Causal Inference for Scientific Discoveries and Fairness-Aware Machine Learning

Yoichi Chikahara

Abstract

Causal inference is a problem of uncovering the mechanism of real-world phenomena that determines how each variable influences another. This problem can be mainly separated into the two tasks. One is causal discovery, which infers the directions and the presence of causal relationships between variables. The other is treatment effect estimation, which estimates the causality strength as how greatly manipulating a cause variable (called a *treatment*) changes its effect (referred to as an *outcome*).

Solving these two causal inference tasks enables us to elucidate the underlying mechanism in scientific phenomena and hence has been actively studied in various fields of science, such as bioinformatics, chemical engineering, meteorology, neuroscience, epidemiology, and economics. In addition, recent causal inference applications include trustworthy machine learning. In particular, improving the fairness of machine learning predictions based on causality has received increasing attention because it widens the range of machine learning applications to decision-making against individuals, such as hiring, loan approval, and child abuse screening.

Therefore, understanding the mechanism via a lens of causality is a promising approach to advancing a step toward accelerating scientific discoveries and making fair machine learning predictions. However, despite many efforts, inferring the causality from observational data still remains a challenging problem. Due to this difficulty, the existing techniques suffer from many methodological limitations, which hinder scientific discoveries and fairness-aware machine learning.

This dissertation is devoted to establishing the causal inference frameworks for accelerating scientific discoveries and improving the reliability of machine learning predictions. To accomplish these two goals, we make the following contributions.

First, we improve the inference accuracy of causal discovery from time series data by developing a data augmentation framework. In particular, we propose a *supervised learning* framework that infers the causal relationship underlying in *test*

data by utilizing *training data*, whose causal relationships are obvious and known. We experimentally show that such a data augmentation framework can effectively deal with the data scarcity issue and the complex nonlinearity between variables, both of which are common in many fields of science, such as bioinformatics.

We then consider to develop an interpretable approach to treatment effect estimation. In particular, to elucidate why the treatment effects are different across individuals, we establish a feature selection framework for discovering the features related to the treatment effect heterogeneity. We formulate a feature importance measure using a distributional discrepancy measure, which enables us to discover a wider range of features than the existing methods. By applying our feature selection framework to the medical survey data, we have successfully found an important feature attribute that could not be detected by the existing method, which demonstrates the effectiveness in accelerating scientific discoveries.

We finally establish a causality-based learning framework for making accurate and fair predictions against individuals. In the field of machine learning and fairness, the causality-based approaches have attracted increasing attention as a promising framework for striking a good balance between fairness and prediction accuracy. However, developing such a framework is challenging due to the difficulty of estimating the causality-based unfairness measure. To overcome this difficulty, we derive the upper bound on the unfairness measure called the probability of individual unfairness (PIU) and imposing a penalty on it to train an accurate and individually fair classifier. We reveal that why such a penalty guarantees individual-level fairness and present several extensions to address the complicated real-world scenarios.

Acknowledgements

First and foremost, I would like to express my deepest gratitude to my supervisor, Hisashi Kashima. Despite his busy schedule, he has provided continued supports and insightful comments, which are valuable throughout the period of my Ph.D. studies.

I would like to show my gratitude to the collaborators of my Ph.D. studies. Akinori Fujino has shared his experience and ideas for improving the paper presentation in terms of clarity and logical flow. Shinsaku Sakaue has a strong enthusiasm for various fields in mathematics and has actively joined in many research discussions with me. These research discussions form a theoretical basis of my Ph.D. studies. Makoto Yamada keeps trying hard even in the research fields that are unfamiliar for him. I have benefited tremendously from such an active attitude, which I believe is essential to become a leading researcher who gets many people involved.

My appreciation goes to the dissertation committee members, Hidetoshi Shimodaira and Akihiro Yamamoto. Their constructive comments are helpful to shape my future work direction.

I am also grateful to my former supervisors: Akira Funahashi, Noriko F. Hiroi, Satoru Miyano, Seiya Imoto, and Rui Yamaguchi. All of them motivated me to the world of scientific studies (in particular, bioinformatics and systems biology), which has provided a starting point of my Ph.D. studies.

Finally, I would like to thank my family and my friends for their heartwarming supports and encouragement.

List of Publications

Publications Included in this Dissertation

This dissertation includes the work of the following five publications [Chikahara and Fujino, 2018a,b; Chikahara *et al.*, 2021, 2022].

Refereed Conference Proceedings

1. **(Chapter 3)**: Yoichi Chikahara and Akinori Fujino.
   Causal Inference in Time Series via Supervised Learning.
   In *Proceedings of the 27-th International Conference on Artificial Intelligence (IJCAI)*, 2042–2048, 2018.
   https://doi.org/10.24963/ijcai.2018/282

2. **(Chapter 4)**: Yoichi Chikahara, Makoto Yamada, and Hisashi Kashima.
   Feature Selection for Discovering Distributional Treatment Effect Modifiers.
   In *Proceedings of the 38-th International Conference on Uncertainty in Artificial Intelligence (UAI)*, 400–410, 2022.
   https://proceedings.mlr.press/v180/chikahara22a.html

3. **(Chapter 5)**: Yoichi Chikahara, Shinsaku Sakaue, Akinori Fujino, and Hisashi Kashima.
   Learning Individually Fair Classifier with Path-Specific Causal-Effect Constraint.
   In *Proceedings of the 24-th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 145–153, 2021.
   https://proceedings.mlr.press/v130/chikahara21a.html

Refereed Journal Articles

1. **(Chapter 3)**: Yoichi Chikahara and Akinori Fujino.
   A Supervised Learning Approach to Granger Causality Inference.
   Information Processing Society of Japan (IPSJ) Transactions on Mathematical Modeling and its Applications (TOM), 11(3), 58–73, 2018.
   http://id.nii.ac.jp/1001/00192870/

2. **(Chapter 5)**: Yoichi Chikahara, Shinsaku Sakaue, Akinori Fujino, and Hisashi Kashima.
   Making Individually Fair Predictions with Causal Pathways.
   Accepted and soon to be published in *Data Mining and Knowledge Discovery. Special Issue on Bias and Fairness in AI* published by Springer Nature.

Contents

Chapter 1

Introduction

1.1 Causal Discovery

Causation can be rephrased as a *mechanism* that determines the real-world phenomena. Elucidating such a mechanism itself contributes to **scientific discoveries**. As such, many scientists have dedicated tremendous efforts for causal discovery, which aims at inferring the directions and the existence of the causal relationships between random variables. The inference targets in science include the gene regulatory network in bioinformatics [Kleinberg and Hripcsak, 2011], the reaction mechanism in chemical engineering [Ting and Barnard, 2022], the atmospheric teleconnections in meteorology [Kretschmer *et al.*, 2021], and the modulation mechanism of neuronal activity in neuroscience [Bergmann and Hartwigsen, 2021].

However, it is challenging to infer the causal relationships from the observed data. This is because as claimed by David Hume, a Scottish philosopher over three centuries ago, these observations only tell us a *constant conjunction* (i.e., correlation) of the events. Unfortunately, *such correlation does not imply the causation* because it might be brought by a third factor called a *confounder*, which influences the observed events, as stated by the Common Cause Principle (CCP) [Reichenbach, 1956].

Although many attempts have been made to deal with the influence of such a third factor [Granger, 1980; Pearl, 2009], causal discovery still remains a challenge because the observed data are often scarce and exhibit complex nonlinearity.

1.2 Treatment Effect Estimation

If the underlying causal directions are obvious, then scientists will be interested in gaining further causal knowledge, in particular, the strength of causal relationship. For instance, in medical science, we know that medical treatment, such as drug administration and vaccination, affects each patient's status; however, it is not obvious how strongly the status is affected by such a treatment.

Once Rubin [1974] offered a formulation of such treatment effects, many researchers have developed the techniques for estimating the treatment effects [Athey and Imbens, 2016; Hill, 2011; Robinson, 1988]. These studies have been accelerated by the appearance of neural-network-based estimation models [Johansson *et al.*, 2016; Shalit *et al.*, 2017]. As a result, a large number of approaches have been established using complex machine learning models [Alaa and van der Schaar, 2017; Hassanpour and Greiner, 2019; Künzel *et al.*, 2019; Nie and Wager, 2021; Yoon *et al.*, 2018].

However, due to the model complexity, these approaches cannot be used to explain why the treatment effects are different across individuals. Such a lack of model interpretability makes it impossible to elucidate the causal mechanism that yields the treatment effect heterogeneity, which hinders scientific discoveries in many fields, such as epidemiology [Jabal *et al.*, 2021] and economics [Taddy *et al.*, 2016].

1.3 Machine Learning and Fairness

The above issue of interpretability is not limited to treatment effect estimation but including many machine learning problems. To resolve this issue, many studies have been dedicated to improve the transparency of machine learning predictions [Heskes *et al.*, 2020; Larsen, 2022; Molnar, 2020].

In this direction, **fairness-aware machine learning** has gained increasing attention because it broadens the scope of machine learning applications. Indeed, machine learning is increasingly being used to make critical decisions that severely affect people's lives, such as loan approval [Khandani *et al.*, 2010], hiring decision [Houser, 2019], child abuse screening [Chouldechova *et al.*, 2018], and recidivism predictions [Angwin *et al.*, 2016]. The huge societal impact of such decisions on people's lives raises concerns about fairness because these decisions may be discriminatory with respect to *sensitive features*, such as race, gender, religion, and sexual orientation.

The fairness is usually defined by the Laws, and many studies have discussed how to formulate such legal definitions as numeric fairness criteria [Dwork *et al.*, 2012; Hardt *et al.*, 2016]. For instance, a well-known fairness criterion called *demographic parity* is formulated as the correlation between sensitive features and the decision outcome. However, the absence of such correlation does not necessarily imply the absence of causation because the correlation can be affected by other factors (i.e., the confounders), as mentioned in Section 1.1. Hence, in the presence of confounders, the correlation-based fairness criteria might wrongly determine the fairness.

A growing number of studies aim to achieve the causality-based fairness criteria [Makhlouf *et al.*, 2020]. However, learning fair predictive models based on causality still remains challenging due to the difficulty of estimating the causal effects.

1.4 Contributions

The goal of this dissertation is to establish the causal inference frameworks for accelerating scientific discoveries and improving the fairness of machine learning predictions. In particular, our contributions are summarized as follows:

Chapter 3 : We propose a data augmentation framework for improving the inference accuracy of causal discovery from time series data. To deal with the data scarcity issue and the complex nonlinearity among time-dependent variables, we establish a *supervised learning* approach that infers the causal relationships using *training data*, whose causal relationships are known.

Chapter 4 : We establish a feature selection approach to elucidating why the treatment effects are different across individuals. We consider a multiple hypothesis test for discovering the features related to the treatment effect heterogeneity. To achieve this, we formulate the feature importance measure based on the distributional discrepancies, which enables us to discover a wider variety of the features than the conventional approaches.

Chapter 5 : Following the existing causality-based fairness criterion, we propose a learning framework that strikes a good balance between fairness and prediction accuracy. Unlike the existing approaches, this framework can make individually fair predictions without making restrictive assumptions on the data.

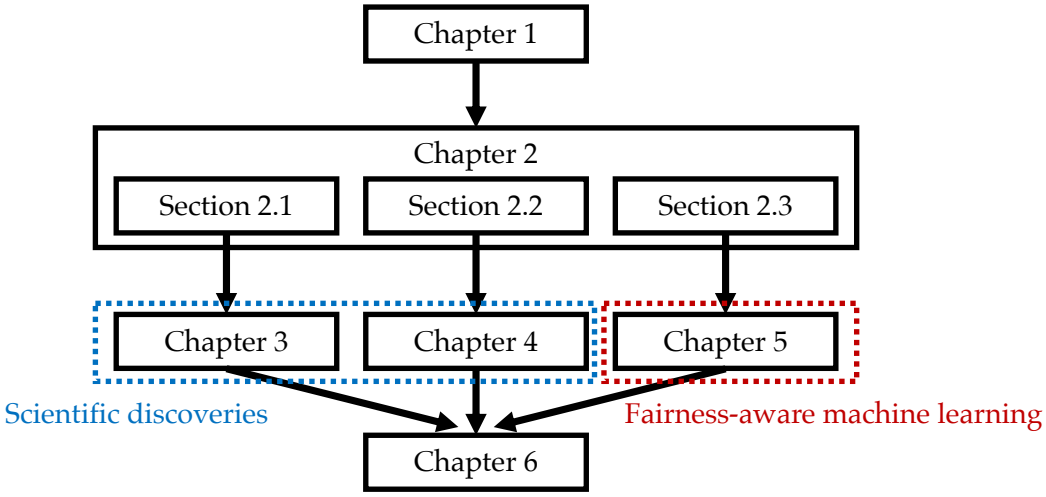

Figure 1.1: Dependence structure of chapters

While Chapters 3 and 4 are devoted to developing the causal inference approaches for making scientific discoveries, Chapter 5 focuses on utilizing causal inference to improve the fairness of machine learning predictions.

1.5 Dissertation Organization

The rest of this dissertation is organized as illustrated in Figure 1.1. Chapter 2 presents the background to the causality concepts: Section 2.1 introduces the causality concepts for time series data, Section 2.2 describes the potential outcome framework, and Section 2.3 offers the formulation of structural equation models. Chapter 3 presents a supervised learning approach to causal discovery from time series data. Chapter 4 presents a feature selection approach for elucidating the treatment effect heterogeneity. Chapter 5 presents a learning framework that guarantees individual-level fairness based on the causality. Chapter 6 concludes this dissertation.

Chapter 2

Preliminaries

This chapter provides an overview of the existing causality concepts. Each causality concept described in this chapter is founded on the following philosophical postulates of causality [Eichler, 2013]:

1. **Temporal precedence**: *A cause precedes its effects in time.*
2. **Interventional influence**: *Manipulating a cause changes its effects.*

This chapter is organized as follows. Section 2.1 presents the temporal causality concepts based on postulate 1: *Granger causality* [Granger, 1969], *Sims causality* [Sims, 1972], and *transfer entropy* [Schreiber, 2000]. Sections 2.2 and 2.3 introduce the potential outcome framework [Rubin, 1974] and the structural equation models [Pearl, 2009], both of which are founded on postulate 2.

2.1 Causality Concepts for Time Series Data

This section introduces a well-known temporal causality concept called Granger causality [Granger, 1969] and provides its comparison with the two temporal causality concepts: Sims causality [Sims, 1972] and transfer entropy [Schreiber, 2000]. The readers who are familiar with these concepts can skip this section.

2.1.1 Granger Causality

Granger causality [Granger, 1969, 1980] is a temporal causality concept, which defines the causal relationship between two time-dependent variables.

Bivariate Granger Causality

To illustrate the definition of Granger causality, consider the bivariate setting where we have a bivariate time series, $X = \{X_t\}$ and $Y = \{Y_t\}$, which are measured at discrete time points $t = 1, 2, \ldots$. Roughly speaking, Granger causality defines X as the cause of Y if the past values of X contain *helpful* information for predicting the future value of Y. Formally, it is defined as follows:

Definition 1 (bivariate Granger causality [Granger, 1969])**.** *Suppose we have a sequence pair of random variables, $X = \{X_t\}$ and $Y = \{Y_t\}$ ($t = 1, 2, \ldots$), where X_t and Y_t are on $\mathcal{X}$ and $\mathcal{Y}$, respectively. Let $S_{X_{:t}}$ and $S_{Y_{:t}}$ be observations of $\{X_1, \ldots, X_t\}$ and $\{Y_1, \ldots, Y_t\}$, respectively.*

Bivariate Granger causality states that X is not the cause of Y if the following holds for all $t = 1, 2, \ldots$

$$\mathrm{P}(Y_{t+1} \mid S_{X_{:t}}, S_{Y_{:t}}) = \mathrm{P}(Y_{t+1} \mid S_{Y_{:t}}); \tag{2.1}$$

otherwise, X is the cause of Y.

Since the equality between two conditional distributions in Eq. (2.1) is equivalent to the conditional independence relation

$$Y_{t+1} \perp\!\!\!\perp S_{X_{:t}} \mid S_{Y_{:t}}, \tag{2.2}$$

Definition 1 states that X is not the cause of Y if Y_{t+1} is conditionally independent of $S_{X_{:t}}$ given $S_{Y_{:t}}$; otherwise, X is the cause of Y.

Bivariate Granger causality cannot define the causal relationships in multivariate time series [Granger, 1969]. This is because when there is no causal relationship between time series X and Y, if these time series are affected by the third time series $Z = \{\boldsymbol{Z}_t\}$, they will be mutually dependent. This indicates that even in the absence of the causal relationship between X and Y, the conditional independence relations (e.g., (2.2)) may not hold, leading to a wrong conclusion that X is the cause of Y, or Y is the cause of X. To resolve this weakness, *conditional Granger causality* [Granger, 1980] has been developed, which is described in the next section.

Conditional Granger Causality

Conditional Granger causality [Granger, 1980] is an extended notion of Granger causality that addresses the influence of the third time series, $Z = \{\boldsymbol{Z}_t\}$, where $\boldsymbol{Z}_t$ is a univariate or multivariate random variable. Formally, it is defined as follows:

Definition 2 (conditional Granger causality [Granger, 1980])**.** *Let $X = \{X_t\}$, $Y = \{Y_t\}$, and $Z = \{\boldsymbol{Z}_t\}$ $(t = 1, 2, \dots)$ be a triplet of random variable sequences, where X_t, Y_t, and $\boldsymbol{Z}_t$ are on $\mathcal{X}$, $\mathcal{Y}$ and $\mathcal{Z}$, respectively. Let $S_{X_{:t}}$, $S_{Y_{:t}}$ and $S_{Z_{:t}}$ be observations of $\{X_1, \dots, X_t\}$, $\{Y_1, \dots, Y_t\}$ and $\{\boldsymbol{Z}_1, \dots, \boldsymbol{Z}_t\}$, respectively.*

Conditional Granger causality states that X is not the cause of Y given Z if the following holds for all $t = 1, 2, \dots$

$$\mathrm{P}(Y_{t+1} \mid S_{X_{:t}}, S_{Y_{:t}}, S_{Z_{:t}}) = \mathrm{P}(Y_{t+1} \mid S_{Y_{:t}}, S_{Z_{:t}});$$

otherwise, X is the cause of Y given Z.

As with the bivariate setting, conditional Granger causality is equivalently represented as the conditional (in)dependence relations: X is not the cause of Y given Z if Y_{t+1} is conditionally independent of $S_{X_{:t}}$ given $S_{Y_{:t}}$ and $S_{Z_{:t}}$, i.e.,

$$Y_{t+1} \perp\!\!\!\perp S_{X_{:t}} \mid S_{Y_{:t}}, S_{Z_{:t}};$$

otherwise, X is the cause of Y given Z.

For this reason, discovering Granger causality from time series data can be formulated as a problem of inferring the conditional (in)dependence relations. This inference problem has a long history in statistics, which is why Granger causality has attracted a lot of attention.

2.1.2 Related Temporal Causality Concepts

After the concept of Granger causality was proposed in Granger [1969], several temporal causality concepts have been introduced [Sims, 1972; Schreiber, 2000]. This section provides a comparison of these concepts with Granger causality to discuss the similarity and the difference between them.

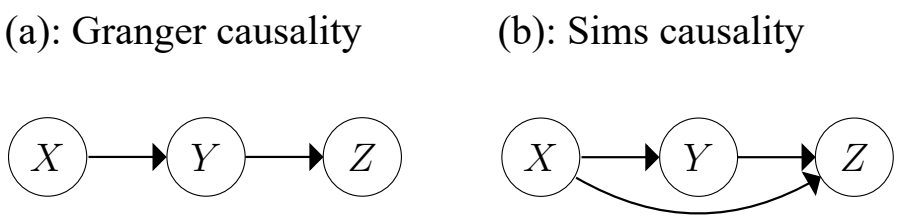

Figure 2.1: Granger causality and Sims causality may lead to different conclusions: If time-dependent variables X, Y, and Z have Granger causality as shown in (a), their Sims causality is given as (b).

Sims Causality

The notion of Sims causality [Sims, 1972] is similar to that of Granger causality. While Granger causality says that X is the cause of Y if the past values of X are helpful for predicting the future value of Y, Sims causality states that X is the cause of Y if the future values of Y are helpful for predicting the present value of X (See Sims [1972]; Florens [2003] for the formal definition).

According to Kuersteiner [2010], in bivariate setting, Granger causality and Sims causality are equivalent. However, in multivariate setting, they are not equivalent. In particular, while the former distinguishes the direct and indirect causal relationships, the latter does not.

To illustrate this difference, consider the trivariate time series X, Y, and Z. Suppose that this time series has the following Granger causality relations: X is the cause of Y given Z, Y is the cause of Z given X, and X is not the cause of Z given Y, as shown in Figure 2.1(a). In this case, as illustrated in Figure 2.1(b), Sims causality draws a different conclusion; unlike Granger causality, X is regarded as the cause of Z given Y due to the indirect influence of X on Z via Y.

This example highlights that Granger causality is a stronger notion of temporal causality in the sense that the presence of Granger causality implies the presence of Sims causality while the converse is not true. Although it is unclear which causality notion is preferable in practice, since Granger causality has been far more widely used than Sims causality [Eichler, 2013], this dissertation focuses on Granger causality.

Transfer Entropy

Transfer entropy [Schreiber, 2000] is an information-theoretic measure of temporal causality [Amblard and Michel, 2013]. This measure uses Shannon entropy [Shannon, 1948] to measure *how greatly the conditional independence relations between time-dependent variables are violated*. Originally, Schreiber [2000] defines it for bivariate time series as follows:

Definition 3 (transfer entropy [Schreiber, 2000])**.** *Suppose we have a pair of random variable sequences* $X = \{X_t\}$ *and* $Y = \{Y_t\}$ $(t = 1, 2, \dots)$*, where* X_t *and* Y_t *are discrete random variables and defined on* $\mathcal{X}$ *and* $\mathcal{Y}$*, respectively. Let* $S_{X_{t-(k-1):t}}$ *and* $S_{Y_{t-(k-1):t}}$ *(for some* $k \in \{1, \dots, t\}$*) be observations of* $\{X_{t-(k-1)}, \dots, X_t\}$ *and* $\{Y_{t-(k-1)}, \dots, Y_t\}$*, respectively.*

Transfer entropy is defined as a difference between the two (conditional) Shannon entropies:

$$\mathrm{T}_{X \to Y} = \mathrm{H}(Y_{t+1} \mid S_{Y_{t-(k-1):t}}) - \mathrm{H}(Y_{t+1} \mid S_{X_{t-(k-1):t}}, S_{Y_{t-(k-1):t}}), \tag{2.3}$$

where H *denotes (conditional) Shannon entropy, which is formulated for two discrete random variables* A *and* B *as*

$$\mathrm{H}(B \mid A) = -\sum_{a,b} \mathrm{P}(A = a, B = b) \log \frac{\mathrm{P}(A = a, B = b)}{\mathrm{P}(A = a)}.$$

While Definition 3 only addresses bivariate discrete-valued time series, it has been extended to multivariate time series [Lizier *et al.*, 2008, 2011] and continuous-valued time series [Cover, 1999].

Transfer entropy can also be used to determine the presence and the direction of Granger causality. In particular, we can infer Granger causality between X and Y by performing a statistical hypothesis test that determines whether transfer entropy $\mathrm{T}_{X \to Y}$ is zero. Barnett *et al.* [2009] show that if the time series follows a multivariate vector autoregressive (VAR) model with Gaussian noises, then the test statistic of such a transfer entropy test is equivalent to that of the Granger causality test for a multivariate linear time series [Geweke, 1982]; hence, in this case, the transfer entropy test is equivalent to the Granger causality test.

Unlike Granger causality, transfer entropy can capture the strength of causality as an information-theoretically intuitive quantity; however, its estimation is not easy. Inferring Shannon entropy H in Eq. (2.3) requires us to estimate the joint distribution, which is challenging, especially when time lag k is large. For this reason, the estimation of transfer entropy often relies on the parametric assumption on the underlying time series (e.g., the assumption that the data follow a VAR model) or requires a large sample size enough to perform nonparametric density estimation. Indeed, the existing methods for inferring Granger causality also suffer from these weaknesses, as described in Chapter 3.

2.2 Potential Outcome Framework

The temporal causality concepts described in Section 2.1 have the two weaknesses:

- They cannot be used if the data are not time series (e.g., i.i.d. data) or if the time series data do not contain their observed time points.
- They cannot quantify how strongly intervening a cause changes its effects.

In particular, the second weakness is a crucial drawback if we are interested in assessing the effects of a *treatment*, such as drug administration [Kosorok and Laber, 2019], education program [Tipton and Olsen, 2018], and advertisement placement [Rzepakowski and Jaroszewicz, 2012].

This section introduces a widely used statistical framework for formulating such treatment effects, which is called the *potential outcome framework* [Rubin, 1974].

2.2.1 Potential Outcomes and Treatment Effects

The potential outcome framework (a.k.a. the Neyman-Rubin causal model) uses the three random variables: treatment A, features $\boldsymbol{X}$ (a.k.a. covariates), and outcome Y. For example, in case of the drug effect evaluation, treatment A stands for the drug administration, features $\boldsymbol{X}$ represent the attributes of each patient, and outcome Y expresses their health condition. Throughout this dissertation, we consider binary treatment $A \in \{0, 1\}$; however, the framework can be extended to categorical and continuous-valued treatments.

The effects of treatment A on outcome Y is formulated using the two random variables called *potential outcomes*, each of which represents the outcome under some treatment. For binary treatment $A \in \{0, 1\}$, these random variables are denoted by Y^0 and Y^1, which express the outcome when an individual gets no treatment ($A = 0$) and the one when the individual receives a treatment ($A = 1$), respectively. Using potential outcomes Y^0 and Y^1, a treatment effect for an individual (a.k.a. **individual treatment effect (ITE)**) is defined as

$$\textbf{ITE:} \quad Y^1 - Y^0. \tag{2.4}$$

An ITE is never observed because Y^0 and Y^1 are never jointly observed.

