Estimating Causal Effects using a Multi-task Deep Ensemble

Ziyang Jiang [1] **Zhuoran Hou** [2] **Yiling Liu** [3] **Yiman Ren** [4] **Keyu Li** [5] **David Carlson** [1 2 5 6]

Abstract

A number of methods have been proposed for causal effect estimation, yet few have demonstrated efficacy in handling data with complex structures, such as images. To fill this gap, we propose Causal Multi-task Deep Ensemble (CMDE), a novel framework that learns both shared and group-specific information from the study population. We provide proofs demonstrating equivalency of CDME to a multi-task Gaussian process (GP) with a coregionalization kernel *a priori*. Compared to multi-task GP, CMDE efficiently handles high-dimensional and multi-modal covariates and provides pointwise uncertainty estimates of causal effects. We evaluate our method across various types of datasets and tasks and find that CMDE outperforms state-of-the-art methods on a majority of these tasks.

1. Introduction

Estimating the causal effect of an action is a fundamental step in determining whether it is significant enough to change human behavior in real-world settings. This is commonly used to help us understand the potential impact of an action and to inform decision-making. For example, governments often judge the impact of an implemented policy by gauging public opinion regarding said policy (Page & Shapiro, 1983). Researchers can assess the efficacy and risk of a medical procedure through the use of both clinical trials, which examine the medical conditions of patients both with and without having the treatment, and observational data, such as electronic health records (Jensen et al., 2012; Zhang et al., 2019). In recent years, people have come up with a variety of approaches that leverage machine learning models to discover causal relationships (Guo et al., 2020; Li & Zhu, 2022). Such methods include, for example, Targeted Maximum Likelihood Estimator (Van Der Laan & Rubin, 2006), Bayesian Additive Regression Trees (Chipman et al., 2010), and Double/Debiased Machine Learning methods (Chernozhukov et al., 2018). Many recent studies focus on learning the individualized treatment effect (ITE) or conditional average treatment effect (CATE) by using deep learning models (see a more comprehensive review in Section 4) or meta-learners (Künzel et al., 2019).

As technology progresses, an increasing number of datasets containing more complex and comprehensive infromation have become available for causal analysis. These datasets often include high-dimensional covariates, such as images, posing unique challenges for causal inference. While previous methods have demonstrated promising performance on a variety of causal inference tasks, including the Infant Health and Development (IHDP (Brooks-Gunn et al., 1992)), Twins (Almond et al., 2005), and Jobs (LaLonde, 1986), few have been specifically evaluated on datasets with high-dimensional and multi-modal structures. In some situations, these complex covariates play a significant role in causal analysis. For instance, neuroimaging is essential in studying how human brain solves problems with multisensory causal inference (Kayser & Shams, 2015), and brain imaging has been used to forecast treatment response in depression (Drysdale et al., 2017), showing that information about causal relationships are embedded in complex data types. This highlights the need for further research on methods that can effectively handle high-dimensional and multi-modal covariates in causal analysis. In this paper, we propose a deep learning framework to address this challenge. Our main contributions are summarized as follows:

- We propose the Causal Multi-task Deep Ensemble (CMDE) framework which estimates the CATE by learning both shared and group-specific information from control and treatment groups in the study population using separate neural networks.
- We demonstrate the relationship between CMDE and

[1]Department of Civil and Environmental Engineering, Duke University, Durham, NC, USA [2]Department of Biostatistics and Bioinformatics, Duke University School of Medicine, Durham, NC, USA [3]Program in Computational Biology and Bioinformatics, Duke University School of Medicine, Durham, NC, USA [4]Department of Economics, University of Michigan Ross School of Business, Ann Arbor, MI, USA [5]Department of Electrical and Computer Engineering, Duke University, Durham, NC, USA [6]Department of Computer Science, Duke University, Durham, NC, USA. Correspondence to: David Carlson <david.carlson@duke.edu>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

multi-task Gaussian process (GP) framework (Alaa & Van Der Schaar, 2017) both through analytical proof and empirical evaluation.

- We propose an alternative configuration of CMDE to handle covariates in a multi-modal setting (e.g., images and tabular data).
- By conducting experiments on semi-synthetic and real-world datasets containing covariates with various number of dimensions and modalities, we show that CMDE outperforms the state-of-the-art methods by improving estimation of the treatment effects.

2. Background

2.1. Problem Setup

We will use lower-case letters (e.g., x, t, y) for individual samples, upper-case letters (e.g., X, T, Y) for random variables, and bold upper-case letters (e.g., $\boldsymbol{X}, \boldsymbol{T}, \boldsymbol{Y}$) for a set of samples throughout the paper. We consider a general setting in causal inference where we assign a specific treatment to a group of individuals. Each individual is represented by a D-dimensional feature vector $X \in \mathcal{X} \subset \mathbb{R}^D$ ($\mathcal{X}$ denotes the training input space) and is associated with a treatment-assignment indicator $T \in \{0, 1\}$. The corresponding *potential outcomes* are denoted by $Y^{(0)} \in \mathbb{R}$ and $Y^{(1)} \in \mathbb{R}$ where the superscripts 0 and 1 represent assignment to the control group and the treatment group, respectively. We assume there exist a joint distribution $P\left(X, T, Y^{(0)}, Y^{(1)}\right)$ which satisfies $0 < P(T = 1|X) < 1$ and the strong ignorability assumption $\left(Y^{(0)}, Y^{(1)}\right) \perp\!\!\!\perp T|X$ as given in the Rubin-Neyman causal model (Rosenbaum & Rubin, 1983; Rubin, 2005). Our goal is to estimate the *conditional average treatment effect* (CATE) from a training dataset containing N data points $\mathcal{D} = \left\{x_i, t_i, y_i^{(t)}\right\}_{i=1}^N$ where CATE can be computed as CATE $:= \mathbb{E}\left[Y^{(1)} - Y^{(0)}|X\right]$. We denote $Y^{(T)}$ and $Y^{(1-T)}$ as *factual* and *counterfactual* outcomes, respectively. That is, we have $Y^{(T)} = (1-T)Y^{(0)} + TY^{(1)}$ and $Y^{(1-T)} = TY^{(0)} + (1-T)Y^{(1)}$.

2.2. Multi-task Gaussian Processes (GPs)

We first introduce the background of multi-task GPs. A GP is a stochastic process that is completely defined by a mean function $\mu : \mathcal{X} \to \mathbb{R}$ and a kernel function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ (Rasmussen, 2003). Without loss of generality, we assume that the mean function $\mu(\boldsymbol{x})$ is zero for simplicity, as is common in the literature. A single-output function $f : \mathcal{X} \to \mathbb{R}$ following a GP is written as

$$f \sim \mathcal{GP}\left(0, k\right). \tag{1}$$

For any finite subset $\boldsymbol{X} = \{x_i\}_{i=1}^N$, $f(\boldsymbol{X})$ follows a multivariate Gaussian distribution with mean zero and covariance matrix $k(\boldsymbol{X}, \boldsymbol{X}) \in \mathbb{R}^{N \times N}$ with entries $k(x_i, x_j)$ where $1 \leq i, j \leq N$. We extend a GP to a multi-task learning scenario by defining a *vector-valued* function $\mathbf{f} : \mathcal{X} \to \mathbb{R}^C$, and we can write the corresponding multi-task GP as

$$\mathbf{f} \sim \mathcal{GP}\left(\mathbf{0}, \mathbf{K}\right), \tag{2}$$

where $\mathbf{K} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^{C \times C}$ denotes a *matrix-valued* kernel function. Again, any finite subset $\mathbf{f}(\boldsymbol{X}) \in \mathbb{R}^{N \times C}$ follows a multivariate Gaussian distribution with mean zero and covariance matrix $\mathbf{K}(\boldsymbol{X}, \boldsymbol{X}) \in \mathbb{R}^{NC \times NC}$ constructed from the set of kernel values $\left(\mathbf{K}\left(x_i, x_j\right)\right)_{c,c'}$ with $1 \leq i, j \leq N$ and $1 \leq c, c' \leq C$, giving the form $\text{vec}(\mathbf{f}(\boldsymbol{X})) \sim \mathcal{N}(\mathbf{0}, \mathbf{K}(\boldsymbol{X}, \boldsymbol{X}))$ where $\text{vec}(\cdot)$ denotes vectorization which transforms $\mathbf{f}(\boldsymbol{X})$ from $\mathbb{R}^{N \times C}$ to $\mathbb{R}^{NC}$.

Alaa & Van Der Schaar (2017) used a multi-task GP for causal inference by setting the number of tasks equal to the number of potential outcomes, which is $C = 2$ for a binary treatment $T \in \{0, 1\}$. Here, $\mathbf{f} = [f_0, f_1]^T$, where f_0 approximates the potential outcome for $T = 0$, and f_1 approximates the potential outcome for $T = 1$. Defining $\mathbf{e} = [-1, 1]^T$, we can then approximate CATE as

$$\begin{aligned} \text{CATE}(x) :=& \mathbb{E}\left[Y^{(1)} - Y^{(0)}|X\right] \\ =& \mathbf{f}^T(x)\mathbf{e} = f_1(x) - f_0(x). \end{aligned} \tag{3}$$

2.3. Coregionalization Models

A common approach to construct a multi-task GP is to use coregionalization models (Alvarez et al., 2012). For example, we can construct the matrix-valued kernel function $\mathbf{K}$ from a single-output kernel by using the *Intrinsic Coregionalization Model* (ICM (Goovaerts et al., 1997)),

$$\mathbf{K}_{\text{ICM}}(x, x') = k(x, x')\mathbf{B}, \tag{4}$$

where $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a scalar-valued kernel function and $\mathbf{B} \in \mathbb{R}^{C \times C}$ is called a *coregionalization matrix*. If a function follows $\mathbf{f} \sim \mathcal{GP}\left(\mathbf{0}, \mathbf{K}_{\text{ICM}}\right)$, then each of its entries $\mathbf{f}_c$ can be expressed as a linear combination of functions sampled from a GP with zero mean and covariance function k. That is, for a coregionalization matrix with $\text{rank}(\mathbf{B}) = R$, we have $\mathbf{f}_c(x) = \sum_{r=1}^R a_c^r u^r(x)$ where $u^r(x) \sim \mathcal{GP}(0, k)$ for all $r \in [1, R]$.

We can construct a more generalized model by using a *Linear Model of Coregionalization* (LMC (Journel & Huijbregts, 1976; Goovaerts et al., 1997)) to define $\mathbf{K}$,

$$\mathbf{K}_{\text{LMC}}(x, x') = \sum_{q=1}^{Q} k_q(x, x')\mathbf{B}_q, \tag{5}$$

where $k_q : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is again a scalar-valued kernel function and $\mathbf{B}_q \in \mathbb{R}^{C \times C}$ for all $q \in [1, Q]$. It is straightforward

to see that the LMC can be viewed as a mixture of Q ICMs. Likewise, if a function follows $\mathbf{f} \sim \mathcal{GP}(\mathbf{0}, \mathbf{K}_{\text{LMC}})$, then we have $\mathbf{f}_c(x) = \sum_{q=1}^{Q} \sum_{r=1}^{R_q} a_{c,q}^r u_q^r(x)$ where $u_q^r(x) \sim \mathcal{GP}(0, k_q)$ for all $r = 1, ..., R_q$ with $\text{rank}(\mathbf{B}_q) = R_q$ and all $q = 1, ..., Q$.

