Disentangled Latent Representation Learning for Tackling the Confounding M-Bias Problem in Causal Inference

Debo Cheng[‡†], Yang Xie[¶†], Ziqi Xu[‡†], Jiuyong Li[‡*], Lin Liu[‡], Jixue Liu[‡], Yinghao Zhang[¶] and Zaiwen Feng[¶*]
[¶] College of Informatics, Huazhong Agricultural University, Wuhan, China
[‡] UniSA STEM, University of South Australia, Adelaide, Australia

Abstract—**In causal inference, it is a fundamental task to estimate the causal effect from observational data. However, latent confounders pose major challenges in causal inference in observational data, for example, confounding bias and M-bias. Recent data-driven causal effect estimators tackle the confounding bias problem via balanced representation learning, but assume no M-bias in the system, thus they fail to handle the M-bias. In this paper, we identify a challenging and unsolved problem caused by a variable that leads to confounding bias and M-bias simultaneously. To address this problem with co-occurring M-bias and confounding bias, we propose a novel Disentangled Latent Representation learning framework for learning latent representations from proxy variables for unbiased Causal effect Estimation (DLRCE) from observational data. Specifically, DLRCE learns three sets of latent representations from the measured proxy variables to adjust for the confounding bias and M-bias. Extensive experiments on both synthetic and three real-world datasets demonstrate that DLRCE significantly outperforms the state-of-the-art estimators in the case of the presence of both confounding bias and M-bias.**

Index Terms—**Causal Inference, Causal Effect Estimation, Confounding Bias, M-bias, Disentangled Representation Learning, Latent Confounders**

I. Introduction

Causal effect estimation is an important approach to understand the underlying causal mechanisms of problems in various areas, such as economics [1], [2], epidemiology [3], medicine [4] and computer science [5], [6]. Randomised control trials (RCTs) are the gold standard for assessing causal effects, but conducting RCTs can often be infeasible or impractical due to ethical considerations, high costs, or time constraints [1], [7]. Therefore, estimating causal effects from observational data has emerged as an important alternative strategy of RCTs [1], [6], [8].

The presence of confounding bias, caused by confounders, creates challenges when using observational data for causal effect estimation [1], [5]. A confounder is a variable that influences both the treatment variable, denoted as W, and the outcome variable, denoted as Y. Many works [9], [10] consider all measured variables, denoted as $\mathbf{X}$ as the set of confounders as shown in Fig. 1 (a), in which $\mathbf{X}$ causes both W and Y simultaneously in the causal DAG. To address the confounding bias, various methods that employ covariate adjustment [5], [8] or confounding balancing [9], [11] have been developed, under the unconfoundedness assumption [1], [6]. For example, Shalit et al. [9] proposed a representation learning based counterfactual regression framework.

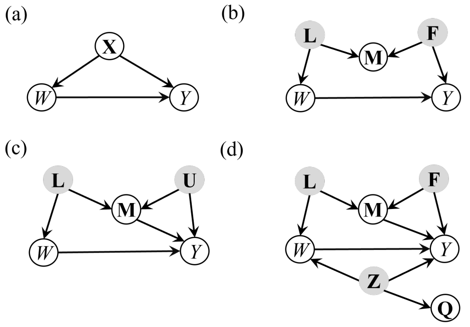


Fig. 1. In the figure, $\mathbf{L}$, $\mathbf{F}$ and $\mathbf{Z}$ each represent a set of latent variables (indicated by shaded circles), and the other variables are measured. W and Y are the treatment and outcome variables respectively. $\mathbf{X}$ is the set of covariates, $\mathbf{M}$ is a set of M-bias variables, and $\mathbf{Q}$ is a set of proxy variables of $\mathbf{Z}$. Note that $\mathbf{M}$ is also a set of proxies of $\mathbf{L}$ and $\mathbf{F}$. In the four causal DAGs, (a) a simple case with confounding bias caused by $\mathbf{X}$ w.r.t. (W, Y); (b) an illustration of M-structure: $W \leftarrow \mathbf{L} \rightarrow \mathbf{M} \leftarrow \mathbf{F} \rightarrow Y$. Conditioning on $\mathbf{M}$ results in M-bias wrt., (W, Y); (c) A DAG illustrating the problem identified in this work, $\mathbf{M}$ serves as a set of M-bias variables and a set of confounders, w.r.t. (W, Y); (d) the DAG showing the problem studied in this work.

Nevertheless, some variables are unmeasured or unobserved due to various uncontrollable factors in many real-world applications [13], [14], consequently, the unconfoundedness assumption is violated due to the presence of the latent variables. These latent variables result in not only confounding bias, but also M-bias. M-bias is introduced by conditioning on a variable that is caused by two latent variables. We call a measured variable an M-bias variable if it is a direct effect variable of two or more latent variables, forming an "M-structure". For example, in Fig. 1 (b), $\mathbf{M}$ is a set of measured

[†] These authors contributed equally.

[*] Corresponding authors: J. Li (Jiuyong.Li@unisa.edu.au) and Z. Feng (Zaiwen.Feng@mail.hzau.edu.cn).

A DAG (directed acyclic graph) is a graph with directed edges only and contains no cycles. More details of graph terminologies can be found in [5].

The unconfoundedness assumption holds when there are no unmeasured confounders for each pair measured variables [1], [12].

variables and is the set of M-bias variables since $\mathbf{L}$ and $\mathbf{F}$ are two sets of latent variables and they are direct causes of $\mathbf{M}$. In this case, when $\mathbf{M}$ is considered as confounders such that it is adjusted for estimating the causal effect of W on Y, a spurious association between W and Y occurs since the path $W \leftarrow \mathbf{L} \rightarrow \mathbf{M} \leftarrow \mathbf{F} \rightarrow Y$ is opened when $\mathbf{M}$ is given (conditioned on). The spurious association along the path causes a biased estimation wrt., (W, Y). Such a bias is known as M-bias in causal effect literature [3], [15], [16].

Excluding the M-bias variable in an adjustment set is a common way of dealing with M-bias [5], [8], [15], [17]. For example, Enter et al. [18] and Cheng et al. [19] use an anchor node to perform conditioning independence/dependence tests for identifying valid adjustment sets to exclude the M-bias variable from an adjustment set.

Dealing with M-bias becomes complex when a variable acts as both an M-bias variable and a confounder, and we call the problem confounding M-bias problem. We call a variable acting both as an M-bias variable and a confounder a *confounding M-bias variable*, and the problem with a confounding M-bias variable the *confounding M-bias problem*, and Fig. 1 (c) shows an example of the problem. $\mathbf{M}$ is a set of M-bias variables based on path $W \leftarrow \mathbf{L} \rightarrow \mathbf{M} \leftarrow \mathbf{F} \rightarrow Y$ and a set of confounders based on path $W \leftarrow \mathbf{L} \rightarrow \mathbf{M} \rightarrow Y$. Using statistical methods, such as Entner et al. 's [18] and Cheng et al. 's [19], whether adjusting for $\mathbf{M}$ or not, leads to a biased causal effect of W on Y. There is no immediate solution to the confounding M-bias problem but the problem is real. We substantiate the example in Fig. 1 (c) by letting W be 'Study time', Y be 'Academic performance', M be 'Personal interests', L be 'Personal experience', and F be 'IQ'. When 'Personal experience' and 'IQ' are unmeasured, it is impossible to estimate the causal effect of 'Study time' on 'Academic performance' since 'Personal interests' is both an M-bias variable and a confounder.

In this paper, we will solve the problem by leveraging the representation learning technique to recover the information of latent variables, and then using observed variables and the learned representations to unbiasedly estimate causal effect in the presence of confounding M-bias variables.

The causal graph of the problem that is considered in this paper is shown in Fig. 1 (d). To make our solution covers a broader range of practical scenarios, we also consider the latent confounders whose proxies are observed. For example, in Fig. 1 (d), if the unobserved confounder Z represents 'Teaching quality', and Q represents 'Schools', then Q can be used as the proxy of Z.

In summary, this paper makes the following contributions:

- We identify the confounding M-bias problem in causal effect estimation, which is a realistic problem. The problem has not been identified or studied previously.
- We propose a solution, the DLRCE (Disentangled Latent Representation learning for unbiased Causal effect Estimation) algorithm to resolve the confounding M-bias problem. To the best of our knowledge, there are no solutions to this problem. Furthermore, we prove the soundness of the solution.
- We conduct an empirical evaluation to assess the performance of the proposed algorithm on both synthetic and real-world datasets, in comparison to state-of-the-art methods. The experimental results reveal that the proposed algorithm effectively mitigates confounding bias and handles M-bias, and demonstrate its superior performance compared to the baseline methods.

II. Preliminaries

Throughout the paper, we use uppercase and lowercase letters to denote variables and their values, respectively. We use bold-faced uppercase and lowercase letters to represent a set of variables and their corresponding values, respectively.

A graph is a Directed Acyclic Graph (DAG) when it consists of directed edges (represented by →) and does not contain cycles. In this paper, we use $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ to denote a DAG, where $\mathbf{V} = \mathbf{X} \cup \mathbf{U} \cup \{W, Y\}$ represents the set of nodes, i.e., $\mathbf{X}$ the set of measured variables, $\mathbf{U}$ the set of latent confounders, W the treatment variable, and Y the outcome variable, and $\mathbf{E} \subseteq \mathbf{V} \times \mathbf{V}$ indicates the set of directed edges.

