Causal Inference and Representation Learning with Observational Data

by

Zhixuan Chu

(Under the Direction of Stephen L. Rathbun and Sheng Li)

Abstract

The dramatically growing availability of observational data is being witnessed in various domains of science and technology, which facilitates the study of causal inference. Compared with randomized controlled trials, causal inference from observational data has become an appealing research direction owing to a large amount of available data and low budget requirement for its collection. In particular, the success of representation learning inspires advanced methods for learning causal effects with observational data. However, some issues around the causal effect estimation are still challenging, such as missing counterfactual outcomes, treatment selection bias, lack of interpretability and explainability, inclusion of various covariate types, hidden confounders, difficulty of continual learning for incrementally available observational data, etc. This dissertation provides a comprehensive review of existing causal inference methods and proposes several novel approaches based on representation learning to solve these issues.

Index words: Treatment Effect Estimation, Deep Learning, Observational Data, Variable Selection, Continual Learning, Basket Trial, Treatment Selection Bias, Hidden Confounder

Causal Inference and Representation Learning with Observational Data

by

Zhixuan Chu

B.S., Huazhong University of Science and Technology, China, 2015

M.S., University of California, Riverside, US, 2016

A Dissertation Submitted to the Graduate Faculty of the

University of Georgia in Partial Fulfillment of the Requirements for the Degree.

Doctor of Philosophy

Athens, Georgia

2021

Causal Inference and Representation Learning with Observational Data

by

Zhixuan Chu

Major Professor:	Stephen L. Rathbun
	Sheng Li
Committee:	Ye Shen
	Hanwen Huang
	Kevin K. Dobbin

Electronic Version Approved:

Ron Walcott

Dean of the Graduate School

The University of Georgia

May 2021

Acknowledgments

Foremost, I would like to express my sincere gratitude to my advisors Dr. Rathbun and Dr. Li for their continuous support of my Ph.D study and research, for their patience, motivation, enthusiasm, and immense knowledge. Their guidance helped me throughout my research and the writing of this dissertation. Besides, my sincere thanks goes to my committee members, Dr. Shen, Dr. Huang, and Dr. Dobbin for their help on reviewing my work and giving valuable suggestions. Moreover, I would like to appreciate the faculty members and staff from the Department of Epidemiology and Biostatistics; they provided constant help during my years at The University of Georgia. Finally, I would like to thank my family for their support.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Causal Inference with Observational Data.

In everyday language, correlation and causality are commonly used interchangeably, although they have quite different interpretations. Correlation indicates a general relationship: two variables are correlated when they display an increasing or decreasing trend (Altman & Krzywinski, 2015). Causality is also referred to as cause and effect where the cause is partly responsible for the effect, and the effect is partly dependent on the cause. Causal inference is the process of drawing a conclusion about a causal connection based on the conditions of the occurrence of an effect. The main difference between causal inference and inference of correlation is that the former analyzes the response of the effect variable when the cause is changed (Pearl, 2009a; Stephen & Christopher, 2007).

In many cases, it seems obvious that one action can cause another; however, there exists also many cases that we cannot easily tease out and make sure the relationship. Therefore, learning causality is one dauntingly challenging problem. The most effective way of inferring causality is to conduct a randomized controlled trial, which randomly assigns participants into a treatment group or a control group. As the randomized study is conducted, the only

expected difference between the control and treatment groups is the outcome variable being studied. However, in reality, randomized controlled trials are always time-consuming and expensive, and thus the study cannot involve many subjects, which may be not representative of the real-world population a treatment/intervention would eventually target. Another issue is that the randomized controlled trials only focus on the average of samples, and it doesn't explain the mechanism or pertain for individual subjects. In addition, ethical issues also need to be considered in most of the randomized controlled trials, which largely limits its applications.

Therefore, instead of the randomized controlled trials, the observational data is a tempting shortcut. Observational data is obtained by the researcher simply observing the subjects without any interfering. That means, the researchers have no control over treatments and subjects, and they just observe the subjects and record data based on their observations. From the observational data, we can find their actions, outcomes, and information about what has occurred, but cannot figure out the mechanism why they took a specific action.

1.2 Causal Inference Approaches.

To solve these problems in causal inference from observational data, researchers develop various frameworks, including the potential outcome framework (Rubin, 1974; Splawa-Neyman et al., 1990) and the structural causal model (Pearl, 1995, 2009b, 2014). The potential outcome framework is also known as the Neyman-Rubin Potential Outcomes or the Rubin Causal Model. The potential outcome framework aims to estimate such potential outcomes and then calculate the treatment effect. Therefore, the treatment effect estimation is one of the central problems in causal inference under the potential outcome framework. Another influential framework in causal inference is the structural causal model (SCM), which includes the causal graph and the structural equations. The structural causal model describes the

causal mechanisms of a system where a set of variables and the causal relationship among them are modeled by a set of simultaneous structural equations. In recent years, the magnificent bloom of the machine learning area enhances the development of the causal inference area. Powerful machine learning methods such as decision tree, ensemble methods, deep neural network, are applied to estimate the potential outcome more accurately. In addition to the amelioration on the outcome estimation model, machine learning methods also provide a new aspect to handle the confounders. Benefiting from the recently representation learning methods, the confounder variables are adjusted by learning the balanced representation for all covariates, so that conditioning on the learned representation, the treatment assignment is independent of the confounder variables. In machine learning, the more data the better. However, in causal inference, the more data alone is not yet enough. Having more data only helps to get more precise estimates, but it cannot make sure these estimates are correct and unbiased.

1.3 Research Challenges of Causal Inference.

For the causal inference with observational data, the core question is how to get the counterfactual outcome. For example, we want to answer this question "would this patient have different results if he received a different medication?" Answering such counterfactual questions is challenging due to two reasons (Schwab et al., 2019): the first one is that we only observe the factual outcome and never the counterfactual outcomes that would potentially have happened if they have chosen a different treatment option. The second one is that treatments are typically not assigned at random in observational data, which may lead the treated population differs significantly from the general population. Therefore, missing counterfactual outcomes and treatment selection bias are two major challenges of causal inference.

Recent causal effect estimation methods (F. Johansson et al., 2016; S. Li & Fu, 2017a; Shalit et al., 2017) have built a strong connection with domain adaptation, by enforcing domain invariance with distributional distances such as the Wasserstein distance and maximum mean discrepancy. Inspired by metric learning, some methods (Yao et al., 2018) use hard samples to learn representations that preserve local similarity information and balance the data distributions. In Y. Zhang et al., 2020, the authors argue that distribution invariance is often too strict a requirement, and they propose to use counterfactual variance to measure the domain overlap. Thus, which is the best measurement for the imbalanced domains remains unsettled and the choice highly relies on the characteristics of the domain distributions (Yao et al., 2020). Besides, despite the empirical success of such methods, enforcing balance can, to various extents, remove predictive information and lead to a loss in predictive power, regardless of which type of domain divergence metric is employed (A. Alaa & Schaar, 2018).

Another question is which covariates should be included in the causal inference model. The most important assumption in causal inference is *ignorability*, under which treatment assignment is independent of the potential outcomes given the observed covariates. It is also known as the "no unmeasured confounders" assumption, which means all of the confounders should be measured and included in the analysis. Although including all of the confounders is important, this does not mean that including more variables is better (Z. Chu et al., 2020; Greenland, 2008; Patrick et al., 2011; Schisterman et al., 2009; Shortreed & Ertefaie, 2017). For example, conditioning on an *instrumental* variable that is associated with the treatment assignment but not with the outcome except through exposure can increase both bias and variance of estimated treatment effects (Myers et al., 2011). Conditioning on an *adjustment* variable, which is predictive of outcomes but not associated with treatment assignment, is unnecessary to remove bias, but can reduce variance in estimated treatment effects (Sauer et al., 2013). Therefore, conditioning on these variables, let alone *spurious* variables that are

not associated with treatment assignment and outcomes, may introduce more impalpable bias into model, especially in scenarios with high dimensional variables.

Due to the fact that identifying and collecting all of the confounders is impossible in practice, as well as the existence of hidden confounders, the strong ignorability assumption is usually untenable. By leveraging big data, it becomes possible to find a proxy for the hidden confounders. Network information, which serves as an efficient structured representation of non-regular data, is ubiquitous in the real world. Advanced by the powerful representation capabilities of various graph neural networks, networked data has recently received increasing attention (Kipf & Welling, 2016; Velickovic et al., 2019; Veličković et al., 2017). Besides, it can be used to help recognize the patterns of hidden confounders.

Besides, the existing methods only focus on source-specific and stationary observational data. Such learning strategies assume that all observational data are already available during the training phase and from the only one source. This assumption is unsubstantial in practice due to two reasons. The first one is based on the characteristics of observational data, which are incrementally available from non-stationary data distributions. For instance, the number of electronic medical records in one hospital is growing every day, or the electronic medical records for one disease may be from different hospitals or even different countries. This characteristic implies that one cannot have access to all observational data at one time point and from one single source. The second reason is based on the realistic consideration of accessibility. For example, when the new observational are available, if we want to refine the model previously trained by original data, maybe the original training data are no longer accessible due to a variety of reasons, e.g., legacy data may be unrecorded, proprietary, too large to store, or subject to privacy constraint (J. Zhang et al., 2020). This practical concern of accessibility is ubiquitous in various academic and industrial applications. That's what it boiled down to: in the era of big data, we face the new challenges in causal inference with observational data: the *extensibility* for incrementally available observational data,

the *adaptability* for extra domain adaptation problem except for the imbalance between treatment and control groups in one source, and the *accessibility* for a huge amount of data.

1.4 The Organization of Dissertation.

The dissertation is organized as follows. In Chapter 2, we define the notations and assumptions under the potential outcome framework. In Chapter 3, we give one comprehensive review of the existing causal inference methods. Then, aiming to solve the above mentioned research challenges of causal inference, our proposed methods and experiments are presented in the following chapters. We close with a conclusion in the last Chapter.

Chapter 2

Background

2.1 Definitions

Here we define the notations under the potential outcome framework (Rubin, 1974; Splawa-Neyman et al., 1990), which is logically equivalent to another framework, the structural causal model framework (Judea Pearl, 2012). The foundation of potential outcome framework is that the causality is tied to treatment (or action, manipulation, intervention), applied to a unit (G. W. Imbens & Rubin, 2015a). The treatment effect is obtained by comparing units' potential outcomes of treatments. In the following, we first introduce three essential concepts in causal inference: unit, treatment, and outcome.

Definition 1. *Unit. A unit is the atomic research object in the treatment effect study.*

A unit can be a physical object, a firm, a patient, an individual person, or a collection of objects or persons, such as a classroom or a market, at a particular time point (G. W. Imbens & Rubin, 2015a). Under the potential outcome framework, the atomic research objects at different time points are different units. One unit in the dataset is a sample of the whole population, so in this survey, the term "sample" and "unit" are used interchangeably.

Definition 2. *Treatment. Treatment refers to the action that applies (exposes, or subjects) to a unit.*

Let T ($T \in \{0, 1, 2, \ldots, N_T\}$) denote the treatment, where $N_T + 1$ is the total number of possible treatments. In the aforementioned medicine example, Medicine A is a treatment. Most of the literatures consider the binary treatment, and in this case, the group of units applied with treatment $T = 1$ is the *treated group*, and the group of units with $T = 0$ is the *control group*.

Definition 3. *Potential outcome. For each unit-treatment pair, the outcome of that treatment when applied on that unit is the potential outcome (G. W. Imbens & Rubin, 2015a).*

The potential outcome of treatment with value t is denoted as $Y(T = t)$.

Definition 4. *Observed outcome. The observed outcome is the outcome of the treatment that is actually applied.*

The observed outcome is also called factual outcome, and we use Y^F to denote it where F stands for "factual". The relation between the potential outcome and the observed outcome is: $Y^F = Y(T = t)$ where t is the treatment actually applied.

Definition 5. *Counterfactual outcome: Counterfactual outcome is the outcome if the unit had taken another treatment.*

The counterfactual outcomes are the potential outcomes of the treatments except the one actually taken by the unit. Since a unit can only take one treatment, only one potential outcome can be observed, and the remaining unobserved potential outcomes are the counterfactual outcome. In the multiple treatment case, let $Y^{CF}(T = t^{'})$ denote the counterfactual outcome of treatment with value $t^{'}$. In the binary treatment case, for notation simplicity, we use Y^{CF} to denote the counterfactual outcome, and $Y^{CF} = Y(T = 1 - t)$, where t is the treatment actually taken by the unit.

In the observational data, besides the chosen treatments and the observed outcome, the units' other information is also recorded, and they can be separated as pre-treatment variables and the post-treatment variables.

Definition 6. *Pre-treatment variables: Pre-treatment variables are the variables that are not affected by the treatment.*

Pre-treatment variables are also named as *background variables*, and they can be patients' demographics, medical history, and etc. Let X denote the pre-treatment variables.

Definition 7. *Post-treatment variables: The post-treatment variables are the variables that are affected by the treatment.*

One example of post-treatment variables is the intermediate outcome, such as the lab test after taking the medicine in the aforementioned medicine example.

Treatment Effect. After introducing the observational data and the key terminologies, the treatment effect can be quantitatively defined using the above definitions. The treatment effect can be measured at the population, treated group, subgroup, and individual levels. To make these definitions clear, here we define the treatment effect under binary treatment, and it can be extended to multiple treatments by comparing their the potential outcomes.

At the population level, the treatment effect is named as the Average Treatment Effect (ATE), which is defined as:

$$\text{ATE} = \mathbb{E}[\mathbf{Y}(T=1) - \mathbf{Y}(T=0)], \tag{2.1.1}$$

where $\mathbf{Y}(T=1)$ and $\mathbf{Y}(T=0)$ are the potential treated and control outcome of the whole population respectively.

For the treated group, the treatment effect is named as Average Treatment effect on the Treated group (ATT), and it is defined as:

$$\text{ATT} = \mathbb{E}[\mathbf{Y}(T=1)|T=1] - \mathbb{E}[\mathbf{Y}(T=0)|T=1], \tag{2.1.2}$$

where $\mathbf{Y}(T=1)|T=1$ and $\mathbf{Y}(T=0)|T=1$ are the potential treated and control outcome of the treated group respectively.

At the subgroup level, the treatment effect is called Conditional Average Treatment Effect (CATE), which is defined as:

$$\text{CATE} = \mathbb{E}[\mathbf{Y}(T=1)|X=x] - \mathbb{E}[\mathbf{Y}(T=0)|X=x], \tag{2.1.3}$$

where $\mathbf{Y}(T=1)|X=x$ and $\mathbf{Y}(T=0)|X=x$ are the potential treated and control outcome of the subgroup with $X=x$, respectively. CATE is a common treatment effect measurement under the case where the treatment effect varies across different subgroups, which is also known as the heterogeneous treatment effect.

At the individual level, the treatment effect is called Individual Treatment Effect (ITE), and the ITE of unit i is defined as:

$$\text{ITE}_i = Y_i(T=1) - Y_i(T=0), \tag{2.1.4}$$

where $Y_i(T=1)$ and $Y_i(T=0)$ are the potential treated and control outcome of unit i respectively. In some literatures (F. Johansson et al., 2016; Shalit et al., 2017), the ITE is viewed equivalent to the CATE.

Objective. For causal inference, our objective is to estimate the treatment effects from the observational data. Formally speaking, given the observational dataset, $\left\{X_i, T_i, Y_i^F\right\}_{i=1}^N$,

where N is the total number of units in the datasets, the goal of the causal inference task is to estimate the treatment effects defined above.

2.2 Assumptions

In order to estimate the treatment effect, the following assumptions are commonly used in the causal inference literature.

Assumption 2.2.1. ***Stable Unit Treatment Value Assumption (SUTVA)****. The potential outcomes for any unit do not vary with the treatment assigned to other units, and, for each unit, there are no different forms or versions of each treatment level, which lead to different potential outcomes.*

This assumption emphasizes two points: The first point is the independence of each unit, that is, there are no interactions between units. In the context of the above illustrative example, one patient's outcome will not affect other patients' outcomes.

The second point is the single version for each treatment. In the above example, Medicine A with different dosages are different treatments under the SUTVA assumption.

Assumption 2.2.2. ***Ignorability.*** *Given the background variable, X, treatment assignment T is independent to the potential outcomes, i.e., $T \perp\!\!\!\perp Y(T=0), Y(T=1)|X$.*

In the context of the illustrative example, this ignorability assumption indicates two folds: First, if two patients have the same background variable X, their potential outcomes should be the same whatever the treatment assignment is, i.e., $p(Y_i(0), Y_i(1)|X = x, T = T_i) = p(Y_j(0), Y_j(1)|X = x, T = T_j)$. Analogously, if two patients have the same background variable value, their treatment assignment mechanism should be same whatever the value of potential outcomes they have, i.e., $p(T|X = x, Y_i(0), Y_i(1)) = p(T|X = x, Y_j(0), Y_j(1))$. The ignorability assumption is also named as unconfoundedness assumption. With this

unconfoundedness assumption, for the units with the same background variable X, their treatment assignment can be viewed as random.

Assumption 2.2.3. ***Positivity***. *For any value of X, treatment assignment is not deterministic:*

$$P(T = t|X = x) > 0, \quad \forall\, t \text{ and } x. \tag{2.2.1}$$

If for some values of X, the treatment assignment is deterministic; then for these values, the outcomes of at least one treatment could never be observed. In this case, it would be unable and meaningless to estimate the treatment effect. To be more specific, suppose there are two treatments: Medicine A and Medicine B. Let's assume that patients with age greater than 60 are always assigned with medicine A, then it will be unable and meaningless to study the outcome of medicine B on those patients. In other words, the positivity assumption indicates the variability, which is important for treatment effect estimation. In (G. W. Imbens & Rubin, 2015a), the ignorability and the positivity assumptions together are called *Strong Ignorability* or *Strongly Ignorable Treatment Assignment*.

With these assumptions, the relationship between the observed outcome and the potential outcome can be rewritten as:

$$\begin{aligned} \mathbb{E}[Y(T = t)|X = x] &= \mathbb{E}[Y(T = t)|T = t, X = x] \text{ (Ignorability)} \\ &= \mathbb{E}[Y^F|T = t, X = x], \end{aligned} \tag{2.2.2}$$

where Y^F is the random variable of the observed outcome, and $Y(T = t)$ is the random variable of the potential outcome of treatment t. If we are interested in the potential outcome of one specific group (either the subgroup, the treated group, or the whole population), the potential outcome can be obtained by taking expectation of the observed outcome over that group.

With the above equation, we can rewrite the treatment effect defined in Section 2.1 as follows:

$$
\begin{aligned}
\text{ITE}_i &= T_i Y_i^F - T_i Y_i^{CF} + (1 - T_i) Y_i^{CF} - (1 - T_i) Y_i^F \\
\text{ATE} &= \mathbb{E}_X \left[\mathbb{E}[Y^F | T = 1, X = x] - \mathbb{E}[Y^F | T = 0, X = x] \right] \\
&= \frac{1}{N} \sum_i \left(Y_i(T = 1) - Y_i(T = 0) \right) = \frac{1}{N} \sum_i \text{ITE}_i \\
\text{ATT} &= \mathbb{E}_{\mathcal{X}_T} \left[\mathbb{E}[Y^F | T = 1, X = x] - \mathbb{E}[Y^F | T = 0, X = x] \right] \\
&= \frac{1}{N_t reat} \sum_{\{i:T_i=1\}} \left(Y_i(T = 1) - Y_i(T = 0) \right) = \frac{1}{N_t reat} \sum_{\{i:T_i=1\}} \text{ITE}_i \\
\text{CATE} &= \mathbb{E}[Y^F | T = 1, X = x] - \mathbb{E}[Y^F | T = 0, X = x] \\
&= \frac{1}{N_x} \sum_{\{i:X_i=x\}} \left(Y_i(T = 1) - Y_i(T = 0) \right) = \frac{1}{N_x} \sum_{\{i:X_i=x\}} \text{ITE}_i
\end{aligned}
\tag{2.2.3}
$$

where $Y_i(T = 1)$ and $Y_i(T = 0)$ are the potential treated/control outcomes of unit i, N is the total number of units in the whole population, $N_t reat$ is the number of units in the treated group, and N_x is the number of units in the group with $X = x$. The second line in the ATE, ATT and CATE equations are their empirical estimations. Empirically, the ATE can be estimated as the average of ITE on the entire population. Similarly, ATT and CATE can be estimated as the average of ITE on the treated group and specific subgroup separately.

Chapter 3

Causal Inference Methods

In this chapter, we introduce existing causal inference methods. We divide these methods into the following categories: (1) Re-weighting methods; (2) Stratification methods; (3) Matching methods; (4) Tree-based methods; (5) Representation based methods.

3.1 Re-weighting Methods

Due to the existence of confounders, the covariate distributions of the treated group and control group are different, which leads to the *selection bias* problem. In other words, the treatment assignment is correlated with covariates in the observational data. Sample re-weighting is an effective approach to overcome the selection bias. By assigning appropriate weight to each unit in the observational data, a pseudo-population can be created on which the distributions of the treated group and control group are similar.

In sample re-weighting methods, a key concept is *balancing score*. Balancing score $b(x)$ is a general weighting score, which is the function of x satisfying: $T \perp\!\!\!\perp x|b(x)$ (G. W. Imbens & Rubin, 2015a), where T is the treatment assignment and x is the background variables. There are various designs of the balancing score, and apparently, the most trivial design of

balancing score is $b(x) = x$ due to the ignorability assumption. Besides, propensity score is also a special case of balancing score.

Definition 8. *Propensity score: The propensity score is defined as the conditional probability of treatment given background variables (Rosenbaum & Rubin, 1983):*

$$e(x) = Pr(T = 1|X = x) \tag{3.1.1}$$

In detail, a propensity score indicates the probability of a unit being assigned to a particular treatment given a set of observed covariates. Balancing scores that incorporate propensity score are the most common approach.

Propensity score based sample re-weighting

Propensity scores can be used to reduce selection bias by equating groups based on these covariates. Inverse propensity weighting (IPW) (Rosenbaum, 1987a; Rosenbaum & Rubin, 1983), also named as inverse probability of treatment weighting (IPTW), assigns a weight r to each sample:

$$r = \frac{T}{e(x)} + \frac{1-T}{1-e(x)}, \tag{3.1.2}$$

where T is the treatment assignment ($T = 1$ denotes being treated group; $T = 0$ denotes the control group), and $e(x)$ is the propensity score defined in Eqn. (3.1.1).

After re-weighting, the IPW estimator of average treatment effect (ATE) is:

$$\hat{\text{ATE}}_{IPW} = \frac{1}{n}\sum_{i=1}^{n}\frac{T_i Y_i^F}{\hat{e}(x_i)} - \frac{1}{n}\sum_{i=1}^{n}\frac{(1-T_i)Y_i^F}{1-\hat{e}(x_i)}, \tag{3.1.3}$$

and its normalized version, which is preferred especially when the propensity scores are obtained by estimation (G. W. Imbens, 2004a):

$$\hat{\text{ATE}}_{IPW} = \sum_{i=1}^{n} \frac{T_i Y_i^F}{\hat{e}(x_i)} \Big/ \sum_{i=1}^{n} \frac{T_i}{\hat{e}(x_i)} - \sum_{i=1}^{n} \frac{(1-T_i) Y_i^F}{1-\hat{e}(x_i)} \Big/ \sum_{i=1}^{n} \frac{(1-T_i)}{1-\hat{e}(x_i)}. \tag{3.1.4}$$

Both large and small sample theory show that adjustment for the scalar propensity score is enough to remove bias due to all observed covariates (Rosenbaum & Rubin, 1983). The propensity score can be used to balance the covariates in the treatment and control groups and therefore reduce the bias through matching, stratification (subclassification), regression adjustment, or some combination of all three. (D'Agostino Jr, 1998) discusses the use of propensity score to reduce the bias, which also provides examples and detailed discussions.

However, in practice, the correctness of the IPW estimator highly relies on the correctness of the propensity score estimation, and slightly misspecification of propensity scores would cause ATE estimation error dramatically (Imai & Ratkovic, 2014). To handle this dilemma, Doubly Robust estimator (DR) (J. M. Robins et al., 1994), also named as Augmented IPW (AIPW), is proposed. DR estimator combines the propensity score weighting with the outcome regression, so that the estimator is robust even when one of the propensity score or outcome regression is incorrect (but not both). In detail, the DR estimator is formalized as:

$$\begin{aligned}
\hat{\text{ATE}}_{DR} &= \frac{1}{n} \sum_{i=1}^{n} \left\{ \left[\frac{T_i Y_i^F}{\hat{e}(x_i)} - \frac{T_i - \hat{e}(x_i)}{\hat{e}(x_i)} \hat{m}(1, x_i) \right] - \left[\frac{(1-T_i) Y_i^F}{1-\hat{e}(x_i)} - \frac{T_i - \hat{e}(x_i)}{1-\hat{e}(x_i)} \hat{m}(0, x_i) \right] \right\} \\
&= \frac{1}{n} \sum_{i=1}^{n} \left\{ \hat{m}(1, x_i) + \frac{T_i (Y_i^F - \hat{m}(1, x_i))}{\hat{e}(x_i)} - \hat{m}(0, x_i) - \frac{(1-T_i)(Y_i^F - \hat{m}(0, x_i))}{1-\hat{e}(x_i)} \right\},
\end{aligned} \tag{3.1.5}$$

where $\hat{m}(1, x_i)$ and $\hat{m}(0, x_i)$ are the regression model estimations of treated and control outcomes. The DR estimator is consistent and therefore asymptotically unbiased, if either the propensity score is correct or the model correctly reflects the true relationship among

exposure and confounders with the outcome (Fan et al., 2016). In reality, one definitely cannot guarantee whether one model can accurately explain the relationship among variables. The combination of outcome regression with weighting by propensity score ensures that the estimators are robust to misspecification of one of these models (Bang & Robins, 2005; J. Robins et al., 2007; J. M. Robins et al., 1994; Scharfstein et al., 1999).

The DR estimator consults outcomes to make the IPW estimator robust when propensity score estimation is not correct. An alternative way is to improve the estimation of propensity scores. In the IPW estimator, propensity score serves as both the probability of being treated and the covariate balancing score, covariate balancing propensity score (CBPS) (Imai & Ratkovic, 2014) is proposed to exploit such dual characteristics. In particular, CBPS estimates propensity scores by solving the following problem:

$$\mathbb{E}\left[\frac{T_i\tilde{x_i}}{e(x_i;\beta)} - \frac{(1-T_i)\tilde{x_i}}{1-e(x_i;\beta)}\right] = 0, \tag{3.1.6}$$

where $\tilde{x_i} = f(x_i)$ is a predefined vector-valued measurable function of x_i. By solving the above problem, CBPS directly constructs the covariate balancing score from the estimated parametric propensity score, which increase the robustness to the misspecification of the propensity score model.

Another drawback of the original IPW estimator is that it might be unstable if the estimated propensity scores are small. If the probability of either treatment assignment is small, the logistic regression model can become unstable around the tails, causing the IPW to also be less stable. To overcome this issue, trimming is routinely employed as a regularization strategy, which eliminates the samples whose propensity scores are less than a pre-defined threshold (B. K. Lee et al., 2011). However, this approach is highly sensitive to the amount of trimming (Ma & Wang, 2010). Also, theoretical results in (Ma & Wang, 2010) show that the small probability of propensity scores and the trimming procedure may result in different

non-Gaussian asymptotic distribution of IPW estimator. Based on this observation, a two-way robustness IPW estimation algorithm is proposed in (Ma & Wang, 2010). This method combines subsampling with a local polynomial regression based trimming bias corrector, so that it is robust to both small propensity score and the large scale of trimming threshold. An alternative approach to overcome the instability of IPW under small propensity scores is to redesign the sample weight so that the weight is bounded. In (F. Li et al., 2018), the overlap weight is proposed, in which each unit's weight is proportional to the probability of that unit being assigned to the opposite group. In detail, the overlap weight $h(x)$ is defined as $h(x) \propto 1 - e(x)$, where $e(x)$ is the propensity score. The overlap weight is bounded within the interval $[0, 0.5]$, and thus it is less sensitive to extreme vale of propensity score. Recent theoretical results show that the overlap weight has the minimum asymptotic variance among all balancing weights (F. Li et al., 2018).

3.2 Stratification Methods

Stratification, also named as *subclassification* or *blocking* (G. W. Imbens & Rubin, 2015a), is a representative method to adjust the confounders. The idea of stratification is to adjust the bias that stems from the difference between the treated group and the control group by splitting the entire group into homogeneous subgroups (blocks). Ideally, in each subgroup, the treated group and the control group are similar under certain measurements over the covariates, therefore, the units in the same subgroup can be viewed as sampled from the data under randomized controlled trials. Based on the homogeneity of each subgroup, the treatment effect within each subgroup (i.e., CATE) can be calculated through the method developed on RCTs data. After having the CATE of each subgroup, the treatment effect over the interested group can be obtained by combining the CATEs of subgroups belonging to that group.

The key component of stratification methods is how to create the blocks and how to combine the created blocks. The equal frequency (Rosenbaum & Rubin, 1983) is a common strategy to create blocks. Equal-frequency approach split the block by the appearance probability, such as the propensity score, so that the covariates have the same appearance probability (i.e., the propensity score) in each subgroup (block). The ATE is estimated by weighted average of each block's CATE, with the weight as the fraction of the units in this block. However, this approach suffers from high variance due to the insufficient overlap between treated and control groups in the blocks whose propensity score is very high or low. To reduce the variance, in (Hullsiek & Louis, 2002), the blocks, which divided according to the propensity score, are re-weighted by the inverse variance of the block-specific treatment effect. Although this method reduces the variance of equal-frequency method, it unavoidably increases the estimation bias.

3.3 Matching Methods

As mentioned previously, missing counterfactuals and confounder bias are two major challenges in treatment effect estimation. Matching based approaches provide a way to estimate the counterfactual and, at the same time, reduce the estimation bias brought by the confounders. In general, the potential outcomes of the i-th unit estimated by matching are (Abadie et al., 2004):

$$\hat{Y}_i(0) = \begin{cases} Y_i & \text{if } T_i = 0, \\ \frac{1}{\#\mathcal{J}(i)} \sum_{l \in \mathcal{J}(i)} Y_l & \text{if } T_i = 1; \end{cases} \quad \hat{Y}_i(1) = \begin{cases} \frac{1}{\#\mathcal{J}(i)} \sum_{l \in \mathcal{J}(i)} Y_l & \text{if } T_i = 0, \\ Y_i & \text{if } T_i = 1; \end{cases} \tag{3.3.1}$$

where $\hat{Y}_i(0)$ and $\hat{Y}_i(1)$ are the estimated control and treated outcome, $\mathcal{J}(i)$ is the matched neighbors of unit i in the opposite treatment group (Austin, 2011).

The analysis of the matched sample can mimic that of an RCT: one can directly compare outcomes between the treated and control group within the matched sample. In the context of an RCT, one expects that, on average, the distribution of covariates will be similar between treated and control groups. Therefore, matching can be used to reduce or eliminate the effects of confounding when using observational data to estimate treatment effects (Austin, 2011).

3.3.1 Distance Metric

Various distances have been adopted to compare the closeness between units (Gu & Rosenbaum, 1993), such as the widely used Euclidean distance (Rubin, 1973) and Mahalanobis distance (Rubin & Thomas, 2000). Meanwhile, many matching methods develop their own distance metrics, which can be abstracted as: $D(\mathbf{x}_i, \mathbf{x}_j) = ||f(\mathbf{x}_i) - f(\mathbf{x}_j)||_2$. The existing distance metrics mainly vary in how to design the transformation function $f(\cdot)$.

Propensity score based transformation. Original covariates of units can be represented by propensity scores. As a result, the similarity between two units can be directly calculated as: $D(\mathbf{x}_i, \mathbf{x}_j) = |e_i - e_j|$, where e_i, and e_j are the propensity scores of $\mathbf{x}_i$ and $\mathbf{x}_j$, respectively. Later, the linear propensity score based distance metric is also proposed, which is defined as $D(\mathbf{x}_i, \mathbf{x}_j) = |\text{logit}(e_i) - \text{logit}(e_j)|$. This improved version is recommended since it can effectively reduce the bias (Stuart, 2010). Furthermore, the propensity score based distance metric can be combined with other existing distance metrics, which provides a fine-grained comparison. In (Rubin & Thomas, 2000), when the difference of two unit's propensity scores is within a certain range, they are further compared with other distances on some key covariates. Under this metric, the closeness of two units contains two criteria: they are relatively close under propensity score measure, and they particularly similar under the comparison of the key covariates (Stuart, 2010).

Other transformations. Propensity score only adopts the covariate information, while some other distance metrics are learned by utilizing both the covariates and the outcome information so that the transformed space can preserve more information. One representative metric is the prognosis score (Hansen, 2008), which is the estimated control outcome. The transformation function is represented as: $f(x) = \hat{Y}_c$. However, the performance of the prognosis score relies on modeling the relationship between the covariates and control outcomes. Moreover, the prognosis score only takes the control outcome into consideration and ignores the treated outcome. The Hilbert-Schmidt Independence Criterion based nearest neighbor matching (HSIC-NNM) proposed in (Chang & Dy, 2017) could overcome the drawbacks of prognosis score. HSIC-NNM learns two linear projections for control outcome estimation task and treated outcome estimation task separately. To fully explore the observed control/treated outcome information, the parameters of linear projection is learned by maximizing the nonlinear dependency between the projected subspace and the outcome: $M_w = \arg\max_{M_w} \text{HSIC}(\mathbf{X}_w M_w, Y_w^F) - \mathcal{R}(M_w)$, where $w = 0, 1$ represent the control group and treated group, respectively. $\mathbf{X}_w M_w$ is the transformed subspace with the transformation function as: $f(x) = xM_w$. Y_w^F is the observed control/treated outcome, and $\mathcal{R}$ is the regularization to avoid overfitting. The objective function ensures the learned transformation functions project the original covariates to an information subspace where similar units will have similar outcomes.

Compared with propensity score based distance metric that focuses on balancing, prognosis score and HSIC-NNM focus on embedding the relationship between the transformed space and the observed outcome. These two lines of methods have different advantages, and some recent work tries to integrate these advantages together. In (S. Li & Fu, 2017b), the balanced and nonlinear representation (BNR) is proposed to project the covariates into a balanced low-dimensional space. In detail, the parameters in the nonlinear transformation function is learned by jointly optimizing the following two objectives: (1) Maximizing the

differences of noncontiguous-class scatter and within-class scatter so that the units with the same outcome prediction shall have similar representations after transformation; (2) Minimizing the maximum mean discrepancy between the transformed control and outcome group in order to get the balanced space after transformation. A series of works that have similar objectives but vary in balancing regularization have been proposed, such as using the conditional generative adversarial network to ensure the transformation function blocks the treatment assignment information (C. Lee et al., 2018; Yao, Li, Li, Xue, et al., 2019).

The methods mentioned above adopt either one or two transformations for treated and control groups separately. Different from the existing method, Randomized Nearest Neighbor Matching (RNNM) (S. Li et al., 2016) adopts a number of random linear projections as the transformation function, and the treatment effects are obtained as the median treatment effect by nearest-neighbor matching in each transformed subspace. The theoretical motivation of this approach is the Johnson-Lindenstrauss (JL) lemma, which guarantees that the pairwise similarity information of the points in the high-dimensional space can be preserved through random linear projection. Powered by the JL lemma, RNNM ensembles the treatment effect estimation results of several linear random transformations.

3.3.2 Choosing a Matching Algorithm

After defining the similarity metric, the next step is to find the neighbors. In (Caliendo & Kopeinig, 2008), existing matching algorithms are divided into four essential approaches, including the nearest neighbor matching, caliper, stratification and kernel. The most straightforward matching estimator is nearest neighbour matching (NNM). In particular, a unit in the control group is chosen as the matching partner for a treated unit, so that they are closest based on a similarity score (e.g., propensity score). The NNM has several variants like NNM with replacement and NNM without replacement. Treated units are matched to one

control, called pair matching or 1-1 matching, or treated units are matched to two controls, called 1-2 matching, and so on. It's a trade-off to determine the number of neighbors, since a large number of neighbors may result in the treatment effect estimator with high bias but low variance, while small number results in low bias but high variance. It is known, however, that the optimal structure is a full matching in which a treated unit may have one or several controls or a control may have one or several treated units (Gu & Rosenbaum, 1993).

NNM may have bad matches if the closest partner is far away. One can set a tolerance level on the maximum propensity score distance (caliper) to avoid this problem. Hence, caliper matching is one form of imposing a common support condition.

The stratification matching is to partition the common support of the propensity score into a set of intervals and then to take the mean difference in outcomes between treated and control observations in order to calculate the impact within each interval. This method is also known as interval matching, blocking and subclassification (Rosenbaum & Rubin, 1985).

The matching algorithms discussed above have in common that only a few observations in the control group are used to create the counterfactual outcome of a treatment observation. Kernel matching (KM) and local linear matching (LLM) are nonparametric matching that use weighted averages of observations in the control group to create the counterfactual outcome. Thus, one major advantage of these approaches is the lower variance, because we use more information to create counterfactual outcome.

Here, we also want to introduce another matching method called Coarsened Exact Matching (CEM) proposed in (Iacus et al., 2012). Because either the 1-k matching or the full matching fails to consider the extrapolation region, where few or no reasonable matches exist in the other treatment group, CEM was proposed to handle this problem. CEM first coarsen the selected important covariate,i.e., discretization, and then perform exact matching on the coarsened covariates. For example, if the selected covariates are age (age > 50 is 1, and others are 0) and gender (female as 1, and male as 0). A female patient with age 50 in

the treated group is represented by the coarsen covariates as $(1, 1)$. She will only match the patients in the treated group with exactly the same coarsened covariates value. After exact matching, the whole data is separated into two subsets. In one subset, every unit has its exact matched neighbors and it is the opposite in the other subset which contains the units in the extrapolation region. The outcomes of units in the extrapolation region are estimated by the outcome prediction model trained on the matched subset. So far, the treatment effect on the two subsets can be estimated separately, and the final step is to combine treatment effect on the two subsets by weighted average.

We have provided several different matching algorithms, but the most important question is how we should select a perfect matching method. Asymptotically all matching methods should yield the same results as the sample size grows and they will become closer to comparing only exact matches (Smith, 2000). When we only have small samples size, this choice will be important (Heckman et al., 1998). There is one trade-off between bias and variance.

3.4 Tree-based Methods

Another popular method in causal inference is based on decision tree learning, which is one of the predictive modeling approaches. Decision tree is a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from data.

Tree models where the target variable is discrete are called classification trees with prediction error measured based on misclassification cost. In these tree structures, leaves represent class labels and branches represent conjunctions of features that lead to those class labels. Decision trees where the target variable is continuous are called regression trees with prediction error measured by the squared difference between the observed and predicted values. The term Classification And Regression Tree (CART) analysis is an umbrella term used to

refer to both of the above procedures (Breiman, 2017). In CART model, the data space is partitioned and a simple prediction model for each partition space is fitted, and therefore every partitioning can be represented graphically as a decision tree (Loh, 2011).

For estimating heterogeneity in causal effects, a data-driven approach (Athey & Imbens, 2016) based on CART is provided to partition the data into subpopulations that differ in the magnitude of their treatment effects. The valid confidence intervals can be created for treatment effects, even with many covariates relative to the sample size, and without "sparsity" assumptions. This approach is different from conventional CART in two aspects. First, it focuses on estimating conditional average treatment effects instead of directly predicting outcomes as in the conventional CART. Second, different samples are used for constructing the partition and estimating effects each subpopulation, which is referred to as the honest estimation. However, in conventional CART, the same samples are used for these two tasks.

In CART, a tree is built up until a splitting tolerance is reached. There is only one tree, and it is grown and pruned as needed. However, BART is an ensemble of trees, so it is more comparable to random forests. A Bayesian "sum-of-trees" model called Bayesian Additive Regression Trees (BART) is developed in (Chipman et al., 2007, 2010). Every tree in BART model is a weak learner, and it is constrained by a regularization prior. Information can be extracted from the posterior by a Bayesian backfitting MCMC algorithm. BART is a nonparametric Bayesian regression model, which uses dimensionally adaptive random basis elements. Let T be a binary tree which has a set of interior node decision rules and terminal nodes, and let $M = \{\mu_1, \mu_2, ..., \mu_B\}$ be parameters associated with each of the B terminal nodes for T. We use $g(x; T, M)$ to assign a $\mu_b \in M$ to input vector x. The sum-of-trees model can be expressed as:

$$Y = g(x; T_1, M_1) + g(x; T_2, M_2) + \cdots + g(x; T_m, M_m) + \varepsilon, \tag{3.4.1}$$

$$\varepsilon \sim N(0, \sigma^2), \tag{3.4.2}$$

BART has a couple of advantages. It is very easy to implement and only needs to plug in the outcome, treatment assignment, and confounding covariates. In addition, it doesn't require any information about how these variables are parametrically related, so that it requires less guess when fitting the model. Moreover, it can deal with a mass of predictors, yield coherent uncertainty intervals, and handle continuous treatment variables and missing data (Hill, 2011).

BART is proposed to estimate average causal effects. In fact, it can also be used to estimate individual-level causal effects. BART not only can easily identify the heterogeneous treatment effects, but also get more accurate estimates of average treatment effects compared to other methods like propensity score matching, propensity score weighting, and regression adjustment in the nonlinear simulation situations examined (Hill, 2011).

In most previous methods, the prior distribution over treatment effects is always induced indirectly, which is difficult to be attained. A flexible sum of regression trees (i.e., a forest) can address this issue by modeling a response variable as a function of a binary treatment indicator and a vector of control variables (Hahn et al., 2017). This approach interpolates between two extremes: entirely and separately modeling the conditional means of treatment and control, or only the treating treatment assignment as another covariate.

Random forest is a classifier consisting of a combination of tree predictors, in which each tree depends on a random vector that is independently sampled and has the identical distribution for all trees (Breiman, 2001). This model can also be extended to estimate heterogeneous treatment effects based on the Breiman's random forest algorithm (Wager & Athey, 2018a). Trees and forests can be considered as nearest neighbor methods with an adaptive neighborhood metric. Tree-based methods seek to find training examples that are close to a point x, but now closeness is defined with respect to a decision tree. And the

closest points to x are those that fall in the same leaf as it. The advantage of using trees is that their leaves can be narrower along the directions where the signal is changing fast and wider along the other directions, potentially leading to a substantial increase in power when the dimension of the feature space is even moderately large.

The tree-based framework also can be extended to uni- or multi-dimensional treatments (P. Wang et al., 2015). Each dimension can be discrete or continuous. A tree structure is used to specify the relationship between user characteristics and the corresponding treatment. This tree-based framework is robust to model misspecification and highly flexible with minimal manual tuning.

