study is the availability of a closed-form solution for the KL divergence, which serves as the ground truth value. We use the MC simulation estimate for the JS divergence as the ground truth. To ensure the reliability of this experiment, we opted for multivariate Gaussian distributions with a dimensionality of 10. This allows us to compute the true divergence values and assess the accuracy of estimating the ratio of the proposed discriminator. The results reveal a critical relation between sample size and estimation accuracy, particularly for the discriminator-based approach. When the number of samples is limited ($M = 20$, $L = 20, 200$), the estimated divergence from the discriminator deviates significantly from both the analytical value and the MC estimate. This suggests a potential overfitting due to insufficient data, leading to underestimating the true divergence. However, as the number of samples increases, the estimated divergence from the discriminator network progressively converges towards the ground truth value, followed by improved confidence interval precision. Figs. 3 and 4 confirm this trend. Both figures illustrate how increasing the number of training samples M and validation samples L reduces the error associated with the estimated divergence ratio. Further support for the correlation between sample size and estimation accuracy comes from the discriminator loss curves depicted in Fig. 5. As detailed in [22], the loss function converges to a constant value approximating the Jensen-Shannon divergence. The figure visually confirms this concept, as the loss curves flatten with increasing training M and validation L samples. In contrast, a low number of samples (e.g., $M = 20$, $L = 20$) results in larger fluctuations in the loss function, potentially indicating discriminator overfitting. These validations demonstrate the effectiveness of our proposed method: with sufficient training data, the discriminator can accurately learn the density ratio and provide reliable estimates of the divergences, particularly for the JS divergence.

Leveraging Experiment 1, we further emphasize the superior robustness of our proposed approach in estimating the JS divergence, regardless of variations in distribution separation. Fig. 6 illustrates the relation between estimated divergences and ground truth values as the separation between two 4-dimensional multivariate Gaussian distributions increases. As observed, the relation for the JS divergence remains almost linear until it approaches 1. While the relation appears less linear for values near 1, it is essential to note that the primary application of this discriminator lies in comparing similar distributions. Therefore, the focus should be on the performance of low JS divergences. Meanwhile, the KL divergence exhibits a nonlinear relation starting from the first comparisons, with the estimated KL deviating from the ground truth probably because of bias introduced by the discriminator.

B. EXPERIMENT 2

This experiment investigates the performance of the proposed divergence estimation method in Gaussian mixture distributions. These distributions offer a higher complexity level than those used in Experiment 1. Each Gaussian mixture distribution comprises two independent isotropic Gaussian

TABLE 3. Impact of training and validating samples on $D_{\mathbb{KL}}$ and $D_{\mathbb{JS}}$ estimation for Experiment 1.

M	L	Analytical $D_{\mathbb{KL}}$	MC Estimated $D_{\mathbb{KL}}$	Discriminator Estimated $D_{\mathbb{KL}}$	MC Estimated $D_{\mathbb{JS}}$	Discriminator Estimated $D_{\mathbb{JS}}$
20	20	1.035	1.168 ± 0.203	0.802 ± 0.152	0.320 ± 0.105	0.213 ± 0.072
	200		1.026 ± 0.221	0.709 ± 0.248	0.301 ± 0.055	0.200 ± 0.038
	2000		1.024 ± 0.193	0.665 ± 0.084	0.288 ± 0.047	0.209 ± 0.028
200	20		1.221 ± 0.422	0.991 ± 0.520	0.320 ± 0.055	0.289 ± 0.053
	200		1.114 ± 0.257	1.099 ± 0.227	0.313 ± 0.055	0.294 ± 0.049
	2000		1.026 ± 0.191	0.955 ± 0.101	0.289 ± 0.038	0.267 ± 0.038
2000	20		0.841 ± 0.190	0.833 ± 0.229	0.219 ± 0.113	0.218 ± 0.111
	200		1.004 ± 0.190	0.978 ± 0.185	0.300 ± 0.047	0.290 ± 0.049
	2000		1.055 ± 0.212	1.060 ± 0.200	0.299 ± 0.051	0.293 ± 0.050

Analytical $D_{\mathbb{KL}}$ along with MC $D_{\mathbb{KL}}$ and $D_{\mathbb{JS}}$ estimations, as well as proposed discriminator estimations for both divergences. Results are displayed for various combinations of training samples M and validation samples L. There is a clear correlation between the number of samples used and the estimation error.