2.4. Relationship between NNs and GPs

As stated by Matthews et al. (2018), a random deep NN with appropriate activation function will converge *in distribution* to a GP. Specifically, let $f_{\text{NN}} : \mathcal{X} \to \mathbb{R}$ be a function implemented by a NN with zero-mean i.i.d. parameters and continuous activation function ϕ which satisfies the following linear envelope property:

$$|\phi(u)| \leq \beta + m|u| \quad \forall u \in \mathbb{R}, \tag{6}$$

if there exist $\beta, m \geq 0$. This property is satisfied by many common nonlinearities (e.g., ReLU, softplus, tanh, etc.). Functions that violate this property (e.g., exponential) will induce heavy-tail behavior in the post activation. Under these conditions, f_{NN} will converge *in distribution* to a GP in the infinite width limit,

$$f_{\text{NN}} \xrightarrow{d} \mathcal{GP}(\mathbf{0}, k_{\text{NN}}), \tag{7}$$

where $k_{\text{NN}} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a NN-implied kernel function and can be numerically estimated in a recursive manner (Lee et al., 2017). However, for a single NN without a Bayesian formulation, its prediction can only be viewed as a *sample* corresponding to a GP prior. To enable a full GP posterior interpretation, we adopt the sample-then-optimize approach proposed by Matthews et al. (2017) by constructing and training a deep ensemble as we will discuss in the following section.

3. Causal Multi-task Deep Ensemble

Here, we formally present our Causal Multi-task Deep Ensemble (CMDE) framework and elaborate on its relationship to coregionalization models. The ensemble's predictions are made by averaging over all its *baselearners*. The architecture of a single baselearner in our ensemble is depicted in Figure 1 where features X are passed into 3 neural networks $f_H, f_T, f_{HT} : \mathcal{X} \to \mathbb{R}$ separately as shown in Figure 1a. Each baselearner follows the same architecture but has a different random initialization, which we will show corresponds to different draws from a multi-task GP prior. We expect f_H and f_T to learn group-specific information from control and treatment group, respectively, and f_{HT} to learn shared information between the two groups. Each baselearner learns a multi-output function $\hat{\mathbf{f}} = [\hat{f}_0, \hat{f}_1]^T$ which generates two outputs $\hat{Y}^{(0)}$ and $\hat{Y}^{(1)}$ representing the *potential outcomes* for treatment assignment $T \in \{0, 1\}$:

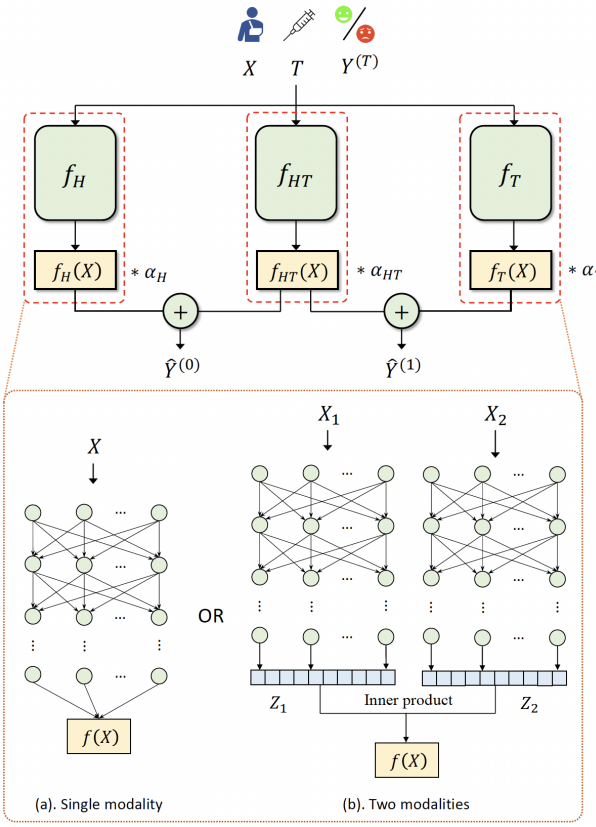

Figure 1. The overall architecture of **a baselearner** in our causal multi-task deep ensemble where f in (a) and (b) can be one of the f_H, f_{HT}, or f_T. Here the treatment assignment indicator T is only used to obtain the corresponding factual outcome $Y^{(T)}$ for training and is not passed into f_H, f_T, or f_{HT} as an input.

$$\hat{Y}^{(0)} = \hat{f}_0(X) := \alpha_H f_H(X) + \alpha_{HT} f_{HT}(X), \tag{8}$$

$$\hat{Y}^{(1)} = \hat{f}_1(X) := \alpha_{HT} f_{HT}(X) + \alpha_T f_T(X), \tag{9}$$

where α_H, α_T, and α_{HT} are trainable parameters that need to be initialized *a priori*. With this formulation, we claim the following theorem.

Theorem 3.1. *If all f_H, f_T, and f_{HT} are neural networks with identical depth, zero-mean i.i.d. parameters with the same variance, and continuous activation function ϕ which satisfies the linear envelope property given in (6), then $\hat{\boldsymbol{f}}$ converges in distribution to a GP with zero mean and ICM kernel in the infinite width limit **a priori**:*

$$\hat{\boldsymbol{f}} \xrightarrow{d} \mathcal{GP}(\mathbf{0}, \boldsymbol{K}_{ICM}), \tag{10}$$

where $\boldsymbol{K}_{ICM}(x, x') = k_{NN}(x, x')\boldsymbol{B}$, $x, x' \in \mathcal{X}$, and k_{NN} is the kernel function implied by f_H, f_T, and f_{HT}, and

$$\boldsymbol{B} = \begin{bmatrix} \alpha_H^2 + \alpha_{HT}^2 & \alpha_{HT}^2 \\ \alpha_{HT}^2 & \alpha_T^2 + \alpha_{HT}^2 \end{bmatrix}. \tag{11}$$

Proof. To prove Theorem 3.1, we calculate the covariance matrix between $\hat{\mathbf{f}}(x)$ and $\hat{\mathbf{f}}(x')$ as below (note that the expectations are taken with respect to the parameters of the functions inside the expectation):

$$\text{cov}\left(\hat{\mathbf{f}}(x), \hat{\mathbf{f}}(x')\right) \tag{12}$$

$$= \mathbb{E}\left[\hat{\mathbf{f}}(x)\hat{\mathbf{f}}(x')^T\right] - \mathbb{E}\left[\hat{\mathbf{f}}(x)\right]\mathbb{E}\left[\hat{\mathbf{f}}(x')\right]^T \tag{13}$$

$$= \begin{bmatrix} \mathbb{E}[\hat{f}_0(x)\hat{f}_0(x')] & \mathbb{E}[\hat{f}_0(x)\hat{f}_1(x')] \\ \mathbb{E}[\hat{f}_1(x)\hat{f}_0(x')] & \mathbb{E}[\hat{f}_1(x)\hat{f}_1(x')] \end{bmatrix}. \tag{14}$$

From (13) to (14), we drop the term $\mathbb{E}\left[\hat{\mathbf{f}}(x)\right]\mathbb{E}\left[\hat{\mathbf{f}}(x')\right]^T$ as all the parameters in f_H, f_T, and f_{HT} are initialized by i.i.d. zero mean random variables. We can then calculate each entry in (14) separately as follows:

$$\mathbb{E}[\hat{f}_0(x)\hat{f}_0(x')] = \alpha_H^2\mathbb{E}[f_H(x)f_H(x')] + \alpha_{HT}^2\mathbb{E}[f_{HT}(x)f_{HT}(x')], \tag{15}$$

$$\mathbb{E}[\hat{f}_0(x)\hat{f}_1(x')] = \alpha_{HT}^2\mathbb{E}[f_{HT}(x)f_{HT}(x')], \tag{16}$$

$$\mathbb{E}[\hat{f}_1(x)\hat{f}_0(x')] = \mathbb{E}[\hat{f}_0(x)\hat{f}_1(x')] = \alpha_{HT}^2\mathbb{E}[f_{HT}(x)f_{HT}(x')], \tag{17}$$

$$\mathbb{E}[\hat{f}_1(x)\hat{f}_1(x')] = \alpha_T^2\mathbb{E}[f_T(x)f_T(x')] + \alpha_{HT}^2\mathbb{E}[f_{HT}(x)f_{HT}(x')]. \tag{18}$$

From (15) to (18), we make use of the fact that the parameters of f_H, f_T, and f_{HT} are independent of each other *a priori*. Also, since f_H, f_T, and f_{HT} share the same depth and initialization strategy, then as elaborated in Section 2.4, we have $\mathbb{E}[f_H(x)f_H(x')] = \mathbb{E}[f_T(x)f_T(x')] = \mathbb{E}[f_{HT}(x)f_{HT}(x')] = k_{\text{NN}}(x, x')$ in the infinite width limit. Therefore, as the width of f_H, f_T, and f_{HT} goes to infinity, we have:

$$\mathbb{E}[\hat{f}_0(x)\hat{f}_0(x')] = (\alpha_H^2 + \alpha_{HT}^2)k_{\text{NN}}(x, x'),$$
$$\mathbb{E}[\hat{f}_0(x)\hat{f}_1(x')] = \mathbb{E}[\hat{f}_1(x)\hat{f}_0(x')] = \alpha_{HT}^2 k_{\text{NN}}(x, x'),$$
$$\mathbb{E}[\hat{f}_1(x)\hat{f}_1(x')] = (\alpha_T^2 + \alpha_{HT}^2)k_{\text{NN}}(x, x').$$

By substituting the equations above back into (14), we get

$$\text{cov}\left(\hat{\mathbf{f}}(x), \hat{\mathbf{f}}(x')\right) = k_{\text{NN}}(x, x')\begin{bmatrix} \alpha_H^2 + \alpha_{HT}^2 & \alpha_{HT}^2 \\ \alpha_{HT}^2 & \alpha_T^2 + \alpha_{HT}^2 \end{bmatrix}, \tag{19}$$

which completes our proof for Theorem 3.1. □

In addition, by following a similar approach, we can also construct $\hat{\mathbf{f}} = [\hat{f}_0, \hat{f}_1]^T$ as below:

$$\hat{f}_0(X) \coloneqq \sum_{q=1}^{Q} \alpha_H^q f_H^q(X) + \alpha_{HT}^q f_{HT}^q(X), \tag{20}$$

$$\hat{f}_1(X) \coloneqq \sum_{q=1}^{Q} \alpha_{HT}^q f_{HT}^q(X) + \alpha_T^q f_T^q(X), \tag{21}$$

where f_H^q, f_T^q, and f_{HT}^q share the same depth and initialization strategy for the same value of q. With this formulation, it can be shown that $\hat{\mathbf{f}}$ converges *in distribution to* a GP with zero mean and LMC kernel in the infinite width limit *a priori* (see detailed proof in Appendix A):

$$\hat{\mathbf{f}} \xrightarrow{d} \mathcal{GP}\left(\mathbf{0}, \mathbf{K}_{\text{LMC}}\right), \tag{22}$$

where $\mathbf{K}_{\text{LMC}}(x, x') = \sum_{q=1}^{Q} k_{\text{NN}}^q(x, x')\mathbf{B}_q$. Here k_{NN}^q is the kernel function implied by f_H^q, f_T^q, or f_{HT}^q and

$$\mathbf{B}_q = \begin{bmatrix} (\alpha_H^q)^2 + (\alpha_{HT}^q)^2 & (\alpha_{HT}^q)^2 \\ (\alpha_{HT}^q)^2 & (\alpha_T^q)^2 + (\alpha_{HT}^q)^2 \end{bmatrix}. \tag{23}$$

In theory, our method can also be extended to the multiple-treatment case $T \in \{1, ..., C\}$, where we can construct $\hat{\mathbf{f}} = \left[\hat{f}_1, ..., \hat{f}_C\right]^T$ as follows:

$$\hat{f}_c(X) \coloneqq \sum_{d=1}^{c-1} \alpha_{dc} f_{dc}(X) + \alpha_c f_c(X) + \sum_{d=c+1}^{C} \alpha_{cd} f_{cd}(X) \quad \forall\, c = 1, ..., C, \tag{24}$$

where f_c learns the group-specific information and f_{dc}, f_{cd} learn the shared information. With this formulation, we can also prove that $\hat{\mathbf{f}}$ converges in distribution to a GP with an ICM kernel as elaborated in Appendix B. Our framework can also be simplified so each baselearner only contains 2 networks as shown in Appendix C. However, we will stick to the 3-network architecture in our experiments as it facilitates the explanation of the role of each network and enhances the clarity of our statement. Note that we only state the equivalence between CMDE and a multi-task GP *a priori*. According to He et al. (2020), the equivalence between a deep ensemble and a GP may still hold *a posteriori* (i.e., after training) if we augment the forward pass of each NN in the baselearner by adding a random and untrainable function $\delta(\cdot)$. However, we do not claim this equivalence, as the parameters of f_H, f_T, and f_{HT} become dependent on each other. Consequently, we are unable to eliminate the cross-terms when calculating $\text{cov}\left(\hat{\mathbf{f}}(x), \hat{\mathbf{f}}(x')\right)$.