In DAG $\mathcal{G}$, two nodes are *adjacent* when there exists a directed edge → between them. In a causal DAG, a directed edge $X_i \rightarrow X_j$ signifies that variable X_i is a cause of variable X_j and X_j is an effect variable of X_i. A path π from X_i to X_k is a directed or causal path if all edges along it are directed towards X_k. If there is a directed path π from X_i to X_k, X_i is known as an ancestor of X_k and X_k is a descendant of X_i. The sets of ancestors and descendants of a node X are denoted as $An(X)$ and $De(X)$, respectively.

A causal DAG $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ is employed to represent the underlying causal mechanism of a system. The following presented Markov property and faithfulness assumptions are often assumed in causal inference with a causal DAG.

Definition 1 (Markov property [5], [20]): Given a DAG $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ and the joint probability distribution $P(\mathbf{V})$, $\mathcal{G}$ satisfies the Markov property if for $\forall V_i \in \mathbf{V}$, V_i is probabilistically independent of all of its non-descendants in $P(\mathbf{V})$, given the parent nodes of V_i.

Definition 2 (Faithfulness [20]): Given a DAG $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ and the joint probability distribution $P(\mathbf{V})$, $\mathcal{G}$ is faithful to a joint distribution $P(\mathbf{V})$ over $\mathbf{V}$ if and only if every independence present in $P(\mathbf{V})$ is entailed by $\mathcal{G}$ and satisfies the Markov property. A joint distribution $P(\mathbf{V})$ over $\mathbf{V}$ is faithful to $\mathcal{G}$ if and only if $\mathcal{G}$ is faithful to $P(\mathbf{V})$.

When the Markov property and faithfulness are satisfied, the dependency/independency relations between variables in the probability distribution $P(\mathbf{V})$ can be inferred from the corresponding causal DAG $\mathcal{G}$ [5], [20]. To determine the conditional independence relationships implied by $\mathcal{G}$, Pearl introduced a graphical criterion, named d-separation.

Definition 3 (d-separation [5]): A path π in a DAG $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ is said to be d-separated (or blocked) by a set of nodes $\mathbf{S}$ if and only if (i) π contains a chain $V_i \rightarrow V_k \rightarrow V_j$ or a fork $V_i \leftarrow V_k \rightarrow V_j$ such that the middle node V_k is in $\mathbf{S}$, or

(ii) π contains a collider V_k such that V_k is not in $\mathbf{S}$ and no descendant of V_k is in $\mathbf{S}$. A set $\mathbf{S}$ is said to d-separate V_i from V_j ($V_i \perp\!\!\!\perp V_j \mid \mathbf{S}$) if and only if $\mathbf{S}$ blocks every path between V_i to V_j. Otherwise, they are said to be d-connected by $\mathbf{S}$, denoted as $V_i \not\perp\!\!\!\perp V_j \mid \mathbf{S}$.

In this work, we assume that the set $\mathbf{X}$ contains pretreatment variables, i.e., all variables in $\mathbf{X}$ are measured before the treatment W is applied and the outcome Y is measured. Based on the potential outcome framework [1], [21], [22], for an individual, with respect to a binary treatment, there are two potential outcomes, $Y(W=1)$ and $Y(W=0)$ respectively. $Y(W=w)$ is the observed outcome when the treatment W is equal to w. Note that, for an individual, we can only observe one of $Y(W=1)$ and $Y(W=0)$ relative to the factual treatment we have applied. The unobserved potential outcome is known as the counterfactual outcome [1], [21], [22]. The individual treatment effect (ITE) for i is defined as:

$$ITE_i = Y_i(W=1) - Y_i(W=0) \tag{1}$$

The average treatment effect (ATE) of W on Y at the population level is defined as:

$$ATE(W,Y) = \mathbb{E}[Y_i(W=1) - Y_i(W=0)] \tag{2}$$

where $\mathbb{E}$ indicates the expectation function. In graphical causal modelling, the ATE is defined as the following using "do" operation introduced by Pearl [5], and defined as:

$$\begin{aligned} &ATE(W,Y) = \\ &\mathbb{E}[Y \mid do(W=1)] - \mathbb{E}[Y \mid do(W=0)] \end{aligned} \tag{3}$$

where $do(\cdot)$ is the do-operator.

The ITE defined in Eq.(1) cannot be obtained from data directly since only one potential outcome is observed for an individual. Instead, a number of data-driven methods have been developed for ATE estimation from data. To estimate the causal effect of W on Y unbiasedly from observational data, covariate adjustment [5], [13], [23] and confounding balance [9] are commonly used method for eliminating the confounding bias. It is critical to discover an adjustment set to eliminate the confounding bias when estimating the causal effect of W on Y. The back-door criterion, a well-known graphical criterion, can be applied to discover such an adjustment set $\mathbf{S} \subseteq \mathbf{X}$ in $\mathcal{G}$.

Definition 4 (Back-door criterion [5]): In a DAG $\mathcal{G} = (\mathbf{V}, \mathbf{E})$, for the pair of variables $(W,Y) \in \mathbf{V}$, a set of variables $\mathbf{S} \subseteq \mathbf{V} \setminus \{W,Y\}$ is said to satisfy the back-door criterion in the given DAG $\mathcal{G}$ if (i) $\mathbf{S}$ does not contain a descendant node of W; and (ii) $\mathbf{S}$ blocks every back-door path between W and Y (the paths between W and Y starting with an arrow into W). A set $\mathbf{S}$ is referred to as a *back-door set* relative to (W,Y) in $\mathcal{G}$ if $\mathbf{S}$ satisfies the back-door criterion relative to (W,Y) in $\mathcal{G}$. Therefore, adjusting for the back-door set $\mathbf{S}$, we have $ATE(W,Y) = \mathbb{E}[Y \mid w=1, \mathbf{S}=\mathbf{s}] - \mathbb{E}[Y \mid w=0, \mathbf{S}=\mathbf{s}]$.

For a sub-population with the same features, the conditional ATE (CATE) (Some researchers use it to approximate ITE [9]) from observational data is as follows:

$$\begin{aligned} CATE(W,Y \mid \mathbf{X}=\mathbf{x}) = \mathbb{E}[Y \mid do(W=1), \mathbf{X}=\mathbf{x}] \\ - \mathbb{E}[Y \mid do(W=0), \mathbf{X}=\mathbf{x}] \end{aligned} \tag{4}$$

where $\mathbf{X}$ contains all factors causing the outcome Y.

When there exists a confounding M-bias variable, the set of measured variables $\mathbf{X}$ are not enough for the identification of $ATE(W,Y)$ and $CATE(W,Y \mid \mathbf{X}=\mathbf{x})$ from data as discussed in Introduction. We will introduce our DLRCE algorithm for solving this challenging problem in Section III.

III. The Proposed DLRCE Algorithm

A. Problem Setting

We assume that the underlying data generation or causal mechanism is represented as a causal DAG $\mathcal{G} = (\mathbf{X} \cup \mathbf{U} \cup \{W,Y\}, \mathbf{E})$ shown in Fig. 1 (d), where $\mathbf{U} = \mathbf{Z} \cup \mathbf{L} \cup \mathbf{F}$ are latent confounders, $\mathbf{X} = \mathbf{Q} \cup \mathbf{M}$ are measured variables, $\mathbf{Q}$ is the set of proxy variables for $\mathbf{Z}$, and $\mathbf{M}$ is the set of proxy variables for both $\mathbf{L}$ and $\mathbf{F}$.

Existing methods cannot be used to obtain an unbiased estimation of the causal effect of W on Y using measured variables since either adjusting or not adjusting for $\mathbf{M}$ results in a biased estimation. The aim of this paper is to unbiasedly estimate the ATE of W on Y, and the CATE of W on Y conditioning on $\mathbf{X}$ from observational data. More precisely, the research problem to be tackled in this paper is as follows.

Problem 1: Given an observational dataset $\mathcal{D}$ of a set measured variables $\{\mathbf{X} = \mathbf{Q} \cup \mathbf{M}, W, Y\}$, and assume that $\mathcal{D}$ is generated from the underlying DAG $\mathcal{G}$=$(\mathbf{X} \cup \mathbf{U} \cup \{W,Y\}, \mathbf{E})$ as shown in Fig. 1 (d). W and Y are the treatment and outcome variables respectively and $\mathbf{X}$ contains pretreatment variables. $\mathbf{Q}$ is a set of proxy confounders and $\mathbf{M}$ represents confounding M-bias variables. The goal is to estimate $ATE(W,Y)$ and $CATE(W,Y \mid \mathbf{X}=\mathbf{x})$ from the dataset $\mathcal{D}$.

B. Theoretical base of the Proposed DLRCE Algorithm

We will leverage the capability of VAE (variational autoencoder) [24], [25] in disentangled representation learning to tackle the confounding M-bias problem. VAEs are generative models, also known as latent variable models, and use a prior distribution and a noise distribution for generating the latent representations of the measured variables. We propose to make use of the VAE technique to learn and disentangle the latent representations of the latent variables in our problem setting. Specifically, we propose to use VAE to learn the representation of $\mathbf{Z}$ from the proxy variables $\mathbf{Q}$, and the latent representations $\mathbf{\Psi}$ from $\mathbf{M}$, and then disentangle $\mathbf{\Psi}$ into representations of $\mathbf{L}$ and $\mathbf{F}$, respectively. As in VAE literature, we use the same letter to denote a latent variable and its learned representation. The learned and disentangled latent representations $\{\mathbf{F}, \mathbf{Z}\}$ and $\mathbf{M}$ are used to obtain unbiased estimation of $CATE(W,Y \mid \mathbf{X}=\mathbf{x})$ and $ATE(W,Y)$ from observational data with latent confounders.