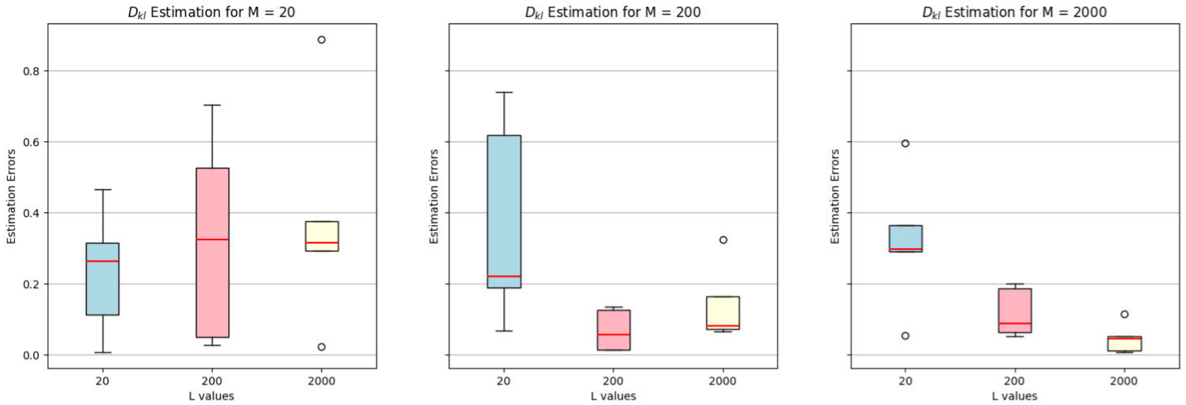


FIGURE 3. Estimation error representation for $D_{\mathbb{KL}}$ in Experiment 1. Results are shown for different combinations of training sample sizes M and validation sample sizes L. As expected, a decrease and precision in the error is observed with increasing values of M and L.

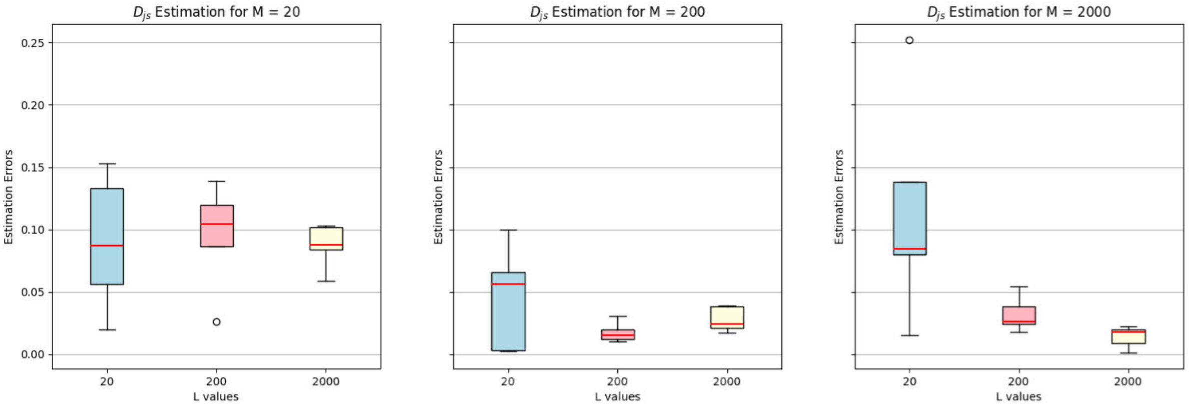


FIGURE 4. Estimation error representation for $D_{\mathbb{JS}}$ in Experiment 1. Results are shown for different combinations of training sample sizes M and validation sample sizes L. As expected, a decrease and precision in the error is observed with increasing values of M and L.

components with distinct mixing probabilities (weights assigned to each component within the mixture). Unlike the prior case, where analytical ground truth for divergence was obtainable, we rely solely on MC approximations as ground truth here and in the following experiment. Similar results were obtained for this experiment compared to the first.

Tables 4 and 5 focus on a practical range for training sample sizes M and validation sample sizes L (both set to $200, 2000$) and present the estimated divergences obtained using our proposed method compared to MC estimates. The results demonstrate that, for both divergences, the proposed method achieves results comparable to the MC estimates. These findings further support the previously observed correlation between sample size and estimation accuracy. Additionally, we observe a narrowing of confidence intervals with increasing sample sizes for both divergence measures. This trend indicates that the estimation becomes progressively more precise as the amount of available data increases. Overall, this experiment's results underscore the proposed method's effectiveness in estimating divergences

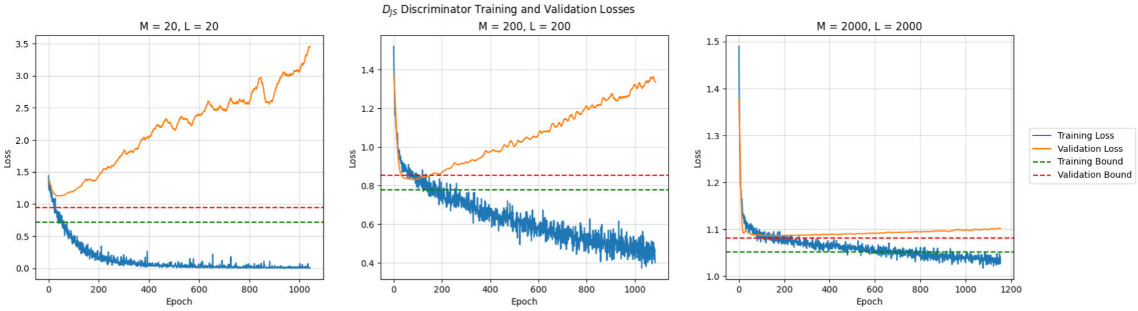


FIGURE 5. Discriminator loss curves for Experiment 1. The loss curves show a clear overfitting due to low sample sizes. Green and red dashed lines represent theoretical convergence values.