However, under several assumptions, we can estimate the average of a treatment effect across individuals; in particular an **average treatment effect (ATE)** and a **conditional average treatment effect (CATE)**:

$$\textbf{ATE:} \quad \mathbb{E}_{Y^0,Y^1}[Y^1 - Y^0], \tag{2.5}$$

$$\textbf{CATE:} \quad \mathbb{E}_{Y^0,Y^1}[Y^1 - Y^0 \mid \boldsymbol{X} = \boldsymbol{x}]. \tag{2.6}$$

While an ATE is the average of a treatment effect over all individuals, a CATE is the average across a subgroup of individuals who have feature attributes $\boldsymbol{X} = \boldsymbol{x}$.

The next section presents a gold standard approach to treatment effect estimation, which is called *randomized controlled trials* (RCTs).

2.2.2 Randomized Controlled Trials (RCTs)

An RCT, which is called A/B testing in the field of marketing design, is an experiment for evaluating treatment effects where the treatments are randomly assigned to individuals. More precisely, in RCTs, treatment A is randomly assigned to be independent of potential outcomes Y^0 and Y^1; in other words,

$$A \perp\!\!\!\perp \{Y^0, Y^1\}. \tag{2.7}$$

Under independence relation (2.7), for instance, the ATE in Eq. (2.5), which equals $\mathbb{E}_{Y^1}[Y^1] - \mathbb{E}_{Y^0}[Y^0]$ because of linearity of expectation, is represented as the

difference between the following two conditional expected values:

$$\mathbb{E}_{Y^1}[Y^1 \mid A = 1] - \mathbb{E}_{Y^0}[Y^0 \mid A = 0]. \tag{2.8}$$

Given the observations of n individuals $\{(a_i, y_i)\}_{i=1}^n$ obtained by the RCT, we can easily formulate the consistent and unbiased estimator of Eq. (2.8) as

$$\frac{1}{n_1} \sum_{i=1;a_i=1}^{n} y_i - \frac{1}{n_0} \sum_{i=1;a_i=0}^{n} y_i, \tag{2.9}$$

where n_0 and n_1 are the number of individuals whose treatment assignment is given as $a_i = 0$ and $a_i = 1$, respectively.

Unfortunately, performing RCTs is highly expensive in terms of time and money and is often impossible due to the legal and ethical reasons. For this reason, instead of RCTs, researchers often rely on *observational data*.

2.2.3 Treatment Effect Estimation from Observational Data

Whereas the RCT data are collected by randomly assigning treatments to individuals, observational data are obtained without any interference, i.e., simply by observing the actions, the features, and the outcomes of individuals.

Although this data-collecting process is cost-effective, it makes independence relation (2.7) violated. This is because in such a data-collecting process, treatment A and outcome Y are usually affected by the variables called **confounders**. For instance, in case of the drug effect evaluation, age is a confounder, which influences drug administration A and health condition Y: individuals with different ages have different treatment preferences and different outcomes. The presence of such a confounder leads to the well-known **selection bias** problem (i.e., the observations are not representative to the inference target distribution), which makes it challenging to estimate ATE and CATE.

A traditional estimation technique for dealing with such selection bias problem is **inverse probability weighting** (IPW) [Rosenbaum and Rubin, 1983], which offers an ATE estimator using importance sampling [Kloek and Van Dijk, 1978]. Specifically, IPW takes an expectation with respect to target distribution $\mathrm{P}(\boldsymbol{X})$ by computing a weighted average of the observations from conditional distribution

$\mathrm{P}(\boldsymbol{X} \mid A = a)$ $(a \in \{0, 1\})$. To achieve this, it uses the following *inverse probability weights*:

$$w^0(A, \boldsymbol{X}) = \frac{\mathbf{I}(A = 0)}{1 - \mathrm{e}(\boldsymbol{X})}, \quad w^1(A, \boldsymbol{X}) = \frac{\mathbf{I}(A = 1)}{\mathrm{e}(\boldsymbol{X})}, \tag{2.10}$$

where $\mathrm{e}(\boldsymbol{X}) \coloneqq \mathrm{P}(A = 1 \mid \boldsymbol{X})$ is the conditional distribution called a *propensity score*, and $\mathbf{I}(A = a)$ is an indicator function that takes 1 if $A = a$; otherwise, 0. Estimating ATE with these weights requires the following two standard assumptions:

Assumption 1 (Conditional ignorability (a.k.a. strong ignorability))**.** *Conditioned on features $\boldsymbol{X}$, treatment A is conditionally independent of potential outcomes Y^0 and Y^1; that is,*

$$A \perp\!\!\!\perp \{Y^0, Y^1\} \mid \boldsymbol{X}. \tag{2.11}$$

Assumption 2 (Positivity)**.** *For any value $\boldsymbol{x}$ of features $\boldsymbol{X}$, propensity score $\mathrm{e}(\boldsymbol{X})$ satisfies the following support condition:*

$$0 < \mathrm{e}(\boldsymbol{x}) < 1. \tag{2.12}$$

Conditional ignorability (Assumption 1) requires features $\boldsymbol{X}$ to contain all confounders, which ensures the equality between the conditional distributions of potential and observed outcomes: $\mathrm{P}(Y^a \mid A = a, \boldsymbol{X}) = \mathrm{P}(Y \mid A = a, \boldsymbol{X})$. By contrast, positivity (Assumption 2) imposes the support condition on propensity score $\mathrm{e}(\boldsymbol{X})$, which guarantees that no zero division occurs in Eq. (2.10). Under these assumptions, the IPW-based estimator of ATE is given as

$$\begin{aligned}
&\mathbb{E}_{Y^1}[Y^1] - \mathbb{E}_{Y^0}[Y^0] \\
&= \mathbb{E}_{\boldsymbol{X}}[\mathbb{E}_{Y^1|\boldsymbol{X}}[Y^1]] - \mathbb{E}_{\boldsymbol{X}}[[\mathbb{E}_{Y^0|\boldsymbol{X}}[Y^0]] \\
&= \mathbb{E}_{\boldsymbol{X}|A=1}\left[\mathbb{E}_{Y|\boldsymbol{X},A=1}\left[\frac{\mathrm{P}(A=1)}{\mathrm{P}(A=1 \mid \boldsymbol{X})}Y\right]\right] - \mathbb{E}_{\boldsymbol{X}|A=0}\left[\mathbb{E}_{Y|\boldsymbol{X},A=0}\left[\frac{\mathrm{P}(A=0)}{\mathrm{P}(A=0 \mid \boldsymbol{X})}Y\right]\right] \\
&= \mathbb{E}_{A,\boldsymbol{X},Y}[w^1(A, \boldsymbol{X})Y] - \mathbb{E}_{A,\boldsymbol{X},Y}[w^0(A, \boldsymbol{X})Y].
\end{aligned} \tag{2.13}$$

Given the observational data about n individuals, $\mathcal{D} = \{(a_i, \boldsymbol{x}_i, y_i)\}_{i=1}^n \overset{i.i.d.}{\sim} \mathrm{P}(A, \boldsymbol{X}, Y)$,

we can estimate the expected values in Eq. (2.13) as

$$\frac{1}{n_1}\sum_{i=1}^{n} w^1(a_i, \boldsymbol{x}_i)y_i - \frac{1}{n_0}\sum_{i=1}^{n} w^0(a_i, \boldsymbol{x}_i)y_i, \tag{2.14}$$

where n_0 and n_1 are the number of individuals whose treatment assignment is given as $a_i = 0$ and $a_i = 1$, respectively.

Recently, a growing number of causal inference methods have been developed to estimate heterogeneous treatment effects, which are expressed as CATEs. Johansson *et al.* [2016]; Shalit *et al.* [2017] have developed a representation learning approach that addresses the selection bias problem by learning a balanced feature representation with neural network. Hassanpour and Greiner [2019] have combined such a representation learning approach with IPW. Hahn *et al.* [2020]; Hill [2011] have formulated nonparametric tree-based models, which use Bayesian inference to quantify the estimation uncertainty. Künzel *et al.* [2019]; Nie and Wager [2021] have established machine learning frameworks called meta-learners, which yield a CATE estimator with the fast convergence rate and hence do not require much data.

Although these methods focus on improving the accuracy of CATE estimation, their estimation models are often too complex to understand why the estimated treatment effects are different across individuals. This is a crucial drawback because elucidating the reason why such treatment effect heterogeneity arises is a common problem in many applications. As an approach that overcomes this drawback, in Chapter 4, we present a feature selection framework for discovering the features related to distributional treatment effect heterogeneity.

2.3 Structural Equation Models (SEMs)

This section introduces structural equation models (SEMs) [Pearl, 2009], which offers an alternative causality formulation to the potential outcome framework [Rubin, 1974]. Below we discuss their difference and similarity; namely,

- Unlike the potential outcome framework, an SEM offers a graphical representation of the causal relationships between variables by a *causal graph* (Section 2.3.1). Such a graphical representation can be used to analyze the causal effects under a complex causal relationships among variables (Section 2.3.3).

- As with the potential outcome framework, an SEM provides an equivalent formulation of potential outcomes (Section 2.3.2).

2.3.1 Causal Graphs and Structural Equations

An SEM is a model for representing the complex causal relationship between observed and unobserved variables. Formally, it is defined as follows:

Definition 4 (Pearl [2009]). *An SEM is triplet* $\mathcal{M} = (\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{F})$*, where*

- $\boldsymbol{U}$ *is a set of* ***exogenous variables****, i.e., unobserved variables that express the external factors, such as measurement errors and unmeasurable quantities.*

- $\boldsymbol{V}$ *is a set of* ***endogenous variables****, i.e., observed variables whose values are determined by* $\boldsymbol{U} \cup \boldsymbol{V}$*.*

- $\boldsymbol{F}$ *is a set of deterministic functions (a.k.a.* ***structural functions****). Each* $f_V \in \boldsymbol{F}$ *is used to express endogenous variable* $V \in \boldsymbol{V}$ *as a structural equation:*

$$V = f_V(\boldsymbol{pa}(V), \boldsymbol{U}_V), \tag{2.15}$$

  where $\boldsymbol{pa}(V) \subseteq \boldsymbol{V} \backslash V$ *is a subset of endogenous variables* $\boldsymbol{V} \backslash V$*, and* $\boldsymbol{U}_V \subseteq \boldsymbol{U}$ *are exogenous variables.*

Structural equation (2.15) determines the values of each observed variable $V \in \boldsymbol{V}$ as an output of deterministic function f_V that takes the two inputs, $\boldsymbol{pa}(V)$ and $\boldsymbol{U}_V$.

Each SEM is associated with a directed acyclic graph (DAG), which is referred to as a causal graph. In each causal graph, the nodes represent endogenous variables $\boldsymbol{V}$, and each edge points from each member of $\boldsymbol{pa}(V)$ to V.

To illustrate a causal graph and an SEM, consider the setup of the potential outcome framework described in Section 2.2. The causal relationships among treatment A, features $\boldsymbol{X}$, and outcome Y can be depicted as the causal graph in Figure 2.2. Suppose that this causal graph structure is associated with SEM $\mathcal{M}^{\text{po}} := (\boldsymbol{U}^{\text{po}}, \boldsymbol{V}^{\text{po}}, \boldsymbol{F}^{\text{po}})$, where $\boldsymbol{V}^{\text{po}} := (A, \boldsymbol{X}, Y)$ is a set of the endogenous vari-

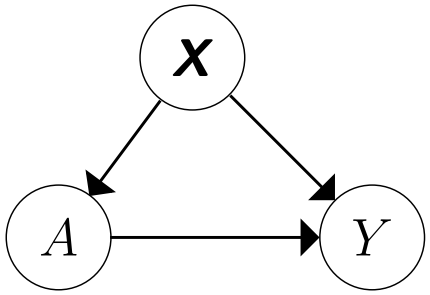

Figure 2.2: Causal graph that describes setting of potential outcome framework: treatment A influences outcome Y, and both of them are affected by features $\boldsymbol{X}$, which include confounders.

ables whose values are determined by the following three structural equations:

$$\boldsymbol{X} = f_{\boldsymbol{X}}(\boldsymbol{U}_{\boldsymbol{X}}), \tag{2.16}$$

$$A = f_A(\boldsymbol{X}, \boldsymbol{U}_A), \tag{2.17}$$

$$Y = f_Y(A, \boldsymbol{X}, \boldsymbol{U}_Y), \tag{2.18}$$

where $f_{\boldsymbol{X}}, f_A, f_Y \in \boldsymbol{F}^{\text{po}}$ are deterministic functions, and $\boldsymbol{U}_{\boldsymbol{X}}$, $\boldsymbol{U}_A, \boldsymbol{U}_Y \subseteq \boldsymbol{U}^{\text{po}}$ are the sets of unobserved variables (e.g., measurement errors and unmeasurable quantities) that influence observed variables $\boldsymbol{X}$, A, and Y, respectively.

Indeed, these structural equations can be used to equivalently formulate the potential outcomes defined in Section 2.2. To do so, we need to modify them with an operation on an SEM, which is called *interventions*.

2.3.2 Interventions and Interventional SEMs

An **intervention** is defined as an operation on an SEM that replaces the structural equation [Pearl, 1994]. For instance, intervention $\text{do}(A = a)$ replaces structural equation (2.17) with constant $A = a$ ($a \in \{0, 1\}$).

This intervention can be used to offer an equivalent formulation of the potential outcomes. By replacing structural equation (2.17) with $A = a$, it yields a modified SEM called an **interventional structural equation model (SEM)**, denoted by

$\mathcal{M}^{\text{po}}_{do(A=a)}$, which contains the following structural equations:

$$\boldsymbol{X} = f_{\boldsymbol{X}}(\boldsymbol{U}_{\boldsymbol{X}}), \tag{2.19}$$

$$A = a, \tag{2.20}$$

$$Y^a \coloneqq f_Y(a, \boldsymbol{X}, \boldsymbol{U}_Y). \tag{2.21}$$

In this interventional SEM, outcome Y is replaced with potential outcome Y^a, i.e., outcome that is observed if treatment is given as $A = a$.

In fact, the notion of an intervention can be used not only to formulate the potential outcomes but also to define the causal effects under the complex causal graph. The analysis of such complex causal effects is called *causal mediation analysis*.

2.3.3 Causal Mediation Analysis

Causal mediation analysis is a methodology for understanding the complex causal relationships by evaluating how greatly one variable *directly* and *indirectly* influences the other. In causal mediation analysis, we address the cases where some part of this influence is mediated by other variables called **mediators**.

Direct and Indirect Effects

To illustrate the direct and indirect effects, consider the causal graph shown in Figure 2.3, which indicates that drug administration A indirectly decreases the risk of heart attack Y by lowering blood pressure M ($A \to M \to Y$) and directly influences Y through another unknown mechanism ($A \to Y$). Following this causal graph, we can perform causal mediation analysis to quantify the direct effects along pathway $A \to Y$ and the indirect effects along pathway $A \to M \to Y$.

To formulate such direct and indirect effects, Pearl [2001] have defined **natural direct effects (NDEs)** and a **natural indirect effects (NIEs)**. Under the causal graph in Figure 2.3, they are defined based on the SEM with the structural equations:

$$A = f_A(\boldsymbol{X}, \boldsymbol{U}_A), \tag{2.22}$$

$$M = f_M(A, \boldsymbol{C}, \boldsymbol{U}_M), \tag{2.23}$$

$$Y = f_Y(A, M, \boldsymbol{X}, \boldsymbol{C}, \boldsymbol{U}_Y), \tag{2.24}$$

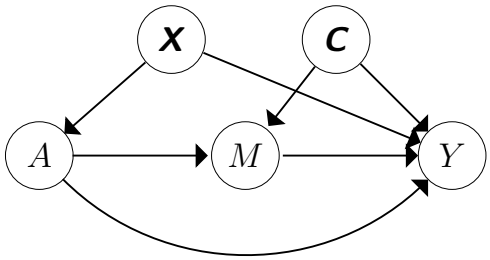

Figure 2.3: Example causal graph that depicts mediation: drug administration A indirectly decreases risk of heart attack Y by lowering blood pressure M ($A \to M \to Y$) and directly influences Y through an unknown mechanism ($A \to Y$). $\boldsymbol{X}$ denotes confounders (e.g., age) that affect A and Y, and $\boldsymbol{C}$ stands for mediator-outcome confounders (e.g., gender) that alter M and Y.

where $\boldsymbol{U}_A$, $\boldsymbol{U}_M$, and $\boldsymbol{U}_Y$ denote the exogenous variables that affect drug administration A, blood pressure M, and heart attack Y, respectively. Given this SEM, NDEs and NIEs are formulated using **potential mediator** M^a ($a \in \{0, 1\}$), which is defined as mediator M under intervention $do(A = a)$ as

$$M^a \coloneqq f_M(a, \boldsymbol{C}, \boldsymbol{U}_M). \tag{2.25}$$

Using potential mediator M^a, an NDE and an NIE are defined as

$$\textbf{NDE:} \quad Y^{1,M^0} - Y^0, \tag{2.26}$$

$$\textbf{NIE:} \quad Y^{0,M^1} - Y^0, \tag{2.27}$$

where Y^{1,M^0} and Y^{0,M^1} are the random variables called **(nested) potential outcomes**, which are defined by the following structural equations:

$$Y^{1,M^0} \coloneqq f_Y(1, M^0, \boldsymbol{X}, \boldsymbol{C}, \boldsymbol{U}_Y), \tag{2.28}$$

$$Y^{0,M^1} \coloneqq f_Y(0, M^1, \boldsymbol{X}, \boldsymbol{C}, \boldsymbol{U}_Y), \tag{2.29}$$

Compared with potential outcome $Y^0 \coloneqq f_Y(0, M^0, \boldsymbol{X}, \boldsymbol{C}, \boldsymbol{U}_Y)$, these nested potential outcomes are formulated by switching input $A = 0$ to $A = 1$ or input M^0 to M^1, both of which are defined two interventional SEMs under $do(A = 0)$ and $do(A = 1)$. By taking a mean difference from these nested potential outcomes, NDEs and NIEs measure the direct effects of A on Y and the indirect effects of A on Y through M.

While we illustrate the formulation of NDEs and NIEs under such a simple causal graph as Figure 2.3, they can be similarly formulated under a more complicated causal graph with multiple causal pathways from A to Y. However, in such cases, NIEs measures the total indirect effects along **all** indirect pathways from A to Y and hence cannot be used to assess the causal effects along **some of** the causal pathways from A to Y. To evaluate such causal effects, we need a generalized notion of causal effects, which is called *path-specific effects*.

Path-Specific Effects

A path-specific effect [Avin *et al.*, 2005] is the notion of causal effects, which measures a causal effect along a given set of causal pathways. This notion corresponds to the generalization of an NDE and an NIE. For instance, letting A and Y denote treatment and outcome, it recovers the NDE of A on Y by setting causal pathways π to the direct pathway (i.e., $\pi = \{A \to Y\}$).

A path-specific effect is a difference between potential outcomes $Y_{A \Leftarrow 0}$ and $Y_{A \Leftarrow 1 \| \pi}$:

$$\textbf{PSE:} \quad Y_{A \Leftarrow 1 \| \pi} - Y_{A \Leftarrow 0}. \tag{2.30}$$

$Y_{A \Leftarrow 0}$ is simply defined as the potential outcome under intervention $do(A = 0)$, which is hence equivalent to potential outcome Y^0 in Eqs. (2.26) and (2.27). By contrast, $Y_{A \Leftarrow 1 \| \pi}$ is defined as a nested potential outcome that is formulated by switching each variable in causal pathways π with a different treatment value, $A = 1$.

Formally, this nested potential outcome is defined with an SEM that is modified by combining two interventional SEMs $\mathcal{M}_{A=0}$ and $\mathcal{M}_{A=1}$. To illustrate this SEM, for each endogenous variable $V \in \boldsymbol{V}$ in original SEM $\mathcal{M}$, consider to partition its parents, $\boldsymbol{pa}(V)$, into two subsets, $\boldsymbol{pa}(V) = \{\boldsymbol{pa}(V)^{\pi}, \boldsymbol{pa}(V)^{\overline{\pi}}\}$, where $\boldsymbol{pa}(V)^{\pi}$ is the members of $\boldsymbol{pa}(V)$ connected with V on pathways π, and $\boldsymbol{pa}(V)^{\overline{\pi}}$ is a complementary set (i.e., $\boldsymbol{pa}(V)^{\overline{\pi}} = \boldsymbol{pa}(V) \backslash \boldsymbol{pa}(V)^{\pi}$). Based on these two subsets, we consider the following structural equation over $V \in \boldsymbol{V}$:

$$V = f_V(\boldsymbol{pa}(V)^{\pi}_{A=1}, \boldsymbol{pa}(V)^{\overline{\pi}}_{A=0}, \boldsymbol{U}_V), \tag{2.31}$$

where $\boldsymbol{pa}(V)^{\pi}_{A=1}$ is the variables in $\boldsymbol{pa}(V)^{\pi}$ whose values are determined by interventional model $\mathcal{M}_{A=1}$, and $\boldsymbol{pa}(V)^{\overline{\pi}}_{A=0}$ is the variables in $\boldsymbol{pa}(V)^{\overline{\pi}}$ whose values are

provided by $\mathcal{M}_{A=0}$. Given an SEM with such a structural equation, nested potential outcome $Y_{A\Leftarrow 1\|\pi}$ is defined as outcome Y defined with structural equation (2.31).

Literature Overview on Causal Mediation Analysis Applications

Traditionally, causal mediation analysis has been applied in various fields, such as epidemiology [Richiardi *et al.*, 2013], econometrics [Heckman and Pinto, 2015], and online marketing [Yin and Hong, 2019].

Recent studies have focused on how to utilize it to improve the interpretability of machine learning predictions. Heskes *et al.* [2020] have formulated each feature's contribution to the predictions by decomposing Shapley values into NDEs and NIEs. Larsen [2022] have employed NIEs to develop an effective and interpretable reward measure in reinforcement learning.

In this research direction, one of the most important application examples of causal mediation analysis is fairness-aware machine learning. Zhang *et al.* [2017, 2018] have utilized causal mediation analysis to analyze the discriminatory bias in the data and to generate a fair dataset. Vig *et al.* [2020] have analyzed the gender bias in the learned natural language processing models. Zhang and Wu [2017]; Nabi and Shpitser [2018]; Chiappa and Gillam [2019]; Xu *et al.* [2019] have developed a framework for learning fair predictive models. Wu *et al.* [2018]; Nabi *et al.* [2019] have extended such a framework to ranking learning and reinforcement learning.

Among these applications, learning fair predictive models is promising because causal mediation analysis enables us to effectively measure the discriminatory bias, which is helpful to strike a good balance between fairness and prediction accuracy. Unfortunately, however, due to the difficulty of estimating the unfairness measures based on path-specific effects, existing methods require restrictive assumptions to develop a fair predictive models. To overcome this limitation, in Chapter 5, we present a learning framework that can effectively learn a fair classifier without making such restrictive assumptions.

Chapter 3

A Supervised Learning Approach to Granger Causality Discovery

In this chapter, we consider the problem of inferring Granger causality from time series data, which is defined as the conditional (in)dependence relations between them, as described in Section 2.1.1. Solving this inference problem allows us to understand the underlying complex causal relationships in time series and hence has diverse importance applications, including economics [Kar *et al.*, 2011], bioinformatics [Yao *et al.*, 2015], and neuroscience [Smith, 2012]. However, as described below, it is not easy to solve the Granger causality inference problem with traditional methods because they require an appropriate selection of the time series regression model, which needs a deep understanding of the time series data analysis. To overcome this limitation, we propose a supervised learning approach to Granger causality inference.

3.1 Introduction

Unraveling the complex causal relationships in time series offers key scientific discoveries in many fields. For this goal, a large body of studies have been devoted to discovering Granger causality, which is one of the central problems in time series analysis. This problem has many important applications in various fields. Application examples include the financial development analysis in economics [Kar *et al.*, 2011], the gene regulatory network discovery in bioinformatics [Yao *et al.*, 2015], and the inference of the brain functional connectivity in neuroscience [Smith, 2012].

As described in Section 2.1.1, Granger causality [Granger, 1969] is defined as the conditional (in)dependence relations between time-dependent variables. Roughly speaking, in case of bivariate setting with two time-dependent variables X and Y, Granger causality defines X as the cause of Y if the past values of X contain *helpful* information for predicting the future values of Y.

To detect such relations, traditional methods for identifying Granger causality use regression models [Bell *et al.*, 1996; Cheng *et al.*, 2014; Granger, 1969; Marinazzo *et al.*, 2008; Sun, 2008]. With these methods, we determine that X is the cause of Y when the prediction errors of Y based only on its past values are significantly reduced by additionally using the past values of X. When the regression model can be well fitted to the data, we can infer correct causal directions. However, in practice, selecting an appropriate regression model for each time series data is difficult and requires a deep understanding of the data analysis. Therefore, it is not easy to identify correct causal directions with these model-based methods.

3.1.1 Contributions

We propose an approach to Granger causality discovery that does not require a deep understanding of the data analysis. To achieve this goal, we develop a supervised learning framework that trains a classifier for assigning a ternary *causal label* ($X \to Y$, $X \leftarrow Y$, or *No Causation*) to each pair of time series. This idea of classification is inspired by recently proposed causal discovery methods for i.i.d. data, which have experimentally worked well [Bontempi and Flauder, 2015; Guyon, 2013; Lopez-Paz *et al.*, 2015, 2017]. A significant advantage of the classification framework over the model-based methods is that apart from the data whose causal relationships we infer, which we call *test data*, it can utilize *training data*, i.e., the data with known causal relationships. Examples of such training data include the synthetic data whose generating processes are obvious and the real-world data whose causal relationships are obvious from the knowledge of domain experts. Based on these training data, the classification framework performs data augmentation, which allows us to achieve high inference accuracy in causal discovery.