We first demonstrate that the latent representations learned and disentangled by the DLRCE algorithm are sound to estimate $ATE(W,Y)$ from the dataset $\mathcal{D}$.

Theorem 1: Given the setting in Problem 1, $ATE(W,Y)$ can be identified if the latent representations $\mathbf{Z}$ and $\mathbf{L}$ are recovered from the dataset $\mathcal{D}$, and we have $ATE(W,Y) = \mathbb{E}[Y \mid W = 1, \mathbf{Z} = \mathbf{z}, \mathbf{L} = \mathbf{l}] - \mathbb{E}[Y \mid W = 0, \mathbf{Z} = \mathbf{z}, \mathbf{L} = \mathbf{l}]$.

Proof 1: In Fig. 1 (d), $\mathbf{Z}$ and $\mathbf{L}$ are the parents of W. We will prove that $\mathbf{Z} \cup \mathbf{L}$ satisfies the back-door criterion (Definition 4) wrt., (W,Y), i.e., $\mathbf{Z} \cup \mathbf{L}$ blocks all the backdoor paths between W and Y. Firstly, $\mathbf{Z}$ and $\mathbf{L}$ do not contain any descendants of W, i.e., the first condition (i) of the back-door criterion holds. Secondly, there are three back-door paths between W and Y, i.e., $W \leftarrow \mathbf{Z} \rightarrow Y$, $W \leftarrow \mathbf{L} \rightarrow \mathbf{M} \rightarrow Y$, and $W \leftarrow \mathbf{L} \rightarrow \mathbf{M} \leftarrow \mathbf{F} \rightarrow Y$. The first back-door path is blocked by $\mathbf{Z}$ and the remaining two back-door paths are blocked by $\mathbf{L}$. Hence $\mathbf{Z}$ and $\mathbf{L}$ block all back-door paths between W and Y, i.e., the second condition (ii) of the back-door criterion holds. Therefore, $\mathbf{Z} \cup \mathbf{L}$ satisfies the back-door criterion and based on Eq. 3 $ATE(W,Y)$ can be identified in the dataset $\mathcal{D}$. Hence, $ATE(W,Y) = \mathbb{E}[Y \mid W = 1, \mathbf{Z} = \mathbf{z}, \mathbf{L} = \mathbf{l}] - \mathbb{E}[Y \mid W = 0, \mathbf{Z} = \mathbf{z}, \mathbf{L} = \mathbf{l}]$.

Theorem 1 presents a theoretical base for $ATE(W,Y)$ estimation. It is worth mentioning that the conditional clause 'if $\mathbf{Z}$ and $\mathbf{L}$ are recovered from the dataset $\mathcal{D}$' in the theorem is a fundamental assumption that is widely made in VAE-based causal inference [26]–[28].

In the following theorem, we will show that the latent representations learned and disentangled by the DLRCE are sound to estimate $CATE(W,Y \mid \mathbf{X} = \mathbf{x})$ from $\mathcal{D}$.

Theorem 2: Given the setting in Problem 1, $CATE(W,Y \mid \mathbf{X} = \mathbf{x})$ can be identified if the latent variables $\mathbf{Z}$ and $\mathbf{F}$ are recovered from the dataset $\mathcal{D}$, and we have $CATE(W,Y \mid \mathbf{X} = \mathbf{x}) = \mathbb{E}[Y \mid W = 1, \mathbf{Z} = \mathbf{z}, \mathbf{M} = \mathbf{m}, \mathbf{F} = \mathbf{f}] - \mathbb{E}[Y \mid W = 0, \mathbf{Z} = \mathbf{z}, \mathbf{M} = \mathbf{m}, \mathbf{F} = \mathbf{f}]$.

Proof 2: We first use the 'do' calculus rules [5] to remove the 'do' operator from the definition of $CATE$, i.e., $CATE(W,Y \mid \mathbf{X} = \mathbf{x}) = \mathbb{E}[Y \mid do(W), \mathbf{Z} = \mathbf{z}, \mathbf{F} = \mathbf{f}, \mathbf{L} = \mathbf{l}, \mathbf{X} = \mathbf{x}] = \mathbb{E}[Y \mid do(W), \mathbf{Z} = \mathbf{z}, \mathbf{F} = \mathbf{f}, \mathbf{L} = \mathbf{l}, \mathbf{M} = \mathbf{m}, \mathbf{Q} = \mathbf{q}]$. Let $\mathcal{G}_{\underline{W}}$ be the manipulated DAG by removing all outgoing edges of W from the causal DAG in Fig. 1 (d), and $\mathcal{G}_{\overline{W}}$ represents the manipulated DAG by eliminating all edges into W. Note that $Y \perp\!\!\!\perp \mathbf{L} \mid \mathbf{M}, \mathbf{F}$ and $Y \perp\!\!\!\perp \mathbf{Q} \mid \mathbf{Z}$ in $\mathcal{G}_{\overline{W}}$. Hence, using Rule 3 of do-calculus, we can remove $\mathbf{L}$ and $\mathbf{Q}$ from the conditioning set, and obtain $CATE(W,Y \mid \mathbf{X} = \mathbf{x}) = \mathbb{E}[Y \mid do(W), \mathbf{Z} = \mathbf{z}, \mathbf{F} = \mathbf{f}, \mathbf{M} = \mathbf{m}]$. Following Rule 2 of do-calculus [5] with the condition $Y \perp\!\!\!\perp W \mid \mathbf{Z}, \mathbf{M}, \mathbf{F}$ in $\mathcal{G}_{\underline{W}}$, we have $CATE(W,Y \mid \mathbf{X} = \mathbf{x}) = \mathbb{E}[Y \mid do(W), \mathbf{F} = \mathbf{f}, \mathbf{M} = \mathbf{m}, \mathbf{Z} = \mathbf{z}) = \mathbb{E}[Y \mid W, \mathbf{Z} = \mathbf{z}, \mathbf{M} = \mathbf{m}, \mathbf{F} = \mathbf{f}]$. Therefore, $\mathbf{Z}$, $\mathbf{F}$ and $\mathbf{M}$ are sufficient for identifying $CATE(W,Y \mid \mathbf{X} = \mathbf{x})$ from the dataset $\mathcal{D}$. Hence, $CATE(W,Y \mid \mathbf{X} = \mathbf{x}) = \mathbb{E}[Y \mid W, \mathbf{Z} = \mathbf{z}, \mathbf{M} = \mathbf{m}, \mathbf{F} = \mathbf{f}]$.

Theorems 1 and 2 provide the ground that learning the latent representations allows us to unbiasedly estimate $CATE$ and ATE from observational data when there exists a set of proxy variables, i.e., $\mathbf{X} = (\mathbf{M}, \mathbf{Q})$. In the next section, we will introduce our proposed DLRCE algorithm for learning these latent representations from observational data.

C. Disentanglement of Latent Representations

In this section, we present the details of our proposed DLRCE algorithm that is built on disentangled representation learning and supported by Theorems 1 and 2. The overall architecture of DLRCE is presented in Fig. 2. DLRCE aims to learn the latent representations $\mathbf{Z}$ from the proxy variables $\mathbf{Q}$, and the latent representations $\mathbf{\Psi}$ from the proxy variables $\mathbf{M}$, and then disentangle $\mathbf{\Psi}$ into two disjoint sets $\mathbf{L}$ and $\mathbf{F}$. Finally, the latent representations $\{\mathbf{F}, \mathbf{Z}\}$ and $\mathbf{M}$ are used for calculating causal effects of W on Y.

To learn and disentangle the representations, the inference model and the generative model are employed to approximate the two posteriors $p_{\varphi_{\mathbf{Q}}}(\mathbf{Q} \mid \mathbf{Z})$ and $p_{\varphi_{\mathbf{M}}}(\mathbf{M} \mid \mathbf{L}, \mathbf{F})$, where $\varphi_{\mathbf{Q}}$ and $\varphi_{\mathbf{M}}$ are the network parameters in the generative model. In the inference model of DLRCE, three separate encoders $q_{\theta_{\mathbf{Z}}}(\mathbf{Z} \mid \mathbf{Q})$, $q_{\theta_{\mathbf{L}}}(\mathbf{L} \mid \mathbf{M})$ and $q_{\theta_{\mathbf{F}}}(\mathbf{F} \mid \mathbf{M})$ are employed to serve as variational posteriors for deducing the three latent representations, for which $\theta_{\mathbf{Z}}$, $\theta_{\mathbf{L}}$ and $\theta_{\mathbf{F}}$ are the network parameters. In the generative model of the DLRCE algorithm, the latent representation $\mathbf{Z}$ is generated from a single encoder $q_{\theta_{\mathbf{Z}}}(\mathbf{Z} \mid \mathbf{Q})$ used by a single decoder $p_{\varphi_{\mathbf{Q}}}(\mathbf{Q} \mid \mathbf{Z})$ to reconstruct $\mathbf{Q}$; the latent representations $\mathbf{L}$ and $\mathbf{F}$ are generated from two separated encoders $q_{\theta_{\mathbf{L}}}(\mathbf{L} \mid \mathbf{M})$ and $q_{\theta_{\mathbf{F}}}(\mathbf{F} \mid \mathbf{M})$ used by a single decoder $p_{\varphi_{\mathbf{M}}}(\mathbf{M} \mid \mathbf{L}, \mathbf{F})$ to reconstruct $\mathbf{M}$.

