Distributional learning via deep copula density weighting

Anonymous author(s)

Abstract

Distribution learning aims at estimating the conditional distribution of a response variable given covariates, which is fundamental in generative learning. However, most existing methods either are not suitable for high-dimensional covariates, or are computationally challenging and not interpretable. We introduce a new perspective of distributional learning via a copula density weighting, where the weighting function is learned via deep neural networks. The proposed copula density weighting is probabilistically interpretable, computationally efficient, and versatile for both synthetic value generation and hot deck imputation. Theoretically, we establish convergence rates of the deep copula density estimator and the conditional distribution estimator based on the U-process techniques. Simulation results demonstrate that our method achieved competitive performance in conditional distribution estimation, with significantly reduced training time compared to popular generative models.

Keywords: Conditional generative learning; Copula density; Deep neural networks; High-dimensional covariates; Hot-deck imputations; U-statistics.

Mathematics Subject Classification (2020): 62E10

1 Introduction

Let $Y \in \mathbb{R}$ be a response variable and $X \in \mathbb{R}^{d_x}$ be a covariate vector, possibly high-dimensional. While standard regression focuses on the conditional expectation, in various applications, it is more desirable to learn the conditional distribution of $Y|X$ to offer a comprehensive depiction of the inherent variability. In this paper, we propose a novel copula density weighting method for the conditional distribution learning, which is interpretable, generative and easy to train.

A variety of methods in statistics and machine learning have been developed for distributional learning, which may be classified into three categories. The first one aims at directly modeling the conditional distribution function either nonparametrically or parametrically, such as the kernel smoothing (Hall et al., 1999) and the conditional transformation model (Hothorn et al., 2014). The second category is to estimate a cascade of quantile regressions, such as Koenker (2005) and Meinshausen (2006). However, parametric methods often face the risk of model misspecification and classic nonparametric methods suffer from the curse of dimensionality. In addition, the quantile regression approach often needs to address the quantile crossing problem (Chernozhukov et al., 2010). A further limitation of the aforementioned methods is the difficulty of the imputation and generation of Y from the estimated conditional distribution, especially for high-dimensional X, which are needed in various applications such as missing data and survey sampling (Kim and Shao, 2021). The third category of distributional learning is via generative models, such as generative adversarial networks (GAN; Goodfellow et al., 2014) and diffusion

models (Ho et al., 2020), which characterise the distribution via a generating process. However, most existing generative models are *not interpretable and computationally challenging*. For instance, the GAN involves solving a min-max optimization known for its training instability, and the diffusion models require iterative sampling over a stochastic process.

To overcome these limitations, we propose a copula density weighting method. Different from existing approaches, which typically estimate the conditional distribution from some goodness-of-fit criterion, our method is motivated by the observation that the conditional distribution function $F_{Y|X}(y|x)$ can be written as $\mathrm{E}\{\mathbb{1}(Y \leq y)|X = x\} = \mathrm{E}\{\mathbb{1}(Y \leq y)\rho_0(Y, x)\}$, where the expectations on the left and right hand sides are taken with respect to the conditional distribution $Y|X = x$ and the marginal distribution of Y, respectively, and ρ_0 is the ratio function of the densities of $P_{X,Y}$ and $P_X \otimes P_Y$, which corresponds to the copula density function of the pair (Y, X) indicating the dependence intensity between X and Y.

A central piece of our framework is the estimation of a copula density $\rho_0(y, x)$. We propose to identify ρ_0 with a dual representation of ϕ-divergences between $P_{X,Y}$ and $P_X \otimes P_Y$, enabling us to estimate ρ_0 via an M-estimation. Deep neural networks are used to model the copula density due to their high expressive capacity. With the estimated copula density, we obtain not only an explicit estimator of the conditional distribution, but also a convenient conditional sampler for either synthetic value generations or hot deck imputations (Kim and Fuller, 2004). More challenging than the existing deep neural network estimation, our estimator is derived via a U-statistic loss. We make theoretical contributions by establishing the convergence rates of the deep copula density estimator and the conditional sampling estimator, and revealing their low-dimensional adaptivity with the U-process techniques and the theory of the deep neural network approximation.

