LEARNING CAUSAL EFFECT FOR HIGH-DIMENSIONAL OBSERVATION DATA WITH UNMEASURED CONFOUNDING

Hedong YAN
Hong Kong Baptist University

ABSTRACT

Learning causal effects from high-dimensional observation data is critical for many realistic applications. The main challenge of causal effect estimation is the confounding problem, which includes both measured and unmeasured confounding, especially in high-dimensional data. In this survey, we have explored various frameworks, objectives, metrics, approaches, datasets, and packages for causal effect learning from observation data. We have also shared our preliminary work and future plans for learning causal effects from high-dimensional observation data. By addressing the confounding problem in causal effect estimation, we can develop better models and algorithms to uncover causal relationships and make more accurate predictions and decisions in real-world scenarios.

1 BACKGROUND

1.1 INTRODUCTION

The estimation of causal effects from observational data is a significant goal in numerous practical applications, including clinical data analysis, medical report assistance, uplift marketing, and AI robustness. This endeavor is aimed at addressing critical questions such as "Would the patient's oxygen saturation have normalized without treatment?" "What should a physician include in the report if the patient had no lung opacity?" "How can one estimate the lift in a user's purchase intention due to an advertisement?" and "How will an AI perform in the face of environmental changes?" Answering these questions is non-trivial, even in low-dimensional cases, due to the confounding problem. This problem arises from that the same observation may be generated by multiple generative models, leading to different estimations of causal effects. The confounding may stem from the fact that treatment assignment is not randomized in observation data, as is the case in random-

ized control experiments (RCTs) Holland (1986). Alternatively, the confounding may arise from the existence of common causes between treatment and outcome, as described by Reichenbach's common cause principle Hofer-Szabó et al. (2013). Therefore, identifying the most effective model for estimating causal effects is crucial for mitigating the confounding problem. Failing to do so may lead to spurious correlations induced by confounders that do not contribute to the causal effect.

In recent years, significant progress has been made in learning causal effect from high dimensional observation data, due in part to the advancement of artificial neural networks and the use of GPU technology. For instance, the Causal Effect Variational Autoencoder (CEVAE) Louizos et al. (2017) leverages a graphical prior to learn a latent confounder from a variational lower bound. It has the ability to learn a latent confounder from high dimensional covariates. Deconfounder Wang & Blei (2019) and Time Series Deconfounder Bica et al. (2020) use a (time-series) factor model to predict individual treatment assignment from a confounder and covariate, and subsequently estimate treatment effect using an outcome model. These approaches are useful for dealing with multiple treatments. However, the current approaches have limitations.

Recent works in causal effect learning can be broadly categorized into two types based on the stage of the causal effect estimation process. The first type, imputation methods, focus on imputing missing values in the data to estimate the causal effect. The second type, dependency methods, rely on modeling the dependency relationships between the treatment, covariates, and outcome to estimate the causal effect.

Imputation approaches consist of four stages, namely counterfactual imputation, balancing representation, learning potential outcome, and learning causal effect. The counterfactual imputation stage is the most important step in these methods as it provides the source of counterfactual knowledge, even though it is often implicit. Balancing representation involves ensuring that units in the treatment group and control group are sampled from the same distribution to avoid confounding induced by observed variables. The potential outcome model is used to predict potential outcomes, and the treatment effect model is used to predict the causal effect. Counterfactual data is imputed by matching before learning.

For example, the Balancing Neural Network (BNN) Johansson et al. (2016) is a two-step model that imputes counterfactual data using the nearest neighbor that was assigned the opposite treatment in the covariate space. The model first learns a balanced representation with a loss function that includes factual error, counterfactual error, and a discrepancy distance. The second step is to learn a linear function that predicts factual outcome and counterfactual outcome from the representation and treatment. On the other hand, the Treatment Agnostic Representation Network (TARNet) Shalit et al.

(2017) is an end-to-end model that does not need nearest neighbor matching in the covariate space. It uses twin neural networks with different parameters to model the treated and control outcomes, respectively. The Counterfactual Regression (CFR) adds an integral probability metric (IPM), such as Wasserstein and Maximum Mean Discrepancy (MMD) distances, for balancing representation between the treated and control groups. Unlike T-learners, the X-learner Künzel et al. (2019) does not learn a balanced representation. Instead, it assumes perfect imputing and learns two Bayesian additive regression trees (BART) as potential outcome models. It computes imputed treatment effect for two linear causal effect models and weighs them using covariates.

However, one limitation of these approaches is that they assume that all confounders have been observed, which may not be true in reality.

The second class of causal inference methods is based on dependencies among variables. These methods typically involve four stages: assuming a causal diagram, abduction, action, and prediction. Counterfactual knowledge in these methods comes from the causal diagram, which specifies the relationships among observed and unobserved variables.

For instance, CEVAE Louizos et al. (2017) is an end-to-end model that uses variational methods. It begins by assuming a causal diagram for the data-generating process. It then uses all observation data to approximate a posterior distribution of the latent variables. Next, it applies do-calculus on the causal diagram to factorize the conditional average treatment effect query into a distribution of all variables, including the latent variables, to avoid confounding problems. Finally, it uses a neural network to learn the factorized distributions to predict potential outcomes. Deconfounder Wang & Blei (2019) is designed to handle multiple treatments. It is a two-stage model that assumes no single common cause between treatments and outcomes. It then uses treatment data to approximate a posterior distribution of proxy variables that ensure that every treatment is independent of the others when conditioned on the proxy variable. If such proxy variables can be found, then hidden confounders can be proven not to exist. Finally, the model uses the proxy variable and treatment to predict the potential outcomes. Time-series deconfounder Bica et al. (2020) is a time-series extension of deconfounder.

However, a limitation of these methods is that the causal diagram (i.e., the dependencies among variables) can be misspecified without careful censorship by domain experts. Additionally, learning the hidden variable from data can be time-consuming, making it difficult to learn a causal diagram for high-dimensional data.

Our interested research problem is how to learn causal effect effectively and efficiently from high dimensional observation data with hidden confounding. We are interested in the acceleration approaches for dependency methods because it takes hidden confounding into consideration. We are facing those challenges.

A key challenge in causal effect learning is dealing with unmeasured confounding, which arises when individuals do not share the same causal diagram and parameters. To address this problem, we plan to classify diagrams into do-equivalent classes and learn parameters separately. We will then develop an individual measurement function to improve the prediction of potential outcomes. The individual fusion function will integrate the learning models for potential outcome prediction. We propose two schemes to approach the function learning problem: the first involves treating function learning as a classification problem, while the second treats it as a module of previous models, allowing us to build an end-to-end model. Notably, we will be the first to learn an individual SCM for causal effect estimation.

The second major challenge we aim to tackle is high dimensional data. Many existing approaches rely on a sparse graph structure assumption, which may not hold for high dimensional data due to the exponentially large number of possible relationships. To address this, we propose a clustering method to group variables into limited clusters with specific constraints. Next, we use identification algorithms to calculate all possible identification results among the clusters. We then learn several models based on these identification results and finally, a fusion model to predict potential outcomes and causal effects. We also plan to incorporate counterfactual loss methods to train the model without directly inferring counterfactual data. The end-to-end model can combine these approaches for efficient causal effect estimation on high dimensional data. Our proposed approach is novel in that it utilizes a combination of clustering, identification algorithms, and counterfactual loss to address the challenges of high dimensional data in causal effect learning.

The third challenge in causal effect estimation is the ability to generalize beyond the observed data distribution. Often, machine learning models are trained on a specific dataset, assuming that the test data will be drawn from the same distribution as the training data. However, in the context of causal effect estimation, the observed data distribution may not represent the entire population or may be influenced by unobserved confounding variables and parameters changes. As a result, models trained on such data may fail to generalize to out-of-distribution data, leading to biased causal effect estimates. Addressing this challenge requires developing methods to identify and handle distribution shifts and unobserved confounding variables, as well as evaluating the model's performance on unseen data to ensure reliable generalization.

1.2 APPLICATIONS

High-dimensional causal effect estimation from observational data has critical applications in practice. This survey introduces some valuable applications related to high-dimensional causal effect estimation from observational data.

1.2.1 MEDICAL REPORT ASSISTANCE

In medical report generation, deep learning models are often trained on real-world reports, such as MIMIC-CXR Johnson et al. (2019a;b). Therefore, their outputs may contain hallucinated references to non-existent priors for an individual. If we regard patients' chest features, such as lung opacity, as treatment and clinic reports as outcome, with chest X-rays as covariate set, we can maximize the causal effect of chest features on the report under consistency constraints, rather than simply fitting a model for factual prediction performance.

1.2.2 UPLIFT MARKETING

In uplift marketing tasks, our aim is to maximize profits with limited advertising or recommendations, rather than simply predicting whether customers or users are interested in certain products. The metric that the company cares about is how much the promotion of products increases the company's income.

1.2.3 SELF-DRIVING ASSISTANCE

In self-driving assistance, the counterfactual outcome is almost always missing because a driver only has one life. However, it is critical for the model to understand the consequences of an action shown in an image, such as pressing the accelerator with high speed towards a tree.

1.2.4 BROAD-SPECTRUM ANTIVIRAL DRUG DISCOVERY

Due to the variety and variation of viruses and the heterogeneity of populations, randomization is not efficient enough for drug discovery. The combination of viruses, potential drugs, and individuals' attributes is exponential. If we can combine electronic health records of individuals and experimental results in the laboratory, it can make drug discovery more efficient.

2 PRELIMINARIES

2.1 FRAMEWORKS

Frameworks are fundamental for understanding the problem of causal effect learning. Two competitive frameworks, potential outcome and structural causal model, do not have overwhelming advantages over each other.