As in the standard VAE [24], [25], we use Gaussian distributions to initialise the prior distributions of $P(\mathbf{Z})$, $P(\mathbf{L})$ and $P(\mathbf{F})$, and defined as:

$$\begin{aligned} &P(\mathbf{Z}) = \prod_{i=1}^{|\mathbf{Z}|} \mathcal{N}(Z_i \mid 0,1); P(\mathbf{L}) = \prod_{i=1}^{|\mathbf{L}|} \mathcal{N}(L_i \mid 0,1); \\ &P(\mathbf{F}) = \prod_{i=1}^{|\mathbf{F}|} \mathcal{N}(F_i \mid 0,1); \end{aligned} \tag{5}$$

In the inference model of DLRCE, the variational posteriors for approximating $q_{\theta_{\mathbf{Z}}}(\mathbf{Z} \mid \mathbf{Q})$, $q_{\theta_{\mathbf{L}}}(\mathbf{L} \mid \mathbf{M})$ and $q_{\theta_{\mathbf{F}}}(\mathbf{F} \mid \mathbf{M})$ are defined as:

$$\begin{aligned} q_{\theta_{\mathbf{Z}}}(\mathbf{Z} \mid \mathbf{Q}) &= \prod_{i=1}^{|\mathbf{Z}|} \mathcal{N}(\mu = \hat{\mu}_{Z_i}, \sigma^2 = \hat{\sigma}^2_{Z_i}); \\ q_{\theta_{\mathbf{L}}}(\mathbf{L} \mid \mathbf{M}) &= \prod_{i=1}^{|\mathbf{L}|} \mathcal{N}(\mu = \hat{\mu}_{L_i}, \sigma^2 = \hat{\sigma}^2_{L_i}); \\ q_{\theta_{\mathbf{F}}}(\mathbf{F} \mid \mathbf{M}) &= \prod_{i=1}^{|\mathbf{F}|} \mathcal{N}(\mu = \hat{\mu}_{F_i}, \sigma^2 = \hat{\sigma}^2_{F_i}); \end{aligned} \tag{6}$$

where $\hat{\mu}_{Z_i}, \hat{\mu}_{L_i}, \hat{\mu}_{F_i}$ and $\hat{\sigma}^2_{Z_i}, \hat{\sigma}^2_{L_i}, \hat{\sigma}^2_{F_i}$ are the estimated means and variances of latent variables Z_i, L_i and F_i, respectively.

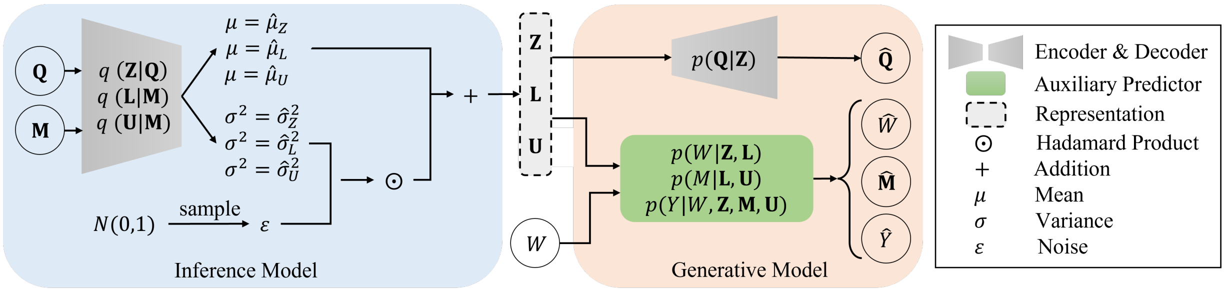


Fig. 2. The architecture of the proposed DLRCE algorithm consists of the inference network and the generative network for learning the three representations from proxy variables. Three auxiliary predictors ensure that the treatment W is predicted by $\mathbf{Z}$ and $\mathbf{L}$, the measured variables $\mathbf{M}$ are predicted by $\mathbf{L}$ and $\mathbf{F}$, and the outcome Y is predicted by $\mathbf{Z}$, $\mathbf{L}$ and $\mathbf{M}$.

In the generative models of DLRCE, $\mathbf{Z}$, $\mathbf{L}$ and $\mathbf{F}$ are used to generate W, $\mathbf{Q}$ and $\mathbf{M}$ as:

$$\begin{aligned} p_{\varphi_W}(W \mid \mathbf{Z}, \mathbf{L}) &= B(\sigma(g_1(\mathbf{Z}, \mathbf{L}))); \\ p_{\varphi_{\mathbf{Q}}}(\mathbf{Q} \mid \mathbf{Z}) &= \prod_{i=1}^{|\mathbf{Q}|} P(Q_i \mid \mathbf{Z}); \\ p_{\varphi_{\mathbf{M}}}(\mathbf{M} \mid \mathbf{L}, \mathbf{F}) &= \prod_{i=1}^{|\mathbf{M}|} P(M_i \mid \mathbf{L}, \mathbf{F}); \end{aligned} \tag{7}$$

where $B(\cdot)$ is the Bernoulli function, $g_1(\cdot)$ is a neural network, $\sigma(\cdot)$ is the logistic function. Notably, $P(Q_i \mid \mathbf{Z})$ and $P(M_i \mid \mathbf{L}, \mathbf{F})$ are the distributions on the i-th variable.

In the generative model of DLRCE, we generate the outcome Y based on its data type.

When Y is a continuous variable, we sample Y from a normal distribution and model it in the treatment and control groups as $P(Y \mid W = 0, \mathbf{Z}, \mathbf{M}, \mathbf{F})$ and $P(Y \mid W = 1, \mathbf{Z}, \mathbf{M}, \mathbf{F})$ respectively. Hence, the generative model for Y is described as:

$$\begin{aligned} p_{\varphi_Y}(Y \mid W, \mathbf{Z}, \mathbf{M}, \mathbf{F}) &= \mathcal{N}(\mu = \hat{\mu}_Y, \sigma^2 = \hat{\sigma}_Y^2); \\ \hat{\mu}_Y &= W \cdot g_2(\mathbf{Z}, \mathbf{M}, \mathbf{F}) + (1 - W) \cdot g_3(\mathbf{Z}, \mathbf{M}, \mathbf{F}); \\ \hat{\sigma}_Y^2 &= W \cdot g_4(\mathbf{Z}, \mathbf{M}, \mathbf{F}) + (1 - W) \cdot g_5(\mathbf{Z}, \mathbf{M}, \mathbf{F}); \end{aligned} \tag{8}$$

where $g_2(\cdot)$, $g_3(\cdot)$, $g_4(\cdot)$ and $g_5(\cdot)$ are the functions parameterised by neural networks.

When Y is a binary variable, we use a Bernoulli distribution to parameterise it:

$$p_{\varphi_Y}(Y \mid W, \mathbf{Z}, \mathbf{M}, \mathbf{F}) = B(\sigma(g_6(W, \mathbf{Z}, \mathbf{M}, \mathbf{F}))), \tag{9}$$

where $g_6(\cdot)$ is a neural network. We optimise these parameters by maximising the evidence lower bound (ELBO) [24]:

$$\begin{aligned} \mathcal{L}_{ELBO} = \; & \mathbb{E}_{q_{\theta_{\mathbf{Z}}}}[\log p_{\varphi_{\mathbf{Q}}}(\mathbf{Q} \mid \mathbf{Z})] + \\ & \mathbb{E}_{q_{\theta_{\mathbf{L}}} q_{\theta_{\mathbf{F}}}}[\log p_{\varphi_{\mathbf{M}}}(\mathbf{M} \mid \mathbf{L}, \mathbf{F})] \\ & - D_{KL}[q_{\theta_{\mathbf{Z}}}(\mathbf{Z} \mid \mathbf{Q}) || P(\mathbf{Z})] \\ & - D_{KL}[q_{\theta_{\mathbf{L}}}(\mathbf{L} \mid \mathbf{M}) || P(\mathbf{L})] \\ & - D_{KL}[q_{\theta_{\mathbf{F}}}(\mathbf{F} \mid \mathbf{M}) || P(\mathbf{F})], \end{aligned} \tag{10}$$

where $D_{KL}[\cdot || \cdot]$ is a KL divergence term.