3.5 Representation Learning Methods

The most basic assumption used in statistical learning theory is that training data and test data are drawn from the same distribution. However, in most practical cases, the test data are drawn from a distribution that is only related, but not identical, to the distribution of the training data. In causal inference, this is also a big challenge. Unlike the randomized control trials, the mechanism of treatment assignment is not explicit in observational data. Therefore, interventions of interest are not independent of the property of the subjects. For example, in an observational study of the treatment effect of a medicine, the medicine is assigned to individuals based on several factors, including the known confounders and some unknown confounders. As a result, the counterfactual distribution will generally be different from the factual distribution. Thus, it is necessary to predict counterfactual outcomes by learning from the factual data, which converts the causal inference problem to a domain adaptation problem.

Extracting effective feature representations is critical for domain adaptation. A model (Ben-David et al., 2007) with a generalization bound is proposed to formalize this intu-

ition theoretically, which can not only explicitly minimize the difference between the source and target domains, but also maximize the margin of the training set. Building on this work (Ben-David et al., 2007), the discrepancy distance between distributions is tailored to adaptation problems with arbitrary loss functions (Mansour et al., 2009). In the following discussions, the discrepancy distance plays an important role in addressing the domain adaptation problem in causal inference.

So far, we can see a clear connection between counterfactual inference and domain adaptation. An intuitive idea is to enforce the similarity between the distributions of different treatment groups in the representation space. The learned representations trade-off three objectives: (1) low-error prediction over the factual representation, (2) low-error prediction over counterfactual outcomes by taking into account relevant factual outcomes, and (3) the distance between the distribution of treatment population and that of control population (F. Johansson et al., 2016). Following this motivation, (Shalit et al., 2017) gives a simple and intuitive generalization-error bound. It shows that the expected ITE estimation error of representation is bounded by a sum of the standard generalization-error of that representation and the distance between the treated and control distributions based on representation. Integral probability metric (IPM) is used to measure the distances between distributions, and explicit bounds are derived for the Wasserstein distance and Maximum Mean Discrepancy (MMD) distance. The goal is to find a representation $\Phi : X \to R$ and hypothesis $h : X \times \{0, 1\} \to Y$ that minimizes the following objective function:

$$\min_{h,\Phi} \frac{1}{n} \sum_{i=1}^{n} r_i \cdot L(h(\Phi(x_i), T_i), y_i) + \lambda \cdot R(h) + \alpha \cdot IPM_G(\{\Phi(x_i)\})_{i:T_i=0}, \{\Phi(x_i)\})_{i:T_i=1}), \quad (3.5.1)$$

where $r_i = \frac{T_i}{2u} + \frac{1-T_i}{2(1-u)}$, $u = \frac{1}{n}\sum_{i=1}^{n} T_i$, and the weight r_i compensates for the difference in treatment group size. R is a model complexity term. Given two probability density functions p, q defined over $S \subseteq R^d$, and a function family G of functions $g : S \to R$, the IPM is defined

as:

$$IPM_G(p,q) := \sup_{g \in G} |\int_S g(s)(p(s) - q(s))ds|. \tag{3.5.2}$$

This model allows for learning complex nonlinear representations and hypotheses with large flexibility. When the dimension of Φ is high, it risks losing the influence of t on h if the concatenation of Φ and T is treated as input. To address this problem, one approach is to parameterize $h_1(\Phi)$ and $h_0(\Phi)$ as two separate "heads" of the joint network. $h_1(\Phi)$ is used to estimate the outcome under treatment and $h_0(\Phi)$ is for the control group. Each sample is used to update only the head corresponding to the observed treatment. The advantage is that statistical power is shared in the common representation layers and the influence of treatment is retained in the separate heads (Shalit et al., 2017). This model can also be extended to any number of treatments, as described in the perfect match (PM) approach (Schwab et al., 2018). Following this idea, a few improved models have been proposed and discussed. For example, (F. D. Johansson et al., 2018) brings together shift-invariant representation learning and re-weighting methods. (Hassanpour & Greiner, 2019) presents a new context-aware weighting scheme based on the importance sampling technique, on top of representation learning, to alleviate the selection bias problem in ITE estimation.

Existing ITE estimation methods mainly focus on balancing the distributions of control and treated groups, but ignore the local similarity information that provides meaningful constraints on the ITE estimation. In (Yao et al., 2018, 2019), a local similarity preserved individual treatment effect (SITE) estimation method is proposed based on deep representation learning. SITE preserves local similarity and balances data distributions simultaneously. The framework of SITE contains five major components: representation network, triplet pairs selection, position-dependent deep metric (PDDM), middle point distance minimization (MPDM), and the outcome prediction network. To improve the model efficiency, SITE takes input units in a mini-batch fashion, and triplet pairs could be selected from every

mini-batch. The representation network learns latent embeddings for the input units. With the selected triplet pairs, PDDM and MPDM can preserve the local similarity information and meanwhile achieve the balanced distributions in the latent space. Finally, the embeddings of mini-batch are fed forward to a dichotomous outcome prediction network to get the potential outcomes. The loss function of SITE is as follows:

$$L = L_{FL} + \beta L_{PDDM} + \gamma L_{MPDM} + \lambda ||M||_2 \tag{3.5.3}$$

where L_{FL} is the factual loss between the estimated and observed factual outcomes. L_{PDDM} and L_{MPDM} are the loss functions for PDDM and MPDM, respectively. The last term is L_2 regularization on model parameters M (except the bias term).

Most models focus on covariates with numerical values, while how to handle covariates with textual information for treatment effect estimation is still an open question. One major challenge is how to filter out the nearly instrumental variables which are the variables more predictive to the treatment than the outcome. Conditioning on those variables to estimate the treatment effect would amplify the estimation bias. To address this challenge, a conditional treatment-adversarial learning based matching (CTAM) method is proposed in (Yao, Li, Li, Xue, et al., 2019). CTAM incorporates the treatment-adversarial learning to filter out the information related to nearly instrumental variables when learning the representations, and then it performs matching among the learned representations to estimate the treatment effects. The CTAM contains three major components: text processing, representation learning, and conditional treatment discriminator. Through the text processing component, the original text is transformed into vectorized representation S. After that, S is concatenated with the non-textual covariates X to construct a unified feature vector, which is then fed into the representation neural network to get the latent representation Z. After learning the representation, Z, together with potential outcomes Y, are fed into the conditional treatment

discriminator. During the training procedures, the representation learner plays a minimax game with the conditional treatment discriminator: By preventing the discriminator from assigning correct treatment, the representation learner can filter out the information related to nearly instrumental variables. The final matching procedure is performed in the representation space Z. The conditional treatment-adversarial learning helps reduce the bias of treatment effect estimation.

Chapter 4

Deep Adaptive Variable Selection Propensity Score

4.1 Introduction

The increasing availability of observational data requires the development of new and innovative statistical methods for causal inference. Instead of a randomized controlled trial, observational studies are tempting and less expensive shortcuts but draws into question what conclusions may be drawn when the researchers do not control treatment options. Due to the fact that the treatments are not assigned at random in observational studies, confounders that influence both the dependent variable and independent variable may cause a spurious association, making treatment effect estimation more difficult.

Various frameworks for causal inference have been proposed to obtain unbiased estimators for treatment effect from observational data (Hernán & Robins, 2006; Pearl, 2014; Rubin, 1974; Splawa-Neyman et al., 1990), many of which are based on the propensity score (Rosenbaum & Rubin, 1983). The propensity score is defined to be the conditional probability of assignment to a particular treatment given the observed covariates (Rosenbaum

& Rubin, 1983). Representative methods based on the propensity score include propensity score matching (Rosenbaum & Rubin, 1983, 1985), propensity score stratification (Cochran, 1968), inverse probability of treatment weighting (Rosenbaum, 1987b), and covariate adjustment via the propensity score (Harder et al., 2010). The common purpose of all of these methods is to reduce the selection bias between the treatment and control groups by controlling for the propensity score.

The calculation of the propensity score involves estimating the probability of treatment assignment conditional on the covariates. The essential question is which covariates should be included in the model. The most important assumption in causal inference is *ignorability*, under which treatment assignment is independent of the potential outcomes given the observed covariates. It is also known as the "no unmeasured confounders" assumption, which means all of the confounders should be measured and included in the analysis. Although including all of the confounders is important, this does not mean that including more variables is better (Z. Chu et al., 2020; Greenland, 2008; Patrick et al., 2011; Schisterman et al., 2009; Shortreed & Ertefaie, 2017). For example, conditioning on *instrumental* variables that are associated with the treatment assignment but not with the outcome except through exposure can increase both bias and variance of estimated treatment effects (Myers et al., 2011). Conditioning on *adjustment* variables that are predictive of outcomes but not associated with treatment assignment is unnecessary to remove bias, but can reduce variance in estimated treatment effects (Sauer et al., 2013). Therefore, conditioning on these variables, let alone *spurious* variables that are not associated with treatment assignment and outcomes, may introduce more impalpable bias into the model, especially in scenarios with high dimensional variables.

Much work (Z. Chu et al., 2020; Shortreed & Ertefaie, 2017; C. Wang et al., 2012; Wilson & Reich, 2014) has been proposed to solve the variable selection problem in causal inference. C. Wang et al., 2012 introduced a Bayesian adjustment for confounding method that

is a parametric variable selection approach. Wilson and Reich, 2014 proposed a decision-theoretic approach to confounder selection and treatment effect estimation. Z. Chu et al., 2020 proposed a feature selection representation matching method based on deep representation learning and matching. Shortreed and Ertefaie, 2017 proposed an outcome adaptive LASSO method that imposes penalties on variables that depend on the relationship between the predictors and outcome when estimating the propensity score.

Even if the confounders and adjustment variables are correctly specified, misspecification of the relationship between treatment assignment and those variables could also introduce bias in estimators of treatment effects. Estimation of the propensity score is typically based on logistic regression, under which the log odds of being assigned to a given treatment is assumed to be a linear function of the covariates. Compared with logistic regression, neural networks are potentially better at describing high dimensional data and estimating the propensity score (Bishop et al., 1995; Westreich et al., 2010). Without a priori decisions regarding the order of the polynomial and the number of interaction terms, deep neural networks can approximate complex polynomial and nonlinear functions of the data (Barron, 1994; Mhaskar, 1996). Due to the severe nonlinearity and unidentifiability of deep neural networks, there has been little work on the inferential properties of neural networks for causal inference (A. M. Alaa et al., 2017; S. Li & Fu, 2017a; Shalit et al., 2017; Yao et al., 2018; Yoon et al., 2018). For the case of normally-distributed outcomes, Dinh and Ho, 2020 have recently demonstrated that deep neural networks with square-error loss and adaptive group LASSO yield consistent estimators that are also consistent with respect to variable selection.

In this chapter, we propose a Deep Adaptive Variable Selection Propensity Score (DAVSPS) based on deep representation learning and outcome adaptive group LASSO. DAVSPS combines the data-driven learning capability of deep representation learning and variable selection consistency of adaptive group LASSO to improve the estimation of the propensity score by selecting out confounders and adjustment variables, while removing instrumental

and spurious variables. The latter extends the approach of Shortreed and Ertefaie, 2017 to deep neural networks. We also extend the results of Dinh and Ho, 2020 to joint modeling of binary outcomes and propensity score estimation.

We organize the rest of this chapter as follows. In Section 2, we provide a brief overview of deep neural networks, variable selection, and causal inference. Our proposed framework is presented in Section 3. In Section 4, we give the proofs about the consistency of outcome prediction estimator with group LASSO and consistency of estimator and variable selection in propensity score estimation with adaptive group LASSO. In Section 5, experiments on simulation data sets are provided. In Section 6, we apply our method to explore the racial disparities in severe maternal morbidity based on a National Inpatient Sample (NIS) data set. We close with our conclusions in Section 7.

4.2 Background

4.2.1 Deep Neural Networks

Linear and generalized linear models may not always adequately describe large data sets with high dimensional covariates, and complex nonlinear relationships among variables. Deep neural networks are one of the most efficient and effective tools for complex learning systems since they can approximate broad classes of continuous functions defined on bounded domains (Barron, 1994; Mhaskar, 1996). Throughout the chapter, we consider a general analytic neural network model. For example, consider predicting an output Y as a function of d_0 covariates $X_1, X_2, \cdots, X_{d_0}$, which we will collect in a vector X and $|X|$ is bounded for all $X \in \mathcal{X}$. As shown in Figure 4.1, a deep neural network is comprised of several layers of interconnected nodes. The first layer is the input layer and is taken to be comprised of the values of the covariates. The last layer is the output layer that produces the results for given

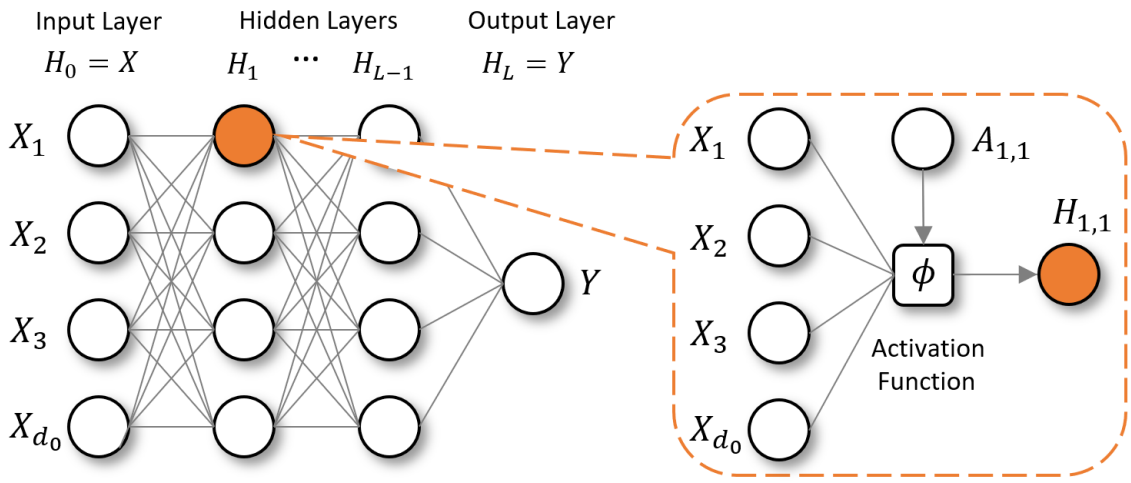

Figure 4.1: The structure of deep neural network.

inputs. The remaining layers are hidden layers. The k-th hidden layer H_k is comprised of d_k nodes, taking values $H_k = (H_{k,1}, \cdots, H_{k,d_k})^T$. We take $H_0 = X$ for notational convenience. The k-th hidden layer is defined as $H_k = \phi(B_k \cdot H_{k-1} + A_k), k = 1, \cdots, L-1$, where the activation function ϕ is analogous to the inverse link function in generalized linear models. Besides, we need the ϕ is analytic. The matrix $B_k \in \mathbf{R}^{d_k \times d_{k-1}}$ is comprised of unknown weights, analogous to the regression coefficients of a multivariate regression, and the $d_k \times 1$ vector A_k may be regarded as a vector of intercepts. In particular, because the deep neural network only interacts with the original covariates through the first hidden layer, the columns $\beta_1, \cdots, \beta_{d_0}$ of B_1 in the first hidden layer are comprised of vectors of parameters associated with input variables X_1, X_2, $\cdots$, and X_{d_0}, respectively. Thus, $||\beta_j||$ may be regarded as a measure of the impact of X_j on the outcome Y, where $j = 1, \cdots, d_0$. The last layer of deep neural networks is output layer H_L which produces the result for given inputs, i.e., $H_L = Y = f(X) = \phi(B_L \cdot H_{L-1} + A_L)$, where H_{L-1} is the last hidden layer and ϕ is analytic activation function that depends on the type of outcome variable. For example, if Y is a

continuous variable, ϕ can be set as constant 1; if it is binary variable, ϕ can be set as an analytic hyperbolic tangent function.

4.2.2 Variable Selection

Least absolute shrinkage and selection operator (LASSO) (Tibshirani, 1996) is a regression analysis method that performs both variable selection and regularization by minimizing the residual sum of squares subject to the regularizing constraint that the sum of the absolute value of the coefficients is less than a constant. Unlike ridge regression, it tends to set the coefficients of some covariates to exactly 0 under the constraint, thus providing a mechanism for model selection. Although the LASSO seems like a very viable procedure and has demonstrably good performance in variable selection in many applications, Meinshausen, Bühlmann, et al., 2006 showed that LASSO selection is consistent only if the underlying model satisfies some conditions; for example, in orthogonal designs, or given proper choice of λ_n when $d = 2$). Zou, 2006 proposed the necessary condition for the consistency of the LASSO selection and pointed out if the necessary condition is not satisfied, LASSO is inconsistent with respect to variable selection. To enjoy oracle properties, the adaptive LASSO (Zou, 2006) was proposed where adaptive weights are used for penalizing different coefficients in the ℓ_1 penalty. The oracle properties mean that the adaptive LASSO with a proper choice of λ_n has variable selection consistency (the nonzero coefficients are selected with probability tending to one) and asymptotic normality (nonzero components are estimated as if the sparse model were known a priori) (Fan & Li, 2001; Shortreed & Ertefaie, 2017; Zou, 2006).

Due to the structure of neural networks, each covariate in the input layer interacts with several nodes in the first hidden layer. Therefore, a covariate can be dropped only if all of the parameters connecting that covariate to all nodes in the first hidden layer are zeros, so

that the standard LASSO and adaptive LASSO are not applicable for neural networks. The Group LASSO may be used to address this concern in neural networks (Scardapane et al., 2017; Zhao et al., 2015; Zhu et al., 2016). The Group LASSO is used to impose sparsity on a group level, where the parameters associated with the same covariate are put in one group, and that all the parameters in the group are shrunk together towards zero.

4.2.3 Causal Inference

The task of causal inference is to estimate the unbiased treatment effect, a measure used to compare treatments (or interventions). At the population level, the treatment effect is named as the Average Treatment Effect (ATE), which is defined as:

$$\text{ATE} = \mathbb{E}[Y(T=1) - Y(T=0)], \tag{4.2.1}$$

where $Y(T=1)$ and $Y(T=0)$ are the potential treated and control outcome of the whole population respectively.

When estimating treatment effects from observational data, we face two major challenges (Yao et al., 2020), i.e., missing counterfactual outcomes and treatment selection bias. Firstly, in real life, we only observe the factual outcome and never all potential outcomes that would potentially have happened had we chosen other different treatment options. Secondly, unlike randomized controlled experiments, treatments are typically not assigned at random in observational data. Due to this treatment assignment bias, the treated population may differ significantly from the general population. These two major issues make treatment effects estimation very challenging.

To overcome these two challenges, *balancing score* is the key concept in most of the approaches. Balancing score $b(x)$ is a general weighting score, which is the function of x satisfying: $T \perp\!\!\!\perp x|b(x)$ (G. W. Imbens & Rubin, 2015a). Therefore, conditioned on

the balancing score, the treatment and control groups can be be directly compared for unbiased estimation of treatment effects (Rosenbaum & Rubin, 1983). Propensity score is a special case of balancing score. It is defined as the conditional probability of treatment given variables, i.e., $e(x) = P(T = 1|X = x)$ (Rosenbaum & Rubin, 1983).

The success of unbiased treatment effete estimation from observational data is based on the following assumptions (G. W. Imbens & Rubin, 2015a), which ensure that the treatment effect can be identified. **Stable Unit Treatment Value Assumption (SUTVA)**: The potential outcomes for any units do not vary with the treatments assigned to other units, and, for each unit, there are no different forms or versions of each treatment level, which lead to different potential outcomes. **Consistency**: The potential outcome of treatment T is equal to the observed outcome if the actual treatment received is T. **Positivity**: For any value of X, treatment assignment is not deterministic, i.e.,$P(T = t|X = x) > 0$, for all t and x. **Ignorability**: Given covariates X, treatment assignment T is independent to the potential outcomes, i.e., $(Y_1, Y_0) \perp\!\!\!\perp T|X$.

4.3 Deep Adaptive Variable Selection Propensity Score

In this section, we firstly present the definitions and notations involved in our model and then describe the proposed framework, i.e., Deep Adaptive Variable Selection Propensity Score (DAVSPS). Finally, we prove the consistency of estimator and variable selection in DAVSPS.

4.3.1 Preliminaries

Suppose that observational data are obtained from n sampling units and each unit received one of two or more treatments. Let t_i denote the treatment assignment for unit i; $i = 1, ..., n$. For binary treatments, $t_i = 1$ for the treatment group, and $t_i = 0$ for the control group. The

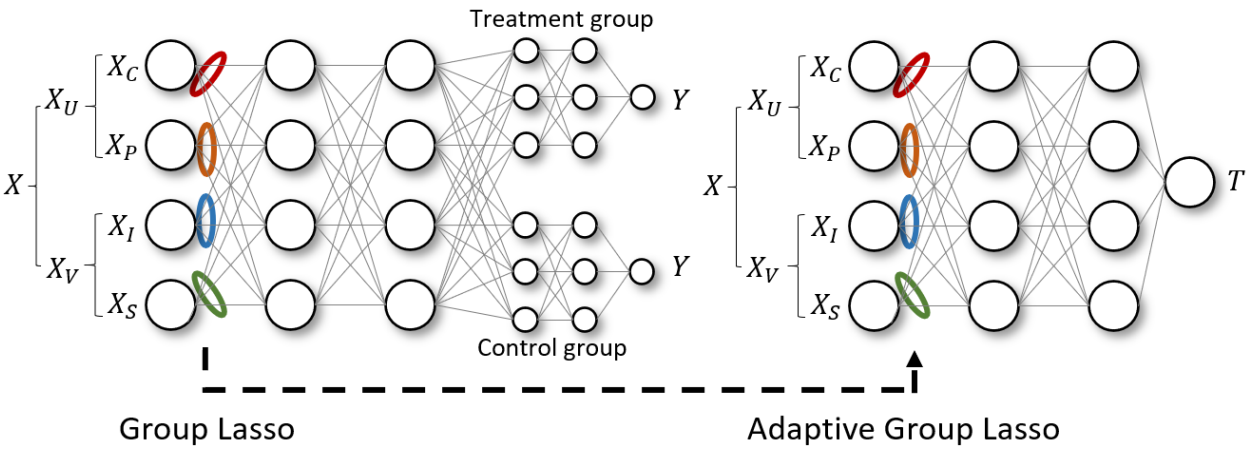

Figure 4.2: The framework of DAVSPS contains two major steps: outcome prediction with group LASSO and propensity score estimation with adaptive group LASSO.

observed outcome for unit i is denoted by Y_i. Let $X \in \mathbf{R}^{d_0}$ denote the vector of all observed variables. There are d_0 covariates, denoted X_j for $j = 1 : d_0$.

According to the different types of covariates, the covariates X can be decomposed into four subsets: $X = [X_{\mathcal{C}}{}^{\intercal}, X_{\mathcal{P}}{}^{\intercal}, X_{\mathcal{I}}{}^{\intercal}, X_{\mathcal{S}}{}^{\intercal}]^{\intercal}$, where $X_{\mathcal{C}}$ are confounders associated with both outcome and treatment assignment; $X_{\mathcal{P}}$ are adjustment variables that predict outcome, but not treatment assignment; $X_{\mathcal{I}}$ are instrumental variables that predict treatment assignment, not outcome; $X_{\mathcal{S}}$ are spurious variables that are not associated with both outcome and treatment assignment. Let $n_{\mathcal{C}}, n_{\mathcal{P}}, n_{\mathcal{I}}$, and $n_{\mathcal{S}}$ denote the cardinalities of $X_{\mathcal{C}}, X_{\mathcal{P}}, X_{\mathcal{I}}$, and $X_{\mathcal{S}}$, respectively, so $d_0 = n_{\mathcal{C}} + n_{\mathcal{P}} + n_{\mathcal{I}} + n_{\mathcal{S}}$. One objective of our analysis is to discern the unknown subset to which each covariate in X belongs.

4.3.2 The Framework of DAVSPS

We propose a deep adaptive variable selection propensity score (DAVSPS) based on deep representation learning and outcome adaptive group LASSO. The key idea of DAVSPS is to

combine the data-driven learning capability of deep representation learning (Bengio et al., 2013; Z. Chu et al., 2020; Shalit et al., 2017; Yao et al., 2018) and selection consistency of adaptive group LASSO (Dinh & Ho, 2020; Zou, 2006) to improve the estimation of propensity score by selecting out confounders and adjustment variables and removing instrumental and spurious variables. The framework of DAVSPS is illustrated in Figure 4.2, which contains two major steps: outcome prediction with group LASSO and propensity score estimation with adaptive group LASSO. More specifically, **Step one** is to use a deep neural network (DNN) based prediction model with group LASSO to predict the outcome and obtain the initial weight estimates for each covariate. **Step two** is to use a DNN based classification model to estimate propensity scores with adaptive group LASSO under which the weighted penalty is based on initial weight estimates (obtained from step one).

Outcome Prediction with Group LASSO. The first component of DAVSPS adopts a deep neural network with group LASSO to predict the outcome, and it is expressed as a nonlinear mapping $f : X \times T \to Y$, where X denotes the original covariates and Y denotes the observed factual outcome. Because the model only interacts with the original covariates through the first hidden layer, we only impose the group LASSO penalty in the first layer. In deep neural network, parameters that connect to the same input covariate are grouped together through the ℓ_2-norm, so that each group is associated with one covariate. Then, each group of parameters is penalized through the ℓ_1-norm. It can be treated as LASSO (ℓ_1 penalty) between groups and ridge (ℓ_2 penalty) within groups.

When the first hidden layer's parameters belonging to one group are shrunk together to zero, the corresponding covariate is removed from the original covariate space. Here, we aim to explore the relationship between the covariates and the outcome conditional on treatment assignment and expect to obtain the initial weight of each covariate, which can help the adaptive group LASSO in the deep classification layer to consistently distinguish the covari-

ates predictive of the outcome (i.e., confounder and adjustment variables) and covariates independent of the outcome (i.e., instrumental and spurious variables). Compared to the simple outcome prediction without penalization in the first step of adaptive LASSO (Shortreed & Ertefaie, 2017), adding penalization can provide more accurate initial estimation for parameters (Dinh & Ho, 2020).

The function $f : X \times T \to Y$ maps the observed covariates and treatment assignment to the corresponding observed outcome. However, when the dimensionality of the input covariates is high, there is a risk of losing the influence of the treatment T on $f : X \times T \to Y$, if the concatenation of X and T are both treated as inputs (Shalit et al., 2017). Besides, no matter which group the subjects belong to, they should share some common properties. Therefore, we first use one common neural network to train all the subjects together and then partition the common neural network into two sub-neural networks respectively corresponding to treatment and control groups:

$$f(x,t) = \begin{cases} f_1(x) & \text{if } t = 0 \\ f_2(x) & \text{if } t = 1. \end{cases} \tag{4.3.1}$$

Here, the first layer $f_1(x)$ is used to estimate the outcome under treatment and the second layer $f_2(x)$ is used to estimate the outcome for the control group. Each sample is only updated in one sub-network corresponding to the observed treatment. Obviously, this model can also be extended to any number of treatments. Let $\hat{y}_i = f_\beta(x)$ denote the predicted observed outcome of unit i corresponding to factual treatment t_i.

Based on the decomposition of X, the parameters in the first hidden layer of outcome prediction model that directly interact with X can also be decomposed into four subsets, i.e., $[\beta_\mathcal{C}, \beta_\mathcal{P}, \beta_\mathcal{I}, \beta_\mathcal{S}]$, where $\beta_\mathcal{C} \in \mathbf{R}^{d_1 \times n_\mathcal{C}}$, $\beta_\mathcal{P} \in \mathbf{R}^{d_1 \times n_\mathcal{P}}$, $\beta_\mathcal{I} \in \mathbf{R}^{d_1 \times n_\mathcal{I}}$, and $\beta_\mathcal{S} \in \mathbf{R}^{d_1 \times n_\mathcal{S}}$.

The estimator for outcome prediction with group LASSO is thus defined by:

$$\widehat{\beta}_n = \arg\min_{\beta} \left\{ \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f_\beta(x_i)) + \lambda_n q(\beta) \right\}, \tag{4.3.2}$$

where $\ell(y_i, f_\beta(x_i))$ denotes the log probability density (mass) function of y_i given $f_\beta(x)$. The penalty function is

$$q(\beta) = \sum_{c=1}^{n_\mathcal{C}} ||\beta_{c(\mathcal{C})}|| + \sum_{p=1}^{n_\mathcal{P}} ||\beta_{p(\mathcal{P})}|| + \sum_{i=1}^{n_\mathcal{I}} ||\beta_{i(\mathcal{I})}|| + \sum_{s=1}^{n_\mathcal{S}} ||\beta_{s(\mathcal{S})}||, \tag{4.3.3}$$

and the tuning parameter $\lambda_n > 0$ controls the trade-off between the outcome prediction and group LASSO. Here, $|| \cdot ||$ is the Euclidean norm. The parameters $\beta_{c(\mathcal{C})}$, $\beta_{p(\mathcal{P})}$, $\beta_{i(\mathcal{I})}$, and$\beta_{s(\mathcal{S})}$ respectively represent the vectors of parameters directly connecting to the c-th confounder $X_\mathcal{C}$, p-th adjustment $X_\mathcal{P}$, i-th instrumental $X_\mathcal{I}$, and s-th spurious variable $X_\mathcal{S}$.

Propensity Score Estimation with Adaptive Group LASSO. The propensity score is the conditional probability of assignment to a particular treatment given the observed covariates (Rosenbaum & Rubin, 1983). The estimation of propensity score for the analysis of the observational data is typically based on the logistic regression. Compared with the logistic regression, neural network is better at dealing with high dimensional data (Bishop et al., 1995; Westreich et al., 2010), without a priori decisions about the order of the polynomial and the number of interaction terms, it can approximate the complex polynomial function (Barron, 1994; Mhaskar, 1996). Therefore, the second component of DAVSPS adopts a deep neural network with adaptive group LASSO to estimate the propensity score, which is expressed as a nonlinear mapping $g : X \to T$.

Based on the discussion in Introduction, a propensity score estimation model should include the confounders $X_\mathcal{C}$ and adjustment variables $X_\mathcal{P}$, and at the same time eliminate instrumental variables $X_\mathcal{I}$ and spurious variables $X_\mathcal{S}$. The regular LASSO forces the coefficients to be equally penalized in the ℓ_1 penalty, regardless of the types of covariates (Zou,

2006), thus it cannot achieve our goal of including $X_{\mathcal{C}}$ and $X_{\mathcal{P}}$, while excluding $X_{\mathcal{I}}$ and $X_{\mathcal{S}}$. To design a penalty function with different regularization strengths according to different types of covariates, we apply the adaptive group LASSO with outcome prediction (Eq.(4.3.2)) as the base estimator into propensity score estimation model.

Based on the decomposition of X, the parameters in the first hidden layer of propensity score estimation model that directly interact with X can also be decomposed into four subsets, i.e., $[\alpha_{\mathcal{C}}, \alpha_{\mathcal{P}}, \alpha_{\mathcal{I}}, \alpha_{\mathcal{S}}]$, where $\alpha_{\mathcal{C}} \in \mathbf{R}^{d_1 \times n_{\mathcal{C}}}$, $\alpha_{\mathcal{P}} \in \mathbf{R}^{d_1 \times n_{\mathcal{P}}}$, $\alpha_{\mathcal{I}} \in \mathbf{R}^{d_1 \times n_{\mathcal{I}}}$, and $\alpha_{\mathcal{S}} \in \mathbf{R}^{d_1 \times n_{\mathcal{S}}}$.

The function g maps the covariates X_i to the corresponding observed treatment assignments T_i by deep neural network with hyperbolic tangent activation function. We assume $p(x_i; \alpha) = 0.5\{1 + \frac{g_\alpha(x_i)}{d}\}$ is the predicted probability that the unit i belongs to treatment group. Subjects with a very small probability close to zero or very large probability close to one can result in a very large weight in the following IPTW. Such weights can increase the variability of the estimated treatment effect (**cole2008constructing**). Thus, we utilize the scale parameter $d > 1$ to control the range of probability. Then, we use the log-likelihood of a Bernoulli to quantify the factual treatment prediction error and define the estimator of propensity score model with adaptive group LASSO by:

$$\widetilde{\alpha}_n = \arg\min_{\alpha} \left\{ \frac{1}{n} \sum_{i=1}^{n} \ell(t_i, p(x_i; \alpha)) + \theta_n q(\alpha) \right\}, \tag{4.3.4}$$

where

$$\ell(t_i, p(x_i; \alpha)) = -\Big(t_i \log(p(x_i; \alpha)) + (1 - t_i) \log(1 - p(x_i; \alpha)) \Big), \tag{4.3.5}$$

$$q(\alpha) = \sum_{c=1}^{n_{\mathcal{C}}} \frac{||\alpha_{c(\mathcal{C})}||}{||\widehat{\beta}_{c(\mathcal{C})}||^{\gamma}} + \sum_{p=1}^{n_{\mathcal{P}}} \frac{||\alpha_{p(\mathcal{P})}||}{||\widehat{\beta}_{p(\mathcal{P})}||^{\gamma}} + \sum_{i=1}^{n_{\mathcal{I}}} \frac{||\alpha_{i(\mathcal{I})}||}{||\widehat{\beta}_{i(\mathcal{I})}||^{\gamma}} + \sum_{s=1}^{n_{\mathcal{S}}} \frac{||\alpha_{s(\mathcal{S})}||}{||\widehat{\beta}_{s(\mathcal{S})}||^{\gamma}}, \tag{4.3.6}$$

and the tuning parameter $\theta_n > 0$ controls the trade-off between the treatment assignment classification and adaptive group LASSO. The power γ is positive. For the covariates removed in the outcome prediction with group LASSO (Eq.(4.3.2)), the $\widehat{\beta} = 0$, so we assume $0/0 = 1$ and the corresponding β in Eq.(4.3.6) will still converge to zero. Here, the adaptive group LASSO uses $\widehat{\beta}_{c(\mathcal{C})}$, $\widehat{\beta}_{p(\mathcal{P})}$, $\widehat{\beta}_{i(\mathcal{I})}$, and $\widehat{\beta}_{s(\mathcal{S})}$ to assign different weights to covariates based on their importance in predicting the outcome variable. In the outcome prediction model with group LASSO (Eq.(4.3.2)), the coefficients of confounders and adjustment variables that are predictive of outcome should be larger than those of instrumental and spurious variables that are not related to outcome. Thus, in the Eq.(4.3.6), the weights ($||\widehat{\beta}_{i(\mathcal{I})}||^{-\gamma}$ and $||\widehat{\beta}_{s(\mathcal{S})}||^{-\gamma}$) for instrumental and spurious variables are inflated to infinity while the weights ($||\widehat{\beta}_{c(\mathcal{C})}||^{-\gamma}$ and $||\widehat{\beta}_{p(\mathcal{P})}||^{-\gamma}$) for confounders and adjustment variables are bounded (Dinh & Ho, 2020; Shortreed & Ertefaie, 2017; Zou, 2006), which will help us to find out the ideal propensity score estimation model.

Treatment Effect Estimation based on DAVSPS. Although our Deep Adaptive Variable Selection Propensity Score is applicable to any methodology based on propensity score, such as matching (G. W. Imbens, 2004b; Rosenbaum & Rubin, 1983), stratification (Rosenbaum & Rubin, 1984), covariate adjustment (Garrido, 2016), and inverse probability reweighting (Hernán & Robins, 2006; Rosenbaum, 1987b), we only consider the inverse probability of treatment weighting as an example in this chapter. The propensity score estimated in Eq.(4.3.6) is defined as $\widehat{e}_i(x_i, \widetilde{\alpha}_n) = P(t_i = 1|x_i)$ and the balancing weight in IPTW estimator is defined by:

$$\widehat{\pi}_i(x_i, t_i) = \frac{t_i}{\widehat{e}_i(x_i, \widetilde{\alpha}_n)} + \frac{1 - t_i}{1 - \widehat{e}_i(x_i, \widetilde{\alpha}_n)}.$$

Although DAVSPS can be applied to individual treatment effect estimation, we only use the average treatment effect (ATE) to describe the treatment effect, denoted as $\tau = E(Y_1) - E(Y_0)$. Here the estimated ATE by IPTW is defined as:

$$\widehat{\tau} = \frac{\sum_{i=1}^{n} \widehat{\pi}_i(x_i, t_i) y_i t_i}{\sum_{i=1}^{n} \widehat{\pi}_i(x_i, t_i) t_i} - \frac{\sum_{i=1}^{n} \widehat{\pi}_i(x_i, t_i) y_i (1 - t_i)}{\sum_{i=1}^{n} \widehat{\pi}_i(x_i, t_i)(1 - t_i)}.$$

Tuning Parameter Selection Criterion. In the outcome prediction with group LASSO (Eq.(4.3.2)), the purpose is to explore the relationship between different types of covariates and the outcome, so we select the tuning parameter λ_n to optimize predictive performance. However, in the propensity score estimation with adaptive group LASSO (Eq.(4.3.6)), our goal is to use the propensity score to balance the selection bias between treatment and control groups, rather than create one classification model for the treatment assignment mechanism (Shortreed & Ertefaie, 2017). Therefore, minimizing the misclassification error is not our priority. We propose a new criterion to help select θ_n by minimizing the imbalance between treatment and control groups in the reweighted conditional distributions by propensity score. The reweighted conditional distributions are defined as:

$$q_1(X|T=1) = \frac{\widehat{\pi}(X,1)p(X|T=1)}{\int \widehat{\pi}(X,1)p(X|T=1)dX} \quad \text{and} \quad q_2(X|T=0) = \frac{\widehat{\pi}(X,0)p(X|T=0)}{\int \widehat{\pi}(X,0)p(X|T=0)dX}.$$

We aim to select θ_n so as to make the empirical reweighted conditional distributions $\widehat{q}_1(X|T=1)$ and $\widehat{q}_2(X|T=0)$ of the covariate space for the treatment and control groups more similar. In particular, we minimize the integral probability metric (IPM), a measure of the divergence between the reweighted conditional distributions of treatment and control groups. The integral probability metric is defined as (Müller, 1997):

$$\text{IPM}_G(\widehat{q}_1, \widehat{q}_2) = \sup_{\phi \in G} \left| \int_X \phi(X)(\widehat{q}_1(X) - \widehat{q}_2(X))dX \right|,$$

which measures the divergence between two distributions $\widehat{q}_1$ and $\widehat{q}_2$ by finding the maximal expected contrast in a function family G including the functions $\phi : X \to \mathbf{R}$. Here, we adopt the $\text{IPM}_G(\widehat{q}_1, \widehat{q}_2)$ defined in the family of 1-Lipschitz functions by setting the function family $G = \{\phi : \|\phi\|_L \leq 1\}$ to be the set of 1-Lipschitz functions, which leads to IPM being the Wasserstein distance (Shalit et al., 2017; Sriperumbudur et al., 2012).

4.4 Variable Selection Consistency of DAVSPS

For propensity score estimation in DAVSPS, the ideal model should include the confounders $X_{\mathcal{C}}$ and adjustment variables $X_{\mathcal{P}}$, while eliminating instrumental variables $X_{\mathcal{I}}$ and spurious variables $X_{\mathcal{S}}$ since the former are predictive of the outcomes, while the latter are not related to the outcomes. Therefore, we first build an outcome prediction model with group LASSO to provide an approximate estimation of the importance of each covariate. Then, in the propensity score estimation, based on the weights obtained from the relationship between covariates and outcomes, we design a regularization term with weights that impose heavier penalties on the variables that are not predictive of outcomes. In this section, we prove the consistency of outcome prediction estimator with group LASSO and consistency of estimator and variable selection in propensity score estimation with adaptive group LASSO.

Dinh and Ho, 2020 demonstrated that for a neural network of the form described in Figure 4.1 and with a continuous outcome, the adaptive group LASSO estimator is consistent and covariate selection consistent. We will extend these results to the classification problem that arises from binary outcome variables in the outcome prediction with group LASSO (step one) and propensity score estimation with adaptive group LASSO (step two). Therefore, regardless of the type of outcome (continuous or binary), the adaptive group LASSO estimator and variable selection of DAVSPS are consistent.

Suppose that the training data $\{(X_i, Y_i)\}_{i=1}^n$ are independent and identically distributed (i.i.d) samples generated from $P^*_{X,Y}$, where $X \in \mathcal{X}$ and $Y \in \{0, 1\}$. Assume further that conditional on covariates X_i, the outcomes Y_i are sampled from Bernoulli distributions with probabilities:

$$p(X_i; \beta) = 0.5\{1 + f_\beta(X_i)/d\}. \tag{4.4.1}$$

Here, the activation function in $f_\beta(X)$ is taken to be an analytic function in the deep neural network and is taken to have range of $(-1, 1)$. For example, it may be the hyperbolic tangent function, i.e., $\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$. We also assume that β belongs to the hypercube $\mathcal{W}$ and that $|X|$ is bounded for all $X \in \mathcal{X}$. The true value β^* for β is assumed to be in the interior of $\mathcal{W}$. The constant $d > 1$ controls the range of $p(X_i; \beta)$, so as to bound it away from zero and one; more specifically, $|p(X; \beta)| \leq 0.5(1 + 1/d)$ for all $\beta \in \mathcal{W}$ and all $X \in \mathcal{X}$. Therefore, the following negative log likelihood is bounded. Then the empirical risk (negative log likelihood) is:

$$R_n(\beta) = -\frac{1}{n} \sum_{i=1}^{n} \Big\{ Y_i \log p(X_i; \beta) + (1 - Y_i) \log(1 - p(X_i; \beta)) \Big\}, \tag{4.4.2}$$

and the risk is:

$$R(\beta) = -E_{P^*_{X,Y}} \Big\{ \{Y \log p(X; \beta) + (1 - Y) \log(1 - p(X; \beta))\} \Big\} \tag{4.4.3}$$

$$= -E_{P_X} \Big\{ p(X; \beta^*) \log p(X; \beta) + (1 - p(X; \beta^*)) \log(1 - p(X; \beta)) \Big\} \tag{4.4.4}$$

Because $\log(1 + x)$ is analytic when $|x| < 1$, and $|p(x; \beta)| < 0.5(1 + 1/d)$, Eqs. (4.4.2) and (4.4.4) are analytic. Therefore, the risk function $R(\beta)$ is also analytic.

To complete the proofs, we make the following assumptions:

Assumption 4.4.1. *The marginal distribution P_X of the covariates has bounded support.*

Assumption 4.4.2. *The outcome prediction model $f_{\beta^*}(x)$ only depends on the set of covariates that are predictive of the outcome, i.e., $X_{\mathcal{U}} \in \mathbf{R}^{n_{\mathcal{U}}} = X_{\mathcal{C}} \bigcup X_{\mathcal{P}}$ and , while being independent of the remaining covariates, i.e., $X_{\mathcal{V}} \in \mathbf{R}^{n_{\mathcal{V}}} = X_{\mathcal{I}} \bigcup X_{\mathcal{S}}$, where $n_{\mathcal{U}} = n_{\mathcal{C}} + n_{\mathcal{P}}$ and $n_{\mathcal{V}} = n_{\mathcal{I}} + n_{\mathcal{S}}$.*

The parameters in the first hidden layer for the outcome model are partitioned into two groups $u = \beta_{\mathcal{C}} \bigcup \beta_{\mathcal{P}}$ and $v = \beta_{\mathcal{I}} \bigcup \beta_{\mathcal{S}}$, where $u \in \mathbf{R}^{d_1 \times n_{\mathcal{U}}}$ and $v \in \mathbf{R}^{d_1 \times n_{\mathcal{V}}}$, and d_1 is the number of nodes in the first hidden layer. This assumption requires that the parameters v of the first hidden layer linked to covariates not predictive of the outcome are equal to zero, i.e., $v = 0$. Conversely, parameters u_j of the first hidden layer linked to each covariate j that is predictive of the outcome are not equal to zero; i.e., $||u_j|| > 0$.