To develop a classification framework for time series data, we formulate a feature representation that provides sufficiently different feature vectors for time series with different causal relationships. The idea for obtaining such feature vectors is founded

on the definition of Granger causality: X is the cause of Y if the following two conditional distributions of the future value of Y are different; one is given the past values of Y and the other is given the past values of X and Y. To build the classifier for Granger causality identification, we utilize the distance between these distributions when preparing feature vectors. To compute the distance, by using *kernel mean embedding*, we map each distribution to a point in the feature space called the reproducing kernel Hilbert space (RKHS) and measure the distance between the points, which is termed the *maximum mean discrepancy* (MMD) [Gretton *et al.*, 2007].

In experiments, our method sufficiently outperformed the model-based Granger causality methods and the supervised learning method for i.i.d. data. Furthermore, we describe how our approach can be extended to multivariate time series and show experimentally that feature vectors have a sufficient difference that depends on Granger causality, which demonstrates the effectiveness of our proposed framework.

3.1.2 Related Work

Model-based methods for Granger Causality Discovery

Many approaches have been dedicated to inferring Granger causality by detecting the conditional (in)dependence relations relations between time-dependent variables.

This is because Granger causality is defined as the conditional (in)dependence relations, as described in Section 2.1.1. To illustrate the definition, consider the bivariate setting with $X = \{X_t\}$ and $Y = \{Y_t\}$, where each of X_t and Y_t is defined on $\mathcal{X}$ and $\mathcal{Y}$, respectively ($t = 1, 2, \dots$). Letting S_X and S_Y denote the observations of $\{X_1, \dots, X_t\}$ and $\{Y_1, \dots, Y_t\}$, the presence and the absence of bivariate Granger causality from X to Y (Definition 1) are defined as the following conditional (in)dependence relations:

$$Y_{t+1} \not\perp\!\!\!\perp S_X \mid S_Y \quad \text{and} \tag{3.1}$$

$$Y_{t+1} \perp\!\!\!\perp S_X \mid S_Y, \tag{3.2}$$

which are equivalent to the (non-)equality between the conditional distributions:

$$\mathrm{P}(Y_{t+1} \mid S_X, S_Y) \neq \mathrm{P}(Y_{t+1} \mid S_Y) \quad \text{and} \tag{3.3}$$

$$\mathrm{P}(Y_{t+1} \mid S_X, S_Y) = \mathrm{P}(Y_{t+1} \mid S_Y), \tag{3.4}$$

respectively.

Since it is challenging to determine whether the equality between two conditional distributions $\mathrm{P}(Y_{t+1} \mid S_X, S_Y)$ and $\mathrm{P}(Y_{t+1} \mid S_Y)$ holds, traditional methods [Bell *et al.*, 1996; Granger, 1969; Marinazzo *et al.*, 2008] develop a hypothesis testing that determines whether the following two conditional expected values are equal:

$$\mathbb{E}[Y_{t+1} \mid S_X, S_Y] = \mathbb{E}[Y_{t+1} \mid S_Y], \tag{3.5}$$

which is much simpler than testing the (non-)equality in Eqs. (3.3) and (3.4).

To represent the conditional expected values in Eq. (3.5), these traditional methods use the two time series regression models: one is fitted to the observations of variables X and Y, and the other is fitted only to the observations of Y. In particular, they use such regression models as the vector autoregressive (VAR) model [Granger, 1969], the generalized additive models (GAM) [Bell *et al.*, 1996], and kernel regression [Marinazzo *et al.*, 2008]. By computing the prediction errors based on the fitted regression models, these methods evaluate the test statistics and determine whether the equality in Eq. (3.5) holds.

The limitation of these model-based methods is that their inference accuracy of Granger causality depend greatly on whether each regression model can be well fitted to the data. Unfortunately, it is difficult in practice to select an appropriate regression model for each time series. To overcome this limitation, we propose a supervised learning approach that allows us to avoid this model selection issue.

Supervised Learning Approaches to Causal Discovery from I.i.d. Data

It is worth noting that several supervised learning approaches have been developed for causal discovery.

The idea of performing supervised learning for discovering the causal relationships has been first introduced in the data analysis competition called ChaLearn [Guyon, 2013]. ChaLearn provided each participant with a large collection of train-

ing data, each of which contained an i.i.d. dataset and a causal label representing the underlying causal direction in the dataset, and the participant attempted to train a classifier that accurately assigns a causal label to each i.i.d. dataset.

While the participants in ChaLearn focused on the laborious task of hand-crafting the features, the subsequent studies have developed more sophisticated feature representation using mutual information [Bontempi and Flauder, 2015], kernel mean embedding [Lopez-Paz *et al.*, 2015], and neural network [Lopez-Paz *et al.*, 2017].

Among these studies, the randomized causation coefficient (RCC) [Lopez-Paz *et al.*, 2015] is closely related with our proposed method because both methods use kernel mean embedding to obtain the features of the distribution that differ depending on the causal relationships. However, RCC and our method are designed to obtain different features of the distribution. In particular, the feature representation of RCC is founded on the postulate of causality, called the *independence of cause and mechanism* (ICM) [Janzing and Scholkopf, 2010], which states that the marginal and conditional distributions differ depending on the underlying causal directions. By contrast, our method is designed to detect Granger causality by measuring the distance between the conditional distributions between time-dependent variables.

All the existing supervised learning approaches to causal discovery are designed for i.i.d. data; hence, these approaches cannot be applied to discover Granger causality from time series data.

3.2 Causal Discovery from Time Series Data via Supervised Learning

3.2.1 Problem Statement

We consider a ternary classification task for inferring Granger causality from time series data. This classification task can be extended to multivariate time series as described in Section 3.2.5.

Let the training data be N time series $S^1, \ldots, S^N$ ($N > 0$), where each time series S^j ($j \in \{1, \ldots, N\}$) has length $T_j > 0$ and consists of the observations of random variables $\{(X_1^j, Y_1^j), \ldots, (X_{T_j}^j, Y_{T_j}^j)\}$. Each time series S^j has a ternary causal label $l^j \in \{+1, -1, 0\}$ that expresses Granger causality between $X^j = (X_1^j, \ldots, X_{T_j}^j)$ and

$Y^j = (Y_1^j, \ldots, Y_{T_j}^j)$ as $X^j \to Y^j$, $X^j \leftarrow Y^j$, or *No Causation*. Suppose that we have feature mapping function $\nu(\cdot)$ that converts each time series S^j to a feature vector.

Using these training data and feature mapping function $\nu(\cdot)$, we train a ternary classifier with $\{(\nu(S^j), l^j)\}_{j=1}^N$. Then the task of discovering Granger causality from another time series S' (i.e., a test data instance) can be rephrased as assigning a causal label to $\nu(S')$ by utilizing the trained classifier.

3.2.2 Classifier Design

To build a classifier that assigns causal labels to time series, we formulate the feature representation $\nu(\cdot)$. In what follows, we describe our ideas for obtaining feature vectors that are sufficiently different depending on Granger causality.

Basic Ideas for Granger Causality Identification

Following the definition of (bivariate) Granger causality (Definition 1),[1] we define causal labels as follows:[2]

$$X \to Y \text{ if } \begin{cases} \mathrm{P}(X_{t+1} \mid S_X, S_Y) & = \mathrm{P}(X_{t+1} \mid S_X) \\ \mathrm{P}(Y_{t+1} \mid S_X, S_Y) & \neq \mathrm{P}(Y_{t+1} \mid S_Y) \end{cases}, \tag{3.6}$$

$$X \leftarrow Y \text{ if } \begin{cases} \mathrm{P}(X_{t+1} \mid S_X, S_Y) & \neq \mathrm{P}(X_{t+1} \mid S_X) \\ \mathrm{P}(Y_{t+1} \mid S_X, S_Y) & = \mathrm{P}(Y_{t+1} \mid S_Y) \end{cases}, \tag{3.7}$$

$$\textit{No Causation} \text{ if } \begin{cases} \mathrm{P}(X_{t+1} \mid S_X, S_Y) & = \mathrm{P}(X_{t+1} \mid S_X) \\ \mathrm{P}(Y_{t+1} \mid S_X, S_Y) & = \mathrm{P}(Y_{t+1} \mid S_Y) \end{cases}, \tag{3.8}$$

where causal label $X \to Y$ states that X is the cause of Y and that Y is **not** the cause of X, and other causal labels are defined in the same way.

To assign causal labels to time series based on Eqs. (3.6), (3.7), and (3.8), it is necessary to determine whether or not the two conditional distributions are identical.

[1] Note that since our approach is founded on Definition 1, which cannot address the case where there are *latent confounders* (i.e., unobserved variables that influence both X and Y), as with the existing methods [Bell *et al.*, 1996; Cheng *et al.*, 2014; Granger, 1969; Marinazzo *et al.*, 2008; Sun, 2008], it does not deal with such a case.

[2] Although we do not consider the case where $\mathrm{P}(X_{t+1} \mid S_X, S_Y) \neq \mathrm{P}(X_{t+1} \mid S_X)$ and $\mathrm{P}(Y_{t+1} \mid S_X, S_Y) \neq \mathrm{P}(Y_{t+1} \mid S_Y)$ (i.e., X is the cause of Y, and Y is also the cause of X), we can straightforwardly address such a case by adding an extra label.

To represent the information about conditional distributions, instead of using regression models, we use kernel mean embedding, which maps a distribution to a point in the feature space called the RKHS. Interestingly, when a *characteristic* kernel (e.g., a Gaussian kernel) is used, the mapping is *injective*: different distributions are not mapped to the same point [Sriperumbudur *et al.*, 2010].

Suppose that kernel mean embedding maps conditional distributions $\mathrm{P}(X_{t+1} \mid S_X, S_Y)$, $\mathrm{P}(X_{t+1} \mid S_X)$ and $\mathrm{P}(Y_{t+1} \mid S_X, S_Y)$, $\mathrm{P}(Y_{t+1} \mid S_Y)$ to the points in the RKHS, $\mu_{X_{t+1}|S_X,S_Y}, \mu_{X_{t+1}|S_Y} \in \mathcal{H}_X$ and $\mu_{Y_{t+1}|S_X,S_Y}, \mu_{Y_{t+1}|S_Y} \in \mathcal{H}_Y$, respectively, where $\mathcal{H}_X$ and $\mathcal{H}_Y$ are the RKHSs defined on $\mathcal{X}$ and $\mathcal{Y}$. Then, with a characteristic kernel, the causal labels in Eqs. (3.6), (3.7), and (3.8) can be rewritten as

$$X \to Y \quad \text{if} \quad \begin{cases} \mu_{X_{t+1}|S_X,S_Y} & = \mu_{X_{t+1}|S_X} \\ \mu_{Y_{t+1}|S_X,S_Y} & \neq \mu_{Y_{t+1}|S_Y} \end{cases}, \tag{3.9}$$

$$X \leftarrow Y \quad \text{if} \quad \begin{cases} \mu_{X_{t+1}|S_X,S_Y} & \neq \mu_{X_{t+1}|S_X} \\ \mu_{Y_{t+1}|S_X,S_Y} & = \mu_{Y_{t+1}|S_Y} \end{cases}, \tag{3.10}$$

$$\textit{No Causation} \quad \text{if} \quad \begin{cases} \mu_{X_{t+1}|S_X,S_Y} & = \mu_{X_{t+1}|S_X} \\ \mu_{Y_{t+1}|S_X,S_Y} & = \mu_{Y_{t+1}|S_Y} \end{cases}, \tag{3.11}$$

To assign causal labels based on Eqs. (3.9), (3.10), and (3.11), we only have to determine whether the two points in the RKHS are the same over time t; equivalently, whether the distance between the two points in the RKHS is zero over time t.

In the literature on kernel methods, such a distance is termed the MMD [Gretton *et al.*, 2007]. Let k_X be a positive-definite kernel function defined on $\mathcal{X}$, $\Phi_X(x) \coloneqq k_X(x, \dot{})$ be a feature mapping function of k_X,[3] and $\mathcal{H}_X$ be the RKHS induced by k_X. Then kernel mean embeddings $\mu_{X_{t+1}|S_X,S_Y}$ and $\mu_{X_{t+1}|S_X}$ are defined as the following conditional expected values of feature mapping Φ_X whose expectations are taken with respect to conditional distributions $\mathrm{P}(X_{t+1} \mid S_X, S_Y)$ and $\mathrm{P}(X_{t+1} \mid S_X)$:

$$\mu_{X_{t+1}|S_X,S_Y} = \mathbb{E}_{X_{t+1}|S_X,S_Y}[\Phi_X(X_{t+1})] \tag{3.12}$$

$$\mu_{X_{t+1}|S_X} = \mathbb{E}_{X_{t+1}|S_X}[\Phi_X(X_{t+1})]. \tag{3.13}$$

[3]For instance, when using Gaussian kernel $k_X(x, x') = \exp(-\gamma\|x - x'\|^2)$ ($\gamma > 0$ is a parameter), the feature mapping is given as $\Phi_X(x) = \exp(-\gamma x^2)[1, \sqrt{2\gamma/1!}\, x, \sqrt{(2\gamma)^2/2!}\, x^2, \dots]^\top$.

Using kernel mean embeddings $\mu_{X_{t+1}|S_X,S_Y}$ and $\mu_{X_{t+1}|S_X}$, the MMD for two conditional distributions $\mathrm{P}(X_{t+1} \mid S_X, S_Y)$ and $\mathrm{P}(X_{t+1} \mid S_X)$ is defined as

$$\mathrm{MMD}^2_{X_{t+1}} \equiv \|\mu_{X_{t+1}|S_X,S_Y} - \mu_{X_{t+1}|S_X}\|^2_{\mathcal{H}_X}. \tag{3.14}$$

In the same way, $\mathrm{MMD}^2_{Y_{t+1}}$ is defined as the distance between $\mu_{Y_{t+1}|S_X,S_Y}$, $\mu_{Y_{t+1}|S_Y} \in \mathcal{H}_Y$, where $\mathcal{H}_Y$ is the RKHS defined by positive-definite kernel function k_Y on $\mathcal{Y}$.

The MMD is a metric between distributions that can be estimated without fitting regression models or performing density estimation. At this point, the MMD is much more attractive than other measures of the distributional discrepancy, such as the Kolmogorov-Smirnov statistic Chen and An [1997] and the Kullback-Leibler divergence [Kullback and Leibler, 1951]: the former requires us to select regression models and the latter requires a density estimation, which is difficult when there are insufficient samples.

However, it is challenging to estimate $\mathrm{MMD}^2_{X_{t+1}}$ and $\mathrm{MMD}^2_{Y_{t+1}}$ because conditional distributions $\mathrm{P}(X_{t+1} \mid S_X, S_Y)$ and $\mathrm{P}(X_{t+1} \mid S_X)$ are conditioned on variable values before time t, and hence we need to address long term dependence among time-dependent variables. Below we describe how we can overcome this challenge.

3.2.3 MMD Estimators

To compute the MMD in Eq. (3.14), we estimate kernel mean embeddings $\mu_{X_{t+1}|S_X,S_Y}$ and $\mu_{X_{t+1}|S_X}$ in Eqs. (3.12) and (3.13). To achieve this, we need to take the expectations of feature mapping $\Phi_X(X_{t+1})$ with respect to the conditional distributions conditioned on the variable values before time t.

We take such expectations by employing the existing time series prediction method called the kernel Kalman filter based on a conditional embedding operator (KKF-CEO) [Zhu *et al.*, 2014]. KKF-CEO is founded on a *state-space model*, which, unlike the first-order Markov processes, can deal with long term dependence among time-dependent variables. Using such a generative model, it performs time series prediction by estimating an expected value of feature mapping function.

In particular, Zhu *et al.* [2014] have formulated the estimator of such an expected value as the weighted sum of the feature mapping function. In case of kernel mean embeddings $\mu_{X_{t+1}|S_X,S_Y}$ and $\mu_{X_{T+1}|S_X}$ in Eqs. (3.12) and (3.13), we can formulate

their estimators as

$$\hat{\mu}_{X_{t+1}|S_X,S_Y} = \sum_{\tau=2}^{t-1} w_\tau^{XY} \Phi_X(x_\tau) \tag{3.15}$$

$$\hat{\mu}_{X_{t+1}|S_X} = \sum_{\tau=2}^{t-1} w_\tau^{X} \Phi_X(x_\tau), \tag{3.16}$$

where $\mathbf{w^{XY}} = [w_2^{XY}, \ldots, w_{t-1}^{XY}]^\top$ and $\mathbf{w^{X}} = [w_2^{X}, \ldots, w_{t-1}^{X}]^\top$ $(t > 3)$ are the real-valued weight vectors. By applying Eqs. (3.15) and (3.16) to Eq. (3.14), the estimator of $\mathrm{MMD}^2_{X_{t+1}}$ is given as

$$\widehat{\mathrm{MMD}}^2_{X_{t+1}} = \sum_{\tau=2}^{t-1}\sum_{\tau'=2}^{t-1}(w_\tau^{XY} w_{\tau'}^{XY} + w_\tau^{X} w_{\tau'}^{X} - 2w_\tau^{XY} w_{\tau'}^{X}) k_X(x_\tau, x_{\tau'}). \tag{3.17}$$

To estimate weight vectors $\mathbf{w^X}$ and $\mathbf{w^{XY}}$, we used the KKF-CEO's estimation algorithm, which is developed in Zhu *et al.* [2014] for performing time series prediction. We can compute weight vector $\mathbf{w^X}$ by directly employing it. As regards weight vector $\mathbf{w^{XY}}$, we simply ran KKF-CEO using product kernel $k_X \dot{k}_Y$. Although computing these weight vectors requires us to set the values of several hyperparameters, they can be appropriately selected for each time series by performing time series prediction with KKF-CEO and minimizing the prediction errors.

3.2.4 Feature Representation

To build a classifier for Granger causality identification, we obtain the feature vectors using MMD pair, denoted by $d_t \coloneqq [\widehat{\mathrm{MMD}}^2_{X_{t+1}}, \widehat{\mathrm{MMD}}^2_{Y_{t+1}}]^\top$.

By designing feature vectors with this MMD pair, we can expect time series with different causal labels to yield sufficiently different feature vectors. This is because as indicated by Eqs. (3.9), (3.10), and (3.11), whether the MMD becomes zero depends on causal labels. Note that each MMD in d_t does not exactly take zero because it is a finite sample estimate. However, we can expect sufficiently different MMD pairs to be estimated from time series with different causal labels, as intuitively shown in Figure 3.1; we experimentally confirm this difference in Section 3.3.2.

The computational difficulty of preparing MMD pair d_t is that the estimator in Eq. (3.17) consists weight vectors $\mathbf{w^X}$ and $\mathbf{w^{XY}}$, whose estimation with KKF-CEO's

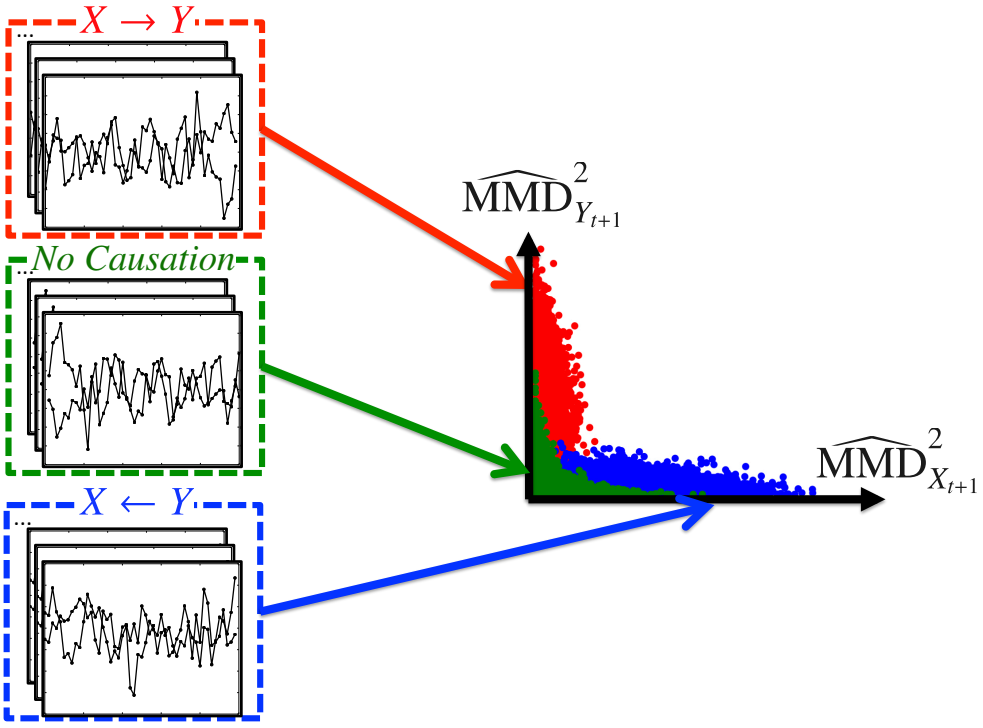

Figure 3.1: Different MMD pairs are estimated from time series with different causal labels. Each dot represents the MMD pair estimated from each time series.

algorithm requires time $O(T^3)$ for time series with length T [Zhu *et al.*, 2014]. To reduce this computation time, we chop a given time series with length T, denoted by $S = \{(x_1, y_1), \cdots, (x_T, y_T)\}$, into the subsequences with length W ($W < T$), i.e., $\{(x_{t-(W-1)}, y_{t-(W-1)}), \cdots, (x_t, y_t)\}$ ($t = W, \cdots, T$). By computing the MMD pairs for each subsequence, we obtain the MMD pairs $\{d_W, \cdots, d_T\}$.

A naive approach to obtaining a feature vector from these MMD pairs is to concatenate them into a single vector. Unfortunately, such a vector has dimensionality $2(T - W + 1)$, which is different depending on time series length T and hence makes it impossible to use the time series with different lengths as training data. Another approach is to formulate a feature vector as the average over the MMD pairs, i.e., $\frac{1}{T-W+1}\sum_{t=W}^{T} d_t$. However, this feature vector can take an identical value between the two different empirical distributions of the MMD pairs if their means are identical.

To avoid mapping different empirical distributions of the MMD pairs to an identical feature vector, we represent a feature vector by utilizing kernel mean embedding again. Let k_D denote a kernel function that measures the similarity between the

MMD pairs. Using kernel function k_D, we define our feature representation as

$$\nu(S) \coloneqq \frac{1}{T-W+1}\sum_{t=W}^{T}\Phi_D(d_t) \quad \text{where} \quad d_t = [\widehat{\text{MMD}}^2_{X_{t+1}}, \widehat{\text{MMD}}^2_{Y_{t+1}}]^\top \tag{3.18}$$

which is an average of feature mapping $\Phi_D(d_t) \coloneqq k_D(d_t, \cdot)$.[4]

To compute feature mapping $\Phi_D(\cdot)$ in Eq. (3.18), we employed random Fourier features (RFF) [Rahimi and Recht, 2007], which approximate the feature mapping as a low-dimensional vector of random features that are sampled from the Fourier transform of the kernel function. In experiments, we set the number of features $m = 100$ and obtained an m-dimensional feature vector for each time series, where we observed no significant improvements in the inference accuracy when using a larger m.

3.2.5 Extensions to Multivariate Time Series

So far, we have presented a supervised learning approach to discovering Granger causality from bivariate time series data. In this section, we describe how our approach can be extended to n-variate time series ($n \geq 3$). In what follows, we first present the feature representation for trivariate time series (i.e., $n = 3$) and then present an extension for addressing the case with $n > 3$.

Trivariate Time Series

Our feature representation for trivariate time series is founded on conditional Granger causality [Geweke, 1984] in Definition 2, which can be applied to multivariate time series unlike bivariate Granger causality in Definition 1.

As described in Section 2.1.1, naively applying bivariate Granger causality (Definition 1) to trivariate time series leads to wrong inference results. To illustrate this, consider the case where there is no causal relationship between X and Y. In this case, if the third variable Z is the common cause of X and Y and yield the correlation between them, we can wrongly conclude that X is the cause of Y or

[4]Note that unlike the estimators in (3.15) and (3.16), the one in (3.18) is formulated as a weighted sum with an identical weight value, i.e., $\frac{1}{T-W+1}$. This formulation difference arises depending on whether the data instances are directly drawn from the target distribution (see Muandet *et al.* [2017] for the detail).

that Y is the cause of X. This is because $\mathrm{P}(Y_{t+1} \mid S_X, S_Y) \neq \mathrm{P}(Y_{t+1} \mid S_Y)$ or $\mathrm{P}(X_{t+1} \mid S_X, S_Y) \neq \mathrm{P}(X_{t+1} \mid S_X)$ might hold due to the influence of Z.

To address the influence of Z, conditional Granger causality compares the two conditional distributions conditioned on Z's observations. Let S_Z be the observations of $\{Z_1, \cdots, Z_t\}$, each of which is defined on $\mathcal{Z}$. Then conditional Granger causality defines X as the cause of Y given Z if $\mathrm{P}(Y_{t+1} \mid S_X, S_Y, S_Z) \neq \mathrm{P}(Y_{t+1} \mid S_Y, S_Z)$ holds; otherwise, X is **not** the cause of Y given Z.

We define the causal labels based on conditional Granger causality. As with Eq. (3.6), we define causal label $X \to Y$ as

$$X \to Y \text{ if } \begin{cases} \mathrm{P}(X_{t+1} \mid S_X, S_Y, S_Z) & = \mathrm{P}(X_{t+1} \mid S_X, S_Z) \\ \mathrm{P}(Y_{t+1} \mid S_X, S_Y, S_Z) & \neq \mathrm{P}(Y_{t+1} \mid S_Y, S_Z), \end{cases}$$

which can be rewritten using kernel mean embedding as

$$X \to Y \quad \text{if } \begin{cases} \mu_{X_{t+1}|S_X,S_Y,S_Z} & = \mu_{X_{t+1}|S_X,S_Z} \\ \mu_{Y_{t+1}|S_X,S_Y,S_Z} & \neq \mu_{Y_{t+1}|S_Y,S_Z} \end{cases},$$

where $\mu_{X_{t+1}|S_X,S_Y,S_Z}$, $\mu_{X_{t+1}|S_X,S_Z}$, $\mu_{Y_{t+1}|S_X,S_Y,S_Z}$, and $\mu_{Y_{t+1}|S_Y,S_Z}$ are the kernel mean embeddings of $\mathrm{P}(X_{t+1} \mid S_X, S_Y, S_Z)$, $\mathrm{P}(X_{t+1} \mid S_X, S_Z)$, $\mathrm{P}(Y_{t+1} \mid S_X, S_Y, S_Z)$, and $\mathrm{P}(Y_{t+1} \mid S_Y, S_Z)$, respectively. Other causal labels, $X \leftarrow Y$ and *No Causation*, can be defined in the same way.