2.1.1 POTENTIAL OUTCOME

In 1923, Neyman published his paper Neyman (1923), but it was not translated into English until 1990 Splawa-Neyman et al. (1990). He used the term "unknown potential yield" to indicate the missing "potential outcome" in his randomization experiment for evaluating crop varieties. The Rubin causal model was named by Holland in 1986 Holland (1986).

In the Rubin causal model, the first step is to define the interested estimand (potential outcome) and then design the assignment mechanism before outcomes are measured. Then, a model is built to analyze the data.

There are some basic assumptions in the potential outcome framework. The Stable Unit Treatment Value Assumption (SUTVA) Rubin (1980) states that units/individuals/samples should be independent of each other, and the treatment effect for an individual is stable. Strong ignorability Rosenbaum & Rubin (1984) means that treatment assignment probability should be positive for every treatment value and every individual, and the assignment mechanism should be independent of potential outcomes. Consistency requires that subjects' response for a specific treatment in an experimental study is the same as the outcome in an observational study.

Recently, people are trying to find weaker assumptions of strong ignorability, such as single strong ignorability Wang & Blei (2019) D'Amour (2019) and sequential single strong ignorability Bica et al. (2020). These assumptions require the number of treatments to be more than one and assume the non-existence of multi-cause hidden confounders.

Other works focus on sensitivity analysis of causal inference to provide confidence intervals Franks et al. (2019). For example, Rosenbaum's sensitivity parameter Rosenbaum (1987) Γ and Bahadur's Bahadur (1971) efficiency were proposed. They try to separate the analysis of exogenous factors from the models.

2.1.2 STRUCTURAL CAUSAL MODEL

The Structural Causal Model (SCM) Pearl (2009) can be seen as a simplified world model. Estimating causal effect is its by-product. It requires the generation of both factual and counterfactual data for all units that share one SCM. In the SCM, we have two kinds of variables for a unit: observed variables and unobserved variables that are outside the observed variable list. The key of the SCM is that it assumes that we completely know how the data of a variable (both observed and unobserved) for a unit is generated. Because all units share an SCM, we know the list of variables $Parent(X)$ that an observed variable X depends on and the list of observed variables $Child(X)$ that depend on X for every unit. For a unit u, the randomness of its observed variables is from the randomness of unobserved variables; the unobserved variables should be measurable so that they can be represented by a distribution.

Also, the concrete generating function of all observed variables and the concrete distribution of the unobserved variable should be given. The requirement of the SCM is too strict, and it is proven almost impossible to learn an SCM from observational data Xia et al. (2021b) if all generating functions of an SCM can be represented by neural networks and unmeasured variables that can be represented by uniform distributions. Although the unobserved confounding problem can be solved efficiently given a SCM, it is clearly a waste and not realistic to learn such a world model for causal effect estimation because many SCMs can produce the same causal effect estimation.

2.2 BASIC NOTATIONS

Definition 1 *Unit. A unit is the object/individual that no isomorphism exists among its subsets in and after the observation study.*

A unit can be a patient, a customer, an individual, a subject, a physical object, or a measurable concept. We denote a unit as u, a unit variable as U, a unit set as $\mathbf{u}$, a unit variable set as $\mathbf{U}$.

Definition 2 *Observation. An observation is the data of measurement for a unit in the observation study.*

Definition 3 *Outcome. An outcome is the attribute set of unit we want to understand, to control, to intervene, or to change after the observation study.*

We denote an outcome as y, an outcome variable as Y, an outcome set as $\mathbf{y}$, an outcome variable set as $\mathbf{Y}$.

Definition 4 *Treatment/Intervention. Treatment is the action set that we want to take on the unit to intervene the outcome for certain purpose after the observation study.*

We denote an treatment as t, an treatment variable as T, an treatment set as $\mathbf{t}$, an treatment variable set as $\mathbf{T}$.

Definition 5 *Potential Outcome. For each treatment given a unit, the outcome that will be observed after the treatment assignment is called potential outcome.*

We denote a potential outcome for outcome Y with treatment $T = t$ as $Y(T = t)$ where potential outcome $Y(T = t)$ is also a variable.

Definition 6 *Observed outcome. For an observation given a unit with assigned treatment, the outcome in the observation study is called observed outcome.*

Definition 7 *Counterfactual outcome. For an observation given a unit with assigned treatment, the outcome if the unit had taken a different treatment in the observation study is called counterfactual outcome.*

The total number of observed outcome and counterfactual outcome given a unit equals the number of potential outcome of this unit. For binary treatment in the observation study, if $T = 1$ for a unit u then the observed outcome is $Y(T = 1)|u$ and the counterfactual outcome is $Y(T = 0)|u$.

Definition 8 *Covariate. For a unit with assigned treatment, if an attribute set has no intersection with treatment and outcome, then it is called covariate.*

Definition 9 *Pre-treatment covariate. For a unit with assigned treatment, the covariate that does not depend on treatment is called pre-treatment covariate.*

Definition 10 *Post-treatment covariate. For a unit with assigned treatment, the covariate that depended on treatment is called post-treatment covariate.*

When referring to covariates, unless otherwise specified, we assume that they are pre-treatment covariates. It is worth noting that pre-treatment is a property of data generation, rather than a sequence of data measurement. For instance, some attributes may have been measured after the treatment but generated before it. Conversely, just because data was measured before the treatment does not mean that the variable it represents was generated before the treatment. This is because there may exist two identical attributes, one measured before and the other after the treatment, which do not change in this study, such as height and age. Therefore, we need to be mindful of the distinction between data measurement and data generation when considering the timing of covariates.

2.3 BASIC OBJECTIVE

For the task of estimating causal effects, we are interested in the following quantities, which we refer to as "estimands" rather than "estimates" because their definitions are based on counterfactual outcomes. In the following basic objective, we assume that $X \cap Y = T \cap Y = X \cap T = \emptyset$ and that treatment is binary. The definitions of these objectives from the individual level to the group level and population level are as follows:

Definition 11 *Individual treatment effect (ITE). For unit i,*

$$ITE(i) \triangleq Y_i(T=1) - Y_i(T=0) \tag{1}$$

Definition 12 *Conditional Average Treatment Effect (CATE). For units $i \in \mathcal{U}$,*

$$\begin{aligned} CATE(\mathcal{U}) &\triangleq E_{i\in\mathcal{U}}(Y_i(T=1) - Y_i(T=0)|X_i = x_i) \\ &\triangleq E(Y(1) - Y(0)|X) \end{aligned} \tag{2}$$

Definition 13 *Conditional Average Treatment effect on the Treated group (CATT) Angrist & Imbens (1995). For units $i \in \mathcal{U}_t$ with $T_i = 1$,*

$$\begin{aligned} CATT(\mathcal{U}_t) &\triangleq E_{i\in\mathcal{U}_t}(Y_i(T=1) - Y_i(T=0)|X_i = x_i) \\ &\triangleq E(Y(1) - Y(0)|X, T=1) \end{aligned} \tag{3}$$

Definition 14 *Conditional Average Treatment effect on the Control group (CATC). For units $i \in \mathcal{U}_c$ with $T_i = 0$,*

$$\begin{aligned} CATC(\mathcal{U}_c) &\triangleq E_{i\in\mathcal{U}_c}(Y_i(T=1) - Y_i(T=0)|X_i = x_i) \\ &\triangleq E(Y(1) - Y(0)|X, T=0) \end{aligned} \tag{4}$$

Definition 15 *(Population) Average Treatment Effect (ATE/PATE). For units $u \in \mathcal{U}$,*

$$\begin{aligned} ATE(\mathcal{U}) &\triangleq E_{i\in\mathcal{U}}(Y_i(T=1) - Y_i(T=0)) \\ &\triangleq E(Y(1) - Y(0)) \end{aligned} \tag{5}$$

Definition 16 *(Population) Average Treatment Effect on the Treated group (ATT/PATT). For units $i \in \mathcal{U}$ with $T_i = 1$,*

$$\begin{aligned} ATT(\mathcal{U}_t) &\triangleq E_{i\in\mathcal{U}_t}(Y_i(T=1) - Y_i(T=0)) \\ &\triangleq E(Y(1) - Y(0)|T=1) \end{aligned} \tag{6}$$

Definition 17 *(Population) Average Treatment Effect on the Control group (ATC/PATC). For units $i \in \mathcal{U}$ with $T_i = 0$,*

$$\begin{aligned} ATC(\mathcal{U}_c) &\triangleq E_{i \in \mathcal{U}_c}(Y_i(T=1) - Y_i(T=0)) \\ &\triangleq E(Y(1) - Y(0)|T=0) \end{aligned} \tag{7}$$

2.4 BASIC MATRICS

Definition 18 *Precision in Estimation of Heterogeneous Effects (PEHE) Hill (2011). For units $i \in \mathcal{U}$ and estimator $\tilde{\tau}(Y_i(1) - Y_i(0)|X_i)$,*

$$\begin{aligned} PEHE(\mathcal{U}, \tilde{\tau}) &\triangleq E_{i \in \mathcal{U}}(||\bar{\tau}_i - \tilde{\tau}_i||_2) \\ &\triangleq \sigma_{\tilde{\tau}}(Y(1) - Y(0)|X) \end{aligned} \tag{8}$$

where $\bar{\tau}_i = E_{f_i}(Y_i(1) - Y_i(0)|X_i)$ and f_i is the sampling function of unit i, and $\tilde{\tau}_i$ is the estimated CATE of unit i.