A necessary and often implicit assumption for investigating the inferential properties of estimators is that the model parameters are identifiable. However, the parameters of deep neural network models are highly unidentifiable. Arbitrary permutations of the hidden nodes and arbitrary scaling under nonlinear activation functions can lead to identical outcomes, etc (Pourzanjani et al., 2017). For example, rearranging hidden nodes has no effect on the model output and these nodes are fully exchangeable because they are sum up together to obtain the output of the next layer. Arbitrarily multiplying all of the incoming weights and biases by a real scalar, a, and then multiplying the outcoming weights by $\frac{1}{a}$, will lead to the same output. Geometrically, it will lead to a high-dimensional hyperbola of equivalent output (Goodfellow et al., 2016). Moreover, in the context of variable selection, when all incoming or outcoming parameters are set to zero, changing the remaining parameters linked to that node has no effect on the outcome.

Following the approach of Dinh and Ho, 2020, we define the set of risk minimizers as:

$$\mathcal{H}^*_\beta := \{\beta : R(\beta) = R(\beta^*)\}.$$

Traditional statistical analyses are often based on Taylor expansion at a local optimum and non-singular Hessian matrix, but this does not work for deep neural networks since H^* may contain subsets of high dimension and the Hessian matrix at an optimum might be singular. Due to its severe unidentifiability and nonlinearity, it is challenging to explore the statistical properties of deep neural networks(Dinh & Ho, 2020). Dinh and Ho, 2020 use Lojasewicz's inequality (Ji et al., 1992) for analytic functions to give an upper bound for the distance between $\hat{\beta}_n$ and $\mathcal{H}_\beta^*$ by the excess risk, thus avoiding the issue of the irregularity of the Hessian matrix.

To study the consistency of the outcome prediction estimator with group LASSO without a full geometric description of the $\mathcal{H}_\beta^*$, we need the following Lemmas. The first lemma states that members of the set $\mathcal{H}_\beta^*$ share hidden layers with activation functions that are are identical almost everywhere on $\mathcal{X}$:

Lemma 4.4.3. *The element $\beta_0 \in \mathcal{H}_\beta^* = \{\beta : R(\beta) = R(\beta^*)\}$ if and only if $f_{\beta_0} = f_{\beta^*}$ almost surely.*

Proof. For each $X \in \mathcal{X}$,

$$-p(X;\beta^*)\log q + 1 - p(X;\beta^*)\log(1-q)$$

is minimized by taking $q = p(X;\beta^*)$. So

$$\begin{aligned}
\inf_{\beta\in W} R(\beta) &= \inf_{\beta\in W} -E_X\Big\{p(X;\beta^*)\log p(X;\beta) - (1-p(X;\beta^*))\log(1-p(X;\beta))\Big\} \\
&\geq -E_X\Big\{\inf_{\beta\in W}\left[p(X;\beta^*)\log p(X;\beta) - (1-p(X;\beta^*))\log(1-p(X;\beta))\right]\Big\} \\
&= -E_X\Big\{\left[p(X;\beta^*)\log p(X;\beta^*) - (1-p(X;\beta^*))\log(1-p(X;\beta^*))\right]\Big\} \\
&= R(\beta^*),
\end{aligned}$$

where $p(X;\beta^*) = 0.5\{1 + f_{\beta^*}(X)/d\}$. Because $\beta_0 \in \mathcal{H}^*_\beta = \{\beta : R(\beta) = R(\beta^*)\}$,

$$\begin{aligned} R(\beta_0) &= -E_{P_X}\Big\{p\left(X;\beta^*\right)\log p\left(X;\beta_0\right) + \left(1 - p\left(X;\beta^*\right)\right)\log\left(1 - p\left(X;\beta_0\right)\right)\Big\} \\ &= \inf_{\beta\in W} R\left(\beta\right), \end{aligned}$$

where $p(X;\beta_0) = 0.5\{1 + f_{\beta_0}(X)/d\}$.

Thus we get $R(\beta^*) \leq \inf_{\beta\in W} R\left(\beta\right) = R(\beta_0)$ with equality if and only if $f_{\beta_0}(X)$ is equal to $f_{\beta^*}(X)$ almost surely on the support of P_X . Since $p_X(x)$ is continuous and positive on its open domain $\mathcal{X}$ and the function $f_\beta(X)$ is analytic, we deduce that $f_{\beta_0}(X) = f_{\beta^*}(X)$ almost everywhere. □

The following lemma describes the set of risk estimators, and is required to demonstrate the consistency of the estimators and variable selection:

Lemma 4.4.4. *For the set of risk minimizers $\mathcal{H}^*_\beta$:*

(i) For $\beta_0 \in \mathcal{H}^*_\beta$, the parameters $u = \beta_\mathcal{C} \bigcup \beta_\mathcal{P}$ of β_0 are bounded away from zero, i.e., $\|u_{k(\beta_0)}\| \geq c_0$ for all $\beta_0 \in \mathcal{H}^*_\beta$ and $k = 1, \ldots, n_u$, where $c_0 > 0$.

(ii) For $\beta_0 \in \mathcal{H}^*_\beta$, the vector $\phi(\beta_0)$ setting the parameters $v = \beta_\mathcal{I} \bigcup \beta_\mathcal{S}$ of β_0 to zero, also belongs to $\mathcal{H}^*_\beta$, i.e., $\phi(\beta_0) \in \mathcal{H}^*_\beta$ for all $\beta_0 \in \mathcal{H}^*_\beta$.

Proof. (i) For the sake of contradiction, we suppose that no such positive c_0 exists. Thus, for $\beta_0 \in \mathcal{H}^*_\beta$, there exists k such that $\|u_{k(\beta_0)}\| = 0$. According to Lemma 4.4.3, the $f_{\beta_0} = f_{\beta^*}$ does not depend on k-th $X_{k(u)}$ covariate that is predictive of outcome. Since this is a contradiction to Assumption 4.4.2, $\|u_{k(\beta_0)}\| \geq c_0$ for all $\beta_0 \in \mathcal{H}^*_\beta$ and $k = 1, \ldots, n_u$.

(ii) Because $\beta_0 \in \mathcal{H}^*$, Lemma 4.4.3 implies that $f_{\beta_0}(X_u, X_v) = f_{\beta^*}(X_u, X_v)$. Based on Assumption 4.4.2, $f_{\beta_0}(X_u, X_v) = f_{\beta_0}(X_u, 0)$. From the definition of vector $\phi(\beta_0)$, $f_{\beta_0}(X_u, 0) =$

$f_{\phi(\beta_0)}(X_u, X_v)$. Therefore, we have $f_{\beta^*}(X_u, X_v) = f_{\phi(\beta_0)}(X_u, X_v)$, which implies that $\phi(\beta_0) \in \mathcal{H}^*_\beta$ for all $\beta_0 \in \mathcal{H}^*_\beta$. □

It remains to prove that the estimator $\hat{\beta}_n$ of outcome prediction with group LASSO belongs to $\mathcal{H}^*_\beta$, and that there exists $\phi(\hat{\beta}_n)$ by setting setting the parameters $v = \beta_\mathcal{I} \bigcup \beta_\mathcal{S}$ of β_0 to zero, which can identify the variables predictive of the outcomes and provide a good estimation for regularization strengths in the next step: propensity score estimation with adaptive group LASSO. To prove the consistency of the outcome prediction estimator with group LASSO, we still need three Lemmas that explore the convergence rate, Lipschitzness of risk function, and generalization bound. The first of these lemmas gives an inequality between the excess risk $R(\beta) - R(\beta^*)$ and the distance between β and $\mathcal{H}^*_\beta$:

Lemma 4.4.5. *There exist $c_2, \nu > 0$ and such that $R(\beta) - R(\beta^*) \geq c_2 d(\beta, \mathcal{H}^*_\beta)^\nu$ for all $\beta \in \mathcal{W}$.*

Proof. Since $R(\beta)$ and $f_\beta(X)$ are analytic in both β and X, the excess risk $r(\beta) = R(\beta) - R(\beta^*)$ is also analytic in β. Therefore $\mathcal{H}^*_\beta = \{\beta : R(\beta) = R(\beta^*)\}$ is the zero level set of the analytic function r. By Lojasewicz's inequality for algebraic varieties (Ji et al., 1992), there exist positive constants C and ν such that $d(\beta, \mathcal{H}^*_\beta)^\nu \leq C|r(\beta)|$ for all $\beta \in \mathcal{W}$, which completes the proof. □

The following lemma proves that the risk $R(\beta)$ and empirical risk $R_n(\beta)$ functions are Lipschitz continuous. This Lemma requires that the output layer of the deep neural net $f_\beta(X)$ is Lipschitz continuous; that is, there exists a finite constant $c_0 > 0$ such that $|f_\beta(X) - f_{\beta'}| \leq c_0||\beta - \beta'||$ for all β and β' in $\mathcal{W}$ and $X \in \mathcal{X}$. Note that theis assumption is satisfied by the hyperbolic tangent function provided that $\mathcal{W}$ and $\mathcal{X}$ are bounded.

Lemma 4.4.6. *Suppose that $f_\beta(X)$ is Lipschitz continuous β and X with $|f_\beta(X)| \in (0, 1)$ for all $X \in (X)$ and all $\beta \in \mathcal{W}$, then the risk $R(\beta)$ is Lipschitz function with Lipschitz*

constant $c_1 > 0$. For any $\delta > 0$, there exists $M_\delta > c_1$ such that $R_n(\beta)$ is an M_δ-Lipschitz function with probability at least $1 - \delta$.

Proof. We first demonstrate that $\log p(X;\beta)$ and $\log(1-p(X;\beta))$ are Lipschitz continuous in X and β. Without loss of generality, assume that the $0 < c_2 \leq p(X_i;\beta) \leq p(X_i;\beta') \leq c_3 < 1$, where $c_2 = 0.5(1 - 1/d)$ and $c_3 = 0.5(1 + 1/d)$ are the bounds of the Bernoulli probabilities (4.4.1) with $d > 1$. Then

$$
\begin{aligned}
|\log p(X_i;\beta) - \log p(X_i;\beta')| &= \log \frac{p(X_i;\beta')}{p(X_i;\beta)} \\
&= \log\left(1 + \left(\frac{p(X_i;\beta')}{p(X_i;\beta)} - 1\right)\right) \\
&\leq \frac{p(X_i;\beta')}{p(X_i;\beta)} - 1 \\
&= \frac{1}{p(X_i;\beta)}\left(p(X_i;\beta') - p(X_i;\beta)\right) \\
&\leq \frac{1}{c_2}\left(p(X_i;\beta') - p(X_i;\beta)\right) \\
&= \frac{1}{c_2}|p(X_i;\beta') - p(X_i;\beta)| \\
&\leq c_4||\beta - \beta'||,
\end{aligned}
$$

where $c_4 = c_0/c_2$. Similarly,

$$\begin{aligned}
|\log(1-p(X_i;\beta)) - \log(1-p(X_i;\beta'))| =& \log\frac{1-p(X_i;\beta)}{1-p(X_i;\beta')} \\
=& \log\left(1+\left(\frac{1-p(X_i;\beta)}{1-p(X_i;\beta')}-1\right)\right) \\
\leq& \frac{1-p(X_i;\beta)}{1-p(X_i;\beta')}-1 \\
=& \frac{1}{1-p(X_i;\beta')}\left(p(X_i;\beta')-p(X_i;\beta)\right) \\
\leq& \frac{1}{1-c_3}\left(p(X_i;\beta')-p(X_i;\beta)\right) \\
=& \frac{1}{1-c_3}|p(X_i;\beta')-p(X_i;\beta)| \\
\leq& c_5||\beta-\beta'||,
\end{aligned}$$

where $c_5 = c_0/(1-c_3)$.

By the triangle and Jensen's inequality,

$$
\begin{aligned}
|R(\beta) - R(\beta')| =& \Big| E\Big\{ \Big\{ p(X;\beta^*) \log p(X;\beta) + (1 - p(X;\beta^*)) \log(1 - p(X;\beta)) \Big\} \\
& - \Big\{ p(X;\beta^*) \log p(X;\beta') + (1 - p(X;\beta^*)) \log(1 - p(X;\beta')) \Big\} \Big\} \Big| \\
=& \Big| E\Big\{ p(X_i;\beta^*) \Big\{ \log p(X;\beta) - \log p(X_i;\beta') \Big\} \\
& + (1 - p(X;\beta^*)) \Big\{ \log(1 - p(X;\beta)) - \log(1 - p(X;\beta')) \Big\} \Big\} \Big| \\
\leq& \Big| E\Big\{ p(X;\beta^*) \Big\{ \log p(X;\beta) - \log p(X;\beta') \Big\} \Big\} \Big| \\
& + \Big| E\Big\{ (1 - p(X;\beta^*)) \Big\{ \log(1 - p(X;\beta)) - \log(1 - p(X;\beta')) \Big\} \Big\} \Big| \\
\leq& E\Big\{ p(X;\beta^*) | \log p(X;\beta) - \log p(X;\beta') | \Big\} \\
& + E\Big\{ (1 - p(X;\beta^*)) | \log(1 - p(X;\beta)) - \log(1 - p(X;\beta')) | \Big\} \\
\leq& E\Big\{ | \log p(X;\beta) - \log p(X;\beta') | \Big\} \\
& + E\Big\{ | \log(1 - p(X;\beta)) - \log(1 - p(X;\beta')) | \Big\} \\
\leq& (c_4 + c_5) ||\beta - \beta'|| \\
=& c_1 ||\beta - \beta'||
\end{aligned}
$$

where $c_1 = c_4 + c_5$.

Therefore, the risk $R(\beta)$ is Lipschitz continuous in β.

Similarly, we can get the empirical risk function $R_n(\beta)$ is an M_δ-Lipschitz function with probability at least $1 - \delta$

$$
\begin{aligned}
|R_n(\beta) - R_n(\beta')| =& \Big| \frac{1}{n} \sum_{i=1}^{n} \Big\{ Y_i \log p(X_i;\beta) + (1-Y_i) \log(1 - p(X_i;\beta)) \Big\} \\
& - \frac{1}{n} \sum_{i=1}^{n} \Big\{ Y_i \log p(X_i;\beta') + (1-Y_i) \log(1 - p(X_i;\beta')) \Big\} \Big| \\
=& \Big| \frac{1}{n} \sum_{i=1}^{n} \Big\{ Y_i \Big\{ \log p(X_i;\beta) - \log p(X_i;\beta') \Big\} \\
& + (1-Y_i) \Big\{ \log(1 - p(X_i;\beta)) - \log(1 - p(X_i;\beta')) \Big\} \Big\} \Big| \\
\leq & \Big| \frac{1}{n} \sum_{i=1}^{n} \Big\{ Y_i \Big\{ \log p(X_i;\beta) - \log p(X_i;\beta') \Big\} \Big\} \Big| \\
& + \Big| \frac{1}{n} \sum_{i=1}^{n} \Big\{ (1-Y_i) \Big\{ \log(1 - p(X_i;\beta)) - \log(1 - p(X_i;\beta')) \Big\} \Big\} \Big| \\
\leq & \frac{1}{n} \sum_{i=1}^{n} \Big\{ Y_i \Big| \log p(X_i;\beta) - \log p(X_i;\beta') \Big| \Big\} \\
& + \frac{1}{n} \sum_{i=1}^{n} \Big\{ (1-Y_i) \Big| \log(1 - p(X_i;\beta)) - \log(1 - p(X_i;\beta')) \Big| \Big\} \\
\leq & \frac{1}{n} \sum_{i=1}^{n} \Big\{ c_4 Y_i ||\beta - \beta'|| + c_5 (1-Y_i) ||\beta - \beta'|| \Big\} \\
\leq & ||\beta - \beta'|| \frac{1}{n} \sum_{i=1}^{n} \Big\{ (c_4 - c_5) Y_i + c_5 \Big\} \\
= & c_6 ||\beta - \beta'|| \Big(c_7 + \frac{1}{n} \sum_{i=1}^{n} Y_i \Big)
\end{aligned}
$$

where $c_6 = c_4 - c_5$ and $c_7 = \frac{c_5}{c_4 - c_5}$.

The proof that $R_n(\beta)$ is a Lipschitz function is completed by noting that by Chebyshev's inequality

$$
Pr\left(\frac{1}{n} \sum_{i=1}^{n} |Y_i| > M_\delta \right) \leq \frac{E|Y_i|}{M_\delta}
$$

that for any $M_\delta > 0$.

□

The following lemma describes the rate at which the empirical risk $R_n(\beta)$ converges to the risk $R(\beta)$ as $n \to \infty$. It extends Lemma 3.3 of Dinh and Ho, 2020 to models for binary outcome variables.

Lemma 4.4.7. *For any $\delta > 0$, there exist $c_1(\delta) > 0$ such that*

$$|R_n(\beta) - R(\beta)| \leq c_1 \frac{\log n}{\sqrt{n}}, \quad \textit{for all } \beta \in \mathcal{W}$$

with probability at least $1 - \delta$.

Proof. Since the hyperbolic tangent function is a bounded Lipschitz function and the weight space $\mathcal{W}$ is bounded, $f_\beta(X_i)$ is bounded. According to the definition of $p(X_i;\beta) = 0.5\{1 + \frac{f_\beta(X_i)}{d}\}$, $\log p(X_i;\beta)$ and $\log(1 - p(X_i;\beta))$ are also bounded when $d > 1$:

$$\log \frac{1 - \frac{1}{d}}{2} \leq \log p(X_i;\beta) \leq \log \frac{1 + \frac{1}{d}}{2};$$

$$\log \frac{1 - \frac{1}{d}}{2} \leq \log(1 - p(X_i;\beta)) \leq \log \frac{1 + \frac{1}{d}}{2};$$

Thus, we have

$$\begin{aligned}
Z_i =& \Big\{ Y_i \log p(X_i;\beta) + (1 - Y_i) \log(1 - p(X_i;\beta)) \Big\} \\
& - \Big\{ p\left(X_i;\beta^*\right) \log p\left(X;\beta\right) + \left(1 - p\left(X_i;\beta^*\right)\right) \log\left(1 - p\left(X_i;\beta\right)\right) \Big\} \\
=& \Big\{ Y_i - p\left(X_i;\beta^*\right) \Big\} \log p\left(X_i;\beta\right) + \Big\{ (1 - Y_i) - \left(1 - p\left(X_i;\beta^*\right)\right) \Big\} \log(1 - p\left(X_i;\beta\right)) \\
\leq& \log p\left(X_i;\beta\right) + \log\left(1 - p\left(X_i;\beta\right)\right) \\
\leq& 2 \log \frac{1 + \frac{1}{d}}{2}
\end{aligned}$$

$$
\begin{aligned}
Z_i =& \Big\{ Y_i - p\left(X_i;\beta^*\right) \Big\} \log p\left(X_i;\beta\right) + \Big\{ (1-Y_i) - \left(1 - p\left(X_i;\beta^*\right)\right) \Big\} \log(1 - p\left(X_i;\beta\right)) \\
&\geq - \log p\left(X_i;\beta\right) - \log\left(1 - p\left(X_i;\beta\right)\right) \\
&\geq - 2\log \frac{1+\frac{1}{d}}{2}
\end{aligned}
$$

Therefore, we get

$$
|Z_i| \leq 2\log \frac{1+\frac{1}{d}}{2} = C_0
$$

for all $i = 1, 2, ..., n$. Applying Bernstein's inequality, we have

$$
Pr\Big\{ |R_n(\beta) - R(\beta)| \geq \frac{\tau}{2} \Big\} \leq 2\exp\Big\{ - \frac{n^2\tau^2}{2n\sigma^2 + \frac{2}{3}C_0 n\tau} \Big\} \leq \exp\{-C_1 n\tau^2\}
$$

where $C_1 = \sigma^- 2$,

$$
\sigma^2 = \mathrm{var}_{P_{X,Y}}\{Y_i \log p(X_i;\beta) + (1-Y_i)\log[1 - p(X_i;\beta)]\}.
$$

and $t \geq 0$. The remainder of the proof follows closely that of Lemma 3.3 of Dinh and Ho, 2020. Define the events

$$
\mathcal{A}(\beta,\tau) = \{|R_n(\beta) - R(\beta)| > \tau/2\},
$$

$$
\mathcal{B}(\beta,\tau) = \{\exists \beta' \in \mathcal{W} \text{ such that} \|\beta' - \beta\| \leq \frac{\tau}{4M_\delta} \text{ and } |R_n(\beta') - R(\beta')| > \tau\},
$$

and

$$
\mathcal{C} = \{|R_n(\beta) - R_n(\beta')| \leq M_\delta \|\beta - \beta'\|, \forall \beta, \beta' \in \mathcal{W}\}.
$$

M_δ in $\mathcal{B}(\beta,\tau)$ and $\mathcal{C}$ is defined in Lemma 4.4.6. Therefore, $\mathcal{B}(\beta,\tau) \cap \mathcal{C} \subset \mathcal{A}(\beta,\tau)$ and $P(\mathcal{C}) \geq 1 - \delta$. Let $m = dim(\mathcal{W})$. Since $\mathcal{W}$ is a compact set, there exist $C_3(m) \geq 1$ and a

finite set $\mathcal{H} \subset \mathcal{W}$ such that

$$\mathcal{W} \subset \bigcup_{\beta \in \mathcal{H}} \mathcal{V}(\beta, \epsilon) \quad \text{and} \quad |\mathcal{H}| \leq C_3/\epsilon^m$$

where $\epsilon = \tau/(4M_\delta)$, $\mathcal{V}(\beta, \epsilon)$ denotes the open ball centered at β with radius ϵ, and $|\mathcal{H}|$ denotes the cardinality of $\mathcal{H}$. By a union bound, the probability that there exists $\beta \in \mathcal{H}$ such that $|R_n(\beta) - R(\beta)| \geq \tau/2$ has upper bound

$$Pr\left[\exists \beta \in \mathcal{H} : |R_n(\beta) - R(\beta)| > \tau/2\right] \leq \frac{C_3(4M_\delta)^m}{\tau^m} e^{-C_2 n \tau^2}.$$

Moreover, since $\mathcal{B}(\beta, \tau) \cap \mathcal{C} \subset \mathcal{A}(\beta, \tau)$ for all $\beta \in \mathcal{H}$, we deduce

$$Pr\left[\{\exists \beta \in \mathcal{W} : |R_n(\beta) - R(\beta)| > \tau\} \cap \mathcal{C}\right] \leq C_4 \tau^{-m} e^{-C_2 n \tau^2}.$$

Hence,

$$Pr\left[\{\exists \beta \in \mathcal{W} : |R_n(\beta) - R(\beta)| > \tau\}\right] \leq C_4 \tau^{-m} e^{-C_2 n \tau^2} + \delta.$$

To complete the proof, we chose τ in such a way that $C_4 \tau^{-m} e^{-C_2 n \tau^2} \leq \delta$. This can be done by choosing $\tau = \mathcal{O}(\log n/\sqrt{n})$. □

Combining Lemmas 4.4.5 and 4.4.7, we have the following Theorem that demonstrates the consistency of outcome prediction estimator under group LASSO:

Theorem 4.4.8. *For any $\delta > 0$ and $\lambda_n = \mathcal{O}(n^{-1/4})$, there exist $C > 0$ and $N_\delta > 0$ such that for all $n \geq N_\delta$,*

$$d(\hat{\beta}_n, \mathcal{H}^*_\beta) \leq C \left(\frac{\log n}{n}\right)^{\frac{1}{4(\nu-1)}} \quad \textit{and} \quad \|\hat{v}_{\hat{\beta}_n}\| \leq C \left(\frac{\log n}{n}\right)^{\frac{1}{4(\nu-1)}}.$$

with probability at least $1 - \delta$.

Proof. The proof of this Theorem is similar to that of Theorem 3.4 in Dinh and Ho, 2020. For completeness, we include the proof below. Let $\phi(\beta)$ denote the weight vector obtained from β by setting the v-components to zero. If we define $\beta_n = \arg\min_{\beta \in \mathcal{H}^*_\beta} \|\hat{\beta}_n - \beta\|$ then $\phi(\beta_n) \in \mathcal{H}^*_\beta$ and $R(\beta_n) = R(\phi(\beta_n))$ by Lemma 4.4.4. Since $q(\beta)$ in Eq.4.3.3 is a Lipschitz function, we have

$$\begin{aligned} c_2 d(\hat{\beta}_n, \mathcal{H}^*_\beta)^\nu &= c_2 \|\beta_n - \hat{\beta}_n\|^\nu \\ &\leq R(\hat{\beta}_n) - R(\beta_n) \\ &\leq 2c_1 \frac{\log n}{\sqrt{n}} + \lambda_n \left(q(\beta_n) - q(\hat{\beta}_n) \right) \\ &\leq 2c_1 \frac{\log n}{\sqrt{n}} + \lambda_n C \|\beta_n - \hat{\beta}_n\| \end{aligned}$$

which implies through Young's inequality that

$$\|\beta_n - \hat{\beta}_n\|^\nu \leq C_1 \lambda_n^{\nu/(\nu-1)} + C_2 \frac{\log n}{\sqrt{n}}.$$

Let G denote the part of the regularization term without the v-component. We note that G is a Lipschitz function, $G(\phi(\beta)) = G(\beta)$ for all β, and $R(\phi(\beta_n)) = R(\hat{\beta}_n)$. Thus,

$$\begin{aligned} \lambda_n \sum_k \|\hat{v}_{k(\beta_n)}\| &\leq R_n(\phi(\beta_n)) - R_n(\hat{\beta}_n) + \lambda_n [G(\phi(\beta_n)) - G(\hat{\beta}_n)] \\ &\leq 2c_1 \frac{\log n}{\sqrt{n}} + R(\phi(\beta_n)) - R(\hat{\beta}_n) + \lambda_n [G(\beta_n) - G(\hat{\beta}_n)] \\ &\leq 2c_1 \frac{\log n}{\sqrt{n}} + \lambda_n C \|\beta_n - \hat{\beta}_n\|. \end{aligned}$$

If $\lambda_n = \mathcal{O}(n^{-1/4})$, then with probability at least $1 - \delta$,

$$d(\hat{\beta}_n, \mathcal{H}^*) \leq C \left(\frac{\log n}{n} \right)^{\frac{1}{4(\nu-1)}} \quad \text{and} \quad \|\hat{v}_n\| \leq C \left(\frac{\log n}{n} \right)^{\frac{1}{4(\nu-1)}}.$$

This completes the proof. □

Similar to the definitions of risk functions in outcome prediction with group LASSO, assume that conditional on covariates X, treatment T is sampled from Bernoulli distributions with probability

$$p(X_i;\alpha) = 0.5\{1 + g_\alpha(X_i)\},$$

where $g_\alpha(X_i)$ is the output layer of deep neural network with analytic activation functions (i.e., hyperbolic tangent function) and α belongs to the hypercube $\mathcal{W}$. The true value α^* for α is assumed to be in the interior of $\mathcal{W}$. Then the negative log likelihood is

$$K_n(\alpha) = -\frac{1}{n}\sum_{i=1}^{n}\Big\{Y_i \log p(X_i;\alpha) + (1-Y_i)\log[1-p(X_i;\alpha)]\Big\}.$$

The risk is defined to be

$$\begin{aligned} K(\alpha) &= -E_{P^*_{X,Y}}\Big\{\{Y\log p(X_i;\alpha) + (1-Y)\log(1-p(X_i;\alpha))\}\Big\} \\ &= -E_{P_X}\Big\{p(X_i;\alpha^*)\log p(X_i;\alpha) + (1-p(X_i;\alpha^*))\log(1-p(X_i;\alpha))\Big\} \end{aligned}$$

Similarly, we define the set of risk minimizers as $\mathcal{H}^*_\alpha := \{\alpha : K(\alpha) = K(\alpha^*)\}$ and can get the Lemmas 4.4.3, 4.4.5, 4.4.6, and 4.4.7. The parameters in the first hidden layer for $g_\alpha(X_i)$ are partitioned into two groups $m = \alpha_\mathcal{I} \bigcup \alpha_\mathcal{S}$ and $e = \alpha_\mathcal{C} \bigcup \alpha_\mathcal{P}$, which correspond to $v = \beta_\mathcal{I} \bigcup \beta_\mathcal{S}$ and $u = \beta_\mathcal{C} \bigcup \beta_\mathcal{P}$.

The following Theorem proves the consistency of estimator and variable selection under Adaptive group LASSO:

Theorem 4.4.9. *Let $\gamma > 0$, $\epsilon > 0$, $\lambda_n = \mathcal{O}(n^{-1/4})$, and $\theta_n = \Omega(n^{-\gamma/(4\nu-4)+\epsilon})$, for any $\delta > 0$. Then there exists N_δ such that for $n > N_\delta$,*

$$d(\widetilde{\alpha}_n, \mathcal{H}^*_\alpha) \leq C \left(\frac{\log n}{n}\right)^{\frac{1}{4(\nu-1)}} \quad \text{and} \quad \|\widetilde{m}_{\widetilde{\alpha}_n}\| = 0$$

with probability at least $1-\delta$, where $\widetilde{m}_{\widetilde{\alpha}_n} = \widetilde{\alpha}_{\mathcal{I}} \bigcup \widetilde{\alpha}_{\mathcal{S}}$.

Proof. For the sake of contradiction, suppose that $\widetilde{m}_{z(\widetilde{\alpha}_n)} \neq 0$ for some z and have the $\phi(\widetilde{\alpha}_n)$ setting the parameter $\widetilde{m}_{z(\widetilde{\alpha}_n)}$ of $\widetilde{\alpha}_n$ to zero and keeping other parameters of $\widetilde{\alpha}_n$. By the definition of $\widetilde{\alpha}_n$, we have

$$\begin{aligned} &K_n(\widetilde{\alpha}_n) + \theta_n \sum_{l=1}^{n_v} \frac{\|\widetilde{m}_{l(\widetilde{\alpha}_n)}\|}{\|\hat{v}_{l(\hat{\alpha}_n)}\|^\gamma} + \theta_n \sum_{k=1}^{n_u} \frac{\|\widetilde{e}_{k(\widetilde{\alpha}_n)}\|}{\|\hat{u}_{k(\hat{\alpha}_n)}\|^\gamma} \\ &\leq K_n(\phi(\widetilde{\alpha}_n)) + \theta_n \sum_{l=1}^{n_v} \frac{\|\widetilde{m}_{l(\phi(\widetilde{\alpha}_n))}\|}{\|\hat{v}_{l(\hat{\alpha}_n)}\|^\gamma} + \theta_n \sum_{k=1}^{n_u} \frac{\|\widetilde{e}_{k(\widetilde{\alpha}_n)}\|}{\|\hat{u}_{k(\hat{\alpha}_n)}\|^\gamma} \end{aligned}$$

After canceling out the same terms, we have

$$K_n(\widetilde{\alpha}_n) + \theta_n \frac{\|\widetilde{m}_{z(\widetilde{\alpha}_n)}\|}{\|\hat{v}_{z(\hat{\beta}_n)}\|^\gamma} \leq K_n(\phi(\widetilde{\alpha}_n)).$$

According to Lemma 4.4.6, there exists M_δ s.t.

$$\theta_n \frac{\|\widetilde{m}_{z(\widetilde{\alpha}_n)}\|}{\|\hat{v}_{z(\hat{\beta}_n)}\|^\gamma} \leq K_n(\phi(\widetilde{\alpha}_n)) - K_n(\widetilde{\alpha}_n) \leq M_\delta \|\phi(\widetilde{\alpha}_n) - \widetilde{\alpha}_n\| = M_\delta \|\widetilde{m}_{z(\widetilde{\alpha}_n)}\|$$

with probability at least $1-\delta$. Since $\widetilde{m}_{z(\widetilde{\alpha}_n)} \neq 0$, we deduce that $\theta_n \frac{1}{\|\hat{v}_{z(\hat{\beta}_n)}\|^\gamma} \leq M_\delta$. This contradicts Theorem 4.4.8, which proves that for n large enough

$$\theta_n \frac{1}{\|\hat{v}_{z(\hat{\beta}_n)}\|^\gamma} \geq C_\delta^{-\gamma}\ \theta_n \left(\frac{n}{\log n}\right)^{\frac{\gamma}{4(\nu-1)}} \geq 2M_\delta$$

with probability at least $1-\delta$.

Since Theorem 4.4.8 and Lemma 4.4.4, we conclude that $\hat{u}_{k(\hat{\alpha}_n)}$ is bounded away from zero as $n \to \infty$. Thus,

$$q_n(\alpha^*) = \sum_{l=1}^{n_v} \frac{\|\widetilde{m}_{l(\widetilde{\alpha}_n)}\|}{\|\hat{v}_{l(\hat{\alpha}_n)}\|^\gamma} + \sum_{k=1}^{n_u} \frac{\|\widetilde{e}_{k(\widetilde{\alpha}_n)}\|}{\|\hat{u}_{k(\hat{\alpha}_n)}\|^\gamma} = \sum_{k=1}^{n_u} \frac{\|\widetilde{e}_{k(\widetilde{\alpha}_n)}\|}{\|\hat{u}_{k(\hat{\alpha}_n)}\|^\gamma} < \infty$$

with probability at least $1-\delta$.

$$c_2 d(\widetilde{\alpha}_n, \mathcal{H}_\alpha^*)^\nu \leq K(\widetilde{\alpha}_n) - K(\alpha^*) \leq 2c_1 \frac{\log n}{\sqrt{n}} + \theta_n \left(q_n(\alpha^*) - q_n(\hat{\alpha}_n)\right) \leq 2c_1 \frac{\log n}{\sqrt{n}} + \theta_n q_n(\alpha^*)$$

Similar to the proof in Theorem 4.4.8, with probability at least $1-\delta$,

$$d(\widetilde{\alpha}_n, \mathcal{H}_\alpha^*) \leq C \left(\frac{\log n}{n}\right)^{\frac{1}{4(\nu-1)}}.$$

This completes the proof.

□

4.5 Simulation Study

In this section, we first conduct simulation experiments on the simulation datasets to evaluate the following aspects: (1) Compared with alternative approaches, our proposed method can improve treatment effect estimation with respect to average treatment effect. (2) The deep adaptive variable selection can accurately select out the confounders and adjustment variables from observational data with high-dimensional variables. (3) The proposed method is robust to different levels of treatment selection bias and the ablation study proves the necessity of tuning parameter selection procedure.

4.5.1 Design for Generating Simulation Data set.

To mimic situations where there are large numbers of variables including instrumental, adjustment, confounding, and irrelevant variables, we generate a synthetic data set that reflects the complexity of observational medical records data. The number of observed variables is set to either 30 or 200, including 5 confounders, 5 adjustment variables, 5 instrumental variables, and either 15 or 185 spurious variables. All of these observed variables are generated from a multivariate Gaussian distribution with zero means, unit variances, and two different exchangeable correlation structures, i.e., moderate correlation ($\rho = 0.1$) and stronger correlation ($\rho = 0.3$). The model used to generate the continuous outcome variable Y in this simulation is the partially linear regression model extending the ideas described in (Robinson, 1988):

$$Y = \eta T + g_{\beta^\star}\Big((X_{\mathcal{C}}{}^\intercal, X_{\mathcal{P}}{}^\intercal)^\intercal\Big) + \epsilon, \tag{4.5.1}$$

where $T \overset{ind.}{\sim} \text{Bernoulli}\Big(e_{\alpha^\star}\big(X_{\mathcal{C}}{}^\intercal, X_{\mathcal{I}}{}^\intercal)^\intercal\big)\Big)$, ϵ are independently sampled from a standard normal distribution, and the true treatment effect η is set to 2. The generating function Eq. (8.4.1), $g_{\beta^\star}\Big((X_{\mathcal{C}}{}^\intercal, X_{\mathcal{P}}{}^\intercal)^\intercal\Big)$ takes one of two forms: The first is a linear of the confounders $X_{\mathcal{C}}$ and the adjustment variables $X_{\mathcal{P}}$. The second is a complex nonlinear model with tangent activation function of the confounders $X_{\mathcal{C}}$ and adjustment variables $X_{\mathcal{P}}$. The parameters $\beta^\star$ are sampled independently from a standard normal distribution. The treatment assignments T are independently sampled from a Bernoulli distribution with probability $e_{\alpha^\star}\big((X_{\mathcal{C}}{}^\intercal, X_{\mathcal{I}}{}^\intercal)^\intercal\big)$, where $e_{\alpha^\star}\big((X_{\mathcal{C}}{}^\intercal, X_{\mathcal{I}}{}^\intercal)^\intercal\big)$ represents the true propensity score. Similar to $g_{\beta^\star}$, $\text{logit}\{e_{\alpha^\star}\}$ also has two forms i.e., simple linear and complex nonlinear. The parameter$\alpha^\star$ are sampled independently from $N(0, 1)$. In summary, we have a total of 6 scenarios; i.e., (a) $\rho = 0.1$, linear, $d = 30$; (b) $\rho = 0.1$, linear, $d = 200$; (c) $\rho = 0.3$, linear, $d = 200$; (d)

$\rho = 0.1$, nonlinear, $d = 30$; (e) $\rho = 0.1$, nonlinear, $d = 200$; (f) $\rho = 0.3$, nonlinear, $d = 200$. For each scenario, we repeat the random sampling procedure to obtain 1000 synthetic data sets. In each dataset, the sample size is 1000.

4.5.2 Comparison of Causal Inference Variable Selection Approaches.

To illustrate the importance of variable selection in causal inference and explore the impact of different types of covariates on the IPTW estimator, we consider three propensity score models: **Pote** includes all variables predictive of the outcome variable or treatment assignment including confounding, adjustment, or instrumental variables (i.e., $X_{\mathcal{C}}$, $X_{\mathcal{P}}$, and $X_{\mathcal{I}}$). **Conf** includes only confounders (i.e., $X_{\mathcal{C}}$). **Targ** includes only confounders and adjustment variables (i.e., $X_{\mathcal{C}}$ and $X_{\mathcal{P}}$). From previous work (Greenland, 2008; Myers et al., 2011; Patrick et al., 2011; Sauer et al., 2013; Schisterman et al., 2009; Shortreed & Ertefaie, 2017), we expect that including instrumental variable will increase the bias and variance, and including adjustment variable will reduce variance but will not have a significant impact on bias. **Targ** is expected to have the smallest bias and standard error. Our DAVSPS is designed to select out the same variables as **Targ**.

We also compare our model with two well-behaved models, the outcome-adaptive LASSO (OAL) of Shortreed and Ertefaie, 2017, and feature selection representation learning (RL) of Z. Chu et al., 2020. OAL selects covariates and estimates corresponding coefficients by incorporating information about the outcome-covariate relationships. OAL is built on simple linear regression, and adaptive LASSO is used to select out the covariates. RL is one method based on the deep representation learning, which maps the original covariate space into a selective, nonlinear and balanced representation space by simultaneously predicting the treatment assignment and outcomes. Due to the structure of fully connected neural

networks, the standard LASSO cannot be used to select out covariates. So, RL contains a sparse one-to-one layer between the input and the first hidden layer thereby permitting the application of LASSO variable selection and regularization. The variable selection at the input level can help select which variables are input into the neural network, which makes the deep neural network more interpretable. The parameters of competing methods (OAL and RL) are set the same as suggested in the original papers. Following the suggestions of Dinh and Ho, 2020, the regularizing parameters λ and θ were selected from the set $\{0.001, 0.01, 0.05, 0.1, 0.5, 1, 2\}$ and $\gamma = 2$. The number of hidden layers is in a range of $\{1, 2, 3, 4\}$ with a range of nodes within each layer $\{20, 30, 40, 50\}$. The numbers of layers and nodes within each layer were chosen based on the average training errors of outcome prediction and treatment classification in propensity score estimation.

Results on Simulation datasets. Figure 4.3 shows the estimates of the average treatment effect (ATE) of our method and competing methods under the six scenarios over 1000 realizations. In each scenario, we provide the box plots of ATE estimates for each model. Under moderate correlation, low dimension of covariates, and the linear model, OAL based on simple linear regression achieves the best performance and the RL and DAVSPS based on deep neural network slightly underperform OAL. Under these scenarios, deep neural network methods with more parameters and nonlinear operations are not expected to come without cost. However, with stronger correlations in scenarios (c) and (f), more covariates in scenarios (d) and (f), and nonlinear relationships between predictor and outcome variables in scenarios (d)-(f), the RL and DAVSPS fully demonstrate the advantages of deep neural network methods. In particular, our DAVSPS significantly outperforms OAL and RL with respect to both bias and variance, and almost performs as well as the ideal model Targ which only includes confounders and adjustment variables.

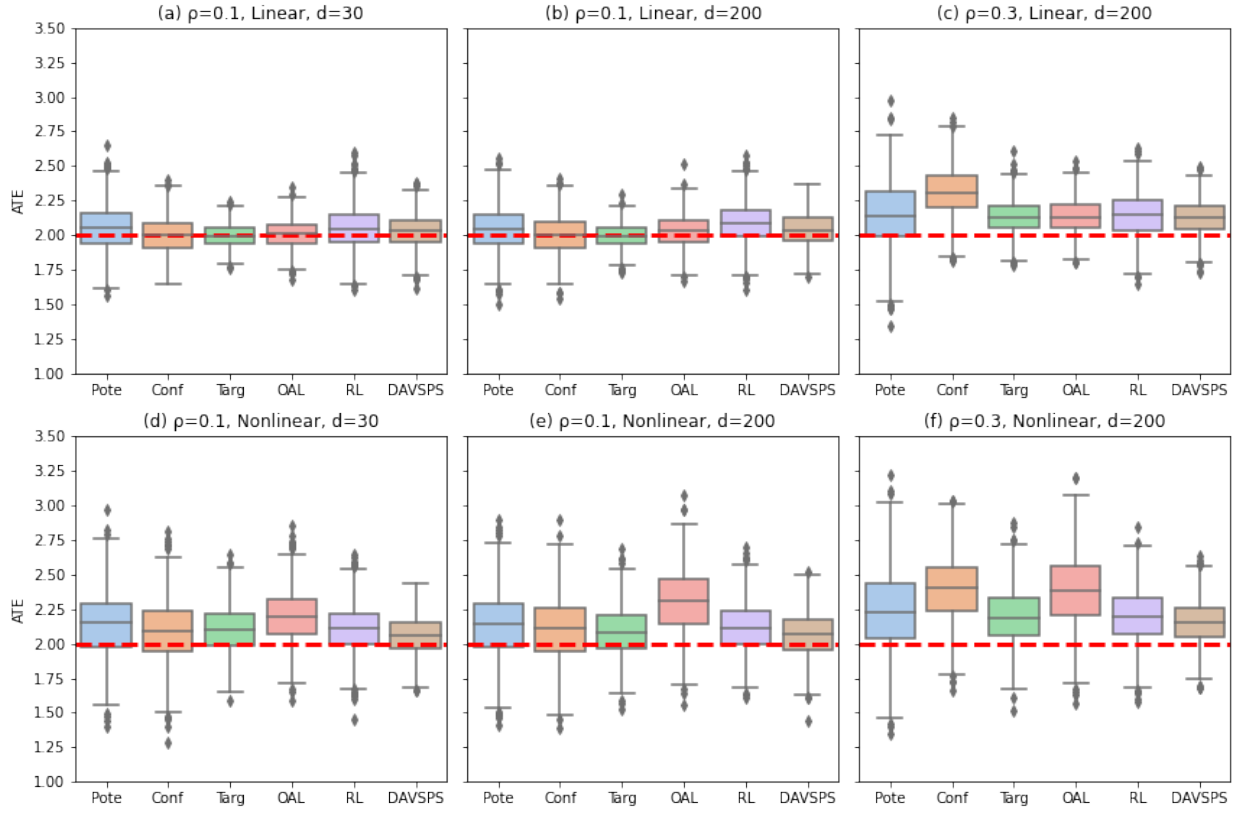

Figure 4.3: Box plots of 1000 inverse probability weighted estimates for the ATE under 6 scenarios.