To assign these causal labels to trivariate time series, we formulate our feature representation by adding the MMD between $\mu_{X_{t+1}|S_X,S_Y,S_Z}$ and $\mu_{X_{t+1}|S_X,S_Z}$ and the one between $\mu_{Y_{t+1}|S_X,S_Y,S_Z}$ and $\mu_{Y_{t+1}|S_Y,S_Z}$. In particular, we modify the formulation of MMD pair d_t in feature representation in Eq. (3.18) to

$$d_t \coloneqq [\widehat{\mathrm{MMD}}^2_{X_{t+1}}, \widehat{\mathrm{MMD}}^2_{Y_{t+1}}, \widehat{\mathrm{MMD}}^2_{X_{t+1}|Z}, \widehat{\mathrm{MMD}}^2_{Y_{t+1}|Z}]^\top,$$

where $\widehat{\mathrm{MMD}}^2_{X_{t+1}|Z}$ and $\widehat{\mathrm{MMD}}^2_{Y_{t+1}|Z}$ denote the estimators of the MMD between $\mu_{X_{t+1}|S_X,S_Y,S_Z}$ and $\mu_{X_{t+1}|S_X,S_Z}$ and the one between $\mu_{Y_{t+1}|S_X,S_Y,S_Z}$ and $\mu_{Y_{t+1}|S_Y,S_Z}$, respectively.

n-variate Time Series ($n > 3$)

Although it is possible to develop the feature representation for n-variate time series ($n > 3$) by adding the extra MMDs to d_t, it is challenging to prepare enough training data to train the classifier since the number of possible combinations of the common cause variables of the variable pair X and Y grows super-exponentially in n.

For this reason, we used the feature representation for trivariate time series. From n-variate time series, we infer a causal relationship between each variable pair X and Y in three steps:

1. For each $v \in \{1, \cdots, n-2\}$, we obtain the feature vector from the observations of the triplet of the variables (X, Y, Z_v).

2. Using each feature vector, we use a trained classifier to compute the probabilities of the three labels ($X \to Y$, $X \leftarrow Y$, and *No Causation*).

3. Finally, we assign the label with the highest average probability.

Addressing the cases where there are more than one common cause variable is left as our future work.

3.3 Experiments

3.3.1 Experimental Settings

Baseline Methods

We compared the performance of our method (hereafter referred to as the supervised inference of Granger causality (**SIGC**)) with the following five baselines:

- **RCC** [Lopez-Paz *et al.*, 2015],[5] the supervised learning method for i.i.d. data
- **GC**$_{VAR}$ [Granger, 1969]:[6] which uses the VAR model to infer Granger causality.
- **GC**$_{GAM}$ [Bell *et al.*, 1996][7], which uses the GAM to identify Granger causality.

[5] https://github.com/lopezpaz/causation_learning_theory
[6] http://people.tuebingen.mpg.de/jpeters/onlineCodeTimino.zip

- **GC**$_{KER}$ [Marinazzo *et al.*, 2008],[7] which performs kernel regression to discover Granger causality.

- **TE** [Schreiber, 2000],[8] which identifies the causal relationships based on transfer entropy (see Definition 3 in Section 2.1.2 for the detail).

For our **SIGC**, we used a random forest classifier[9] to make a fair comparison with **RCC**, which has achieved better performance with the random forest classifier than with the SVM [Lopez-Paz *et al.*, 2015]. To prepare feature vectors, we used the Gaussian kernel as k_X, k_Y, and k_D and set the kernel parameter using the median heuristic, which is a well-known heuristic for selecting it [Scholkopf and Smola, 2001]. We set the parameter W in our method and the parameters in the existing methods to provide the best performance for each method in our synthetic data experiments. For our method, we selected $W = 12$.

Classifier Training

To evaluate the performance of the supervised learning methods (i.e., **SIGC** and **RC**), we trained a classifier using synthetic training data. This is because as described in Lopez-Paz *et al.* [2015], there are few real-world data where the causal relationships are known.

As training data of bivariate time series, we generated $15,000$ pairs of synthetic time series with length $T = 42$ so that there were $5,000$ instances each with causal labels $X \to Y$, $X \leftarrow Y$, and *No Causation*. Here, we used the following *linear* and *nonlinear* time series:

- Linear time series were sampled from the VAR model:

$$\begin{bmatrix} X_t \\ Y_t \end{bmatrix} = \frac{1}{P} \sum_{\tau=1}^{P} A_\tau \begin{bmatrix} X_{t-\tau} \\ Y_{t-\tau} \end{bmatrix} + \begin{bmatrix} E_{X_t} \\ E_{Y_t} \end{bmatrix} \tag{3.19}$$

  where $\tau = 1, \ldots, P$ ($P \in \{1, 2, 3\}$) and E_{X_t}, E_{Y_t} denote noise variables, which were sampled from the Gaussian distribution $\mathcal{N}(0, 1)$. To obtain time series

[7] https://github.com/danielemarinazzo/KernelGrangerCausality
[8] https://github.com/Healthcast/TransEnt
[9] The number of trees is selected from $\{100, 200, 500, 1000, 2000\}$ via 5-fold cross validation.

with $X \to Y$, we used the following coefficient matrix

$$A_\tau = \begin{bmatrix} a_\tau & 0.0 \\ c_\tau & d_\tau \end{bmatrix}$$

where a_τ, d_τ were drawn from the uniform distribution $\mathcal{U}(-1, 1)$, and $c_\tau \in \{-1, 1\}$. Similarly, we prepared time series with $X \leftarrow Y$, and *No Causation*.

- Nonlinear time series were also similarly generated by using the VAR model with a standard sigmoid function $g(x) = 1/(1 + \exp(-x))$. For instance, we prepared time series with $X \to Y$ so that Y_t depended on $\{[g(X_{t-\tau}), Y_{t-\tau}]^\top\}_{\tau=1}^P$ while X_t depended only on $\{X_{t-\tau}\}_{\tau=1}^P$.

- Finally, we scaled each time series with mean 0 and variance 1.

To test the performance on multivariate time series data, we prepared synthetic trivariate time series in a similar way.

3.3.2 Experiments on Bivariate Time Series Data

Synthetic Time Series

We tested our method using synthetic time series data. As test data, we employed the following linear and nonlinear test data:

- *Linear Test Data*: We prepared 300 pairs of linear time series so that the numbers of time series with $X \to Y$, $X \to Y$, and *No Causation* were 100. As with the linear time series in the training data, each time series was sampled from the VAR model (3.19) although several parameter settings were different (e.g., the noise variance was given as $p \in \{0.5, 1.0, 1.5, 2.0\}$).

- *Nonlinear Test Data*: We used 300 pairs of nonlinear time series, where there were 100 time series with $X \to Y$, $X \to Y$, and *No Causation* in each dataset. We generated nonlinear time series with $X \to Y$ by

$$X_t = 0.2X_{t-1} + 0.9E_{X_t} \tag{3.20}$$

$$Y_t = -0.5 + \exp(-(X_{t-1} + X_{t-2})^2) + 0.7\cos(Y_{t-1}^2) + 0.3E_{Y_t} \tag{3.21}$$

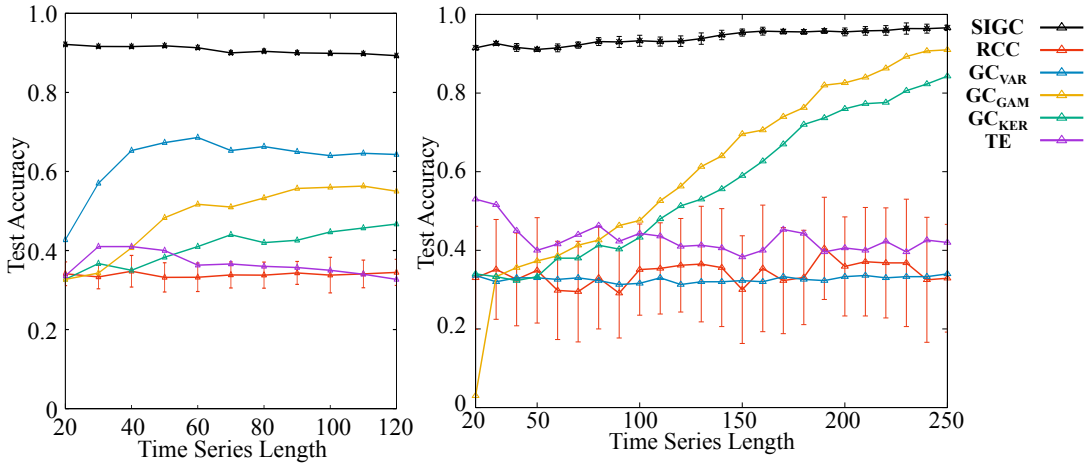

Figure 3.2: Test accuracy for 300 pairs of time series against time series length (left: linear test data; right: nonlinear test data). Means and standard deviations (error bars) are shown for **SIGC** and **RCC** based on 20 runs with different training data.

where the noise variables E_{X_t}, E_{Y_t} were sampled from $\mathcal{N}(0,1)$. Similarly, we prepared nonlinear time series with $X \leftarrow Y$. To prepare time series with *No Causation*, we simply ignored the exponential term in Eq. (3.21).

Using linear and nonlinear test data, we compared the performance of our method with that of the existing methods. Figure 3.2 shows the test accuracy. Note that for **SIGC** and **RCC**, we show the means and the standard deviations (error bars) in 20 experiments with different training data since these methods use randomly generated training data.

As expected, the performance of Granger causality methods $\mathbf{GC}_{VAR}$, $\mathbf{GC}_{GAM}$, and $\mathbf{GC}_{KER}$ depended on whether or not the regression model could be well fitted to the data. For instance, since the VAR model could be well fitted to linear test data, $\mathbf{GC}_{VAR}$ performed well on linear test data although it worked badly on nonlinear test data. In addition, with nonlinear test data, $\mathbf{GC}_{KER}$ and **TE** were less accurate than $\mathbf{GC}_{GAM}$ because the time series was too short to perform kernel regression or density estimation.

By contrast, our method worked sufficiently well on linear and nonlinear test data. The main reason for the good performance lies in our feature representation. This can be seen from a comparison with the supervised learning method **RCC** since it prepares training data in the same way as our method.

To verify our feature representation, we confirmed that feature vectors are suffi-

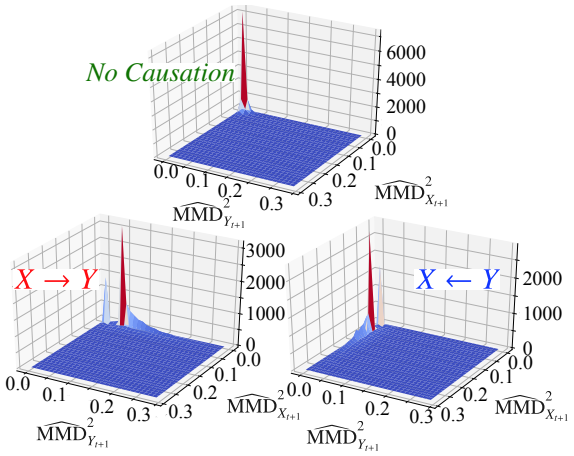

Figure 3.3: Histogram of MMDs computed with linear test data with $X \to Y$ (top left), $X \leftarrow Y$ (top right), and *No Causation* (bottom)

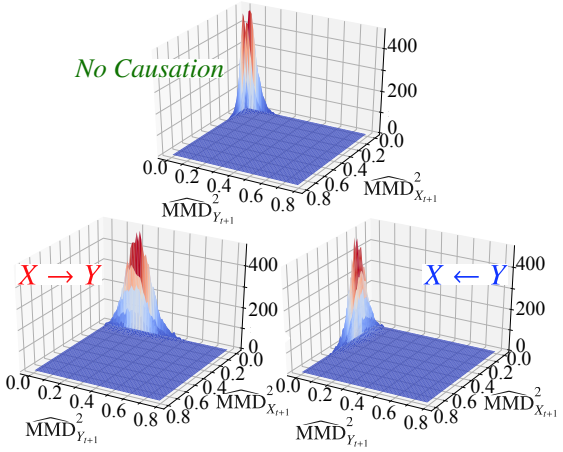

Figure 3.4: Histogram of MMDs computed with nonlinear test data with $X \to Y$ (top left), $X \leftarrow Y$ (top right), and *No Causation* (bottom)

ciently different depending on causal labels. To do so, we used linear and nonlinear test data to plot a histogram of the MMD pairs $\{d_t\}$ that were used to compute the feature vector for each time series. Figure 3.4 show the results. Since each MMD in d_t is a finite sample estimate, no MMD becomes exactly zero. However, the MMDs were sufficiently different depending on the causal labels.

From such a large difference among the MMD pairs, one may consider the following naive *unsupervised* approach that assigns causal labels in two steps:

1. Perform the hypothesis tests that determines whether the mean of $\widehat{\mathrm{MMD}}^2_{X_{t+1}}$ is zero and whether the mean of $\widehat{\mathrm{MMD}}^2_{Y_{t+1}}$ is zero.

2. Assign causal labels (i.e., $X \to Y$, $X \leftarrow Y$, or *No Causation*) to each time series using the p-values of the above hypothesis tests and some threshold value (i.e., significance level).

However, we confirmed that the performance of such an unsupervised approach depended greatly on the threshold value. What is worse, it was less accurate than our method; for instance, its test accuracy was 0.810 on nonlinear test data with length $T = 250$ (not shown in Figure 3.2) while our method achieved 0.966. These results suggest the effectiveness of our supervised learning approach, which can obtain the decision boundary needed to determine the causal label by training a classifier.

Real-world Time Series

We tested our method using real-world time series data. To improve the reliability of the experiment, we used the following two test datasets:

- The first test dataset was composed of five pairs of bivariate time series downloaded from the Cause-Effect Pairs database [Mooij *et al.*, 2016], whose true causal relationships are reported in Mooij *et al.* [2016] as $X \to Y$ for three pairs and as $X \leftarrow Y$ for the others. For instance, the *River Runoff* is a bivariate time series concerning average precipitation X and average river runoff Y, and the true causal relationship is regarded as $X \to Y$.

- Using the aforementioned five real-world time series, we prepared the second test dataset as a collection of the subsequences in each time series, each of which has length $T = 200$.

As regards training data, we used synthetic time series that we prepared in the same way as those for synthetic data experiments.

Tables 3.1 and 3.2 show the results for each test dataset. Our **SIGC** outperformed the other existing methods regardless of time series length T. Although training data were randomly generated, the inference results of **SIGC** were always correct throughout 20 experiments, while those of **RCC** changed depending on the randomness (for this reason, we have omitted the results of **RCC** in Table 3.1).

Table 3.1: Causal relationships inferred from first real-world test dataset. ✓and ✗ denote correct and incorrect results, respectively.

	SIGC	**GC**$_{VAR}$	**GC**$_{GAM}$	**GC**$_{KER}$	**TE**
Temperature ($T = 16382$)	✓	✗	✓	✓	✗
Radiation ($T = 8401$)	✓	✓	✓	✓	✓
Internet ($T = 498$)	✓	✓	✗	✗	✓
Sun Spots ($T = 1632$)	✓	✗	✗	✗	✓
River Runoff ($T = 432$)	✓	✓	✓	✗	✓

Table 3.2: Test accuracy on second real-world test dataset. Means and standard deviations are shown for **SIGC** and **RCC** based on 20 runs.

	SIGC	**RCC**	**GC**$_{VAR}$	**GC**$_{GAM}$	**GC**$_{KER}$	**TE**
Temperature ($T = 200$)	**0.961** (0.011)	0.432 (0.242)	0.950	0.848	0.234	0.492
Radiation ($T = 200$)	**0.987** (0.053)	0.515 (0.345)	0.156	0.0	0.782	0.394
Internet ($T = 200$)	**1.0** (0.0)	0.478 (0.222)	0.157	0.387	0.261	0.498
Sun Spots ($T = 200$)	**1.0** (0.0)	0.435 (0.182)	0.908	0.704	0.076	0.522
River Runoff ($T = 200$)	**0.958** (0.058)	0.399 (0.193)	0.684	0.406	0.155	0.485

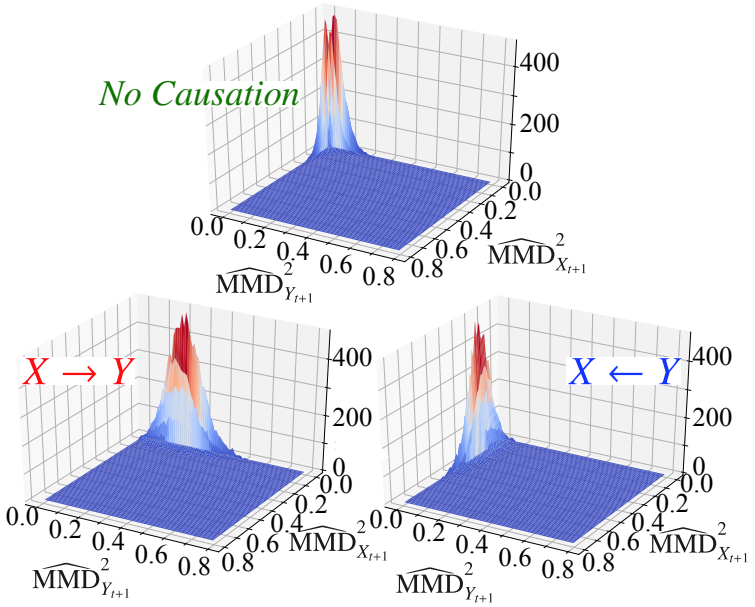

Figure 3.5: Histogram of MMDs used to compute the feature vector for each time series in second real-world dataset with $X \to Y$ (left) and $X \leftarrow Y$ (right)

In addition, we confirmed that the feature vectors obtained from these real-world data are also sufficiently different depending on causal labels. Figure 3.5 shows the results on the second real-world dataset. As expected, they demonstrate that the feature vectors are sufficiently different depending on causal labels.

3.3.3 Experiments on Multivariate Time Series Data

Using synthetic and real-world data, we evaluated the performance of $\mathbf{SIGC}_{tri}$, which utilizes a feature representation for trivariate time series.

Synthetic Data Experiments

We tested $\mathbf{SIGC}_{tri}$ using synthetic trivariate time series data. We generated the test data from the following *three logistic map* [Ott, 2002]:

$$\begin{aligned} X_t &= 0.8(1 - aX_{t-1}^2) + 0.2(1 - aY_{t-1}^2) + sE_{X_t} \\ Y_t &= 1 - aY_{t-1}^2 + sE_{Y_t} \\ Z_t &= 0.8(1 - aZ_{t-1}^2) + 0.2(1 - aX_{t-1}^2) + sE_{Z_t}, \end{aligned} \tag{3.22}$$

where $a = 1.8$, $s = 0.01$, and noise variables E_{X_t}, E_{Y_t}, and E_{Z_t} were drawn from standard Gaussian distribution $\mathcal{N}(0, 1)$. In particular, we prepared 100 trivariate time series test data with length $T = 1000$ by sampling 100 triplets of initial values X_1, Y_1, and Z_1 from Uniform distribution $\mathcal{U}(0, 1)$. The true causal labels are $X \leftarrow Y$ for X and Y, $X \rightarrow Z$ for X and Z, and *No Causation* for Y and Z.

Using such test data, we compared the performance of $\mathbf{SIGC}_{tri}$ with the proposed method for bivariate time series data, denoted by $\mathbf{SIGC}_{bi}$. Table 3.3 presents the test accuracy of each method.

With these synthetic time series, $\mathbf{SIGC}_{tri}$ and $\mathbf{GC}_{KER}$ always correctly inferred the Granger causality because they are founded on conditional Granger causality and thus can be applied to trivariate time series data. By contrast, bivariate causal discovery methods, including $\mathbf{SIGC}_{bi}$, worked poorly. These results demonstrate that our $\mathbf{SIGC}_{tri}$ can effectively deal with the influence of the third variable using the feature representation for trivariate time series data.

Table 3.3: Test accuracy on trivariate synthetic time series data. Means and standard deviations are shown for **SIGC** and **RCC** based on 10 runs.

	Methods for trivariate time series			Methods for bivariate time series			Method for i.i.d. data
	$\mathbf{SIGC}_{tri}$	$\mathbf{GC}_{VAR}$	$\mathbf{GC}_{KER}$	$\mathbf{SIGC}_{bi}$	$\mathbf{GC}_{GAM}$	**TE**	**RCC**
Test accuracy	**1.0** (0.0)	0.72	1.0	0.0 (0.0)	0.0	0.0	0.0 (0.0)

Table 3.4: Macro and micro-averaged F1 scores. Means and standard deviations are shown for our methods and **RCC** based on 10 runs.

	$\mathbf{SIGC}_{tri}$	$\mathbf{SIGC}_{bi}$	**RCC**	$\mathbf{GC}_{VAR}$	$\mathbf{GC}_{GAM}$	$\mathbf{GC}_{KER}$	**TE**
Macro F1	**0.483** (0.0)	0.431 (0.007)	0.407 (0.096)	0.457	0.437	0.351	0.430
Micro F1	**0.637** (0.0)	0.578 (0.011)	0.567 (0.161)	0.567	0.513	0.436	0.449

Real-world Data Experiments

Finally, we evaluated the performance of $\mathbf{SIGC}_{tri}$ using the time series gene expression data. In particular, we used the *Saccharomyces cerevisiae* (yeast) cell cycle gene expression dataset collected by [Spellman *et al.*, 1998]. By combining four short time series that were measured in different microarray experiments, we prepared a time series with the length $T = 57$, where the number of genes was $n = 14$. Following the gene network database called Kyoto encyclopedia of genes and genomes (KEGG).[10], we determined the true causal relationships between the genes.

To evaluate the performance of each method, we used the macro and micro-averaged F1 scores because the number of non-causally-related gene pairs was much larger than the number of causally-related gene pairs.

Table 3.4 shows the results. Since the data were measured in different microarray experiments, all the methods could not sufficiently work well. However, our $\mathbf{SIGC}_{tri}$ worked better than the existing Granger causality methods. It also performed better than $\mathbf{SIGC}_{bi}$, which uses the feature representation for bivariate time series, thus indicating that it is important to consider the influence of the common cause variable as described in Section 3.2.5.

[10] https://www.genome.jp/kegg/

3.4 Conclusion

We have proposed a classification approach to Granger causality identification. Whereas the performance of the model-based methods depended hugely on whether the regression model could be well fitted to the data, our method performed sufficiently well by using the same feature representation and the same classifier (random forest classifier). Furthermore, we demonstrated experimentally the reason for such good performance by showing a sufficient difference between the feature vectors that depends on Granger causality. These results demonstrate the effectiveness of classification approaches to Granger causality identification.

Addressing complicated real-world scenarios (e.g., inferring the causal directions that change over time) constitutes our future work.

Chapter 4

Feature Selection for Discovering Distributional Treatment Effect Modifiers

In this chapter, we consider the problem of finding the features relevant to the difference in treatment effects. This problem is essential to elucidate the reason why the treatment effects are different across individuals, which leads to deep understanding of the underlying causal mechanisms. Existing methods seek the features relevant to treatment effect heterogeneity by measuring how greatly the feature attributes affect the degree of the *conditional average treatment effect* (CATE). However, these methods may overlook important features because CATE, a measure of the average treatment effect, cannot detect the differences in distribution parameters other than the mean (e.g., variance). To resolve this limitation of the existing methods, we propose a feature selection framework for discovering *distributional treatment effect modifiers*, i.e., the features related to *distributional* treatment effect heterogeneity.

4.1 Introduction

When the effects of a treatment (e.g., drug administration) differ across individuals, elucidating why such heterogeneity exists is critical in many applications such as precision medicine [Lee *et al.*, 2018], personalized education [Schochet *et al.*, 2014],

and targeted advertising [Taddy *et al.*, 2016]. A popular approach to explaining treatment effect heterogeneity is to identify the features of an individual that are relevant to the degree of a treatment effect. For instance, to unveil the mechanism of COVID-19 vaccines, recent medical studies have sought the features related to the degree of vaccine-acquired immunity [Jabal *et al.*, 2021].

To find such features, we need to measure how greatly the attributes of each feature influence the degree of a treatment effect. To this end, the existing methods use the *conditional average treatment effect* (CATE) that is conditioned on each feature, i.e., an average treatment effect across the individuals who have an identical attribute of each feature [Imai and Ratkovic, 2013; Tian *et al.*, 2014; Zhao *et al.*, 2022]. However, this average cannot capture distribution parameters other than the mean, such as the variance. As a result, if the attributes of a feature do not affect the average treatment effect but influence other distribution parameters, these mean-based methods will incorrectly conclude that the feature is unrelated to heterogeneity of the treatment effect.

The goal of this paper is to propose a feature selection framework for discovering *distributional treatment effect modifiers*. To achieve this goal, we develop a feature importance measure that quantifies how greatly the attributes of each feature influence the discrepancy between the distributions of *potential outcomes*, i.e., the outcomes when an individual is treated and not treated. We formulate this measure as a variance of the maximum mean discrepancy (MMD) [Gretton *et al.*, 2012] between the conditional potential outcome distributions conditioned on each feature. We derive its computationally efficient estimator using a kernel approximation technique and establish a feature selection algorithm that can control the type I error rate (i.e., the proportion of false-positive results) to the desired level.