3.1. Extension to Multi-modal Covariates

In some cases, the covariates X contain multiple modalities (e.g., $X = \{X_1, X_2\}$ where X_1 is an image and X_2 is in

a tabular format). As illustrated in Figure 1b, we adapt CMDE to such situations by introducing an inner product between the extracted representations from each modalities of X. Specifically, let Z_m be the extracted representation from the mth input modality by a neural network. We can construct f as follows

$$f(X) := \sum_j \prod_m (Z_m)_j, \tag{25}$$

where $(Z_m)_j$ represents the jth entry in Z_m and f is one of the f_H, f_{HT}, or f_T. As proved by Lee et al. (2017) and Jiang et al. (2022), this mechanism yields a multiplicative kernel $k_{\text{mul}} = \prod_m (k_{\text{NN}})_m$ where $(k_{\text{NN}})_m$ is the kernel function implied by the neural network used to extract the representation Z_m from the input modality X_m. With this formulation, the multi-output function $\hat{\mathbf{f}}$ still converges to an ICM or LMC kernel (depending on how we construct $\hat{\mathbf{f}}$) except that we replace k_{NN} with k_{mul}.

3.2. Training of CMDE

A common goal in causal inference is to minimize the *precision in estimating heterogeneous effect* (PEHE (Hill, 2011)) loss, which is defined as

$$\hat{\mathcal{L}}\left(\hat{\mathbf{f}}; \boldsymbol{Y}^{(\boldsymbol{T})}, \boldsymbol{Y}^{(\mathbf{1}-\boldsymbol{T})}\right) = \frac{1}{N}\sum_{i=1}^{N}\left(\hat{\mathbf{f}}^T(x_i)\mathbf{e} - (1-2t_i)\left(y_i^{(1-t_i)} - y_i^{(t_i)}\right)\right)^2, \tag{26}$$

where $\mathbf{e} = [-1, 1]^T$, $\boldsymbol{Y}^{(\boldsymbol{T})} = \left\{y_i^{(t_i)}\right\}_{i=1}^N$ are the factual outcomes, and $\boldsymbol{Y}^{(\mathbf{1}-\boldsymbol{T})} = \{y_i^{(1-t_i)}\}_{i=1}^N$ are the counterfactual outcomes. To train CMDE, we want to minimize the following regularized empirical loss,

$$\hat{\mathbf{f}}^* = \arg\min_{\hat{\mathbf{f}} \in \mathcal{H}_{\mathbf{K}}} \hat{\mathcal{L}}\left(\hat{\mathbf{f}}; \boldsymbol{Y}^{(\boldsymbol{T})}, \boldsymbol{Y}^{(\mathbf{1}-\boldsymbol{T})}\right) + \lambda \|\hat{\mathbf{f}}\|^2_{\mathcal{H}_{\mathbf{K}}}, \tag{27}$$

where $\mathcal{H}_{\mathbf{K}}$ is a *vector-valued Reproducing Kernel Hilbert Space* (vvRKHS) equipped with an inner product $\langle \cdot\,,\,\cdot \rangle_{\mathcal{H}_{\mathbf{K}}}$, and reproducing kernel $\mathbf{K} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^{2\times 2}$. The regularization term smooths the loss based on the GP prior.

However, we cannot compute PEHE directly because we only observe the factual outcomes. Instead, we minimize the following empirical Bayesian PEHE risk with respect to $\hat{\mathbf{f}}$ using stochastic gradient descent (SGD) with L_2 regularization:

$$\hat{\mathcal{R}} = \frac{1}{N}\sum_{i=1}^{N}\left(y_i^{(t_i)} - \mathbb{E}\left[\hat{y}_i^{(t_i)} | x_i\right]\right)^2 + \left\|\text{Var}\left[\hat{y}_i^{(1-t_i)} | x_i\right]\right\|_1, \tag{28}$$

where $\hat{y}_i^{(t_i)}$ and $\hat{y}_i^{(1-t_i)}$ are the potential outcomes estimated by each estimator in CMDE as given in (8) and (9). The empirical mean and variance in (28) are computed over all the estimators in the ensemble. It has been proved by Alaa & Van Der Schaar (2017) that minimizing this risk is equivalent to minimizing the expectation of $\hat{\mathcal{L}}\left(\hat{\mathbf{f}}; \boldsymbol{Y}^{(\boldsymbol{T})}, \boldsymbol{Y}^{(\mathbf{1}-\boldsymbol{T})}\right)$ with respect to the posterior distribution of the counterfactual outcomes, which leads to a kernel that considers not only factual errors but also generalization to counterfactuals.

4. Related Work

There exist several previous studies that focus on learning the potential outcomes with deep models or ensemble models, including Balancing Counterfactual Regression (Johansson et al., 2016), the Counterfactual Regression Network (CFRNet (Shalit et al., 2017)), and Bayesian Additive Regression Trees (BART) (Chipman et al., 2010). These works generally attempt to learn a function $f : \mathcal{X} \times \{0, 1\} \to \mathbb{R}$ which takes both the covariates and the treatment indicator as inputs. The treatment effect for an individual x can thus be estimated as $\tau(x) \approx \hat{\tau}_f(x) = f(x, t = 1) - f(x, t = 0)$. However, the representation power of the treatment indicator t can be significantly diluted when the dimension of the covariates x becomes high, which can negatively affect the estimation of potential outcomes (Alaa & Van Der Schaar, 2017; Alaa et al., 2017). Besides this line of work, Alaa & Van Der Schaar (2017) proposed a multi-task GP framework that directly outputs the potential outcomes for both $t = 0$ and $t = 1$ by learning a multi-output function $\mathbf{f} : \mathcal{X} \to \mathbb{R}^2$. The treatment effect in this case is estimated as $\tau(x) \approx \hat{\tau}_{\mathbf{f}}(x) = \mathbf{f}(x)^T\mathbf{e}$ where $\mathbf{e} = [-1, 1]^T$. Similar to CMGP, our deep ensemble model also learns a multi-output function while we replace the GPs with NNs to handle larger datasets and high-dimensional covariates, especially in cases where NNs outperform traditional GP approaches (e.g., images).

To the best of our knowledge, there is limited research on utilizing deep ensemble learning for causal inference in the existing literature. One notable study in this domain is the work by Hartford et al. (2021), which focuses on treatment effect estimation using an ensemble of deep network-based instrumental variable estimators. Ensembles are commonly employed in machine learning and have been recognized for their effectiveness in reducing variance. In our approach, we adopt the ensemble method to approximate the full multi-task Gaussian process (GP) prior, whereby each baselearner can be viewed as a stochastic draw from this prior.

It is also worth noting that CFRNet learns a *balanced* representation such that the induced distributions for control and treatment groups look similar. Specifically, CFRNet consists of a network Φ which learns a representation from the covariates followed by two networks h which learns hypotheses h_0 and h_1. This concept was employed in several subse-

quent studies on deep causal inference, including the Causal Effect Variational Autoencoder (CEVAE, (Louizos et al., 2017)), the Deep Counterfactual Network (DCN (Alaa et al., 2017)), and the Deep Orthogonal Networks (DONUT (Hatt & Feuerriegel, 2021)). In contrast, our CMDE framework learns 3 representations which contains both group-specific and shared information from control and treatment groups. This provides more modeling flexibility, especially when the control and treatment groups are highly imbalanced.

As we explain in Section 3.1, CMDE can be extended to handle multi-modal covariates. One recent study that focused on multi-modal causal inference is the Deep Multi-modal Structural Equations (DMSE) (Deshpande et al., 2022). A key distinction between DMSE and our method is that our causal graph does not include any latent variables. While we acknowledge the potential benefits of latent-variable-based approaches in certain scenarios, their performance heavily relies on the complexity and accurate specification of the latent variable distribution, as emphasized by Rissanen and Marttinen (2021).

The equivalence between NNs and GPs is also highly relevant to our method. Neal first proved that single-hidden-layer NNs become GPs as the width of the hidden layer goes to infinity (Neal, 1996; 2012). This proof was then extended to deep neural networks (DNNs) by Lee et al. (2017) who designed an efficient implementation to calculate the NN-implied kernel and Matthews et al. (2018) who empirically evaluated the convergence rate via maximum mean discrepancy (MMD). Garriga-Alonso et al. (2018) also proved that convolutional neural networks (CNNs) are GPs in the limit of infinite number of channels. These findings allow us to simulate GP behavior using the outputs from NNs.

5. Experimental Results

We conduct experiments on a total of 6 datasets: one purely synthetic dataset, 3 benchmark datasets, and 2 datasets with multiple input modalities[1]. The detailed experimental setup is given in Appendix D.

5.1. Synthetic Dataset

To empirically show the convergence of CMDE to its GP counterpart as the width of NNs goes to infinity, we first test CMDE on a synthetic dataset (see detailed data generation process in Appendix D.1) and compare it to a causal multi-task GP (CMGP) (Alaa & Van Der Schaar, 2017) with an

[1]The code to replicate all experiments is available at: https://github.com/jzy95310/ICK/tree/main/experiments/causal_inference

	Twins: $\sqrt{\hat{\epsilon}_{\text{PEHE}}}$		Jobs: $\mathcal{R}_{\text{pol}}(\pi)$	
	In-sample	Out-of-sample	In-sample	Out-of-sample
CMDE	$.32 \pm .00$	$\mathbf{.32 \pm .01}$	$\mathbf{.05 \pm .01}$	$\mathbf{.26 \pm .02}$
CMGP	$.44 \pm .00$	$.44 \pm .01$	$.12 \pm .02$	$.30 \pm .02$
CEVAE	$.32 \pm .00$	$\mathbf{.32 \pm .01}$	$.11 \pm .03$	$.29 \pm .03$
GANITE	$.33 \pm .00$	$.33 \pm .01$	$.10 \pm .02$	$.30 \pm .01$
X-RF	$\mathbf{.30 \pm .00}$	$.33 \pm .01$	N/A	N/A
X-BART	$.32 \pm .00$	$\mathbf{.32 \pm .01}$	N/A	N/A
CFR-Wass	$.32 \pm .00$	$\mathbf{.32 \pm .01}$	$.09 \pm .03$	$.28 \pm .02$
CFR-MMD	$.32 \pm .00$	$\mathbf{.32 \pm .01}$	$.08 \pm .04$	$.28 \pm .03$
DONUT	$.32 \pm .00$	$\mathbf{.32 \pm .01}$	$.09 \pm .05$	$.27 \pm .03$

Table 1. Performance of CATE estimation on the Twins (left) and the Jobs (right) datasets for both in-sample and out-of-sample settings. Lower $\sqrt{\hat{\epsilon}_{\text{PEHE}}}$ or $\mathcal{R}_{\text{pol}}(\pi)$ is better. For Jobs, we do not report the results of X-learner as it directly estimates the individual treatment effect (ITE) instead of $y^{(t)}$.