Definition 19 *Precision in Estimation of Heterogeneous effects on the Treated group (PEHT). For units $i \in \mathcal{U}_t$ with $T_i = 1$ and and estimator $\tilde{\tau}(Y_i(1) - Y_i(0)|X_i, T_i = 1)$,*

$$\begin{aligned} PEHT(\mathcal{U}_t, \tilde{\tau}) &\triangleq E_{i \in \mathcal{U}_t}(||\bar{\tau}_i - \tilde{\tau}_i||_2) \\ &\triangleq \sigma_{\tilde{\tau}}(Y(1) - Y(0)|X, T=1) \end{aligned} \tag{9}$$

where $\bar{\tau}_i = E_{f_i}(Y_i(1) - Y_i(0)|X_i, T_i = 1)$ and f_i is the sampling function of unit i, and $\tilde{\tau}_i$ is the estimated CATT of unit i.

Definition 20 *Precision in Estimation of Heterogeneous effects on the Control group (PEHC). For units $i \in \mathcal{U}_c$ with $T_i = 0$ and and estimator $\tilde{\tau}(Y_i(1) - Y_i(0)|X_i, T_i = 0)$,*

$$\begin{aligned} PEHC(\mathcal{U}_c, \tilde{\tau}) &\triangleq E_{i \in \mathcal{U}_c}(||\bar{\tau}_i - \tilde{\tau}_i||_2) \\ &\triangleq \sigma_{\tilde{\tau}}(Y(1) - Y(0)|X, T=0) \end{aligned} \tag{10}$$

where $\bar{\tau}_i = E_{f_i}(Y_i(1) - Y_i(0)|X_i, T_i = 0)$ and f_i is the sampling function of unit i, and $\tilde{\tau}_i$ is the estimated CATC of unit i.

2.5 LEARNING CAUSAL EFFECT FORM HIGH-DIMENSIONAL DATA

In machine learning, we care about the effectiveness and efficiency of our causal model (causal estimator) for high dimensional data.

The main differences between a causal model and traditional predictive model is that the missing of counterfactual outcome is inevitable in both training and validation procedures but counterfactual outcome is indispensable in testing procedures. In the training and validation procedure, we can use data of covariate, treatment, and factual outcome of units for model learning. In the testing procedure, the missing counterfactual outcome is well-defined and valuable because the unit could have been assigned to a different treatment policy rather than a predestinate policy due to randomization of treatment, subjective initiative of unit, and inaccessibility of deterministic predictive model of treatment assignment.

In this report, the meaning of high dimensional causal effect estimation is: the dimension of outcome is larger than 100, the potential treatment assignment is larger than 100, the dimension of covariate is larger than 1000. In order to extend the basic objective for binary treatment to high dimensional cases while remaining the essential properties, the definition and objective of causal effect for machine learning will be reconsidered in this part. We hope machine's response for causal effect estimation is consistent with reality as much as possible after learning from the observation data and performing the recommend actions.

Definition 21 *Counterfactual imputation. Given factual observations* $(t_i, Y_i(t), x_i)$ *of* $u_{i=1}^n$*, the task to impute counterfactual outcome* $Y_i(\mathcal{T}_i/t_i)$ *is called counterfactual imputation where* $\mathcal{T}_i/t_i$ *is other potential treatments that* u_i *could have been assigned to.*

We remark that counterfactual imputation is an unsupervised task because counterfactual outcome can never been observed in both train dataset and validation dataset. The counterfactual imputation task is the fundamental challenge of causal inference.

Definition 22 *Generalized Individual Treatment Effect. For unit* i *with multiple treatment, the treatment effect of treatment* t_i *is*

$$GITE = Y_i(t_i) - E(Y_i(\mathcal{T}_i/t_i)) \tag{11}$$

The definition of GITE is based on such consideration. It may be controversial to image all potential treatments and their potential outcomes of an observation. And it is also meaningless to list all possible treatments and potential outcomes of this observation because the same observation will seldom be observed after the observation study. However, it seems spontaneously for us to image expectation of potential outcomes of other different treatments for a certain observation of a unit. Other targets of causal effect can be extended to high dimensions following similar way and those definitions are very useful.

The first case is model learning. Given generative model of outcome and treatment, it is reasonable to maximize the GITE which is constrained by observation consistency because we know that most treatment actions in ICU are decided by experienced and knowledgeable doctors and our model is obviously not as good as doctor at the start of learning.

The second case is optimal treatment discovery. If we want to find the optimal treatment to maximize the outcome, we could search the treatment with maximum GITE. Also, we can minimize the GITE by treatment assignment model's parameters and chose treatment with maximum GITE.

3 RELATED WORKS

The primary challenge in estimating causal effects is the confounding problem, which can arise from both observed and unobserved variables or mechanisms. In this survey, we assume that treatment assignment is probabilistic and individualistic Imbens & Rubin (2015).

To address the issue of confounding, the unconfoundedness (ignorability) assumption is often used, which assumes that there are no hidden confounding variables. We can divide causal learning approaches into two categories based on whether they make this assumption or not. However, the presence of hidden confounding can affect the performance of algorithms that assume unconfoundedness. The performance of causal learning algorithms depends on several factors, including the learning stage, counterfactual imputation, balancing regularization, potential outcome prediction, estimand modeling, and the presence of hidden confounding. These factors are summarized in Table 1.

Table 1: Algorithms of causal effect learning from observation data. BLR/BNN: Shalit et al. (2017);TARNet/CFR-MMD/CFR-Wasserstein: Johansson et al. (2016);Dargonet: Shi et al. (2019);X-learner: Künzel et al. (2019);CEVAE: Louizos et al. (2017);Deconfounder: Wang & Blei (2019);GANITE: Yoon et al. (2018);SITE: Yao et al. (2018);DRNets: Schwab et al. (2020);VCNets: Nie et al. (2021).

Algorithms	Learning Stage	Counterfactual Imputation	Balancing Regularization	Potential Outcome Prediction	Estimand Modeling	Hidden Confounding
BLR BNN	Two-stage	Nearest Neighbor	Moment's Difference	Linear Neural Network	None	None
TARNet CFR-MMD CFR-Wasserstein Dargonnet	End-to-end	Perfect Counterfactual	None MMD Wasserstein CrossEntropy	Twin Neural Networks	None	None
X-Learner	Three-stage	Perfect Counterfactual	None	Twin BARTs	Yes	None
CEVAE	End-to-End	Perfect Counterfactual	Bayesian Variational Inference Network	Model Network	None	Proxy variables
Deconfounder	Two-stage	Perfect Counterfactual	Posterior Predictive Check of Factor Model	Linear	None	Proxy variables
GANITE	Two-stage	Counterfactual GAN	None	ITE GAN	None	None
SITE	End-to-end	PDDM Similarity	Middle Point Distance	Neural Network	None	None
DRNets VCNets	End-to-end	Nearest Neighbor	None	Treatment-Dose Networks Varying Coefficient Network	None	None

3.1 OBSERVED CONFOUNDING

To address observed confounding, many causal effect learning approaches rely on the unconfoundedness assumption Imbens & Rubin (2015). This assumption implies that the treatment assignment t_i is independent of the potential outcomes $(Y_i(t_i), Y_i(\mathcal{T}/t_i))$, given the covariates X_i. However, the fundamental difficulty of observed confounding is that a unit with treatment can be measured only once, so the ratio of missing counterfactual data is inevitably very high (at least 50%), as compared to standard regression tasks. This makes it challenging to estimate the causal effect accurately.

3.1.1 COUNTERFACTUAL IMPUTATION

In the context of causal effect estimation, comparing identical individuals/units with different treatments is an intuitive idea to estimate causal effect, but it is almost impossible to find exact matches due to the exponential requirement of units with the increase of covariates' dimensionality.

Balancing is an alternative approach based on randomization, which does not require identical units with different treatments. It is based on the fact that for a certain deviation tolerance of the treatment effect (or fixed deviation tolerance of covariates), the probability to reject the match is exponentially decreasing with the number of units in each group with different treatments if all units were sampled from the same distribution (Gaussian and sub-Gaussian) and assigned to different treatment groups randomly Rosenbaum & Rubin (2022).

To simulate the randomness of treatment assignment, a balancing score $b(\mathbf{X})$ is used to make no significant difference between the covariates of the treatment and control groups. A balancing score is any function that satisfies $\forall \mathbf{x}(\mathbf{t} \perp \mathbf{x}|b(\mathbf{x}))$. As a consequence, $Pr(\mathbf{x}|\mathbf{t_1}, b(x)) = Pr(\mathbf{x}|\mathbf{t_2}, b(\mathbf{x}))$ and $Pr(\mathbf{t}|\mathbf{x_1}, b(\mathbf{x_1})) = Pr(\mathbf{t}|\mathbf{x_2}, b(\mathbf{x_2}))$, where $b(\mathbf{x_1}) = b(\mathbf{x_2})$. Examples of balancing scores include exact matching ($b(\mathbf{x}) = \mathbf{x}$), propensity score matching ($b(\mathbf{x}) = Pr(\mathbf{t}|\mathbf{x})$), and principal unobserved covariate matching ($b(\mathbf{x}) = Pr(\mathbf{t}|\mathbf{x}, \mathbf{Y}(\mathbf{t}))$) Rosenbaum & Rubin (1983).

Regardless of the balancing score used for matching, we use the matched unit's observed outcome, which has a different treatment but the same balancing score, to impute the counterfactual outcome of the current unit because the difference in outcomes between those matched units is only from the source of treatment assignment probability's randomness.