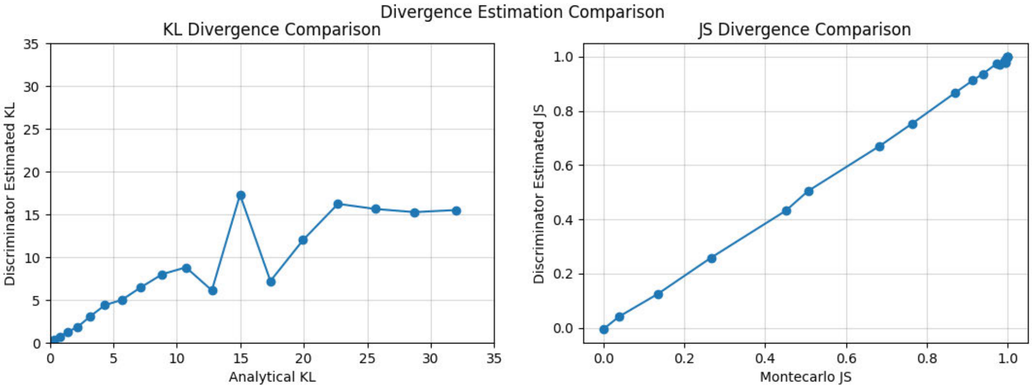


FIGURE 6. Relation between ground truth and estimated divergences as distribution separation increases in Experiment 1.

TABLE 4. Impact of training and validating samples on $D_{\mathbb{KL}}$ estimation for Experiment 2.

M	L	MC Estimated $D_{\mathbb{KL}}$	Discriminator Estimated $D_{\mathbb{KL}}$
200	200	2.669 ± 0.118	2.686 ± 0.294
	2000	2.887 ± 0.045	3.037 ± 0.517
2000	200	2.710 ± 0.100	2.628 ± 0.357
	2000	2.850 ± 0.071	2.913 ± 0.129

MC and proposed method $D_{\mathbb{KL}}$ estimation for various combinations of training and validation samples M and L. The results highlight a clear correlation between the number of samples used and the estimation error.

TABLE 5. Impact of training and validating samples on $D_{\mathbb{JS}}$ estimation for Experiment 2.

M	L	MC Estimated $D_{\mathbb{JS}}$	Discriminator Estimated $D_{\mathbb{JS}}$
200	200	0.392 ± 0.013	0.374 ± 0.020
	2000	0.428 ± 0.009	0.412 ± 0.009
2000	200	0.415 ± 0.020	0.412 ± 0.019
	2000	0.423 ± 0.014	0.420 ± 0.013

MC and proposed method $D_{\mathbb{JS}}$ estimation for various combinations of training and validation samples M and L. The results highlight a clear correlation between the number of samples used and the estimation error.

for more intricate distributions, such as Gaussian mixtures. The remaining analysis can be found in the Appendix.

C. EXPERIMENT 3

The following experiments introduce the concept of a generative process, where synthetic data are generated based on N samples from a particular data distribution. We investigate the influence of the quality of the GM in terms of the number of samples used to generate. This specific experiment uses a Gaussian mixture with two components as the real distribution. We employ a Gaussian mixture model (GMM) as the GM. The trained GMM then generates synthetic data that serve as an approximation of the real data. Finally, we estimate the divergence using our approach with a sufficient number of samples to achieve a low confidence interval and compare it to the MC simulation estimation. We analyze the behavior of the GMM under different training configurations, varying the number of training samples N. Since previous experiments demonstrated that a high number of samples for both M and L achieve better divergence estimations, we have fixed $M = 2000$ and $L = 2000$ for this experiment. We compare the estimated divergence errors for this combination of training samples M and validation L used by the divergence estimator.