The advantages of the copula density weighting are threefold. First, as the copula density describes the dependence between Y and X, the method assigns greater weights to values of Y exhibiting stronger tendencies to occur jointly with $X = x$. Therefore, it approach offers enhanced probabilistic interpretability compared with most existing methods focusing on fitting performances. Second, the copula density weighting can be conveniently employed as a generative method to produce synthetic values of Y and avoids the computational cost of sampling procedures such as rejection sampling. Moreover, the proposed generation strategy can have desirable properties of the hot deck imputation as discussed in Andridge and Little (2010), while existing generative models are not capable of. Another advantage is that its computation only requires optimizing a standard multi-layer perception neural network that models the copula density function, whose training cost can be significantly reduced compared with other generative models.

The paper is organized as follows. Section 2 introduces the proposed copula density weighting strategy for the conditional distribution estimation. Section 3 details a variational divergence minimization method for copula density estimation. Section 4 presents a two-step procedure for the conditional generation with the copula density weighting. Section 5 reports results from simulations that evaluate the performance of the proposed method and compare it with other generative models.

2 Conditional distribution estimation with copula density weighting

Suppose that the observed data $\{Z_i = (X_i, Y_i)\}_{i=1}^n$ are independent copies of $(X, Y) = Z$, where $X \in \mathcal{X} \subset \mathbb{R}^d$ and $Y \in \mathcal{Y} \subset \mathbb{R}$. We denote F_Z and f_Z as its cumulative distribution function and density function, respectively.

Different from most existing methods, our method is inspired by the principle of importance weighting. The main reason that the estimation of $F_{Y|X=x}(y)$ is more challenging than that of $F_Y(y)$ is because the latter is essentially $\mathrm{E}\{\mathbb{1}(Y \le y)\}$ and can be straightforwardly estimated by $n^{-1}\sum_{i=1}^n \mathbb{1}(Y_i \le y)$. In contrast, few repeated observations following $F_{Y|X=x}$ are available at a given x. To address such a challenge, for every $(x, y) \in \mathcal{X} \times \mathcal{Y}$, we transform $F_{Y|X=x}(y)$ as

$$F_{Y|X=x}(y) = \mathrm{E}\{\mathbb{1}(Y \le y)|X = x\} = \mathrm{E}_Y\left\{\mathbb{1}(Y \le y)\frac{f_{(X,Y)}(x, Y)}{f_X(x)f_Y(Y)}\right\}, \tag{1}$$

where the expectation on the right-hand side is taken with respect to the F_Y. Define

$$\rho_0(x, y) = \frac{f_{(X,Y)}(x, y)}{f_X(x)f_Y(y)} \quad \text{for every } (x, y) \in \mathcal{X} \times \mathcal{Y}.$$

Then, the conditional distribution $F_{Y|X=x}(y)$ can be transformed into an expectation form $\mathrm{E}\{\mathbb{1}(Y \le y)\rho_0(x, Y)\}$, enabling one to directly sample averages to estimate the right-hand side of (1). Suppose that $\widehat{\rho}$ is an estimator of ρ_0, then $F_{Y|X=x}(y)$ can be simply estimated with

$$\widehat{F}_{Y|X=x}(y; \widehat{\rho}) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}(Y_i \le y)\widehat{\rho}(x, Y_i). \tag{2}$$