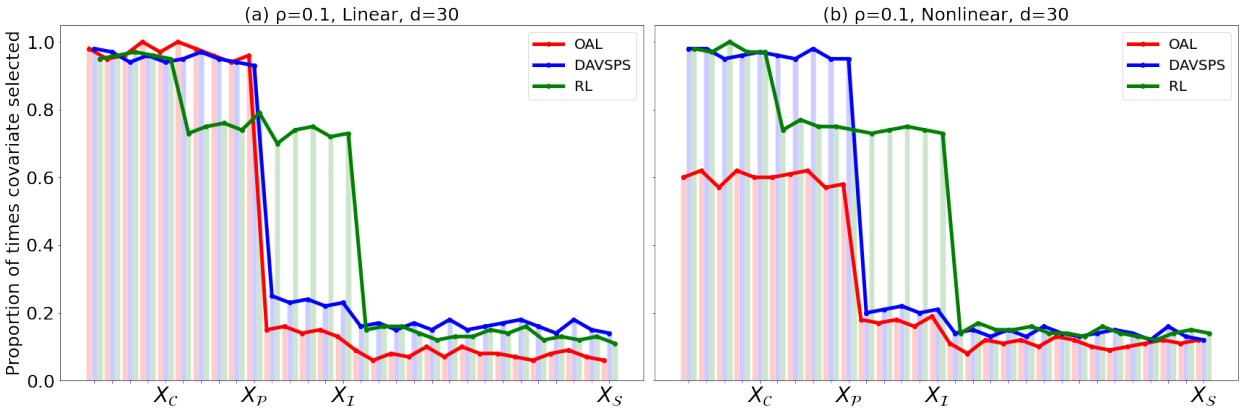

Figure 4.4: Proportion of times covariates were selected over 1000 simulations for scenarios (a) and (d).

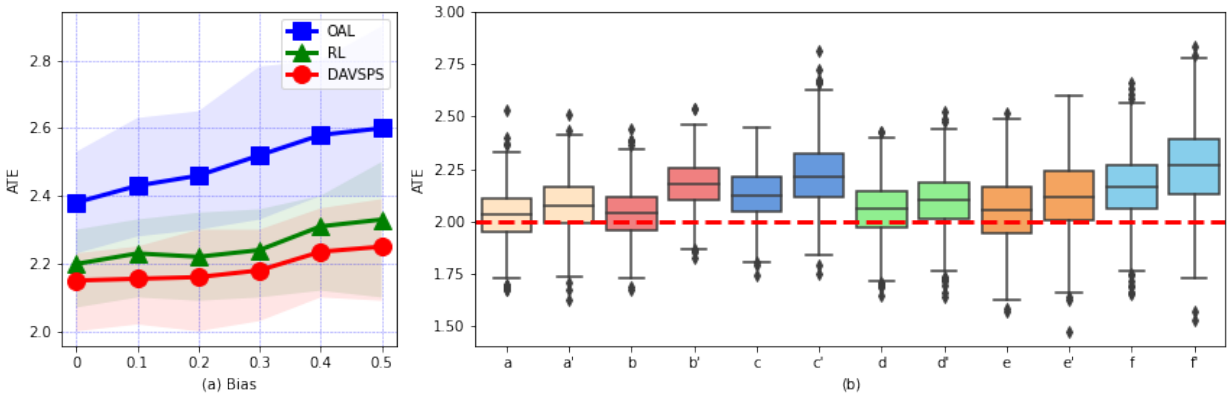

Figure 4.5: (a) ATE performance on simulation dataset with different degrees of treatment selection bias. (b) is the ablation study on integral probability metric.

Figure 4.4 describes the variable selection results for OAL, RL, and DAVSPS under scenarios (a) and (d) with 30 covariates. OAL accurately identifies the confounders $X_{\mathcal{C}}$ and adjustment variables $X_{\mathcal{P}}$ under the linear model, but performs poorly when the relationship between the outcome and covariates is nonlinear. RL has relatively stable performance in both scenarios where it can identify the confounders $X_{\mathcal{C}}$, but cannot discriminate the adjustment valuables $X_{\mathcal{P}}$ and instrumental variables $X_{\mathcal{I}}$. DAVSPS, consistently selected the confounders $X_{\mathcal{C}}$ and adjustment variables $X_{\mathcal{P}}$.

To evaluate the robustness to different levels of treatment selection bias, we carried out a study of the impact of selection bias. Although in our simulation procedure, the treatment selection bias has been taken in account based on their own propensity score $e_{\alpha^\star}\left(X_{\mathcal{C}}{}^\intercal, X_{\mathcal{I}}{}^\intercal)^\intercal\right)$, we use conditional sampling from treatment and control groups to increase the treatment selection bias. If the propensity score $e_{\alpha^\star}$ is equal to constant 0.5, it means no matter what the confounders and instrumental variables are, the unit is randomly assigned to either the treatment or the control group with the same probability, so that there is no treatment selection bias. The greater $|e_{\alpha^\star}\left(X_{\mathcal{C}}{}^\intercal, X_{\mathcal{I}}{}^\intercal)^\intercal\right) - 0.5|$ is, the larger selection bias will end up

getting. Following the idea in Shalit et al., 2017, with probability $1-q$, we randomly draw the treatment and control units; with probability q, we draw the treatment and control units that have the greatest $|e_{\alpha^\star}\left(X_{\mathcal{C}}{}^\intercal, X_{\mathcal{I}}{}^\intercal)^\intercal\right)^\intercal) - 0.5|$. Thus, the higher the q is, the larger the selection bias is. We run OAL, RL, and DAVSPS on the simulation datasets with q from 0 to 0.5, and show the results in Figure 4.5 (a). We can observe that our method consistently outperforms the competing methods under different levels of divergence and is robust to a high level of treatment assignment bias.

To demonstrate the necessity of IPM tuning parameter selection criterion that minimizing the treatment selection bias, we adopt the standard tuning parameter selection criterion for neural network model, i.e., minimizing training error, instead of using the IPM. As shown in Figure 4.5 (b), under six different scenarios, the performance (letter with prime) becomes poor after replacing the IPM tuning parameter selection compared to the original DAVSPS (letter without prime).

4.6 Racial disparities in severe maternal morbidity based on National Inpatient Sample

A severe maternal morbidity (SMM) is any complication of labor and delivery that can result in death or other significant consequences to a woman's well-being (CDCReproductiveHealth, 2020). Women who nearly died but survived these complications have been studied as surrogates of maternal deaths making SMM an important risk factor for maternal mortality (WHOMaternalMortality, 2020). SMMs such as severe cardiovascular conditions, hemorrhage, and other complications have been identified as the leading causes of maternal mortality, accounting for nearly 75% of all maternal deaths (WHOMaternalMortality,

2020). Most maternal deaths are preventable; therefore, it is important to understand the predictors and correlates associated with SMM to aid in prevention and early diagnosis.

There are significant racial and ethnic disparities in both maternal morbidity and mortality in the United States (CDCReproductiveHealth, 2020; Fingar et al., 2018; Geller et al., 2018). Black women are three to four times more likely to experience a maternal death than White women (CDCReproductiveHealth, 2020; Geller et al., 2018). In addition, Black women are also significantly more likely to experience severe maternal morbidity than White women (Fingar et al., 2018). Given the known racial disparities in maternal morbidity, it is important to understand which factors can predict a woman's chance of experiencing an SMM. Understanding the predictors of SMM could help to prevent and manage SMMs and improve care for pregnant women, laboring women, and women in postpartum.

This study uses the Healthcare Cost and Utilization Project (HCUP) National Inpatient Sample (NIS) from 2016 to 2018. The NIS is a largely publicly available database that provides administrative health care data. NIS is designed to produce regional and national estimates for a variety of healthcare related topics such as medical conditions, procedures, hospital characteristics, among others. Our analytic sample consists of $1,412,179$ patient discharge records and it can be categorized by race and SMM as Table 4.1. Variables that are known to be associated with differences in the rate of SMM were observed. These variables were maternal age, community income, type of insurance, location of residence, hospital teaching status, region of hospital, and ownership of hospital. Each of these variables was defined in previous studies. All the multi-value categorical variables were converted to dummy variables to assist selection of important covariates. Therefore, 24 covariates were taken into variable selection consideration. The detailed descriptions of each covariate distribution are provided in Table 4.2.

To capture the true racial disparities between Black and White in severe maternal morbidity, we need to control for the relationship between race and demographic variables. For

Table 4.1: A two-way crosstabulation table by SMM (absent or present) and race (White and Black).

		SMM		
Race		Absent	Present	Total
white	number	1086387	14113	1100500
	percentage	98.72%	1.28%	100%
Black	number	303531	8148	311679
	percentage	97.39%	2.61%	100%
Total	number	1389918	22261	1412179

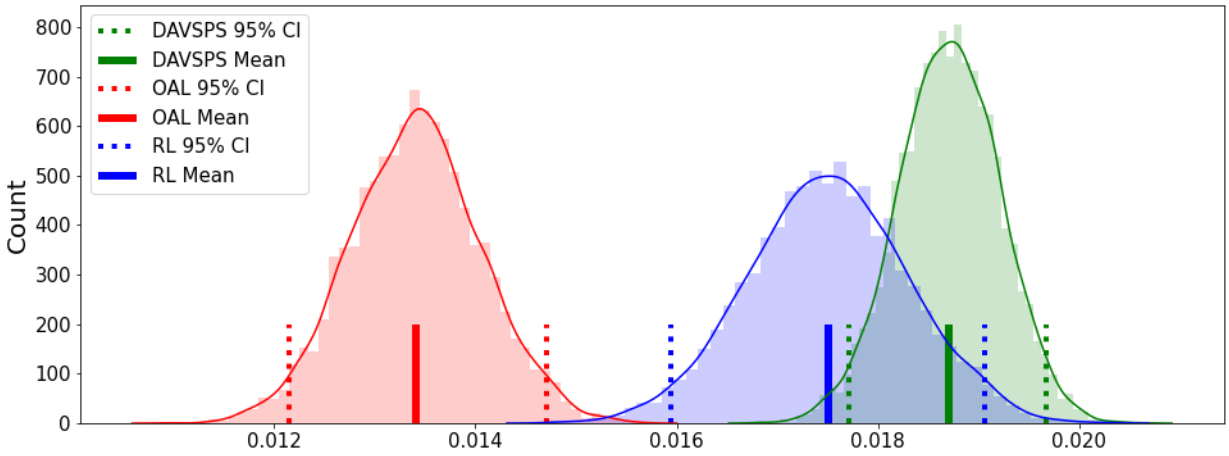

Figure 4.6: Racial disparities in SMM (Black-White) bootstrap distribution with 5000 bootstrap iterations.

Table 4.2: Covariate distribution grouped by race. Last three columns report percent of times each covariate was selected.

	Race		Percentage of Selected		
Covariates	White (n=1100500)	Black (n=311679)	DAVSPS	OAL	RL
Weight (discharge-level weight)	4.99 (0.00)	4.99 (0.00)	65.24	98.16	55.12
Age in years at admission	29.17 (5.58)	27.52 (5.98)	23.18	4.62	18.26
Total charges	18528.43 (16451.58)	21509.04 (20220.47)	5.98	23.61	8.12
Community income					
Quartile 1 (poorest)	21.60 (237659)	49.49 (154256)	82.14	33.84	58.78
Quartile 2	26.63 (293040)	23.46 (73111)	85.66	20.12	65.12
Quartile 3	26.99 (296984)	17.09 (53276)	78.90	35.64	68.62
Quartile 4 (wealthiest)	24.79 (272817)	9.96 (31036)	88.36	39.96	77.50
Type of insurance					
Medicaid	31.08 (342000)	63.73 (198621)	58.92	23.12	41.84
Private	63.43 (698086)	30.45 (94906)	70.48	58.88	50.12
Uninsured (self-pay/no charge)	1.54 (16898)	2.22 (6929)	76.96	30.92	55.76
Other (other public insurance)	3.95 (43516)	3.60 (11223)	62.42	48.22	42.10
Location of residence					
Large metropolitan	48.94 (538553)	65.23 (203317)	23.90	21.12	55.62
Small metropolitan	32.52 (357935)	26.19 (81616)	65.68	44.84	68.18
Micropolitan	10.85 (119352)	5.15 (16038)	21.84	26.18	63.56
Rural	7.69 (84660)	3.44 (10708)	15.66	30.18	76.22
Hospital teaching status					
Non-Teaching Hospital	35.83 (394357)	22.61 (70473)	33.62	21.82	20.16
Teaching Hospital	64.17 (706143)	77.39 (241206)	64.86	42.52	44.12
Region of hospital					
Northeast	17.18 (189074)	14.68 (45755)	96.72	52.86	95.16
Midwest or North Central	26.75 (294329)	19.22 (59895)	95.84	42.66	90.34
South	37.93 (417436)	57.99 (180746)	98.26	48.84	88.76
West	18.14 (199661)	8.11 (25283)	98.70	38.34	92.14
Control/ownership of hospital					
Government, nonfederal	10.12 (111372)	13.80 (43025)	67.56	23.16	84.92
Private, not profit	78.04 (858790)	72.30 (225352)	62.92	42.62	79.66
Private, invest own	11.84 (130338)	13.89 (43302)	94.66	48.22	90.10

Table 4.3: ATE estimates of racial disparities in SMM (Black-White) together with 95% confidence intervals.

Method	Racial disparities in SMM (Black-White)	95% Confidence Interval	
Direct calculate	1.332%		
OAL	1.341%	(1.215%,	1.471%)
RL	1.751%	(1.595%,	1.906%)
DAVSPS	1.871%	(1.771%,	1.967%)

some demographic variables such as community income and type of insurance coverage, there exist big differences between White and Black, which maybe leads to the severe bias when estimating the racial disparities in SMM. Therefore, we apply OAL, RL, and our method DAVSPS to this data set with 5000 bootstrap iterations to estimate the average racial disparities and 95% confidence intervals (Figure 4.6) conditional on the potential confounders. As shown in Table 4.3, if we do not control for potential confounders, we can observe a significant difference 1.33% (2.61% − 1.28%) in SMM between Whites and Blacks. The racial disparity discovered from our data set (from 2016-2018) is consistent with the conclusion in Fingar et al., 2018 from previous data (2006-2015). From Table 4.3, when OAL is used, ATE is estimated to be 1.341% with a 95% CI of $(1.215\%, 1.471\%)$; when RL is used, it is 1.751% with a 95% CI of $(1.595\%, 1.906\%)$; and when DAVSPS is used, it is 1.871% with a 95% CI of $(1.771\%, 1.967\%)$. Compared with the direct calculation without controlling for confounders, RL and DAVSPS discover a larger racial disparity in SMM between Blacks and Whites than OAL. The last three columns in Table 4.2 provides the percentage of times covariate selected in the model, which represents the importance in the prediction of SMM. Community income, type of insurance, region of hospital, and control/ ownership of hospital are relatively more frequently selected as predictors of SMM.

4.7 Summary

In this chapter, we present a novel deep adaptive variable selection propensity score (DAVSPS) based on deep representation learning and outcome adaptive group LASSO. Experimental results on simulated datasets show that DAVSPS provides more accurate estimation of average treatment effect with respect to bias and variance, and is more highly adaptable to complicated observational data. Although in this work, the deep adaptive variable selection is only used to estimate the propensity score, it can be connected to any causal inference

approach seamlessly to help select out the target variables by imposing different penalties obtained from the relationship of interest. Besides, the treatment effect may also depend on individuals, thus the treatment effect heterogeneity should be taken into account. In this case, our deep adaptive variable selection can also be applied to estimate the individual treatment effect due to the strong approximation properties of deep learning.

Chapter 5

Adversarial Learning for Estimating Treatment Effects in Basket Trials

5.1 Introduction

With the rapid development of next generation sequencing and comprehensive genomic profiling, genomic characterization informs treatment of a variety cancers. Some genetic mutations have been linked to multiple cancer types; for example BRCA1 abd BRCA2 are associated with an increased risk of breast, ovarian and pancreatic cancers (Mersch et al., 2015). Traditional clinical trials focusing on patients with a single cancer type are time-consuming, expensive, and frequently fail, so they are not sufficient for the development of genomic technologies. Patients are generally classified by their primary cancer and randomized controlled trials are conducted to create standard therapies for each cancer type. It is unrealistic to conduct separate clinical trials for each sub-population based on molecular

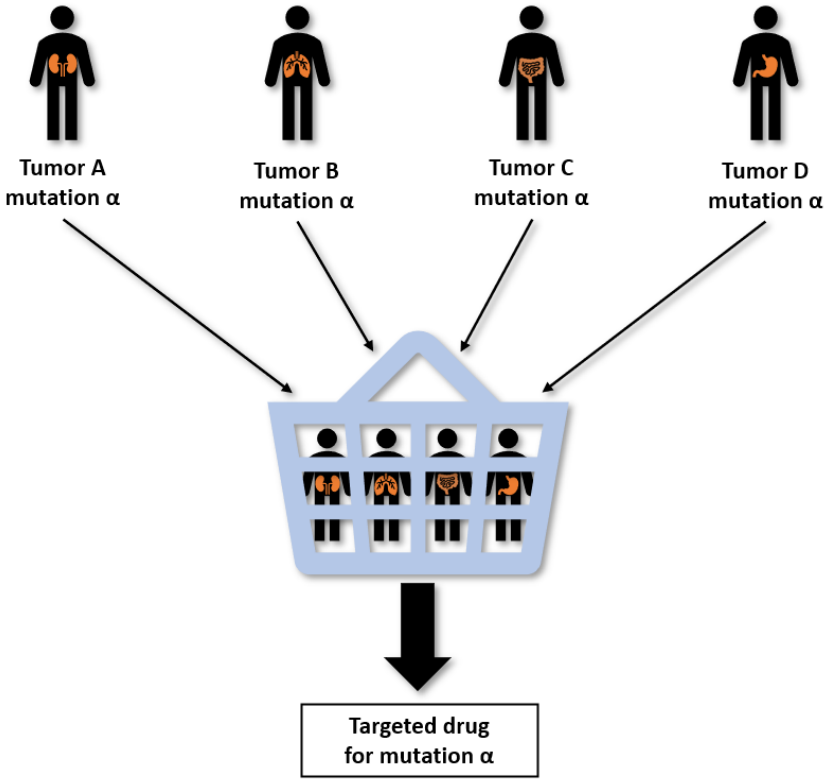

Figure 5.1: A basket trial is usually a non-randomized and single-arm trial.

subtypes or detailed classification of tumors (Hirakawa et al., 2018). Therefore, a new-style clinical trial protocol is in urgent demand in oncology.

A novel clinical design called basket trial has been developed based on the presence of a specific genomic mutation, irrespective of histology (Astsaturov, 2017; Simon, 2017; Tao et al., 2018). Unlike traditional clinical trials which test a drug against a specific cancer type, the core organizing principle of basket trials is a common genomic mutation. A basket trial is usually a non-randomized, single-arm trial so that all patients with the specified genomic mutation receive the same treatment. Treatment selection only depends on genomic mutation type, instead of tumor type. An example of basket trial is shown in Fig. 5.1, where

the term "basket" arises from each collection of patients sharing a particular mutation, and sub-studies for the same drug are conducted by tumor groups within the whole "basket". Patients enrolled in a basket trial are heterogeneous with respect to tumor type, histologic type, and patient characteristics, so the treatment effects are sensitive to population heterogeneity. Therefore, the absence of a control group becomes a limitation of treatment effect evaluation (Hirakawa et al., 2018). Ignoring the heterogeneity of tumors may lead to failure to detect treatment effects and the inability to produce scientifically reliable findings (Strzebonska & Waligora, 2019). Besides, focusing only on molecular therapy targeting single mutation without considering the complexity of tumor biology may introduce bias.

In this chapter, we apply causal inference models to basket trials. Estimating causal effects from observational data has become an appealing research direction owing to the availability of data and low budget requirements compared with randomized controlled trials. This chapter is the first to apply machine learning and causal inference to basket trials and explore the relationship between the traditional multiple treatments design and the basket trial design. In particular, we propose a multi-task adversarial learning (MTAL) method incorporating feature selection, multi-task representation learning and adversarial learning to remove selection bias (tumor type heterogeneity) introduced by confounders. Our method generates all potential outcomes for each unit across all tumor types, regardless of heterogeneity from different tumor types, so that the sample size may be increased in basket trials for rare tumor types, increasing statistical power. We also define targeted group treatment effects to better describe treatment effects among sub-groups in a basket trial. We present the practicality and advantages of our MTAL method for synthetic basket trials, evaluate the proposed estimator on the IHDP and News benchmark datasets, and demonstrate its superiority over state-of-the-art methods.

5.2 Related Work

The landscape for oncology clinical trials is changing dramatically due to the advent of genomic characterization. Among diverse master protocols (Park et al., 2019), a basket trial evaluates the treatment effect of a targeted therapy on patients with the same genomic mutation, regardless of tumor types. Bayesian hierarchical modeling has been proposed to adaptively borrow strength across cancer types to improve the statistical power of basket trials (Berry et al., 2013; Simon, 2017). To avoid inflated type I errors in Bayesian hierarchical modeling, calibrated Bayesian hierarchical modeling has been proposed to evaluate the treatment effect in basket trials (Y. Chu & Yuan, 2018). As an alternative to Bayesian hierarchical modeling, we will apply powerful machine learning tools to basket trials.

Embracing the rapid developments in machine learning, various treatment effect estimation methods for observational data have been proposed for causal inference. Balancing neural networks (BNNs) (F. Johansson et al., 2016) and counterfactual regression networks (CFRNET) (Shalit et al., 2017) have been proposed to balance covariate distributions across treatment and control groups by regarding the problem of counterfactual inference as a domain adaptation problem. These models may be extended to any number of treatments even with continuous parameters, as described in the perfect match (PM) approach (Schwab et al., 2018) and DRNets (Schwab et al., 2019). Li and Fu (S. Li & Fu, 2017b) regard counterfactual prediction as a classification problem and conduct matching based on balanced and nonlinear representations. GANITE (Yoon et al., 2018) uses Generative Adversarial Nets for individual treatment estimation. To the best of our knowledge, our model is the first to introduce machine learning and causal inference to the task of estimating treatment effects for basket trials.

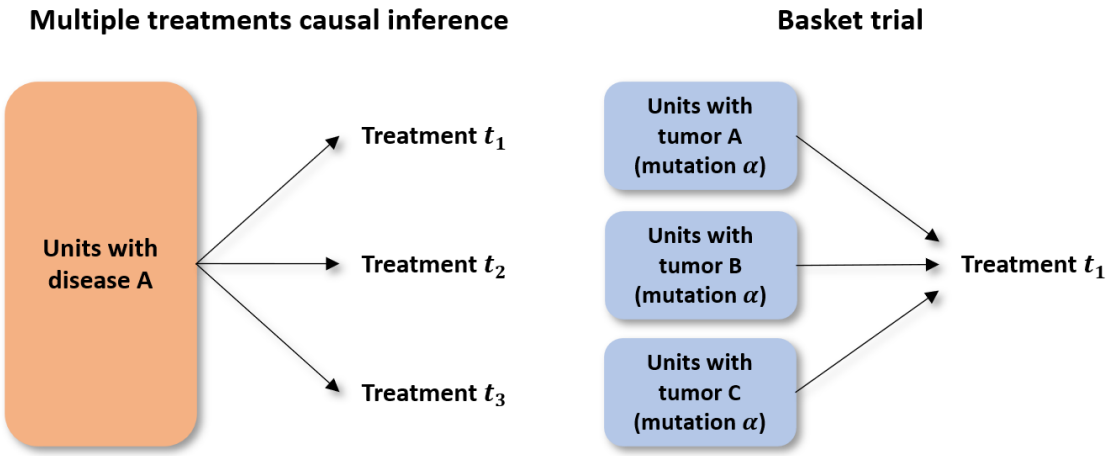

Figure 5.2: The relationship between conventional multiple treatment causal inference (top) and basket trial (bottom).

5.3 The Proposed Framework

5.3.1 Problem Statement

Clarification on the New Problem Setup. In traditional causal inference for observational data, researchers consider binary or multiple treatments for a set of experimental units. For example, a person who has cancer may be offered a choice between two treatment therapies. We can observe the outcome of the chosen treatment but not the potential outcome of the treatment not selected. It is impossible to see the potential outcomes of both therapies; one of the potential outcomes is always missing. The potential outcome framework (Rubin, 1974; Splawa-Neyman et al., 1990) aims to estimate unobserved potential outcomes and then calculate the treatment effect. The basket trial tests how well a new drug works in patients who have different types of cancer with the same mutation. Patients with the same genetic mutations are put in one "basket" and are divided into different subgroups according

to their cancer types. The differences in study design for potential outcome framework and basket trials are illustrated in Fig. 5.2. For the potential outcome framework, there is one population and several treatments, but in basket trials, there are several sub-populations and only one treatment.

Clarification on the Challenges. In the potential outcome framework (Rubin, 1974; Splawa-Neyman et al., 1990), we mainly face two challenges: missing unobserved counterfactual outcomes for each patient under alternative treatments not received and treatment selection bias. In basket trials, we have similar challenges: missing unobserved counterfactual outcomes for each patient under alternative cancers not contracted and cancer selection bias where the distributions of predictors differ among cancer types. In traditional causal inference for observational data, confounders are variables that affect both the treatment assignment and the outcome. Similarly, in basket trials, there still exist confounders that are associated with both cancer type and treatment outcome. These variables can explain why some patients with the same mutation have different types of cancer and can also influence treatment outcome. Due to the confounders, it is difficult in a basket trial to estimate the true treatment effects of a drug targeting the mutation of interest and the true treatment effect of drug for a specific type cancer type. If a significant treatment effect is not found, analysis of basket trial without appropriate causal inference cannot determine whether failure is due to uselessness of drug, particularity of a cancer type, or individual characteristics of patients.

Clarification on Treatment Effects Estimation. Because in this new setting, there is no control group, we do not care about the traditional treatment effects estimation between treated and control groups, e.g., average treatment effect (ATE) or individual treatment effect (ITE). We only focus on the counterfactual outcome inference problem, which is the core problem regardless of new setting or traditional treatment effects estimation setting. Most basket trials are conducted as single-arm trials without a control group and a primary

endpoint is given by an objective response rate (ORP). We proposed a new metric named targeted group response rate (TGOR) to better describe treatment effects in basket trials. TGOR describes the overall objective response rate for a given mutation or a given tumor type. It can evaluate the treatment effect of the drug for the whole targeted population with the same mutation and the effect of drug in the sub-population with different specific tumors type. Our MTAL method can help remove heterogeneity across tumor types when estimating the treatment effect for targeted mutation and remove heterogeneity across patients with the same tumor when estimating treatment effect for a targeted tumor.

Problem Setup. Suppose a basket trial is conducted as a one-arm phase II trial that tests how well a new drug works in patients who have different types of tumours but share the same genetic mutation. Data are available on an outcome for n participants. Let $t_i \in \{1, ..., k\}$ denote the type of tumour for unit i; $i = 1, ..., n$. The primary endpoint is the objective response rate (ORR) (Food, Administration, et al., 2007), determined by tumor assessments from radiological tests or physical examinations. Let y_t^i denote the potential outcome of the unit i $(i = 1, ..., n)$ with the tumour $t \in \{1, ..., k\}$. The observed outcome, called factual outcome is denoted by y^f and remaining unobserved potential outcomes are called counterfactual outcomes denoted by y^{cf}. The estimated potential, factual, and counterfactual outcomes are $\hat{y}$, $\hat{y}^f$, and $\hat{y}^{cf}$, respectively. Let $X \in \mathbb{R}^d$ denote all observed covariates. We extend the potential outcome framework (Rubin, 1974; Splawa-Neyman et al., 1990) to analysis of basket trial data. The following assumptions ensure that the treatment effects can be identified: **Consistency**: The potential outcome of treatment t is equal to the observed outcome if the actual treatment received is t. **Positivity**: For any value of X, treatment assignment is not deterministic, i.e.,$P(T = t|X = x) > 0$, for all t and x. **Ignorability**: Given covariates X, treatment assignment t is independent to the potential outcomes, i.e., $(y_1, y_0) \perp\!\!\!\perp t|X$.

5.3.2 Model Architecture

We propose a multi-task adversarial learning (MTAL) method to analyze basket trial data or observational data in basket trials, which can remove heterogeneity across tumor types when estimating the treatment effects for a targeted mutation, and remove heterogeneity among patients with one type of tumor when estimating the treatment effect for the targeted tumor and estimating the personalized treatment effect for individual patients. Our method is also useful for studying rare cancers and cancers with rare genetic mutations by inferring the outcome of existing patients with counterfactual cancers to increase sample size and statistical power.

Our method contains two major components: outcome generator and true or false discriminator (TF discriminator), as shown in Fig. 8.1. In the outcome generator, we use feature selection multi-task deep learning to estimate the potential outcomes for units across all tumor types. Because different types of tumor may have different predictor variables, which may be components of all observed covariates, a deep feature selection model including a sparse one-to-one layer between the input and the first hidden layer, and an elastic net regularization term throughout the fully-connected representation layers is an essential foundation for potential outcome estimation. Our TF discriminator can tell whether the outcome given the covariates and tumor type is factual outcome. At the beginning, the TF discriminator can easily find out which outcome is factual outcome and which one is our inferred counterfactual outcome under alternative tumor types not contracted by those patients. The outcome generator attempts to generate counterfactual outcomes in such a way that the TF discriminator cannot accurately determine which is the factual outcome. The two models are trained together in a zero-sum game and they are adversarial until the TF discriminator model is fooled by the generator, which means that the outcome generator

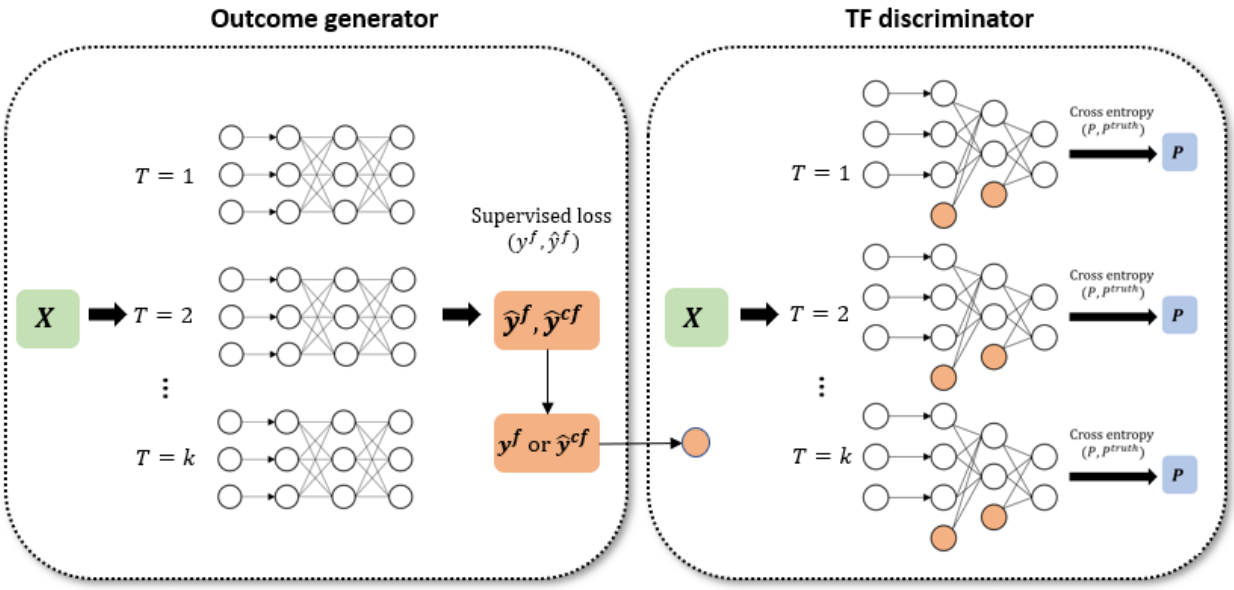

Figure 5.3: The framework of our multi-task adversarial learning net (MTAL).

is generating plausible examples. At this time, we have removed the tumor type selection bias and obtained all potential outcomes for each patient across all kinds of tumors.

Outcome Generator

Our goal is to correctly predict potential outcomes for each patient across all tumors types by a function $g : x \times t \rightarrow y$, which is parameterized by a feed-forward deep neural network structured by multiple hidden layers with non-linear activation functions. Deep neural networks can often dramatically increase prediction accuracy, describe complex relationships, and generate structured high-level representation of features when compared to parametric models. The function $g : x \times t \rightarrow y$ uses features x and tumor type t as inputs to predict potential outcomes. The output of g estimates potential outcomes across k tumors including

single factual outcome $\hat{y}^f$ and $k-1$ counterfactual outcomes $\hat{y}^{cf}$. The factual outcomes y^f are used to minimize the loss of prediction $\hat{y}^f$.

The function $g(x,t)$ maps the features and tumor type to the corresponding potential outcomes. However, when the dimension of the observed variables is high, there is a risk of losing the influence of t on $g(x,t)$ if the concatenation of x and t is treated as input (Shalit et al., 2017). To address this problem, $g(x,t)$ is partitioned into multiple head nets $g_t(x); t = \{1, ..., k\}$ corresponding to each cancer type. For each cancer type, there is one independent head net for predicting the potential outcome under this tumor. Each unit is used to update only the head net corresponding to the observed tumor type. We aim to minimize the mean squared error in predicting factual outcomes by $g(x,t)$, i.e., $\mathcal{L}_Y = \frac{1}{n}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2$, where $\hat{y}_i = g(x_i, t_i)$ denotes the inferred observed outcome of unit i corresponding to factual treatment t_i.

Due to the peculiarities of different tumor types, only a subset of all observed covariates might be predictors for each tumor type. To accommodate this, we add a deep feature selection net (Z. Chu et al., 2020; Y. Li et al., 2016) to each head net $g_t(x), t = \{1, 2, ..., k\}$, which enables variable selection in deep neural networks. This model takes advantage of deep structures to capture non-linearity and conveniently selects a subset of features of the data at the input level. In this model, the feature selection layer is a sparse one-to-one layer between the input and the first hidden layer. Feature selection at the input level can help select which variables are input into the neural network and used for representing pre-treatment variables, which makes the deep neutral network more interpretable.

In the feature selection layer, every input variable only connects to its corresponding node where the input variable is weighted. We use a 1-1 layer instead of a fully connected layer. To select input features, weights w in the feature selection layer and the following representation layers have to be sparse and only the features with nonzero weights are selected to enter the following layers. We first considered LASSO (Tibshirani, 1996) for this purpose. LASSO

is a penalized least squares method imposing the L_1-penalty on the regression coefficients by $\Re(w) = \|w\|_1$. However, for observational data with high dimensional variables, LASSO cannot remove enough variables before it saturates. To overcome this limitation, the elastic net (Zou & Hastie, 2005) is adopted in our model, which adds a quadratic term $\|w\|_2^2$ to the penalty i.e., $\Re(w) = \lambda\|w\|_2^2 + \alpha\|w\|_1$, where λ and α are tuning parameters. After combining the mean squared error in predicting factual outcomes and elastic net regularization term, we minimize the objective function in the outcome generator module:

$$\mathcal{L}_g = \frac{1}{n}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2 + \lambda\sum_{t=1}^{k}\sum_{s=1}^{S_t}\|w^{(s)}\|_2^2 + \alpha\sum_{t=1}^{k}\sum_{s=1}^{S_t}\|w^{(s)}\|_1, \tag{5.3.1}$$

where S_t is the number of deep feature selection layers for the t-th head net including the feature selection layer and the representation layers. The $w^{(s)}$ are the parameters of deep neural network in the s-th layer of outcome generator. The $\lambda \geq 0$ and $\alpha \geq 0$ are hyperparameters that not only control the trade-off between the regularization term and the following objective terms, but also controls the trade-off between smoothness and sparsity of the weights in the feature selection layer (Y. Li et al., 2016).

True or False Discriminator

The true or false (TF) discriminator is intended to remove tumor type bias and thus improve the prediction accuracy of potential outcomes inferred in the outcome generator for each unit across all types of tumors by adversarial learning. We define one TF discriminator as $\phi : x \times t \times (y^f \text{or } \hat{y}^{cf}) \to P$ where P is the TF discriminator's judgement, i.e., probability that this outcome for unit i given x and t is factual outcome, which is defined as:

$$P = \begin{cases} P(\text{TF judges } y^f \text{as factual outcome}|x,t) \text{if } t \text{ is factual type} \\ P(\text{TF judges } \hat{y}^{cf} \text{as factual outcome}|x,t) \text{if } t \text{ is not factual type.} \end{cases} \tag{5.3.2}$$

To improve the accuracy of prediction and avoid risk of losing the influence of tumor type t and potential outcomes (y^for $\hat{y}^{cf}$) due to high dimensional features x, we adopt the same architecture as outcome generator, which has multiple head nets for different tumor types. In each head net, a deep feature selection net is added to select the appropriate predictors for a each type of tumor. To improve the influence of $(y^f, \hat{y}^{cf})$ in the TF discriminator, we add (y^for $\hat{y}^{cf}$) into each layer after one on one feature selection layer and the dimension of each layer in TF discriminator decreases layer by layer.

The TF discriminator is a binary classification task, which puts one label (i.e., true or false factual outcome) on the vector concatenating the representation vector of x and potential outcome (y^for $\hat{y}^{cf}$) under each type of tumor head net, so the loss of discrimination is measured by the cross-entropy with truth probability where $P^{\text{truth}} = 1$ if y^f is input and $P^{\text{truth}} = 0$ if $\hat{y}^{cf}$ is input. In each iteration of training, we make sure to input the same number of units in each tumor type to ensure that there exist factual units in each head net. When there are several types of tumors, we face the imbalanced classification issue. If there are k types of tumor and n units in each tumor type are input into the model training procedure, then in one each head net, n units are factual outcomes and $(k-1)n$ units are inferred counterfactual outcomes. As k increases, the imbalance of factual outcome numbers and inferred counterfactual outcome numbers in each head net will aggravate. The weighted cross entropy is used to reduce this imbalance. Because inputs of TF discriminator are generated by the outcome generator $g(x,t)$, the weighted cross entropy of TF discriminator and elastic net are defined as:

$$\begin{aligned}\mathcal{L}_{\phi,g} = &- \frac{1}{n \times k} \sum_{t=1}^{k} \sum_{i=1}^{n} (w_0 p_{ti}^{\text{truth}} \log(p_{ti}) + w_1 (1 - p_{ti}^{\text{truth}}) \log(1 - p_{ti})) \\ &+ \lambda \sum_{t=1}^{k} \sum_{r=1}^{r_t} \|w^{(r)}\|_2^2 + \alpha \sum_{t=1}^{k} \sum_{r=1}^{r_t} \|w^{(r)}\|_1,\end{aligned} \tag{5.3.3}$$

where p_{ti}^{truth} is the probability that this input outcome for unit i under tumor t is the observed factual outcome or inferred outcome from generator module, i.e., 1 or 0, separately. P_{ti} is the probability judged by TF discriminator that this input outcome for unit i under tumor t is factual outcome. The w_1 and w_0 are the proportions of factual outcomes and counterfactual outcomes in total outcomes. Because during training, the same number of units in each tumor type are input, $w_1 = \frac{1}{k}$ and $w_0 = \frac{k-1}{k}$ in each head net. The number of deep feature selection layers for the t-th head net is denoted by r_t, and $w^{(r)}$ are the weights for the deep neural network in the r-th layer of the TF discriminator. $\lambda \geq 0$ and $\alpha \geq 0$ are the same as those in the outcome generator.

Adversarial Learning

Thus far, we have described an outcome generator to estimate potential outcomes for each unit across all types of tumor and a TF discriminator to determine if each potential outcome given unit's features under different tumor types is factual. In the initial iterations of the training algorithm, the outcome generator may generate potential outcomes that are very different from factual outcomes as determined by the TF discriminator. As the model is trained further, the TF discriminator may no longer be able to discriminate between the generated potential outcome and the factual outcome. At this point, we have attained all potential outcomes for each unit under all tumor types. The training procedure optimizing the outcome generator and TF discriminator uses the minimax decision rule:

$$\min_g \max_\phi \ (\mathcal{L}_g - \beta \mathcal{L}_{\phi,g}), \tag{5.3.4}$$

where β is a hyper-parameter controlling the trade-off between the outcome generator and discriminator. Compared to the deep regression task in the outcome generator, the TF discriminator is a relatively simple binary classification, which is easier to optimize. In every

optimization iteration, in order to get more accurate inferred potential outcomes to fool the discriminator based on the discriminator's current ability of telling which is factual outcome and which is counterfactual outcome, we can optimize $\min_g (\mathcal{L}_g - \beta\mathcal{L}_{\phi,g})$ several times after we optimize $\max_\phi (-\beta\mathcal{L}_{\phi,g})$ one time.

5.3.3 Targeted Group Analysis

The proposed MTAL method can generate all potential outcomes for each unit under all tumor types, which can help basket trials increase sample size and thus increase statistical power, and remove the influence of heterogeneity among different tumor types.