To encourage the disentanglement of latent representations and ensure that $\mathbf{M}$ can be recovered by $\mathbf{L}$ and $\mathbf{F}$, and to ensure W can be predicted by $\mathbf{Z}$ and $\mathbf{L}$, and Y can be predicted by $\mathbf{Z}$, $\mathbf{M}$ and $\mathbf{F}$, three auxiliary predictors are added to the variational ELBO. Finally, the objective of DLRCE can be described as:

$$\begin{aligned} \mathcal{L}_{DLRCE} = & - \mathcal{L}_{ELBO} + \alpha \mathbb{E}_{q_{\theta_{\mathbf{L}}} q_{\theta_{\mathbf{F}}}}[\log q(\mathbf{M} \mid \mathbf{L}, \mathbf{F})] \\ & + \beta \mathbb{E}_{q_{\theta_{\mathbf{Z}}} q_{\theta_{\mathbf{L}}}}[\log q(W \mid \mathbf{Z}, \mathbf{L})] \\ & + \gamma \mathbb{E}_{q_{\theta_{\mathbf{Z}}} q_{\theta_{\mathbf{F}}}}[\log q(Y \mid W, \mathbf{Z}, \mathbf{M}, \mathbf{F})], \end{aligned} \tag{11}$$

where α, β and γ are the weights for balancing the three auxiliary predictors.

To estimate the CATEs of individuals conditioning on their measured variables $\mathbf{X}$, we employ the three encoders $q_{\theta_{\mathbf{Z}}}(\mathbf{Z} \mid \mathbf{Q})$, $q_{\theta_{\mathbf{L}}}(\mathbf{L} \mid \mathbf{M})$ and $q_{\theta_{\mathbf{F}}}(\mathbf{F} \mid \mathbf{M})$ to sample the approximated posteriors, and average the predicted potential outcomes using the classifier $q(Y \mid W, \mathbf{Z}, \mathbf{M}, \mathbf{F})$. Finally, by utilising Theorems 1 and 2, DLRCE is able to estimate the $ATE(W, Y)$ and $CATE(W, Y \mid \mathbf{X} = \mathbf{x})$ from the dataset $\mathcal{D}$.

IV. Experiments

In this section, we conduct experiments on both synthetic and real-world datasets to evaluate the performance of DLRCE for estimating ATE and $CATE$ from observational data with latent confounders. For the synthetic datasets, we use the causal DAG in Fig. 1 (d) to generate synthetic datasets with ground truths of ATE and $CATE$ for evaluating the performance of DLRCE. For the experiments on real-world datasets, we choose three benchmark datasets, Schoolingreturns [29], Cattaneo2 [30] and Sachs [31] where the empirical causal effects are available in the literature.

A. Experiment Setup

Baseline causal effect estimators. We compare our proposed DLRCE algorithm with nine state-of-the-art causal effect estimators that are widely used to estimate ATE and CATE from observational data. The seven estimators can be divided into two groups, Machine Learning based estimators and VAE based estimators. The Machine learning based estimators include (1) LinearDML (LDML) [32]: It is to solve the reverse causal metric bias by applying a cross-fitting strategy; (2) SparseLinearDML (SLDML) [33]: The loss function of the LinearDML estimator is modified by incorporating L_1 regularisation; (3) KernelDML [34]: It combines dimensionality

TABLE I

Estimation bias (mean±standard deviation) over 30 independently repeated experiments on the synthetic datasets with Y_{linear}. The best result is marked in boldface. Our proposed DLRCE algorithm obtains the smallest bias.

Method	Sample sizes				
	2k	4k	6k	8k	10k
LDML	30.82±0.31	28.56±0.19	28.36±0.17	28.65±0.12	28.08±0.05
SLDML	39.98±0.38	28.57±0.18	28.50±0.16	28.54±0.11	28.10±0.05
KernelDML	39.71±0.46	41.06±0.19	41.00±0.22	41.98±0.14	43.09±0.09
X-learner	23.99±0.51	22.24±0.17	22.15±0.14	21.86±0.11	21.92±0.10
R-learner	28.37±1.04	39.73±0.30	29.04±0.21	29.95±0.13	28.86±0.05
LDRlearner	48.57±0.51	47.48±0.17	46.50±0.18	46.73±0.12	47.22±0.09
CFDML	39.53±0.49	35.51±0.15	33.16±0.12	32.50±0.08	31.81±0.05
CEVAE	31.03±0.79	45.73±0.39	35.60±0.56	29.47±1.10	23.27±0.68
TEDVAE	40.59±0.40	34.59±0.18	31.82±0.14	31.27±0.11	29.75±0.07
DLRCE	**12.62±1.20**	**14.09±1.06**	**15.20±1.23**	**12.54±0.74**	**13.59±0.54**

TABLE II

Estimation PEHE (mean±standard deviation) over 30 independently repeated experiments on the synthetic datasets with Y_{linear}. The best result is marked in boldface. Our proposed DLRCE algorithm obtains the smallest PEHE.

Method	Sample sizes				
	2k	4k	6k	8k	10k
LDML	1.06±0.03	0.93±0.02	0.90±0.01	0.91±0.01	0.88±0.01
SLDML	1.07±0.03	0.93±0.02	0.91±0.01	0.90±0.01	0.89±0.01
KernelDML	1.25±0.05	1.22±0.02	1.24±0.02	1.27±0.01	1.48±0.01
X-learner	3.61±0.01	3.56±0.01	3.53±0.01	3.52±0.01	3.52±0.01
R-learner	7.69±2.21	5.00±0.49	4.13±0.37	3.91±0.25	3.39±0.14
LDRlearner	1.61±0.04	1.51±0.01	1.46±0.01	1.46±0.01	1.48±0.01
CFDML	1.34±0.03	1.19±0.01	1.15±0.01	1.12±0.01	1.10±0.01
CEVAE	1.17±0.06	1.36±0.05	1.18±0.06	0.94±0.06	0.82±0.07
TEDVAE	1.39±0.03	1.09±0.01	0.99±0.01	0.95±0.01	0.94±0.01
DLRCE	**0.46±0.12**	**0.50±0.12**	**0.58±0.13**	**0.47±0.07**	**0.56±0.10**

reduction techniques and kernel methods; (4) Mete-learners (including X-learner and R-learner) [35]; (5) LinearDRLearner (LDRlearner) [36]: It is based on double neural networks for addressing the bias in causal effect estimation; (6) CausalForestDML (CFDML) [37]: It employs two random forests for causal estimations for predicting two potential outcomes respectively. The VAE based estimators include: (1) causal effect variational autoencoder (CEVAE) [26] and (2) treatment effect by disentangled variational autoencoder (TEDVAE) [28].

Evaluation metrics. We employ the estimation bias $\Big|(A\hat{T}E - ATE)/ATE\Big| * 100$ (%) to evaluate the performance of all estimators, where ATE is the true causal effect and $A\hat{T}E$ is the estimated causal effect. We utilise the Precision of the Estimation of Heterogeneous Effect (PEHE) for the quality of CATE estimation [26], [38] defined as $\sqrt{\varepsilon_{PEHE}} = \sqrt{\mathbb{E}(((y_1 - y_0) - (\hat{y}_1 - \hat{y}_0))^2)}$, where y_1, y_0 represent the true potential outcomes and $\hat{y}_1, \hat{y}_0$ represent the predicted potential outcomes. Note that PEHE is widely employed for assessing CATE estimations in causal inference [6]. To mitigate random noise, we repeat the experiments multiple times and report the average and the standard deviation. For the three real-world datasets, since there is no ground truth causal effects available, we evaluate all estimators against the reference causal effects found in the literature.

Implementation details. We use *Python* and the libraries including *pytorch* [39], *pyro* [40] and *econml* to implement our proposed DLRCE algorithm. The implementation of DLRCE is available at the anonymous site https://anonymous.4open.science/r/DLRCE-385A. The implementations of LDML, SLDML, KernelDML, LDRLearner and CFDML are from the *Python* package *encoml* [41]. The implementations of X-learner and R-learner are from the *Python* package *CausalML* [42]. The implementation of CEVAE is based on the *Python* library *pyro* [40] and the implementations of TEDVAE is from the authors' GitHub.

B. Evaluations on Synthetic Datasets

We use the causal DAG in Fig. 1 (d) to generate the synthetic datasets with sample sizes, 2k, 4k, 6k, 8k, and 10k for our experiments. In the causal DAG $\mathcal{G}$, $\mathbf{M}$ and $\mathbf{X}$ are two set of proxy variables. L, F and $\mathbf{Z}$ are latent confounders. Similar to [10], [43], L, F and $\mathbf{Z}$ are generated from Bernoulli distribution. For an element $M \in \mathbf{M}$, it is generated from the two latent confounders L and F by using $M = \eta_1 * L + \eta_2 * F$, where η_1 and η_2 are two coefficients. For an element $X \in \mathbf{X}$, it is generated from the latent confounder Z by using $X \sim N(Z, \eta_3 * Z)$, where η_3 is a coefficient. For generating the treatment W, we use Bernoulli distribution with the conditional probability $P(W = 1 \mid L, Z, \mathbf{M}) = [1 + exp\{1 + 0.25 * L + 0.25 * Z\}]$.

In this work, we generate two types of potential outcomes $Y(W)$, namely a linear function Y_{linear} and a nonlinear function $Y_{nonlinr}$ as $Y(W) = 2 + 3 * W + 3 * \mathbf{M} + 2 * F * \mathbf{M} + 3 * Z + \epsilon_w$, where ϵ_w is an error term, and $Y(W) =$

https://github.com/WeijiaZhang/TEDVAE

TABLE III

Estimation bias (mean±standard deviation) over 30 independently repeated experiments on the synthetic datasets with Y_{nonlin}. The best result is marked in boldface. Our proposed DLRCE algorithm obtains the smallest bias.