One appealing property of (2) lies in its interpretability. Let $C(u, v)$ be the copula function for (X, Y) with the density function $c(u, v) = \partial^2 C(u, v)/\partial u \partial v$. Sklar's theorem (Nelsen, 2006) implies that $F_{(X,Y)}(x, y) = C(F_X(x), F_Y(y))$ for every (x, y). Differentiating both sides yields

$$f_{(X,Y)}(x, y) = \frac{\partial^2 F_{(X,Y)}(x, y)}{\partial x \partial y} = \frac{\partial^2 C(u, v)}{\partial u \partial v}\frac{\partial F_X(x)}{\partial x}\frac{\partial F_y(x)}{\partial y} = c(u, v)f_X(x)f_Y(y),$$

where $u = F_X(x)$ and $v = F_Y(y)$. This means that $\rho_0(x, y)$ essentially corresponds to the copula density function $c(u, v)$. Since a high value of the copula density indicates strong dependency between X and Y, the estimator (2) assigns a larger weight to Y_i when the joint occurrence of $X = x$ and $Y = Y_i$ is more probable. Compared with the locally smoothing methods, whose weights only depend on the distance between X_i and x, the copula weighting function is determined by the dependence intensity between Y and X, enabling the proposed estimator to have a more probabilistic interpretation. In addition, instead of depending on the distance between X_i and x, the weight function ρ_0 can be learnt with more flexible and expressive machine learning methods such as the deep neural networks, as we will introduce in the following section.

3 Copula density estimation with deep neural networks

In this section, we propose a variational divergence minimization method based on neural networks for the copula density estimation and establish its theoretical properties.

To present our method, we first introduce the notion of ϕ-divergences. Let P and Q be any two distributions with P being absolutely continuous with respect to Q. The ϕ-divergence between P and Q is defined as $D_\phi(P\|Q) = \mathrm{E}_Q\left\{\phi\left(\mathsf{d}P/\mathsf{d}Q\right)\right\}$, where ϕ is convex and satisfies $\phi(1) = 0$, and density ratio between P and Q. Let $\phi_*(t) = \sup_{u\in \mathrm{Dom}_\phi}\{ut - \phi(u)\}$ be the Fenchel dual of ϕ. Then, $D_\phi(P\|Q)$ has a variational representation introduced in Nguyen et al. (2010).

Lemma 3.1 *Let $\mathcal{F}$ be the class of non-negative functions, then*

$$D_\phi(P\|Q) \geq \sup_{f\in\mathcal{F}} \left(\mathrm{E}_P\{f(Z)\} - \mathrm{E}_Q[\phi_*\{f(Z)\}]\right),$$

and the supremum is attained at $f_0 = \phi'(\mathsf{d}P/\mathsf{d}Q)$.

According to Lemma 3.1, the supremum of $\mathrm{E}_P\{f(Z)\} - \mathrm{E}_Q[\phi_*\{f(Z)\}]$ has a one-to-one correspondence to the density ratio function $r_0 := \mathsf{d}P/\mathsf{d}Q$. More explicitly, let $\ell_1(r) = \phi_*\{\phi'(r)\}$ and $\ell_2(r) = \phi'(r)$, then r_0 can be identified as the solution to the following optimization problem:

$$r_0 = \arg\min_{r\in\mathcal{F}} L(r) \quad \text{with} \quad L(r) = \mathrm{E}_Q\{\ell_1(r)\} - \mathrm{E}_P\{\ell_2(r)\}, \tag{3}$$

where $\mathcal{F}$ contains all non-negative functions. Taking $P = P_{(X,Y)}$ and $Q = P_X \otimes P_Y$. In the light of (3), the copula density function ρ_0 can be identified as

$$\rho_0 = \arg\min_{\rho\in\mathcal{F}} L(\rho) \quad \text{with} \quad L(\rho) = \mathrm{E}_{P_{\boldsymbol{X}}\otimes P_{\boldsymbol{Y}}}\{\ell_1(\rho)\} - \mathrm{E}_{P_{(X,Y)}}\{\ell_2(\rho)\}. \tag{4}$$