In basket trials, we must consider different configurations of effectiveness. For example, the drug may truly work for only one type of tumor due to the heterogeneity of tumors. Alternatively, it may actually work for all types of tumors, which means it works for the mutation regardless the type of tumor. Each of these configurations can lead to markedly different statistical properties (Cunanan et al., 2017). Therefore, we not only want to evaluate the treatment effect of the drug for the mutation (the whole population in study), but also want to evaluate the effect of drug for specific tumors (the sup-population in study). In addition, most basket trial are conducted as single-arm trials without a control group and a primary endpoint is given by an objective response rate (ORP). We propose a new metric named targeted group response rate (TGOR) to better describe treatment effects in basket trials. TGOR describes the overall objective response rate for a given mutation or a given tumor type, which is defined as:

$$\text{TGOR}_{\text{mutation}} = \frac{1}{n \times k} \sum_{t=1}^{k} \sum_{i=1}^{n} y^{ti} \quad \text{and} \quad \text{TGOR}_{\text{tumor}} = \frac{1}{n_c} \sum_{i=1}^{n_c} y^{i}. \tag{5.3.5}$$

where n is the number of patients with that mutation, and n_c is the the sub-sample who have that mutation and that specific cancer, a subset of mutation sample n.

Our MTAL method can help remove heterogeneity across tumor types in basket trials when estimating the treatment effect for targeted mutation $\text{TGOR}_{\text{mutation}}$, remove heterogeneity across patient with one type of tumor when estimating the treatment effect for a targeted tumor $\text{TGOR}_{\text{tumor}}$ and estimate the individualized treatment effects for an individual patient. Our method is also useful for studying rare cancers and cancers with rare genetic mutations by borrowing strength from more common cancers sharing the same mutation to infer the potential outcomes of existing patients under counterfactual cancer to increase sample size and statistical power.

5.4 Experiments and Analysis

Because our method is the first model for estimating treatment effects for basket trials, no other baseline methods are available. To evaluate our model's estimation performance, we modify our model (by removing the deep feature selection module) to coordinate the settings in traditional treatment effect estimation (binary and multiple treatments) and use benchmarks (*IHDP* and *News*) to demonstrate our estimation performance on the counterfactual outcomes. We also use one synthetic basket trial dataset to demonstrate our method's stability in basket trial.

5.4.1 Performance Evaluation on Estimating the Counterfactual Outcomes

We coordinate our model to be compatible with the settings in traditional treatment effect estimation and conduct experiments on binary treatment benchmark *IHDP* and multiple treatment benchmark *News* with $2, 4, 8,$ and 16 treatment options.

Datasets. *IHDP.* The IHDP data set is a commonly adopted benchmark collected by the Infant Health and Development Program (Brooks-Gunn et al., 1992). These data are generated based on a randomized controlled trial where intensive high-quality care and specialist home visits were provided to low-birth-weight and premature infants. There are a total of 25 pre-treatment covariates and 747 units, including 608 control units and 139 treatment units. The outcome is the infants' cognitive test scores which can be simulated using the pre-treatment covariates and the treatment assignment information through the NPCI package [1]. In the IHDP data set, a biased subset of the treatment group is removed to simulate the selection bias (Shalit et al., 2017). We repeat these procedures 1000 times so as to conduct evaluations of uncertainty of estimates. *News.* The News data set was first introduced for binary treatments counterfactual estimation by (F. Johansson et al., 2016) and extended to multiple treatment benchmarks by (Schwab et al., 2018). The News benchmark includes 5000 randomly sampled news articles from the NY Times corpus and the opinions of a media consumer exposed to multiple news items. Each unit is a news item and the features are word counts. The outcome represents the reader's opinion of the news item. The treatment options are various devices used for viewing news items, e.g. smartphone, tablet, desktop, television or others. We use the extended multiple treatment data set according to the specification in (Schwab et al., 2018). A topic model is trained on the whole NY Times corpus to model consumers preferences towards reading given news items on specific devices, where $k+1$ centroids are randomly picked in the topic space where k represents the number of treatment options and the remaining is the control group. We use four different variants of this data set with 5000 units, 2870 features and $k = 2, 4, 8,$ and 16 treatment options.

Baselines. To evaluate the accuracy of our model's treatment effects estimates, we compare our multi-task adversarial learning net model with the following methods: k-nearest

[1] https://github.com/vdorie/npci

Table 5.1: Hyperparameters and ranges.

Hyperparameter	IHDP	News
β	0, $\{10^k\}_{k=-6}^{2}$	0, $\{10^k\}_{k=-6}^{2}$
λ, α	0, $\{10^k\}_{k=-6}^{-1}$	0, $\{10^k\}_{k=-6}^{-1}$
No. of hidden layers without feature selection layer	2, 3, 4, 5	2, 3, 4, 5
Dim. of first hidden layer	50, 100, 150	50, 100, 150
Batch size	50×2, 75×2, 100×2	$30 \times k$, $40 \times k$, $50 \times k$

neighbor (kNN) (D. E. Ho et al., 2007), Causal forests (CF) (Wager & Athey, 2018b),Random forests (RF) (Breiman, 2001), Bayesian additive regression trees (BART) (Chipman et al., 2010), Generative adversarial nets for inference of ITE (GANITE) (Yoon et al., 2018), Propensity score matching with logistic regression (PSM) (D. E. Ho et al., 2011), Treatment-agnostic representation network (TARNET) (Shalit et al., 2017), Counterfactual regression network ($\text{CFRNET}_{\text{wass}}$) (Shalit et al., 2017), local similarity preserved individual treatment effect estimation method (SITE) (Yao et al., 2018), and Perfect match (PM) (Schwab et al., 2018).

Parameter Settings. To ensure a fair comparison, we follow a standardised implementation [2] to realize hyperparameter optimisation for IHDP and News data sets and extend the original binary treatment effect estimation methods to multiple treatments according to specifications in (Schwab et al., 2018). The hyperparameters of our method are chosen based on performance on the validation data set, and the searching range as shown in Table 6.5. MTAL is implemented using standard feed-forward neural networks with Dropout (Srivastava et al., 2014) and the ReLU activation function. Adam (Kingma & Ba, 2014) is adopted to optimize the objective function.

[2] https://github.com/d909b/perfect_match

Table 5.2: Performance on IHDP and News data sets. We present mean ± standard deviation of $\sqrt{\epsilon_{\text{PEHE}}}$ and $\sqrt{\epsilon_{\text{mPEHE}}}$ on the test sets.

	IHDP	News-2	News-4	News-8	News-16
Method	$\sqrt{\epsilon_{\text{PEHE}}}$	$\sqrt{\epsilon_{\text{PEHE}}}$	$\sqrt{\epsilon_{\text{mPEHE}}}$	$\sqrt{\epsilon_{\text{mPEHE}}}$	$\sqrt{\epsilon_{\text{mPEHE}}}$
kNN	6.66 ± 6.89	18.14 ± 1.64	27.92 ± 2.44	26.20 ± 2.18	27.64 ± 2.40
PSM	2.70 ± 3.85	17.40 ± 1.30	37.26 ± 2.28	30.50 ± 1.70	28.17 ± 2.02
RF	4.54 ± 7.09	17.39 ± 1.24	26.59 ± 3.02	23.77 ± 2.14	26.13 ± 2.48
CF	4.47 ± 6.55	17.59 ± 1.63	23.86 ± 2.50	22.56 ± 2.32	21.45 ± 2.23
BART	2.57 ± 3.97	18.53 ± 2.02	26.41 ± 3.10	25.78 ± 2.66	27.45 ± 2.84
GANITE	5.79 ± 8.35	18.28 ± 1.66	24.50 ± 2.27	23.58 ± 2.48	25.12 ± 3.53
PD	5.14 ± 6.55	17.52 ± 1.62	20.88 ± 3.24	21.19 ± 2.29	22.28 ± 2.25
TARNET	1.32 ± 1.61	17.17 ± 1.25	23.40 ± 2.20	22.39 ± 2.32	21.19 ± 2.01
$\text{CFRNET}_{\text{wass}}$	0.88 ± 1.25	16.93 ± 1.12	22.65 ± 1.97	21.64 ± 1.82	20.87 ± 1.46
PM	0.84 ± 0.61	16.76 ± 1.26	21.58 ± 2.58	20.76 ± 1.86	20.24 ± 1.46
SITE	$\mathbf{0.81 \pm 1.22}$	16.87 ± 1.34	22.33 ± 2.08	21.84 ± 2.21	20.88 ± 1.52
MTAL	1.06 ± 1.28	$\mathbf{16.58 \pm 1.20}$	$\mathbf{20.42 \pm 1.88}$	$\mathbf{19.98 \pm 2.01}$	$\mathbf{19.32 \pm 1.76}$

Table 5.3: Performance on IHDP and News data sets of MTAL and competing methods.

	IHDP	News-2	News-4	News-8	News-16
Method	ϵ_{ATE}	ϵ_{ATE}	ϵ_{mATE}	ϵ_{mATE}	ϵ_{mATE}
kNN	3.19 ± 1.49	7.83 ± 2.55	19.40 ± 3.12	15.11 ± 2.34	17.27 ± 2.10
PSM	0.49 ± 0.81	4.89 ± 2.39	30.19 ± 2.47	22.09 ± 1.98	18.81 ± 1.74
RF	0.64 ± 1.25	5.50 ± 1.20	18.03 ± 3.18	12.40 ± 2.29	15.91 ± 2.00
CF	0.65 ± 1.24	4.02 ± 1.33	13.54 ± 2.48	9.70 ± 1.91	8.37 ± 1.76
BART	0.53 ± 1.02	5.40 ± 1.53	17.14 ± 3.51	14.80 ± 2.56	17.50 ± 2.49
GANITE	0.98 ± 1.90	4.65 ± 2.12	13.84 ± 2.69	11.20 ± 2.84	13.20 ± 3.28
PD	1.37 ± 1.65	4.69 ± 3.17	8.47 ± 4.51	7.29 ± 2.97	10.65 ± 2.22
TARNET	0.24 ± 0.29	4.58 ± 1.29	13.63 ± 2.18	9.38 ± 1.92	8.30 ± 1.66
$\text{CFRNET}_{\text{wass}}$	0.20 ± 0.24	4.54 ± 1.48	12.96 ± 1.69	8.79 ± 1.68	8.05 ± 1.40
PM	0.24 ± 0.20	3.99 ± 1.01	10.04 ± 2.71	6.51 ± 1.66	5.76 ± 1.33
SITE	$\mathbf{0.18 \pm 0.23}$	4.53 ± 1.32	12.75 ± 1.88	9.01 ± 1.86	8.63 ± 1.41
MTAL	0.34 ± 0.28	$\mathbf{3.88 \pm 1.11}$	$\mathbf{8.01 \pm 1.43}$	$\mathbf{5.97 \pm 1.58}$	$\mathbf{5.12 \pm 1.31}$

Results and Analysis. We adopt two commonly used evaluation metrics. The first one is the error in average treatment effect (ATE) estimation defined as $\epsilon_{\text{ATE}} = |\text{ATE} - \widehat{\text{ATE}}|$,

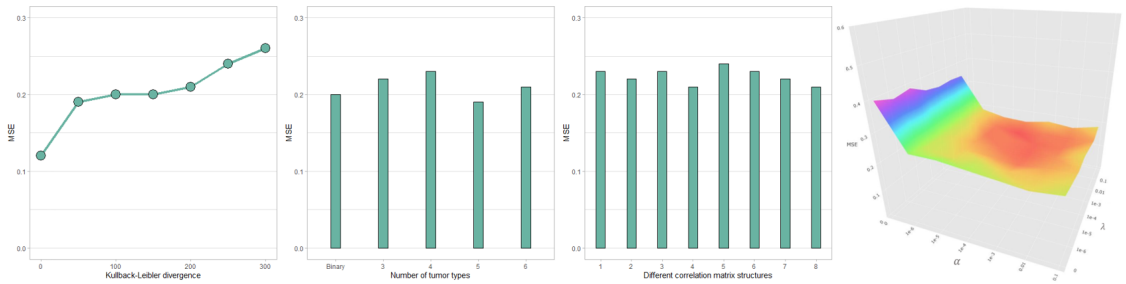

Figure 5.4: The results for synthetic basket trial data sets.

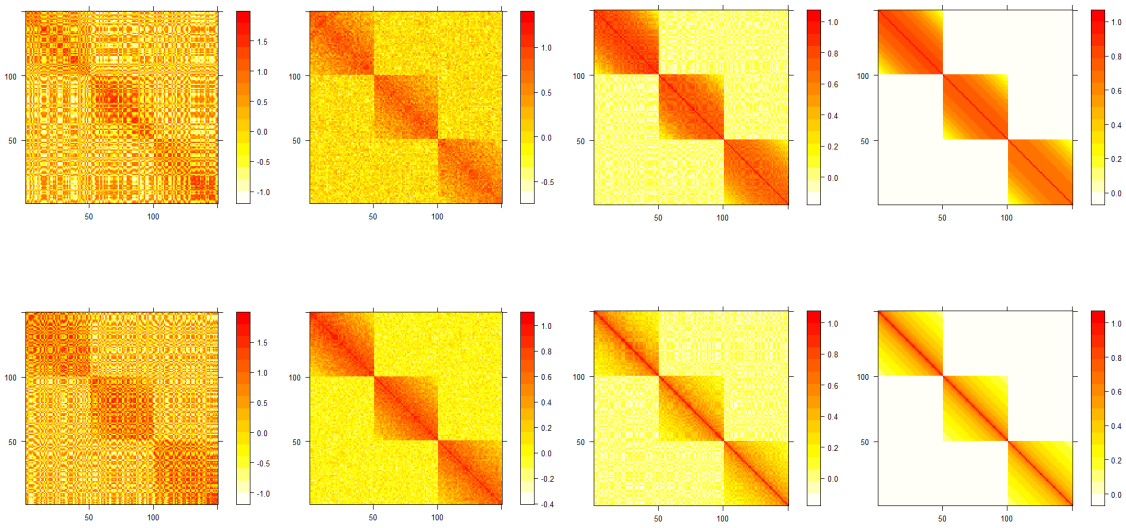

Figure 5.5: Different covariates correlation structures.

where $\text{ATE} = \frac{1}{n}\sum_{i=1}^{n}(Y_1^i - Y_0^i)$ and $\widehat{\text{ATE}}$ is an estimated ATE. The second one is the error of expected precision in estimation of heterogeneous effect (PEHE) (Hill, 2011), which is defined as $\epsilon_{\text{PEHE}} = \frac{1}{n}\sum_{i=1}^{n}(\text{ITE}_i - \widehat{\text{ITE}}_i)^2$, where $\text{ITE}_i = Y_1^i - Y_0^i$ and $\widehat{\text{ITE}}_i$ is an estimated ITE for unit i. In addition, for multiple treatment evaluations, we follow the definitions in (Schwab et al., 2018), where both ϵ_{PEHE} and ϵ_{ATE} can be extended to multiple treatments by averaging PEHE and ATE between every possible pair of treatments. They are defined as $\epsilon_{\text{mPEHE}} = \frac{1}{\binom{k}{2}}\sum_{t=1}^{k}\sum_{j=1}^{t}\epsilon_{\text{PEHE},t,j}$ and $\epsilon_{\text{mATE}} = \frac{1}{\binom{k}{2}}\sum_{t=1}^{k}\sum_{j=1}^{t}\epsilon_{\text{ATE},t,j}$. Table 5.2 and table 5.3 show the performance of our method and baseline methods on the IHDP and

News data sets. MTAL achieves the best performance with respect to PEHE and ATE for News data sets with different numbers of treatment options. For the IHDP data set, MTAL still has competitive performance when compared to the best baseline methods with respect to PEHE and ATE. The results on these two benchmarks for conventional binary and multiple treatments effects estimation can demonstrate that our method is capable of precisely estimating the treatment effects.

5.4.2 Synthetic Basket Trial Data Set

Dataset. To evaluate of our model's performance for basket trials, we simulate one synthetic data set which mimics the characteristics of a basket trial. Because different types of tumors may have different predictor variables, which may be subset of all observed covariates, we use different subsets of the observable covariates to generate the outcomes for different tumor types. To mimic the real situation further, we consider different covariance matrices in the covariates simulation. For example, the covariates predicting outcomes in each tumor type are taken to have stronger correlations than covariates predicting outcomes for other tumor types.

We generate a set of synthetic data sets which reflects the complexity of observational medical records data. The sample size for tumor type k is n_k, where $k = 1, 2, ..., K$. So, the total sample size is $n = \sum_{k=1}^{K} n_k$ units. The predictor variables for tumor type k are $x_k \in \mathbb{R}^d$. The potential outcomes y_k for tumor type k are generated as $y_k|x_k \sim \cos\left((w_k^\intercal x_k + n)^2\right)$, where $w_k \sim Uniform((-1, 1)^{d\times 1})$. The vector of all observed covariates $x = (x_1^\intercal, x_2^\intercal, ..., x_K^\intercal)^\intercal$ is sampled from a multivariate normal distribution with mean μ_k and different random positive definite covariance matrices Σ. By varying the value of μ_k, data with different levels of selection bias are generated (Yao et al., 2018; Yoon et al., 2018). Let D be the diagonal matrix with the square roots of the diagonal entries of Σ on its diagonal, i.e.,

$D = \sqrt{diag(\sigma)}$, then the correlation matrix is given by $R = D^{-1}\Sigma D^{-1}$. We simulate correlation matrix to better explain the relationship of covariates among and within different tumor types, instead of directly simulating covariates matrix. We use algorithm 3 in (Hardin et al., 2013) to simulate positive definite correlation matrices consisting of different within tumor type correlations and between tumor type correlations. Our correlation matrices are based on the hub correlation structure which has a known correlation between a hub variable and each of remaining variables (Langfelder et al., 2008; B. Zhang & Horvath, 2005). Each variable in a tumor type is correlated with the hub-variable with decreasing strength from specified maximum correlation to minimum correlation and different tumor types are generated independently or with weaker correlation among tumor types. Defining the first variable as the hub, for the ith variable ($i = 2, 3, ..., d$), the correlation between it and the hub-variable in one tumor type is given by $R_{i,1} = \rho_{\max} - \left(\frac{i-2}{d-2}\right)^{\gamma} (\rho_{\max} - \rho_{\min})$, where $\rho_{\max}$ and $\rho_{\min}$ are specified maximum and minimum correlations, and the rate γ controls rate at which correlations decay.

After specifying the relationship between the hub variable and remaining variables in one tumor type, we use Toeplitz structure to fill out the remainder of the hub correlation matrix and get the hub-Toeplitz correlation matrix R_k for tumor type k. Here, R is the $d \times d$ matrix having the blocks $R_1, R_2, ..., R_K$ along the diagonal and zeros at off-diagonal elements. This yields a correlation matrix with nonzero correlations within tumor types and zero correlation among tumor types. The amount of correlation among tumor types which can be added to the positive-definite correlation matrix R is determined by its smallest eigenvalue.

Results and Analysis. The mean squared error is used as the performance metric to evaluate our model under the settings of binary or multiple tumor types, different selection bias, and different correlation matrix for observed covariates. The mean squared error is defined as MSE $= \frac{1}{N \times K} \sum_{i=1}^{N} \sum_{k=1}^{K} \left(y_k(x_i) - \hat{y}_k(x_i)\right)^2$, where $y_k(x_i)$ and $\hat{y}_k(x_i)$ are factual and estimated outcomes for unit i with features x_i and tumor type k, respectively.

We simulate 5 data sets with $2, 3, 4, 5,$ and 6 tumor types, separately. From the second figure in Fig. 5.4, our MTAL performs relatively steadily for binary and multiple tumor types. To evaluate the performance with respect to selection bias, Kullback-Leibler divergence is adopted to quantify selection bias among different tumor types. Here, we use the data sets with binary tumor types. The all observed covariates in two tumor types are generated by multivariate normal distribution with mean 0 and different mean μ_1 for the remaining tumor type, so different values of μ_1 represent different Kullback-Leibler divergences; i.e., selection bias between two tumor types. From the first figure in Fig. 5.4, for the MTAL method, MSE increases modestly with increasing selection bias. To evaluate the sensitivity of the MTAL method to the correlation structure of the covariates, we generate 8 different correlation matrices with different levels of correlation for variables within each tumor type and among different tumor types in Fig. 5.5. From the third figure in Fig. 5.4, we find that the MSE owing to the feature selection layers in our MTAL method are not sensitive to the structure of the correlation matrices. In addition, from the fourth figure in Fig. 5.4, the performance of our model, with respect to MSE, is significantly improved compared with the models without L_1 and L_2 penalties. Also, the overall performance on different combinations of hyperparameters of L_1 and L_2 penalties is stable over a large range of tuning parameter values, which confirms the effectiveness and robustness of deep feature selection in our MTAL method.

5.5 Summary

In this chapter, we propose a multi-tasks adversarial learning (MTAL) method by incorporating feature selection multi-task deep learning and adversarial learning to remove heterogeneity of tumor types in basket trials. To the best of our knowledge, our model is the first work introducing machine learning and causal inference to the task of analyzing basket trial

data. It not only improves the basket trial analysis, but also has its superiority over state-of-the-art methods in estimating multiple treatment effects for observational data. In future work, we will follow this direction to apply causal inference models and machine learning methods into more medical practical applications, such as umbrella and platform trials.

Chapter 6

Graph Infomax Adversarial Learning for Treatment Effect Estimation with Networked Observational Data

6.1 Introduction

A further understanding of cause and effect beyond observational data is critical across many domains including statistics, computer science, education, public policy, economics, and health care. Although randomized controlled trials (RCT) are usually considered as the gold standard for causal inference, estimating causal effects from observational data has received growing attention owing to the increasing availability of data and the low costs compared to RCT.

When estimating treatment effects from observational data, we face two major issues, i.e., missing counterfactual outcomes and treatment selection bias. The foremost challenge

for solving these two issues is the existence of confounders, which are the variables that affect both treatment assignment and outcome. Unlike RCT, treatments are typically not assigned at random in observational data. Due to the confounders, subjects would have a preference for a certain treatment option, which leads to a bias of the distribution for the confounders among different treatment options. This phenomenon exacerbates the difficulty of counterfactual outcome estimation. For most of existing methods (A. M. Alaa & van der Schaar, 2017; Z. Chu et al., 2020; Hill, 2011; S. Li & Fu, 2017a; Shalit et al., 2017; Wager & Athey, 2018b; Yao et al., 2018; Yao, Li, Li, Xue, et al., 2019), the *strong ignorability* assumption is the most important prerequisite. Given covariates, it assumes that the treatment assignment is independent of the potential outcomes and for any value of covariates, treatment assignment is not deterministic. Strong ignorability is also known as the *no unmeasured confounders* assumption. This assumption requires that all the confounders be observed and sufficient to characterize the treatment assignment mechanism. Moreover, strong ignorability is a sufficient condition for the individual treatment effect (ITE) function to be identifiable (G. W. Imbens & Wooldridge, 2009).

However, due to the fact that identifying and collecting all of the confounders is impossible in practice, as well as the existence of hidden confounders, the strong ignorability assumption is usually untenable. By leveraging big data, it becomes possible to find a proxy for the hidden confounders. Network information, which serves as an efficient structured representation of non-regular data, is ubiquitous in the real world. Advanced by the powerful representation capabilities of various graph neural networks, networked data has recently received increasing attention (Kipf & Welling, 2016; Velickovic et al., 2019; Veličković et al., 2017). Besides, it can be used to help recognize the patterns of hidden confounders. A network deconfounder (Guo et al., 2019) is proposed to recognize hidden confounders by combining the graph convolutional networks (Kipf & Welling, 2016) and counterfactual regression (Shalit et al., 2017).

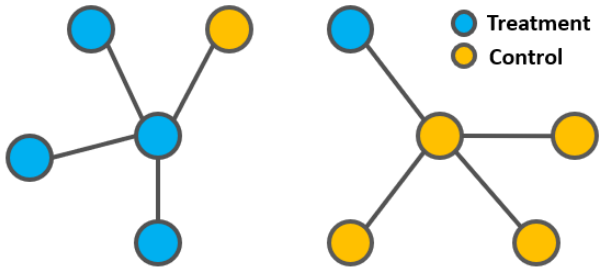

Figure 6.1: Example of the imbalance of network structure.

The networked observational data consists of two components, node features and network structures. Due to the confounding bias in causal inference problem, the imbalance not only exists in distributions of feature variables in treatment and control groups but also in network structures. For example, in social networks, the links are more likely to appear among more similar people, so the subjects are more likely to follow other subjects in the same group as shown in Fig. 6.1, which will aggravate the imbalance in the representation space learned by graph neural networks. Fig. 6.2 shows the existence of imbalanced network structures in the benchmarks of causal inference with networked data (*BlogCatalog* and *Flickr*). Unlike the networked data in traditional graph learning tasks, such as node classification and link detection, the networked data under the causal inference problem has its particularity, i.e., imbalanced network structure. For most existing work on networked observational data, they did not consider this peculiarity of graph structure under causal inference settings. Directly applying graph neural networks designed for traditional graph learning tasks cannot capture all of the information from imbalanced networked data.

To fully exploit the information in the networked data with the imbalanced network structure, we propose a Graph Infomax Adversarial Learning method (GIAL) to estimate the treatment effects from networked observational data. In our model, structure mutual

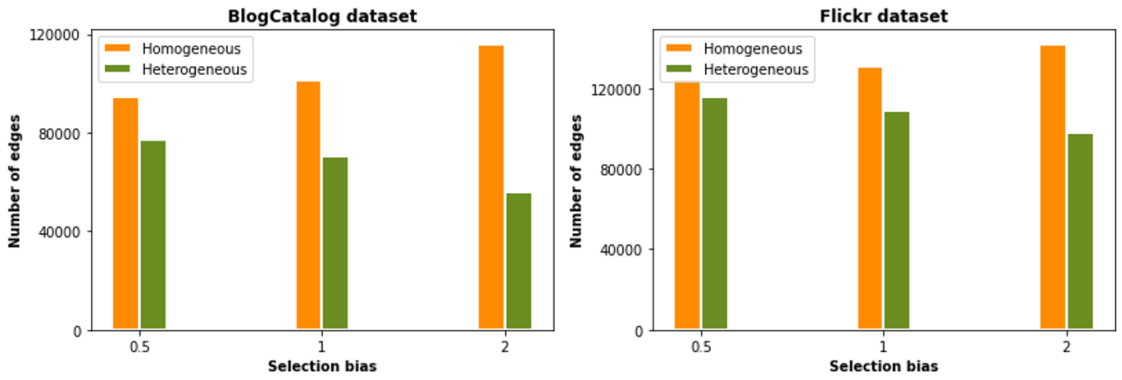

Figure 6.2: In the benchmarks of causal inference with networked data (*BlogCatalog* and *Flickr*), the homogeneous edges are consistently greater than heterogeneous edges for both datasets. Besides, as the selection bias increases, the difference between homogeneous and heterogeneous edges gets larger.

information is maximized to help graph neural networks to extract a representation space, which best represents observed and hidden confounders from the networked data with the imbalanced structure. Also, adversarial learning is applied to balance the learned representation distributions of treatment and control groups and to generate the potential outcomes for each unit across two groups. Overall, GIAL can make full use of network structure to recognize patterns of hidden confounders, which has been validated by extensive experiments on benchmark datasets.

We organize the rest of this chapter as follows. Technical backgrounds including notations and assumptions are introduced in Section 2. Our proposed framework is presented in Section 3. In Section 4, experiments on networked observational data are provided. Section 5 reviews related work. Section 6 concludes the chapter.

6.2 Background

Suppose that the observational data contain n units and that each unit received one of two or more treatments. Let t_i denote the treatment assignment for unit i; $i = 1, ..., n$. For binary treatments, $t_i = 1$ is for the treatment group, and $t_i = 0$ for the control group. The outcome for unit i is denoted by Y_t^i when treatment t is applied to unit i; that is, Y_1^i is the potential outcome of unit i in the treatment group and Y_0^i is the potential outcome of unit i in the control group. For observational data, only one of the potential outcomes is observed according to the actual treatment assignment of unit i. The observed outcome is called the factual outcome, and the remaining unobserved potential outcomes are called counterfactual outcomes. Let $X \in \mathbb{R}^d$ denote all observed variables of a unit.

Let $\mathcal{G}(\mathcal{V}, \mathcal{E})$ denote an undirected graph, where $\mathcal{V}$ represents n nodes in $\mathcal{G}$ and $\mathcal{E}$ is a set of edges between nodes. According to the adjacency relationships in $\mathcal{E}$, the corresponding adjacent matrix $A \in \mathbb{R}^{n \times n}$ of the graph $\mathcal{G}$ can be defined as follows. If $(v_i, v_j) \in \mathcal{E}$, $A_{ij} = 1$, otherwise $A_{ij} = 0$. When edges have different weights, A_{ij} can be assigned to a real value.

In this chapter, we explore the observational data as networks. In particular, the graph $\mathcal{G}$ is the networked observational data. Every node in $\mathcal{V}$ is one unit in observational data, an edge in $\mathcal{V}$ describes the relationship between a pair of units, and adjacent matrix A represents the whole network structure. Therefore, the observational data can be denoted as $(\{x_i, t_i, y_i\}_{i=1}^n, A)$. We follow the potential outcome framework for estimating treatment effects (Rubin, 1974). The individual treatment effect (ITE) for unit i is the difference between the potential treated and control outcomes, which is defined as: $\text{ITE}_i = Y_1^i - Y_0^i, \quad (i = 1, ..., n)$. The average treatment effect (ATE) is the difference between the mean potential treated and control outcomes, which is defined as $\text{ATE} = \frac{1}{n} \sum_{i=1}^n (Y_1^i - Y_0^i), \quad (i = 1, ..., n)$. The success of the potential outcome framework is based on the strong ignorability

assumption, which ensures that the treatment effect can be identified (G. W. Imbens & Rubin, 2015b; Yao et al., 2020).

Assumption 6.2.1. ***Strong Ignorability****: Given covariates X, treatment assignment T is independent of the potential outcomes, i.e., $(Y_1, Y_0) \perp\!\!\!\perp T|X$ and for any value of X, treatment assignment is not deterministic, i.e.,$P(T = t|X = x) > 0$, for all t and x.*

In our model, we relax the strong ignorability and allow the existence of hidden confounders. We aim to use network structure information to recognize the hidden confounders and then estimate treatment effects based on the learned confounder representations.

6.3 The Proposed Framework

6.3.1 Motivation

The foremost challenge of causal inference from observational data is how to recognize hidden confounders. Recently, leveraging the powerful representation capabilities of various graph neural networks, network structures can be utilized to help recognize the patterns of hidden confounders in networked observational data.

Due to the particularity of the causal inference problem, the networked data in causal inference is different from that in traditional graph learning tasks such as node classification and link detection. As network information is incorporated into the model, we face a new imbalance issue,i.e., imbalance of network structure in addition to the imbalance of observed covariate distributions. A link has a larger probability of appearing between two more similar people. It implies that one unit is more likely to be connected to other units in the same group (treatment or control). Therefore, directly applying traditional graph learning methods to learn the representation of networked data could not fully exploit the useful information for causal inference. It is essential to design new methods that can capture the

representation of hidden confounders, implied from the imbalanced network structure and observed confounders existed in the covariates simultaneously.

To solve this problem, we propose the Graph Infomax Adversarial Learning method (GIAL) to estimate the treatment effects from the networked observational data, which can recognize patterns of hidden confounders from imbalanced network structure.

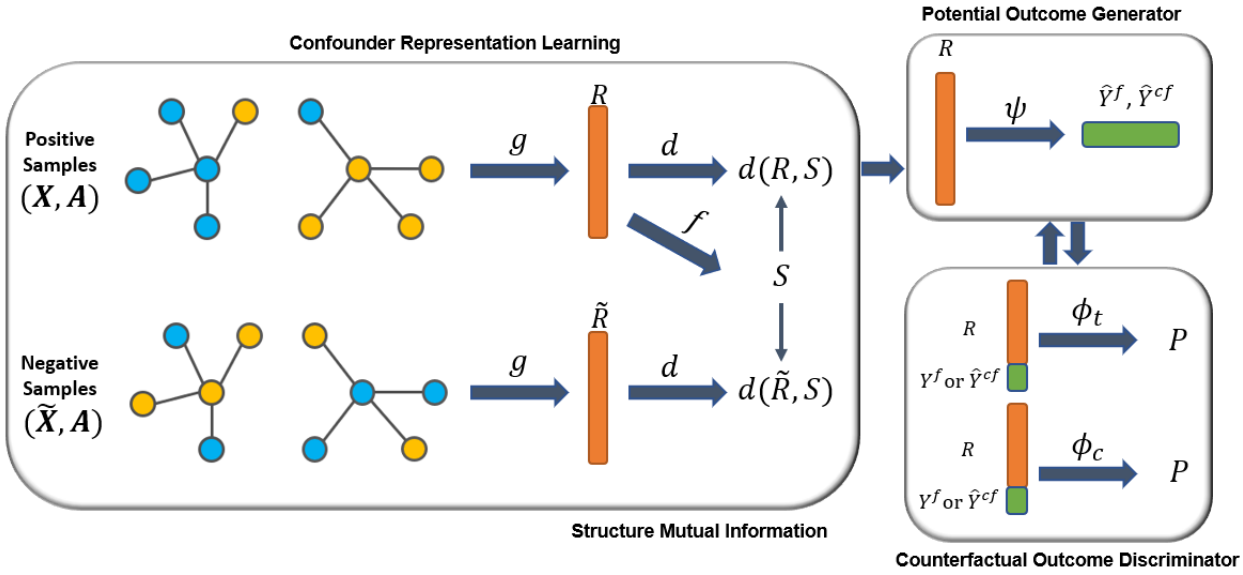

Figure 6.3: Framework of our Graph Infomax Adversarial Learning method (GIAL).

6.3.2 Model Architecture

As shown in Fig. 7.1, our GIAL consists of four main components, i.e., confounder representation learning, structure mutual information maximization, potential outcome generator, and counterfactual outcome discriminator. Firstly, we utilize the graph neural network and structure mutual information to learn the representations of hidden confounders and observed confounders, by mapping the feature covariates and network structure simultaneously into a representation space. Then the potential outcome generator is applied to infer the potential outcomes of units across treatment and control groups based on the learned representation

space and treatment assignment. At the same time, the counterfactual outcome discriminator is incorporated to remove the imbalance in the learned representations of treatment and control groups, and thus it improves the prediction accuracy of potential outcomes inferred in the outcome generator by playing a minimax game. In the following, we present the details of each component.

Confounder Representation Learning. Based on the graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$, our goal is to learn the representation of confounders by a function $g : X \times A \rightarrow R, R \in \mathbb{R}^d$, which is parameterized by a graph neural network. For node i, this function maps the feature covariates of node i and the adjacency matrix, which encodes the network structure, into a d-dimensional representation space $R = \{r_1, r_2, ..., r_n\}$, in order to represent the confounders of node i. To better capture information resided in the networked data, we separately adopt two powerful graph neural network methods, i.e., the graph convolutional network (GCN) (Kipf & Welling, 2016) and graph attention network layers (GAT) (Veličković et al., 2017), to learn the representation space. For these two models, their effectiveness of the learned representations has been verified in various graph learning tasks. The major difference between GCN and GAT is how the information from the one-hop neighborhood is aggregated. For GCN, a graph convolution operation is used to produce the normalized sum of the node features of neighbors. GAT introduces the attention mechanism to better quantify the importance of each edge. Here, we want to find out which model is better to unravel patterns of hidden confounders from the networked data with imbalanced covariate and imbalanced network structure.

For the graph convolutional network (GCN) model, the representation learning function $g : X \times A \rightarrow R$ is parameterized with the following layer-wise propagation rule:

$$r^{(l+1)} = \sigma(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}r^{(l)}W^{(l)}), \tag{6.3.1}$$

where $\tilde{A} = A + I_n$ is the adjacency matrix of graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$ with inserted self-loops, i.e., the identity matrix I_n. $\tilde{D}$ is its corresponding degree matrix, i.e., $\tilde{D}_{ii} = \sum_j \tilde{A}_{ij}$ and $W^{(l)}$ is a layer-specific trainable weight matrix. $\sigma(\cdot)$ denotes an activation function and here we apply the parametric ReLU (PReLU) function (He et al., 2015). A number of GCN layers can be stacked to approximate the function $g : X \times A \to R$.

For the graph attention network (GAT) model, the representation of confounder for the i-th node is a function of its covariates and receptive field. Here, we define the i-th node v_i and its one-hop neighbor nodes as the receptive field $\mathcal{N}(v_i)$. The representation learning function $g : X \times A \to R$ is parameterized with the following equation:

$$r_i^{(l+1)} = \sigma\left(\sum_{j \in \mathcal{N}(i)} \alpha_{ij}^{(l)} W^{(l)} r_j^{(l)}\right), \tag{6.3.2}$$

where $W^{(l)}$ is the learnable weight matrix and $W^{(l)} r_j^{(l)}$ is a linear transformation of the lower layer representation $r_j^{(l)}$. $\sigma(\cdot)$ is the activation function for nonlinearity. In Eq. (6.3.2), the representation of the i-th node and its neighbors are aggregated together, scaled by the normalized attention scores $\alpha_{ij}^{(l)}$.

$$\alpha_{ij}^{(l)} = \frac{\exp(\text{LeakyReLU}({a^{(l)}}^T (W^{(l)} r_i^{(l)} || W^{(l)} r_j^{(l)})))}{\sum_{k \in \mathcal{N}(i)} \exp(\text{LeakyReLU}({a^{(l)}}^T (W^{(l)} r_i^{(l)} || W^{(l)} r_k^{(l)})))}, \tag{6.3.3}$$

where softmax is used to normalize the attention scores on each node's incoming edges. The pair-wise attention score between two neighbors is calculated by LeakyReLU function $({a^{(l)}}^T (W^{(l)} r_i^{(l)} || W^{(l)} r_j^{(l)}))$. Here, it first concatenates the linear transformation of the lower layer representations for two nodes, i.e., $W^{(l)} r_i^{(l)} || W^{(l)} r_j^{(l)}$, where $||$ denotes concatenation, and then it takes a dot product of itself and a learnable weight vector $a^{(l)}$. Finally, the LeakyReLU function is applied.

To stabilize the learning process, a multi-head attention mechanism is employed. We compute multiple different attention maps and finally aggregate all the learned representations. In particular, K independent attention mechanisms execute the transformation of Eq. (6.3.2), and then their outputs are merged in two ways:

$$\text{concatenation} : r_i^{(l+1)} = ||_{k=1}^{K} \sigma\left(\sum_{j \in \mathcal{N}(i)} \alpha_{ij}^k W^k r_j^{(l)}\right) \tag{6.3.4}$$

or

$$\text{average} : h_i^{(l+1)} = \sigma\left(\frac{1}{K}\sum_{k=1}^{K}\sum_{j \in \mathcal{N}(i)} \alpha_{ij}^k W^k h_j^{(l)}\right) \tag{6.3.5}$$

When performing the multi-head attention on the final layer of the network, concatenation is no longer sensible. Thus, we use the concatenation for intermediary layers and the average for the final layer. An arbitrary number of GAT layers can be stacked to approximate the function $g : X \times A \to R$.

Structure Mutual Information Maximization. Inspired by a recent successful unsupervised graph learning method (Velickovic et al., 2019), we maximize structure mutual information to capture the imbalanced graph structure with respect to treatment and control nodes in the networked observational data. We aim to learn representations that can capture the imbalanced structure of the entire graph. Specifically, we utilize a structure summary function, $f : R \to S, S \in \mathbb{R}^d$, to summarize the learned representation into an entire graph structure representation, i.e., $s = f(g(X, A))$. From the observations in empirical evaluations, the structure summary function could be defined as $s = \sigma(\frac{1}{n}\sum_{i=1}^{n} r_i)$ to best capture the entire graph structure, where σ is the logistic sigmoid activation function.

Here, our purpose is to learn a representation vector, which can capture the entire graph structure encoded by the graph structure summary vector s and also reflect the abnormal

imbalance in the graph structure. Therefore, we aim at maximizing the mutual information between the learned representation vector r_i and the structure summary vector s.

Mutual information is a fundamental quantity for measuring the relationship between random variables. For example, the dependence of two random variables W and Z is quantified by mutual information as (Belghazi, Baratin, Rajeshwar, et al., 2018):

$$I(W;Z) = \int_{\mathcal{W}\times\mathcal{Z}} \log \frac{d\mathbb{P}_{WZ}}{d\mathbb{P}_W \otimes \mathbb{P}_Z} d\mathbb{P}_{WZ}, \tag{6.3.6}$$

where $\mathbb{P}_{WZ}$ is the joint probability distribution, and $\mathbb{P}_W = \int_{\mathcal{W}} d\mathbb{P}_{WZ}$ and $\mathbb{P}_Z = \int_{\mathcal{Z}} d\mathbb{P}_{WZ}$ are the marginals. However, mutual information has historically been difficult to compute. From the viewpoint of Shannon information theory, mutual information can be estimated as Kullback-Leibler divergence:

$$I(W;Z) = H(W) - H(W|Z) = D_{KL}(\mathbb{P}_{WZ}||\mathbb{P}_W \otimes \mathbb{P}_Z). \tag{6.3.7}$$

Actually, in our model, it is unnecessary to use the exact KL-based formulation of MI, as we only want to maximize the mutual information between representation vector r_i and structure summary vector s. A simple and stable alternative based on the Jensen-Shannon divergence (JSD) can be utilized. Thus, we follow the intuitions from deep infomax (Hjelm et al., 2018) and deep graph infomax (Velickovic et al., 2019) to maximize the mutual information.

To act as an agent for maximizing the mutual information, one discriminator $d : R \times S \rightarrow P, P \in \mathbb{R}$ is employed. The discriminator is formulated by a simple bilinear scoring function with nonlinear activation: $d(r_i, s) = \sigma({r_i}^T W s)$, which estimates the probability of the i-th node representation contained within the graph structure summary s. W is a learnable scoring matrix. To implement the discriminator, we also need to create the negative samples

compared with original samples and then use the discriminator to distinguish which one is from positive samples (original networked data) and which one is from the negative samples (created fake networked data), such that the original graph structure information could be correctly captured. The choice of the negative sampling procedure will govern the specific kinds of structural information that is desirable to be captured (Velickovic et al., 2019). Here, we focus on the imbalance between the edges that link nodes in the same group and those that link nodes in the different groups, i.e., treatment unit to treatment unit, treatment unit to control unit, and control unit to control unit. Therefore, our discriminator is designed to force the representations to capture this imbalanced structure by creating negative samples where the original adjacency matrix A is preserved, whereas the negative samples $\tilde{X}$ are obtained by row-wise shuffling of X. That is, the created fake networked data consists of the same nodes as the original graph, but they are located in different places in the same structure. Thus, the nodes at both ends of the edges may change the treatment choices like from treatment to control, from control to treatment, or remain unchanged. Then we also conduct the confounder representation learning for the created fake networked data $(\tilde{X}, A)$ to get the $\tilde{r}_i$. With the proposed discriminator, we could have $d(r_i, s)$ and $d(\tilde{r}_i, s)$, which indicate the probabilities of containing the representations of the i-th positive sample and negative sample in the graph structure summary, respectively.