4.1.1 Contributions

Our contributions are summarized as follows:

- We formulate an MMD-based feature importance measure for discovering distributional treatment effect modifiers (Section 4.3.2). We derive its computationally efficient weighted estimator using a kernel approximation technique (Section 4.3.3).

- We develop an algorithm that selects distributional treatment effect modifiers while controlling the type I error rate (Section 4.3.4). To evaluate the significance, we perform multiple hypothesis tests based on the p-values computed with the conditional resampling scheme.
- We experimentally show that our method successfully finds the features related to treatment effect heterogeneity and outperforms the existing mean-based method.

4.2 Background

4.2.1 Problem Setup

Suppose that we have a sample of n individuals $\mathcal{D} = \{(a_i, \boldsymbol{x}_i, y_i)\}_{i=1}^n \overset{i.i.d.}{\sim} \mathrm{P}(A, \boldsymbol{X}, Y)$ for $i = 1, \ldots, n$. Here $A \in \{0, 1\}$ is a binary treatment ($A = 1$ if an individual is treated; otherwise, $A = 0$), $\boldsymbol{X} = [X_1, \ldots, X_d]^\top$ is d-dimensional features (a.k.a. covariates), where each feature $X_m \in \mathcal{X}$ $(m = 1, \ldots, d)$ takes either discrete or continuous values, and $Y \in \mathbb{R}$ is a continuous-valued outcome.[1] Here we assume that (1) features $\boldsymbol{X}$ are measured before applying the treatment and observing outcome Y (i.e., features $\boldsymbol{X}$ are *pretreatment variables* and not *mediators* or *colliders* [Elwert and Winship, 2014]) and that (2) features $\boldsymbol{X}$ contain all *confounders*, i.e., the variables that affect treatment A and outcome Y. Note that these assumptions are standard in the existing work [Imai and Ratkovic, 2013; Zhao *et al.*, 2022].

Given sample $\mathcal{D}$, we solve the problem of selecting the features in $\boldsymbol{X}$ that influence the effect of treatment A on outcome Y. In this problem, which features should be selected depends on the measurement scale of the treatment effect [Hernán and Robins, 2020, Chapter 4]. There are two measurement scales: additive scale $Y^1 - Y^0$ and multiplicative scale Y^1/Y^0, where Y^0 and Y^1 are random variables that are referred to as potential outcomes, each of which represents the outcome when $A = 0$ and when $A = 1$, respectively [Rubin, 1974]. In this study, we define the treatment effect for each individual on an additive scale as $Y^1 - Y^0$ because this scale is standard and widely used in numerous applications [Lee *et al.*, 2018; Schochet *et al.*, 2014; Taddy *et al.*, 2016].

[1] We assume $Y \in \mathbb{R}$ to use the kernel approximation technique [Rahimi and Recht, 2007], which is described in Section 4.3.3.

Unfortunately, we cannot observe treatment effect $Y^1 - Y^0$. This is because we cannot jointly observe two potential outcomes Y^0 and Y^1; we only observe either Y^0 or Y^1, which is obtained as $Y = (1 - A)Y^0 + AY^1$ ($A \in \{0, 1\}$). For this reason, existing methods use the average treatment effect across individuals, which can be estimated from the data.

4.2.2 Mean-based Approaches

Many existing methods [Tian *et al.*, 2014; Zhao *et al.*, 2022] seek the features whose attributes affect the degree of the average treatment effect called CATE, which is defined for each feature's attribute, $X_m = x$ ($m = 1, \ldots, d$), as follows:

$$\begin{aligned} T_m(x) &\coloneqq \mathbb{E}[Y^1 - Y^0 \mid X_m = x] \\ &= \mathbb{E}[Y^1 \mid X_m = x] - \mathbb{E}[Y^0 \mid X_m = x]. \end{aligned} \tag{4.1}$$

CATE $T_m(x)$ is an average treatment effect over the individuals who share an identical attribute, $X_m = x$. Note that this CATE is different from the one conditioned on all features $\boldsymbol{X}$, which is an inference target of the recent causal inference methods [Chang and Dy, 2017; Hassanpour and Greiner, 2019; Hill, 2011; Künzel *et al.*, 2019; Nie and Wager, 2021; Shalit *et al.*, 2017; Yoon *et al.*, 2018].

Using CATE T_m ($m = 1, \ldots, d$), the features that influence the degree of the average treatment effect are defined as the following *treatment effect modifiers*:

Definition 5 (Rothman *et al.* [2008])**.** *Feature X_m is said to be a treatment effect modifier if there are at least two values of X_m, x_m and $x_m^\star$ ($x_m \neq x_m^\star$), such that CATE T_m in Eq. (4.1) takes different values, i.e., $T_m(x_m) \neq T_m(x_m^\star)$.*

Definition 5 states that feature X_m is a treatment effect modifier if CATE $T_m(x)$ is not a constant with respect to value $X_m = x$. Roughly speaking, when we group individuals by their X_m's values and compute the average treatment effect in each group of the individuals, if there are at least two groups with different averages, then feature X_m is a treatment effect modifier [VanderWeele, 2009].

The existing methods seek such treatment effect modifiers by fitting a regression model that is linear in treatment A with a sparse regularizer [Imai and Ratkovic, 2013; Sechidis *et al.*, 2021; Tian *et al.*, 2014; Zhao *et al.*, 2022].

Table 4.1: Joint probability tables of potential outcomes in Example 1. Nonzero probabilities are shown in bold. Total expresses marginal potential outcome probabilities.

$\mathrm{P}(Y^0, Y^1 \mid X = 0)$				
Y^0 \ Y^1	-1	0	1	Total
-1	0	0	0	0
0	**0.5**	0	**0.5**	**1.0**
1	0	0	0	0
Total	**0.5**	0	**0.5**	**1.0**

$\mathrm{P}(Y^0, Y^1 \mid X = 1)$				
Y^0 \ Y^1	-1	0	1	Total
-1	0	0	0	0
0	0	**1.0**	0	**1.0**
1	0	0	0	0
Total	0	**1.0**	0	**1.0**

4.2.3 Weakness of Mean-based Approaches

Since the aforementioned existing methods rely on the average treatment effect, they cannot detect the features whose attributes do not influence the average treatment effect but do affect other functionals of the joint distribution of potential outcomes, such as the covariance between potential outcomes and the treatment effect variance [Russell, 2021]. To illustrate such a feature, consider the following toy example:

Example 1. *Let* $Y^0, Y^1 \in \{-1, 0, 1\} \subset \mathbb{R}$ *be the potential outcomes and let* $X \in \{0, 1\}$ *be a binary feature. Suppose that joint distribution* $\mathrm{P}(Y^0, Y^1 \mid X)$ *is given as Table 4.1. Then feature* X*'s values are irrelevant to the average treatment effect and the covariance between potential outcomes but relevant to the treatment effect variance:*

$$\mathbb{E}[Y^1 - Y^0 \mid X = 0] = \mathbb{E}[Y^1 - Y^0 \mid X = 1] = 0$$
$$\mathrm{Cov}[Y^0, Y^1 \mid X = 0] = \mathrm{Cov}[Y^0, Y^1 \mid X = 1] = 0$$
$$\mathrm{Var}[Y^1 - Y^0 \mid X = 0] = 1; \quad \mathrm{Var}[Y^1 - Y^0 \mid X = 1] = 0.$$

Joint distribution $\mathrm{P}(Y^0, Y^1 \mid X)$ presented in Table 4.1 shows that feature X is related to a difference in treatment effects: While no individual with attribute $X = 1$ receives any treatment effect, those with $X = 0$ get positive or negative effects. However, since the CATE values do not depend on X, the existing mean-based methods will incorrectly conclude that feature X is unrelated to the treatment effect heterogeneity. This implies that using CATE is insufficient to capture such *distributional* treatment effect heterogeneity and might lead to overlooking important features.

4.3 Discovering Distributional Treatment Effect Modifiers

4.3.1 Detecting Distributional Heterogeneity

We propose a feature selection framework for discovering the features related to distributional treatment effect heterogeneity. To find such features, we consider the problem of determining whether the values of each feature X_m $(m = 1, \ldots, d)$ influence the functionals of the joint distribution of potential outcomes $\mathrm{P}(Y^0, Y^1 \mid X_m)$, such as the average treatment effect, the treatment effect variance, and the covariance between potential outcomes. [2] This problem is challenging because we cannot infer joint distribution $\mathrm{P}(Y^0, Y^1 \mid X_m)$ since we can never jointly observe potential outcomes Y^0 and Y^1, as described in Section 4.2.1.

To overcome this challenge, we propose measuring the importance of each feature X_m $(m = 1, \ldots, d)$ by quantifying how greatly X_m's values influence the discrepancy between conditional distributions $\mathrm{P}(Y^0 \mid X_m)$ and $\mathrm{P}(Y^1 \mid X_m)$. This idea is motivated by the following fact: *if the discrepancy between* $\mathrm{P}(Y^0 \mid X_m)$ *and* $\mathrm{P}(Y^1 \mid X_m)$ *varies with* X_m*'s values, then joint distribution* $\mathrm{P}(Y^0, Y^1 \mid X_m)$ *is also changeable depending on* X_m*'s values, and some functionals of the joint distribution depend on* X_m. This fact can be easily proved by taking its contraposition, as shown in Section 4.7.1.

Such an idea enables us to detect feature X in Example 1, whose values influence the treatment effect variance. This is because, in this example, the discrepancy between conditional potential outcome distributions $\mathrm{P}(Y^0 \mid X)$ and $\mathrm{P}(Y^1 \mid X)$ changes depending on X's values.

Note, however, that our idea does not always work well. This is because there are counterexamples where feature X_m's values do not affect the discrepancy between conditional distributions $\mathrm{P}(Y^0 \mid X_m)$ and $\mathrm{P}(Y^1 \mid X_m)$ but influence joint distribution $\mathrm{P}(Y^0, Y^1 \mid X_m)$. We take a simple counterexample in Section 4.7.2 and present the

[2] Identifying which functionals of the joint distribution are affected by each feature's values is extremely challenging due to the impossibility of inferring the joint distribution. One possible solution is to use techniques for estimating the lower and upper bounds on the functionals [Chen *et al.*, 2016; Russell, 2021; Shingaki and Kuroki, 2021]. Although such bounds require several additional assumptions, they have been successfully applied in several fields, including fairness-aware machine learning [Chikahara *et al.*, 2021].

empirical performances in such cases in Section 4.5.4. Nevertheless, compared with the existing methods, we can detect a wider variety of features relevant to treatment effect heterogeneity, which leads to a better understanding of the underlying causal mechanisms.

4.3.2 Feature Importance Measure

To express the importance of each feature X_m $(m = 1, \ldots, d)$, we measure the discrepancy between distributions $\mathrm{P}(Y^0 \mid X_m)$ and $\mathrm{P}(Y^1 \mid X_m)$ using the MMD [Gretton *et al.*, 2012].

In fact, there are several MMD-based metrics for measuring the discrepancy between potential outcome distributions [Bellot and van der Schaar, 2021; Muandet *et al.*, 2021; Park *et al.*, 2021]. However, these metrics cannot be applied in our setting because they are not designed for the conditional distributions conditioned on a single feature; we give details of this reason in Section 4.4.2.

Consequently, we develop an MMD-based metric for conditional distributions $\mathrm{P}(Y^0 \mid X_m)$ and $\mathrm{P}(Y^1 \mid X_m)$. Let $k_Y \colon \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ be a positive-definite kernel function. Then the squared MMD between the conditional distributions conditioned on a feature value, $X_m = x$, is defined as

$$\begin{aligned} &D_m^2(x) \coloneqq \mathrm{MMD}^2(\mathrm{P}(Y^0 \mid X_m = x), \mathrm{P}(Y^1 \mid X_m = x)) \\ =&\,\mathbb{E}_{Y^0, Y^{0\prime} \mid X_m = X_m' = x}[k_Y(Y^0, Y^{0\prime})] + \mathbb{E}_{Y^1, Y^{1\prime} \mid X_m = X_m' = x}[k_Y(Y^1, Y^{1\prime})] \\ &- 2\,\mathbb{E}_{Y^0, Y^1 \mid X_m = x}[k_Y(Y^0, Y^1)], \end{aligned} \tag{4.2}$$

where superscript prime $'$ denotes an independent copy of each random variable, and expectation $\mathbb{E}_{Y^0, Y^{0\prime} \mid X_m = X_m' = x}$ is taken with respect to $\mathrm{P}(Y^0, Y^{0\prime} \mid X_m = X_m' = x)$; other expectations are taken in a similar manner. This metric has the following property: If k_Y belongs to the class of kernel functions called *characteristic kernels* [Gretton *et al.*, 2012], then squared MMD is $D_m^2(x) = 0$ if and only if $\mathrm{P}(Y^0 \mid X_m = x) = \mathrm{P}(Y^1 \mid X_m = x)$. Examples of characteristic kernels include the Gaussian kernel.

Based on squared MMD D_m^2, we define the features related to distributional treatment effect heterogeneity as the following *distributional treatment effect modifiers*:

Definition 6. *Feature X_m is said to be a distributional treatment effect modifier if there are at least two values of X_m, x_m and $x_m^\star$ ($x_m \neq x_m^\star$), such that squared MMD D_m^2 in Eq. (4.2) takes different values, i.e., $D_m^2(x_m) \neq D_m^2(x_m^\star)$.*

In other words, feature X_m is a distributional treatment effect modifier if the squared MMD between $\mathrm{P}(Y^0 \mid X_m)$ and $\mathrm{P}(Y^1 \mid X_m)$ varies depending on X_m's values.

To detect such a variation, we formulate the importance of each feature X_m as the variance of the squared MMD:

$$I_m \coloneqq \mathrm{Var}[D_m^2(X_m)]. \tag{4.3}$$

4.3.3 Estimator of Feature Importance

To estimate feature importance measure I_m in Eq. (4.3), we need to compute the expected values in Eq. (4.2) whose expectations can be represented as the ones over conditional distributions $\mathrm{P}(Y^0 \mid X_m = x)$ and $\mathrm{P}(Y^1 \mid X_m = x)$.

However, we cannot directly compute them because we have no access to the observations from these conditional distributions. To overcome this difficulty, we develop a weighted estimator that can be computed from the observed data.

Weighted Conditional MMD (WCMMD)

To infer squared MMD $D_m^2(x)$ in Eq. (4.2), we develop an estimator of the expected value over conditional distribution $\mathrm{P}(Y^a \mid X_m = x)$ ($a \in \{0, 1\}$) using a weighting-based estimation technique called importance sampling.

To derive such an estimator, we use weight functions called inverse probability weights [Rosenbaum and Rubin, 1983]:

$$w^0(A, \boldsymbol{X}) = \frac{\mathbf{I}(A = 0)}{1 - \mathrm{e}(\boldsymbol{X})}, \quad w^1(A, \boldsymbol{X}) = \frac{\mathbf{I}(A = 1)}{\mathrm{e}(\boldsymbol{X})}, \tag{4.4}$$

where $\mathrm{e}(\boldsymbol{X}) \coloneqq \mathrm{P}(A = 1 \mid \boldsymbol{X})$ is the conditional distribution called a *propensity score*, and $\mathbf{I}(A = a)$ is an indicator function that takes 1 if $A = a$; otherwise 0. In addition, we make the two standard assumptions: *positivity*, which imposes support condition $0 < \mathrm{e}(\boldsymbol{x}) < 1$ for all $\boldsymbol{x}$ [Rosenbaum and Rubin, 1983], and *conditional*

ignorability (a.k.a. *strong ignorability*), which requires conditional independence relation $\{Y^0, Y^1\} \perp\!\!\!\perp A \mid \boldsymbol{X}$; this relation is satisfied if features $\boldsymbol{X}$ are pretreatment variables, contain no mediator or collider, and include all confounders (See e.g., Elwert and Winship [2014] for the details).

Under these assumptions, for instance, expected value $\mathbb{E}_{Y^1|X_m=x}[Y^1]$ can be reformulated as

$$\begin{aligned}
&\mathbb{E}_{Y^1|X_m=x}[Y^1] \\
=&\mathbb{E}_{\boldsymbol{X}_{-m}|X_m=x}[\mathbb{E}_{Y^1|\boldsymbol{X}_{-m},X_m=x}[Y^1]] \\
=&\mathbb{E}_{\boldsymbol{X}_{-m}|X_m=x,A=1}\left[\mathbb{E}_{Y|\boldsymbol{X}_{-m},X_m=x,A=1}\left[\frac{\mathrm{P}(A=1)}{\mathrm{P}(A=1 \mid \boldsymbol{X})}Y\right]\right] \\
=&\mathbb{E}_{A,\boldsymbol{X}_{-m},Y|X_m=x}[w^1(A,\boldsymbol{X})Y],
\end{aligned}$$

where $\boldsymbol{X}_{-m} \coloneqq \boldsymbol{X} \backslash X_m$ denotes the features with X_m removed.

To estimate squared MMD $D_m^2(x)$ in Eq. (4.2) in the same way, we formulate the following estimator, which we call a *weighted conditional MMD* (WCMMD):

$$\begin{aligned}
&\mathrm{WCMMD}^2_{X_m=x} \\
\coloneqq&\mathbb{E}_{A,A',\boldsymbol{X}_{-m},\boldsymbol{X}'_{-m},Y,Y'|X_m=X'_m=x}[w^0(A,\boldsymbol{X})w^0(A',\boldsymbol{X}')k_Y(Y,Y')] \\
+&\mathbb{E}_{A,A',\boldsymbol{X}_{-m},\boldsymbol{X}'_{-m},Y,Y'|X_m=X'_m=x}[w^1(A,\boldsymbol{X})w^1(A',\boldsymbol{X}')k_Y(Y,Y')] \\
-2&\,\mathbb{E}_{A,A',\boldsymbol{X}_{-m},\boldsymbol{X}'_{-m},Y,Y'|X_m=X'_m=x}[w^0(A,\boldsymbol{X})w^1(A',\boldsymbol{X}')k_Y(Y,Y')]. \qquad (4.5)
\end{aligned}$$

We can show that this WCMMD equals $D_m^2(x)$ under conditional ignorability and positivity assumptions:

Proposition 1. *Suppose that conditional ignorability and positivity hold. Then* $D_m^2(x) = \mathit{WCMMD}^2_{X_m=x}$.

See Section 4.7.3 for the proof. Hence, WCMMD has the same property with squared MMD $D_m^2(x)$: If k_Y is a characteristic kernel, $\mathrm{WCMMD}^2_{X_m=x} = 0$ if and only if $\mathrm{P}(Y^0 \mid x) = \mathrm{P}(Y^1 \mid x)$.

Empirical Estimator of WCMMD

To infer MMD $D_m^2(x)$ with Eq. (4.5), we estimate the conditional expected values conditioned on $X_m = x$ using sample $\mathcal{D} = \{(a_i, \boldsymbol{x}_i, y_i)\}_{i=1}^n \overset{i.i.d.}{\sim} \mathrm{P}(A, \boldsymbol{X}, Y)$.

If feature X_m takes discrete values, we only have to take the averages over the individuals with $X_m = x$. Formally, by letting $\omega_i^{a,x}$ for $i = 1, \ldots, n$ and $a \in \{0, 1\}$ be

$$\omega_i^{a,x} = \frac{\mathbf{I}(x_{m,i} = x)}{\sum_{l=1}^{n} \mathbf{I}(x_{m,l} = x)} w^a(a_i, \boldsymbol{x}_i), \tag{4.6}$$

we can estimate the expected values in Eq. (4.5) by

$$\widehat{D}_m^2(x) \coloneqq \sum_{i=1}^{n}\sum_{j=1}^{n} \left(\omega_i^{0,x}\omega_j^{0,x} + \omega_i^{1,x}\omega_j^{1,x}\right) k_Y(y_i, y_j) - 2\sum_{i=1}^{n}\sum_{j=1}^{n} \omega_i^{0,x}\omega_j^{1,x} k_Y(y_i, y_j). \tag{4.7}$$

For continuous-valued feature X_m, we smoothen indicator function $\mathbf{I}$ in Eq. (4.6) by applying the kernel smoothing technique [Nadaraya, 1964; Watson, 1964] as

$$\omega_i^{a,x} = \frac{\frac{1}{h_{X_m}} k_{X_m}(x_{m,i}, x)}{\sum_{l=1}^{n} \frac{1}{h_{X_m}} k_{X_m}(x_{m,l}, x)} w^a(a_i, \boldsymbol{x}_i), \tag{4.8}$$

where the similarity between X_m's values is measured by kernel function k_{X_m} with bandwidth h_{X_m}; in our experiments, we formulate k_{X_m} as the Gaussian kernel:

$$k_{X_m}(x_m, x_m^\star) = \exp\left(-\frac{\|x_m - x_m^\star\|^2}{h_{X_m}^2}\right).$$

In both cases where $\omega_i^{a,x}$ is given as Eqs. (4.6) and (4.8), we can show the consistency of estimator $\widehat{D}_m^2(x)$ in the limit of infinite sample size:

Theorem 1. *Suppose that weight $\omega_i^{a,x}$ is given as (4.6) or (4.8). Then under the assumptions presented in Section 4.7.4, we have $\widehat{D}_m^2(x) \xrightarrow{p} D_m^2(x)$ as $n \to \infty$.*

See Section 4.7.4 for the proof. In practice, we need to estimate $\omega_i^{a,x}$ by inferring propensity score $\mathrm{e}(\boldsymbol{X}) \coloneqq \mathrm{P}(A = 1 \mid \boldsymbol{X})$ using a regression model, such as logistic regression and neural network.

A drawback of estimator $\widehat{D}_m^2(x)$ in Eq. (4.7) is that it needs computation time $O(n^2)$ for sample size n, implying that estimating $D_m^2(x)$ for each $x = x_{m,1}, \ldots, x_{m,n}$ requires $O(n^3)$, which is impractical for large n. To resolve this issue, in what follows, we develop a computationally efficient variant of $\widehat{D}_m^2(x)$.

Computationally Efficient Empirical Estimator

To reduce the time of computing the squared MMD estimator $\widehat{D}_m^2(x)$ in Eq. (4.7), we use a well-known kernel approximation technique called random Fourier features (RFFs) [Rahimi and Recht, 2007].

With RFFs, we approximate kernel function $k_Y(y_i, y_j)$ in Eq. (4.7) as an inner product of two feature vectors:

$$k_Y(y_i, y_j) \approx \widetilde{k}_Y(y_i, y_j) = \langle \boldsymbol{z}(y_i), \boldsymbol{z}(y_j) \rangle_{\mathbb{R}^r}, \tag{4.9}$$

where $\boldsymbol{z}\colon \mathbb{R} \to \mathbb{R}^r$ is a mapping that outputs a vector of the r features, where r is a hyperparameter. These r features are randomly sampled from the Fourier transform of kernel function k_Y. We formulate k_Y as a Gaussian kernel with bandwidth h_Y; in this case, feature mapping $\boldsymbol{z}$ is given as $\boldsymbol{z}(y) = [\sqrt{2}\cos(\lambda_1 y + \zeta_1), \ldots, \sqrt{2}\cos(\lambda_r y + \zeta_r)]^\top$, where $\lambda_1, \ldots, \lambda_r$ are drawn from Gaussian distribution $\mathcal{N}(0, 2h_Y)$, and $\zeta_1, \ldots, \zeta_r$ are sampled from uniform distribution $\mathrm{Unif}(0, 2\pi)$ [Rahimi and Recht, 2007].

Based on (4.9), we approximate estimator $\widehat{D}_m^2(x)$ in Eq. (4.7) as

$$\widetilde{D}_m^2(x) \coloneqq \langle \widetilde{\mu}_{Y^0|x}, \widetilde{\mu}_{Y^0|x} \rangle_{\mathbb{R}^r} + \langle \widetilde{\mu}_{Y^1|x}, \widetilde{\mu}_{Y^1|x} \rangle_{\mathbb{R}^r} - 2\langle \widetilde{\mu}_{Y^0|x}, \widetilde{\mu}_{Y^1|x} \rangle_{\mathbb{R}^r} \tag{4.10}$$

where $\widetilde{\mu}_{Y^0|x}$ and $\widetilde{\mu}_{Y^1|x}$ are the following weighted averages of the r-dimensional random feature vector:

$$\widetilde{\mu}_{Y^0|x} = \sum_{i=1}^{n} \omega_i^{0,x} \boldsymbol{z}(y_i); \;\; \widetilde{\mu}_{Y^1|x} = \sum_{i=1}^{n} \omega_i^{1,x} \boldsymbol{z}(y_i).$$

Using (4.10), we estimate our feature importance measure as

$$\widetilde{I}_m = \frac{1}{n-1} \sum_{\iota=1}^{n} \left(\widetilde{D}_m^2(x_{m,\iota}) - \frac{1}{n} \sum_{\varsigma=1}^{n} \widetilde{D}_m^2(x_{m,\varsigma}) \right)^2. \tag{4.11}$$

Computing this estimator requires $O(rn^2)$, which is feasible by setting hyperparameter r to be a moderate value.

4.3.4 Feature Selection with Conditional Randomization Test (CRT)

Using estimated importance measures $\widetilde{I}_1, \ldots, \widetilde{I}_d$, we select distributional treatment effect modifiers. To achieve this, we perform multiple hypothesis tests where for each $m = 1, \ldots, d$, we consider the following null and alternative hypotheses:

$$\mathcal{H}_{0,\mathrm{m}} \colon I_m = 0 \quad \text{and} \quad \mathcal{H}_{1,\mathrm{m}} \colon I_m > 0. \tag{4.12}$$

To decide whether to reject each null hypothesis $\mathcal{H}_{0,\mathrm{m}}$, we compute p-value p_m, i.e., the probability of obtaining test statistic I_m such that $I_m \geq \widetilde{I}_m$ under null hypothesis $\mathcal{H}_{0,\mathrm{m}}$. Evaluating this p-value requires the distribution of test statistic I_m under $\mathcal{H}_{0,\mathrm{m}}$. However, analytically deriving this distribution is extremely difficult because the asymptotic distributions of data-dependent weights $\omega_i^{0,x}$ and $\omega_i^{1,x}$ in feature importance measure $\widetilde{I}_m$ are unclear.