Re-weighting can be seen as a method to achieve the same balancing score for units with different treatment groups. For example, Inverse Probability of Treatment Weighting (IPTW) Rosenbaum & Rubin (1983) first uses the propensity score as the initial balancing score for all units. Then IPTW

uses a re-weighting function to create a new balancing score for different treatment groups, which forces those groups to have the same balancing score. Finally, Average Treatment Effect (ATE) can be computed by IPTW.

Stratification can be seen as a method to achieve the same balancing score with both different treatment groups and similar covariates. For example, the equal frequency approach Rosenbaum & Rubin (1983) splits the sub-groups by propensity score. It can be used to compute Conditional Average Treatment Effect (CATE).

Recent works focus on learning low-dimensional balancing representations and distribution distance metrics by neural networks Clivio et al. (2022) Wang et al. (2021) rather than probabilities whose dimension is the same as the number of treatment assignments.

3.1.2 Estimand Modeling

In causal inference, an estimand is a parameter that describes the causal effect of a treatment on an outcome. Common examples of estimands include the average treatment effect (ATE) and the conditional average treatment effect (CATE). While these parameters can be estimated using individual counterfactual predictions, it is often more efficient to build a model that directly estimates the estimand.

For example, in the CATE estimation problem with binary treatment, several approaches have been proposed to model the estimand. S-learner learns a single model that takes both the treatment assignment and the covariates as input and predicts the outcome. T-learner learns two separate models for the treatment and control groups and uses them to estimate the difference in outcomes. X-learner Künzel et al. (2019) is a more flexible approach that builds an estimand model to handle unbalanced treatment assignment. It learns two outcome models for the treatment and control groups and then calculates two imputed treatment effects from the observed and counterfactual outcomes. The two imputed treatment effects are then weighted and combined using a chosen weight function Kallus (2020), such as the propensity score, to produce an estimate of the CATE.

The performance of these meta-learners depends heavily on the choice of base models. In practice, tree-based models such as Bayesian Additive Regression Trees (BART) Chipman et al. (2010) have been found to work well. Estimand modeling is a powerful approach to estimating causal effects, and its use is becoming increasingly popular in the machine learning and causal inference communities.

3.2 UNOBSERVED CONFOUNDING

In reality, it is impossible to collect data on all background variables and even define them well sometimes. As a result, unobserved confounding is not a rare occurrence; it is a common phenomenon and sometimes we may be more confident in the existence of hidden confounders than in the family of data generating functions or distributions. For example, patients' mental state, social status, and doctors' knowledge may all be unobserved confounders.

3.2.1 IDENTIFICATION AND APPROXIMATION

Causal effect identification can transform the query about the interested effect to operational intervention and observable observation even when hidden confounding exists. If the causal effect is not identifiable, then approximation methods can be used to get a bound on the causal effect. These approximation methods can be regarded as a fast inference tool for Structural Causal Models (SCMs). The precondition of such approaches is that the real data generating procedure can be represented by our presumed SCM. The limitation of such approaches is the learning difficulty of SCM Xia et al. (2021b).

Identification formulas for causal diagrams were developed in the last 30 years based on the definition of Pearl's structural causal model. Back-door adjustment, front-door adjustment, and do-calculus for Directed Acyclic Graphs (DAGs) were named, and the proof of those theorems was given in Pearl (1993) and Pearl (1995). However, the approach of such identification does not consider unobservable confounders or automatic identification algorithms, and the completeness of such identification methods was also not given. In 2002, Tian & Pearl (2002) proposed a complete criterion "c-factorization" for singleton treatment and singleton outcome. Huang & Valtorta (2006) and Shpitser & Pearl (2006a) proposed complete identification algorithms (Huang's algorithm and Shpiser's **ID** algorithm) to transform causal effect queries without condition variables into functions of observation distribution automatically for multiple treatments and outcomes in Bayesian networks with hidden variables and semi-Markovian models, respectively. And Shpitser & Pearl (2006b) proposed the **IDC** algorithm for causal effect queries with condition and proved the completeness. However, all these identification methods do not consider the undirected edges (stable symmetric relationships). In 2019, Sherman & Shpitser (2018) proposed a complete identification algorithm for segregated graphs to address such patterns. There are also other identification algorithms for causal diagrams with loops Forré & Mooij (2020).

However, it is also meaningful to not assume that any intervention on those variables is impossible after observational studies because active intervention will introduce information that observation cannot provide. Bareinboim & Pearl (2012) defined z-identifiability and proposed the complete $\mathbf{ID}^{\mathbf{z}}$

algorithm to address the problem that any combination of experiments on $\mathbf{Z}$ can be performed, and observable distribution is known for queries without condition variables. Lee et al. (2019) defined g-identifiability and proposed the **gID** algorithm. It can factorize the original causal effect query into the expression of intervention distribution of $\mathbf{Z}$, and it does not need any observational data.

Recently, researchers have begun to notice that solving the identification problem in SCM (structure causal model) directly is not the only way. Lee & Bareinboim (2021) revealed the connection between matrix theory and traditional identification and proposed an algorithm that leverages proxy-based methods and traditional methods. Neural identification was first proposed and theoretically analyzed in Xia et al. (2021a), who also proved the completeness of their neural identification algorithm. However, such neural identification requires retraining models if the assignment values of T and Y are changed.

Compared to do-calculus based algorithms for structure causal model, po-calculus Malinsky et al. (2019) with single world intervention graph (SWIG) Richardson & Robins (2013) is a useful complete identification method in the potential outcome framework.

In cases where identification is not possible, we can still give a bound to the intervention query from observation data. Balke & Pearl (1997) give the tightest bound to a graph with instrument variables. Recently, Zhang & Bareinboim (2021) gave a tighter bound than the natural bound for a general DAG by utilizing observation data.

3.2.2 PROXY VARIABLES

Proxy variables approaches assume that the joint distribution of hidden confounders and observed variables $p(Z, X, T, Y)$ can be approximated from observed data (X, T, Y). Goodfellow et al. (2016) listed some cases where such approximation is possible.

One example of a proxy variable approach is the Counterfactual Variational AutoEncoder (CEVAE) Shalit et al. (2017). CEVAE uses a non-parametric causal diagram prior to factorize the causal effect into observation probability $p(Z, X, T, Y)$. It uses $E(Y(T)) = E_{p(Z|X)}(Y|Z, T, X)$ to estimate factual outcome and counterfactual outcome. The limitation of CEVAE is that the Z CEVAE learned may contain mediators if the causal diagram prior was misspecified. Additionally, CEVAE assumes that all individuals were generated by a single diagram.

Another proxy variable approach is the Time-series Deconfounder Bica et al. (2020), which is a time-series neural network version of Deconfounder Wang & Blei (2019) that is focused on multiple treatments and linear models. Time-series Deconfounder assumes no hidden common cause

between a single treatment variable and potential outcomes. Thus, if treatment variables are independent of each other given some substitute confounder, the hidden multi-cause confounder cannot exist because if it existed, such independence would not hold due to d-separation. Time-series Deconfounder first learns a factor model $p(Z, T) = p(Z)p(T|Z)$, where Z is the substitute confounder, and then does a predictive check (similar to a generalization ability check) of $p(T|Z)$. It then learns an outcome model which inputs are treatment and the substitute confounder to predict potential outcomes.

3.3 DIMENSIONALITY REDUCTION

In order to deal with high-dimensional covariates, some dimensionality reduction approaches may be helpful for causal effect estimation tasks.

Current dimensionality reduction research can be divided into three classes according to the reduction target. The first class is to determine the dimension based on information loss. For example, Xia et al. (2009) minimize regression mean squared error (MSE) from cross-validation for a linear model with a kernel. Dong & Gao (2021) propose a Lagrange loss with a binary mask π for variational autoencoders (VAE) and prove its convergent dimension is a local minimum. However, the hidden distribution is usually in Gaussian space, which is often regarded as an "uninteresting" signal noise due to the *central limit theorem*. The second class evaluates the non-Gaussianity of latent space. For example, Darnell et al. (2017) assign a stability score to the principal component and regard the change point with the smallest p-value as an indicator. Non-Gaussian component analysis (NGCA) Blanchard et al. (2006); Bean (2014); Goyal & Shetty (2019) assumes Gaussian noise is independent of the non-Gaussian subspace, and they discard the Gaussian component to determine the signal space. However, the algorithm is either exponential related to the dimension of the non-Gaussian subspace due to the error of accumulation Goyal & Shetty (2019) or the polynomial time is unacceptable. Therefore, it cannot be applied directly to general high-dimensional data. The third class is the end-to-end approach for a specified task. For example, Ding & Li (2007) and Luo et al. (2018) search for the most discriminative subspace for clustering. Recently, Baggenstoss & Kay (2022) propose a general approach based on probability density function (PDF) estimation without assumption about data structure, although the choice of hidden dimension is empirical. Tavory (2019) use normalized maximum likelihood to determine the principal component cardinality. Table 2 illustrates the assumptions of representative dimensionality reduction methods.

PCA Pearson (1901)Hotelling (1933) is a widely used linear dimensionality reduction technique. It uses orthogonal transformation to obtain the uncorrelated principal components. Autoencoder Kramer (1991) Kingma & Welling (2013), on the other hand, is a non-linear dimensionality re-

Table 2: Dimensionality reduction assumptions. G: Gaussian; I: independent; nG: non-Gaussian; $\perp$: orthogonal; $\rightarrow$: generate; ANN: additive normal noise; DAG: directed acyclic graph.