ICM kernel,

$$\mathbf{K}_{\text{ICM}}(x, x') = k_{\text{NN}}(x, x')\mathbf{B},$$
$$k_{\text{NN}}(x, x') = \frac{2}{\pi}\sin^{-1}\left(\frac{2\tilde{x}^T\Sigma\tilde{x}'}{\sqrt{(1+2\tilde{x}^T\Sigma\tilde{x})(1+2\tilde{x}'^T\Sigma\tilde{x}')}}\right),$$

where $\tilde{x} = [1, x_1, x_2, ..., x_D]$ (e.g., a constant concatenated to the feature vector) and $\Sigma \in \mathbb{R}^{D\times D}$ is a pre-defined parameter representing the covariance of the weights in a single-hidden-layer NN. To evaluate how well we estimate the treatment effect, we use the PEHE metric

$$\epsilon_{\text{PEHE}} = \frac{1}{N}\sum_{i=1}^{N}\left(\mathbb{E}_{y_i^{(0)}, y_i^{(1)}\sim\mathcal{Y}}\left[y_i^{(1)} - y_i^{(0)}\right] - \left(\hat{y}_i^{(1)} - \hat{y}_i^{(0)}\right)\right)^2, \tag{29}$$

where $y^{(0)}, y^{(1)}$ are true outcomes and $\hat{y}_i^{(0)}, \hat{y}_i^{(1)}$ are predicted outcomes. As shown in Figure 2, the two methods yield very similar mean predictions and PEHE values except that CMDE tends to extrapolate with less confidence (i.e. higher standard deviation) where there exist fewer observed samples. We attribute this effect to only using 10 estimators for CMDE. We also plot the group-specific and shared components $\alpha_H f_H$, $\alpha_T f_T$, and $\alpha_{HT} f_{HT}$ in CMDE, which reveals that f_{HT} learns the overall shape shared by the two response surfaces while f_H and f_T learn the magnitude of difference between the two surfaces.

5.2. Benchmark Datasets

We then evaluate CMDE on 3 frequently used benchmark datasets in the existing causal inference literature: a dataset acquired from the Atlantic Causal Inference Conference held in 2019 (ACIC2019 (Dorie et al., 2019)), the Twins

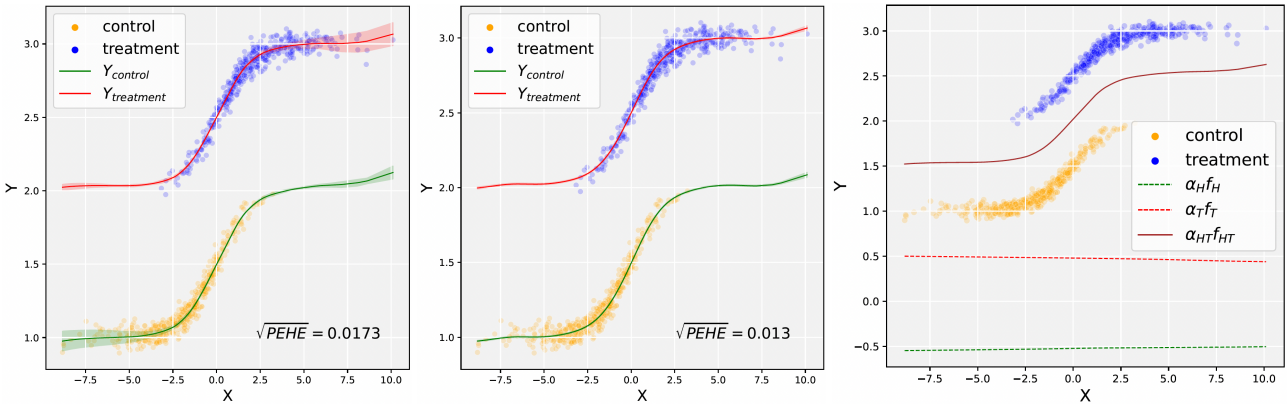

Figure 2. Predictions for the control and treatment groups on the synthetic dataset by CMDE (left) and multi-task GP with ICM kernel (middle) where dots represent observed samples, lines represent mean predictions, and shaded regions represent predicted values within 2 standard deviations. In addition, we also plot the contribution of group-specific and shared components for CMDE (right). It can be observed that f_{HT} learns the shared features between the control and treatment groups and f_H and f_T learns the group-specific features.

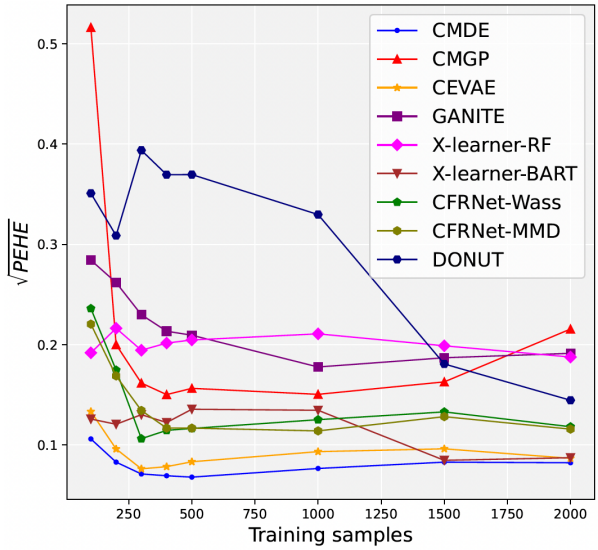

Figure 3. Performance of CATE estimation ($\sqrt{\epsilon_{\text{PEHE}}}$) on a dataset acquired from the Atlantic Causal Inference Conference held in 2019 (ACIC2019). Lower $\sqrt{\epsilon_{\text{PEHE}}}$ is better. Most of the methods exhibit saturated performance with > 500 training samples.

dataset containing twins birth in the United States from 1989 to 1991 (Almond et al., 2005), and the Jobs dataset studied by LaLonde (1986) which is composed of randomized data based on state-supported work programs and non-randomized data from observational studies (see data preprocessing details in Appendix D.2). For evaluation metrics, we use ϵ_{PEHE} as given in (29) for ACIC2019 since we know the true expected values of $Y^{(0)}$ and $Y^{(1)}$. For the Twins dataset, since we observe both the factual and counterfactual outcomes (i.e. $y_i^{(0)}$ and $y_i^{(1)}$) on the paired data but do not know the underlying distribution $\mathcal{Y}$, we use the following empirical PEHE:

$$\hat{\epsilon}_{\text{PEHE}} = \frac{1}{N}\sum_{i=1}^{N}\left(\left(y_i^{(1)} - y_i^{(0)}\right) - \left(\hat{y}_i^{(1)} - \hat{y}_i^{(0)}\right)\right)^2. \tag{30}$$

For the Jobs dataset, only the factual outcomes are observed, so we use a metric called *policy risk*,

$$\mathcal{R}_{\text{pol}}(\pi_f) = 1 - \frac{\sum_{i=1}^{N} y_i^{(t_i)} \mathbb{1}[\pi_f(x_i)=t_i]}{\sum_{i=1}^{N} \mathbb{1}[\pi_f(x_i)=t_i]}, \tag{31}$$

where we let the policy π_f of a model f to be $\pi_f(x) = 1$ if $f(x, t = 1) > f(x, t = 0)$ and $\pi_f(x) = 0$ otherwise. We compare CMDE with a total of 8 benchmark models: multi-task GP (CMGP (Alaa & Van Der Schaar, 2017)), CEVAE (Louizos et al., 2017), Generative Adversarial Nets (GANITE (Yoon et al., 2018)), X-learner (Künzel et al., 2019) of which the base learners are random forests (RF) and BART (Chipman et al., 2010), Counterfactual Regression Network (CFRNet (Shalit et al., 2017)) with 2-Wasserstein distance and Maximum Mean Discrepancy (MMD), and Deep Orthogonal Networks (DONUT (Hatt & Feuerriegel, 2021)). For the ACIC2019 dataset, we vary the training set size to compare algorithms in Figure 3, and observe that CMDE gives the lowest error (ϵ_{PEHE}) on the treatment effect. Furthermore, as shown in Table 1, CMDE demonstrates competitive performance compared to other benchmark models in terms of PEHE on the Twins dataset, and outperforms all other benchmark models in terms of policy risk on Jobs.

5.3. Datasets with Multi-modal Covariates

To demonstrate CMDE's strength in terms of handling high-dimensional and multi-modal covariates as described in Section 3.1, we further adopt 2 datasets: a semi-synthetic COVID-19 dataset built upon a collection of patients' chest X-ray images and their corresponding demographic infor-

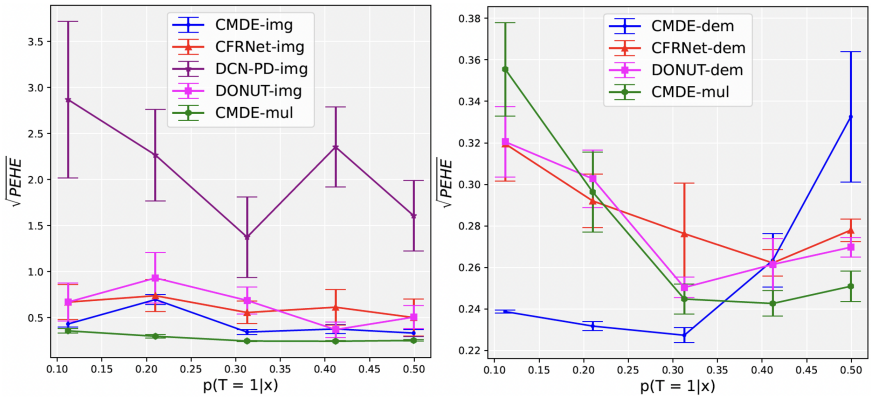

Figure 4. Results of CATE estimation ($\sqrt{\epsilon_{\text{PEHE}}}$) on the semi-synthetic COVID-19 dataset with different propensities where the covariates are either X-ray images (left) or demographic information (right). CMDE with multi-modal covariates (both images and demographic information) are marked as CMDE-mul in both figures. The lines and error bars represent mean and half of the standard deviation of $\sqrt{\epsilon_{\text{PEHE}}}$, respectively. The error of DCN-PD with demographic information is too high so we do not show it in the right figure for better visualization.

mation and diagnosis (e.g., COVID-19 or other viral pneumonia, bacterial pneumonia, fungal pneumonia, etc.) (Cohen et al., 2020) and a real-world dataset from the Student-Teacher Achievement Ratio (STAR) experiment (Word et al., 1990) with some features replaced by images with corresponding characteristics from the UTK dataset (Zhang et al., 2017). The details such as dataset pre-processing and model architectures can be found in Appendix D.3.