While similar methods have been used in Sugiyama et al. (2012) and Kato and Teshima (2021) for density ratio estimation, a challenge in (4) is that there are no direct realizations for the decoupled pair $(X, Y) \sim P_X \otimes P_Y$. To address the challenge, we introduce a U-statistic loss defined as

$$\widehat{L}_n(\rho) = \binom{n}{2}^{-1} \sum_{1\leq i<j\leq n} h(Z_i, Z_j; \rho), \tag{5}$$

where for every $z_1 = (x_1, y_1)$ and $z_2 = (x_2, y_2)$, the symmetric kernel h is defined as

$$h(z_1, z_2; \rho) = \frac{1}{2}\left[\ell_1\{\rho(x_1, y_2)\} + \ell_1\{\rho(x_2, y_1)\} - \ell_2\{\rho(x_1, y_1)\} - \ell_2\{\rho(x_2, y_2)\}\right].$$

In the supplement material (SM), we provide more discussions and properties of the proposed loss function (5). It provides a general framework for copula density estimation, including the least squares and logistic loss as specific cases. Using the above U-statistic loss, ρ_0 can be estimated via a standard optimization problem. We specify the candidate function class as non-negative deep neural networks with a multi-layer perception architecture, which can be expressed as $f(z) = \psi_{L+1} \circ \sigma_L \circ \psi_L \circ \cdots \circ \sigma_1 \circ \phi_1(z)$, where $\psi_i(z) = W_i z + b_i$ with $W_i \in \mathbb{R}^{d_i \times d_{i-1}}$,

$b_i \in \mathbb{R}^{d_i}$ with $d_0 = d_x + 1$, and σ_i is the ReLU activation function. The width of the network is defined as $D = \max_{i=1,\cdots,L} d_i$. The collection of the neural network f with depth L and width D is denoted as $\mathcal{M}_n(L, D)$. To ensure the estimator is non-negative, we employ an exponential transformation and let $\mathcal{F}_n(L, D) = \{\exp(f), f \in \mathcal{M}_n(L, D)\}$. The deep copula density function estimator is

$$\widehat{\rho}_n = \underset{\rho \in \mathcal{F}_n(L,D)}{\arg\min} \ \widehat{L}_n(\rho). \tag{6}$$

To evaluate the theoretical properties of $\widehat{\rho}$, the following conditions are imposed. Specifically, we assume that ρ_0 is a smooth function with some low intrinsic dimension, while its nominal dimension $(d + 1)$ can be high and grows with n. Let $Z = (X, Y)$ and $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. Denote $\mathcal{H}^\beta(\mathcal{Z}, B)$ as a Hölder space, where $\beta > 0$ denotes the smoothness of ρ_0, and $B > 0$ is **a bound on** the magnitude of its derivatives, and let $\mathcal{M}$ be a manifold (Lee, 2006) on $\mathcal{X} \times \mathcal{Y}$ where ρ_0 is supported on. More details on $\mathcal{H}^\beta(\mathcal{Z}, B)$ and $\mathcal{M}$ are given in the SM.

Condition 3.1 *(i) The covariate space $\mathcal{X}$ is compact. (ii) There exist some constants $c_1 > 0$ such that $\mathrm{P}(|Y| > t) \leq e^{-c_1 t}$ for every $t > 0$. (iii) $\rho_0 \in \mathcal{H}^\beta(\mathcal{Z}, B)$ for some constants $\beta, B > 0$ and is supported on a neighborhood of $\mathcal{M}$ whose intrinsic dimension is $d_\mathcal{M}$, where $2 \leq d_\mathcal{M} \leq d$.*

Condition 3.2 *For every $\rho \in \mathcal{F}_n(L, D) \cup \{\rho_0\}$ and every $z \in \mathcal{Z}$, it holds that $\rho(z) \leq B$ and $\phi''\{\rho(z)\} \in (c_2^{-1}, c_2)$ for some constant $c_2 > 1$.*

Theorem 3.1 *Under Conditions 1 and 2, the mean squares error of $\widehat{\rho}$ satisfies*

$$\|\widehat{\rho} - \rho_0\|_{L_2(Z)} = O_p\left(C_{d,\mathcal{M}} \log(n) n^{-\frac{\beta}{d_\star + 2\beta}}\right),$$

where $d_\star = O(d_\mathcal{M} \log(d))$ and $C_{d,\mathcal{M}}$ depends linearly on d and polynomially on $d_\mathcal{M}$, with the width D and L in (6) satisfying $D = O(C_{d,\mathcal{M}})$ and $L = O(n^{\frac{d_\star}{2(d_\star + 2\beta)}} \log(n))$, respectively.