We optimize the discriminator to maximize mutual information between r_i and s based on the Jensen Shannon divergence via a noise-contrastive type objective with a standard binary cross-entropy (BCE) loss (Hjelm et al., 2018; Velickovic et al., 2019):

$$\mathcal{L}_m = \frac{1}{2n}\Big(\sum_{i=1}^{n}\mathbb{E}_{(X,A)}[\log d(r_i, s)] + \sum_{j=1}^{n}\mathbb{E}_{(\tilde{X},A)}[\log\left(1 - d(\tilde{r}_i, s)\right)]\Big). \tag{6.3.8}$$

Potential Outcome Generator. So far, we have learned the representation space of confounders from networked data with the imbalanced network structure and imbalanced

covariates. The function $\Psi : R \times T \rightarrow Y$ maps the representation of hidden confounders and observed confounders as well as a treatment to the corresponding potential outcome, which is parameterized by a feed-forward deep neural network with multiple hidden layers and non-linear activation functions. The function $\Psi : R \times T \rightarrow Y$ uses representations and treatment options as inputs to predict potential outcomes. The output of Ψ estimates potential outcomes across treatment and control groups, including the estimated factual outcome $\hat{y}^f$ and the estimated counterfactual outcomes $\hat{y}^{cf}$. The factual outcomes y^f are used to minimize the loss of prediction $\hat{y}^f$. We aim to minimize the mean squared error in predicting factual outcomes:

$$\mathcal{L}_\Psi = \frac{1}{n}\sum_{i=1}^{N}(\hat{y}_i^f - y_i^f)^2, \tag{6.3.9}$$

where $\hat{y}_i = \Psi(r_i, t_i)$ denotes the inferred observed outcome of unit i corresponding to the factual treatment t_i.

Counterfactual Outcome Discriminator. The counterfactual outcome discriminator is intended to remove the imbalance of confounder representations between treatment and control groups, and thus it could improve the prediction accuracy of potential outcomes inferred by the outcome generator. We define the counterfactual outcome discriminator as $\Phi : R \times T \times (Y^f \text{or } \hat{Y}^{cf}) \rightarrow P$, where P is the discriminator's judgement, i.e., probability that this outcome for unit i given R and T is factual outcome. P is defined as:

$$P = \begin{cases} P(\text{judges } y^f \text{as factual} | x, t) \text{ if } t \text{ is factual treatment choice} \\ P(\text{judges } \hat{y}^{cf} \text{as factual} | x, t) \text{ if } t \text{ is not factual treatment choice.} \end{cases} \tag{6.3.10}$$

To improve the accuracy of prediction and avoid risk of losing the influence of treatment t and potential outcomes (y^for $\hat{y}^{cf}$) due to high dimensional representation vector, we adopt separate head networks for treatment and control groups (Shalit et al., 2017). Besides, to

improve the influence of $(y^f, \hat{y}^{cf})$ in the discriminator, we add $(y^f \text{or } \hat{y}^{cf})$ into each layer of the neural network, repetitively.

The discriminator deals with a binary classification task, which assigns one label (i.e., factual outcome or counterfactual outcome) to the vector concatenating the representation vector r and potential outcome $(y^f \text{or } \hat{y}^{cf})$ under the treatment head network and control head network, respectively. Thus, the loss of discrimination is measured by the cross-entropy with truth probability, where $P^{\text{truth}} = 1$ if y^f is input, and $P^{\text{truth}} = 0$ if $\hat{y}^{cf}$ is input. In each iteration of training, we make sure to input the same number of units in the treatment and control groups to ensure that there exist the same number of factual outcomes as counterfactual outcomes in each head network to overcome the imbalanced classification. The inputs of discriminator are generated by the outcome generator $\Psi(R, T)$, and then the cross entropy loss of the counterfactual outcome discriminator is defined as:

$$\mathcal{L}_{\Phi,\Psi} = -\frac{1}{2n}\sum_{t=0}^{1}\sum_{i=1}^{n}(p_{ti}^{\text{truth}}\log(p_{ti}) + (1 - p_{ti}^{\text{truth}})\log(1 - p_{ti})), \tag{6.3.11}$$

where p_{ti}^{truth} is the indicator that this input outcome for unit i under treatment option t is the observed factual outcome or inferred outcome from generator module, i.e., p_{ti}^{truth} equals 1 or 0, separately. P_{ti} is the probability judged by discriminator that how likely this input outcome for unit i under treatment option t is a factual outcome.

Thus far, we have introduced the outcome generator to estimate potential outcomes for each unit across treatment and control groups, and the discriminator to determine if the potential outcome is factual, given a unit's confounder representation under treatment or control group. In the initial iterations of the model training, the outcome generator may generate potential outcomes that are very different from factual outcomes as determined by the discriminator. As the model is trained further, the discriminator may no longer be able to distinguish the generated counterfactual outcome and the factual outcome. At this

point, we have attained all potential outcomes for each unit under treatment and control groups. For the training procedure of optimizing the outcome generator and discriminator, the minimax game is adopted. Putting all of the above together, the objective function of our Graph Infomax Adversarial Learning (GIAL) method is:

$$\min_{\Psi}\max_{\Phi,m} \left(\mathcal{L}_{\Psi} + \alpha\mathcal{L}_{m} - \beta\mathcal{L}_{\Phi,\Psi}\right), \tag{6.3.12}$$

where α and β are the hyper-parameters controlling the trade-off among the outcome generator, mutual information, and discriminator.

6.3.3 Overview of GIAL

The proposed Graph Infomax Adversarial Learning method (GIAL) can estimate the treatment effects from networked observational data, which utilizes the graph neural network (GCN or GAT) and structure mutual information to learn the representations of hidden confounders and observed confounders, by mapping the feature covariates and network structure simultaneously into a representation space. Adversarial learning is also employed to mitigate the representation imbalance between treatment and control groups and to predict the counterfactual outcomes. After obtaining the counterfactual outcomes, GIAL can estimate the treatment effects.

We summarize the procedures of GIAL as follows:

1. Create the negative samples $(\tilde{X}, A)$ by row-wise shuffling of X and keeping the original adjacency matrix A.

2. Learn the representation space R for the positive samples (X, A) by function $g : X \times A \to R$ by a graph neural network.

3. Learn the representation space $\tilde{R}$ for the negative samples $(\tilde{X}, A)$ by function $g : \tilde{X} \times A \to \tilde{R}$ by the same graph neural network as Step 2.

4. Utilize a structure summary function $f : R^{n \times d} \to S$ to summarize the learned representation into a graph-level structure representation, i.e., $s = f(g(X, A))$.

5. Employ a discriminator $d : R \times S \to P$ to obtain $d(r_i, s)$ and $d(\tilde{r}_i, s)$, which are the probabilities that the representations of i-th positive and negative samples are contained within the original graph structure summary s.

6. Utilize functions g, f and d to maximize mutual information between R and S.

7. Use potential outcome generator $\Psi : R \times T \to Y$ to estimate the potential outcomes.

8. Apply counterfactual discriminator $\Phi : R \times T \times (Y^f \text{or } \hat{Y}^{cf}) \to P$ to remove imbalance of confounder representations between treatment and control group.

9. Here, Steps 6, 7, and 8 in the procedure are jointly trained together by optimizing minimax rule Eq. (6.3.12) about $\mathcal{L}_m$, $\mathcal{L}_\Psi$, and $\mathcal{L}_{\Phi,\Psi}$ to update parameters in g, f, d, Φ, and Ψ.

6.4 Experiments

In this section, we conduct experiments on two semi-synthetic networked datasets, including the BlogCatalog and Flickr, to evaluate the following aspects: (1) Our proposed method can improve treatment effect estimation with respect to average treatment effect and individualized treatment effect compared to the state-of-the-art methods. (2) The structure mutual information can help representations capture more hidden confounder information, and thus increase the predictive accuracy for counterfactual outcomes. (3) The proposed method is robust to the hyperparameters.

6.4.1 Dataset

BlogCatalog. BlogCatalog is a social blog directory that manages the bloggers and their blogs. In this dataset, each unit is a blogger and each edge represents the social relationship between two bloggers. The features are bag-of-words representations of keywords in bloggers' descriptions. We follow the assumptions and procedures of synthesizing the outcomes and treatment assignments in (Guo et al., 2019). In this semi-synthetic networked dataset, the outcomes are the opinions of readers on each blogger and the treatment options are mobile devices or desktops on which blogs are read more. If the blogger's blogs are read more on mobile devices, the blogger is in the treatment group; rather, they are read more on desktops, the blogger is in the control group. We also assume that the topics of bloggers with the social relationship can causally affect their treatment assignment and readers' opinions on them. To model readers' preference on reading some topics from mobile devices and others from desktops, one LDA topic model (Guo et al., 2019) is trained. Three settings of datasets are created with $k = 0.5, 1, \text{and } 2$ that represent the magnitude of the confounding bias in the dataset. $k = 0$ means the treatment assignment is random and there is no selection bias, and greater k means larger selection bias. The simulation procedures are repeated 10 times for each setting of $k \in 0.5, 1, 2$.

Flickr. Flickr is a popular photo-sharing and hosting service, and it supports an active community where people can share each other's photos. In the Flickr dataset, each unit is a user and each edge represents the social relationship between two users. The features of each user represent a list of tags of interest. The same settings and simulation procedures as BlogCatalog dataset are adopted here. The simulation procedures are repeated 10 times for each setting of $k \in 0.5, 1, 2$. Table 6.1 presents an overview of these two datasets.

Table 6.1: Properties of BlogCatalog and Flickr datasets.

Datasets	**BlogCatalog**	**Flickr**
Nodes	5,196	7,575
Features	8,189	12,047
Edges	171,743	239,738
Treatments	2	2

6.4.2 Baseline Methods

We compare the proposed Graph Infomax Adversarial Learning method (GIAL) method with the following baseline methods. **Network Deconfounder (ND)** (Guo et al., 2019) utilizes the graph convolutional networks and integral probability metric to learn balanced representations to recognize patterns of hidden confounders from the network dataset. **Counterfactual Regression (CFRNET)** (Shalit et al., 2017) maps the original features into a latent representation space by minimizing the error in predicting factual outcomes and imbalance measured by integral probability metric between the treatment representations and the control representations. **Treatment-agnostic Representation Networks (TARNet)** (Shalit et al., 2017) is a variant of counterfactual regression without balance regularization. **Causal Effect Variational Autoencoder (CEVAE)** (Louizos et al., 2017) is based on Variational Autoencoders (VAE), which follows the causal structure of inference to simultaneously estimate the unknown latent space summarizing the confounders and the causal effect. **Causal Forests (CF)** (Wager & Athey, 2018b) is a nonparametric forest-based method for estimating heterogeneous treatment effects by extending Breiman's random forest algorithm. **Bayesian Additive Regression Trees (BART)** (Chipman et al., 2010) is a nonparametric Bayesian regression model, which uses dimensionally adaptive random basis elements.

6.4.3 Descriptive Data Analysis

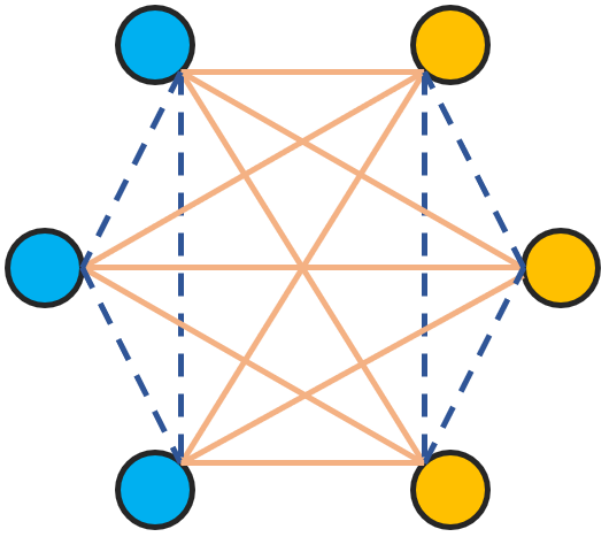

Figure 6.4: Example of complete graph. The solid line represents heterogeneous edge and the dashed line means homogeneous edge.

Before estimating the treatment effects from these two networked datasets, we provide the descriptive data analysis to demonstrate the existence of network structural imbalance in the networked data for causal inference problems.

According to graph theory, in the complete graph which is a simple undirected graph where every pair of distinct nodes is connected by a unique edge, there are $\frac{n\times(n-1)}{2}$ edges for n nodes. We assume that the n nodes are evenly divided into treatment group and control group with the same $\frac{n}{2}$ nodes in each group, and also each node has the same possibility to have an edge (relationship) with another node regardless of the node's treatment assignment. Then, this graph is still a complete graph with $\frac{n\times(n-1)}{2}$ edges. Now the edges in this graph are put into two categories: (a) the homogeneous group including the edges that link the nodes in the same group (treatment-treatment or control-control); (b) the heterogeneous group including the edges that link the nodes in different groups (treatment-control). Under

Table 6.2: Summary of homogeneous edges and heterogeneous edges for the BlogCatalog datasets and Flickr datasets.

Dataset	k	Homogeneous	Heterogeneous
BlogCatalog	0.5	**94524.5**	77218.5
	1	**101102.8**	70640.2
	2	**116031.8**	55711.2
Flickr	0.5	**124320.9**	115417.1
	1	**130978.5**	108759.5
	2	**141957.3**	97780.7

Table 6.3: Performance comparison on BlogCatalog and Flickr datasets with different $k \in 0.5, 1, 2$.

	BlogCatalog						Flickr					
	k=0.5		k=1		k=2		k=0.5		k=1		k=2	
Method	$\sqrt{\epsilon_{\text{PEHE}}}$	ϵ_{ATE}	$\sqrt{\epsilon_{\text{PEHE}}}$	ϵ_{ATE}	$\sqrt{\epsilon_{\text{PEHE}}}$	ϵ_{ATE}	$\sqrt{\epsilon_{\text{PEHE}}}$	ϵ_{ATE}	$\sqrt{\epsilon_{\text{PEHE}}}$	ϵ_{ATE}	$\sqrt{\epsilon_{\text{PEHE}}}$	ϵ_{ATE}
BART	4.808	2.680	5.770	2.278	11.608	6.418	4.907	2.323	9.517	6.548	13.155	9.643
CF	7.456	1.261	7.805	1.763	19.271	4.050	8.104	1.359	14.636	3.545	26.702	4.324
CEVAE	7.481	1.279	10.387	1.998	24.215	5.566	12.099	1.732	22.496	4.415	42.985	5.393
TARNet	11.570	4.228	13.561	8.170	34.420	13.122	14.329	3.389	28.466	5.978	55.066	13.105
$\text{CFRNET}_{\text{MMD}}$	11.536	4.127	12.332	5.345	34.654	13.785	13.539	3.350	27.679	5.416	53.863	12.115
$\text{CFRNET}_{\text{Wass}}$	10.904	4.257	11.644	5.107	34.848	13.053	13.846	3.507	27.514	5.192	53.454	13.269
ND	4.532	0.979	4.597	0.984	9.532	2.130	4.286	0.805	5.789	1.359	9.817	2.700
GIAL_{GAT} (Ours)	4.215	0.912	4.258	0.937	9.119	1.982	4.015	0.773	5.432	1.2312	9.428	2.586
GIAL_{GCN} (Ours)	**4.023**	**0.841**	**4.091**	**0.883**	**8.927**	**1.780**	**3.938**	**0.682**	**5.317**	**1.194**	**9.275**	**2.245**

the assumption that each node has the same possibility to be connected with another node regardless of node's treatment assignment, we can find that in the homogeneous group, there are $\frac{n^2}{4} - \frac{n}{2}$ edges and in the heterogeneous group, there are $\frac{n^2}{4}$ edges. The number of edges in heterogeneous group should be greater than that in homogeneous group. For example, as shown in Fig. 6.4, there is one complete graph with 6 nodes including 3 treatment nodes and 3 control nodes. The heterogeneous group has 9 edges, while the homogeneous group has 6 edges.

Table 6.4: Summary of results in ablation studies.

	k=0.5		k=1		k=2	
	$\sqrt{\epsilon_{\text{PEHE}}}$	ϵ_{ATE}	$\sqrt{\epsilon_{\text{PEHE}}}$	ϵ_{ATE}	$\sqrt{\epsilon_{\text{PEHE}}}$	ϵ_{ATE}
BlogCatalog						
GIAL	4.023	0.841	4.091	0.883	8.927	1.780
GIAL (w/o SMI)	4.422	0.982	4.481	0.981	9.315	2.142
GIAL (w/o CD)	4.482	0.987	4.951	1.023	13.598	3.215
Flickr						
GIAL	3.938	0.682	5.317	1.194	9.275	2.245
GIAL (w/o SMI)	4.158	0.792	5.694	1.375	9.673	2.661
GIAL (w/o CD)	4.284	0.812	6.127	1.435	11.524	3.564

We separately calculate the average numbers of homogeneous edges and heterogeneous edges for the BlogCatalog datasets and Flickr datasets, and report them in Table 6.2. We can observe that the homogeneous edges are consistently greater than heterogeneous edges for both datasets with different k. This result totally agrees with our expectation that, in the causal inference problem, the network structure is imbalanced. Therefore, the relationship is more likely to appear among people who are in the same group. This is the major difference between traditional graph learning tasks and the causal inference task on networked data, which is also the motivation of our proposed model.

6.4.4 Experimental Settings

In the following experiments, we randomly sample 60% and 20% of the units as the training set and validation set, and use the remaining 20% units to form the test set. For each dataset with different imbalance k, the simulation procedures are repeated 10 times and we report the average mean.

GIAL. By using different graph neural networks to learn the representation space from networked dataset, the proposed GIAL method has two variants denoted as GIAL_{GCN} and GIAL_{GAT}, which adopt the original implementation of graph convolutional network (Kipf & Welling, 2016) and graph attention network (GAT) (Velickovic et al., 2019), respectively. Besides, a squared l_2 norm regularization with hyperparameter 10^{-4} is added into our model to mitigate the overfitting issue. The hyperparameters of our method are chosen based on performance on the validation dataset, and the searching range is shown in Table 6.5. The Adam SGD optimizer (Kingma & Ba, 2014) is used to train the final objective function Eq. (6.3.12) with an initial learning rate of 0.001 and an early stopping strategy with a patience of 100 epochs.

Table 6.5: Hyperparameters and ranges.

Hyperparameter	Range
α, β	0, 10^{-4},10^{-3} ,10^{-2} ,10^{-1}
Dim. of confounder representation	50, 100, 150, 200
No. of GCN and GAT layers	1, 2, 3
No. of attention heads in GAT	1, 2, 3, 4
No. of outcome generator layer	1, 2, 3, 4

Baseline Methods. BART, CF, CEVAE, TARNet, and CFRNET are not originally designed for the networked observational data, so they cannot directly utilize the network information. To be fair, we concatenate the corresponding row of adjacency matrix to the original features (Guo et al., 2019), but this strategy cannot effectively improve the performance of baselines due to the curse of dimensionality. Besides, we adopt their default hyperparameter settings.

6.4.5 Results

For the BlogCatalog and Flickr datasets, we adopt two commonly used evaluation metrics to evaluate the performance of our method and baselines. The first one is the error of ATE estimation, which is defined as $\epsilon_{\text{ATE}} = |\text{ATE} - \widehat{\text{ATE}}|$, where ATE is the true value and $\widehat{\text{ATE}}$ is an estimated ATE. The second one is the error of expected precision in estimation of heterogeneous effect (PEHE) (Hill, 2011), which is defined as $\epsilon_{\text{PEHE}} = \frac{1}{n}\sum_{i=1}^{n}(\text{ITE}_i - \widehat{\text{ITE}}_i)^2$, where ITE_i is the true ITE for unit i and $\widehat{\text{ITE}}_i$ is an estimated ITE for unit i.

Table 6.3 shows the performance of our method and baseline methods on the BlogCatalog and Flickr datasets over 10 realizations. We report the average results of $\sqrt{\epsilon_{\text{PEHE}}}$ and ϵ_{ATE} on the test sets. GIAL_{GCN} achieves the best performance with respect to $\sqrt{\epsilon_{\text{PEHE}}}$ and ϵ_{ATE} in all cases of both datasets. Although the GIAL_{GAT} also has obvious improvements compared to baseline methods, it is outperformed by GIAL_{GCN}. GCN demonstrates clear superiority over GAT when recognizing patterns of hidden confounders from imbalanced network structure. Because $k = 0.5, 1, \text{and}\, 2$ is used to represent the magnitude of the confounding bias in both datasets, results show that GIAL consistently outperforms the baseline methods under different levels of divergence, and our method is robust to a high level of confounding bias. Compared to baseline methods (e.g., CFRNET) only relying on observed confounders but without utilizing the network information, our model is capable of recognizing the patterns of hidden confounders from the network structure. Compared to baseline methods with learning network information (e.g., ND), our model has significant performance advantages, which demonstrates our model can capture more information from imbalanced network structure. The reason is that our method maximizes the structure mutual information, instead of directly adopting graph learning method without considering the specificity of networked data in the causal inference problem.

6.4.6 Model Evaluation

Experimental results on both datasets show that GIAL obtains a more accurate estimation of the ATE and ITE than the state-of-the-art methods. We further evaluate the performance of GIAL from two perspectives, including the effectiveness of each component, and its robustness to hyper-parameters.

We perform two ablation studies of GIAL_{GCN} on both datasets. The first one is GIAL (w/o SMI) where structure mutual information maximizing module is removed. We directly adopt graph neural networks to learn the representation space without considering structural imbalance of networked data. The second ablation study is GIAL (w/o CD) where the counterfactual outcome discriminator is removed and there is not any restriction on the divergence between the representation distributions of treatment and control groups.

As shown in Table 6.4, the performance becomes poor after removing either the structure mutual information or counterfactual outcome discriminator, compared to the original GIAL. More specifically, after removing the structure mutual information, $\sqrt{\epsilon_{\text{PEHE}}}$ and ϵ_{ATE} increase dramatically and have similar performance to other baseline methods. Besides, as the bias (k) increases, the difference between the performance of GIAL (w/o CD) and performance of original GIAL increases further. Therefore, the structure mutual information and counterfactual outcome discriminator are essential components of our model.

Next, we explore the model's sensitivity to the most important parameters α and β, which control the ability to capture the graph structure and handle the confounding bias when estimating the potential outcomes. We show the results of $\sqrt{\epsilon_{\text{PEHE}}}$ and ϵ_{ATE} on BlogCatalog dataset with different k in Fig. 6.5. We observe that the performance is stable over a large parameter range. It confirms the effectiveness and robustness of structure mutual information and counterfactual outcome discriminator in GIAL, which is consistent with our ablation studies, i.e., GIAL (w/o SMI) and GIAL (w/o CD).

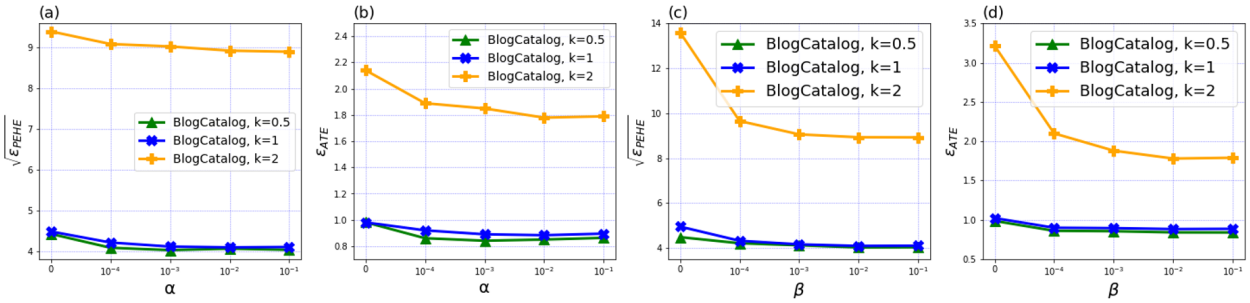

Figure 6.5: Sensitivity analysis for α and β of structure mutual information and counterfactual outcome discriminator.

6.5 Related Work

The related work is presented along with two directions: learning causal effects from observational data and graph neural networks.

Embracing the rapid developments in machine learning and deep learning, various causal effect estimation methods for observational data have sprung up. Balancing neural networks (BNNs) (F. Johansson et al., 2016) and counterfactual regression networks (CFRNET) (Shalit et al., 2017) are proposed to balance covariate distributions across treatment and control groups by formulating the problem of counterfactual inference as a domain adaptation problem. For most of existing methods, the strong ignorability assumption is the most important prerequisite. However, this assumption might be untenable in practice. A series of methods have been proposed to relax the strong ignorability assumption. A latent variable is inferred as a substitute for unobserved confounders and then uses that substitute to perform causal inference (Y. Wang & Blei, 2019). Variational autoencoder has been used to infer the relationships between the observed confounders based on the assumption joint distribution of the latent confounders and the observed confounders can be approximately recovered

solely from the observations (Louizos et al., 2017). However, they still rely on the assumption that the model is capable of extracting the latent information to represent confounders from observational data. Recently, some work aims to relax the strong ignorability assumption to estimate the causal inference in the presence of unobserved confounding, where the network connecting the units is a proxy is for the unobserved confounding. The network deconfounder (Guo et al., 2019) learns representations of confounders from network data by adopting the graph convolutional networks and reduces the selection bias by minimizing Wasserstein distance. Another work utilizes graph attention networks to learn representations and mitigates confounding bias by representation balancing and treatment prediction, simultaneously (Guo, Li, Li, et al., 2020). Causal network embedding (CNE) (Veitch et al., 2019) is proposed to learn node embeddings from network data to represent confounders by reducing the causal estimation problem to a semi-supervised prediction of both the treatments and outcomes. For the existing methods about networked data, they do not dig deeply on what is the essential difference between the networked data under causal inference problem and the networked data for traditional graph learning tasks such as node classification, link detection, etc. This is the reason why we propose this Graph Informax Adversarial Learning model, instead of directly adopting the GCN or GAT to learn the representation from the networked data.

Graph learning is increasingly becoming fascinating as more and more real-world data can be modeled as networked data. Graph convolutional network (Kipf & Welling, 2016) is an effective approach for semi-supervised learning on networked data, via a localized first-order approximation of spectral graph convolutions. Graph attention network (GAT) (Veličković et al., 2017) is an attention-based architecture leveraging masked self-attentional layers where nodes are able to attend over their neighborhoods' features. Deep graph infomax (DGI) (Velickovic et al., 2019) is one approach for learning node representations within networked data in an unsupervised manner, which relies on maximizing mutual information between

patch representations and high-level summaries of graphs. In our model, we extend the idea in DGI originally aimed for unsupervised learning to representation learning under the causal inference setting. Utilizing the structure mutual information can help representations capture the imbalanced structure that is specific to the causal inference problem.

6.6 Summary

In this chapter, we propose the Graph Infomax Adversarial Learning method (GIAL) to catch hidden confounders and estimate the treatment effects from networked observational data. GIAL makes full use of the network structure to capture more information by recognizing the imbalance in network structure. Our work clarifies the greatest particularity of networked data under the causal inference problem compared with traditional graph learning tasks, that is, the structural imbalance due to confounding bias between treatment and control groups. Extensive experimental results on two benchmark datasets show the effectiveness and advantages of the proposed GIAL method.

Chapter 7

Continual Lifelong Causal Effect Inference with Observational Data

7.1 Introduction

Causal effect inference is a critical research topic across many domains, such as statistics, computer science, public policy, and economics. Randomized controlled trials (RCT) are usually considered as the gold-standard for causal effect inference, which randomly assigns participants into a treatment or control group. As the RCT is conducted, the only expected difference between the treatment and control groups is the outcome variable being studied. However, in reality, randomized controlled trials are always time-consuming and expensive, and thus the study cannot involve many subjects, which may be not representative of the real-world population the intervention would eventually target. Nowadays, estimating causal effects from observational data has become an appealing research direction owing to a large amount of available data and low budget requirements, compared with RCT (Yao et al., 2020). Researchers have developed various strategies for causal effect inference with observational data, such as tree-based methods (Chipman et al., 2010; Wager & Athey, 2018a),

representation learning methods (Z. Chu et al., 2020; F. Johansson et al., 2016; S. Li & Fu, 2017b; Shalit et al., 2017), adapting Bayesian algorithms (A. M. Alaa & van der Schaar, 2017), generative adversarial nets (Yoon et al., 2018), variational autoencoders (Louizos et al., 2017) and so on.

Although significant advances have been made to overcome the challenges in causal effect estimation with observational data, such as missing counterfactual outcomes and selection bias between treatment and control groups, the existing methods only focus on source-specific and stationary observational data. Such learning strategies assume that all observational data are already available during the training phase and from the only one source. This assumption is unsubstantial in practice due to two reasons. The first one is based on the characteristics of observational data, which are incrementally available from non-stationary data distributions. For instance, the number of electronic medical records in one hospital is growing every day, or the electronic medical records for one disease may be from different hospitals or even different countries. This characteristic implies that one cannot have access to all observational data at one time point and from one single source. The second reason is based on the realistic consideration of accessibility. For example, when the new observational are available, if we want to refine the model previously trained by original data, maybe the original training data are no longer accessible due to a variety of reasons, e.g., legacy data may be unrecorded, proprietary, too large to store, or subject to privacy constraint (J. Zhang et al., 2020). This practical concern of accessibility is ubiquitous in various academic and industrial applications. That's what it boiled down to: in the era of big data, we face the new challenges in causal inference with observational data: the **extensibility** for incrementally available observational data, the **adaptability** for extra domain adaptation problem except for the imbalance between treatment and control groups in one source, and the **accessibility** for a huge amount of data.

Existing causal effect inference methods, however, are unable to deal with the aforementioned new challenges, i.e., extensibility, adaptability, and accessibility. Although it is possible to adapt existing causal inference methods to address the new challenges, these adapted methods still have inevitable defects. Three straightforward adaptation strategies are described as follows. (1) If we directly apply the model previously trained based on original data to new observational data, the performance on new task will be very poor due to the domain shift issues among different data sources; (2) If we utilize newly available data to re-train the previously learned model, adapting changes in the data distribution, old knowledge will be completely or partially overwritten by the new one, which can result in severe performance degradation on old tasks. This is the well-known *catastrophic forgetting* problem (French, 1999; McCloskey & Cohen, 1989); (3) To overcome the catastrophic forgetting problem, we may rely on the storage of old data and combine the old and new data together, and then re-train the model from scratch. However, this strategy is memory inefficient and time-consuming, and it brings practical concerns such as copyright or privacy issues when storing data for a long time (Samet et al., 2013). Our empirical evaluations in Section 4 demonstrate that any of these three strategies in combination with the existing causal effect inference methods is deficient.

To address the above issues, we propose a **C**ontinual Causal **E**ffect **R**epresentation **L**earning method (CERL) for estimating causal effect with incrementally available observational data. Instead of having access to all previous observational data, we only store a limited subset of feature representations learned from previous data. Combining the selective and balanced representation learning, feature representation distillation, and feature transformation, our method preserves the knowledge learned from previous data and update the knowledge by leveraging new data, so that it can achieve the continual causal effect estimation for new data without compromising the estimation capability for previous data. To summarize, our main contributions include: Our work is the first to introduce the con-

tinual lifelong causal effect inference problem for the incrementally available observational data and three corresponding evaluation criteria, i.e., extensibility, adaptability, and accessibility; We propose a new framework for continual lifelong causal effect inference based on deep representation learning and continual learning; Extensive experiments demonstrate the deficiency of existing methods when facing the incrementally available observational data and our model's outstanding performance.

7.2 Background and Problem Statement

Suppose that the observational data contain n units collected from d different domains and the d-th dataset D_d contains the data $\{(x, y, t)|x \in X, y \in Y, t \in T\}$ collected from d-th domain, which contains n_d units. Let X denote all observed variables, Y denote the outcomes in the observational data, and T is a binary variable. Let $D_{1:d} = \{D_1, D_2, ..., D_d\}$ be the set of combination of d dataset, separately collected from d different domains. For d datasets $\{D_1, D_2, ..., D_d\}$, they have the common observed variables but due to the fact that they are collected from different domains, they have different distributions with respect to X, Y, and T in each dataset. Each unit in the observational data received one of two treatments. Let t_i denote the treatment assignment for unit i; $i = 1, ..., n$. For binary treatments, $t_i = 1$ is for the treatment group, and $t_i = 0$ for the control group. The outcome for unit i is denoted by y_t^i when treatment t is applied to unit i; that is, y_1^i is the potential outcome of unit i in the treatment group and y_0^i is the potential outcome of unit i in the control group. For observational data, only one of the potential outcomes is observed. The observed outcome is called the factual outcome and the remaining unobserved potential outcomes are called counterfactual outcomes.

In this chapter, we follow the potential outcome framework for estimating treatment effects (Rubin, 1974; Splawa-Neyman et al., 1990). The individual treatment effect (ITE)

for unit i is the difference between the potential treated and control outcomes, and is defined as $\text{ITE}_i = y_1^i - y_0^i$. The average treatment effect (ATE) is the difference between the mean potential treated and control outcomes, which is defined as $\text{ATE} = \frac{1}{n}\sum_{i=1}^{n}(y_1^i - y_0^i)$.

The success of the potential outcome framework is based on the following assumptions (G. W. Imbens & Rubin, 2015b), which ensure that the treatment effect can be identified. **Stable Unit Treatment Value Assumption (SUTVA)**: The potential outcomes for any units do not vary with the treatments assigned to other units, and, for each unit, there are no different forms or versions of each treatment level, which lead to different potential outcomes. **Consistency**: The potential outcome of treatment t is equal to the observed outcome if the actual treatment received is t. **Positivity**: For any value of x, treatment assignment is not deterministic, i.e.,$P(T = t|X = x) > 0$, for all t and x. **Ignorability**: Given covariates, treatment assignment is independent to the potential outcomes, i.e., $(y_1, y_0) \perp\!\!\!\perp t|x$.

The goal of our work is to develop a novel continual causal inference framework, given new available observational data D_d, to estimate the causal effect for newly available data D_d as well as the previous data $D_{1:(d-1)}$ without having access to previous training data in $D_{1:(d-1)}$.

7.3 The Proposed Framework

The availability of "real world evidence" is expected to facilitate the development of causal effect inference models for various academic and industrial applications. How to achieve continual learning from incrementally available observational data from non-stationary data domains is a new direction in causal effect inference. Rather than only focusing on handling the selection bias problem, we also need to take into comprehensive consideration three

aspects of the model, i.e., the **extensibility** for incrementally available observational data, the **adaptability** for various data sources, and the **accessibility** for a huge amount of data.

We propose the **C**ontinual Causal **E**ffect **R**epresentation **L**earning method (CERL) for estimating causal effect with incrementally available observational data. Based on selective and balanced representation learning for treatment effect estimation, CERL incorporates feature representation distillation to preserve the knowledge learned from previous observational data. Besides, aiming at adapting the updated model to original and new data without having access to the original data, and solving the selection bias between treatment and control groups, we propose one representation transformation function, which maps partial original feature representations into new feature representation space and makes the global feature representation space balanced with respect to treatment and control groups. Therefore, CERL can achieve the continual causal effect estimation for new data and meanwhile preserve the estimation capability for previous data, without the aid of original data.

7.3.1 Model Architecture

To estimate the incrementally available observational data, the framework of CERL is mainly composed of two components: (1) the baseline causal effect learning model is only for the first available observational data, and thus we don't need to consider the domain shift issue among different data sources. This component is equivalent to the traditional causal effect estimation problem; (2) the continual causal effect learning model is for the sequentially available observational data, where we need to handle more complex issues, such as knowledge transfer, catastrophic forgetting, global representation balance, and memory constraint. We present the details of each component as follows.

The Baseline Causal Effect Learning Model

We first describe the baseline causal effect learning model for the initial observational dataset and then bring in subsequent datasets. For causal effect estimation in the initial dataset, it can be transformed into the traditional causal effect estimation problem. Motivated by the empirical success of deep representation learning for counterfactual inference (Z. Chu et al., 2020; Shalit et al., 2017), we propose to learn the selective and balanced feature representations for treated and control units, and then infer the potential outcomes based on learned representation space.

Learning Selective and Balanced Representation. Firstly, we adopt a deep feature selection model that enables variable selection in one deep neural network, i.e., $g_{w_1} : X \to R$, where X denotes the original covariate space, R denotes the representation space, and w_1 are the learnable parameters in function g. The elastic net regularization term (Zou & Hastie, 2005) is adopted in our model, i.e., $L_{w_1} = \|w_1\|_2^2 + \|w_1\|_1$. Elastic net throughout the fully connected representation layers assigns larger weights to important features. This strategy can effectively filter out the irrelevant variables and highlight the important variables.

Due to the selection bias between treatment and control groups and among the sequential different data sources, the magnitudes of confounders may be significantly different. To effectively eliminate the imbalance caused by the significant difference in magnitudes between treatment and control groups and among different data sources, we propose to use cosine normalization in the last representation layer. In the multi-layer neural networks, we traditionally use dot products between the output vector of the previous layer and the incoming weight vector, and then input the products to the activation function. The result of dot product is unbounded. Cosine normalization uses cosine similarity instead of simple dot products in neural networks, which can bound the pre-activation between -1 and 1. The result could be even smaller when the dimension is high. As a result, the variance can

be controlled within a very narrow range (Luo et al., 2018). Cosine normalization is defined as $r = \sigma(r_{norm}) = \sigma\big(\cos(w, x)\big) = \sigma(\frac{w \cdot x}{|w||x|})$, where r_{norm} is the normalized pre-activation, w is the incoming weight vector, x is the input vector, and σ is nonlinear activation function.

Motivated by Shalit et al., 2017, we adopt integral probability metrics (IPM) when learning the representation space to balance the treatment and control groups. The IPM measures the divergence between the representation distributions of treatment and control groups, so we want to minimize the IPM to make two distributions more similar. Let $P(g(x)|t = 1)$ and $Q(g(x)|t = 0)$ denote the empirical distributions of the representation vectors for the treatment and control groups, respectively. We adopt the IPM defined in the family of 1-Lipschitz functions, which leads to IPM being the Wasserstein distance (Shalit et al., 2017; Sriperumbudur et al., 2012). In particular, the IPM term with Wasserstein distance is defined as $\text{Wass}(P, Q) = \inf_{k \in \mathcal{K}} \int_{g(x)} \|k(g(x)) - g(x)\| P(g(x)) d(g(x))$, where γ denotes the hyper-parameter controlling the trade-off between $\text{Wass}(P, Q)$ and other terms in the final objective function. $\mathcal{K} = \{k | Q(k(g(x))) = P(g(x))\}$ defines the set of push-forward functions that transform the representation distribution of the treatment distribution P to that of the control Q and $g(x) \in \{g(x)_i\}_{i:t_i=1}$.

Inferring Potential Outcomes. We aim to learn a function $h_{\theta_1} : R \times T \to Y$ that maps the representation vectors and treatment assignment to the corresponding observed outcomes, and it can be parameterized by deep neural networks. To overcome the risk of losing the influence of T on R, $h_{\theta_1}(g_{w_1}(x), t)$ is partitioned into two separate tasks for treatment and control groups, respectively. Each unit is only updated in the task corresponding to its observed treatment. Let $\hat{y}_i = h_{\theta_1}(g_{w_1}(x), t)$ denote the inferred observed outcome of unit i corresponding to factual treatment t_i. We minimize the mean squared error in predicting factual outcomes: $L_Y = \frac{1}{n_1} \sum_{i=1}^{n_1} (\hat{y}_i - y_i)^2$.

Putting all the above together, the objective function of our baseline causal effect learning model is: $L = L_Y + \alpha Wass(P, Q) + \lambda L_{w_1}$, where α and λ denote the hyper-parameters controlling the trade-off among $Wass(P, Q)$, L_w, and L_Y in the objective function.

The Sustainability of Model Learning

By far, we have built the baseline model for causal effect estimation with observational data from a single source. To avoid catastrophic forgetting when learning new data, we propose to preserve a subset of lower dimensional feature representations rather than all original covariates. We also can adjust the number of preserved feature representations according to the memory constraint.

After the completion of baseline model training, we store a subset of feature representations $R_1 = \{g_{w_1}(x)|x \in D_1\}$ and the corresponding $\{Y, T\} \in D_1$ as memory M_1. The size of stored representation vectors can be reduced to satisfy the pre-specified memory constraint by a herding algorithm (Rebuffi et al., 2017; Welling, 2009). The herding algorithm can create a representative set of samples from distribution and requires fewer samples to achieve a high approximation quality than random subsampling. We run the herding algorithm separately for treatment and control groups to store the same number of feature representations from treatment and control groups. At this point, we only store the memory set M_1 and model g_{w_1}, without the original data (D_1).

The Continual Causal Effect Learning Model

For now, we have stored memory M_1 and baseline model. To continually estimate the causal effect for incrementally available observational data, we incorporate feature representation distillation and feature representation transformation to estimate causal effect for all seen data based on balanced global feature representation space. The framework of CERL is shown in Fig. 7.1.

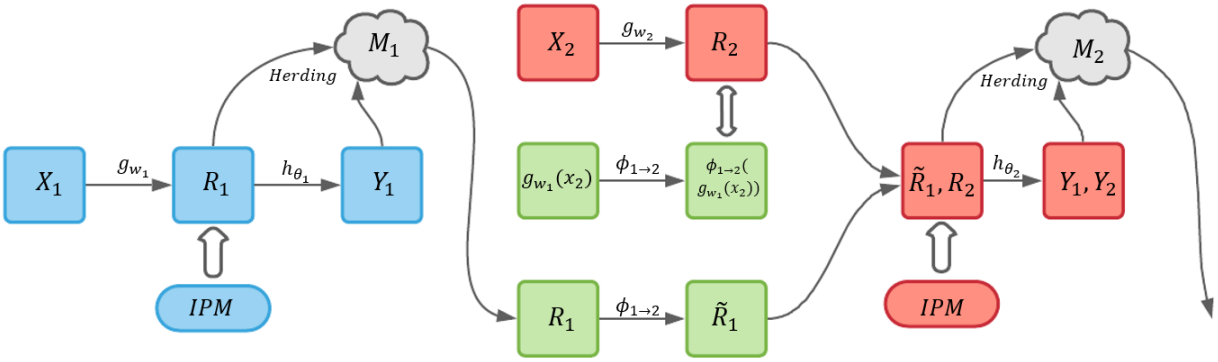

Figure 7.1: The framework of continual causal effect learning model.