We first conduct experiments on the semi-synthetic COVID-19 dataset under different propensity score settings. The results of CATE estimation for CMDE and benchmark deep causal models (i.e. CFRNet (Shalit et al., 2017), Deep Counterfactual Network with Propensity Dropout (DCN-PD (Alaa et al., 2017)), and DONUT (Hatt & Feuerriegel, 2021)) are shown in Figure 4. It can be observed that, with multi-modal covariates (i.e. both X-ray images and demographic information), CMDE-mul achieves the lowest PEHE compared to other benchmarks with only the X-ray images as covariates (or model inputs). In addition, CMDE-mul demonstrates superior performance compared to all other benchmarks with demographic information as covariates when the propensity score (i.e. $P(T = 1|x)$) is close to 0.5. However, when the propensity score ranges from 0.1 to 0.3, CMDE with only demographic information yields better results. In other words, CMDE seems to benefit from using simpler covariate information when the control and treatment groups are relatively imbalanced. We attribute this phenomenon to a bias-variance tradeoff. Specifically, we note that the estimation variance tends to be larger when using simpler covariates compared to more complex ones. Conversely, incorporating additional information (e.g., bias reduction) through the use of more comprehensive covariates can lead to more accurate predictions. In situations where there is insufficient overlap between the control and treatment groups, the overall variance will increase. However, the increase in variance will be more substantial in cases involving complex covariates. Therefore, in scenarios with limited overlap, it may be advantageous to reduce the set of covariates, as the full estimation process may be dominated by error stemming from the variance term.

We also compare CMDE with the same 3 benchmark models on another real-world dataset from the STAR experiment which studied the effect of class size on the students' performance and test scores. Since the original dataset corresponds to a randomized control trial and the true average treatment effect (ATE) can be estimated directly, we use the following ATE error as our evaluation metric in this experiment:

$$\epsilon_{\text{ATE}} = \left| \text{ATE}_{\text{true}} - \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}\left[\hat{y}_i^{(1)}|x_i\right] - \mathbb{E}\left[\hat{y}_i^{(0)}|x_i\right] \right|. \quad (32)$$

The results of CATE estimation are visualized as bar plots as shown in Figure 5. We can see that CMDE outperforms other benchmark models with only the images or the students' information as covariates. Furthermore, with multi-modal covariates (i.e., both images and students' information), CMDE-mul yields the smallest ATE error with the lowest predictive uncertainty.

6. Discussion

NN architectures in CMDE As we show in Section 3, f_H, f_T, and f_{HT} need to have the same depth and initialization strategy to guarantee CMDE's convergence to a multi-task GP with ICM kernel. While this requirement is essential for theoretical considerations, it does not pose a significant practical limitation. In practice, if the distributions of control and treatment groups exhibit significant difference, it is recommended either to use LMC kernel or to employ different NN architectures for each function.

Limitations While CMDE has shown excellent results in our experiments, we identify some potential limitations to be addressed in future work. For example, the hyperparameters (e.g., width, depth, initial parameter values, etc.) of NNs in CMDE can be hard to tune for specific tasks. We also

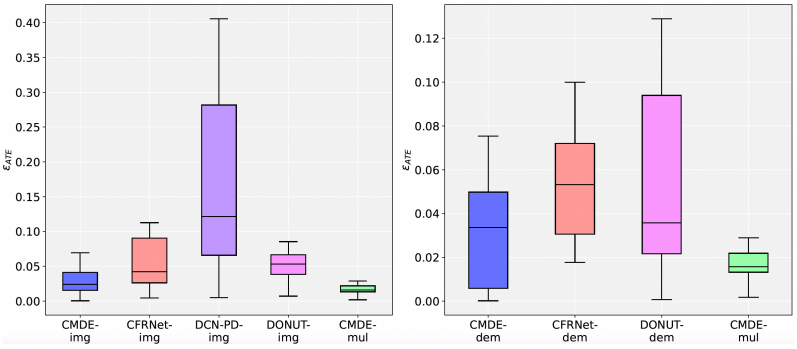

Figure 5. Box plots of CATE estimation (ϵ_{ATE}) on the STAR dataset where the covariates are either images (left) or the students' information (right). CMDE with multi-modal covariates (both images and students' information) are marked as CMDE-mul on the x-axis in both figures. The boxes extend from the 1st quantile to the 3rd quantile of ϵ_{ATE} with a line at the median. We do not show DCN-PD-dem on the right for cleaner visualization.

find the performance of CMDE, in some cases, is sensitive to the initial values of the coefficients applied to NNs (e.g., α_H, α_T, and α_{HT}), although we set these coefficients to be trainable. This requires us to have some prior knowledge about which type of information, group-specific or shared, is more dominant in specific datasets. A detailed discussion is given in Appendix E.

Applications We believe CMDE is applicable to a variety of real-world causal inference scenarios involving high-dimensional and multi-modal covariates, such as using medical records and images to estimate treatment effects in observational studies, or using A/B testing to determine the efficacy of a new version of user interface.

Societal Impact Currently we are not aware of any new potential negative societal impacts of our work; however, like all machine learning methods that could be applied in the wild, the societal impact will depend on the task at hand. For example, the STAR dataset uses pictures of individuals to estimate causal effects; image processing can encode unwanted biases and checks should be in place before the deployment of any such system.

7. Conclusion

We present a framework for estimating the causal effect of a treatment using a multi-task deep ensemble which learns both group-specific and shared information from control and treatment groups using separate neural networks. Theoretically, we demonstrate that our framework converges to a multi-task GP with an ICM/LMC kernel. We also provide empirical evidence of this relationship and visualize the contribution of each neural network components in our framework. Experimental results on various types of datasets demonstrate superior performance of CMDE compared to state-of-the-art approaches.

Acknowledgments

Research reported in this publication was supported by the National Institute of Biomedical Imaging and Bioengineering of the National Institutes of Health and the the National Institute of Mental Health under Award Number R01EB026937. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

Alaa, A. M. and Van Der Schaar, M. Bayesian inference of individualized treatment effects using multi-task gaussian processes. *Advances in neural information processing systems*, 30, 2017.

Alaa, A. M., Weisz, M., and Van Der Schaar, M. Deep counterfactual networks with propensity-dropout. *arXiv preprint arXiv:1706.05966*, 2017.

Almond, D., Chay, K. Y., and Lee, D. S. The costs of low birth weight. *The Quarterly Journal of Economics*, 120 (3):1031–1083, 2005.

Alvarez, M. A., Rosasco, L., Lawrence, N. D., et al. Kernels for vector-valued functions: A review. *Foundations and Trends® in Machine Learning*, 4(3):195–266, 2012.

Brooks-Gunn, J., Liaw, F.-r., and Klebanov, P. K. Effects of early intervention on cognitive function of low birth weight preterm infants. *The Journal of pediatrics*, 120(3): 350–359, 1992.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. Double/debiased machine learning for treatment and structural parameters, 2018.

Chipman, H. A., George, E. I., and McCulloch, R. E. Bart: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1):266–298, 2010.

Cohen, J. P., Morrison, P., Dao, L., Roth, K., Duong, T. Q., and Ghassemi, M. Covid-19 image data collection: Prospective predictions are the future. *arXiv preprint arXiv:2006.11988*, 2020.

Deshpande, S., Wang, K., Sreenivas, D., Li, Z., and Kuleshov, V. Deep multi-modal structural equations for causal effect estimation with unstructured proxies. In *Advances in Neural Information Processing Systems*, 2022.

Dorie, V., Hill, J., Shalit, U., Scott, M., and Cervone, D. Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition. *Statistical Science*, 34(1):43–68, 2019.

Drysdale, A. T., Grosenick, L., Downar, J., Dunlop, K., Mansouri, F., Meng, Y., Fetcho, R. N., Zebley, B., Oathes, D. J., Etkin, A., et al. Resting-state connectivity biomarkers define neurophysiological subtypes of depression. *Nature medicine*, 23(1):28–38, 2017.

Garriga-Alonso, A., Rasmussen, C. E., and Aitchison, L. Deep convolutional networks as shallow gaussian processes. *arXiv preprint arXiv:1808.05587*, 2018.

Goovaerts, P. et al. *Geostatistics for natural resources evaluation*. Oxford University Press on Demand, 1997.

Guo, R., Cheng, L., Li, J., Hahn, P. R., and Liu, H. A survey of learning causality with data: Problems and methods. *ACM Computing Surveys (CSUR)*, 53(4):1–37, 2020.

Hartford, J. S., Veitch, V., Sridhar, D., and Leyton-Brown, K. Valid causal inference with (some) invalid instruments. In *International Conference on Machine Learning*, pp. 4096–4106. PMLR, 2021.

Hatt, T. and Feuerriegel, S. Estimating average treatment effects via orthogonal regularization. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pp. 680–689, 2021.

He, B., Lakshminarayanan, B., and Teh, Y. W. Bayesian deep ensembles via the neural tangent kernel. *Advances in neural information processing systems*, 33:1010–1022, 2020.

Hill, J. L. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011.

Jensen, P. B., Jensen, L. J., and Brunak, S. Mining electronic health records: towards better research applications and clinical care. *Nature Reviews Genetics*, 13(6):395–405, 2012.

Jiang, Z., Zheng, T., and Carlson, D. Incorporating prior knowledge into neural networks through an implicit composite kernel. *arXiv preprint arXiv:2205.07384*, 2022.

Johansson, F., Shalit, U., and Sontag, D. Learning representations for counterfactual inference. In *International conference on machine learning*, pp. 3020–3029. PMLR, 2016.

Journel, A. G. and Huijbregts, C. J. *Mining geostatistics*. The Blackburn Press, 1976.

Kayser, C. and Shams, L. Multisensory causal inference in the brain. *PLoS biology*, 13(2):e1002075, 2015.

Künzel, S. R., Sekhon, J. S., Bickel, P. J., and Yu, B. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the national academy of sciences*, 116(10):4156–4165, 2019.

LaLonde, R. J. Evaluating the econometric evaluations of training programs with experimental data. *The American economic review*, pp. 604–620, 1986.

Lee, J., Bahri, Y., Novak, R., Schoenholz, S. S., Pennington, J., and Sohl-Dickstein, J. Deep neural networks as gaussian processes. *arXiv preprint arXiv:1711.00165*, 2017.

Li, Z. and Zhu, Z. A survey of deep causal model. *arXiv preprint arXiv:2209.08860*, 2022.

Louizos, C., Shalit, U., Mooij, J. M., Sontag, D., Zemel, R., and Welling, M. Causal effect inference with deep latent-variable models. *Advances in neural information processing systems*, 30, 2017.

Matthews, A. G. d. G., Hron, J., Turner, R. E., and Ghahramani, Z. Sample-then-optimize posterior sampling for bayesian linear models. In *NeurIPS Workshop on Advances in Approximate Bayesian Inference*, 2017.

Matthews, A. G. d. G., Rowland, M., Hron, J., Turner, R. E., and Ghahramani, Z. Gaussian process behaviour in wide deep neural networks. *arXiv preprint arXiv:1804.11271*, 2018.

Neal, R. M. Priors for infinite networks. In *Bayesian Learning for Neural Networks*, pp. 29–53. Springer, 1996.

Neal, R. M. *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media, 2012.

Page, B. I. and Shapiro, R. Y. Effects of public opinion on policy. *American political science review*, 77(1):175–190, 1983.

Rasmussen, C. E. Gaussian processes in machine learning. In *Summer school on machine learning*, pp. 63–71. Springer, 2003.

Rissanen, S. and Marttinen, P. A critical look at the consistency of causal estimation with deep latent variable models. *Advances in Neural Information Processing Systems*, 34:4207–4217, 2021.

Rosenbaum, P. R. and Rubin, D. B. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

Rubin, D. B. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331, 2005.