Theorem 3.1 establishes that the convergence rate of $\widehat{\rho}$ is primarily governed by $d_\mathcal{M}$ and β, the intrinsic dimension and the smoothness of ρ_0, respectively, revealing a key advantage of the deep copula density estimator over classical nonparametric methods. Notably, if X is high-dimensional but Y depends only on a small subset of X, then the effective dimension $d_\star$ remains low, and $\widehat{\rho}$ can adaptively attain a fast convergence rate. More theoretically challenging than most existing works of deep neural network estimation, the loss function (5) is a U-statistic instead of a sum of independent variables. Our proof presented in the SM is based on the U-process techniques (Clémençon et al., 2008 and De la Pena and Giné, 2012) and the local complexity analysis.

4 Generation and imputation with copula density weighting

Generative learning aims to transform a known distribution, such as the uniform distribution, to the target distribution $F_{Y|X=x}$. Via the copula density weighting and a marginal distribution estimator $\widehat{F}_Y$ for F_Y, the conditional generation can be easily decomposed into two steps.

Step 1. Generate $\{U_m\}_{m=1}^M$ for some positive integer M independently from the uniform distribution $U(0,1)$. Define $\widetilde{Y}_m = \widehat{F}_Y^{-1}(U_m)$ for $m = 1, \ldots, M$.
Step 2. For any given x, draw $\{\widehat{Y}_v\}_{v=1}^K$ for some positive integer K independently from distribution

$$\widetilde{F}_{Y|X=x}(y; \widehat{\rho}, \widehat{F}_Y) = \sum_{m=1}^{M} \mathbb{1}(\widetilde{Y}_m \leq y) \frac{\widehat{\rho}(x, \widetilde{Y}_m)}{\sum_{j=1}^M \widehat{\rho}(x, \widetilde{Y}_j)}. \tag{7}$$

With the above two steps, we obtain a set $\{\widehat{Y}_v\}_{v=1}^K$ sampled from the estimated conditional distribution of $Y|X = x$. The copula density weights are normalzied to ensure (7) is a valid distribution. In addition, it can be interpreted as a self-normalized importance sampling estimator (Hesterberg, 1995) and improves numerical stability.

Notably, (7) can not only generate new samples, but also can be used for hot deck imputations, which refers to imputing missing values using observed responses, thus avoiding the use of artificial values. It is particularly desirable in applications such as household surveys (Andridge and Little, 2010). In our framework, the hot deck imputation can be achieved by setting $\widehat{F}_Y$ as the empirical distribution of the observed $\{Y_i\}_{i=1}^n$, ensuring that all sampled values are among the observed data points. On the other hand, if $\widehat{F}_Y$ is continuous, such as the kernel smoothing estimator, then $\{\widetilde{Y}_m\}_{m=1}^M$ in Step 1 will not be restricted to the observed data, and new synthetic values can be generated in (7). In the following, we establish the convergence rate of $\widetilde{F}_{Y|X}$ for $\widehat{F}_Y$ chosen as either the empirical distribution estimator $\widehat{F}_n(y) = n^{-1}\sum_{i=1}^n \mathbb{1}(Y_i \leq y)$ or the kernel smoothing estimator of $\widehat{F}_h(y) = n^{-1}\sum_{i=1}^n \int_{-\infty}^y K((Y_i - y)/h)$.