Feature Representation Distillation. For next available dataset $D_2 = \{(x, y, t)|x \in X, y \in Y, t \in T\}$ collected from second domain, we adopt the same selective representation learning $g_{w_2} : X \to R_2$ with elastic net regularization (L_{w_2}) on new parameters w_2. Because we expect our model can estimate causal effect for both previous and new data, we want the new model to inherit some knowledge from previous model. In continual learning, knowledge distillation (Hinton et al., 2015; Z. Li & Hoiem, 2017) is commonly adopted to alleviate the catastrophic forgetting, where knowledge is transferred from one network to another network by encouraging the outputs of the original and new network to be similar. However, for the continual causal effect estimation problem, we focus more on the feature representations, which are required to be balanced between treatment and control, and among different data domains. Inspired by Dhar et al., 2019; Hou et al., 2019; Iscen et al., 2020, we propose feature representation distillation to encourage the representation vector $\{g_{w_1}(x)|x \in D_2\}$ based on baseline model to be similar to the representation vector $\{g_{w_2}(x)|x \in D_2\}$ based on new model by Euclidean distance. This feature distillation can help prevent the learned representations from drifting too much in the new feature representation space. Because we apply the cosine normalization to feature representations and $\|A - B\|^2 = (A-B)^\intercal(A-B) =$

$\|A\|^2 + \|B\|^2 - 2A^\intercal B = 2\big(1 - cos(A, B)\big)$, the feature representation distillation is defined as $L_{FD}(x) = 1 - cos\big(g_{w_1}(x), g_{w_2}(x)\big)$, where $x \in D_2$.

Feature Representation Transformation. We have previous feature representations R_1 stored in M_1 and new feature representations R_2 extracted from newly available data. R_1 and R_2 lie in different feature representation space and they are not compatible with each other because they are learned from different models. In addition, we cannot learn the feature representations of previous data from the new model g_{w_2}, as we no longer have access to previous data. Therefore, to balance the global feature representation space including previous and new representations between treatment and control groups, a feature transformation function is needed from previous feature representations R_1 to transformed feature representations $\tilde{R}_1$ compatible with new feature representations space R_2. We define a feature transformation function as $\phi_{1\to2} : R_1 \to \tilde{R}_1$. We also input the feature representations of new data D_2 learned from old model, i.e., $g_{w_1}(x)$, to get the transformed feature representations of new data, i.e., $\phi_{1\to2}(g_{w_1}(x))$. To keep the transformed space compatible with the new feature representation space, we train the transformation function $\phi_{1\to2}$ by making the $\phi_{1\to2}(g_{w_1}(x))$ and $g_{w_2}(x)$ similar, where $x \in D_2$. The loss function is defined as $L_{FT}(x) = 1 - cos\big(\phi_{1\to2}(g_{w_1}(x)), g_{w_2}(x)\big)$, which is used to train the function $\phi_{1\to2}$ to transform feature representations between different feature spaces. Then, we can attain the transformed old feature representations $\tilde{R}_1 = \phi_{1\to2}(R_1)$, which is in the same space as R_2.

Balancing Global Feature Representation Space. We have obtained a global feature representation space including the transformed representations of stored old data and new representations of new available data. We adopt the same integral probability metrics as baseline model to make sure that the representation distributions are balanced for treatment and control groups in the global feature representation space. In addition, we define a potential outcome function $h_{\theta_2} : (\tilde{R}_1, R_2) \times T \to Y$. Let $\hat{y}_i^M = h_{\theta_2}\big(\phi_{1\to2}(r_i), t\big)$, where $r_i \in M_1$, and $\hat{y}_j^D = h_{\theta_2}\big(g_{w_2}(x_j), t\big)$, where $x_j \in D_2$ denote the inferred observed outcomes.

We aim to minimize the mean squared error in predicting factual outcomes for global feature representations including transformed old feature representations and new feature representations: $L_G = \frac{1}{\tilde{n}_1}\sum_{i=1}^{\tilde{n}_1}(\hat{y}_i^M - y_i^M)^2 + \frac{1}{n_2}\sum_{j=1}^{n_2}(\hat{y}_j^D - y_j^D)^2$, where $\tilde{n}_1$ is the number of units stored in M_1 by herding algorithm, $y_i^M \in M_1$, and $y_j^D \in D_2$.

In summary, the objective function of our continual causal effect learning model is $L = L_G + \alpha Wass(P,Q) + \lambda L_{w_2} + \beta L_{FD} + \delta L_{FT}$, where α, λ, β, and δ denote the hyper-parameters controlling the trade-off among $Wass(P,Q)$, L_{w_2}, L_{FD}, L_{FT}, and L_G in the final objective function.

7.3.2 Overview of CERL

In the above sections, we have provided the baseline and continual causal effect learning models. When the continual causal effect learning model for the second data is trained, we can extract the $R_2 = \{g_{w_2}(x)|x \in D_2\}$ and $\tilde{R}_1 = \{\phi_{1\to2}(r)|r \in M_1\}$. We define a new memory set as $M_2 = \{R_2, Y_2, T_2\} \cup \phi_{1\to2}(M_1)$, where $\phi_{1\to2}(M_1)$ includes $\tilde{R}_1$ and the corresponding $\{Y, T\}$ stored in M_1. Similarly, to satisfy the pre-specified memory constraint, M_2 can be reduced by conducting the herding algorithm to store the same number of feature representations from treatment and control groups. We only store the new memory set M_2 and new model g_{w_2}, which are used to train the following model and balance the global feature representation space. It is unnecessary to store the original data (D_1 and D_2) any longer.

We follow the same procedure for the subsequently available observational data. When we obtain the new observational data D_d, we can train $h_{\theta_d}(g_{w_d})$ and $\phi_{d-1\to d} : R_{d-1} \to \tilde{R}_{d-1}$ based on the continual causal effect learning model. Besides, the new memory set is defined as: $M_d = \{R_d, Y_d, T_d\} \cup \phi_{d-1\to d}(M_{d-1})$. So far, our model $h_{\theta_d}(g_{w_d})$ can estimate causal effect

for all seen observational data regardless of the data source and it doesn't require access to previous data. As shown in Algorithm 1, we summarize the procedures of CERL as follows:

Algorithm 1: Continual Causal Effect Representation Learning

Data: Given d incrementally available observational data from D_1 to D_d
if $\{x, y, t\} \in D_1$ **then**
    *** Train baseline causal effect model $h_{\theta_1}(g_{w_1})$ ***
    $w_1, \theta_1 = \text{OPTIMIZE}(L_Y + \alpha Wass(P, Q) + \lambda L_{w_1})$
    $R_1 = \{g_{w_1}(x) | x \in D_1\}$
    $M_1 = \text{HERDING}\{R_1, Y_1, T_1\}$
else
    for $\{x, y, t\} \in D_2, ..., D_d$ **do**
        *** Train continual causal effect model $h_{\theta_d}(g_{w_d})$ ***
        $w_d, \theta_d, \phi_{d-1 \to d} = \text{OPTIMIZE}(L_G + \alpha Wass(P, Q) + \lambda L_{w_2} + \beta L_{FD} + \delta L_{FT})$
        $\tilde{R}_{d-1} = \phi_{d-1 \to d}(R_{d-1})$
        $R_d = \{g_{w_d}(x) | x \in D_d\}$
        $M_d = \text{HERDING}\big(\{R_d, Y_d, T_d\} \cup \{\tilde{R}_{d-1}, Y_{d-1} \in M_{d-1}, T_{d-1} \in M_{d-1}\}\big)$
    end
end

7.4 Experiments

We adapt the traditional benchmarks, i.e., News (F. Johansson et al., 2016; Schwab et al., 2018) and BlogCatalog (Guo, Li, & Liu, 2020) to continual causal effect estimation. Specifically, we consider three scenarios to represent the different degrees of domain shifts among the incrementally available observational data, including the substantial shift, moderate shift, and no shift. Besides, we generate a series of synthetic datasets and also conduct ablation studies to demonstrate the effectiveness of our model on multiple sequential datasets. The model performance with different numbers of preserved feature representations, and the robustness to hyperparameters are also evaluated.

7.4.1 Dataset Description

We utilize two semi-synthetic benchmarks for the task of continual causal effect estimation, which are based on real-world features, synthesized treatments and outcomes.

News. The News dataset consists of 5000 randomly sampled news articles from the NY Times corpus[1]. It simulates the opinions of media consumers on news items. The units are different news items represented by word counts $x_i \in \mathbb{N}^V$ and outcome $y(x_i) \in \mathbb{R}$ is the news item. The intervention $t \in \{0, 1\}$ represents the viewing device, desktop ($t = 0$) or mobile ($t = 1$). We extend the original dataset specification in F. Johansson et al., 2016; Schwab et al., 2018 to enable the simulation of incrementally available observational data with different degrees of domain shifts. Assuming consumers prefer to read certain media items on specific viewing devices, we train a topic model on a large set of documents and define $z(x)$ as the topic distribution of news item x. We define one topic distribution of a randomly sampled document as centroid z_1^c for mobile and the average topic representation of all document as centroid z_0^c for desktop. Therefore, the reader's opinion of news item x on device t is determined by the similarity between $z(x)$ and z_t^c, i.e., $y(x_i) = C(z(x)^\intercal z_0^c + t_i \cdot z(x)^\intercal z_1^c) + \epsilon$, where $C = 60$ is a scaling factor and $\epsilon \sim N(0, 1)$. Besides, the intervention t is defined by $p(t = 1|x) = \frac{e^{k \cdot z(x)^\intercal z_1^c}}{e^{k \cdot z(x)^\intercal z_0^c} + e^{k \cdot z(x)^\intercal z_1^c}}$, where $k = 10$ indicates an expected selection bias. In the experiments, 50 LDA topics are learned from the training corpus and 3477 bag-of-words features are in the dataset. To generate two sequential datasets with different domain shifts, we combine the news items belonging to LDA topics from 1 to 25 into first dataset and the news items belonging to LDA topics from 26 to 50 into second dataset. There is no overlap of the LDA topics between the first dataset and second dataset, which is considered as *substantial domain shift*. In addition, the news items belonging to LDA topics from 1 to 35 and items belonging to from 16 to 50 are used to construct the first dataset and second

[1] https://archive.ics.uci.edu/ml/datasets/bag+of+words

dataset, respectively, which is regarded as *moderate domain shift*. Finally, randomly sampled items from 50 LDA topics compose the first and second dataset, resulting in *no domain shift*, because they are from the same distribution. Under each domain shift scenario and each dataset, we randomly sample 60% and 20% of the units as the training set and validation set and let the remaining be the test set.

BlogCatalog. BlogCatalog (Guo, Li, & Liu, 2020) is a blog directory that manages the bloggers and their blogs. In this semi-synthetic dataset, each unit is a blogger and the features are bag-of-words representations of keywords in bloggers' descriptions collected from real-world source. We adopt the same settings and assumptions to simulate the treatment options and outcomes as we do for the News dataset. 50 LDA topics are learned from the training corpus. 5196 units and 2160 bag-of-words features are in the dataset. Similar to the generation procedure of News datasets with domain shifts, we create two datasets for each of the three domain shift scenarios. Under each domain shift scenario and each dataset, we randomly sample 60% and 20% of the units as the training set and validation set and let the remaining be the test set.

7.4.2 Results and Analysis

Evaluation Metrics. We adopt two commonly used evaluation metrics. The first one is the error of ATE estimation, which is defined as $\epsilon_{\text{ATE}} = |\text{ATE} - \widehat{\text{ATE}}|$, where ATE is the true value and $\widehat{\text{ATE}}$ is an estimated ATE. The second one is the error of expected precision in estimation of heterogeneous effect (PEHE) Hill, 2011, which is defined as $\epsilon_{\text{PEHE}} = \frac{1}{n}\sum_{i=1}^{n}(\text{ITE}_i - \widehat{\text{ITE}}_i)^2$, where ITE_i is the true ITE for unit i and $\widehat{\text{ITE}}_i$ is an estimated ITE for unit i.

We employ three strategies to adapt traditional causal effect estimation models to incrementally available observational data: (A) directly apply the model previously trained based on original data to new observational data; (B) utilize newly available data to fine-tune the

Table 7.1: Performance on two sequential data and M=500.

		News					**BlogCatalog**				
		Previous data			New data		Previous data			New data	
	Strategy	$\sqrt{\epsilon_{\text{PEHE}}}$	ϵ_{ATE}		$\sqrt{\epsilon_{\text{PEHE}}}$	ϵ_{ATE}	$\sqrt{\epsilon_{\text{PEHE}}}$	ϵ_{ATE}		$\sqrt{\epsilon_{\text{PEHE}}}$	ϵ_{ATE}
Substantial shift	CFR-A	2.49	0.80		**3.62**	**1.18** ↑	9.92	4.25		**13.65**	**6.21** ↑
	CFR-B	**3.23**	**1.06** ↑		2.71	0.91	**14.21**	**6.98** ↑		9.77	4.11
	CFR-C	2.51	0.82		2.70	0.92	9.93	4.24		9.77	4.12
	CERL	2.55	0.84		2.71	0.91	9.96	4.25		9.78	4.12
Moderate shift	CFR-A	2.58	0.85		**3.06**	**1.02** ↑	9.89	4.22		**11.26**	**5.03** ↑
	CFR-B	**2.98**	**0.99** ↑		2.65	0.92	**12.35**	**5.67** ↑		9.83	4.18
	CFR-C	2.56	0.85		2.63	0.90	9.88	4.21		9.81	4.16
	CERL	2.59	0.86		2.66	0.92	9.90	4.24		9.82	4.17
No shift	CFR-A	2.58	0.87		2.62	0.88	9.86	4.20		9.85	4.19
	CFR-B	2.60	0.88		2.60	0.87	9.85	4.18		9.83	4.18
	CFR-C	2.58	0.87		2.59	0.87	9.84	4.18		9.83	4.18
	CERL	2.59	0.87		2.60	0.87	9.85	4.19		9.83	4.18

Table 7.2: Performance on two sequential data and $M = 10000$.

	Previous data			New data		
Strategy	$\sqrt{\epsilon_{\text{PEHE}}}$	ϵ_{ATE}		$\sqrt{\epsilon_{\text{PEHE}}}$	ϵ_{ATE}	
CFR-A	1.47	0.35		2.51	0.73	↑
CFR-B	1.82	0.47	↑	1.63	0.45	
CFR-C	1.49	0.36		1.62	0.44	
CERL	1.49	0.37		1.63	0.44	
CERL (w/o FRT)	1.71	0.43	↑	1.63	0.44	
CERL (w/o herding)	1.57	0.40	↑	1.63	0.44	
CERL (w/o cosine norm)	1.51	0.38	↑	1.65	0.44	

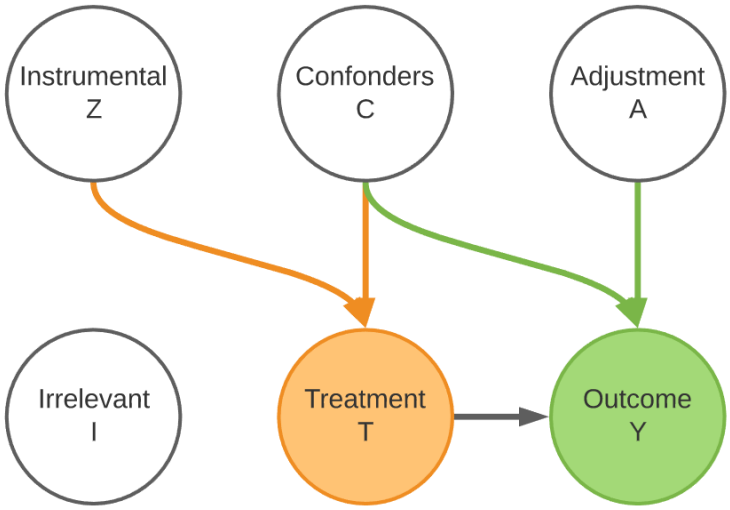

Figure 7.2: The relationship among different types of covariates.

previously learned model; (C) store all previous data and combine with new data to re-train the model from scratch. Among these three strategies, (C) is expected to be the best performer and get the ideal performance with respect to ATE and PEHE, although it needs to take up the most resources (all the data from previous and new dataset). We implement the three strategies based on the counterfactual regression model (CFR) (Shalit et al., 2017), which is a representative causal effect estimation method.

As shown in Table 7.1, under no domain shift scenario, the three strategies and our model have the similar performance on the News and BlogCatalog datasets, because the previous and new data are from the same distribution. CFR-A, CFR-B, and CERL need less resources than CFR-C. Under substantial shift and moderate shift scenarios, we find strategy CFR-A performs well on previous data, but significantly declines on new dataset; strategy CFR-B shows the catastrophic forgetting problem where the performance on previous dataset is poor; strategy CFR-C performs well on both previous and new data, but it re-trains the whole model using both previous and new data. However, if there is a memory constraint

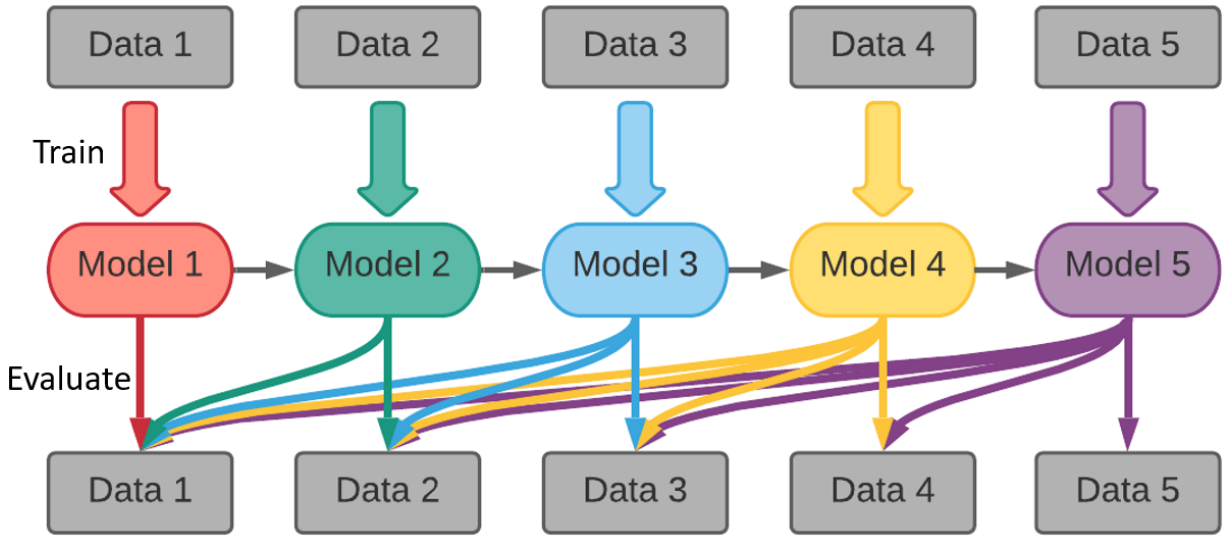

Figure 7.3: The work flow of continual learning task.

or a barrier to accessing previous data, the strategy CFR-C cannot be conducted. Our CERL has a similar performance to strategy CFR-C, while CERL does not require access to previous data. Besides, by comparing the performance under substantial and moderate shift scenarios, the larger domain shift leads to worse performance of CFR-A and CFR-B. However, no matter what the domain shift is, the performance of our model CERL is consistent with the ideal strategy CFR-C.

7.4.3 Model Evaluation

Synthetic Dataset. Our synthetic data include confounders, instrumental, adjustment, and irrelevant variables. The interrelations among these variables, treatments, and outcomes are illustrated in Figure 8.4. We totally simulate five different data sources with five different multivariate normal distributions to represent the incrementally available observational data. In each data source, we randomly draw 10000 samples including treatment units and control

units. Therefore, for five datasets, they have different selection bias, magnitude of covariates, covariance matrices for variables, and number of treatment and control units. To ensure a robust estimation of model performance, for each data source, we repeat the simulation procedure 10 times and obtain 10 synthetic datasets.

Results. Similar to the experiments for News and BlogCatalog benchmarks, we still utilize two sequential datasets to compare our model with CFR under three strategies on the more complex synthetic data. As shown in Table 8.1, the result is consistent with the conclusions on News and BlogCatalog. Our model's performance demonstrates its superiority over CFR-A and CFR-B. CERL is comparable with CFR-C, while it does not need to have access to the raw data from previous dataset. Besides, we also conduct three ablation studies to test the effectiveness of the important components in CERL, i.e., CERL (w/o FRT), CERL (w/o herding), and CERL (w/o cosine norm). CERL (w/o FRT) is the simplified CERL without the feature representation transformation, which is based on traditional continual learning with knowledge distillation. Because the previous feature representation is not stored or transformed into new feature space, we only utilize new data to balance the bias between treatment and control groups. CERL (w/o herding) adopts random subsampling strategy to select samples into memory, instead of herding algorithm. CERL (w/o cosine norm) removes the cosine normalization in the last representation layer. Table 8.1 shows that the performance becomes poor after removing anyone in the feature representation transformation, herding, or cosine normalization modules compared to the original CERL. More specifically, after removing the feature representation transformation, $\sqrt{\epsilon_{\text{PEHE}}}$ and ϵ_{ATE} increase dramatically, which demonstrates that the knowledge distillation always used in continual learning task is not enough for the continual causal effect estimation. Also, using herding to select a representative set of samples from treatment and control distributions is crucial for the feature representation transformation.

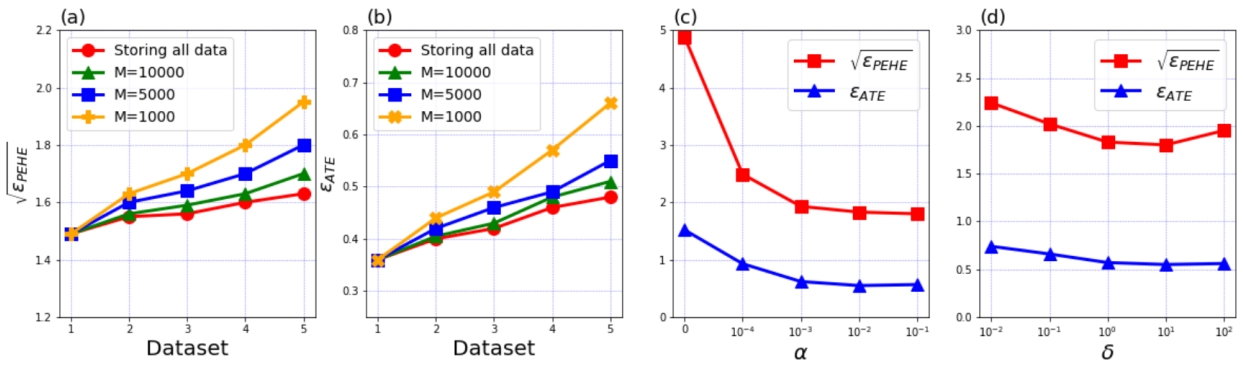

Figure 7.4: Performance of CERL under different settings.

CERL Performance Evaluation. As illustrated in Figure 7.3, the five observational data are incrementally available in sequence, and the model will continue to estimate the causal effect without having access to previous data. We further evaluate the performance of CERL from three perspectives, i.e., the impact of memory constraint, effeteness of cosine normalization, and its robustness to hyper-parameters. As shown in Figure 7.4 (a) and (b), as the model continually learns a new dataset, every time when finishing training one new dataset, we report the $\sqrt{\epsilon_{\text{PEHE}}}$ and ϵ_{ATE} on test sets composed of previous data and new data. Our model with memory constraints has a similar performance to the ideal situation, where all data are available to train the model from scratch. However, our model can effectively save memory space, e.g., when facing the fifth dataset, our model only stores 1000, 5000, or 10000 feature representations, but the ideal situation needs to store $5 \times 10000 = 50000$ observations with all covariates. For the cosine normalization, we perform an ablation study of CERL (M=5000, 5 datasets), where we remove cosine normalization in the representation learning procedure. We find the $\sqrt{\epsilon_{\text{PEHE}}}$ increases from 1.80 and 1.92 and ϵ_{ATE} from 0.55 to 0.61. Next, we explore the model's sensitivity to the most important parameter α and δ, which controls the representation balance and representation transformation. From Fig. 7.4

(c) and (d), we observe that the performance is stable over a large parameter range. In addition, the parameter β for feature representation distillation is set to 1 (Iscen et al., 2020; Rebuffi et al., 2017).

7.5 Summary

It is the first time to propose the continual lifelong causal effect inference problem and the corresponding evaluation criteria. As the real world evidence is becoming more prominent, how to integrate and utilize these powerful data for causal effect estimation becomes a new research challenge. To address this challenge, we propose the Continual Causal Effect Representation Learning method for estimating causal effect with observational data, which are incrementally available from non-stationary data distributions. Extensive experiments demonstrate the superiority of our method over baselines for continual causal effect estimation.

Chapter 8

Learning Informative and Domain-Independent Representations for Causal Effect Inference

8.1 Introduction

Nowadays, the study of causal effect estimation is much facilitated by the dramatically growing availability of observational data. Although a huge amount of observational data is accumulated to conduct treatment effect estimation, it brings a new challenge, i.e., missing counterfactual outcomes, compared with randomized controlled trials (RCT). As the RCT is conducted, the only expected difference between the treatment and control groups is the outcome variable being studied. However, when estimating causal effects with observational data, we only observe one factual outcome and never all potential outcomes that would potentially have happened had we chosen other treatment options. Due to the hallmark of

observational data, subjects would have a preference for a certain treatment option, which leads to a bias of the distribution for the covariates among different treatment options. The selection bias makes the distribution of the covariates in the treatment group different from the control group, and such a huge discrepancy between the treatment and control groups exacerbates the difficulty of counterfactual outcome estimation. Therefore, how to handle the selection bias is a challenging problem in causal effect estimation.

Recent causal effect estimation methods (F. Johansson et al., 2016; S. Li & Fu, 2017a; Shalit et al., 2017) have built a strong connection with domain adaptation, by enforcing domain invariance with distributional distances such as the Wasserstein distance and maximum mean discrepancy. Inspired by metric learning, some methods (Yao et al., 2018) use hard samples to learn representations that preserve local similarity information and balance the data distributions. In (Y. Zhang et al., 2020), the authors argue that distribution invariance is often too strict a requirement, and they propose to use counterfactual variance to measure the domain overlap. Thus, which is the best measurement for the imbalanced domains remains unsettled and the choice highly relies on the characteristics of the domain distributions (Yao et al., 2020). Besides, despite the empirical success of such methods, enforcing balance can, to various extents, remove predictive information and lead to a loss in predictive power, regardless of which type of domain divergence metric is employed (A. Alaa & Schaar, 2018).

Besides, when handling the selection bias, there is another issue that can lead to poor potential outcome estimation, i.e., the types of observed variables. The major drawback of existing causal inference methods is that they always treat all observed variables as pre-treatment variables, which are not affected by treatment assignments but may be predictive of outcomes. This assumption is not tenable for observational data. If all observed variables are directly used to estimate treatment effects, more impalpable bias may be introduced into the model. For example, conditioning on an instrumental variable, which is associated with

the treatment assignment but not with the outcome, can increase both bias and variance of estimated treatment effects (Myers et al., 2011).

To sum up, successfully estimating the causal effect needs three desiderata, i.e., filtering out information about instrumental and irrelevant variables, capturing the predictive information, and mitigating the covariate imbalance between treatment and control groups. To achieve these three desiderata simultaneously, we propose an Informative and Domain-Independent Representation Learning (IDRL) method to estimate the causal effects with observational data by seeking a representation space, which not only contains the common predictive information about potential outcome estimation but also excludes the domain-dependent information. IDRL relies on two mutual information structures: one is to maximize the mutual information between global summary representation and individual feature representation, which can maximally capture the common predictive information for both treatment and control groups and filter out the noise only for specific individual or group; the other is to minimize the mutual information between feature representation vector and treatment options, which makes feature representations independent from treatment option domains.

Our main contributions are summarized in the following: Our work utilizes the global summary representation to capture the common predictive information for both treatment and control groups; Circumventing the strategy of enforcing balance between treatment and control groups (adopting various domain divergence metrics), our IDRL method learns the domain-independent representation to solve the selection bias problem in causal effect estimation.

8.2 Background

Suppose that the observational data contain n units and that each unit received one of two or more treatments. Let t_i denote the treatment assignment for unit i; $i = 1, ..., n$. For binary treatments, $t_i = 1$ is for the treatment group, and $t_i = 0$ for the control group. The outcome for unit i is denoted by Y_t^i when treatment t is applied to unit i; that is, Y_1^i is the potential outcome of unit i in the treatment group and Y_0^i is the potential outcome of unit i in the control group. For observational data, only one of the potential outcomes is observed as the actual treatment assignment of unit i. The observed outcome is called the factual outcome, and the remaining unobserved potential outcomes are called counterfactual outcomes. Let $X \in \mathbb{R}^d$ denote all observed variables of a unit. Then the observational data can be denoted as $\{x_i, t_i, y_i\}_{i=1}^n$.

We follow the potential outcome framework for estimating treatment effects (Rubin, 1974) and the strong ignorability assumption, which ensures that the treatment effect can be identified (G. W. Imbens & Rubin, 2015b).

Assumption 8.2.1. ***Strong Ignorability****: Given covariates X, treatment assignment T is independent of the potential outcomes, i.e., $(Y_1, Y_0) \perp\!\!\!\perp T|X$ and for any value of X, treatment assignment is not deterministic, i.e., $P(T = t|X = x) > 0$, for all t and x.*

8.3 Proposed Framework

8.3.1 Motivation

The most challenging issue in causal effect estimation from observational data is how to properly handle the covariate shift between treatment and control groups caused by treatment selection bias. This phenomenon exacerbates the difficulty of counterfactual outcome

estimation. In the traditional distribution balancing strategies, different domain divergence metrics are sensitive to the characteristics of data distributions and the predictive information may be inadvertently removed when enforcing the balance procedure. Besides, the instrumental and irrelevant variables also can bring bias and variance into the counterfactual outcome estimation. To address the above issues, we propose the **Informative and Domain-Independent Representation Learning (IDRL)** method.

8.3.2 Model Architecture

As shown in Fig. 8.1, our IDRL method consists of four main components, including feature representation learning, information maximization learning, domain-independent learning, and potential outcome generator. In the feature representation learning, IDRL first learns an individual representation vector for each subject via the standard feed-forward deep neural network. At the same time, the information maximization learning and domain-independent learning are incorporated into the representation learning procedure to filter out domain-dependent information, solve the selection bias, and preserve the common predictive information for treatment and control groups. Finally, the potential outcomes can be inferred by the outcome generator based on the learned representation.

Feature Representation Learning. This step is to learn the feature representations of observed covariates by a function $g : X \to R, R \in \mathbb{R}^d$, which is parameterized by a deep neural network. The function $g(\cdot)$ maps the original covariate space X into a d-dimensional representation space $R = \{r_1, r_2, ..., r_n\}$.

Information Maximization Learning. Inspired by a recent unsupervised representation learning method that exploits individual and global information (Hjelm et al., 2018; Velickovic et al., 2019), we maximize the mutual information between the individual representation and global representation, such that the representation space could capture the

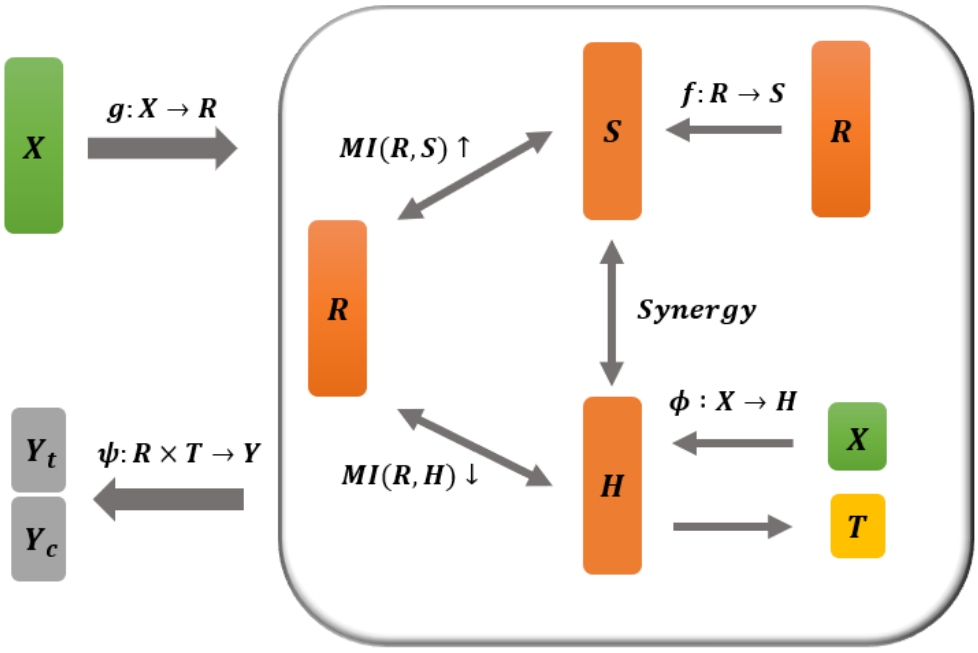

Figure 8.1: The framework of the proposed IDRL.

common predictive information for treatment and control groups. Specifically, we utilize a summary function, $f : R^{n\times d} \rightarrow S, S \in \mathbb{R}^d$, which summarizes the learned individual representation into a global representation vector, i.e., $s = f(g(X))$. From the observations in empirical evaluations, the summary function could be defined as weighted averaging of all the subjects' representations: $s = \sigma(\frac{1}{2n_t}\sum_{i\in n_t} r_i + \frac{1}{2n_c}\sum_{i\in n_c} r_i)$ to best capture the global representation, where σ is the logistic sigmoid activation function, n_t and n_c are the subject numbers of treatment and control groups, respectively. Our goal is to make all the subjects' representations preserve the common predictive information used to estimate the potential outcomes for treatment and control groups. Therefore, we aim at maximizing the mutual information $MI(r_i, s)$ between the learned individual representation r_i and global summary representation s, where the feature representation learning can pick and choose what type of information in the original covariates is preserved into the learned representation vector.

If the feature representation learning passes information specific to only some individuals or certain treatment options, this does not increase the mutual information with any of the other subjects. This encourages the feature representation learning to prefer information that is shared across the subjects.

Domain-Independent Learning. To handle the selection bias, we incorporate a domain-independent learning module, which helps the subject's feature representation to be independent of its domain, instead of enforcing the domain invariance by various metrics (e.g., Wasserstein distance, maximum mean discrepancy). When the feature representations are independent of the domains, we cannot tell which domain the subject is from and thus filter out the information about the treatment assignments. Because the mutual information is small when the two variables are statistically independent, while it is large when two variables preserve the same information content, we employ the mutual information to measure the independence between feature representations and domains. To give full expression to the treatment domain information, we utilize the treatment domain prediction H to represent the treatment domain by function $\phi : X \to H \to T$, rather than directly using treatment domain indicator T. Therefore, we aim at minimizing the mutual information $MI(r_i, h_i)$ between learned representation space r_i and treatment domain prediction h_i.

Potential Outcome Generator. So far, we have learned the feature representation space from feature representation learning, along with information maximization learning and domain-independent learning. The function $\psi : R \times T \to Y$ maps the representation vectors as well as the treatment assignment to the corresponding potential outcome, which is parameterized by a feed-forward deep neural network with multiple hidden layers and non-linear activation functions. To avoid the risk of losing the influence of T when the dimension of representation space is high, $\psi : R \times T \to Y$ is partitioned into two head layers for treatment and control groups, separately. The output of ψ estimates potential outcomes across treatment and control groups, including the estimated factual outcome $\hat{y}^f$ and the

estimated counterfactual outcomes $\hat{y}^{cf}$. The factual outcomes y^f are used to minimize the loss of prediction $\hat{y}^f$. We aim to minimize the mean squared error in predicting factual outcomes:

$$\mathcal{L}_Y = \frac{1}{n}\sum_{i=1}^{N}(\hat{y}_i^f - y_i^f)^2, \tag{8.3.1}$$

where $\hat{y}_i = \psi(r_i, t_i)$ denotes the inferred observed outcome of unit i corresponding to the factual treatment t_i.

Mutual Information Estimation. Mutual information is a fundamental quantity for measuring the relationship between random variables. For example, the dependence of two random variables W and Z is quantified by mutual information as (Belghazi, Baratin, Rajeswar, et al., 2018):

$$MI(W;Z) = \int_{\mathcal{W}\times\mathcal{Z}} \log\frac{d\mathbb{P}_{WZ}}{d\mathbb{P}_W \otimes \mathbb{P}_Z} d\mathbb{P}_{WZ}, \tag{8.3.2}$$

where $\mathbb{P}_{WZ}$ is the joint probability distribution, and $\mathbb{P}_W \otimes \mathbb{P}_Z$ are the product of marginals $\mathbb{P}_W = \int_{\mathcal{W}} d\mathbb{P}_{WZ}$ and $\mathbb{P}_Z = \int_{\mathcal{Z}} d\mathbb{P}_{WZ}$. The mutual information is small when the two variables W and Z are statistically independent, while is large when two variables preserve the same information content.

However, mutual information has historically been difficult to compute. From Shannon information theory, mutual information can be estimated as the Kullback-Leibler divergence (D_{KL}) between the joint distribution $\mathbb{P}_{WZ}$ and the product of their marginal distributions $\mathbb{P}_W \otimes \mathbb{P}_Z$:

$$MI(W;Z) = D_{KL}(\mathbb{P}_{WZ}||\mathbb{P}_W \otimes \mathbb{P}_Z). \tag{8.3.3}$$

Actually, in our method, it is unnecessary to use the exact KL-based formulation of MI, as we only want to maximize the mutual information $MI(r_i, s)$ between individual representation r_i and global representation s, and minimize the mutual information $MI(r_i, h_i)$

between individual representation r_i and treatment domain prediction h_i. A simple and stable alternative based on the Jensen-Shannon divergence (JSD) can be utilized. Some recent work (Belghazi, Baratin, Rajeswar, et al., 2018; Hjelm et al., 2018) has proved that an implicit estimation of mutual information can be attained with an encoder-discriminator architecture. Thus, we follow the intuitions from Deep Infomax (Hjelm et al., 2018) to optimize the mutual information involved in our method. To act as an agent for optimizing the mutual information, one discriminator is employed, which relies on a sampling strategy that draws positive and negative samples from the joint distribution and the marginal product, respectively. To implement the discriminator, we need to create the negative samples compared with the original samples, and then use the discriminator to distinguish which one is from positive samples (original data) and which one is from the negative samples (created fake data). The choice of the negative sampling procedure will govern the specific kinds of information that is desirable to be captured (Velickovic et al., 2019). Under causal inference settings, our main challenge is the covariate shift caused by selection bias, so we independently shuffle feature variables of positive samples X to generate negative samples $\tilde{X}$ as shown in Fig. 8.2, which can break the imbalanced feature variable patterns in original X. Thus, for each feature variable in $\tilde{X}$, there is no imbalance with respect to treatment options.

For $MI(r_i, s)$, one discriminator $d_s : R \times S \rightarrow P, P \in \mathbb{R}$ is employed. The discriminator is formulated by a simple bilinear scoring function with nonlinear activation: $d_s(r_i, s) = \sigma({r_i}^T W s)$, which estimates the probability of the i-th subject representation contained within the global representation s. W is a learnable scoring matrix. We also conduct the feature representation learning for the negative samples $\tilde{X}$ to get the $\tilde{r}_i$. With the proposed discriminator, we could have $d_s(r_i, s)$ and $d_s(\tilde{r}_i, s)$, which indicate the probabilities of containing the representations of the i-th positive sample and negative sample in the global summary representation, respectively. We optimize the discriminator d_s to maximize

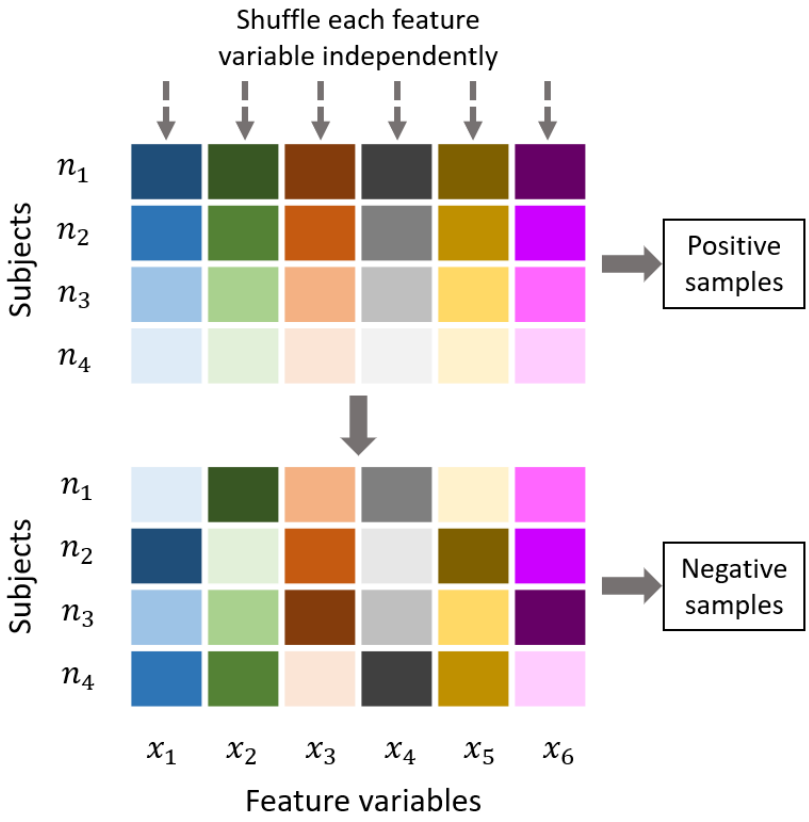

Figure 8.2: The strategy of generating negative samples.

mutual information between r_i and s based on the Jensen Shannon divergence via a noise-contrastive type objective with a standard binary cross-entropy (BCE) loss (Hjelm et al., 2018; Velickovic et al., 2019). The $\mathcal{L}_{MI(r_i,s)}$ is defined as:

$$\frac{1}{2n}\Big(\sum_{i=1}^{n}\mathbb{E}_X[\log d(r_i,s)]+\sum_{j=1}^{n}\mathbb{E}_{\tilde{X}}[\log(1-d(\tilde{r}_j,s))]\Big). \tag{8.3.4}$$

For $MI(r_i,h_i)$, one discriminator $d_h : R\times H\to P, P\in\mathbb{R}$ is adopted. Our method aims at optimizing the discriminator d_h, i.e., minimizing the mutual information between learned representation space r_i and treatment domain prediction h_i. Similarly, the discriminator

is formulated by a simple bilinear scoring function with nonlinear activation: $d_h(r_i, h_i) = \sigma({r_i}^T W h_i)$, where r_i is the i-th subject's representation learned in feature representation learning procedure and h_i is the i-th subject's treatment domain prediction learned from function ϕ. We also attain the treatment domain prediction $\tilde{h}_i$ for negative samples $\tilde{X}$ by function ϕ. Therefore, in discriminator d_h, we could have $d_h(r_i, h_i)$ and $d_h(r, \tilde{h}_i)$, so the $\mathcal{L}_{MI(r_i,h_i)}$ is defined as:

$$\frac{1}{2n}\Big(\sum_{i=1}^{n}\mathbb{E}_X[\log d(r_i, h_i)] + \sum_{j=1}^{n}\mathbb{E}_{\tilde{X}}[\log(1 - d(r_j, \tilde{h}_j))]\Big), \tag{8.3.5}$$

8.3.3 Overview of IDRL

The proposed IDRL method leverages the synergy between two mutual information modules to filter out the treatment domain information and noise, and thus capture the common predictive information for both treatment and control groups. In this way, our method can effectively increase the capability of predicting potential outcomes. As shown in Fig. 8.3, we summarize the procedures of IDRL as follows:

1. Create the negative samples $\tilde{X}$ by independently shuffling feature variables of positive samples X.

2. Learn the representation space R for the positive samples X and $\tilde{R}$ for the negative samples $\tilde{X}$ by function $g : X \to R$ and $g : \tilde{X} \to \tilde{R}$, respectively.