Method	Sample sizes				
	2k	4k	6k	8k	10k
LDML	45.39±0.71	47.02±0.49	46.90±0.30	45.58±0.32	43.97±0.16
SLDML	46.79±0.76	47.14±0.42	47.12±0.30	45.79±0.33	44.03±0.16
KernelDML	54.30±0.94	61.93±0.74	63.45±0.50	61.72±0.39	62.34±0.26
X-learner	33.08±0.69	30.12±0.71	34.62±0.20	33.80±0.27	30.60±0.23
R-learner	26.13±0.43	23.73±0.32	25.64±0.21	25.60±0.12	23.44±0.13
LDRlearner	69.19±0.98	72.61±0.80	71.94±0.48	70.28±0.33	69.55±0.21
CFDML	59.94±0.84	57.61±0.41	53.87±0.37	51.52±0.20	48.86±0.16
CEVAE	24.15±3.08	61.91±2.30	46.21±3.71	47.07±4.44	41.37±5.48
TEDVAE	59.96±1.21	59.15±0.63	54.94±0.40	52.26±0.27	48.39±0.17
DLRCE	**15.52±0.89**	**16.58±7.70**	**19.32±3.37**	**10.57±0.59**	**10.32±0.65**

TABLE IV

Estimated PEHE (mean±standard deviation) over 30 independently repeated experiments on the synthetic datasets with Y_{nonlin} for different methods. The best result is marked in boldface. Our proposed DLRCE algorithm obtains the smallest PEHE.

Method	Samples				
	2k	4k	6k	8k	10k
LDML	1.65±0.07	1.52±0.04	1.53±0.02	1.46±0.02	1.39±0.01
SLDM	1.62±0.07	1.53±0.05	1.54±0.02	1.45±0.02	1.39±0.01
KernelDML	1.69±0.12	1.78±0.10	1.92±0.05	1.87±0.03	1.88±0.01
X-learner	6.23±0.04	6.22±0.03	6.17±0.01	6.18±0.01	6.10±0.01
R-learner	6.91±0.04	4.81±0.21	3.76±0.01	3.25±0.01	2.83±0.01
LDRlearner	2.37±0.09	2.27±0.09	2.31±0.03	2.24±0.03	2.20±0.02
CFDML	2.04±0.06	1.92±0.03	1.89±0.02	1.84±0.01	1.76±0.01
CEVAE	1.46±0.21	1.90±0.18	1.51±0.29	1.35±0.34	1.56±0.42
TEDVAE	2.07±0.09	1.82±0.09	1.75±0.03	1.63±0.02	1.54±0.02
DLRCE	**0.70±0.09**	**0.76±0.64**	**0.95±0.35**	**0.55±0.08**	**0.54±0.04**

$2+3*W+L*\mathbf{M}+\mathbf{M}+2*F+3*Z+\epsilon_w$, respectively. Based on the data generation process, all synthetic datasets have both potential outcomes, i.e., the true ITE for an individual is known. In our simulation study, the true ATE is 3. To evaluate the performance of our DLRCE algorithm, we conduct the experiments 30 times independently for each setting.

We report the estimation bias and PEHE for the synthetic datasets generated from Y_{linear} in Tables I and II, and for the synthetic datasets generated from Y_{nonlin} in Tables III and IV. **Results.** From the experimental results, we have the following observations: (1) Machine learning based estimators, LDML, SLDML, KernelDML, X-learner, R-learner, LDRlearner and CFDML have a large estimation bias and PEHE on both types of synthetic datasets since these estimators rely on the assumption of unconfoundedness and cannot learn a valid representation from proxy variables to block all back-door paths between W and Y. (2) VAE based estimators, TEDVAE and CEVAE methods have a large estimation bias and PEHE on both types of synthetic datasets since both methods fail to deal with the confounding M-bias variable studied in this work. (3) The proposed DLRCE algorithm obtains the smallest estimation bias and PEHE among all methods on both types of synthetic datasets since our DLRCE algorithm learns and disentangles three latent representations $\mathbf{Z}$, $\mathbf{L}$ and $\mathbf{F}$ from proxy variables $(\mathbf{X}, \mathbf{M})$ to effectively block all back-door paths between W and Y. The smallest estimation bias and PEHE further confirm the correctness of our DLRCE algorithm in learning three latent representations $\mathbf{Z}$, $\mathbf{L}$ and $\mathbf{F}$ from proxy variables. (4) The compared algorithms, machine learning based and VAE based estimators achieve better performance compared to the synthetic datasets generated from Y_{linear} and relatively poorer performance on the synthetic datasets generated from Y_{nonlin}. Our proposed DLRCE algorithm consistently produces good performance across both types of datasets.

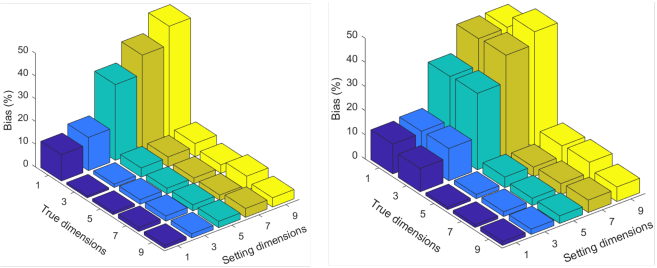
In sum, the simulation studies demonstrate that the proposed DLRCE algorithm effectively addresses the problem of confounding M-bias when estimating ATE and CATE from observational data in the presence of latent confounders. It further provides evidence that DLRCE is capable of recovering latent variable representations from proxy variables.

C. *Parameters Analysis*

In our DLRCE algorithm, there are three tuning parameters, namely α, β, and γ, used to balance $\mathcal{L}_{ELBO}$ and the three classifiers during the training process. We consider setting $\{\alpha, \beta, \gamma\}$ = $\{0.1, 0.5, 1, 1.5, 2\}$ to analyse the sensitivity of the three parameters on synthetic datasets with a sample size of 10k, generated using the same data generation process described in Section IV-B. We report the estimation bias of DLRCE algorithm in Table V. From Table V, we observe that the three parameters α, β, γ have a low sensitivity to the estimation bias of the DLRCE algorithm in ATE estimation. In summary, it is recommended to set the three tuning parameters, α, β, and γ, to small values for our DLRCE algorithm.

TABLE V
THE ESTIMATION BIAS WITH THE DIFFERENT SETTING OF TUNNING PARAMETERS α, β AND γ.

Weight	Dataset	
	Linear	Nonlinear
$\{\alpha, \beta, \gamma\}$ = 0.1	14.23±0.54	9.31±0.57
$\{\alpha, \beta, \gamma\}$ = 0.5	14.46±1.26	12.92±1.33
$\{\alpha, \beta, \gamma\}$ = 1	13.59±0.54	10.32±0.65
$\{\alpha, \beta, \gamma\}$ = 1.5	11.60±0.88	18.18±2.12
$\{\alpha, \beta, \gamma\}$ = 2	15.15±0.59	11.03±1.00



(a) Synthetic datasets with Y_{linear} (b) Synthetic datasets with Y_{nonlin}

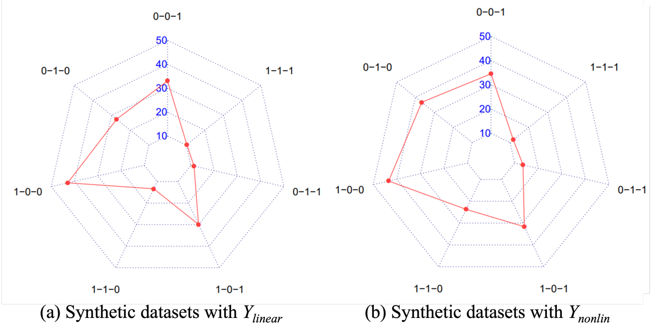
Fig. 3. Estimation bias on both types of synthetic datasets. 'True dimensions' refer to the dimensions of $\mathbf{L}$, $\mathbf{F}$, and $\mathbf{Z}$ in the data, and 'Setting dimensions' correspond to the parameters of $|\mathbf{L}|$, $|\mathbf{F}|$, and $|\mathbf{Z}|$ in the DLRCE algorithm.

D. A Study on the Dimensionality of Latent Representations

In our simulation studies, we set the dimensions of $\mathbf{L}$, $\mathbf{F}$, and $\mathbf{Z}$ to 1, respectively. We conducted a study on the dimensionality of latent representations to demonstrate the effectiveness of this setting. To achieve this goal, we fixed the sample size to 10k for all synthetic datasets and repeated the experiments 30 times independently to minimise random noise for each setting. Following the data generation process described in Section IV-B, we generated a set of synthetic datasets with dimensions of the three latent variables $(\mathbf{L}, \mathbf{F}, \mathbf{Z})$ set to $\{1, 3, 5, 7, 9\}$ respectively. In our DLRCE algorithm, we set three parameters $(|\mathbf{L}|, |\mathbf{F}|, |\mathbf{Z}|)$ to $\{1, 3, 5, 7, 9\}$ respectively to conduct experiments on these synthetic datasets. The estimation bias of the DLRCE algorithm on these datasets is displayed in Fig. 3. From Fig. 3, we observe that the estimation bias of the DLRCE algorithm is the smallest on both types of synthetic datasets when $(|\mathbf{L}|, |\mathbf{F}|, |\mathbf{Z}|)$ is set to (1, 1, 1) regardless of the true dimensions of $\mathbf{L}$, $\mathbf{F}$, and $\mathbf{Z}$ in the data. Hence, this finding suggests that setting $|\mathbf{L}|$, $|\mathbf{F}|$, and $|\mathbf{Z}|$ to 1 is reasonable.