Fig. 7 shows the estimated divergences as N increases. The results demonstrate a clear correlation between the size of the training data for the GMM, N, and the estimated divergence errors. When the GMM is poorly trained due to a limited number of samples ($N = 10$), it cannot effectively capture the underlying data patterns. This leads to generated data that significantly deviate from the real distribution. Consequently, the real divergence values increase significantly, and a large estimation error occurs, particularly for unbounded $D_{\mathbb{KL}}$. $D_{\mathbb{JS}}$, which is inherently bounded between 0 and 1, exhibits

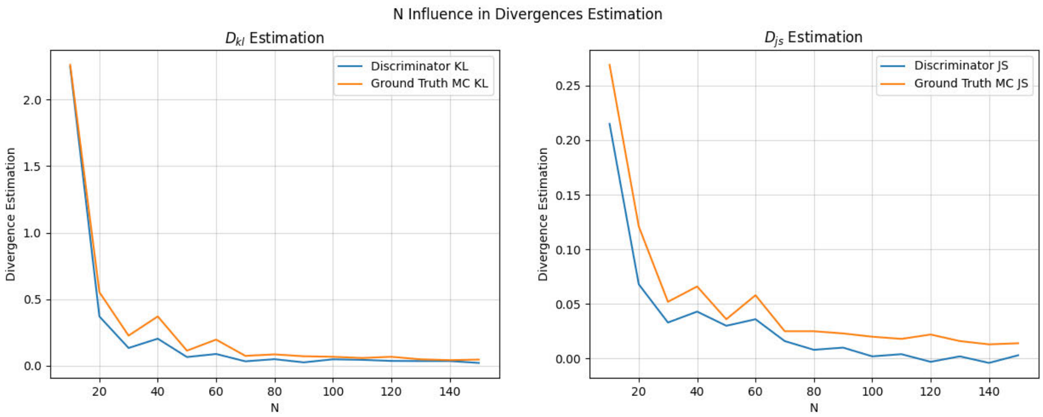


FIGURE 7. Estimated $D_{\mathbb{KL}}$ and $D_{\mathbb{JS}}$ compared to ground truth values for varying sample sizes of N in Experiment 3. Note the independent y-axes for each divergence measure due to their inherent scale differences, with $D_{\mathbb{JS}}$ exhibiting lower values than $D_{\mathbb{KL}}$.

TABLE 6. Comparison of generative models for real-world data.

CTGAN $D_{\mathbb{KL}}$	VAE $D_{\mathbb{KL}}$	CTGAN $D_{\mathbb{JS}}$	VAE $D_{\mathbb{JS}}$
0.342 ± 0.016	0.185 ± 0.030	0.138 ± 0.002	0.099 ± 0.002

Similarity comparison between real data and their synthetic counterparts generated by each model using estimated divergences, $D_{\mathbb{KL}}$ and $D_{\mathbb{JS}}$. Lower values in both metrics indicate a greater similarity between real and synthetic data.

less extreme error values. As the number of training samples for the GMM increases ($N = 200$ and $N = 2000$), the generated data become more representative of the real distribution. This results in lower divergence values and reduced estimation errors for $D_{\mathbb{KL}}$ and $D_{\mathbb{JS}}$. This finding supports the validity of our proposed validation technique for generative processes involving Gaussian mixture distributions.

D. EXPERIMENT 4

The final experiment introduces a more realistic setting by incorporating a real-world dataset along with its synthetic counterpart generated using a GM. This scenario retains the concept of a generative process, but the target distribution becomes the real-world data itself. The dataset chosen is Adult. As described in [26], it contains information on $32,561$ individuals used to predict their annual income exceeding \$50,000. It represents an extract from the 1994 U.S. Census database [27]. For our experiments and after the previous analysis was done, we used a subset of $10,000$ samples obtained from [28]. This dataset comprises 14 features encompassing various data types, including categorical, binary, and integer values. Based on the previous analysis of the influence of the sample size, we fixed the hyperparameters for this experiment at $N = 10000$, $M = 7500$, and $L = 1000$. To evaluate GMs, we compared the performance of two state-of-the-art approaches: the widely used CTGAN [12] and the VAE-based generator proposed in [25] (VAE), which achieved superior results based on several validation criteria.

Table 6 confirms this, obtaining lower divergence results for [25]. To explain why the VAE may outperform the CTGAN in our experiments, we refer to [25], which compares the VAE generator with CTGAN and TVAE using metrics similar to ours. The results support our observations. We suggest that VAEs perform better due to their stability in managing complex data distributions, unlike GANs, which struggle with convergence. This stability in VAEs contributes to their enhanced ability to generate more realistic synthetic data, highlighting significant differences in model architecture and distribution handling. This experiment focuses on evaluating the ability of the divergence estimator to assess the similarity between the real data and its synthetic counterpart. This experiment provides valuable insights into the practical application of the proposed method in real-world scenarios where measuring the similarity between real and synthetic data is crucial.