Method	Mapping	$p(\mathbf{z})$	$p(\mathbf{x})$
PCA	Linear	IG	IG
ICA	Linear	InG	InG+G
t-SNE	Nonlinear	Local continuty	Local continuty
βVAE	Nonlinear	IG with β	\
NGCA	Linear	G$\perp$nG	ANN
LinGAM	Linear	G$\rightarrow$nG	ANN with DAG

duction technique that typically uses neural networks and gradient-based optimization to learn the parameters for efficient computation. In autoencoders, the reconstruction error is an important part of the loss function.

CausalVAE Yang et al. (2021) introduces causal effect learning among latent variables through labeled data and prior distribution of labels. The key to their success in learning the DAG over labels is the difference in distributions between the causal and anti-causal directions.

3.4 Evaluation

3.4.1 Benchmark Datasets

The use of large datasets and benchmarks in research has been shown to be significant, as demonstrated by the impact of ImageNet. Benchmarking on datasets can help to evaluate hypotheses, algorithms, and models. However, there are few large datasets collected from reality for causal effect learning tasks. There are two main challenges to benchmarking causal algorithms and models that are different from traditional correlation data benchmarking. Firstly, evaluating interventions often requires far more time and money than predictions for algorithms and models. Sometimes interventions can even be unethical. For example, we cannot encourage or force someone to smoke or make someone sick. Secondly, counterfactual data can never be collected in the real world, and there is a lack of credible methodology and sufficient representative research to transform real datasets into counterfactual datasets. Table 3 provides some datasets that may be useful.

The benchmark datasets used for evaluation often rely on simulation after randomized experiments (SaRE) or well-matched twins (MT). For example, IHDP Hill (2011), Jobs Smith & Todd (2005), and Simulated GWAS data Song et al. (2015) (Mendelian randomization) are based on the National Supported Work Demonstration experiment (begun in 1986), the Infant Health and Development Program (begun in 1985), and the Northern Finland Birth Cohort data (published in 2009), respectively. The outcome models (response surfaces) are often linear/generalized linear models. Otherwise, matched twins data is used for evaluation. For instance, the Twins dataset Louizos et al.

(2017) is from twin births data from 1989 to 1991 in the USA. In the Twins dataset, the counterfactual mortality of one twin was regarded as the mortality of the other twin.

The approaches to generate counterfactual outcomes, which will never be observed, cannot be trivially applied to high-dimensional data. For the SaRE approach, the high-dimensional treatment set requires exponentially-scaled randomized experiments. For the MT approach, the high-dimensional covariate set requires strict balancing inside twins, which means that the covariate set must be sampled from the same distribution. The high-dimensional outcome set requires us to understand the exponential-scale entanglement inside the outcome set.

There is also an approach to evaluate causal effects that requires intervention after learning. For example, the uplift model learned in the development and test environment will be uploaded to the online environment for further evaluation Zhao & Harinen (2019). However, the cost of such evaluation is too high to follow for developing causal effect estimation algorithms rapidly.

As an alternative, factual outcome prediction performance will also be considered as a metric of causal effect estimation sometimes. However, we will not take it seriously in this survey because such alternatives ignore the fundamental problem in causal effect estimation: missing counterfactual outcome data.

3.4.2 CAUSAL PACKAGES

Another perspective for building an experimental platform is to maintain unified packages in the causal toolbox. This can help researchers to propose and test novel ideas quickly, thus promoting the development of causal science. There are many packages that implement pipelines for causal learning or reasoning. Some of them provide standard and state-of-the-art learning and reasoning algorithms, such as the causal-learn package. Related work about causal packages is illustrated in Table 4.

Table 3: Causal Dataset. Causeme: 202; JustCause: Hawkins & Kim (2021); e-CARE: Du et al. (2022); IHDP: Hill (2011); News: Johansson et al. (2016); Twins: Louizos et al. (2017); Jobs: Shalit et al. (2017); Movies: Wang & Blei (2019); GWAS: Song et al. (2015).

Type	Name	Introduction	Website
Benchmark	Causeme	time-series	https://causeme.uv.es/
Benchmark	JustCause	support IHDP, ACIC etc.	https://justcause.readthedocs.io/en/latest/
Benchmark	e-CARE	reasoning and explanation for NLP	https://scir-sp.github.io
Dataset	IHDP	home visits and IQ testing	https://github.com/vdorie/npci
Dataset	News	New York Times corpus	https://archive.ics.uci.edu/ml/datasets/Bag+of+Words
Dataset	Twins	birth weight and mortality	http://www.nber.org/data/linked-birth-infant-death-data-vital-statistics-data.html
Dataset	Jobs	labor earnings	https://users.nber.org/ rdehejia/data/.nswdata3.html
Dataset	Movies	Movie income and stars	https://www.kaggle.com/tmdb
Dataset	GWAS	genome-wide association studies	https://github.com/StoreyLab/gcatest
Competition	ACIC 2022	conference challenge	https://acic2022.mathematica.org/data
Competition	PCIC 2022	conference challenge	https://pattern.swarma.org/pcic/competition.html

Table 4: Causal Packages. Tetrad: Ramsey et al. (2018); CausalDiscoveryToolbox: Kalainathan & Goudet (2019); Ananke: Nabi et al. (2020), Lee & Shpitser (2020), Bhattacharya et al. (2020); EconML: Keith Battocchi (2019); dowhy: Sharma et al. (2019); causalml: Chen et al. (2020); Causal-Curve: Kobrosly (2020); grf: Athey et al. (2019); dosearch: Tikka et al. (2021); causaleffect: Tikka & Karvanen (2017); dagitty: Textor et al. (2016).

Motivation	Toolbox	Support Team	Introduction
Causal Learning	causal-learn	CMU, DMIR, Gong Mingming team, Shouhei Shimizu team	python version of Tetrad
	Tetrad	CMU	Java
	CausalDiscoveryToolbox	FenTechSolutions	python, DAG/Pair, dataset, independence, structure learning, metrics
	gCastle	Huawei Noah	python, data generation and process, causal structure learning, metrics
	tigramite	Jakob Runge	python, learning from time-series data
Causal Reasoning	Ananke	Ilya Shpitser team	python, support do-calculus
	EconML	Microsoft	python, Econometrics
	dowhy	Microsoft	python
	causalml	Uber	python, campaign target optimization, personalized engagement
	CausalImpact	Google	R, time-series, adertisement and click
	WhyNot	John Miller	python, simulator and environment
	Causal-Curve	Kobrosly, R.W.	python, continuous variable such as price, time and income
	grf	grf-lab of Standford	R
	dosearch	Santtu Tikka	R
	causaleffect	Santtu Tikka	R
	dagitty	\	R, support adjustment formula
End-to-End	causalnex	QuantumBlack	python, 0.11.1, structure learning, domain knowledge, estimation
	Y-learn	CSDN	python, published in June 2022

4 OUR PRELIMINARY WORK

4.1 PACKAGES OF IDENTIFICATION ALGORITHM AND SCM

Learning causal effects for high-dimensional data is a challenging task that requires the use of automated identification algorithms to efficiently identify the causal effect before the learning process. To this end, we implemented Shpitser's complete identification algorithm, as existing open-source codes such as causaleffect, Ananke, dowhy, and dagitty Textor et al. (2016) did not provide the complete identified mathematical expression required for the algorithm. Our implementation is based on the Python programming language, with a causal diagram as input and a mathematical expression using LaTeX language as output.

We conducted experiments to identify the Average Treatment Effect (ATE) and Conditional Average Treatment Effect (CATE) given data distribution of covariate, treatment, and outcome. Our identification results are available on the following website: "https://github.com/herdonyan/EstimandIdentification/blob/main/3variablesfigs/graphid_web.html". The website can be previewed by adding "https://htmlpreview.github.io/?" before the website name.

Moreover, we have developed a software package capable of generating samples from a given structural causal model (SCM) consisting of a diagram, distributions, and functions. This package facilitates the simulation of SCM datasets and can be utilized to evaluate the performance of causal inference methods. The package is put into the same repository with ID algorithms and identified graphs.

4.2 HIDDEN CONFOUNDING AND OUT-OF-DISTRIBUTION GENERALIZATION

From the identification result, we can train the prediction model and compute causal effect following the factorization results. However, we wondered what would happen if we did not do identification but just prediction. For example, the identification result of figure 1 is $P(C|do(S)) = \frac{\sum_d P(d)P(S,C|d,B)}{\sum_d P(d)P(S|d,B)}$. We choose $C^* = \arg_c \max P(c|do(S))$ as prediction value. The conditional prediction is $E(C|S,D,B)$. The average prediction is $E(C)$. In the following, we use X_1 denote dopamine, X_2 denote brain, T denote smoking, and Y denote lung cancer.

The experimental properties we are interested in about our model and algorithm after identification is OOD generalization under parametric interventions from correct identification comparing with pure prediction. It can be measured in two aspects: OOD unbiasness and variance. If the estimand is $E(Y_i(1)-Y_i(0))$, then we can use ATE and PEHE as unbiasness and variance measurement respectively.