E. Ablation Study

Next, we examine the impact of three latent representations $\mathbf{L}$, $\mathbf{F}$, and $\mathbf{Z}$ on the performance of DLRCE. To do this, we set the dimensions of $(\mathbf{L}, \mathbf{F}, \mathbf{Z})$ to (1,0,0), (0,1,0), (0,0,1), (1,1,0), (1,0,1), (0,1,1), and (1,1,1), respectively. We conduct a series of experiments on both types of synthetic datasets with a sample size of 10k, generated using the same data generation process described in Section IV-B. Figure 4 illustrates the capability of each latent representation in terms of estimation bias using a radar chart. For example, in Figure 4 (a), the DLRCE performances achieve the smallest estimation bias when the dimensions of $(\mathbf{L}, \mathbf{F}, \mathbf{Z})$ are set to (1,1,1). It is worth noting that $(\mathbf{L}, \mathbf{F}, \mathbf{Z}) = (0, 1, 1)$ yields the second smallest estimation bias, consistent with the conclusion in Theorem 2. Moreover, all three latent representations contribute to bias reduction, with $\{\mathbf{Z}, \mathbf{F}\}$ contributing the most.



(a) Synthetic datasets with Y_{linear} (b) Synthetic datasets with Y_{nonlin}

Fig. 4. Radar charts for DLRCE's capability in learning and disentangling the latent representations. Each vertex on the polygons denotes the latent representations' dimensions. For example, 0-1-1 implies that $(\mathbf{L}, \mathbf{F}, \mathbf{Z})$=(0, 1, 1), i.e., $|\mathbf{L}|$ = 0.

F. Experiments on Three Real-World Datasets

In this section, we assess the performance of DLRCE against the above-mentioned comparisons on three real-world datasets, Schoolingreturns [29], Cattaneo2 [30] and Sachs [31] for which the empirical causal effects are available in the literature. The details of the three datasets are described below.

Schoolingreturns Dataset. This dataset consists of 3,010 records and 19 variables [29]. The treatment variable is the education level of a person. The outcome variable is raw wages in 1976 (in cents per hour). The goal of collecting this dataset is to study the causal effect of the education level on wages. The estimated $ATE(W, Y)$ = 0.1329 with 95% confidence interval (0.0484, 0.2175) from the works [44] as the reference causal effect.

Cattaneo2 Dataset. The Cattaneo 2 [30] is widely employed to investigate the ATE of maternal smoking status during pregnancy (W) on a baby's birth weight (in grams). Cattaneo2 consists of the birth weights of 4,642 singleton births in Pennsylvania, USA [30], [45]. Cattaneo2 contains 864 smoking mothers (W=1) and 3,778 nonsmoking mothers (W=0). The dataset contains several covariates: mother's age, mother's marital status, an indicator for the previous infant where the newborn died, mother's race, mother's education, father's education, number of prenatal care visits, months since last birth, an indicator of firstborn infant and an indicator of alcohol consumption during pregnancy. The authors [45] found a strong negative effect of maternal smoking on the weights of babies, namely about $200g$ to $250g$ lighter for a baby with a mother smoking during pregnancy by statistical analysis on all covariates.

Sachs Dataset. The dataset contains 853 samples and 11 variables [31]. The treatment is Erk (the manipulation of

It can be downloaded from the site: http://www.stata-press.com/data/r13/cattaneo2.dta

TABLE VI
THE ESTIMATED $\hat{ATE}$ ON THE THREE REAL-WORLD DATASETS. NOTE THAT THE ESTIMATED $\hat{ATE}$ BY OUR DLRCE FALLS IN THE EMPIRICAL CAUSAL EFFECT INTERVAL ON ALL THREE REAL-WORLD DATASETS. THE ESTIMATED $\hat{ATE}$ IS WITHIN THE EMPIRICAL INTERVAL MARKED IN BOLDFACE.

Method	Datasets		
	Schoolingreturns	Cattaneo2	Sachs
LDML	-0.045	-170.179	36.118
SLDML	-0.504	-153.859	152.900
KDML	-0.021	-146.824	19.360
X-Learner	**0.161**	**-230.61**	18.661
R-Learner	-0.020	**-234.96**	24.072
LDRLearner	-0.020	-179.853	37.400
CFDML	-0.040	**-241.436**	25.774
CEVAE	0.026	**-221.234**	**0.254**
TEDVAE	0.231	**-235.325**	**0.255**
DLRCE	**0.101**	**-226.448**	**1.278**

concentration levels of a molecule). The outcome is the concentration of Akt. In this work, we take the reported $ATE(W, Y)$ = 1.4301 with 95% confidence interval (0.05, 3.23) in the work [46] as the reference causal effect.

Results. We report the results on the three real-world datasets in Table VI. From Table VI, we can see that (1) the estimated $\hat{ATE}$s by DLRCE on three real-world datasets are within the empirical intervals respectively. (2) The estimated $\hat{ATE}$ by X-learner on Schoolingreturns, by X-learner, R-learner, CFDML, CEVAE and TEDVAE on Cattaneo2, and by CEVAE and TEDVAE on Sachs are within the empirical intervals, but these methods do not produce estimates within the confidence intervals for all three data sets. The other methods fail to obtain an estimation within the empirical intervals on any of the three datasets. (3) The estimates of LDML, SLDML, KDML, R-learner, LDRLearner and CFDML on Schoolingreturns are negative which is opposite to a positive estimate in the literature [44]. (4) The estimated $\hat{ATE}$s on Sachs by Machine learning based estimators, such as LDML, SLDML, KDML, X-learner, R-learner, LDRLearner and CFDML, are far away from the empirical interval (0.05, 3.23).

In a word, the proposed DLRCE algorithm performs better than the stat-of-the-art causal effect estimators on the three real-world datasets. This further confirms the potential applicability of DLRCE in real-world applications.

V. RELATED WORK

Machine learning for causal effect estimation. Causal effect estimations from observational data have received extensive attention from the artificial intelligence and statistics communities [1], [5], [6], [8], [47]. For instance, matching methods [48], [49] and tree-based methods [37], [50]–[52] have been developed to address confounding bias in causal effect estimation from observational data. Additionally, meta-learners [35] have also been studied for estimating the average treatment effect (ATE) and conditional average treatment effect (CATE) from observational data.

Representation learning for causal effect estimation. Recently, representation learning methods [6], [9], [53] have been applied to causal effect estimation, but they often rely on the unconfoundedness assumption [1]. For example, Shalit et al. proposed a balanced representation learning method for counterfactual regression (CFRNet) [9]. Yoon et al. first used a GAN model to learn representations for causal effect estimation. Different from these methods, our proposed DLRCE algorithm addresses the challenging problem of confounding M-bias variable in causal effect estimation.

Proxy variables for causal effect estimation. Proxy variables are the measured covariates that are at best of the true underlying confounding mechanism [26], [54], [55]. Kallus et al. [54] proposed to infer the confounders from proxy variables by using matrix factorisation. Miao et al. [55] proposed the general conditions for causal effects identification using more general proxies, but they did not propose a practical data-driven method. CEVAE [26] uses the VAE model to learn the representations from proxy variables for causal effect estimation. However, CEVAE fails to deal with the confounding M-bias problem in data studied in this work as shown in our experiments. To the best of our knowledge, our DLRCE algorithm is the first work to solve the problem of confounding M-bias variable using the disentanglement of representation learning techniques.

VI. CONCLUSION

In this paper, we identify a challenging problem in estimating causal effects from observational data in the presence of latent confounders, i.e., the problem of confounding M-bias as shown in the causal DAG in Fig. 1 (c). Existing methods tackle confounding bias through balanced representation learning or covariate adjustment, but are unable to handle the problem of confounding M-bias, and lead to biased causal effect estimation as shown in our experiments. To address this problem, we propose a novel disentangled representation learning framework, the DLRCE algorithm for causal effect estimation from observational data in the presence of latent confounders. DLRCE learns three sets of latent representations from proxy variables to adjust for both confounding bias and M-bias. Extensive experiments on synthetic and three real-world datasets demonstrate that DLRCE outperforms existing causal effect estimation methods for ATE and CATE estimation in datasets with both types of biases. The proposed method shows promise in causal effect estimation in real-world datasets and opens up avenues for addressing complex confounding scenarios in causal inference.

VII. ACKNOLEDGEMENT

This work was supported by the Chinese Central Universities (grant number: 2662023XXPY004) and the Australian Research Council (grant number: DP230101122).

REFERENCES

[1] G. W. Imbens and D. B. Rubin, *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

[2] D. B. Rubin, "Causal inference using potential outcomes: Design, modeling, decisions," *J Am Stat Assoc*, vol. 100, no. 469, pp. 322–331, 2005.

[3] S. Greenland, “Quantifying biases in causal models: classical confounding vs collider-stratification bias,” *Epidemiology*, vol. 14, no. 3, pp. 300–306, 2003.

[4] A. F. Connors, T. Speroff *et al.*, “The effectiveness of right heart catheterization in the initial care of critically iii patients,” *Journal of the American Medical Association*, vol. 276, no. 11, pp. 889–897, 1996.