V. CONCLUSION

This research proposes a novel and practical approach for validating synthetic tabular data generated by various models. The core of this method lies in using a divergence estimator based on a probabilistic classifier to capture the discrepancies between the real and synthetic data distributions. This approach overcomes the limitations associated with traditional marginal divergence comparisons by considering the joint distribution. While marginal comparisons assess the similarity of individual features between real and synthetic data, they can be misleading. Even if the marginal distributions of each feature appear similar, the joint distribution, which captures the relations between features, may differ significantly. This can lead to unrealistic synthetic data, where individual features appear plausible, but their co-occurrences do not represent the real data. By considering the joint distribution, our proposed method provides a more comprehensive assessment of the quality of synthetic tabular data.

The efficacy of the proposed method is comprehensively evaluated through a series of experiments with progressively increasing complexity. The initial phase establishes a solid

TABLE 7. Impact of training and validating samples on $D_{\mathbb{KL}}$ and $D_{\mathbb{JS}}$ estimation for Experiment 2.

M	L	MC Estimated $D_{\mathbb{KL}}$	Discriminator Estimated $D_{\mathbb{KL}}$	MC Estimated $D_{\mathbb{JS}}$	Discriminator Estimated $D_{\mathbb{JS}}$
20	20	3.245 ± 0.820	1.832 ± 0.912	0.490 ± 0.083	0.371 ± 0.078
	200	2.699 ± 0.095	1.193 ± 0.430	0.439 ± 0.017	0.328 ± 0.058
	2000	2.833 ± 0.032	1.512 ± 0.381	0.423 ± 0.006	0.331 ± 0.039
200	20	2.639 ± 0.404	3.101 ± 0.779	0.405 ± 0.066	0.340 ± 0.114
	200	2.669 ± 0.118	2.686 ± 0.294	0.392 ± 0.013	0.374 ± 0.020
	2000	2.887 ± 0.045	3.037 ± 0.517	2.887 ± 0.045	3.037 ± 0.517
2000	20	2.974 ± 0.626	2.771 ± 0.751	0.481 ± 0.090	0.474 ± 0.096
	200	2.710 ± 0.100	2.628 ± 0.357	0.415 ± 0.020	0.412 ± 0.019
	2000	2.850 ± 0.071	2.913 ± 0.129	0.423 ± 0.014	0.420 ± 0.013

Analytical $D_{\mathbb{KL}}$ along with MC $D_{\mathbb{KL}}$ and $D_{\mathbb{JS}}$ estimations, as well as proposed discriminator estimations for both divergences. Results are displayed for various combinations of training samples M and validation samples L. There is a clear correlation between the number of samples used and the estimation error.

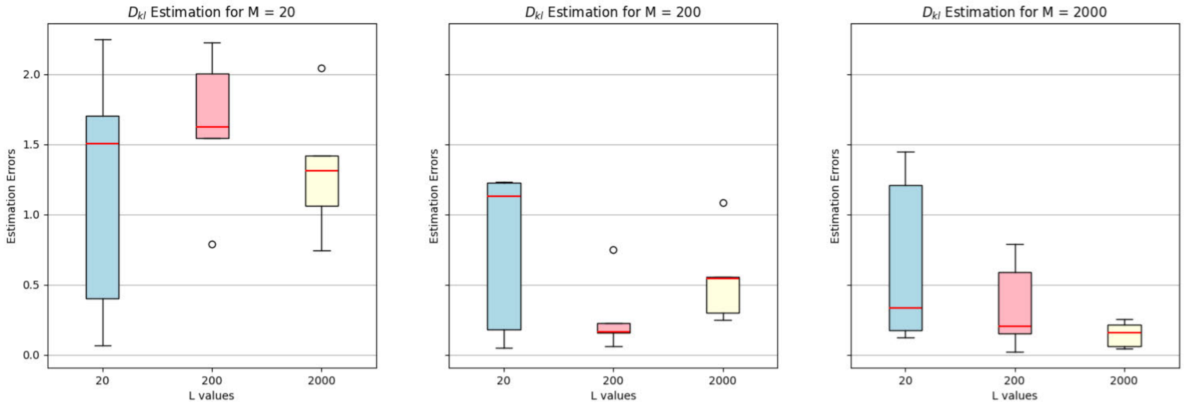


FIGURE 8. Estimation error representation for $D_{\mathbb{KL}}$ in Experiment 2. Results are shown for different combinations of training sample sizes M and validation sample sizes L. As expected, a decrease and precision in the error is observed with increasing values of M and L.

foundation by analyzing the performance in controlled scenarios with well-defined theoretical distributions. The results demonstrate that the accuracy of divergence estimates is highly dependent on the amount of training data available for both the GM and the divergence estimator network. Subsequent experiments explore more intricate scenarios involving Gaussian mixture distributions and real-world datasets. The findings consistently support the effectiveness of the proposed method in approximating the true divergences. Our emphasis in this paper has been to show the advantages of using divergences as synthetic data validation for tabular data.