3. Learn the treatment domain prediction H for the positive samples (X, T) by function $\phi : X \to H \to T$ and $\tilde{H}$ for the negative samples by plugging $\tilde{X}$ into function ϕ.

4. Utilize a summary function $f : R^{n\times d} \to S$ to summarize the learned representation into a global summary representation, i.e., $s = f(g(X))$.

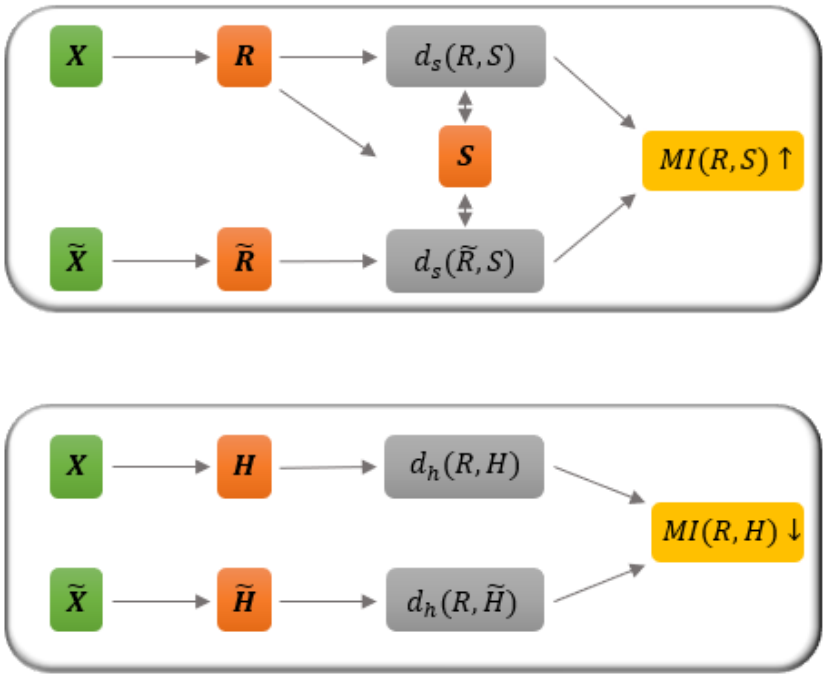

Figure 8.3: The procedures of IDRL.

5. Employ the discriminator $d_s : R \times S \to P$ to obtain $d_s(r_i, s)$ and $d_s(\tilde{r}_i, s)$, and the discriminator $d_h : R \times H \to P$ to obtain $d_h(h_i, r_i)$ and $d_h(\tilde{h}_i, r_i)$.

6. Update parameters of g, f, d_s, and d_h to maximize mutual information between R and S and minimize mutual information between R and H, by applying gradient descent to maximize Eq. (8.3.4) and minimize Eq. (8.3.5).

7. Use potential outcome generator $\psi : R \times T \to Y$ to estimate the potential outcomes by minimizing Eq. (8.3.1)

8.4 Experiments

In this section, we conduct experiments on three benchmarks, including the IHDP, Jobs, and News with multiple treatment options, to compare our proposed IDRL method with the state-of-the-art causal effect estimation methods. We also experiment with the synthetic datasets with different settings to validate the following aspects: (1) Capability of reliably predicting potential outcomes when facing data with different characteristics of the domain distributions and complicated variable types. (2) Robustness with respect to different levels of treatment selection bias. (3) The contribution of each component of the proposed IDRL method.

8.4.1 Experiments on Benchmark Datasets

Datasets and Settings. To evaluate our method and baselines on treatment effect estimation, we use the binary treatment benchmarks, i.e., IHDP and Jobs datasets, and a multiple treatment benchmark News dataset (with 2, 4, 8, and 16 treatment options). We compare our IDRL method with the following baseline methods: kNN (D. E. Ho et al., 2007), CF (Wager & Athey, 2018b), RF (Breiman, 2001), BART (Chipman et al., 2010), TARNET (Shalit et al., 2017), $\text{CFRNET}_{\text{wass}}$ (Shalit et al., 2017), SITE (Yao et al., 2018), PM (Schwab et al., 2018), CMGP (A. M. Alaa & van der Schaar, 2017). For IHDP and Jobs, we report in-sample and out-of-sample performance with $\sqrt{\epsilon_{\text{PEHE}}}$ and ϵ_{ATE}, and the policy risk R_{pol} and ϵ_{ATT}, respectively. For News dataset with multiple treatments, we only report the performance on the test sets with $\sqrt{\epsilon_{\text{mPEHE}}}$ and ϵ_{mATE}.

Results and Analysis. Table 8.1 and 8.2 show the performance of our method and baseline methods on the IHDP, Jobs, and News with different treatment options. Our method significantly outperforms all competing algorithms on the Jobs dataset and News datasets.

Table 8.1: Performance on IHDP and Jobs of IDRL and competing methods.

	IHDP				Jobs			
	In-sample		Out-sample		In-sample		Out-sample	
Method	$\sqrt{\epsilon_{\text{PEHE}}}$	ϵ_{ATE}	$\sqrt{\epsilon_{\text{PEHE}}}$	ϵ_{ATE}	R_{pol}	ϵ_{ATT}	R_{pol}	ϵ_{ATT}
kNN	2.1	0.14	4.1	0.79	0.23	0.02	0.26	0.13
RF	4.2	0.73	6.6	0.96	0.23	0.03	0.28	0.09
CF	3.8	0.18	3.8	0.4	0.19	0.03	0.2	0.07
BART	2.1	0.23	2.3	0.34	0.23	0.02	0.25	0.08
TARNET	0.88	0.26	0.95	0.28	0.17	0.05	0.21	0.11
CFRNET	0.71	0.25	0.76	0.27	0.17	0.04	0.21	0.08
PM	n.r.	n.r.	0.84	0.24	n.r.	n.r.	0.18	0.16
SITE	0.69	0.22	0.75	0.24	0.17	0.04	0.21	0.09
CMGP	**0.65**	**0.11**	0.77	**0.13**	0.22	0.06	0.24	0.09
IDRL (Ours)	0.68	0.18	**0.73**	0.20	**0.13**	**0.02**	**0.16**	**0.04**

On the IHDP dataset, our method has the best performance in the out-sample case and achieves comparable results with the best baselines, such as CFRNET, SITE, and CMGP in the in-sample case. Besides, the encouraging results on the News datasets with multiple treatments show that our method is capable of handling the treatment selection bias from multiple domains.

8.4.2 Experiments on Synthetic Datasets

We further evaluate the performance of our method when facing data with different characteristics of the domain distributions, complicated variable types, and severe covariate imbalance. In addition, we evaluate the contribution of each component in our method.

Synthetic Dataset. To reflect the complexity of observational data, our synthetic data include confounders C, instrumental variables Z, adjustment A, and irrelevant variables I. The interrelations among these variables, treatments and outcomes are illustrated in

Table 8.2: Performance on News with 2, 4, 8, and 16 treatments of IDRL and competing methods.

	News-2		News-4		News-8		News-16	
Method	$\sqrt{\epsilon_{\text{PEHE}}}$	ϵ_{ATE}	$\sqrt{\epsilon_{\text{mPEHE}}}$	ϵ_{mATE}	$\sqrt{\epsilon_{\text{mPEHE}}}$	ϵ_{mATE}	$\sqrt{\epsilon_{\text{mPEHE}}}$	ϵ_{mATE}
kNN	18.14	7.83	27.92	19.40	26.20	15.11	27.64	17.27
RF	17.39	5.5	26.59	18.03	23.77	12.4	26.13	15.91
CF	17.59	4.02	23.86	13.54	22.56	9.7	21.45	8.37
BART	18.53	5.4	26.41	17.14	25.78	14.8	27.45	17.5
TARNET	17.17	4.58	23.40	13.63	22.39	9.38	21.19	8.3
CFRNET	16.93	4.54	22.65	12.96	21.64	8.79	20.87	8.05
PM	16.76	3.99	21.58	10.04	20.76	6.51	20.24	5.76
IDRL (Ours)	**16.41**	**3.23**	**21.12**	**9.33**	**19.98**	**5.83**	**19.64**	**4.66**

Fig. 8.4. The model used to generate the continuous outcome variable Y in this simulation is the partially linear regression model (Eq. (8.4.1)), extending the ideas described in (Jacob et al., 2019; Robinson, 1988):

$$Y = \tau((C^\intercal, A^\intercal)^\intercal)T + g((C^\intercal, A^\intercal)^\intercal) + \epsilon, \tag{8.4.1}$$

where $T \overset{ind.}{\sim} \text{Bernoulli}(e_0((C^\intercal, Z^\intercal)^\intercal))$.

Results and Analysis. As shown in Table 8.3, our method significantly outperforms competitive baselines, such as TARNET, SITE, CFRNET, and CMGP. These baseline methods all rely on the assumption that all observed variables are pre-treatment variables. However, this assumption is not tenable for observational data in practice, which may include instrumental, adjustment, confounding, and irrelevant variables, as in our simulated dataset. Besides, we conduct ablation studies and report the performance of our method without the Information Maximization Module or Domain-Independent Module, respectively. From the

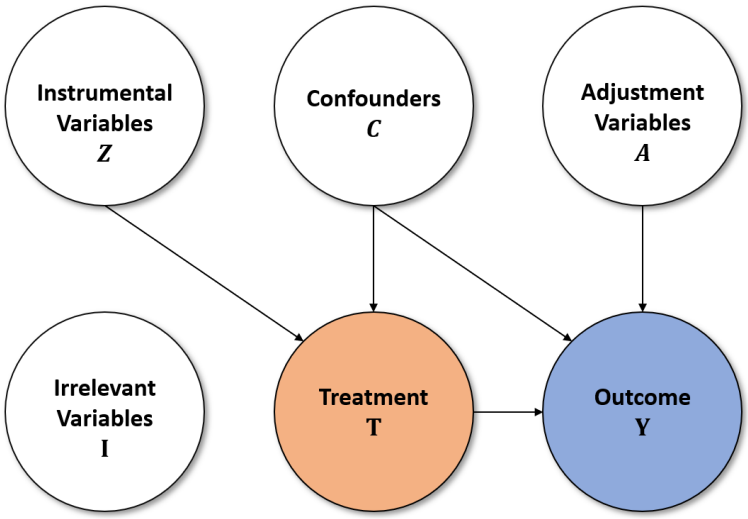

Figure 8.4: The types of observed variables.

results, the performance suffers when either is left out, which demonstrates the effectiveness of these two modules in our method.

Besides, we evaluate the robustness of our proposed method with respect to different levels of treatment selection bias. If the propensity score e_0 is equal to constant 0.5, there is no treatment selection bias. The greater $|e_0((C^\intercal, Z^\intercal)^\intercal) - 0.5|$ is, the larger selection bias will end up getting. Following the setting in (Shalit et al., 2017), with probability $1 - q$, we randomly draw the treatment and control units; with probability q, we draw the treatment and control units that have the greatest $|e_0((C^\intercal, Z^\intercal)^\intercal) - 0.5|$. Thus, the higher the q is, the larger the selection bias is. We run CFRNET and our method on the simulation datasets with q from 0 to 1, and show the results in Fig. 8.5. We can observe that our method consistently outperforms the baseline methods under different levels of divergence and is robust to a high level of treatment bias.

To evaluate the sensitivity of our method to different characteristics of the domain distributions and the capability of handling the instrumental and irrelevant variables, we gen-

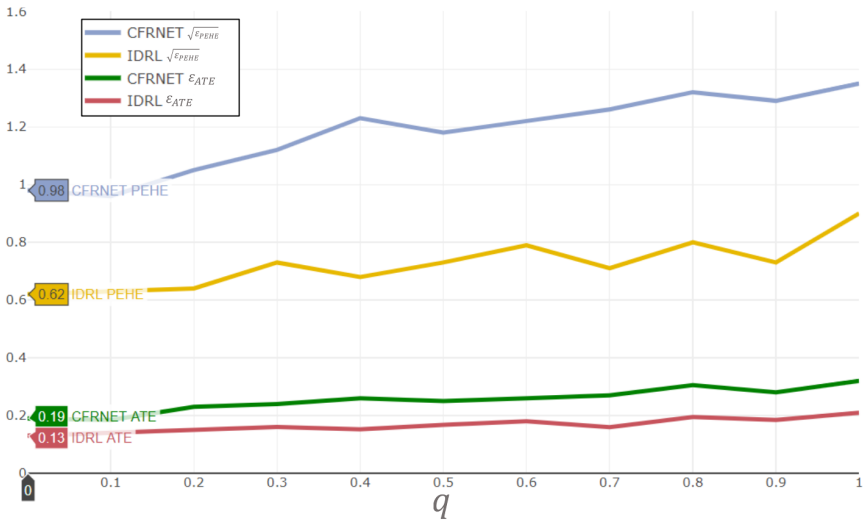

Figure 8.5: Performance on simulation dataset with q from 0 to 1.

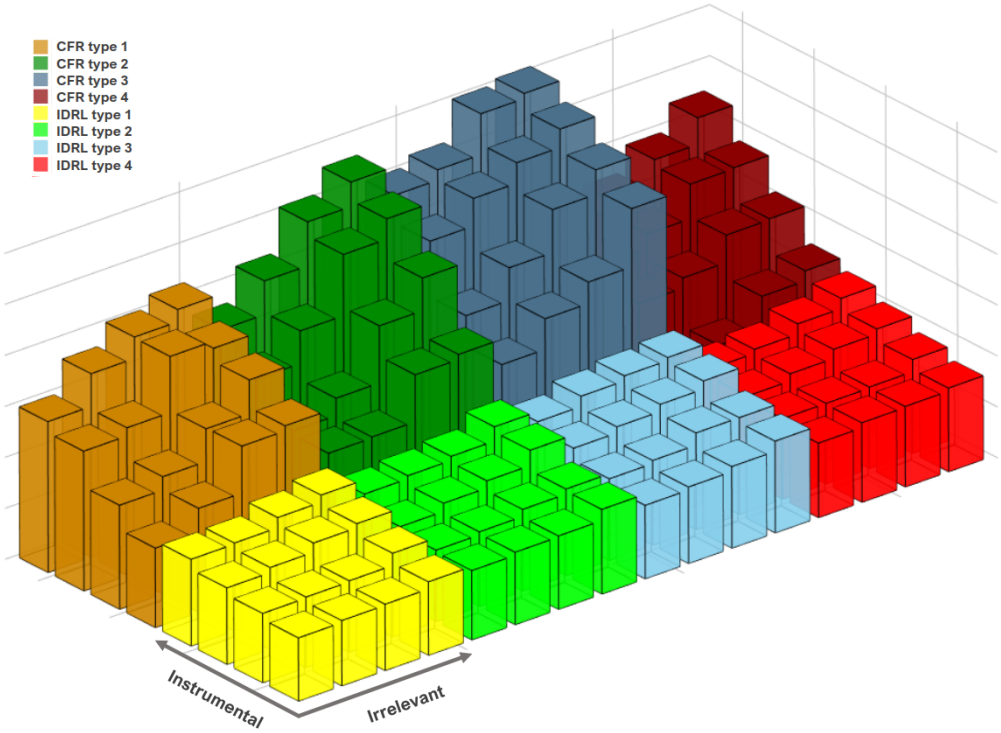

Figure 8.6: Performance of CFRNET (Top row) and IDRL (Bottom row) on simulated dataset under different settings.

Table 8.3: Performance on simulated dataset of IDRL and competing methods.

Method	$\sqrt{\epsilon_{\text{PEHE}}}$	ϵ_{ATE}
TARNET	1.36	0.31
SITE	1.15	0.25
CFRNET	0.98	0.19
CMGP	0.83	0.17
IDRL w/o $MI(r_i, s)$	0.89	0.17
IDRL w/o $MI(r_i, h_i)$	1.24	0.27
IDRL	0.62	0.13

erate four different domain distributions with different correlation matrices for all observed variables in the simulated dataset and gradually increase the numbers of instrumental and irrelevant variables, respectively. As shown in Fig. 8.6, we can observe our IDRL consistently outperforms the baseline CFRNET under any situation. Our method is robust to four types of different domain distributions, and the increase of instrumental and irrelevant variables.

8.5 Related Work

Learning from observational data requires adjusting for the covariate shift that exists between treatment and control groups. Balancing neural networks (BNNs) (F. Johansson et al., 2016) and counterfactual regression networks (CFRNET) (Shalit et al., 2017) are proposed to balance covariate distributions across treatment and control groups by formulating the problem of counterfactual inference as a domain adaptation problem and by enforcing domain invariance with distributional distances such as Wasserstein distance and Maximum Mean Discrepancy. A local similarity preserved individualized treatment effect (SITE) estimation method (Yao et al., 2018) is proposed to use hard samples to learn representations that preserve local similarity information and balance the data distributions. In (Y. Zhang et

al., 2020), the authors argue that domain invariance is often too strict a requirement and use counterfactual variance to measure the distributional overlap. In (A. M. Alaa & van der Schaar, 2017), the covariate shift problem caused by selection bias is alleviated via a risk-based empirical Bayes method by minimizing the empirical error in factual outcomes and the uncertainty in counterfactual outcomes.

8.6 Summary

In this chapter, we propose the Informative and Domain-independent Representation Learning (IDRL) method for treatment effect estimation with observational data. IDRL offers a new thought in probing into the covariate imbalance problem in causal inference. Circumventing the traditional strategy of enforcing balance between treatment and control groups, IDRL leverages the mutual information to capture the common predictive information and handle the selection bias, which has been verified by extensive experiments on multiple benchmark datasets.

Chapter 9

Conclusion

Most studies across many fields, such as statistics, computer science, public policy, and economics, aim to solve causal rather than associative problems. Causal inference has been an attractive research topic for a long time as it provides an effective way to uncover causal relationships in real-world problems. The representative method is randomized controlled trial where the assignment of treatment or control is random. Therefore, The only expected difference between the treatment and control groups is the outcome variable being studied. A well-blinded randomized controlled trial is often considered the gold standard for studying causal relationships. However, in reality, randomized controlled trials are always time-consuming and expensive, and thus the study cannot involve many subjects, which may be not representative of the real-world population a treatment/intervention would eventually target. Another issue is that the randomized controlled trials only focus on the average of samples, and it does not explain the mechanism or pertain for individual subjects. In addition, ethical issues also need to be considered in most of the randomized controlled trials, which largely limits its applications. Therefore, instead of the randomized controlled trials, the observational data is a tempting shortcut. Observational data is obtained by the researcher simply observing the subjects without any interfering. That means, the researchers

have no control over treatments and subjects, and they just observe the subjects and record data based on their observations.

In this dissertation, we provide a comprehensive review of the methods under the well-known potential outcome framework and propose several novel models to deal with the challenges of the causal inference with observational data. Extensive experiment results and theory analysis demonstrate the superiority of our methods while facing some causal inference problems. Despite substantial progress in the area of causal inference with observational data, there are still many important problems to be solve. For example the causal interpretation and explainability based on the deep learning models (Moraffah et al., 2020), most of these models are like the black-boxes, so it is hard to understand how the decisions are made in the models, which makes the models unreliable and untrustworthy. Besides, the existing methods only focus on source-specific and stationary observational data (Z. Chu et al., 2021). Such learning strategies assume that all observational data are already available during the training phase and from the only one source. This assumption is unsubstantial in practice due to two reasons. The first one is based on the characteristics of observational data, which are incrementally available from non-stationary data distributions. The second reason is based on the realistic consideration of accessibility, e.g., legacy data may be unrecorded, proprietary, too large to store, or subject to privacy constraint (J. Zhang et al., 2020). Therefore, how to solve the continual learning of causal inference is one open question. In addition, unlike the traditional treatment effect defined on individual experimental units, it maybe focuses on a group of units. For example, teachers may choose groups of students who do not know each other to teach a new curriculum. This is called treatment entanglement (Toulis et al., 2018). It is a special case where individual treatments depend on a common population quantity and how to deal with this new problem is also challenging.

Bibliography

Abadie, A., Drukker, D., Herr, J. L., & Imbens, G. W. (2004). Implementing matching estimators for average treatment effects in stata. *The stata journal*, *4*(3), 290–311.

Alaa, A., & Schaar, M. (2018). Limits of estimating heterogeneous treatment effects: Guidelines for practical algorithm design. *International Conference on Machine Learning*, 129–138.

Alaa, A. M., & van der Schaar, M. (2017). Bayesian inference of individualized treatment effects using multi-task gaussian processes. *Advances in Neural Information Processing Systems*, 3424–3432.

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

Altman, N., & Krzywinski, M. (2015). Points of significance: Association, correlation and causation. *Nature Methods*, *12*(10), 899–900.

Astsaturov, I. (2017). Future clinical trials: Genetically driven trials. *Surgical Oncology Clinics*, *26*(4), 791–797.

Athey, S., & Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, *113*(27), 7353–7360.

Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate behavioral research*, *46*(3), 399–424.

Bang, H., & Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics*, *61*(4), 962–973.

Barron, A. R. (1994). Approximation and estimation bounds for artificial neural networks. *Machine learning*, *14*(1), 115–133.

Belghazi, M. I., Baratin, A., Rajeshwar, S., Ozair, S., Bengio, Y., Courville, A., & Hjelm, D. (2018). Mutual information neural estimation. *International Conference on Machine Learning*, 531–540.

Belghazi, M. I., Baratin, A., Rajeswar, S., Ozair, S., Bengio, Y., Courville, A., & Hjelm, R. D. (2018). Mine: Mutual information neural estimation.

Ben-David, S., Blitzer, J., Crammer, K., & Pereira, F. (2007). Analysis of representations for domain adaptation. *Advances in neural information processing systems*, 137–144.

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, *35*(8), 1798–1828.

Berry, S. M., Broglio, K. R., Groshen, S., & Berry, D. A. (2013). Bayesian hierarchical modeling of patient subpopulations: Efficient designs of phase ii oncology clinical trials. *Clinical Trials*, *10*(5), 720–734.

Bishop, C. M. et al. (1995). *Neural networks for pattern recognition*. Oxford university press.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5–32.

Breiman, L. (2017). *Classification and regression trees*. Routledge.

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

Caliendo, M., & Kopeinig, S. (2008). Some practical guidance for the implementation of propensity score matching. *Journal of economic surveys*, *22*(1), 31–72.

CDCReproductiveHealth. (2020). *Centers for disease control and prevention. severe maternal morbidity in the united states 2020*.

Chang, Y., & Dy, J. G. (2017). Informative subspace learning for counterfactual inference. *Thirty-First AAAI Conference on Artificial Intelligence*.

Chipman, H. A., George, E. I., & McCulloch, R. E. (2007). Bayesian ensemble learning. *Advances in neural information processing systems*, 265–272.

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

Chu, Y., & Yuan, Y. (2018). A bayesian basket trial design using a calibrated bayesian hierarchical model. *Clinical Trials*, *15*(2), 149–158.

Chu, Z., Rathbun, S., & Li, S. (2021). Continual lifelong causal effect inference with real world evidence [under review]. *Submitted to International Conference on Learning Representations*. https://openreview.net/forum?id=IOqr2ZyXHz1

Chu, Z., Rathbun, S. L., & Li, S. (2020). Matching in selective and balanced representation space for treatment effects estimation. *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, 205–214.

Cochran, W. G. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics*, 295–313.

Cunanan, K. M., Gonen, M., Shen, R., Hyman, D. M., Riely, G. J., Begg, C. B., & Iasonos, A. (2017). Basket trials in oncology: A trade-off between complexity and efficiency. *Journal of Clinical Oncology*, *35*(3), 271.

D'Agostino Jr, R. B. (1998). Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Statistics in medicine*, *17*(19), 2265–2281.

Dhar, P., Singh, R. V., Peng, K.-C., Wu, Z., & Chellappa, R. (2019). Learning without memorizing. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5138–5146.

Dinh, V., & Ho, L. S. T. (2020). Consistent feature selection for analytic deep neural networks. *arXiv preprint arXiv:2010.08097*.

Fan, J., Imai, K., Liu, H., Ning, Y., & Yang, X. (2016). *Improving covariate balancing propensity score: A doubly robust and efficient approach* (tech. rep.). Technical report, Princeton Univ.

Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, *96*(456), 1348–1360.

Fingar, K. R., Hambrick, M. M., Heslin, K. C., & Moore, J. E. (2018). Trends and disparities in delivery hospitalizations involving severe maternal morbidity, 2006–2015: Statistical brief# 243.

Food, U., Administration, D. et al. (2007). Guidance for industry: Clinical trial endpoints for the approval of cancer drugs and biologics. *Federal Register*, *72*.

French, R. M. (1999). Catastrophic forgetting in connectionist networks. *Trends in cognitive sciences*, *3*(4), 128–135.

Garrido, M. M. (2016). Covariate adjustment and propensity score. *Jama*, *315*(14), 1521–1522.

Geller, S. E., Koch, A. R., Garland, C. E., MacDonald, E. J., Storey, F., & Lawton, B. (2018). A global view of severe maternal morbidity: Moving beyond maternal mortality. *Reproductive health*, *15*(1), 31–43.

Goodfellow, I., Bengio, Y., Courville, A., & Bengio, Y. (2016). *Deep learning* (Vol. 1). MIT press Cambridge.

Greenland, S. (2008). Invited commentary: Variable selection versus shrinkage in the control of multiple confounders. *American journal of epidemiology*, *167*(5), 523–529.

Gu, X. S., & Rosenbaum, P. R. (1993). Comparison of multivariate matching methods: Structures, distances, and algorithms. *Journal of Computational and Graphical Statistics*, *2*(4), 405–420.

Guo, R., Li, J., Li, Y., Candan, K. S., Raglin, A., & Liu, H. (2020). Ignite: A minimax game toward learning individual treatment effects from networked observational data.

Guo, R., Li, J., & Liu, H. (2019). Learning individual treatment effects from networked observational data. *arXiv preprint arXiv:1906.03485*.

Guo, R., Li, J., & Liu, H. (2020). Learning individual causal effects from networked observational data. *Proceedings of the 13th International Conference on Web Search and Data Mining*, 232–240.

Hahn, P. R., Murray, J. S., & Carvalho, C. (2017). Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects.

Hansen, B. B. (2008). The prognostic analogue of the propensity score. *Biometrika*, *95*(2), 481–488.

Harder, V. S., Stuart, E. A., & Anthony, J. C. (2010). Propensity score techniques and the assessment of measured covariate balance to test causal associations in psychological research. *Psychological methods*, *15*(3), 234.

Hardin, J., Garcia, S. R., & Golan, D. (2013). A method for generating realistic correlation matrices. *The Annals of Applied Statistics*, 1733–1762.

Hassanpour, N., & Greiner, R. (2019). Counterfactual regression with importance sampling weights. *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, 5880–5887.

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. *Proceedings of the IEEE international conference on computer vision*, 1026–1034.

Heckman, J. J., Ichimura, H., & Todd, P. (1998). Matching as an econometric evaluation estimator. *The review of economic studies*, *65*(2), 261–294.

Hernán, M. A., & Robins, J. M. (2006). Estimating causal effects from epidemiological data. *Journal of Epidemiology & Community Health*, *60*(7), 578–586.

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

Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Hirakawa, A., Asano, J., Sato, H., & Teramukai, S. (2018). Master protocol trials in oncology: Review and new trial designs. *Contemporary clinical trials communications*, *12*, 1–8.

Hjelm, R. D., Fedorov, A., Lavoie-Marchildon, S., Grewal, K., Bachman, P., Trischler, A., & Bengio, Y. (2018). Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*.

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political analysis*, *15*(3), 199–236.

Ho, D. E., Imai, K., King, G., Stuart, E. A., et al. (2011). Matchit: Nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software, http://gking. harvard. edu/matchit*.

Hou, S., Pan, X., Loy, C. C., Wang, Z., & Lin, D. (2019). Learning a unified classifier incrementally via rebalancing. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 831–839.

Hullsiek, K. H., & Louis, T. A. (2002). Propensity score modeling strategies for the causal analysis of observational data. *Biostatistics*, *3*(2), 179–193.

Iacus, S. M., King, G., & Porro, G. (2012). Causal inference without balance checking: Coarsened exact matching. *Political analysis*, *20*(1), 1–24.

Imai, K., & Ratkovic, M. (2014). Covariate balancing propensity score. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *76*(1), 243–263.

Imbens, G. W. (2004a). Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and statistics*, *86*(1), 4–29.

Imbens, G. W. (2004b). Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and statistics*, *86*(1), 4–29.

Imbens, G. W., & Rubin, D. B. (2015a). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Imbens, G. W., & Rubin, D. B. (2015b). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Imbens, G. W., & Wooldridge, J. M. (2009). Recent developments in the econometrics of program evaluation. *Journal of economic literature*, *47*(1), 5–86.

Iscen, A., Zhang, J., Lazebnik, S., & Schmid, C. (2020). Memory-efficient incremental learning through feature adaptation. *arXiv preprint arXiv:2004.00713*.

Jacob, D., Härdle, W. K., & Lessmann, S. (2019). Group average treatment effects for observational studies. *arXiv preprint arXiv:1911.02688*.

Ji, S., Kollár, J., & Shiffman, B. (1992). A global łojasiewicz inequality for algebraic varieties. *Transactions of the American Mathematical Society*, *329*(2), 813–818.

Johansson, F., Shalit, U., & Sontag, D. (2016). Learning representations for counterfactual inference. *International conference on machine learning*, 3020–3029.

Johansson, F. D., Kallus, N., Shalit, U., & Sontag, D. (2018). Learning weighted representations for generalization across designs. *arXiv preprint arXiv:1802.08598*.

Judea Pearl. (2012). Judea pearl on potential outcomes.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kipf, T. N., & Welling, M. (2016). Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Langfelder, P., Zhang, B., & Horvath, S. (2008). Defining clusters from a hierarchical cluster tree: The dynamic tree cut package for r. *Bioinformatics*, *24*(5), 719–720.

Lee, B. K., Lessler, J., & Stuart, E. A. (2011). Weight trimming and propensity score weighting. *PloS one*, *6*(3), e18174.

Lee, C., Mastronarde, N., & van der Schaar, M. (2018). Estimation of individual treatment effect in latent confounder models via adversarial learning.

Li, F., Morgan, K. L., & Zaslavsky, A. M. (2018). Balancing covariates via propensity score weighting. *Journal of the American Statistical Association*, *113*(521), 390–400.

Li, S., & Fu, Y. (2017a). Matching on balanced nonlinear representations for treatment effects estimation. *Advances in Neural Information Processing Systems*, 929–939.

Li, S., & Fu, Y. (2017b). Matching on balanced nonlinear representations for treatment effects estimation. *Advances in Neural Information Processing Systems*, 929–939.

Li, S., Vlassis, N., Kawale, J., & Fu, Y. (2016). Matching via dimensionality reduction for estimation of treatment effects in digital marketing campaigns. *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, 3768–3774.

Li, Y., Chen, C.-Y., & Wasserman, W. W. (2016). Deep feature selection: Theory and application to identify enhancers and promoters. *Journal of Computational Biology*, *23*(5), 322–336.

Li, Z., & Hoiem, D. (2017). Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, *40*(12), 2935–2947.

Loh, W.-Y. (2011). Classification and regression trees. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *1*(1), 14–23.

Louizos, C., Shalit, U., Mooij, J. M., Sontag, D., Zemel, R., & Welling, M. (2017). Causal effect inference with deep latent-variable models. *Advances in Neural Information Processing Systems*, 6446–6456.

Luo, C., Zhan, J., Xue, X., Wang, L., Ren, R., & Yang, Q. (2018). Cosine normalization: Using cosine similarity instead of dot product in neural networks. *International Conference on Artificial Neural Networks*, 382–391.

Ma, X., & Wang, J. (2010). Robust inference using inverse probability weighting. *Journal of the American Statistical Association*, 1–10.

Mansour, Y., Mohri, M., & Rostamizadeh, A. (2009). Domain adaptation: Learning bounds and algorithms. *arXiv preprint arXiv:0902.3430*.

McCloskey, M., & Cohen, N. J. (1989). Catastrophic interference in connectionist networks: The sequential learning problem. *Psychology of learning and motivation* (pp. 109–165). Elsevier.

Meinshausen, N., Bühlmann, P. et al. (2006). High-dimensional graphs and variable selection with the lasso. *Annals of statistics*, *34*(3), 1436–1462.

Mersch, J., Jackson, M. A., Park, M., Nebgen, D., Peterson, S. K., Singletary, C., Arun, B. K., & Litton, J. K. (2015). Cancers associated with brca 1 and brca 2 mutations other than breast and ovarian. *Cancer*, *121*(2), 269–275.

Mhaskar, H. N. (1996). Neural networks for optimal approximation of smooth and analytic functions. *Neural computation*, *8*(1), 164–177.

Moraffah, R., Karami, M., Guo, R., Raglin, A., & Liu, H. (2020). Causal interpretability for machine learning-problems, methods and evaluation. *ACM SIGKDD Explorations Newsletter*, *22*(1), 18–33.

Müller, A. (1997). Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 429–443.

Myers, J. A., Rassen, J. A., Gagne, J. J., Huybrechts, K. F., Schneeweiss, S., Rothman, K. J., Joffe, M. M., & Glynn, R. J. (2011). Effects of adjusting for instrumental variables on bias and precision of effect estimates. *American journal of epidemiology*, *174*(11), 1213–1222.

Park, J. J., Siden, E., Zoratti, M. J., Dron, L., Harari, O., Singer, J., Lester, R. T., Thorlund, K., & Mills, E. J. (2019). Systematic review of basket trials, umbrella trials, and platform trials: A landscape analysis of master protocols. *Trials*, *20*(1), 1–10.

Patrick, A. R., Schneeweiss, S., Brookhart, M. A., Glynn, R. J., Rothman, K. J., Avorn, J., & Stürmer, T. (2011). The implications of propensity score variable selection strategies in pharmacoepidemiology: An empirical illustration. *Pharmacoepidemiology and drug safety*, *20*(6), 551–559.

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

Pearl, J. (2009a). Causal inference in statistics: An overview. *Statistics surveys*, *3*, 96–146.

Pearl, J. (2009b). *Causality*. Cambridge university press.

Pearl, J. (2014). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. Elsevier.

Pourzanjani, A. A., Jiang, R. M., & Petzold, L. R. (2017). Improving the identifiability of neural networks for bayesian inference. *NIPS Workshop on Bayesian Deep Learning*, *4*, 29.

Rebuffi, S.-A., Kolesnikov, A., Sperl, G., & Lampert, C. H. (2017). Icarl: Incremental classifier and representation learning. *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2001–2010.

Robins, J., Sued, M., Lei-Gomez, Q., & Rotnitzky, A. (2007). Comment: Performance of double-robust estimators when" inverse probability" weights are highly variable. *Statistical Science*, *22*(4), 544–559.

Robins, J. M., Rotnitzky, A., & Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, *89*(427), 846–866.

Robinson, P. M. (1988). Root-n-consistent semiparametric regression. *Econometrica: Journal of the Econometric Society*, 931–954.

Rosenbaum, P. R. (1987a). Model-based direct adjustment. *Journal of the American Statistical Association*, *82*(398), 387–394.

Rosenbaum, P. R. (1987b). Model-based direct adjustment. *Journal of the American Statistical Association*, *82*(398), 387–394.

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

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

Rosenbaum, P. R., & Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, *39*(1), 33–38.

Rubin, D. B. (1973). Matching to remove bias in observational studies. *Biometrics*, 159–183.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, *66*(5), 688.

Rubin, D. B., & Thomas, N. (2000). Combining propensity score matching with additional adjustments for prognostic covariates. *Journal of the American Statistical Association*, *95*(450), 573–585.

Samet, S., Miri, A., & Granger, E. (2013). Incremental learning of privacy-preserving bayesian networks. *Applied Soft Computing*, *13*(8), 3657–3667.

Sauer, B. C., Brookhart, M. A., Roy, J., & VanderWeele, T. (2013). A review of covariate selection for non-experimental comparative effectiveness research. *Pharmacoepidemiology and drug safety*, *22*(11), 1139–1145.

Scardapane, S., Comminiello, D., Hussain, A., & Uncini, A. (2017). Group sparse regularization for deep neural networks. *Neurocomputing*, *241*, 81–89.

Scharfstein, D., Rotnitzky, A., & Robins, J. (1999). Comments and rejoinder. *Journal of the American Statistical Association*, *94*(448), 1121–1146.

Schisterman, E. F., Cole, S. R., & Platt, R. W. (2009). Overadjustment bias and unnecessary adjustment in epidemiologic studies. *Epidemiology (Cambridge, Mass.)*, *20*(4), 488.

Schwab, P., Linhardt, L., Bauer, S., Buhmann, J. M., & Karlen, W. (2019). Learning counterfactual representations for estimating individual dose-response curves. *arXiv preprint arXiv:1902.00981*.

Schwab, P., Linhardt, L., & Karlen, W. (2018). Perfect match: A simple method for learning representations for counterfactual inference with neural networks. *arXiv preprint arXiv:1810.00656*.

Shalit, U., Johansson, F. D., & Sontag, D. (2017). Estimating individual treatment effect: Generalization bounds and algorithms. *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, 3076–3085.

Shortreed, S. M., & Ertefaie, A. (2017). Outcome-adaptive lasso: Variable selection for causal inference. *Biometrics*, *73*(4), 1111–1122.

Simon, R. (2017). Critical review of umbrella, basket, and platform designs for oncology clinical trials. *Clinical Pharmacology & Therapeutics*, *102*(6), 934–941.

Smith, J. (2000). *A critical survey of empirical methods for evaluating active labor market policies* (tech. rep.). Research Report.

Splawa-Neyman, J., Dabrowska, D. M., & Speed, T. (1990). On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science*, 465–472.

Sriperumbudur, B. K., Fukumizu, K., Gretton, A., Schölkopf, B., Lanckriet, G. R., et al. (2012). On the empirical estimation of integral probability metrics. *Electronic Journal of Statistics*, *6*, 1550–1599.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *The journal of machine learning research*, *15*(1), 1929–1958.

Stephen, M., & Christopher, W. (2007). *Counterfactuals and causal inference: Methods and principles for social research*. Cambridge University Press Cambridge, UK.

Strzebonska, K., & Waligora, M. (2019). Umbrella and basket trials in oncology: Ethical challenges. *BMC medical ethics*, *20*(1), 58.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical science: a review journal of the Institute of Mathematical Statistics*, *25*(1), 1.

Tao, J. J., Schram, A. M., & Hyman, D. M. (2018). Basket studies: Redefining clinical trials in the era of genome-driven oncology. *Annual review of medicine*, *69*, 319–331.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288.

Toulis, P., Volfovsky, A., & Airoldi, E. M. (2018). Propensity score methodology in the presence of network entanglement between treatments. *arXiv preprint arXiv:1801.07310*.

Veitch, V., Wang, Y., & Blei, D. (2019). Using embeddings to correct for unobserved confounding in networks. *Advances in Neural Information Processing Systems*, 13769–13779.

Velickovic, P., Fedus, W., Hamilton, W. L., Liò, P., Bengio, Y., & Hjelm, R. D. (2019). Deep graph infomax. *ICLR (Poster)*.

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., & Bengio, Y. (2017). Graph attention networks. *arXiv preprint arXiv:1710.10903*.

Wager, S., & Athey, S. (2018a). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, *113*(523), 1228–1242. https://doi.org/10.1080/01621459.2017.1319839

Wager, S., & Athey, S. (2018b). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, *113*(523), 1228–1242.

Wang, C., Parmigiani, G., & Dominici, F. (2012). Bayesian effect estimation accounting for adjustment uncertainty. *Biometrics*, *68*(3), 661–671.

Wang, P., Sun, W., Yin, D., Yang, J., & Chang, Y. (2015). Robust tree-based causal inference for complex ad effectiveness analysis. *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, 67–76.

Wang, Y., & Blei, D. M. (2019). The blessings of multiple causes. *Journal of the American Statistical Association*, *114*(528), 1574–1596.

Welling, M. (2009). Herding dynamical weights to learn. *Proceedings of the 26th Annual International Conference on Machine Learning*, 1121–1128.

Westreich, D., Lessler, J., & Funk, M. J. (2010). Propensity score estimation: Machine learning and classification methods as alternatives to logistic regression. *Journal of clinical epidemiology*, *63*(8), 826.

WHOMaternalMortality. (2020). *World health organization. maternal mortality 2020*.

Wilson, A., & Reich, B. J. (2014). Confounder selection via penalized credible regions. *Biometrics*, *70*(4), 852–861.

Yao, L., Chu, Z., Li, S., Li, Y., Gao, J., & Zhang, A. (2020). A survey on causal inference. *arXiv preprint arXiv:2002.02770*.

Yao, L., Li, S., Li, Y., Huai, M., Gao, J., & Zhang, A. (2018). Representation learning for treatment effect estimation from observational data. *Advances in Neural Information Processing Systems*, 2633–2643.

Yao, L., Li, S., Li, Y., Huai, M., Gao, J., & Zhang, A. (2019). Ace: Adaptively similarity-preserved representation learning for individual treatment effect estimation. *2019 IEEE International Conference on Data Mining*, 1432–1437.

Yao, L., Li, S., Li, Y., Xue, H., Gao, J., & Zhang, A. (2019). On the estimation of treatment effect with text covariates. *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, 4106–4113.

Yoon, J., Jordon, J., & van der Schaar, M. (2018). GANITE: estimation of individualized treatment effects using generative adversarial nets. *6th International Conference on Learning Representations*.

Zhang, B., & Horvath, S. (2005). A general framework for weighted gene co-expression network analysis. *Statistical applications in genetics and molecular biology*, *4*(1).

Zhang, J., Zhang, J., Ghosh, S., Li, D., Tasci, S., Heck, L., Zhang, H., & Kuo, C.-C. J. (2020). Class-incremental learning via deep model consolidation. *The IEEE Winter Conference on Applications of Computer Vision*, 1131–1140.

Zhang, Y., Bellot, A., & van der Schaar, M. (2020). Learning overlapping representations for the estimation of individualized treatment effects. *arXiv preprint arXiv:2001.04754*.

Zhao, L., Hu, Q., & Wang, W. (2015). Heterogeneous feature selection with multi-modal deep neural networks and sparse group lasso. *IEEE Transactions on Multimedia*, *17*(11), 1936–1948.

Zhu, W., Lan, C., Xing, J., Zeng, W., Li, Y., Shen, L., & Xie, X. (2016). Co-occurrence feature learning for skeleton based action recognition using regularized deep lstm networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, *30*(1).

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association*, *101*(476), 1418–1429.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, *67*(2), 301–320.