Theorem 4.1 *Under Conditions 1 and 2 and taking M large enough such that $n = O(M)$,*

$$\mathrm{E}_X[\{\sup_{y\in\mathcal{Y}} |\widetilde{F}_{Y|X}(y; \widehat{\rho}; \widehat{F}_Y) - F_{Y|X}(y)|\}^2] = O_p\left(\zeta_n + C_{d,\mathcal{M}} \log(n) n^{-\frac{\beta}{d_\star + 2\beta}}\right),$$

where $\zeta_n = n^{-\frac{2s}{2s+1}}$ for $\widehat{F}_Y = \widehat{F}_h$ with $h = O(n^{-\frac{1}{2s+1}})$ under Condition A in the SM, where s is the smoothness of the marginal density f_Y, and $\zeta_n = 0$ for $\widehat{F}_Y = \widehat{F}_n$.

Theorem 4.1 reveals that the estimation error of $\widetilde{F}_{Y|X}$ is of the same order as that of $\widehat{\rho}$ with $\widehat{F}_Y$ being the empirical distribution $\widehat{F}_n$ for hot deck imputations. When the kernel smoothing is used to generate new synthetic values in Step 1, an extra factor of the order $O_p(n^{-\frac{2s}{2s+1}})$ will be introduced. Overall, the estimation error of $\widetilde{F}_{Y|X}$ is largely governed by that of $\widehat{\rho}$ and achieves a fast convergence rate when the intrinsic dimension of ρ_0 is low.

The proposed copula density weighting generative model has several advantages over existing models. First, the method provides a versatile framework for both synthetic value generations and hot-deck imputations. In contrast, most existing deep generative models are limited to generating synthetic values and lack the ability for hot-deck imputations, and traditional hot-deck imputation methods rely on parametric or semiparametric assumptions as reviewed in Kim and Shao (2021), which is not as expressive as the deep neual networks. Second, whereas many deep generative methods rely on complex and black-box transformations, our approach decomposes this mapping into *two explicit and interpretable steps*: marginal inverse transform sampling and copula-based sampling. Third, most existing methods are computationally intensive and often

numerically unstable. For example, the GAN requires solving a non-convex min-max optimization, whose solution is difficult to converge. Diffusion models (Ho et al., 2020) and flow matching models (Lipman et al., 2022) involve iterative sampling steps. In contrast, our method depends on only two estimators $\widehat{F}_Y$ and $\widehat{\rho}(x, y)$, where the latter is modelled as a standard fully connected feedforward neural network whose computations are substantially simpler.

5 Simulations

We report results from a simulation study that was designed to evaluate the performance of the copula density weighting method. In the simulation, the covariates $\{X_i = (X_{1i}, \cdots, X_{ip})'\}_{i=1}^n$ were generated independently from the uniform distribution $U(-1, 1)^p$ with $p = 100$. We considered three model scenarios for the generation mechanism of Y_i:

$$\begin{aligned}
&\text{Scenario A:} \quad Y_i \sim U(0,2)\mathbb{1}(X_{1i} \le 0) + U(-2,0)\mathbb{1}(X_{1i} > 0) \quad \text{(mean shift)} \\
&\text{Scenario B:} \quad Y_i \sim N(0, X_{1i} + 1.5) \quad \text{(variance shift)} \\
&\text{Scenario C:} \quad Y_i \sim N(0,2)\mathbb{1}(X_{1i} \le 0) + U(-1,1)\mathbb{1}(X_{1i} > 0) \quad \text{(distribution family shift)}
\end{aligned}$$

While the covariate X_i was of large dimension, we specified Y_i only depended on X_{1i}, which implies the copula density ρ_0 had a low intrinsic dimension. The training sample sizes were $n = 2000, 5000$ and 10000. The proposed method was employed for the estimation of the conditional distribution estimation and generation based on the procedures outlined in Sections 2 and 4, respectively. We obtained the copula density function estimator $\widehat{\rho}(x, y)$ based on (6), where the ϕ function was specified to be $\phi(x) = (x - 1)^2$, corresponding to the χ^2-divergence.

In addition, we also considered two deep generative models, the conditional variational auto-encoder (Sohn et al., 2015) and the diffusion model. Implementation details of the three methods are reported in the SM. Fig 1 displays the conditional generation performances of the three models, showing that the samples generated with the proposed method were the most similar to the true data under all scenarios. It successfully captured the dependence heterogeneity between Y and X_1, and unlike the other two models, it did not generate data outside the true data support.