Shalit, U., Johansson, F. D., and Sontag, D. Estimating individual treatment effect: generalization bounds and algorithms. In *International Conference on Machine Learning*, pp. 3076–3085. PMLR, 2017.

Van Der Laan, M. J. and Rubin, D. Targeted maximum likelihood learning. *The international journal of biostatistics*, 2(1), 2006.

Word, E. et al. *Student/Teacher Achievement Ratio (STAR) Tennessee's K-3 Class Size Study. Final Summary Report 1985-1990.* Tennessee State Department of Education, 1990, 1990.

Yoon, J., Jordon, J., and Van Der Schaar, M. Ganite: Estimation of individualized treatment effects using generative adversarial nets. In *International Conference on Learning Representations*, 2018.

Zhang, L., Wang, Y., Ostropolets, A., Mulgrave, J. J., Blei, D. M., and Hripcsak, G. The medical deconfounder: assessing treatment effects with electronic health records. In *Machine Learning for Healthcare Conference*, pp. 490–512. PMLR, 2019.

Zhang, Z., Song, Y., and Qi, H. Age progression/regression by conditional adversarial autoencoder. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5810–5818, 2017.

A. Proof of Convergence to LMC Kernel

As stated in Section 3, by constructing $\hat{\mathbf{f}}$ as below:

$$\hat{f}_0(X) \coloneqq \sum_{q=1}^{Q} \alpha_H^q f_H^q(X) + \alpha_{HT}^q f_{HT}^q(X), \tag{33}$$

$$\hat{f}_1(X) \coloneqq \sum_{q=1}^{Q} \alpha_{HT}^q f_{HT}^q(X) + \alpha_T^q f_T^q(X), \tag{34}$$

where f_H^q, f_T^q, and f_{HT}^q share the same depth and initialization strategy for the same value of q, we can again separately compute each term in cov $\left(\hat{\mathbf{f}}(x), \hat{\mathbf{f}}(x')\right)$ as follows (again, note that the expectations are taken with respect to the parameters of the functions inside the expectation):

$$\begin{aligned}
\mathbb{E}[\hat{f}_0(x)\hat{f}_0(x')] &= \mathbb{E}\left[\left(\sum_{q=1}^{Q} \alpha_H^q f_H^q(x) + \alpha_{HT}^q f_{HT}^q(x)\right)\left(\sum_{q=1}^{Q} \alpha_H^q f_H^q(x') + \alpha_{HT}^q f_{HT}^q(x')\right)\right] \\
&= \sum_{q=1}^{Q} \left(\alpha_H^q\right)^2 \mathbb{E}\left[f_H^q(x) f_H^q(x')\right] + \left(\alpha_{HT}^q\right)^2 \mathbb{E}\left[f_{HT}^q(x) f_{HT}^q(x')\right], \qquad (35)\\
\mathbb{E}[\hat{f}_0(x)\hat{f}_1(x')] &= \mathbb{E}\left[\left(\sum_{q=1}^{Q} \alpha_H^q f_H^q(x) + \alpha_{HT}^q f_{HT}^q(x)\right)\left(\sum_{q=1}^{Q} \alpha_{HT}^q f_{HT}^q(x') + \alpha_T^q f_T^q(x')\right)\right] \\
&= \sum_{q=1}^{Q} \left(\alpha_{HT}^q\right)^2 \mathbb{E}\left[f_{HT}^q(x) f_{HT}^q(x')\right], \qquad (36)\\
\mathbb{E}[\hat{f}_1(x)\hat{f}_0(x')] &= \mathbb{E}\left[\left(\sum_{q=1}^{Q} \alpha_{HT}^q f_{HT}^q(x) + \alpha_T^q f_T^q(x)\right)\left(\sum_{q=1}^{Q} \alpha_H^q f_H^q(x') + \alpha_{HT}^q f_{HT}^q(x')\right)\right] \\
&= \sum_{q=1}^{Q} \left(\alpha_{HT}^q\right)^2 \mathbb{E}\left[f_{HT}^q(x) f_{HT}^q(x')\right], \qquad (37)\\
\mathbb{E}[\hat{f}_1(x)\hat{f}_1(x')] &= \mathbb{E}\left[\left(\sum_{q=1}^{Q} \alpha_{HT}^q f_{HT}^q(x) + \alpha_T^q f_T^q(x)\right)\left(\sum_{q=1}^{Q} \alpha_{HT}^q f_{HT}^q(x') + \alpha_T^q f_T^q(x')\right)\right] \\
&= \sum_{q=1}^{Q} \left(\alpha_{HT}^q\right)^2 \mathbb{E}\left[f_{HT}^q(x) f_{HT}^q(x')\right] + \left(\alpha_T^q\right)^2 \mathbb{E}\left[f_T^q(x) f_T^q(x')\right]. \qquad (38)
\end{aligned}$$

For (35) to (38), we get rid of the cross terms based on the fact that the parameters of different neural networks all have zero mean and are independent of each other. Note that f_H^q, f_T^q, and f_{HT}^q share the same depth and initialization strategy as stated in Theorem 3.1 for the same value of q, indicating that $\mathbb{E}[f_H^q(x)f_H^q(x')] = \mathbb{E}[f_T^q(x)f_T^q(x')] = \mathbb{E}[f_{HT}^q(x)f_{HT}^q(x')] = k_{\text{NN}}^q(x, x')$ for $q = 1, 2, ..., Q$ in the infinite width limit *a priori*. Therefore, we have:

$$\begin{aligned}
\mathbb{E}[\hat{f}_0(x)\hat{f}_0(x')] &= \sum_{q=1}^{Q} \left(\left(\alpha_H^q\right)^2 + \left(\alpha_{HT}^q\right)^2\right) k_{\text{NN}}^q(x, x'), \\
\mathbb{E}[\hat{f}_0(x)\hat{f}_1(x')] &= \mathbb{E}[\hat{f}_1(x)\hat{f}_0(x')] = \sum_{q=1}^{Q} \left(\alpha_{HT}^q\right)^2 k_{\text{NN}}^q(x, x'), \\
\mathbb{E}[\hat{f}_1(x)\hat{f}_1(x')] &= \sum_{q=1}^{Q} \left(\left(\alpha_{HT}^q\right)^2 + \left(\alpha_T^q\right)^2\right) k_{\text{NN}}^q(x, x').
\end{aligned}$$

Substituting the expressions above back into $\text{cov}\left(\hat{\mathbf{f}}(x), \hat{\mathbf{f}}(x')\right)$ as given in (14), we get:

$$\text{cov}\left(\hat{\mathbf{f}}(x), \hat{\mathbf{f}}(x')\right) = \sum_{q=1}^{Q} k_{\text{NN}}^{q}(x, x')\mathbf{B}_q \quad \text{where} \quad \mathbf{B}_q = \begin{bmatrix} (\alpha_H^q)^2 + (\alpha_{HT}^q)^2 & (\alpha_{HT}^q)^2 \\ (\alpha_{HT}^q)^2 & (\alpha_{HT}^q)^2 + (\alpha_T^q)^2 \end{bmatrix}. \tag{39}$$

This proves that $\hat{\mathbf{f}}$ will converge in distribution to a GP with zero mean and LMC kernel in the infinite width limit *a priori*.

B. Proof of Convergence to ICM Kernel for the Multiple-Treatment Case

As elaborated in Section 3, for multiple-treatment case $T \in \{1, ..., C\}$, we can construct $\hat{\mathbf{f}} = \left[\hat{f}_1, ..., \hat{f}_C\right]^T$ as follows:

$$\hat{f}_c(X) \coloneqq \sum_{d=1}^{c-1} \alpha_{dc} f_{dc}(X) + \alpha_c f_c(X) + \sum_{d=c+1}^{C} \alpha_{cd} f_{cd}(X) \; \forall \; c = 1, ..., C-1, \tag{40}$$

where f_c learns the group-specific information and f_{dc}, f_{cd} learn the shared information. Similar to Appendix A, we can calculate each separate term in $\text{cov}\left(\hat{\mathbf{f}}(x), \hat{\mathbf{f}}(x')\right)$ (again, note that the expectations are taken with respect to the parameters of the functions inside the expectation). For diagonal terms, we have:

$$\begin{aligned}
&\mathbb{E}[\hat{f}_c(x)\hat{f}_c(x')] \\
&= \mathbb{E}\left[\left(\sum_{d=1}^{c-1} \alpha_{dc} f_{dc}(x) + \alpha_c f_c(x) + \sum_{d=c+1}^{C} \alpha_{cd} f_{cd}(x)\right)\left(\sum_{d=1}^{c-1} \alpha_{dc} f_{dc}(x') + \alpha_c f_c(x') + \sum_{d=c+1}^{C} \alpha_{cd} f_{cd}(x')\right)\right] \\
&= \sum_{d=1}^{c-1} \alpha_{dc}^2 \mathbb{E}\left[f_{dc}(x) f_{dc}(x')\right] + \alpha_c^2 \mathbb{E}[f_c(x) f_c(x')] + \sum_{d=c+1}^{C} \alpha_{cd}^2 \mathbb{E}\left[f_{cd}(x) f_{cd}(x')\right].
\end{aligned} \tag{41}$$

For off-diagonal terms, we have:

$$\begin{aligned}
&\mathbb{E}[\hat{f}_c(x)\hat{f}_{c'}(x')] \\
&= \mathbb{E}\left[\left(\sum_{d=1}^{c-1} \alpha_{dc} f_{dc}(x) + \alpha_c f_c(x) + \sum_{d=c+1}^{C} \alpha_{cd} f_{cd}(x)\right)\left(\sum_{d=1}^{c'-1} \alpha_{dc'} f_{dc'}(x') + \alpha_c' f_{c'}(x') + \sum_{d=c'+1}^{C} \alpha_{c'd} f_{c'd}(x')\right)\right] \\
&= \alpha_{cc'}^2 \mathbb{E}[f_{cc'}(x) f_{cc'}(x')],
\end{aligned} \tag{42}$$

where $c < c'$. For $c > c'$, we have $\alpha_{cc'} = \alpha_{c'c}$ (i.e. the covariance matrix is symmetric). Here we again get rid of the cross terms based on the fact that the parameters of different neural networks all have zero mean and are independent of each other. If all neural networks in this formulation (i.e. a total of $C(C+1)/2$ networks) share the same depth and initialization strategy as stated in Theorem 3.1, indicating that $\mathbb{E}[f_c(x)f_c(x')] = k_{\text{NN}}(x, x') \; \forall \; c = 1, ..., C$ and $\mathbb{E}[f_{cc'}(x)f_{cc'}(x')] = k_{\text{NN}}(x, x') \; \forall \; c < c'$ and $c' = 1, ..., C$ in the infinite width limit *a priori*, then we can further write the diagonal and off-diagonal terms in $\text{cov}\left(\hat{\mathbf{f}}(x), \hat{\mathbf{f}}(x')\right)$ as:

$$\begin{aligned}
\mathbb{E}[\hat{f}_c(x)\hat{f}_c(x')] &= \left(\alpha_c^2 + \sum_{d=1}^{c-1} \alpha_{dc}^2 + \sum_{d=c+1}^{C} \alpha_{cd}^2\right) k_{\text{NN}}(x, x'), \\
\mathbb{E}[\hat{f}_c(x)\hat{f}_{c'}(x')] &= \alpha_{cc'}^2 k_{\text{NN}}(x, x').
\end{aligned}$$

With this, we derive:

$$\text{cov}\left(\hat{\mathbf{f}}(x), \hat{\mathbf{f}}(x')\right) = k_{\text{NN}}(x, x') \begin{bmatrix} \alpha_1^2 + \sum_{d=2}^{C} \alpha_{1d}^2 & \alpha_{12}^2 & \cdots & \alpha_{1C}^2 \\ \alpha_{12}^2 & \alpha_{12}^2 + \alpha_2^2 + \sum_{d=3}^{C} \alpha_{2d}^2 & \cdots & \alpha_{2C}^2 \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_{1C}^2 & \alpha_{2C}^2 & \cdots & \sum_{d=1}^{C-1} \alpha_{dC}^2 + \alpha_C^2 \end{bmatrix}. \tag{43}$$

This proves that $\hat{\mathbf{f}}$ will converge in distribution to a GP with zero mean and ICM kernel in the infinite width limit *a priori* when there exists a total of C treatments, i.e. $T \in \{1, ..., C\}$.