This research offers significant contributions that extend beyond the specific tabular data validation domain. The proposed methodology facilitates better validation practices for various fields dependent on GMs. Its key strength lies in capturing complex relations between different data distributions, leading to more robust and reliable validation processes. In addition to the positive results, the study also highlights the importance of the quality of the GM. When it is inadequately trained due to insufficient data, it can significantly impact the accuracy of the divergence estimation. This emphasizes the need to consider GM training procedures carefully to ensure reliable validation results.

Several promising avenues exist for future research. One direction involves exploring the potential to extend the proposed approach to more complex data structures, such as images or time series data. Additionally, investigating the integration of this validation technique within GM training pipelines could enable the development of self-improving GMs that can automatically adjust their parameters to generate data that closely resemble the target distribution. Moreover, addressing the impact of changing the assumption of prior probabilities for the source distributions to be compared, where the number of real and synthetic samples may vary, presents a significant challenge. Techniques to mitigate the effects of unbalanced classes in the regression framework could be explored. These techniques may include adaptive sampling strategies, class weighting, or methods specifically tailored to handle imbalanced data distributions. Implementing such approaches could enhance the robustness and applicability of the validation technique across diverse datasets and real-world applications. Finally, developing a robust methodology to estimate divergences when limited samples are available presents a crucial challenge. Addressing this challenge would further enhance the versatility and practicality of the proposed validation technique in real-world scenarios with restricted data availability. Furthermore, it would be beneficial to investigate alternative density

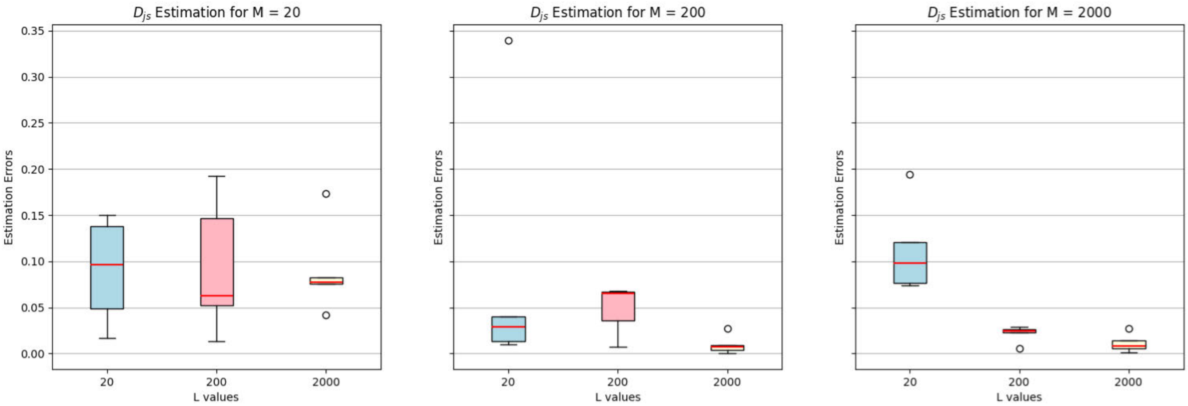


FIGURE 9. Estimation error representation for $D_{\mathbb{JS}}$ in Experiment 2. Results are shown for different combinations of training sample sizes M and validation sample sizes L. As expected, a decrease and precision in the error is observed with increasing values of M and L.

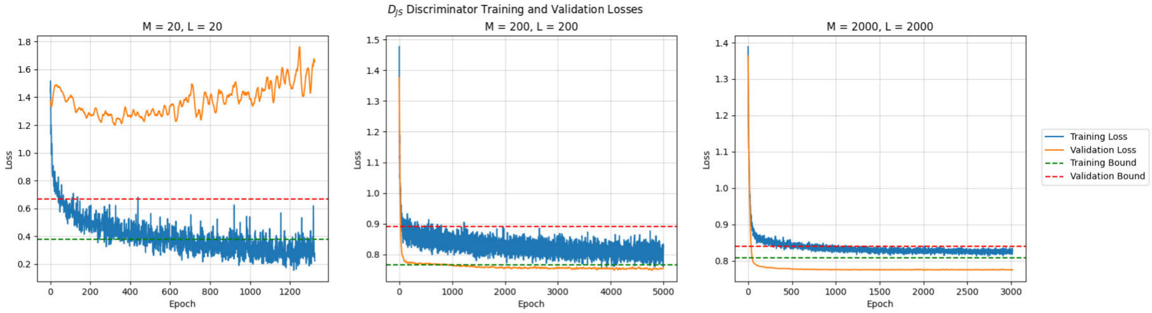


FIGURE 10. Discriminator loss curves for Experiment 2. The loss curves show a clear overfitting due to low sample sizes. Green and red dashed lines represent theoretical convergence values.

ratio estimation techniques, such as those presented in [29] and [30]. These methods may offer advantages regarding sample size requirements, estimation error, or other relevant properties.

APPENDIX

Leveraging the validation process established in Experiment 1, Experiment 2 employs the same methodology. The remaining analyses for Experiment 2 are presented as follows.