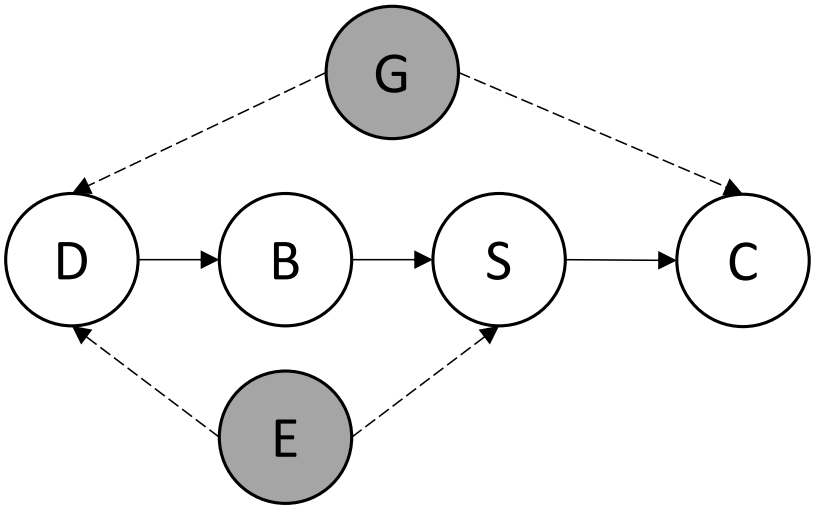

Figure 1: Example of four variables. D means dopamine; B means senior brain activity (frontal lobe); G means unobserved gene/physique; E means social environment not easy to measure. S means smoking behaviour, and C means cancer. For example, $E \rightarrow D$ may represent some life pressures, and $E \rightarrow S$ may be unconscious mimic nature.

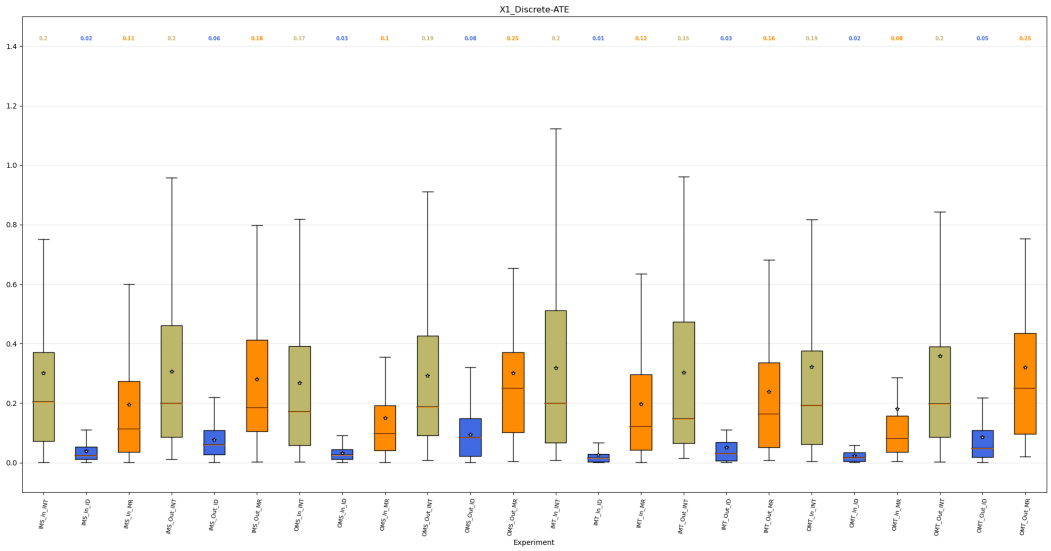

Figure 2: Experiment error for ATE estimation where X1 is discrete. Star is median value. Red line is average value. 'I' means inner mechanisms, and 'O' means outer mechanisms. 'S' means the parametric intervention is mechanism shifting, and 'T' means the parametric intervention is random transformation of mechanism.

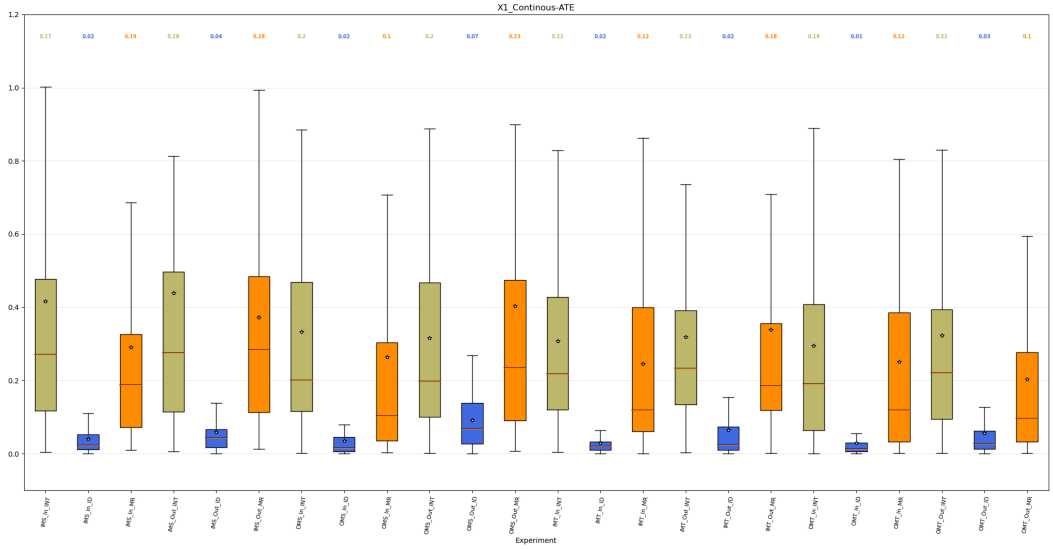

Figure 3: Experiment error for ATE estimation where X1 is continuous. Star is median value. Red line is average value. 'I' means inner mechanisms, and 'O' means outer mechanisms. 'S' means the parametric intervention is mechanism shifting, and 'T' means the parametric intervention is random transformation of mechanism.

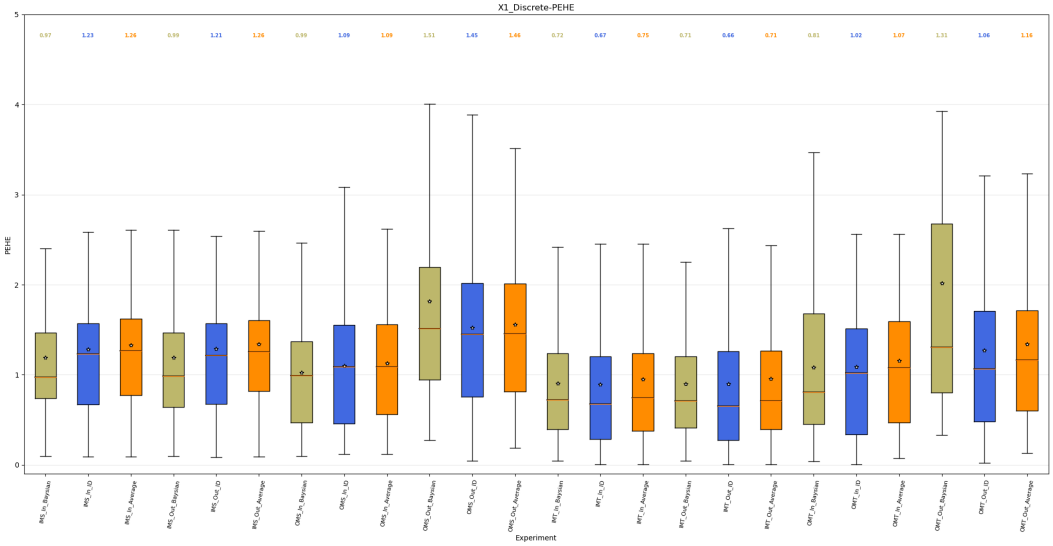

Figure 4: Experiment error for PEHE estimation where X1 is discrete. Star is median value. Red line is average value. 'I' means inner mechanisms, and 'O' means outer mechanisms. 'S' means the parametric intervention is mechanism shifting, and 'T' means the parametric intervention is random transformation of mechanism.

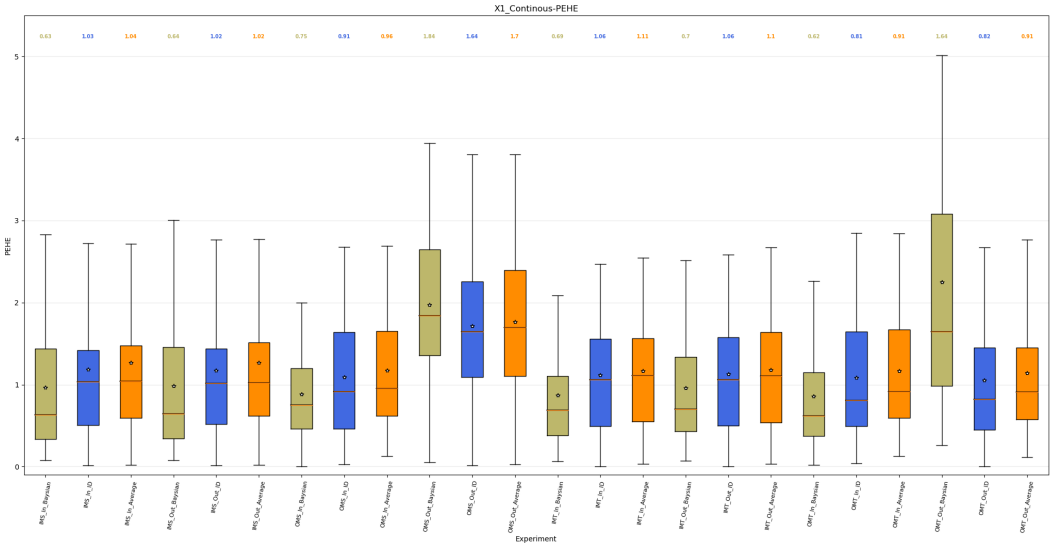

Figure 5: Experiment error for PEHE estimation where X1 is continuous. Star is median value. Red line is average value. 'I' means inner mechanisms, and 'O' means outer mechanisms. 'S' means the parametric intervention is mechanism shifting, and 'T' means the parametric intervention is random transformation of mechanism.