Fig 2 displays box-plots for the mean squares errors (MSE) and the training time in the conditional distribution estimation for the three models evaluated on a testing sample of size is 2000. The figure reveals that the proposed method achieved the best estimation accuracy among the three methods, with the MSEs decreasing with the increase of the sample size. The diffusion model had comparable performance to the proposed method under Scenario A. However, its MSE failed to converge in the other two model scenarios. Notably, the training time of the proposed method was substantially smaller than the other two methods, revealing the computational advantage of the copula density weighting, which only requires solving a standard empirical risk minimization problem (6). In contrast, the conditional variational auto-encoder method requires to train both the encoder and decoder networks during training, and the diffusion model is also computationally intensive due to its iterative denoising process during sampling.

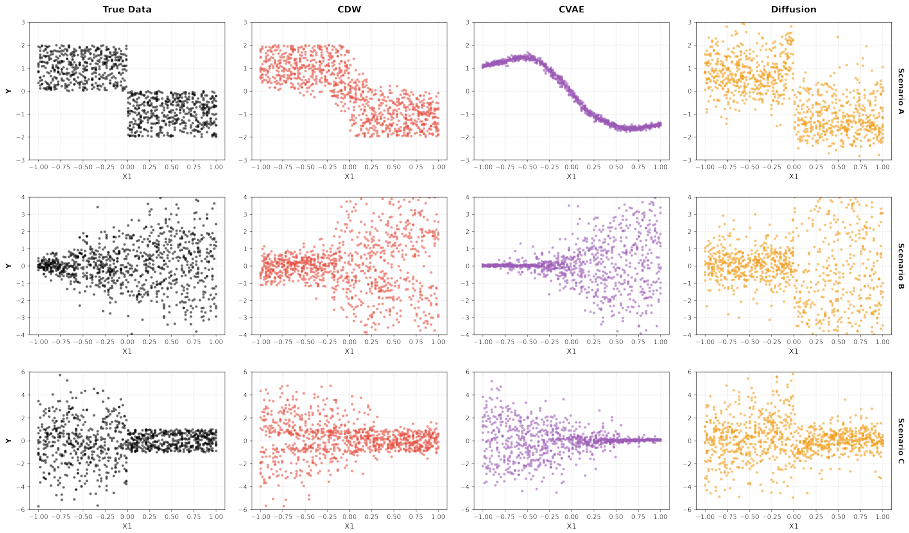

Figure 1: Visualization for conditional data generation performances. The three rows corresponds to the three scenarios, and the four columns display the generation of the true data generation model and three estimated models, including the proposed copula density weighting (CDW), the conditional variational auto-encoder (CVAE) and the diffusion, which were trained with the sample size $n = 10000$.