Table 7 extends the findings of the previous experiment by investigating the performance of our proposed approach with a restricted number of training and validation samples for the divergence estimator. The results indicate a trend of convergence towards the known ground truth divergence values as the number of samples utilized for both training and validation is progressively increased.

Figs. 8 and 9 provide further substantiation for this observed trend.

Additionally, Fig. 10 delves into the behavior of the discriminator's loss function across various sample size combinations.

ACKNOWLEDGMENT

The authors declare that they have no known competing financial interests or personal relations that could have appeared to influence the work reported in this article.

(Patricia A. Apellániz and Ana Jiménez contributed equally to this work.)

REFERENCES

[1] A. Jadon and S. Kumar, "Leveraging generative AI models for synthetic data generation in healthcare: Balancing research and privacy," in *Proc. Int. Conf. Smart Appl., Commun. Netw. (SmartNets)*, Jul. 2023, pp. 1–4.

[2] H. Caliskan, O. F. Yayla, and Y. Genc, "A comparative analysis of synthetic data generation with VAE and CTGAN models on financial credit loan offer data," in *Proc. 8th Int. Conf. Comput. Sci. Eng. (UBMK)*, Sep. 2023, pp. 212–217.

[3] D. Talwar, S. Guruswamy, N. Ravipati, and M. Eirinaki, "Evaluating validity of synthetic data in perception tasks for autonomous vehicles," in *Proc. IEEE Int. Conf. Artif. Intell. Test. (AITest)*, Aug. 2020, pp. 73–80.

[4] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training GANs," 2016, *arXiv:1606.03498*.

[5] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local Nash equilibrium," 2017, *arXiv:1706.08500*.

[6] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "BERTScore: Evaluating text generation with BERT," 2019, *arXiv:1904.09675*.

[7] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proc. 40th Annu. Meeting Assoc. Comput. Linguistics*. Baltimore, MD, USA: Association for Computational Linguistics, 3115, pp. 311–318.

[8] Z. Che, Y. Cheng, S. Zhai, Z. Sun, and Y. Liu, "Boosting deep learning risk prediction with generative adversarial networks for electronic health records," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, Nov. 2017, pp. 787–792.

[9] F. Yang, Z. Yu, Y. Liang, X. Gan, K. Lin, Q. Zou, and Y. Zeng, "Grouped correlational generative adversarial networks for discrete electronic health records," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Nov. 2019, pp. 906–913.

[10] S. Rashidian, F. Wang, R. Moffitt, V. Garcia, A. Dutt, W. Chang, V. Pandya, J. Hajagos, M. Saltz, and J. Saltz, "SMOOTH-GAN: Towards sharp and smooth synthetic EHR data generation," in *Artificial Intelligence in Medicine*, M. Michalowski and R. Moskovitch, Eds., Cham, Switzerland: Springer, 1007, pp. 37–48.

[11] A. Goncalves, P. Ray, B. Soper, J. Stevens, L. Coyle, and A. P. Sales, "Generation and evaluation of synthetic patient data," *BMC Med. Res. Methodol.*, vol. 20, no. 1, pp. 1–40, Dec. 2020.

[12] L. Xu, M. Skoularidou, A. Cuesta-Infante, and K. Veeramachaneni, "Modeling tabular data using conditional GAN," 2019, *arXiv:1907.00503*.

[13] S. Dash, A. Yale, I. Guyon, and K. P. Bennett, "Medical time-series data generation using generative adversarial networks," in *Artificial Intelligence in Medicine*, M. Michalowski and R. Moskovitch, Eds., Cham, Switzerland: Springer, 2020, pp. 382–391.

[14] J. Snoke, G. Raab, B. Nowok, C. Dibben, and A. Slavkovic, "General and specific utility measures for synthetic data," 2016, *arXiv:1604.06651*.

[15] M. Hernadez, G. Epelde, A. Alberdi, R. Cilla, and D. Rankin, "Synthetic tabular data evaluation in the health domain covering resemblance, utility, and privacy dimensions," *Methods Inf. Med.*, vol. 62, pp. e19–e38, Jun. 2023.

[16] A. Hu, R. Xie, Z. Lu, A. Hu, and M. Xue, "TableGAN-MCA: Evaluating membership collisions of GAN-synthesized tabular data releasing," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, 2021, pp. 2096–2112.

[17] S. Wang, C. Rudolph, S. Nepal, M. Grobler, and S. Chen, "PART-GAN: Privacy-preserving time-series sharing," in *Artificial Neural Networks and Machine Learning—ICANN 2020*, I. Farkaš, P. Masulli, and S. Wermter, Eds., Cham, Switzerland: Springer, 1007, pp. 578–593.

[18] Z. Zhao, A. Kunar, R. Birke, and L. Y. Chen, "CTAB-GAN: Effective table data synthesizing," in *Proc. Asian Conf. Mach. Learn.*, 2021, pp. 97–112.