For this reason, we approximate the distribution of the test statistic under null hypothesis $\mathcal{H}_{0,\mathrm{m}}$, where feature X_m is irrelevant to treatment effect heterogeneity. To this end, we simulate such an irrelevant feature for each X_m without changing joint distribution $\mathrm{P}(\boldsymbol{X})$ so that the joint distribution of this synthetically generated dummy feature and other observed features $\boldsymbol{X}_{-m} \coloneqq \boldsymbol{X} \backslash X_m$ is equal to the original joint distribution, $\mathrm{P}(\boldsymbol{X})$. To achieve this, following the resampling scheme called *conditional randomization test* (CRT) [Candes *et al.*, 2018, Section F], we sample new X_m's values from the conditional distribution, $\mathrm{P}(X_m \mid \boldsymbol{X}_{-m})$, without looking at the values of treatment A and outcome Y.

Our CRT proceeds as illustrated in Algorithm 1. We first estimate conditional distribution $\mathrm{P}(X_m \mid \boldsymbol{X}_{-m})$ by fitting a generative model $\mathcal{L}$ to the data; in our experiments, we employ a widely-used deep generative model called the conditional variational autoencoder (CVAE) [Sohn *et al.*, 2015]. Then, using fitted generative model $\mathcal{L}$, we prepare B datasets, each of which contains different values of the synthetic dummy features drawn from $\mathcal{L}$. In particular, for each $b = 1, \ldots, B$, we repeat the two steps: sampling n values of feature X_m as $x_{m,i}^{(b)} \sim \mathcal{L}(X_m \mid \boldsymbol{x}_{-m,i})$ $(i = 1, \ldots, n)$ and using these values to compute test statistic $\widetilde{I}_m^{(b)}$. By repeating these steps, we

Algorithm 1 Conditional Randomization Test (CRT)

Input: sample $\mathcal{D} = \{(a_i, \boldsymbol{x}_i, y_i)\}_{i=1}^n$, estimated statistic $\widetilde{I}_m$
Output: p-value $\hat{p}_m$

Fit generative model $\mathcal{L}$ to sample $\mathcal{D}$.
for $b = 1, \ldots, B$ **do**
  for $i = 1, \ldots, n$ **do**
    Draw $x_{m,i}^{(b)} \sim \mathcal{L}(X_m \mid \boldsymbol{x}_{-m,i})$.
    $\boldsymbol{x}_i^{(b)} \leftarrow x_{m,i}^{(b)} \cup \boldsymbol{x}_{-m,i}$
  end for
  Compute test statistic $\widetilde{I}_m^{(b)}$ using $\{(a_i, \boldsymbol{x}_i^{(b)}, y_i)\}_{i=1}^n$.
end for
Compute p-value $\hat{p}_m$ by Eq. (4.13).
return $\hat{p}_m$

Algorithm 2 Proposed feature selection framework

Input: sample $\mathcal{D} = \{(a_i, \boldsymbol{x}_i, y_i)\}_{i=1}^n$, significance level α
Output: feature index set $\hat{S} \subseteq \{1, \ldots, d\}$

for $m = 1, \ldots, d$ **do**
  Compute test statistic $\widetilde{I}_m$ with sample $\mathcal{D}$.
  Compute p-value as $\hat{p}_m \leftarrow \mathrm{CRT}(\mathcal{D}, \widetilde{I}_m)$.
end for
Adjust p-values as $\hat{p}_1^*, \ldots, \hat{p}_d^*$ using a multiple testing procedure.
Select feature index set as $\hat{S} = \{m \colon \hat{p}_m^* \leq \alpha\}$.
return $\hat{S}$

obtain an empirical distribution of the test statistic and compute a p-value as

$$\hat{p}_m = \frac{1}{B} \sum_{b=1}^{B} \mathbf{I}\left(\widetilde{I}_m^{(b)} \geq \widetilde{I}_m\right). \tag{4.13}$$

After computing p-values $\hat{p}_1, \ldots, \hat{p}_d$, we perform multiple hypothesis tests. Since the chance of obtaining false positives increases with the number of hypotheses tested, we control such false positives by adjusting the p-values; we used the Benjamini-Hochber (BH) adjustment procedure [Benjamini and Hochberg, 1995] in our experiments. We summarize our feature selection framework in Algorithm 2.

One of the advantages of applying CRT is that if the fitted generative model equals the true conditional distribution (i.e., $\mathcal{L}(X_m \mid \boldsymbol{X}_{-m}) = \mathrm{P}(X_m \mid \boldsymbol{X}_{-m})$ for all $m = 1. \ldots, d$), it can precisely control the the type I error rate to be at most significance level α [Candes *et al.*, 2018, Section F]. Although learning such gener-

ative models is difficult, we experimentally confirmed that our method successfully controlled the type I error rate to be close to α (Section 4.5.2).

As a disadvantage, performing CRT is computationally expensive: It requires computing test statistic B times for each feature. Although this computation is embarrassingly parallelizable, it needs $O(Bdrn^2)$ in total, even with our computationally efficient estimator of the test statistic. Our future work will investigate how to further reduce the computation time; for instance, the CRT's computationally efficient variants (e.g., Liu *et al.* [2021]) might be helpful.

4.4 Related Work

4.4.1 Interpreting Treatment Effect Heterogeneity

A growing number of causal inference methods have been developed to accurately estimate heterogeneous treatment effects using neural networks [Johansson *et al.*, 2016; Shalit *et al.*, 2017; Yoon *et al.*, 2018], tree-based models [Hahn *et al.*, 2020; Hill, 2011], and machine learning frameworks called meta-learners [Künzel *et al.*, 2019; Nie and Wager, 2021].

However, few are designed to elucidate a causal mechanism that yields the treatment effect heterogeneity. The Causal Rule Ensemble method [Lee *et al.*, 2020] seeks the important features by learning a rule-based model that emulates the input-output relationship of a fitted treatment effect estimation model. Gilad *et al.* [2021] considered a hypothesis test for discovering the treatment effect modifiers from social network data. However, none of these methods can find the features related to distributional treatment effect heterogeneity because they are also based on the average treatment effect and cannot find the features related to other functionals of the joint distribution of potential outcomes.

To overcome this limitation of the existing mean-based methods, we established a feature selection framework for discovering the important features related to the functionals of the joint distribution of potential outcomes.

4.4.2 MMD between Potential Outcome Distributions

To find distributional treatment effect modifiers, we formulated a weighted estimator of the MMD that measures the discrepancy between conditional potential outcome distributions.

Our estimator has a clear advantage in that it can consistently estimate the MMD between the conditional distributions conditioned on a single feature, $\mathrm{P}(Y^0 \mid X_m)$ and $\mathrm{P}(Y^1 \mid X_m)$ $(m = 1, \ldots, d)$, by addressing the confounders in features $\boldsymbol{X}$.

The existing estimators cannot consistently estimate such an MMD. The kernel treatment effect (KTE) [Muandet *et al.*, 2021] and the weighted MMD (WMMD) [Bellot and van der Schaar, 2021] are designed to quantify the discrepancy between marginal distributions $\mathrm{P}(Y^0)$ and $\mathrm{P}(Y^1)$; hence they cannot address the conditional distributions. Although the conditional distributional treatment effect (CoDiTE) [Park *et al.*, 2021] measures the MMD between conditional distributions $\mathrm{P}(Y^0 \mid \boldsymbol{X})$ and $\mathrm{P}(Y^1 \mid \boldsymbol{X})$, we cannot naively apply it by considering the setting where features $\boldsymbol{X}$ only contain a single feature (i.e., $\boldsymbol{X} = \{X_m\}$). This is because this measure only addresses the confounders that are included in the conditioning variables, and if setting $\boldsymbol{X} = \{X_m\}$, we cannot eliminate the influence of the confounders in $\boldsymbol{X}_{-m}$.

To consistently estimate the MMD between conditional distributions $\mathrm{P}(Y^0 \mid X_m)$ and $\mathrm{P}(Y^1 \mid X_m)$, we derived an IPW-based estimator by regarding the MMD as a function of features $\boldsymbol{X}$ and then averaging out unwanted features $\boldsymbol{X}_{-m}$ (by taking an integral with respect to $\mathrm{P}(\boldsymbol{X}_{-m} \mid X_m)$).

4.5 Experiments

4.5.1 Setup

We compared our proposed framework with the following two baselines: (1) the existing mean-based method called the selective inference method for effect modification (SI-EM) [Zhao *et al.*, 2022] and (2) a naive variant of our method (Naive), which samples the values of a synthetic dummy feature corresponding to X_m $(m = 1, \ldots, d)$ from (empirical) marginal distribution $\mathrm{P}(X_m)$.

We ran all methods with significance level $\alpha = 0.05$. As regards our method and Naive, we set the number of RFFs to $r = 1000$, selected the values of kernel

bandwidths $h_{X_1}, \ldots, h_{X_d}$ and h_Y using a well-known heuristic called median heuristic [Schölkopf *et al.*, 2002], and inferred propensity score $\mathrm{e}(\boldsymbol{X})$ by fitting a feed-forward neural network that contains two linear layers with 50 neurons and Rectified Linear Unit (ReLU) activation functions. With our method, we performed a CRT by setting the number of resampled datasets to $B = 100$. Here we formulated generative model $\mathcal{L}(X_m \mid \boldsymbol{X}_{-m})$ for each $m = 1, \ldots, d$ as a CVAE whose encoders and decoders are given as the feed-forward neural networks that contain two linear layers with 128 neurons and ReLU functions. We confirmed that the number of neurons did not greatly affect the performance in Section 4.5.4.

4.5.2 Synthetic Data Experiments

Data: We prepared synthetic datasets as follows. We drew treatment A from the Bernoulli distribution and features $\boldsymbol{X} = [X_1, \ldots, X_d]^\top$ $(d = 30)$ from the Gaussian distributions:

$$
\begin{aligned}
&A \sim \mathrm{Ber}(0.5), \\
&\boldsymbol{X} \mid A = 0 \sim \mathcal{N}(-\mu, \boldsymbol{\Sigma}), \quad \text{and} \quad \boldsymbol{X} \mid A = 1 \sim \mathcal{N}(\mu, \boldsymbol{\Sigma}),
\end{aligned}
$$

where Ber and $\mathcal{N}$ denote the Bernoulli and Gaussian distributions, respectively, $\mu = [0.2, \ldots, 0.2]^\top$ is a d-dimensional vector, and $\boldsymbol{\Sigma}$ is a $d \times d$ covariance matrix whose (i, j)-th element is $\Sigma_{i,j} = \sigma^{|i-j|}$ $(\sigma = 0.2)$ for each $i, j \in \{1, \ldots, d\}$. We sampled outcome $Y = (1 - A)Y^0 + AY^1$ by generating potential outcomes Y^0 and Y^1 with the following four generation processes where five features $X_1, \ldots, X_5$ are distributional treatment effect modifiers:

- **LinMean**:

$$
Y^0 \sim \mathcal{N}(-f(X_1, \ldots, X_5), 1); Y^1 \sim \mathcal{N}(f(X_1, \ldots, X_5), 1),
$$

- **NonlinMean**:

$$
Y^0 \sim \mathcal{N}(-g(X_1, \ldots, X_5), 1); Y^1 \sim \mathcal{N}(g(X_1, \ldots, X_5), 1),
$$

- **LinVar**:

$$Y^0 \sim \mathcal{N}(-5, 1); Y^1 \sim \mathcal{N}(0, h(f(X_1, \ldots, X_5))^2),$$

- **NonlinVar**:

$$Y^0 \sim \mathcal{N}(-5, 1); Y^1 \sim \mathcal{N}(0, h(g(X_1, \ldots, X_5))^2),$$

where f, g and h are the following functions:

$$\begin{aligned} f(X_1, \ldots, X_5) &= 4X_1 + 2X_2 + X_3 + 2X_4 + 4X_5, \\ g(X_1, \ldots, X_5) &= \sum_{j=1}^{5} (X_j - 0.5)^3 + 3 \sum_{j=1}^{5} X_j - 6, \\ h(v) &= \max(v, 1). \end{aligned}$$

Under LinMean and NonlinMean, features $X_1, \ldots, X_5$ affect the average treatment effect while under LinVar and NonlinVar, they influence the treatment effect variance.

Results: Using these synthetic datasets, we evaluated the performance of each method. We computed a true positive rate (TPR) and a false positive rate (FPR), defined as $\frac{d_{\mathrm{TP}}}{d_{\mathrm{T}}}$ and $\frac{d_{\mathrm{FP}}}{d - d_{\mathrm{T}}}$, where $d_{\mathrm{T}} = 5$ is the number of truly relevant features, and d_{TP} and d_{FP} are the number of truly relevant features that are correctly selected as such and the number of irrelevant features that are wrongly selected as the relevant ones, respectively. For each method, we performed 50 experiments with different synthetic datasets generated with different random numbers and computed the average and the standard deviation of TPRs and FPRs over 50 runs.

Figure 4.1 presents the results on the LinMean, NonlinMean, LinVar and NonlinVar datasets. With all of them, our method successfully achieved high TPRs while controlling FPRs to be close to $\alpha = 0.05$. Although SI-EM yielded high TPRs with the LinMean and NonlinMean datasets, since this method is not designed to detect the features related to treatment effect variance, it failed to find important features from the LinVar and NonlinVar datasets. With Naive, not only the TPRs but also the FPRs were higher than our method (especially with the LinMean and LinVar datasets), indicating that it selected many features; however, many of these were false positives, which is problematic in practice.

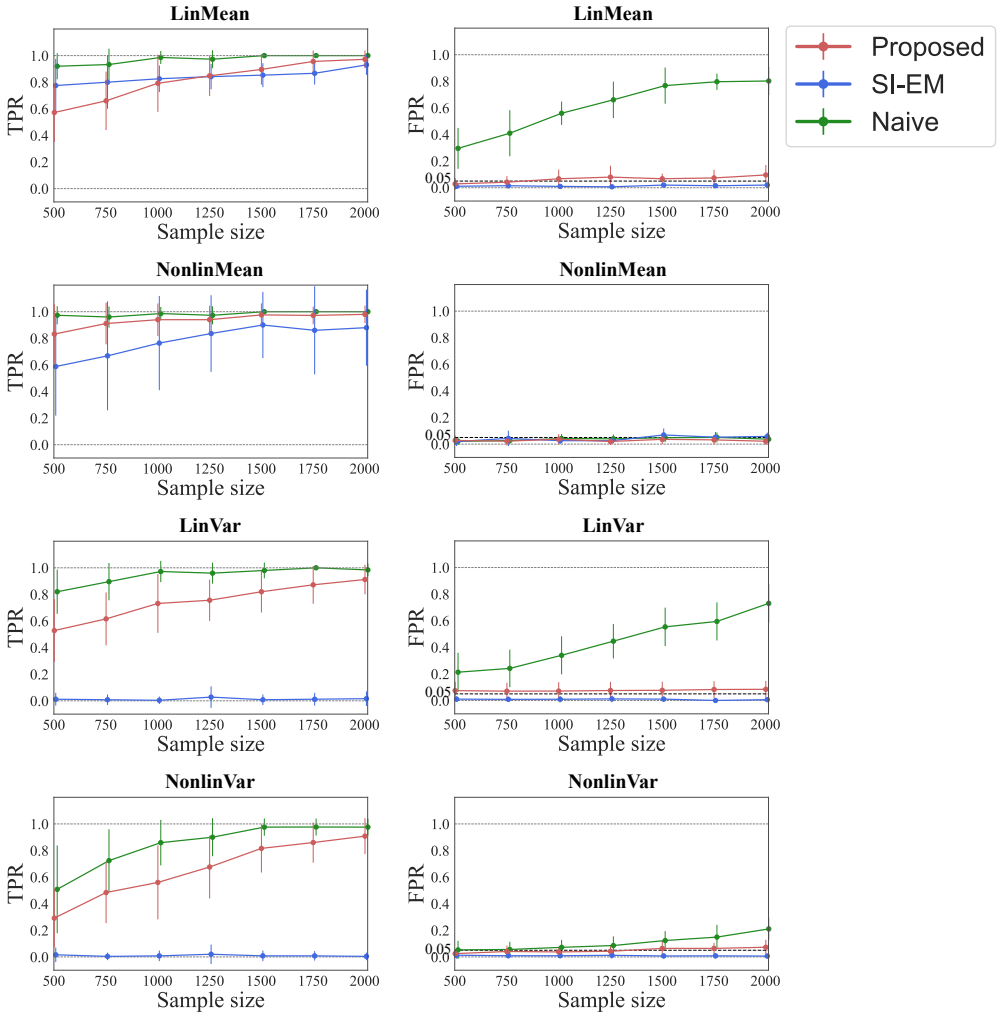

Figure 4.1: TPRs (left) and FPRs (right) of each method on synthetic data with sample sizes $n = 500, 750, \ldots, 2000$. Mean and standard deviation (error bars) over 50 runs with different datasets are shown.

To further illustrate the difference between our method and Naive, consider how each method approximates the p-value of each feature X_m $(m = 1, \ldots, d)$. Both methods compute the p-value by sampling a synthetic dummy feature that is irrelevant to treatment effect heterogeneity; however, its sampling distribution is different. While our method samples it from (estimated) conditional distribution $\mathrm{P}(X_m \mid \boldsymbol{X}_{-m})$ in the CRT, Naive employs (empirical) marginal distribution $\mathrm{P}(X_m)$ without looking at the values of features $\boldsymbol{X}_{-m}$. The latter generation process *unnecessarily* changes joint distribution $\mathrm{P}(\boldsymbol{X})$: The joint distribution of a synthetic feature and observed features $\boldsymbol{X}_{-m}$ is greatly different from that of the original features $\boldsymbol{X}$; this difference

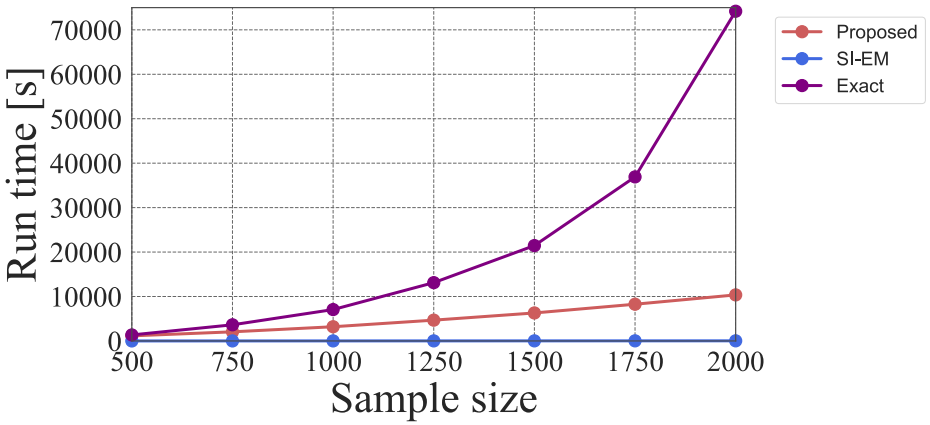

Figure 4.2: Run time comparison among proposed method (red), SI-EM (blue), and Exact (purple) on LinMean dataset with sample sizes $n = 500, 750, \ldots, 2000$

is much larger than with our method. Due to such a large change in $\mathrm{P}(\boldsymbol{X})$, Naive failed to approximate the test statistic's distribution and yielded high FPRs. By contrast, by avoiding greatly changing joint distribution $\mathrm{P}(\boldsymbol{X})$ with the CRT, our method effectively evaluated the statistical significance of each feature.

Meanwhile, the use of the CRT requires considerable computation time, as discussed in Section 4.3.4. To confirm this, we compared the run time of our method with two baselines: SI-EM and the variant of our method (Exact), which computes the feature importance measure by Eq. (4.7) without any approximation. Regarding our method and Exact, we evaluated the total run time, including the training time of the propensity score model and the CVAE. We ran all methods on a 64-bit CentOS machine with 2.10 GHz Xeon Gold 6130 (x2) CPUs and 256-GB RAM.

Figure 4.2 shows the run time on the LinMean dataset with sample sizes $n = 500, 750, \ldots, 2000$. When $n = 2000$, SI-EM and our method required 27 and $10,360$ seconds, respectively, thus exhibiting a notable difference. However, our method needed far less time than Exact, demonstrating the effectiveness of kernel approximation with RFFs.

In summary, these results show the following findings:

- Our method poses a computational challenge; however, it successfully found the features related to average treatment effect and treatment effect variance.
- SI-EM does not need much time; however, it failed to detect the features related to treatment effect variance.

Table 4.2: p-values of features selected by our method from NHANES dataset: Mean and standard deviation are shown for all features with mean p-values less than $\alpha = 0.05$.

Feature	Adjusted p-value
Age	0.0075 ± 0.0305
Gender	0.0046 ± 0.0269
Number of cigarettes smoked	0.0 ± 0.0

Thus, our proposed feature selection framework has made a significant step toward discovering the features related to distributional treatment effect heterogeneity, which, to the best of our knowledge, is the first attempt in causal inference studies. A further reduction of computation time is left as our future work, as described in Section 4.3.4.

4.5.3 Real-World Data Experiments

Data: We used the health records from the National Health and Nutrition Examination Survey (NHANES).[3] Following Zhao *et al.* [2022], we collected the records of $n = 9677$ individuals. Each record contains $d = 20$ features, such as age, gender, race, income, and past medical history (e.g., asthma, gout, stroke, and heart disease); 3 of them take continuous values, and the others are discrete.

With this dataset, we investigated which features modify the effects of obesity on low-grade systemic inflammation by regarding whether body mass index (BMI) exceeds 25 as treatment A and serum C-reactive protein (CRP) level as outcome Y. Discovering such features has important medical implications because low-grade inflammation increases the risk of various chronic diseases, such as cancers and cardiovascular disease [Rodríguez-Hernández *et al.*, 2013].

Since the truly relevant features are unknown, we cannot evaluate the TPRs and FPRs. For this reason, we compared the features selected by our method and SIEM. To take into account the randomness of the CRT, we computed the mean of the adjusted p-values over 50 runs and used this mean p-value to select the features.

Results: Table 4.2 presents the adjusted p-values for all features that are selected by our proposed method.

[3] https://wwwn.cdc.gov/nchs/nhanes/

Both our method and SI-EM successfully selected age and gender, which were reported as important in the previous medical studies [Visser *et al.*, 1999]. Although SI-EM selected only these two features, our method concluded that the number of cigarettes smoked is also statistically significant. Selecting this feature is interesting and seems reasonable because the synergistic effect of obesity and smoking on systemic inflammation has been reported in previous studies [Ólafsdóttir *et al.*, 2005].

4.5.4 Additional Experimental Results

In what follows, we present several additional synthetic data experiments to further evaluate the performance of our method. Section 4.5.4 shows the performance on the data where the truly relevant features do not affect the discrepancy between marginal potential outcome distributions, which is our inference target. Section 4.5.4 displays the results when using different neural network architectures in the models of propensity score and CVAE.

Examining Counterexamples

This section presents the performance of our method on the synthetic data where the features do not influence the discrepancy between conditional distributions $\mathrm{P}(Y^0 \mid X_m)$ and $\mathrm{P}(Y^1 \mid X_m)$ but affect joint distribution $\mathrm{P}(Y^0, Y^1 \mid X_m)$. With such data, our method does not work well because it relies on the discrepancy between $\mathrm{P}(Y^0 \mid X_m)$ and $\mathrm{P}(Y^1 \mid X_m)$, as described in Section 4.3.1.

To evaluate the performance, we prepared synthetic data in a similar manner to Section 4.5.2, which only differs in the generation process of potential outcomes Y^0 and Y^1. Here we set the sample size to $n = 2000$ and sampled the values of Y^0 and Y^1 from the following 2-dimensional Gaussian distributions:

- **LinCovar**:

$$\begin{bmatrix} Y^0 \\ Y^1 \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} -5 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 1 - \frac{1}{h(f(X_1,\ldots,X_5))} \\ 1 - \frac{1}{h(f(X_1,\ldots,X_5))} & 1 \end{bmatrix} \right), \tag{4.14}$$

Table 4.3: TPRs and FPRs of our method on LinCovar and NonlinCovar datasets. Mean and standard deviation over 50 runs are shown.

	TPR	FPR
LinCovar	0.02 ± 0.06	0.02 ± 0.02
NonlinCovar	0.04 ± 0.08	0.02 ± 0.02

- **NonlinCovar**:

$$\begin{bmatrix} Y^0 \\ Y^1 \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} -5 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 1 - \frac{1}{h(g(X_1,\ldots,X_5))} \\ 1 - \frac{1}{h(g(X_1,\ldots,X_5))} & 1 \end{bmatrix}\right), \tag{4.15}$$

where functions f, g, and h are presented in Section 4.5.2. Under LinCovar and NonlinCovar, features $X_1, \ldots, X_5$ only influence the covariance between potential outcomes Y^0 and Y^1 and do not affect any functionals of the marginal distributions.

We performed 50 experiments and evaluated their mean and standard deviation of TPRs and FPRs. Table 4.3 presents the results. As expected, our method could not correctly select features $X_1, \ldots, X_5$ because their values do not affect the discrepancy between conditional potential outcome distributions.

Note, however, that selecting these features is extremely challenging because it is impossible to estimate the covariance since we cannot infer the joint distribution of potential outcomes, as described in Section 4.3.1. Due to this difficulty, all the existing mean-based methods also fail, and compared with such methods, ours can detect a wider variety of features.

Performance Evaluation with Different Neural Network Architectures

Since our method relies on two neural network models to represent propensity function $\mathrm{e}(\boldsymbol{X})$ and CVAE $\mathcal{L}(X_m \mid \boldsymbol{X}_{-m})$ $(m = 1, \ldots, d)$, we confirmed how greatly the neural network architectures affect the overall feature selection performance.

To confirm this, we performed additional synthetic data experiments with sample size $n = 1000$. We evaluated the mean and standard deviation of TPRs and FPRs over 50 runs by changing the number of neurons of each layer in the two-layered neural network models, which is fixed to 50 for propensity score and to 128 for CVAE in the experiments in Section 4.5.2.