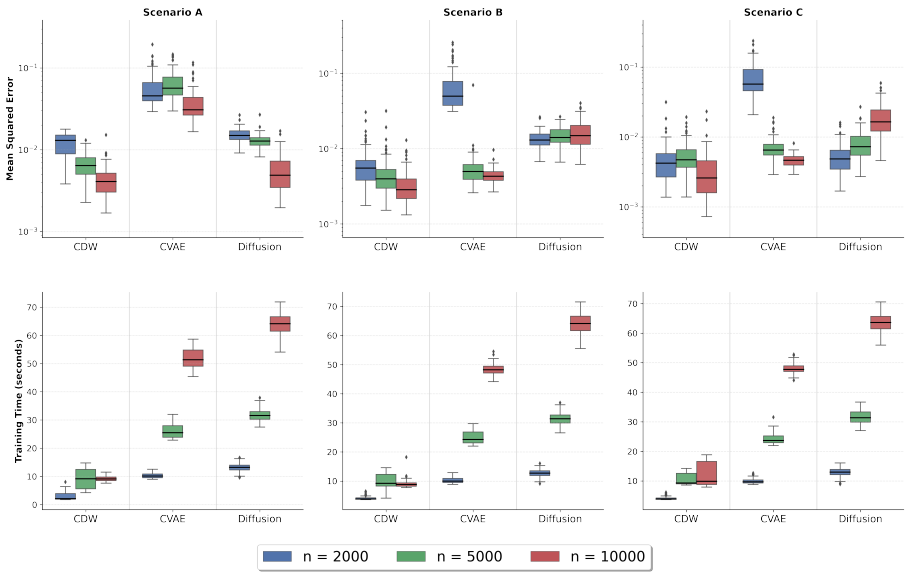

Figure 2: Boxplots of the mean squared errors and training time of the three models for conditional distribution estimation, including the proposed copula density weighting (CDW), the conditional variational auto-encoder (CVAE) and the diffusion under the three scenarios based on 300 replications.

6 Discussion

This study proposes a method for conditional distribution estimation and synthetic data generation based on the copula density weighting. Different from classic conditional distribution estimation methods, the proposed method leverages the high expressiveness of the deep neural networks and can facilitate convenient generation for high-dimensional covariates. Compared with existing generative models in machine learning, the proposed method is more interpretable with substantially reduced computational burden, and can be used for hot deck imputations.

References

Rebecca R Andridge and Roderick JA Little. A review of hot deck imputation for survey non-response. *International statistical review*, 78(1):40–64, 2010.

Victor Chernozhukov, Iván Fernández-Val, and Alfred Galichon. Quantile and probability curves without crossing. *Econometrica*, 78(3):1093–1125, 2010.

Stéphan Clémençon, Gábor Lugosi, and Nicolas Vayatis. Ranking and Empirical Minimization of U-statistics. *The Annals of Statistics*, 36(2):844 – 874, 2008.

Victor De la Pena and Evarist Giné. *Decoupling: from dependence to independence*. Springer Science & Business Media, 2012.

Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, 2014.

Peter Hall, Rodney CL Wolff, and Qiwei Yao. Methods for estimating a conditional distribution function. *Journal of the American Statistical association*, 94(445):154–163, 1999.

Tim Hesterberg. Weighted average importance sampling and defensive mixture distributions. *Technometrics*, 37(2):185–194, 1995.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Torsten Hothorn, Thomas Kneib, and Peter Bühlmann. Conditional transformation models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 76(1):3–27, 2014.

Masahiro Kato and Takeshi Teshima. Non-negative bregman divergence minimization for deep direct density ratio estimation. In *International Conference on Machine Learning*, pages 5320–5333. PMLR, 2021.

Jae Kwang Kim and Wayne A. Fuller. Fractional hot deck imputation. *Biometrika*, 91(3):559–578, 2004. doi: 10.1093/biomet/91.3.559. URL https://doi.org/10.1093/biomet/91.3.559.

Jae Kwang Kim and Jun Shao. *Statistical methods for handling incomplete data*. Chapman and Hall/CRC, 2021.

Roger Koenker. *Quantile regression*, volume 38. Cambridge university press, 2005.

John M Lee. *Riemannian manifolds: an introduction to curvature*, volume 176. Springer Science & Business Media, 2006.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2022.

Nicolai Meinshausen. Quantile regression forests. *Journal of Machine Learning Research*, 7(35): 983–999, 2006.

Roger B Nelsen. *An introduction to copulas*. Springer, 2006.

XuanLong Nguyen, Martin J. Wainwright, and Michael I Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11):5847–5861, 2010.

Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. In *Advances in Neural Information Processing Systems*, volume 28, 2015.

Masashi Sugiyama, Taiji Suzuki, and Takafumi Kanamori. Density-ratio matching under the bregman divergence: a unified framework of density-ratio estimation. *Annals of the Institute of Statistical Mathematics*, 64:1009–1044, 2012.

Supplemental Material

This supplement material provides additional discussions, definitions, and all proofs of the theorems of the main article. It also includes some additional details of the simulations.