[19] A. Tucker, Z. Wang, Y. Rotalinti, and P. Myles, "Generating high-fidelity synthetic patient data for assessing machine learning healthcare software," *NPJ Digit. Med.*, vol. 3, no. 1, pp. 1–13, Nov. 2020.

[20] W.-X. Liu et al., "Effect of age as a continuous variable on survival outcomes and treatment selection in patients with extranodal nasal-type NK/T-cell lymphoma from the China lymphoma collaborative group (CLCG)," *Aging*, vol. 11, no. 19, pp. 8463–8473, Oct. 2019.

[21] M. S. Walia, T. Brendan, and S. Mckeever, "Synthesising tabular datasets using Wasserstein conditional GANs with gradient penalty (WCGAN-GP)," in *Proc. 28th Irish Conf. Artif. Intell. Cogn. Sci. (AICS)*, Dublin, Ireland, 2020.

[22] L. C. Tiao, "Density ratio estimation for KL divergence minimization between implicit distributions," Univ. Sydney, Sydney, NSW, Australia, 2018. [Online]. Available: https://tiao.io/post/density-ratio-estimation-for-kl-divergence-minimization-between-implicit-distributions/

[23] E. Cech and M. Katetov, *Point Sets*. San Francisco, CA, USA: Academia, 1969.

[24] V. N. Vapnik, "Setting of the learning problem," in *The Nature of Statistical Learning Theory*, 2nd ed., New York, NY, USA: Springer, 2000, ch. 1, pp. 17–33.

[25] P. A. Apellániz, J. Parras, and S. Zazo, "An improved tabular data generator with VAE-GMM integration," 2024, *arXiv:2404.08434*.

[26] B. Becker and R. Kohavi, "Adult," UCI Mach. Learn. Repository, Univ. California, Berkeley, CA, USA, 1996, doi: 10.24432/C5XW20.

[27] R. Kohavi, "Census income," UCI Mach. Learn. Repository, Univ. California, Berkeley, CA, USA, 1996, doi: 10.24432/C5GP7S.

[28] N. Patki, R. Wedge, and K. Veeramachaneni, "The synthetic data vault," in *Proc. IEEE Int. Conf. Data Sci. Adv. Analytics (DSAA)*, Oct. 2016, pp. 399–410.

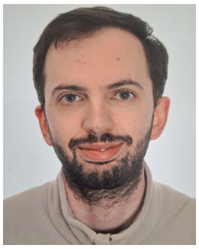
[29] K. Choi, C. Meng, Y. Song, and S. Ermon, "Density ratio estimation via infinitesimal classification," in *Proc. 25th Int. Conf. Artif. Intell. Statist.*, vol. 151, G. Camps-Valls, F. J. R. Ruiz, and I. Valera, Eds., May 2022, pp. 2552–2573.

[30] B. Rhodes, K. Xu, and M. U. Gutmann, "Telescoping density-ratio estimation," 2020, *arXiv:2006.12204*.

PATRICIA A. APELLÁNIZ received the bachelor's degree in telecommunication engineering from the Universidad Autónoma de Madrid, Madrid, Spain, in 2018, and the master's degree in telecommunication engineering from the Universidad Politécnica de Madrid (UPM), Madrid, where she is currently pursuing the Ph.D. degree in deep learning algorithms for medical applications. She is also a Researcher with UPM. Her research interest includes audio and image signal processing to different deep learning applications.

ANA JIMÉNEZ received the bachelor's and M.Sc. degrees in signal processing for big data and the M.Sc. degree in telecommunications engineering from the Universidad Politecnica de Madrid (UPM). Her research interest includes artificial intelligence applied to the biomedical field.

BORJA ARROYO GALENDE received the degree in natural environment engineering and the degree in computer engineering from the Universidad Politécnica de Madrid (UPM), and the master's degree in engineering and data science from the U.N.E.D. He is currently developing his thesis in the context of machine learning in federated environments. He is also a Software Developer/Architect with UPM.



JUAN PARRAS received the B.S. degree in telecommunications engineering from the Universidad de Jaén, in 2014, and the M.Sc. and Ph.D. degrees in telecommunications engineering from the Universidad Politécnica de Madrid (UPM), in 2016 and 2020, respectively. He is currently an Assistant Professor with UPM. His research interests include deep generative models, deep reinforcement learning, game theory, and optimization with health and communications applications.

SANTIAGO ZAZO is currently a Dr. Engineer with the Universidad Politécnica de Madrid (UPM), in 1995. In 1998, he joined UPM, where he is currently a Full Professor in signal theory and communications. More recently, he has been mostly focused on deep learning with medical applications, distributed optimization, game theory, and reinforcement learning. He is the author/co-author of more than 40 journal articles and about 200 conference papers. His main research interest includes signal processing.