In our experiment, we use the linear model (same structure with figure 1) as a real-world model to generate data and test the out-of-distribution generalization ability. Each predictor of our association layer model is linear regression or classification model. To keep the consistency with X-learner, we also use two models for treatment and control group separately. We use random transformation and shifting of mechanisms as parametric intervention to test the robustness of our framework. For every setting, we run 50 independent experiments to evaluate the result where there are 1000 samples totally in each experiment.

The train sample number is 800, and the train/valid splitting is 640:160. The test sample number is 200. In algorithm 2 and 3, the sampling numbers of X_1 and (Y, T) are both 100. The dimension of every variable is 1. In optimization, the max epoch is 100000, and we will stop if there is no decrease of loss above 20 and 100 epochs for continuous and discrete testing, respectively. The loss function is MSE loss for regression and Cross Entropy loss for classification; the learning rate is 0.001. When positivity is not satisfied or the joint distribution is zero, we will resample data. The T are discrete variables and X_2 and Y are continuous variables. X_1 can be continuous or discrete variable. We don't use variational method to fitting function of error variance, and use prior noted in the paper directly due to convenience. All the experiment are independent. Figure 6 shows some continuous data. In those figures, left part is train data, and right part is testing data. Yellow and purple means different treatment assignments. And z-axis is value of Y.

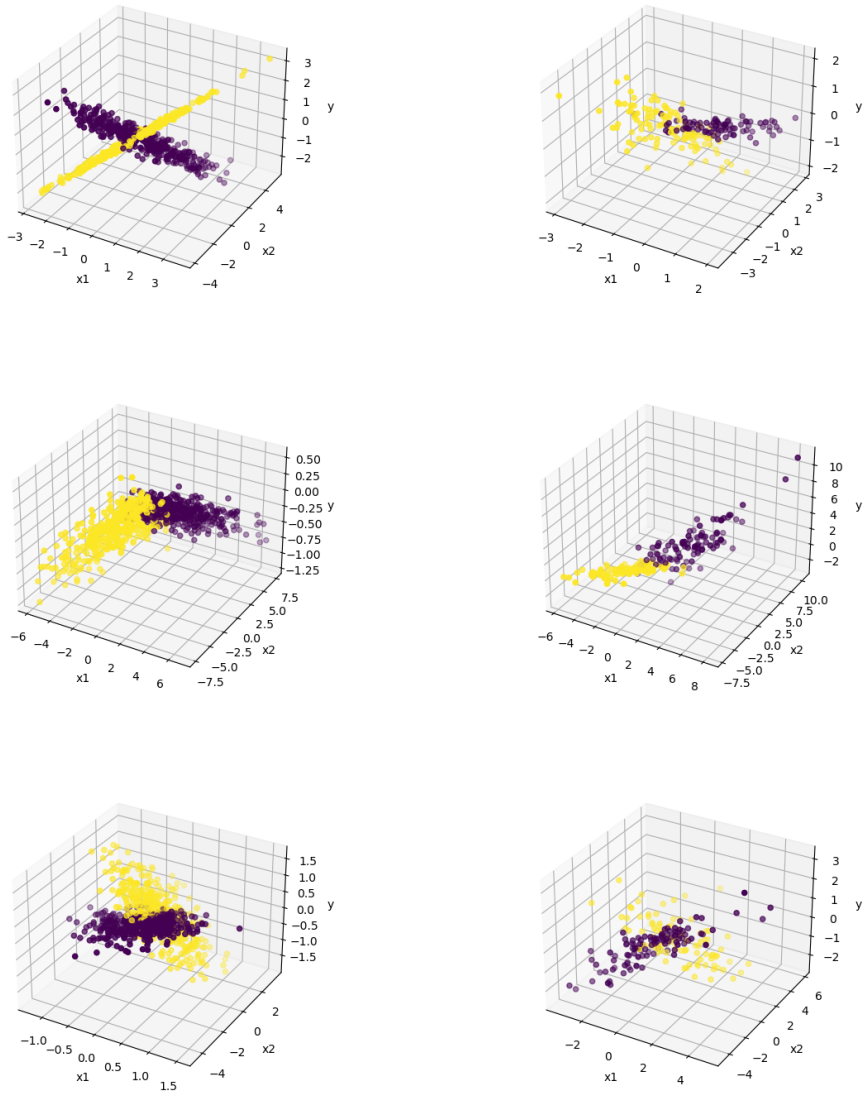

Figure 6: Some samples of generated data

Although nonlinear model is not used in our experiments, it can still work if there are nonlinear predictors and environments.

Figure 2, 3, 4, and 5 show the experiment results. We should notice that in-sample testing is not only IID testing due to the missing counterfactual data, and our out-sample testing is under those parametric interventions. In unbiasness testing, estimations after identification are more unbiased than MR Freedman (2008) and INT Lin (2013) from ATE estimation result in both discrete and continuous cases. Considering estimation variance, it got better performance when outer mechanisms are changed.

REFERENCES

Causeme: An online system for benchmarking causal discovery methods.

Joshua Angrist and Guido Imbens. Identification and estimation of local average treatment effects, 1995.

Susan Athey, Julie Tibshirani, and Stefan Wager. Generalized random forests. *The Annals of Statistics*, 47(2):1148–1178, 2019.

Paul M Baggenstoss and Steven Kay. Nonlinear dimension reduction by pdf estimation. *IEEE Transactions on Signal Processing*, 70:1493–1505, 2022.

Raghu Raj Bahadur. *Some limit theorems in statistics*. SIAM, 1971.

Alexander Balke and Judea Pearl. Bounds on treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association*, 92(439):1171–1176, 1997.

E. Bareinboim and J. Pearl. Causal inference by surrogate experiments: z-identifiability. In N. Freitas and K. Murphy (eds.), *Proceedings of the 28th Conference on Uncertainty in Artificial Intelligence*, pp. 113–120, Catalina Island, CA, Aug 2012. AUAI Press.

Derek Merrill Bean. *Non-Gaussian component analysis*. University of California, Berkeley, 2014.

Rohit Bhattacharya, Razieh Nabi, and Ilya Shpitser. Semiparametric inference for causal effects in graphical models with hidden variables. *arXiv preprint arXiv:2003.12659*, 2020.

Ioana Bica, Ahmed Alaa, and Mihaela Van Der Schaar. Time series deconfounder: Estimating treatment effects over time in the presence of hidden confounders. In *International Conference on Machine Learning*, pp. 884–895. PMLR, 2020.

Gilles Blanchard, Motoaki Kawanabe, Masashi Sugiyama, Vladimir Spokoiny, Klaus-Robert Müller, and Sam Roweis. In search of non-gaussian components of a high-dimensional distribution. *Journal of Machine Learning Research*, 7(2), 2006.

Huigang Chen, Totte Harinen, Jeong-Yoon Lee, Mike Yung, and Zhenyu Zhao. Causalml: Python package for causal machine learning. *arXiv preprint arXiv:2002.11631*, 2020.

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

Oscar Clivio, Fabian Falck, Brieuc Lehmann, George Deligiannidis, and Chris Holmes. Neural score matching for high-dimensional causal inference. In *International Conference on Artificial Intelligence and Statistics*, pp. 7076–7110. PMLR, 2022.

Alexander D'Amour. On multi-cause causal inference with unobserved confounding: Counterexamples, impossibility, and alternatives. *arXiv preprint arXiv:1902.10286*, 2019.

Gregory Darnell, Stoyan Georgiev, Sayan Mukherjee, and Barbara E Engelhardt. Adaptive randomized dimension reduction on massive data. *Journal of Machine Learning Research*, 2017.

Chris Ding and Tao Li. Adaptive dimension reduction using discriminant analysis and k-means clustering. In *Proceedings of the 24th international conference on Machine learning*, pp. 521–528, 2007.

Yiran Dong and Chuanhou Gao. An adaptive dimension reduction algorithm for latent variables of variational autoencoder. *arXiv preprint arXiv:2111.08493*, 2021.

Li Du, Xiao Ding, Kai Xiong, Ting Liu, and Bing Qin. e-care: a new dataset for exploring explainable causal reasoning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 432–446, 2022.

Patrick Forré and Joris M Mooij. Causal calculus in the presence of cycles, latent confounders and selection bias. In *Uncertainty in Artificial Intelligence*, pp. 71–80. PMLR, 2020.

AlexanderM Franks, Alexander D'Amour, and Avi Feller. Flexible sensitivity analysis for observational studies without observable implications. *Journal of the American Statistical Association*, 2019.

David A Freedman. On regression adjustments to experimental data. *Advances in Applied Mathematics*, 40(2):180–193, 2008.

Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT Press, 2016.

Navin Goyal and Abhishek Shetty. Non-gaussian component analysis using entropy methods. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, pp. 840–851, 2019.

Ty Hawkins and Andrew Kim. Just cause. In *Just War Theory and Literary Studies*, pp. 55–83. Springer, 2021.

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

Gábor Hofer-Szabó, Miklós Rédei, and László E Szabó. *The principle of the common cause*. Cambridge University Press, 2013.

Paul W Holland. Statistics and causal inference. *Journal of the American statistical Association*, 81 (396):945–960, 1986.

Harold Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417, 1933.

Yimin Huang and Marco Valtorta. Pearl's calculus of intervention is complete. In *UAI '06, Proceedings of the 22nd Conference in Uncertainty in Artificial Intelligence, Cambridge, MA, USA, July 13-16, 2006*. AUAI Press, 2006.

Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

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

Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):1–8, 2019a.

Alistair EW Johnson, Tom J Pollard, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Yifan Peng, Zhiyong Lu, Roger G Mark, Seth J Berkowitz, and Steven Horng. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*, 2019b.