C. Two-Network Architecture for CMDE

The architecture of each baselearner in CMDE as presented in Section 3 can be simplified to the following two-network architecture (i.e. f_A and f_B):

$$\hat{Y}^{(0)} = \hat{f}_0(X) \coloneqq \alpha_0 f_A(X) + \beta_0 f_B(X), \tag{44}$$

$$\hat{Y}^{(1)} = \hat{f}_1(X) \coloneqq \alpha_1 f_A(X) + \beta_1 f_B(X). \tag{45}$$

Following a similar procedure as given in Appendices A and B, we have:

$$\mathbb{E}[\hat{f}_0(x)\hat{f}_0(x')] = \alpha_0^2\mathbb{E}[f_A(x)f_A(x')] + \beta_0^2\mathbb{E}[f_B(x)f_B(x')] = (\alpha_0^2 + \beta_0^2)k_{\text{NN}}(x, x'), \tag{46}$$

$$\mathbb{E}[\hat{f}_0(x)\hat{f}_1(x')] = \alpha_0\alpha_1\mathbb{E}[f_A(x)f_A(x')] + \beta_0\beta_1\mathbb{E}[f_B(x)f_B(x')] = (\alpha_0\alpha_1 + \beta_0\beta_1)k_{\text{NN}}(x, x'), \tag{47}$$

$$\mathbb{E}[\hat{f}_1(x)\hat{f}_0(x')] = \mathbb{E}[\hat{f}_0(x)\hat{f}_1(x')] = (\alpha_0\alpha_1 + \beta_0\beta_1)k_{\text{NN}}(x, x'), \tag{48}$$

$$\mathbb{E}[\hat{f}_1(x)\hat{f}_1(x')] = \alpha_1^2\mathbb{E}[f_A(x)f_A(x')] + \beta_1^2\mathbb{E}[f_B(x)f_B(x')] = (\alpha_1^2 + \beta_1^2)k_{\text{NN}}(x, x'). \tag{49}$$

The covariance function then becomes:

$$\text{cov}\left(\hat{\mathbf{f}}(x), \hat{\mathbf{f}}(x')\right) = k_{\text{NN}}(x, x') \begin{bmatrix} \alpha_0^2 + \beta_0^2 & \alpha_0\alpha_1 + \beta_0\beta_1 \\ \alpha_0\alpha_1 + \beta_0\beta_1 & \alpha_1^2 + \beta_1^2 \end{bmatrix}. \tag{50}$$

Therefore, $\hat{\mathbf{f}}$ will still converge in distribution to a GP with zero mean and ICM kernel in the infinite width limit *a priori* with this two-network architecture.

D. Details of Experimental Setup

D.1. Synthetic Dataset

We construct the synthetic dataset in Section 5.1 by following the steps below. For $i = 1, 2, ..., N$, do

$$\begin{aligned}
x_i &\sim \mathcal{N}\left(0, \sigma_x^2\right), \\
t_i &\sim \text{Bern}(p_i) \quad \text{where} \quad p_i = \frac{1}{1 + \exp(-x_i)}, \\
\xi_i &\sim \mathcal{N}\left(0, \sigma_\xi^2\right), \\
\mu_i^{(0)} &= 1 + \frac{1}{1 + \exp(-x_i)}, \\
\mu_i^{(1)} &= 2 + \frac{1}{1 + \exp(-x_i)}, \\
y_i^{(0)} &= \mu_i^{(0)} + \xi_i, \\
y_i^{(1)} &= \mu_i^{(1)} + \xi_i, \\
y_i &= y_i^{(0)} \quad \text{if} \quad t_i = 0 \quad \text{else} \quad y_i^{(1)}.
\end{aligned}$$

For our experiment, we set $\sigma_x^2 = 9$ and $\sigma_\xi^2 = 0.0025$ and sample $N = 3000$ data points. The CMDE model consists of 10 estimators where f_H, f_T, and f_{HT} in each estimator are single-hidden-layer neural networks with ReLU activation and 2048 units in the hidden layer. We set the initial values of α_H, α_T, and α_{HT} to be $\alpha_H = 0$, $\alpha_T = 0$, and $\alpha_{HT} = 1$. All weight and bias parameters in f_H, f_T, and f_{HT} are independently drawn from a normal distribution $\mathcal{N}(0, \sigma_w^2 I)$ *a priori* and $\sigma_w^2 = 0.1$.

D.2. Benchmark Datasets

We elaborate the data pre-processing details in the sub-sections below. The model hyperparameter details are listed in Table 2. Also, note that for Twins and Jobs dataset, we use both the training and validation set to evaluate the models for in-sample setting and just the test set to evaluate the models for out-of-sample setting. We repeat the experiments on Twins and Jobs dataset 10 times and report the mean and the standard deviation as given in Table 1.

	ACIC	Twins	Jobs
CMDE	number of estimators $= 10$ LMC kernel ($Q = 2$), depth $= 2$, width $= 512$, $\alpha_H^1 = \alpha_T^1 = \alpha_{HT}^1 = 1$, $\alpha_H^2 = \alpha_T^2 = \alpha_{HT}^2 = 1$, softplus activation	number of estimators $= 10$ LMC kernel ($Q = 2$), depth $= 2$, width $= 512$, $\alpha_H^1 = \alpha_T^1 = 1, \alpha_{HT}^1 = 0.1$, $\alpha_H^2 = \alpha_T^2 = 1, \alpha_{HT}^2 = 0.1$, tanh activation	number of estimators $= 10$ LMC kernel ($Q = 2$), depth $= 2$, width $= 512$, $\alpha_H^1 = \alpha_T^1 = 1, \alpha_{HT}^1 = 0.1$, $\alpha_H^2 = \alpha_T^2 = 1, \alpha_{HT}^2 = 0.1$, tanh activation
CMGP	LMC kernel with RBF base kernel	N/A	LMC kernel with RBF base kernel
CEVAE	†	†	†
GANITE	$k_G = k_I = 256$, depth $= 0$, $h_{dim} = 100$, $\alpha = 0.1$, $\beta = 0$	$k_G = k_I = 128$, depth $= 5$, $h_{dim} = 8$, $\alpha = 2$, $\beta = 2$	$k_G = k_I = 128$, depth $= 3$, $h_{dim} = 4$, $\alpha = 1$, $\beta = 5$
X-learner-RF	number of estimators = 100	number of estimators = 100	N/A
X-learner-BART	number of estimators = 100	number of estimators = 100	N/A
CFRNet-Wass	depth (ϕ and h) $= 2$, width (ϕ and h) $= 512$, $\alpha = 0.1$, ReLU activation	depth (ϕ and h) $= 2$, width (ϕ and h) $= 512$, $\alpha = 1$, tanh activation	depth (ϕ and h) $= 2$, width (ϕ and h) $= 512$, $\alpha = 1$, tanh activation
CFRNet-MMD	depth (ϕ and h) $= 2$, width (ϕ and h) $= 512$, $\alpha = 0.1$, ReLU activation	depth (ϕ and h) $= 2$, width (ϕ and h) $= 512$, $\alpha = 1$, tanh activation	depth (ϕ and h) $= 2$, width (ϕ and h) $= 512$, $\alpha = 1$, tanh activation
DONUT	depth (ϕ and h) $= 2$, width (ϕ and h) $= 512$, ReLU activation	depth (ϕ and h) $= 2$, width (ϕ and h) $= 512$, tanh activation	depth (ϕ and h) $= 2$, width (ϕ and h) $= 512$, tanh activation

Table 2. Model hyperparameters used for CMDE and other benchmark models in the ACIC2019, Twins, and Jobs experiments. † To save space, for CEVAE, please refer to the source code for hyperparameter details.

D.2.1. Atlantic Causal Inference Conference (ACIC) Dataset

The covariates in ACIC2019 dataset are either simulated or drawn from publicly available datasets. We take the high-dimensional version (where we have 185 covariates in total) for our experiments. The full training and test sets contain a total of 6.4M and 16K data points, respectively. Due to time and memory constraints, we pick a small subset containing 2000 data points from each of the training and test sets. The download link is provided below:
https://sites.google.com/view/acic2019datachallenge/data-challenge?pli=1

D.2.2. Twins Dataset

The Twins dataset contains the information of twin births in the United States from 1989 to 1991. It contains 40 covariates pertaining to pregnancy, twin births, and parents. The treatment is defined as $T = 1$ as being the heavier twin and $T = 0$ as being the lighter twin. The outcome is defined as the 1-year mortality. The full dataset contains a total of 11400 data points and we average over 10 train-validation-test splits with a ratio of 56:24:20.

D.2.3. Jobs Dataset

The Jobs dataset studied by LaLonde is a widely used benchmark where the treatment T is job training and the outcome Y is the individual's income in 1975. The covariates include 8 variables such as age, education, race, and income in 1974. The dataset consists of a randomized portion based on the National Supported Work program (722 samples) and a non-randomized portion acquired from observational studies (2490 samples). Before conducting the experiment, we convert Y (income in 1975) into binary outcomes (i.e. employed/unemployed or $\mathbb{1}[Y = 0]$). The test set is sampled *only from the randomized portion* and we average over 10 train-validation-test splits with a ratio of 56:24:20.

D.3. Datasets with Multi-modal Covariates

We give the details of data generation and pre-processing for each experiment in the sub-sections below. For CMDE and benchmark deep causal models, we use a convolutional neural network (CNN) architecture when we have images as covariates and a fully connected neural network architecture when we have tabular data (e.g., demographic information in COVID-19 dataset) as covariates. The details of model architectures are displayed in Table 3. We repeat the experiments on both datasets 10 times and report the corresponding statistics as given in Figures 4 and 5.

D.3.1. Data Generation Procedure of the Semi-synthetic COVID-19 Dataset

We create a semi-synthetic dataset based on a publicly available COVID-19 X-ray dataset. The dataset includes 951 images and some other demographic and image-related information, collected from several public sources (Cohen et al., 2020). After data cleaning and imputation, 857 samples are used in our analysis and we use a train-validation-test splito ratio of 40:20:40. The variables we include are patient id, offset, sex, age, RT-PCR-positive, survival, intubated, intubation-present, went-icu, in-icu.