Table 4.4: TPRs and FPRs of our method with different numbers of neurons in propensity score model. Mean and standard deviation over 50 runs are shown.

		Number of neurons in propensity score model			
		25	50	100	200
LinMean	TPR	0.80±0.21	0.79±0.22	0.84±0.14	0.84±0.16
	FPR	0.06±0.06	0.06±0.07	0.08±0.06	0.08±0.06
NonlinMean	TPR	0.95±0.10	0.94±0.12	0.98±0.06	0.97±0.08
	FPR	0.04±0.04	0.04±0.04	0.03±0.03	0.05±0.04
LinVar	TPR	0.71±0.19	0.73±0.19	0.77±0.16	0.76±0.18
	FPR	0.08±0.07	0.07±0.08	0.10±0.07	0.09±0.07
NonlinVar	TPR	0.64±0.25	0.62±0.25	0.63±0.26	0.64±0.25
	FPR	0.04±0.04	0.04±0.04	0.04±0.04	0.04±0.04

Table 4.5: TPRs and FPRs of our method with different numbers of neurons in CVAE model. Mean and standard deviation over 50 runs are shown.

		Number of neurons in CVAE model			
		16	64	128	256
LinMean	TPR	0.82±0.18	0.82±0.17	0.79±0.22	0.83±0.16
	FPR	0.08±0.06	0.07±0.06	0.06±0.07	0.10±0.07
NonlinMean	TPR	0.96±0.09	0.98±0.06	0.94±0.12	0.94±0.05
	FPR	0.04±0.04	0.03±0.03	0.04±0.04	0.05±0.04
LinVar	TPR	0.68±0.19	0.66±0.17	0.73±0.19	0.70±0.16
	FPR	0.07±0.05	0.06±0.05	0.07±0.08	0.08±0.07
NonlinVar	TPR	0.58±0.25	0.56±0.25	0.62±0.25	0.60±0.20
	FPR	0.02±0.03	0.03±0.03	0.04±0.04	0.04±0.05

Tables 4.4 and 4.5 display the results. With all synthetic datasets, the number of neurons in propensity score and CVAE did not greatly affect the performance.

4.6 Conclusions

We proposed a feature selection framework for discovering the features related to the distributional treatment effect heterogeneity. The key advantage of our framework is that it can identify the features whose values influence the functionals of the joint distribution of potential outcomes if the feature values also affect the discrepancy between conditional potential outcome distributions. To the best of our knowledge, this is the first feature selection approach to revealing the causal mechanism that yields the distributional treatment effect heterogeneity. We experimentally show that our feature selection framework successfully selected important features and outperformed the existing method.

4.7 Proofs

4.7.1 Relationship between Marginal and Joint Distributions

To confirm that our feature importance measure is reasonable, we consider the following two relationships:

- *If the discrepancy between marginal potential outcome distributions* $\mathrm{P}(Y^0 \mid X_m)$ *and* $\mathrm{P}(Y^1 \mid X_m)$ *varies with feature* X_m*'s values, then joint distribution* $\mathrm{P}(Y^0, Y^1 \mid X_m)$ *is also changeable depending on* X_m*'s values.*
- *If joint distribution* $\mathrm{P}(Y^0, Y^1 \mid X_m)$ *changes depending on feature* X_m*'s values, then some functionals of the joint distribution depend on* X_m*'s values.*

Since the second relationship is obvious, in this section, we show that the first relationship holds. For simplicity, we consider binary feature $X_m \in \{0, 1\}$; however, the following discussion also holds for discrete-valued and continuous-valued X_m.

To prove the first relationship, it is sufficient to show that its contraposition holds: If $\mathrm{P}(Y^0, Y^1 \mid X_m = 0) = \mathrm{P}(Y^0, Y^1 \mid X_m = 1)$, then the discrepancy between $\mathrm{P}(Y^0 \mid X_m = 0)$ and $\mathrm{P}(Y^1 \mid X_m = 0)$ equals the one between $\mathrm{P}(Y^0 \mid X_m = 1)$ and $\mathrm{P}(Y^1 \mid X_m = 1)$. We can easily prove this contraposition. From the equality of the joint distributions, we have $\mathrm{P}(Y^0 \mid X_m = 0) = \mathrm{P}(Y^0 \mid X_m = 1)$ and $\mathrm{P}(Y^1 \mid X_m = 0) = \mathrm{P}(Y^1 \mid X_m = 1)$. These equalities imply that the discrepancy between $\mathrm{P}(Y^0 \mid X_m = 0)$ and $\mathrm{P}(Y^1 \mid X_m = 0)$ equals the one between $\mathrm{P}(Y^0 \mid X_m = 1)$ and $\mathrm{P}(Y^1 \mid X_m = 1)$. Thus we proved the first relationship.

4.7.2 Counterexamples

As described in Section 4.3.1, there are several counterexamples where our method cannot find the features related to the functionals of the joint distribution of potential outcomes.

Let Y^0 and Y^1 be the potential outcomes and $X \in \{0, 1\}$ be a binary feature. Suppose that the discrepancy between marginal distributions $\mathrm{P}(Y^0 \mid X)$ and $\mathrm{P}(Y^1 \mid X)$ is measured as the MMD [Gretton *et al.*, 2012]. Then we can represent such

counterexamples as the cases where the following relations hold:

$$\mathrm{P}(Y^0, Y^1 \mid X = 0) \neq \mathrm{P}(Y^0, Y^1 \mid X = 1)$$
$$\mathrm{MMD}^2(\mathrm{P}(Y^0 \mid X = 0), \mathrm{P}(Y^1 \mid X = 0)) = \mathrm{MMD}^2(\mathrm{P}(Y^0 \mid X = 1), \mathrm{P}(Y^1 \mid X = 1)).$$

Letting the potential outcomes be $Y^0, Y^1 \in \{-1, 0, 1\} \subset \mathbb{R}$, we take an example of joint probability tables that satisfies the above relations in Table 4.6. In this example, the MMD between marginal distributions remains unchanged:

$$\mathrm{MMD}^2(\mathrm{P}(Y^0 \mid X = 0), \mathrm{P}(Y^1 \mid X = 0)) = \mathrm{MMD}^2(\mathrm{P}(Y^0 \mid X = 1), \mathrm{P}(Y^1 \mid X = 1)) = 0.$$

By contrast, the joint distribution changes depending on X's values, as illustrated in Table 4.6. As a result, although the average treatment effect does not change, the treatment effect variance and the covariance between potential outcomes vary as follows:

$$\mathbb{E}[Y^1 - Y^0 \mid X = 0] = \mathbb{E}[Y^1 - Y^0 \mid X = 1] = 0$$
$$\mathrm{Cov}[Y^0, Y^1 \mid X = 0] = 1; \quad \mathrm{Cov}[Y^0, Y^1 \mid X = 1] = -1$$
$$\mathrm{Var}[Y^1 - Y^0 \mid X = 0] = 0; \quad \mathrm{Var}[Y^1 - Y^0 \mid X = 1] = 4.$$

In this example, since we cannot detect any change in the MMD between marginal distributions, our method fails to find that feature X is related to treatment effect heterogeneity. Note, however, that the existing mean-based approaches would also fail because the average treatment effect remains unchanged.

Addressing such counterexamples is extremely difficult. It requires us to estimate the functionals of the joint potential outcome distribution; however, inferring such a joint distribution is impossible, as described in Section 4.3.1. One possible solution is to utilize several techniques for estimating the lower and upper bounds on these functionals by making additional assumptions [Chen *et al.*, 2016; Russell, 2021; Shingaki and Kuroki, 2021]. Establishing a feature selection framework that utilizes such lower and upper bounds remains our future work.

Table 4.6: Joint probability tables of potential outcomes. Nonzero probabilities are shown in bold. Total expresses marginal potential outcome probabilities.

$\mathrm{P}(Y^0, Y^1 \mid X = 0)$				
Y^0 \ Y^1	-1	0	1	Total
-1	**0.5**	0	0	**0.5**
0	0	0	0	0
1	0	0	**0.5**	**0.5**
Total	**0.5**	0	**0.5**	**1.0**

$\mathrm{P}(Y^0, Y^1 \mid X = 1)$				
Y^0 \ Y^1	-1	0	1	Total
-1	0	0	**0.5**	**0.5**
0	0	0	0	0
1	**0.5**	0	0	**0.5**
Total	**0.5**	0	**0.5**	**1.0**

4.7.3 Proposition 1

Proof. Recall the following definition of $\mathrm{WCMMD}^2_{X_m=x}$:

$$
\begin{aligned}
&\mathrm{WCMMD}^2_{X_m=x} \\
&:= \mathbb{E}_{A,A',\boldsymbol{X}_{-m},\boldsymbol{X}'_{-m},Y,Y'|X_m=X'_m=x}[w^0(A,\boldsymbol{X})w^0(A',\boldsymbol{X}')k_Y(Y,Y')] \\
&+ \mathbb{E}_{A,A',\boldsymbol{X}_{-m},\boldsymbol{X}'_{-m},Y,Y'|X_m=X'_m=x}[w^1(A,\boldsymbol{X})w^1(A',\boldsymbol{X}')k_Y(Y,Y')] \\
&-2\,\mathbb{E}_{A,A',\boldsymbol{X}_{-m},\boldsymbol{X}'_{-m},Y,Y'|X_m=X'_m=x}[w^0(A,\boldsymbol{X})w^1(A',\boldsymbol{X}')k_Y(Y,Y')]. \qquad (4.5)
\end{aligned}
$$

We show that the first term in Eq. (4.5) equals the one in $D^2_m(x)$ in Eq. (4.2). Using conditional ignorability and positivity assumptions, we have

$$
\begin{aligned}
&\mathbb{E}_{A,A',\boldsymbol{X}_{-m},\boldsymbol{X}'_{-m},Y,Y'|X_m=X'_m=x}[w^0(A,\boldsymbol{X})w^0(A',\boldsymbol{X}')k_Y(Y,Y')] \\
&= \mathbb{E}_{\boldsymbol{X}_{-m},\boldsymbol{X}'_{-m}|X_m=X'_m=x}\left[\mathbb{E}_{A,A',Y,Y'|\boldsymbol{X}_{-m},\boldsymbol{X}'_{-m},X_m=X'_m=x}\left[\frac{\mathbf{I}(A=0)}{1-\mathrm{e}(\boldsymbol{X})}\frac{\mathbf{I}(A'=0)}{1-\mathrm{e}(\boldsymbol{X}')}k_Y(Y,Y')\right]\right] \\
&= \mathbb{E}_{\boldsymbol{X}_{-m},\boldsymbol{X}'_{-m}|X_m=X'_m=x}[\mathbb{E}_{Y^0,Y^{0\prime}|\boldsymbol{X}_{-m},\boldsymbol{X}'_{-m},X_m=X'_m=x}[k_Y(Y^0,Y^{0\prime})]] \\
&= \mathbb{E}_{Y^0,Y^{0\prime}|X_m=x,X'_m=x}[k_Y(Y^0,Y^{0\prime})].
\end{aligned}
$$

Similarly, the second and third terms in Eq. (4.5) equal those in $\mathrm{MMD}^2(\mathrm{P}(Y^0 \mid x), \mathrm{P}(Y^1 \mid x))$ in Eq. (4.2). Thus we proved Proposition 1. □

4.7.4 Theorem 1

From Proposition 1, we only have to show that $\widehat{D}^2_m(x) \xrightarrow{p} \mathrm{WCMMD}^2_{X_m=x}$ $(n \to \infty)$ under the assumptions of conditional ignorability and positivity:

Assumption 3 (Conditional ignorability)**.** *For treatment A, features $\boldsymbol{X}$, and potential outcomes Y^0 and Y^1, the following conditional independence relation holds:*

$$\{Y^0, Y^1\} \perp\!\!\!\perp A \mid \boldsymbol{X}.$$

Assumption 4 (Positivity)**.** *For any value $\boldsymbol{x}$ of features $\boldsymbol{X}$, propensity score* $\mathrm{e}(\boldsymbol{X})$ *satisfies the following support condition:*

$$0 < \mathrm{e}(\boldsymbol{x}) < 1.$$

To prove $\widehat{D}^2_m(x) \xrightarrow{p} \mathrm{WCMMD}^2_{X_m=x}$ $(n \to \infty)$, we make several additional assumptions and impose the condition that the following symmetric function is square integrable:

$$\begin{aligned} &K((A, \boldsymbol{X}, Y), (A', \boldsymbol{X}', Y')) \\ :=&(w^0(A, \boldsymbol{X})w^0(A', \boldsymbol{X}') + w^1(A, \boldsymbol{X}, Y)w^1(A', \boldsymbol{X}', Y') \\ &- w^0(A, \boldsymbol{X})w^1(A', \boldsymbol{X}') - w^1(A, \boldsymbol{X})w^0(A', \boldsymbol{X}'))k_Y(Y, Y'). \end{aligned}$$

Assumption 5. *Symmetric function K is square integrable:*

$$\mathbb{E}_{A,A',\boldsymbol{X},\boldsymbol{X}',Y,Y'}[K((A, \boldsymbol{X}, Y), (A', \boldsymbol{X}', Y'))] < \infty.$$

When X_m is continuous-valued, and $\omega^{a,x}$ is given by (4.8), we make the following standard assumptions on kernel function k_{X_m}:

Assumption 6. *Let K_{X_m} be the following kernel function that measures the similarity between two values x_m and $x^\star_m$ on $\mathcal{X}$:*

$$K_{X_m}(x_m - x^\star_m) := \frac{1}{h_{X_m}} k_{X_m}(x_m, x^\star_m).$$

Then the order of function $K_{X_m}(u)$ is given by integer $\delta \geq 2$; in other words, the

following holds:

$$\int u^{\delta} K_{X_m}(u) du < \infty.$$

Assumption 7. *Bandwidth h_{X_m} of kernel function k_{X_m} satisfies*

$$h_{X_m} \to 0 \quad \textit{and} \quad n h_{X_m} \to \infty. \quad (n \to \infty)$$

In addition, we impose the smoothness conditions on marginal distribution $\mathrm{P}(X_m)$ and the joint distribution of features $\mathrm{P}(\boldsymbol{X})$:

Assumption 8. *Density functions $\mathrm{P}(X_m)$ and $\mathrm{P}(\boldsymbol{X})$ are δ times continuously differentiable.*

Using these assumptions, we prove Theorem 1:

Proof. **The case where weight $\omega_i^{a,x}$ is given by Eq. (4.6):**
Let $K_{i,j} \coloneqq K((a_i, \boldsymbol{x}_i, y_i), (a_j, \boldsymbol{x}_j, y_j))$ for $i, j \in \{1, \ldots, n\}$ and $n_x \coloneqq \sum_{i=1}^{n} \mathbf{I}(x_{m,i} = x)$. Then empirical estimator $\widehat{D}_m^2(x)$ is given as

$$\begin{aligned}
\widehat{D}_m^2(x) &= \frac{1}{n_x^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \mathbf{I}(x_{m,i} = x)\, \mathbf{I}(x_{m,j} = x) K_{i,j} \\
&= \left(\frac{n}{n_x}\right)^2 \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \mathbf{I}(x_{m,i} = x)\, \mathbf{I}(x_{m,j} = x) K_{i,j} \\
&= \left(\frac{n}{n_x}\right)^2 V_n^x,
\end{aligned}$$

where

$$V_n^x \coloneqq \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \mathbf{I}(x_{m,i} = x)\, \mathbf{I}(x_{m,j} = x) K_{i,j}$$

is a V-statistic whose corresponding U-statistic is given as

$$U_n^x \coloneqq \frac{1}{{}_n\mathrm{C}_2} \sum_{i<j} \mathbf{I}(x_{m,i} = x)\, \mathbf{I}(x_{m,j} = x) K_{i,j}.$$

We prove the consistency of $\widehat{D}_m^2(x)$ by showing the following three relations:

$$U_n^x \xrightarrow{a.s.} \mathbb{E}_{A,A',\boldsymbol{X},\boldsymbol{X}',Y,Y'}[\mathbf{I}(X_m = x)\,\mathbf{I}(X_m = x)K((A,\boldsymbol{X},Y),(A',\boldsymbol{X}',Y'))] \tag{4.16}$$

$$\left(\frac{n}{n_x}\right)^2 U_n^x \xrightarrow{a.s.} \mathrm{WCMMD}^2_{X_m=x} \tag{4.17}$$

$$U_n^x - V_n^x \xrightarrow{p} 0. \tag{4.18}$$

Relation (4.16) holds from the Strong Law of Large Numbers for U-statistics [Hoeffding, 1961]. By combining this relation with the fact that $\frac{n_x}{n} = \frac{1}{n}\sum_{i=1}^n \mathbf{I}(x_{m,i} = x) \xrightarrow{a.s.} \mathrm{P}(X_m = x)$, we can derive the relation in Eq. (4.17). The relation in Eq. (4.18) can be shown as follows. Under Assumption 5, since $\mathbb{E}[K((A,\boldsymbol{X},Y),(A',\boldsymbol{X}',Y'))] \leq \mathbb{E}[K((A,\boldsymbol{X},Y),(A,\boldsymbol{X},Y))] < \infty$, by employing Lemma 5.7.3 in Serfling [2009], we have $\mathbb{E}[|U_n^x - V_n^x|] = O(n^{-1})$, and thus by applying Markov's inequality, we have

$$\mathrm{P}(|U_n^x - V_n^x| \geq \epsilon) \leq \frac{\mathbb{E}[|U_n^x - V_n^x|]}{\epsilon} \to 0 \quad \text{as } n \to \infty,$$

which is sufficient to prove the relation in Eq. (4.18).

By combining Eqs. (4.16), (4.17), and (4.18), we have $\widehat{D}_m^2(x) \xrightarrow{p} \mathrm{WCMMD}^2_{X_m=x}$ as $n \to \infty$. Since Proposition 1 holds under Assumptions 3 and 4, we have $\widehat{D}_m^2(x) \xrightarrow{p} D_m^2(x)$ as $n \to \infty$. Thus we prove the consistency of $\widehat{D}_m^2(x)$.

The case where weight $\omega_i^{a,x}$ is given by Eq. (4.8):

In this case, empirical estimator $\widehat{D}_m^2(x)$ is given as

$$\widehat{D}_m^2(x) = \frac{\frac{1}{n^2 h_{X_m}^2}\sum_{i=1}^n\sum_{j=1}^n k_{X_m}(x_{m,i},x)k_{X_m}(x_{m,j},x)K_{i,j}}{\frac{1}{n^2 h_{X_m}^2}\sum_{i=1}^n\sum_{j=1}^n k_{X_m}(x_{m,i},x)k_{X_m}(x_{m,j},x)}. \tag{4.19}$$

From the Strong Law of Large Numbers, as $n \to \infty$, the numerator in Eq. (4.19) converges to the following expected value:

$$\mathbb{E}_{A,A',\boldsymbol{X},\boldsymbol{X}',Y,Y'}\left[\frac{1}{h_{X_m}^2}K_{X_m}\left(\frac{X_m - x}{h_{X_m}}\right)K_{X_m}\left(\frac{X_m' - x}{h_{X_m}}\right)K((A,\boldsymbol{X},Y),(A',\boldsymbol{X}',Y'))\right].$$

Under Assumptions 6 and 8, we can reformulate this expected value by performing

a Taylor expansion as follows:

$$\begin{aligned}
&\mathbb{E}_{A,A',\boldsymbol{X},\boldsymbol{X}',Y,Y'}\left[\frac{1}{h_{X_m}^2}K_{X_m}\left(\frac{X_m-x}{h_{X_m}}\right)K_{X_m}\left(\frac{X'_m-x}{h_{X_m}}\right)K((A,\boldsymbol{X},Y),(A',\boldsymbol{X}',Y'))\right] \\
&=\mathbb{E}_{U=u,V=v}[\mathbb{E}_{A,A',\boldsymbol{X}_{-m},\boldsymbol{X}'_{-m},Y,Y'|X_m=x+h_{X_m}u,X'_m=x+h_{X_m}v}[\\
&\quad \mathrm{P}(X_m=x+h_{X_m}u)\,\mathrm{P}(X'_m=x+h_{X_m}v)K_{X_m}(u)K_{X_m}(v)K((A,\boldsymbol{X},Y),(A',\boldsymbol{X}',Y'))]] \\
&=\mathbb{E}_{A,A',\boldsymbol{X}_{-m},\boldsymbol{X}'_{-m},Y,Y'|X_m=x,X'_m=x}[\mathrm{P}^2(X_m=x)K((A,\boldsymbol{X},Y),(A',\boldsymbol{X}',Y'))]+O_p\left(h_{X_m}^{\delta}\right).
\end{aligned} \tag{4.20}$$

Regarding the denominator in Eq. (4.19), from the consistency results of the kernel density estimator in Wied and Weißbach [2012], we have

$$\frac{1}{nh_{X_m}}\sum_{j=1}^{n}k_{X_m}(x_{m,j},x)\xrightarrow{a.s.}\mathrm{P}(X_m=x). \tag{4.21}$$

By combining Eqs. (4.20) and (4.21), under Assumption 7, we have $\widehat{D}_m^2(x)\xrightarrow{p}\mathrm{WCMMD}^2_{X_m=x}$ as $n\to\infty$. Using Proposition 1, we have $\widehat{D}_m^2(x)\xrightarrow{p}D_m^2(x)$ as $n\to\infty$. Thus we proved the consistency of $\widehat{D}_m^2(x)$.

□

Chapter 5

Making Individually Fair Predictions with Causal Pathways

In this chapter, we consider the problem of learning a *fair* predictive model by utilizing the causal relationships underlying in the data. Machine learning is being increasingly used to make algorithmic decisions that have strong societal impact on people's lives. Due to their huge societal impact, such algorithmic decisions need to be accurate and fair with respect to sensitive features, including race, gender, religion, and sexual orientation. To achieve a good balance between prediction accuracy and fairness, causality-based methods have been proposed, which utilize a *causal graph* with *unfair pathways*. However, as described in this chapter, none of these methods can ensure fairness for each individual without making restrictive functional assumptions about the data generating processes, which are not satisfied in many cases. To overcome such a weakness of the existing methods, we propose a far more practical causality-based framework for learning an individually fair classifier.

5.1 Introduction

Algorithmic decision-making systems based on machine learning have become ubiquitous in our societies. These systems are increasingly used to make decisions that severely affect people's lives, e.g., granting loans [Khandani *et al.*, 2010], employment decisions [Houser, 2019], child abuse assessment [Chouldechova *et al.*, 2018], and recidivism predictions [Angwin *et al.*, 2016]. Since these decisions often have a huge

societal impact on each individual, machine learning predictions that support them must be both accurate and fair with respect to such sensitive features as gender, race, religion, disabilities, and sexual orientation.

To make fair predictions, many methods have been developed that remove the correlation between sensitive features and predictions [Dwork *et al.*, 2012; Feldman *et al.*, 2015; Hardt *et al.*, 2016]. However, in complex real-world scenarios, these methods may impose unnecessary fairness constraints and decrease the prediction accuracy. Consider hiring decisions for male and female applicants. When applied to this scenario, traditional correlation-based methods reject male and female applicants at the same rate [Feldman *et al.*, 2015]. However, in complex real-world scenarios, the presence of gender differences in rejection rates may sometimes be justified. For instance, consider hiring decisions for physically demanding jobs. Because the jobs require physical strength, it is sometimes **not** discriminatory to reject applicants due to a lack of physical strength. Since physical strength is often affected by gender, such rejections produce a gender difference in rejection rates, which is fair and does not need to be removed. Nevertheless, the aforementioned methods aim to eliminate such fair differences, thus imposing unnecessary fairness constraints. This implies that when we have a male and female applicant, even if the man has a much more physical strength than the woman, these methods may reject him and accept her, which greatly reduces the prediction accuracy.

To achieve high prediction accuracy, several causality-based methods have been proposed that avoid imposing unnecessary fairness constraints [Chiappa and Gillam, 2019; Kusner *et al.*, 2017; Nabi and Shpitser, 2018; Zhang *et al.*, 2017]. These methods measure the unfairness of predictions using a DAG called a causal graph that contains *unfair pathways*. For instance, in the case of hiring decisions for physically demanding jobs, a causal graph may be expressed as shown in Figure 5.1, where A, Q, M, and Y represent gender, qualifications, physical strength, and prediction, respectively. With this causal graph, we can express our consensus that prediction Y is unfair if it is based on gender A and fair

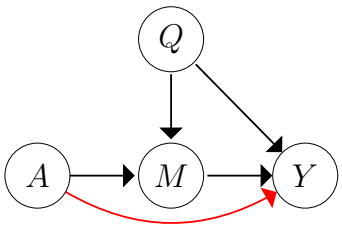

Figure 5.1: Causal graph representing a scenario of hiring decisions for physically demanding jobs. Red edge $A \to Y$ is regarded as unfair pathway π.

if it is based on physical strength M. To express such a consensus, we regard $A \to Y$ as unfair and $A \to M \to Y$ as fair, expressed by unfair pathway $\pi = \{A \to Y\}$. By utilizing this unfair pathway, we can effectively measure unfair gender differences as unfair path-specific effects [Avin *et al.*, 2005] of gender A on prediction Y via unfair pathway π.

However, due to the difficulty of estimating such an unfair effect, no existing method can ensure fairness for each individual without making restrictive functional assumptions on the data. To guarantee individual-level fairness, the *path-specific counterfactual fairness* (PSCF) method [Chiappa and Gillam, 2019] requires an assumption that the data are generated by a restricted class of functions called *additive noise models* [Hoyer *et al.*, 2009]. Unfortunately, this functional assumption is not satisfied in many cases. Although several existing methods such as the *fair inference on outcome* (FIO) method [Nabi and Shpitser, 2018] do not require such demanding functional assumptions, they cannot ensure fairness for each individual.