[5] J. Pearl, *Causality*. Cambridge university press, 2009.

[6] R. Guo, L. Cheng *et al.*, “A survey of learning causality with data: Problems and methods,” *ACM Computing Surveys (CSUR)*, vol. 53, no. 4, pp. 1–37, 2020.

[7] A. Deaton and N. Cartwright, “Understanding and misunderstanding randomized controlled trials,” *Social Science & Medicine*, vol. 210, pp. 2–21, 2018.

[8] D. Cheng, J. Li *et al.*, “Data-driven causal effect estimation based on graphical causal modelling: A survey,” vol. abs/2208.09590, 2022.

[9] U. Shalit, F. D. Johansson, and D. A. Sontag, “Estimating individual treatment effect: generalization bounds and algorithms,” in *Proceedings of the 34th International Conference on Machine Learning, ICML*, 2017, pp. 3076–3085.

[10] N. Hassanpour and R. Greiner, “Counterfactual regression with importance sampling weights,” in *IJCAI*, 2019, pp. 5880–5887.

[11] S. Athey, G. W. Imbens, and S. Wager, “Approximate residual balancing: debiased inference of average treatment effects in high dimensions,” *J R Stat Soc Series B (Stat Methodol)*, vol. 80, no. 4, pp. 597–623, 2018.

[12] D. B. Rubin, “Estimating causal effects of treatments in randomized and nonrandomized studies.” *J. Educ. Psychol*, vol. 66, no. 5, p. 688, 1974.

[13] E. Perković, J. Textor *et al.*, “Complete graphical characterization and construction of adjustment sets in Markov equivalence classes of ancestral graphs,” *J. Mach. Learn. Res*, vol. 18, no. 1, pp. 8132–8193, 2018.

[14] B. van der Zander, M. Liśkiewicz, and J. Textor, “Separators and adjustment sets in causal graphs: Complete criteria and an algorithmic framework,” *Artificial Intelligence*, vol. 270, pp. 1–40, 2019.

[15] J. Pearl, “Myth, confusion, and science in causal analysis,” *Tech. Rep. R-348*, 2009, los Angeles, CA: University of California.

[16] P. Ding and L. W. Miratrix, “To adjust or not to adjust? sensitivity analysis of m-bias and butterfly-bias,” *Journal of Causal Inference*, vol. 3, no. 1, pp. 41–57, 2015.

[17] W. H. Greene, *Econometric analysis*. Pearson Education India, 2003.

[18] D. Entner, P. Hoyer, and P. Spirtes, “Data-driven covariate selection for nonparametric estimation of causal effects,” in *AISTATS*, 2013, pp. 256–264.

[19] D. Cheng, J. Li *et al.*, “Local search for efficient causal effect estimation,” *IEEE Transactions on Knowledge & Data Engineering*, no. 01, pp. 1–14, 2022.

[20] P. Spirtes, C. N. Glymour *et al.*, *Causation, prediction, and search*. MIT press, 2000.

[21] P. R. Rosenbaum and D. B. Rubin, “The central role of the propensity score in observational studies for causal effects,” *Biometrika*, vol. 70, no. 1, pp. 41–55, 1983.

[22] D. B. Rubin, “Using multivariate matched sampling and regression adjustment to control bias in observational studies,” *Journal of the American Statistical Association*, vol. 74, no. 366a, pp. 318–328, 1979.

[23] X. De Luna, I. Waernbaum, and T. S. Richardson, “Covariate selection for the nonparametric estimation of an average treatment effect,” *Biometrika*, vol. 98, no. 4, pp. 861–875, 2011.

[24] D. P. Kingma and M. Welling, “Auto-encoding variational bayes,” in *2nd International Conference on Learning Representations, ICLR*, 2014.

[25] D. P. Kingma, M. Welling *et al.*, “An introduction to variational autoencoders,” *Foundations and Trends® in Machine Learning*, vol. 12, no. 4, pp. 307–392, 2019.

[26] C. Louizos, U. Shalit, J. M. Mooij, D. Sontag, R. Zemel, and M. Welling, “Causal effect inference with deep latent-variable models,” in *Advances in Neural Information Processing Systems*, 2017, pp. 6446–6456.

[27] N. Hassanpour and R. Greiner, “Learning disentangled representations for counterfactual regression,” in *International Conference on Learning Representations*, 2019, pp. 1–11.

[28] W. Zhang, L. Liu, and J. Li, “Treatment effect estimation with disentangled latent factors,” in *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI*, 2021, pp. 10 923–10 930.

[29] D. Card, “Using geographic variation in college proximity to estimate the return to schooling,” National Bureau of Economic Research, Inc, NBER Working Papers 4483, 1993.

[30] M. D. Cattaneo, “Efficient semiparametric estimation of multi-valued treatment effects under ignorability,” *Journal of Econometrics*, vol. 155, no. 2, pp. 138–154, 2010.

[31] K. Sachs, O. Perez *et al.*, “Causal protein-signaling networks derived from multiparameter single-cell data,” *Science*, vol. 308, no. 5721, pp. 523–529, 2005.

[32] V. Chernozhukov, D. Chetverikov *et al.*, “Double/debiased machine learning for treatment and structural parameters,” *The Econometrics Journal*, vol. 21, no. 1, pp. C1–C68, 2018.

[33] V. Chernozhukov, M. Goldman *et al.*, “Orthogonal machine learning for demand estimation: High dimensional causal inference in dynamic panels,” *arXiv*, pp. arXiv–1712, 2017.

[34] X. Nie and S. Wager, “Quasi-oracle estimation of heterogeneous treatment effects,” *Biometrika*, vol. 108, no. 2, pp. 299–319, 2021.

[35] S. R. Künzel, J. S. Sekhon *et al.*, “Metalearners for estimating heterogeneous treatment effects using machine learning,” *PNAS*, vol. 116, no. 10, pp. 4156–4165, 2019.

[36] D. J. Foster and V. Syrgkanis, “Orthogonal statistical learning,” *arXiv preprint arXiv:1901.09036*, 2019.

[37] S. Athey, J. Tibshirani, and S. Wager, “Generalized random forests,” *The Annals of Statistics*, vol. 47, no. 2, pp. 1148–1178, 2019.

[38] J. L. Hill, “Bayesian nonparametric modeling for causal inference,” *J Comput Graph Stat*, vol. 20, no. 1, pp. 217–240, 2011.

[39] A. Paszke, S. Gross *et al.*, “Pytorch: An imperative style, high-performance deep learning library,” in *Advances in Neural Information Processing Systems 32, NeurIPS*, 2019, pp. 8024–8035.

[40] E. Bingham, J. P. Chen *et al.*, “Pyro: Deep universal probabilistic programming,” *The Journal of Machine Learning Research*, vol. 20, no. 1, pp. 973–978, 2019.

[41] K. Battocchi, E. Dillon *et al.*, “EconML: A Python Package for ML-Based Heterogeneous Treatment Effects Estimation,” https://github.com/microsoft/EconML, 2019.

[42] H. Chen, T. Harinen *et al.*, “Causalml: Python package for causal machine learning,” *arXiv preprint arXiv:2002.11631*, 2020.

[43] D. Cheng, J. Li *et al.*, “Toward unique and unbiased causal effect estimation from data with hidden variables,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, pp. 6108 – 6120, 2022.

[44] M. Verbeek, *A Guide to Modern Econometrics*. John Wiley & Sons, 2008.

[45] D. Almond, K. Y. Chay, and D. S. Lee, “The costs of low birth weight,” *The Quarterly Journal of Economics*, vol. 120, no. 3, pp. 1031–1083, 2005.

[46] D. Cheng, J. Li *et al.*, “Discovering ancestral instrumental variables for causal inference from observational data,” *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–11, 2023.

[47] M. A. Hernán and J. M. Robins, *Causal Inference*. CRC Boca Raton, FL;, 2010.

[48] D. B. Rubin and N. Thomas, “Matching using estimated propensity scores: relating theory to practice,” *Biometrics*, pp. 249–264, 1996.

[49] A. Abadie and G. W. Imbens, “Large sample properties of matching estimators for average treatment effects,” *econometrica*, vol. 74, no. 1, pp. 235–267, 2006.

[50] H. A. Chipman, E. I. George, R. E. McCulloch *et al.*, “Bart: Bayesian additive regression trees,” *The Annals of Applied Statistics*, vol. 4, no. 1, pp. 266–298, 2010.

[51] S. Athey and G. Imbens, “Recursive partitioning for heterogeneous causal effects,” *PNAS*, vol. 113, no. 27, pp. 7353–7360, 2016.

[52] S. Wager and S. Athey, “Estimation and inference of heterogeneous treatment effects using random forests,” *J Am Stat Assoc*, vol. 113, no. 523, pp. 1228–1242, 2018.

[53] J. Yoon, J. Jordon, and M. van der Schaar, “GANITE: estimation of individualized treatment effects using generative adversarial nets,” in *6th International Conference on Learning Representations, ICLR*, 2018.

[54] N. Kallus, X. Mao, and M. Udell, “Causal inference with noisy and missing covariates via matrix factorization,” in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 2018, pp. 6921–6932.

[55] W. Miao, Z. Geng, and E. J. Tchetgen Tchetgen, “Identifying causal effects with proxy variables of an unmeasured confounder,” *Biometrika*, vol. 105, no. 4, pp. 987–993, 2018.