We generate potential outcomes using both demographic and image information. For image, we categorize the diagnosis of the X-ray into the following categories: viral pneumonia, bacterial pneumonia, fungal pneumonia, pneumonia caused by other causes (lipoid and aspiration), pneumonia by unknown cause, and tuberculosis. We use these categories to represent image information. The potential outcome Y is defined as general overall severity of diseases and T is the binary treatment. Larger value of Y indicates worse prognosis. The potential outcomes are generated by the following equations:

$$
\begin{aligned}
Y(0) &= \beta_0 + \beta_1^T X_{\text{d}} + \beta_2^T X_{\text{im}} + \beta_3^T X_{\text{d}} \otimes X_{\text{d}} + \epsilon_0, \\
Y(1) &= \beta_0 + \beta_1^T X_{\text{d}} + \beta_2^T X_{\text{im}} + \beta_3^T X_{\text{d}} \otimes X_{\text{d}} + \beta_{t1}^T X_{\text{im}} + \beta_{t2}^T X_{\text{d}} \otimes X_{\text{im}} + \epsilon_1,
\end{aligned}
$$

where ϵ_0 ,$\epsilon_1 \sim N(0, 0.1)$, X_{d} is the vector of demographic variables, X_{im} is the vector of diagnosis categories and $\otimes$ is the symbol of Kronecker product. We assign values of β in a clinically meaningful way and β_3 and β_{t2} are sparse matrices in the sense that only a few variables would interact with each other. More details of the data generating process such as the exact values of β could be found in our source code.

For treatment, we consider two main scenarios: observational study randomized study. The results of observational study setting are shown in Figure 4, where treatment depends on some covariates:

$$
\begin{aligned}
p(x_i) &= \frac{1}{1 + \exp(-\beta_t x_i)}, \\
T &\sim \text{Bern}\left(\frac{p_2 p(x_i)}{p(X)/N}\right),
\end{aligned}
$$

where $X = (X_{\text{d}}, X_{\text{im}}), p_2 = \{0.1, 0.2, 0.3, 0.4, 0.5\}$. The $\frac{p_2 p(x_i)}{p(X)/N}$ term is to control the mean of $p(x_i)$ to mimic an unbalanced assignment mechanism and $P(T = 1|X)$ is called the propensity score. In randomized study setting, treatment does not depend on any covariates. Specifically, we set $T \sim \text{Bern}(p_1)$ where $p_1 = \{0.1, 0.2, 0.3, 0.4, 0.5\}$ to mimic an unbalanced assignment mechanism. The results of CATE estimation in this setting are presented in Figure D1. It turns out that the conclusions derived from this figure are very similar to the ones we state in Section 5.3.

D.3.2. Data Pre-processing Procedure of the STAR Dataset

The effect of class size on student's achievement is an important topic in the American K-12 education system. To study the effect, the State Department of Education in Tennessee conducted a four-year longitudinal, class-size randomized study called The Student/Teacher Achievement Ratio (STAR) from 1985 to 1989. Using the first graders' data from the STAR project, we estimate the effect of class size on class-level mathematical performance on a standardized test. Since the original dataset corresponds to a randomized study, the true effect of class size can be estimated directly. Following similar covariate selection as Deshpande et al. (2022), we use highest degree obtained by teacher, career ladder position of teacher, number of years of experience of teacher, and teacher's race as numerical features. To construct multi-modal covariates, the students' gender and ethnicity are replaced by images of corresponding characteristics from the UTK dataset (Zhang et al., 2017). All categorical covariates are converted into one-hot encoding. Test scores and all continuous covariates are normalized using min-max normalization. Observations with any missing values are filtered out, resulting in a total of 6563 data points. The train-validation-test split ratio is again set to be 40:20:40.

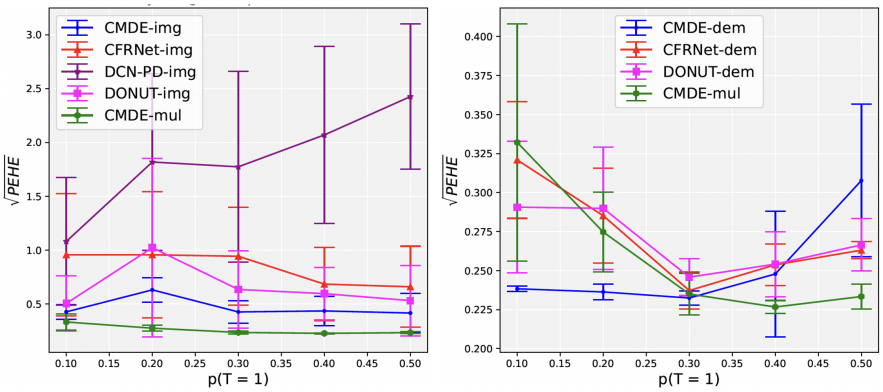

Figure D1. Results of CATE estimation ($\sqrt{\epsilon_{\text{PEHE}}}$) on the semi-synthetic COVID-19 dataset in randomized study setting where the covariates are either X-ray images (left) or demographic information (right). CMDE with multi-modal covariates (both images and demographic information) are marked as CMDE-mul in both figures. The lines and error bars represent mean and half of the standard deviation of $\sqrt{\epsilon_{\text{PEHE}}}$, respectively. The error of DCN-PD with demographic information is too high so we do not show it in the right figure for better visualization.

	COVID-19 dataset		STAR dataset	
	Image covariates	Tabular data covariates	Image covariates	Tabular data covariates
CMDE	#estimators $= 10$ ICM kernel, depth $= 2$, #channels in hidden layers $= 64$, filter size $= 3$, stride $= 1$ $\alpha_H = \alpha_T = 1$, $\alpha_{HT} = 0.5$, softplus activation	#estimators $= 10$ ICM kernel, depth $= 2$, width $= 512$, $\alpha_H = \alpha_T = 1$, $\alpha_{HT} = 0.5$, softplus activation	#estimators $= 10$ ICM kernel, depth $= 2$, #channels in hidden layers $= 64$, filter size $= 3$, stride $= 1$ $\alpha_H = \alpha_T = 1$, $\alpha_{HT} = 0.1$, softplus activation	#estimators $= 10$ ICM kernel, depth $= 2$, width $= 512$, $\alpha_H = \alpha_T = 1$, $\alpha_{HT} = 0.1$, softplus activation
CFRNet	depth (ϕ and h) $= 2$, #channels in hidden layers of ϕ and $h = 64$, filter size $= 3$, stride $= 1$, $\alpha = 0.01$, softplus activation	depth (ϕ and h) $= 2$, width (ϕ and h) $= 512$, $\alpha = 0.01$, softplus activation	depth (ϕ and h) $= 2$, #channels in hidden layers of ϕ and $h = 64$, filter size $= 3$, stride $= 1$, $\alpha = 0.01$, softplus activation	depth (ϕ and h) $= 2$, width (ϕ and h) $= 512$, $\alpha = 0.01$, softplus activation
DCN-PD	depth (shared and idiosyncratic networks) $= 2$, #channels in hidden layers of shared and idiosyncratic networks $= 64$, filter size $= 3$, stride $= 1$, softplus activation	depth (shared and idiosyncratic networks) $= 2$, width (shared and idiosyncratic networks) $= 512$, softplus activation	depth (shared and idiosyncratic networks) $= 2$, #channels in hidden layers of shared and idiosyncratic networks $= 64$, filter size $= 3$, stride $= 1$, softplus activation	depth (shared and idiosyncratic networks) $= 2$, width (shared and idiosyncratic networks) $= 512$, softplus activation
DONUT	depth (ϕ and h) $= 2$, #channels in hidden layers of ϕ and $h = 64$, filter size $= 3$, stride $= 1$, softplus activation	depth (ϕ and h) $= 2$, width (ϕ and h) $= 512$, softplus activation	depth (ϕ and h) $= 2$, #channels in hidden layers of ϕ and $h = 64$, filter size $= 3$, stride $= 1$, softplus activation	depth (ϕ and h) $= 2$, width (ϕ and h) $= 512$, softplus activation

Table 3. Model hyperparameters used for CMDE and other deep causal benchmark models in the COVID-19 X-ray image and STAR experiments

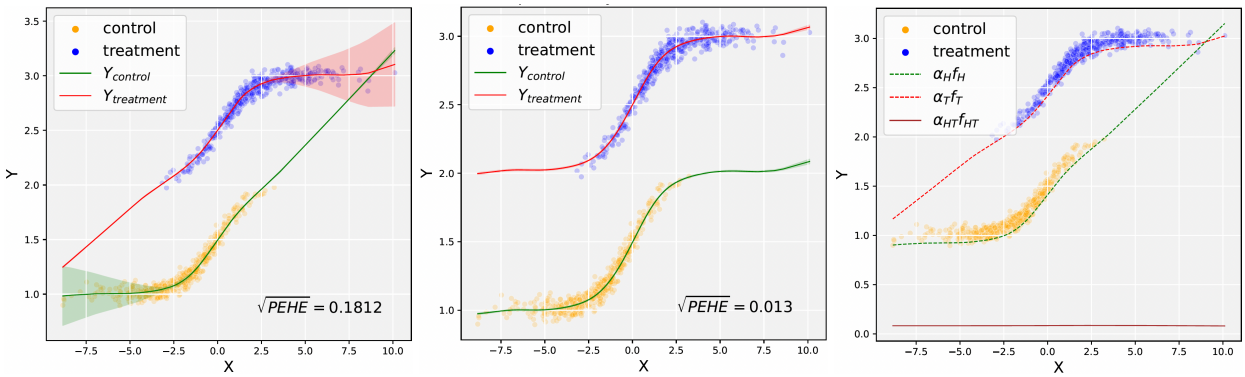

Figure E1. Predictions for the control and treatment groups on the synthetic dataset by CMDE (left) with $\alpha_H = 1, \alpha_T = 1, \alpha_{HT} = 0$ and multi-task GP with ICM kernel (middle) where dots represent observed samples, lines represent mean predictions, and shaded regions represent predicted values within 2 standard deviations. In addition, we also plot the contribution of group-specific and shared components for CMDE (right). It can be observed that with inappropriate coefficient initializaiton, CMDE fails to capture the response surfaces of the control and treatment groups and yields a much larger error on treatment effect estimation than multi-task GP.

E. Discussion on Fine-tuning of Coefficients

As mentioned in the Discussion, CMDE is sensitive to the initial values of the coefficients applied in NNs (e.g. α_H, α_T, and α_{HT}), thus requiring us to have some prior knowledge about which type of information, group-specific or shared, is more dominant. This is dependent on the task at hand. For example, in the synthetic data example in Section 5.1, the shared information between the two groups (i.e, shape of the two curves) is important for estimating the potential outcomes when there is no group-specific information. Note that we refer to the group-specific information as the distinct features of the covariates X in each group, and here the two curves are only off by a constant instead of anything related to the covariate X (see Appendix D.1 for the data generation procedure). Therefore, we initialize the coefficients to be $\alpha_H = \alpha_T = 0$ and $\alpha_{HT} = 1$. To verify this point, we also repeat this experiment by setting $\alpha_H = \alpha_T = 1$ and $\alpha_{HT} = 0$. We realize that without correctly capturing the shape of the two curves, CMDE performs much worse on the treatment effect estimation as shown in Figure E1. On the contrary, the Twins dataset contains 40 covariates pertaining to pregnancy, twin births, and parents. The treatment is defined as $T = 1$ as being the heavier twin and $T = 0$ as being the lighter twin. The outcome is defined as the 1-year mortality rate. Here, our prior knowledge is that the unique features of each child will contribute more to the mortality rate, so we initialize the coefficients to be $\alpha_H = \alpha_T = 1$ and $\alpha_{HT} = 0.1$.

F. Accessibility of the Datasets

All datasets used in our experiments are available at `https://github.com/jzy95310/ICK/tree/main/data` and are released under the MIT license. For ACIC, Twins, Jobs, and STAR, the original datasets have an open-access license and are publicly available. For COVID-19 experiment, the original dataset is available at: `https://github.com/ieee8023/covid-chestxray-dataset` where each image has a specified license including Apache 2.0, CC BY-NC-SA 4.0, and CC BY 4.0. All other files and scripts are released under a CC BY-NC-SA 4.0 license.