Diviyan Kalainathan and Olivier Goudet. Causal discovery toolbox: Uncover causal relationships in python. *arXiv preprint arXiv:1903.02278*, 2019.

Nathan Kallus. Generalized optimal matching methods for causal inference. *J. Mach. Learn. Res.*, 21:62–1, 2020.

Maggie Hei Greg Lewis Paul Oka Miruna Oprescu Vasilis Syrgkanis Keith Battocchi, Eleanor Dillon. EconML: A Python Package for ML-Based Heterogeneous Treatment Effects Estimation. https://github.com/microsoft/EconML, 2019. Version 0.x.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Roni W Kobrosly. causal-curve: A python causal inference package to estimate causal dose-response curves. *Journal of Open Source Software*, 5(52):2523, 2020.

Mark A Kramer. Nonlinear principal component analysis using autoassociative neural networks. *AIChE journal*, 37(2):233–243, 1991.

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

Jaron JR Lee and Ilya Shpitser. Identification methods with arbitrary interventional distributions as inputs. *arXiv preprint arXiv:2004.01157*, 2020.

S. Lee and E. Bareinboim. Causal identification with matrix equations. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

S. Lee, J. Correa, and E. Bareinboim. General identifiability with arbitrary surrogate experiments. In *Proceedings of the 35th Conference on Uncertainty in Artificial Intelligence*, Tel Aviv, Israel, 2019. AUAI Press.

Winston Lin. Agnostic notes on regression adjustments to experimental data: Reexamining freedman's critique. *The Annals of Applied Statistics*, 7(1):295–318, 2013.

Christos Louizos, Uri Shalit, Joris Mooij, David Sontag, Richard Zemel, and Max Welling. Causal effect inference with deep latent-variable models. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 6449–6459, 2017.

Tingjin Luo, Chenping Hou, Feiping Nie, and Dongyun Yi. Dimension reduction for non-gaussian data by adaptive discriminative analysis. *IEEE transactions on cybernetics*, 49(3):933–946, 2018.

Daniel Malinsky, Ilya Shpitser, and Thomas Richardson. A potential outcomes calculus for identifying conditional path-specific effects. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 3080–3088. PMLR, 2019.

Razieh Nabi, Rohit Bhattacharya, and Ilya Shpitser. Full law identification in graphical models of missing data: Completeness results. In *International Conference on Machine Learning*, pp. 7153–7163. PMLR, 2020.

Jersey Neyman. Sur les applications de la théorie des probabilités aux experiences agricoles: Essai des principes. *Roczniki Nauk Rolniczych*, 10:1–51, 1923.

Lizhen Nie, Mao Ye, Qiang Liu, and Dan Nicolae. Vcnet and functional targeted regularization for learning causal effects of continuous treatments. *arXiv preprint arXiv:2103.07861*, 2021.

Judea Pearl. [bayesian analysis in expert systems]: comment: graphical models, causality and intervention. *Statistical Science*, 8(3):266–269, 1993.

Judea Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 1995.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Karl Pearson. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, 2(11):559–572, 1901.

Joseph D Ramsey, Kun Zhang, Madelyn Glymour, Ruben Sanchez Romero, Biwei Huang, Imme Ebert-Uphoff, Savini Samarasinghe, Elizabeth A Barnes, and Clark Glymour. Tetrad—a toolbox for causal discovery. In *8th International Workshop on Climate Informatics*, 2018.

Thomas S Richardson and James M Robins. Single world intervention graphs (swigs): A unification of the counterfactual and graphical approaches to causality. *Center for the Statistics and the Social Sciences, University of Washington Series. Working Paper*, 128(30):2013, 2013.

P R Rosenbaum and D B Rubin. Propensity scores in the design of observational studies for causal effects. *Biometrika*, 09 2022. ISSN 1464-3510. doi: 10.1093/biomet/asac054. URL `https://doi.org/10.1093/biomet/asac054`. asac054.

Paul R Rosenbaum. Sensitivity analysis for certain permutation inferences in matched observational studies. *Biometrika*, 74(1):13–26, 1987.

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

Paul R Rosenbaum and Donald B Rubin. Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American statistical Association*, 79(387):516–524, 1984.

Donald B Rubin. Randomization analysis of experimental data: The fisher randomization test comment. *Journal of the American statistical association*, 75(371):591–593, 1980.

Patrick Schwab, Lorenz Linhardt, Stefan Bauer, Joachim M Buhmann, and Walter Karlen. Learning counterfactual representations for estimating individual dose-response curves. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 5612–5619, 2020.

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

Amit Sharma, Emre Kiciman, et al. DoWhy: A Python package for causal inference. https://github.com/microsoft/dowhy, 2019.

Eli Sherman and Ilya Shpitser. Identification and estimation of causal effects from dependent data. *Advances in neural information processing systems*, 31, 2018.

Claudia Shi, David Blei, and Victor Veitch. Adapting neural networks for the estimation of treatment effects. *Advances in neural information processing systems*, 32, 2019.

Ilya Shpitser and Judea Pearl. Identification of joint interventional distributions in recursive semi-markovian causal models. In *Proceedings, The Twenty-First National Conference on Artificial Intelligence and the Eighteenth Innovative Applications of Artificial Intelligence Conference, July 16-20, 2006, Boston, Massachusetts, USA*, pp. 1219–1226. AAAI Press, 2006a. URL `http://www.aaai.org/Library/AAAI/2006/aaai06-191.php`.

Ilya Shpitser and Judea Pearl. Identification of conditional interventional distributions. In *UAI '06, Proceedings of the 22nd Conference in Uncertainty in Artificial Intelligence, Cambridge, MA, USA, July 13-16, 2006*. AUAI Press, 2006b.

Jeffrey A Smith and Petra E Todd. Does matching overcome lalonde's critique of nonexperimental estimators? *Journal of econometrics*, 125(1-2):305–353, 2005.

Minsun Song, Wei Hao, and John D Storey. Testing for genetic associations in arbitrarily structured populations. *Nature genetics*, 47(5):550–554, 2015.

Jerzy Splawa-Neyman, Dorota M Dabrowska, and TP Speed. On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science*, pp. 465–472, 1990.

Ami Tavory. Determining principal component cardinality through the principle of minimum description length. In *International Conference on Machine Learning, Optimization, and Data Science*, pp. 655–666. Springer, 2019.

Johannes Textor, Benito van der Zander, Mark S Gilthorpe, Maciej Liśkiewicz, and George TH Ellison. Robust causal inference using directed acyclic graphs: the r package 'dagitty'. *International journal of epidemiology*, 45(6):1887–1894, 2016.

Jin Tian and Judea Pearl. A general identification condition for causal effects. In Rina Dechter, Michael J. Kearns, and Richard S. Sutton (eds.), *Proceedings of the Eighteenth National Conference on Artificial Intelligence and Fourteenth Conference on Innovative Applications of Artificial Intelligence, July 28 - August 1, 2002, Edmonton, Alberta, Canada*, pp. 567–573. AAAI Press / The MIT Press, 2002. URL http://www.aaai.org/Library/AAAI/2002/aaai02-085.php.

Santtu Tikka and Juha Karvanen. Identifying causal effects with the r package causaleffect. *Journal of Statistical Software*, 76:1–30, 2017.

Santtu Tikka, Antti Hyttinen, and Juha Karvanen. Causal effect identification from multiple incomplete data sources: A general search-based approach. *Journal of Statistical Software*, 99:1–40, 2021.

Tianyu Wang, Marco Morucci, M Usaid Awan, Yameng Liu, Sudeepa Roy, Cynthia Rudin, Alexander Volfovsky, et al. Flame: A fast large-scale almost matching exactly approach to causal inference. *J. Mach. Learn. Res.*, 22:31–1, 2021.

Yixin Wang and David M Blei. The blessings of multiple causes. *Journal of the American Statistical Association*, 114(528):1574–1596, 2019.

K. Xia, K. Lee, E. Bengio, and E. Bareinboim. The causal-neural connection: Expressiveness, learnability, and inference. In *Advances in Neural Information Processing Systems*, volume 34, 2021a.

Kevin Xia, Kai-Zhan Lee, Yoshua Bengio, and Elias Bareinboim. The causal-neural connection: Expressiveness, learnability, and inference. *Advances in Neural Information Processing Systems*, 34:10823–10836, 2021b.

Yingcun Xia, Howell Tong, Wai Keung Li, and Li-Xing Zhu. An adaptive estimation of dimension reduction space. In *Exploration of A Nonlinear World: An Appreciation of Howell Tong's Contributions to Statistics*, pp. 299–346. World Scientific, 2009.

Mengyue Yang, Furui Liu, Zhitang Chen, Xinwei Shen, Jianye Hao, and Jun Wang. Causalvae: Disentangled representation learning via neural structural causal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9593–9602, 2021.

Liuyi Yao, Sheng Li, Yaliang Li, Mengdi Huai, Jing Gao, and Aidong Zhang. Representation learning for treatment effect estimation from observational data. *Advances in neural information processing systems*, 31, 2018.

Jinsung Yoon, James Jordon, and Mihaela Van Der Schaar. Ganite: Estimation of individualized treatment effects using generative adversarial nets. In *International conference on learning representations*, 2018.

Junzhe Zhang and Elias Bareinboim. Non-parametric methods for partial identification of causal effects. Technical report, Technical Report Technical Report R-72, Columbia University, Department of . . . , 2021.

Zhenyu Zhao and Totte Harinen. Uplift modeling for multiple treatments with cost optimization. In *2019 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pp. 422–431. IEEE, 2019.