In this chapter, we propose a learning framework that guarantees individual-level fairness without making impractical functional assumptions. We train an individually fair classifier based on an unfairness measure that can be estimated without making strong functional assumptions on data. To obtain such an unfairness measure, we consider the *probability of individual unfairness* (PIU), i.e., the probability that an unfair effect takes non-zero values, and derive its upper bound that can be estimated from data. We train an individually fair classifier by forcing this upper bound value to be nearly zero, which can be achieved by solving the penalized optimization problem.

5.1.1 Contributions

Our contributions are summarized as follows:

- We establish a learning framework that guarantees fairness for each individual without restrictive functional assumptions on data (Table 5.1). To achieve this, we force the PIU value to be close to zero by imposing a penalty on its upper bound, which we can estimate from data.

- We elucidate why imposing such a penalty guarantees individual-level fairness in Section 5.4. To do so, we compare our penalty with the constraint of the

Table 5.1: Comparison with existing methods

Method	Individually fair	Functional assumptions on data
Our method	Yes	Unnecessary
PSCF [Chiappa and Gillam, 2019]	Yes	Necessary
FIO [Nabi and Shpitser, 2018]	No	Unnecessary

existing FIO method, which does not ensure individual-level fairness, from a viewpoint of feasible regions of (constrained) optimization problems.

- We show how our learning framework can be extended to address challenging real-world scenarios in Section 5.5. We provide two extensions that allow us to address cases where there are unobserved variables called *latent confounders* (Section 5.5.1) and where the true causal graph is uncertain (Section 5.5.2).

- We experimentally show that our method makes much fairer predictions for each individual than the existing methods at a slight expense of prediction accuracy.

The rest of this chapter is organized as follows. Section 5.2 describes our problem setting, several basic concepts of causality, and the weaknesses of the existing methods. Section 5.3 presents our proposed learning framework that resolves the weaknesses of the existing methods. Section 5.4 illustrates why our framework guarantees individual-level fairness. Section 5.5 introduces two extensions of our framework for dealing with challenging real-world scenarios. Section 5.6 shows the performance of our learning framework (Section 5.3) and its extensions (Section 5.5). Section 5.7 concludes this chapter. All proofs are provided in Section 5.8.

5.2 Background

5.2.1 Problem Statement

We consider a binary classification task that takes the following two inputs. One is training data that contain the observations of decision outcome $Y \in \{0, 1\}$ and the features of each individual $\boldsymbol{X}$, including sensitive feature $A \in \{0, 1\}$. The other is a DAG called a causal graph, whose nodes and edges express random variables in $\{\boldsymbol{X}, Y\}$ and causal relationships, respectively [Pearl, 2009]. In most of this chapter,

we assume that such a graph structure can be depicted by domain experts or inferred from data using existing causal discovery methods [Glymour *et al.*, 2019]. Although this assumption is widely used in the existing methods [Chiappa and Gillam, 2019; Kusner *et al.*, 2017; Nabi and Shpitser, 2018; Zhang and Wu, 2017], it can be violated if the true causal graph is uncertain, which is possible in practice. To address such cases, in Section 5.5.2, we consider a variant of the classification task where multiple candidates of causal graphs are given as input.

Taking training data and causal graph(s) as input, we train classifier h_θ that predicts decision outcome Y from features $\boldsymbol{X}$. We seek classifier parameter θ that achieves a good balance between prediction accuracy and fairness with respect to sensitive feature A. Let L_θ be a loss function that measures prediction errors, and let G_θ be a penalty function that quantifies the unfairness of predictions. Given n training instances $\{(\boldsymbol{x}_i, y_i)_{i=1}^n\}$, we consider the following optimization problem:

$$\min_\theta \quad \frac{1}{n}\sum_{i=1}^{n} L_\theta(\boldsymbol{x}_i, y_i) + \lambda G_\theta(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n), \tag{5.1}$$

where $\lambda \geq 0$ is a hyperparameter that controls the penalty on unfairness.

To achieve high prediction accuracy, we must appropriately design penalty function G_θ in (5.1) to avoid unnecessary penalizations. For instance, in the example of hiring decisions for physically demanding jobs (Section 5.1), imposing a penalty by gender differences in rejection rates might be unnecessary because this gender difference is yielded not only by the rejections of applicants based on gender A but also based on physical strength M; the latter rejection is sometimes not discriminatory since the job requires physical strength.

To avoid imposing such an unnecessary penalization, we utilize prior knowledge about discrimination (e.g., the consensus that rejecting applicants because of the lack of physical strength is fair only if physical strength is necessary for the job). To express such prior knowledge, we use unfair pathways π in the input causal graph, as illustrated in the next section.

5.2.2 Unfair Pathway Examples

We take examples of unfair pathways based on two scenarios of hiring decisions. We consider the following features of each applicant: gender A, qualification Q,

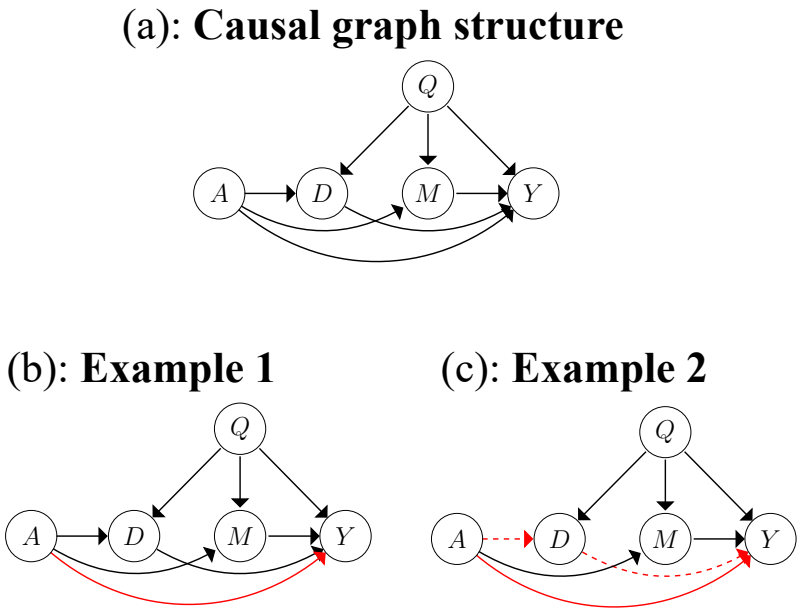

Figure 5.2: Each hiring decision scenario is expressed by causal graph structure in (a), where A, Q, M, D, and Y denote gender, qualification, physical strength, academic background, and prediction. Our prior knowledge about discrimination is expressed as unfair pathways. (b): In **Example 1**, solid red edge $A \to Y$ is unfair. (c): In **Example 2**, $A \to Y$ and dashed red pathway $A \to D \to Y$ are unfair.

physical strength M, and academic background D (e.g., field of study). The causal relationships between these features and prediction Y are represented by the causal graph structure displayed in Figure 5.2(a).

From the pathways from A to Y in this causal graph, we choose unfair pathways π, whose choice depends on the consensus on discrimination in each scenario. In what follows, we present the simplest choice of unfair pathways.

Example 1. Let direct pathway $A \to Y$ be unfair (i.e., $\pi = \{A \to Y\}$), as illustrated in Figure 5.2(b). With this choice of unfair pathway π, we can express our consensus that only *direct discrimination* (i.e., treating someone unfairly because of gender A) should be prohibited, thus indicating that prediction Y is discriminatory only if it is based on gender A.

As described in Kusner *et al.* [2017, Section S4], when only direct pathway $\pi = \{A \to Y\}$ is regarded as unfair as illustrated in **Example 1**, we can make fair predictions by training a predictive model without sensitive feature A. Since several input features (i.e., academic background D and physical strength M) are affected

by A, the trained predictive model may still exhibit a large difference of rejection rates between men and women; however, this gender difference is regarded as fair in this scenario. There might be cases where a job needs expertise in a particular academic background and much physical strength.

In more complicated scenarios than **Example 1**, we cannot achieve fairness only by removing sensitive feature A from inputs of a predictive model. For instance, consider the cases where unfair pathways π contain multiple pathways from A to Y:

Example 2. Let unfair pathways be $\pi = \{A \to Y, A \to D \to Y\}$ (Figure 5.2(c)). These unfair pathways imply that we should avoid not only direct discrimination ($A \to Y$) but also *indirect discrimination* based on academic background D ($A \to D \to Y$). To forbid such indirect discrimination, we need to eliminate gender differences in rejection rates that are yielded by rejecting applicants based on whether they have academic expertise in a particular domain (e.g., computer science) whose number of students exhibits a large gender difference.

In **Example 2**, although it is unfair to reject applicants based on academic background D, it is fair to decline their job applications based on physical strength M because pathway $A \to M \to Y$ is not included in unfair pathways π. This might correspond to hiring decision scenarios for jobs that require much physical strength (but no particular academic expertise).

A naive approach to ensure fairness in these scenarios is to predict without gender A or academic background D. This approach, however, may unnecessarily decrease the prediction accuracy if D is only slightly influenced by gender A and largely affected by such unobserved important features as logical thinking skills.

To achieve high prediction accuracy, we need to design penalty function G_θ in (5.1) by measuring the unfairness from data based on unfair pathways π.

5.2.3 Measuring Unfairness from Data

To measure the unfairness of predictions based on data, we utilize the concept of path-specific effects [Avin *et al.*, 2005], which express how largely an observed variable affects another variable via pathways in a causal graph. This unfairness measure, which we call an unfair effect, enables us to quantify how greatly sensitive feature A influences prediction Y via unfair pathways π.

Such an unfair effect is defined for each individual as a difference between the two predicted decision outcomes that are obtained using different feature attributes. For instance, when making a hiring decision for a female applicant, an unfair effect for her is the difference between the two predictions, each of which is made using the following different CVs. One is the original CV, which includes her feature attributes. The other is a *counterfactual* CV where some of her attributes are modified as if she were male. This modification of feature attributes depends on unfair pathways π. For example, in the case of **Example 2** with $\pi = \{A \to Y, A \to D \to Y\}$, gender A is changed to male, and academic background D is modified to a counterfactual one that she would choose if she were male; on the other hand, physical strength M is left unchanged because it does not appear in π, which means that it is irrelevant to discrimination.

To express such a counterfactual attribute, we need to formulate how each feature attribute is determined and express what attributes would be obtained if sensitive feature attribute were changed. To formulate them, we need an SEM and an interventional SEM that express generating processes of (observed) data and counterfactual data, respectively.

SEMs and Interventional SEMs

An SEM, which expresses how each random variable value is determined. As described in Definition 4 in Section 2.3.1, it is defined as triplet $\mathcal{M} = (\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{F})$, where $\boldsymbol{U}$ denotes unobserved variables called exogenous variables, $\boldsymbol{V}$ denotes observed variables called endogenous variables, $\boldsymbol{F}$ denotes deterministic functions. Each observed variable $V \in \boldsymbol{V}$ is determined by the following structural equation:

$$V = f_V(\boldsymbol{pa}(V), \boldsymbol{U}_V), \tag{2.15}$$

where $\boldsymbol{pa}(V) \subseteq \boldsymbol{V} \backslash V$ are the observed variables that are the parents of V in the causal graph, and $\boldsymbol{U}_V \subseteq \boldsymbol{U}$ are unobserved variables that represent external factors, such as measurement errors and unmeasurable quantities.

In our problem setting, we consider the following SEM, $\mathcal{M}^p$, where superscript p denotes prediction. With this SEM, by letting $\boldsymbol{V} = \{\boldsymbol{X}, Y\}$, we express how the values of features $\boldsymbol{X}$ and prediction Y are determined. Since prediction Y is represented using classifier h_θ, structural equation over Y is given by $Y = h_\theta(\boldsymbol{X})$

if it is deterministic; otherwise, $Y = h_\theta(\boldsymbol{X}, \boldsymbol{U}_Y)$, where $\boldsymbol{U}_Y \subseteq \boldsymbol{U}$ denotes the unobserved random noises used in the classifier. These formulations of structural equations can be regarded as a special case of (2.15), where function $f_Y \in \boldsymbol{F}$ is replaced with classifier h_θ. By contrast, structural equations over features $\boldsymbol{X}$ are formulated without considering such a special case. To take an example, consider the causal graph in Figure 5.2(a), where $\boldsymbol{X} = \{A, Q, D, M\}$. The parents of each feature variable are $\boldsymbol{pa}(A) = \phi$, $\boldsymbol{pa}(Q) = \phi$, $\boldsymbol{pa}(D) = \{A, Q\}$, and $\boldsymbol{pa}(M) = \{A, Q\}$, respectively. Based on these parental relationships between variables, we can express various data generating processes as structural equations. For instance, using univariate unobserved noises U_A, U_Q, U_M, and U_D, the following structural equations over A, Q, M, and D might be possible:

$$A = U_A, \quad Q = U_Q, \quad M = 3A + 0.5Q + U_M, \quad D = A + U_D Q. \tag{5.2}$$

If such an SEM is given, we can formulate how attributes for each individual are determined under a counterfactual situation where their sensitive feature attribute is changed. To do so, we replace the structural equation over sensitive feature A (e.g., $A = U_A$ in Eq. (5.2)) with $A = a$, where $a \in \{0, 1\}$ is a constant. As described in Section 2.3.2, this replacement of structural equations is called intervention $do(A = a)$, which forces A's value to be constant $A = a$ for all individuals. This indicates that even when some individuals have attribute $A \neq a$, which is randomly determined by an original SEM, it is changed to a different one, $A = a$, expressing the counterfactual situation. The data generating processes under such counterfactual situations are characterized by an SEM modified by intervention $do(A = a)$, called an interventional SEM, denoted by $\mathcal{M}^p_{A=a}$. For instance, when the original SEM is expressed with structural equations (5.2), interventional SEM $\mathcal{M}^p_{A=a}$ is formulated using the following structural equations:

$$A = a, \quad Q = U_Q, \quad M(a) = 3a + 0.5Q + U_M, \quad D(a) = a + U_D Q, \tag{5.3}$$

where gender A is fixed to constant a, and attributes $M(a)$ and $D(a)$ are affected by this constant. For individuals with (observed) gender $A \neq a$, attributes $M(a)$ and $D(a)$ are not observed; they represent the counterfactual attributes of physical strength M and academic background D if their gender A were changed to a. By

contrast, for individuals whose (observed) gender A is expressed as a, attributes $M(a)$ and $D(a)$ are identical to the observed attributes: M and D.

Using such attributes as $M(a)$ and $D(a)$ in Eq. (5.3), we formulate the unfair effects in the next section.

Unfair Effects

An unfair effect is the difference between two predictions, $Y_{A\Leftarrow 1\|\pi} - Y_{A\Leftarrow 0}$, where $Y_{A\Leftarrow 0}$ and $Y_{A\Leftarrow 1\|\pi}$ are called potential outcomes.

Potential outcome $Y_{A\Leftarrow 0}$ is defined as prediction Y under interventional model $\mathcal{M}^p_{A=0}$ with intervention $do(A = 0)$. By contrast, $Y_{A\Leftarrow 1\|\pi}$ is formulated using two interventional models, $\mathcal{M}^p_{A=0}$ and $\mathcal{M}^p_{A=1}$, whose formulation depends on unfair pathways π.[1]

For instance, for **Example 1** in Section 5.2.2, these potential outcomes are formulated to measure unfair effects based on direct pathway $\pi = \{A \to Y\}$. When classifier $h_\theta(A, Q, M, D)$ is deterministic,[2] potential outcomes $Y_{A\Leftarrow 0}$ and $Y_{A\Leftarrow 1\|\pi}$ are expressed:

$$Y_{A\Leftarrow 0} = h_\theta(0, Q, M(0), D(0)) \quad \text{and} \tag{5.4}$$

$$Y_{A\Leftarrow 1\|\pi} = h_\theta(1, Q, M(0), D(0)), \tag{5.5}$$

where $M(0)$ and $D(0)$ denote features M and D under intervention $do(A = 0)$, as presented in Eq. (5.3). In Eq. (5.4), input $A = 0$ is used. By contrast in Eq. (5.5), it is switched to $A = 1$ without changing the other input features (i.e., Q, $M(0)$, and $D(0)$). By taking difference $Y_{A\Leftarrow 1\|\pi} - Y_{A\Leftarrow 0}$ based on these inputs, we can measure the unfairness as path-specific effects via direct pathway $\pi = \{A \to Y\}$; such path-specific effects correspond to the measure of direct effects called NDEs [Pearl, 2001], which is described in Section 2.3.

Unlike NDEs, path-specific effects can quantify unfairness based on multiple pathways from A to Y. For example, they can measure the unfairness in **Example 2** in Section 5.2.2, where unfair pathways are $\pi = \{A \to Y, A \to D \to Y\}$ by formulating

[1] Here $A = 0$ can be regarded as a baseline for measuring an unfair effect, and this baseline can be switched to $A = 1$, which yields potential outcomes $Y_{A\Leftarrow 1}$ and $Y_{A\Leftarrow 0\|\pi}$.

[2] When classifier h_θ is not deterministic, potential outcomes are formulated in the same way using random noise that is employed in the classifier.

potential outcomes as follows:

$$Y_{A\Leftarrow 0} = h_\theta(0, Q, M(0), D(0)) \quad \text{and} \tag{5.4}$$

$$Y_{A\Leftarrow 1\|\pi} = h_\theta(1, Q, M(0), D(1)). \tag{5.6}$$

Here potential outcome $Y_{A\Leftarrow 0}$ is given in the same way as in **Example 1** because it does not depend on unfair pathways π. By contrast, potential outcome $Y_{A\Leftarrow 1\|\pi}$ is formulated by modifying inputs $A = 0$ and $D(0)$ in Eq. (5.4) to $A = 1$ and $D(1)$. By modifying inputs in this way and taking difference $Y_{A\Leftarrow 1\|\pi} - Y_{A\Leftarrow 0}$, we can measure the unfairness based on unfair pathways $\pi = \{A \to Y, A \to D \to Y\}$.

Intuitively, in the hiring decision scenario, switching several input feature attributes corresponds to modifying a CV, and by computing $Y_{A\Leftarrow 1\|\pi} - Y_{A\Leftarrow 0}$ based on such inputs, we measure the unfair differences of predicted hiring decision outcomes.

In practice, however, we cannot compute such unfair differences because some input features are not observed for each individual. For instance, to compute potential outcomes based on Eqs. (5.4) and (5.6), we need both input features $D(0)$ and $D(1)$. They are, however, not jointly observed for each individual as already described, which makes it impossible to compute an unfair effect for each individual.

For this reason, to learn fair predictive models, existing methods [Chiappa and Gillam, 2019; Nabi and Shpitser, 2018; Zhang *et al.*, 2017] use the (conditional) mean unfair effect, which can be estimated from observed data under several assumptions.

Mean Unfair Effects and Conditional Mean Unfair Effects

Formally, the (conditional) mean unfair effect is defined as the (conditional) expected value of unfair effects:

Definition 7. *For unfair pathways π and potential outcomes $Y_{A\Leftarrow 0}, Y_{A\Leftarrow 1\|\pi} \in \{0, 1\}$, the mean unfair effect is given by*

$$\mathbb{E}_{Y_{A\Leftarrow 0}, Y_{A\Leftarrow 1\|\pi}}[Y_{A\Leftarrow 1\|\pi} - Y_{A\Leftarrow 0}] = \mathrm{P}(Y_{A\Leftarrow 1\|\pi} = 1) - \mathrm{P}(Y_{A\Leftarrow 0} = 1). \tag{5.7}$$

Definition 8. *For unfair pathways π and potential outcomes $Y_{A\Leftarrow 0}, Y_{A\Leftarrow 1\|\pi} \in \{0, 1\}$, the conditional mean unfair effect (a.k.a. the path-specific counterfactual effect [Wu*

et al., *2019b]) conditioned on input features* $\boldsymbol{X} = \boldsymbol{x}$ *is given by*

$$\begin{aligned} &\mathbb{E}_{Y_{A\Leftarrow 0}, Y_{A\Leftarrow 1\|\pi}}[Y_{A\Leftarrow 1\|\pi} - Y_{A\Leftarrow 0} \mid \boldsymbol{X} = \boldsymbol{x}] \\ =& \mathrm{P}(Y_{A\Leftarrow 1\|\pi} = 1 \mid \boldsymbol{X} = \boldsymbol{x}) - \mathrm{P}(Y_{A\Leftarrow 0} = 1 \mid \boldsymbol{X} = \boldsymbol{x}). \end{aligned} \tag{5.8}$$

Roughly speaking, while the mean unfair effect is an average of the unfair effects over **all** individuals, the conditional mean unfair effect is an average in the **subgroup** of individuals who have identical feature attributes $\boldsymbol{X} = \boldsymbol{x}$.

To estimate the mean unfair effect, we need marginal probabilities $\mathrm{P}(Y_{A\Leftarrow 0} = 1)$ and $\mathrm{P}(Y_{A\Leftarrow 1\|\pi} = 1)$ in Eq. (5.7), which can be inferred using several assumptions such as conditional independence conditions called sequential ignorabilities (See Section 5.3.3 for details).

By contrast, as pointed out by Wu *et al.* [2019b], computing the conditional mean unfair effect requires additional demanding functional assumptions due to the difficulty of estimating conditional probabilities $\mathrm{P}(Y_{A\Leftarrow 0} = 1 \mid \boldsymbol{X} = \boldsymbol{x})$ and $\mathrm{P}(Y_{A\Leftarrow 1\|\pi} = 1 \mid \boldsymbol{X} = \boldsymbol{x})$ in Eq. (5.8). Suppose that conditioned feature values $\boldsymbol{X} = \boldsymbol{x}$ contain sensitive feature value $A = a$ ($a \in \{0, 1\}$). To compute conditional probabilities $\mathrm{P}(Y_{A\Leftarrow 0} = 1 \mid \boldsymbol{X} = \boldsymbol{x})$ and $\mathrm{P}(Y_{A\Leftarrow 1\|\pi} = 1 \mid \boldsymbol{X} = \boldsymbol{x})$, we need to infer the distribution of features over individuals with $A = a$; however, such a distribution is unavailable because some features are not observed for these individuals. For instance, in **Example 2**, either $D(0)$ or $D(1)$ is not observed for individuals with $A = a$ ($a \in \{0, 1\}$), which prevents us from estimating the conditional probabilities in Eq. (5.8). If the underlying SEM can be expressed by such simple functions as additive noise model $V = f_V(\boldsymbol{pa}(V)) + U_V$ [Hoyer *et al.*, 2009], we can approximate these conditional probabilities. However, assuming such a simple SEM is too restrictive because it cannot express most data generating processes. For instance, additive noise models cannot represent a structural equation over D in Eq. (5.2) due to multiplicative noise U_D. Therefore, in many cases, we cannot correctly estimate the conditional mean unfair effect.

However, to make individually fair predictions, we need to force the conditional mean unfair effect value to be zero, as described in the next section.

5.2.4 Individual-Level Fairness

To achieve fairness for each individual, it is insufficient to make the mean unfair effect value (close to) zero. This is because even when it is zero, depending on the attributes of each individual, $\boldsymbol{X} = \boldsymbol{x}$, unfair effect $Y_{A \Leftarrow 1 \| \pi} - Y_{A \Leftarrow 0}$ might be positive or negative, which indicates that predictions are discriminatory for individuals with attributes $\boldsymbol{x}$. We cannot simply resolve this issue using, e.g., the mean of the absolute values of the unfair effects because such a quantity requires a joint distribution of $Y_{A \Leftarrow 0}$ and $Y_{A \Leftarrow 1 \| \pi}$; unfortunately, this joint distribution is unavailable because we cannot obtain both $Y_{A \Leftarrow 0}$ and $Y_{A \Leftarrow 1 \| \pi}$ for each individual without an SEM.

For this reason, using the conditional mean unfair effect conditioned on features $\boldsymbol{X}$, Wu *et al.* [2019b] defines individual-level fairness as follows:

Definition 9 (Wu *et al.* [2019b])**.** *Given unfair pathways π in a causal graph, classifier h_θ achieves a* ***(path-specific) individual-level fairness*** *if*

$$\mathbb{E}_{Y_{A \Leftarrow 0}, Y_{A \Leftarrow 1 \| \pi}}[Y_{A \Leftarrow 1 \| \pi} - Y_{A \Leftarrow 0} \mid \boldsymbol{X} = \boldsymbol{x}] = 0 \tag{5.9}$$

holds for any value of $\boldsymbol{x}$ of input features $\boldsymbol{X}$.

Definition 9 states that classifier h_θ is individually fair if the conditional mean unfair effect is zero for any subgroup of individuals with identical feature attributes $\boldsymbol{X} = \boldsymbol{x}$. Here the conditional mean unfair effect value depends on classifier parameter θ since $Y_{A \Leftarrow 0}$ and $Y_{A \Leftarrow 1 \| \pi}$ are expressed using classifier h_θ.

To make individually fair predictions, we need to find appropriate θ values that satisfy (5.9). However, finding such θ values is extremely challenging due to the difficulty of estimating the conditional mean unfair effect in (5.9), which requires the restrictive functional assumptions described in Section 5.2.3.

For this reason, no existing methods can make individually fair predictions without restrictive functional assumptions, as described in the next section.

5.2.5 Related Work

Motivated by recent developments in inferring causal graph structures [Chikahara and Fujino, 2018a; Glymour *et al.*, 2019], many causality-based approaches to fairness have been proposed [Chiappa and Gillam, 2019; Kilbertus *et al.*, 2017; Kusner *et al.*,

2017, 2019; Nabi and Shpitser, 2018; Nabi *et al.*, 2019; Russell *et al.*, 2017; Salimi *et al.*, 2019; Wu *et al.*, 2018, 2019a; Xu *et al.*, 2019; Zhang *et al.*, 2017; Zhang and Wu, 2017; Zhang *et al.*, 2018; Zhang and Bareinboim, 2018a,b].

As discussed by Makhlouf *et al.* [2020], compared with correlation-based approaches [Dwork *et al.*, 2012; Feldman *et al.*, 2015; Hardt *et al.*, 2016], causality-based approaches provide more intuitive fairness interpretations because they can determine whether the correlation between sensitive features and predictions arises from causation or spurious correlation. Moreover, several causality-based approaches that utilize path-specific effects can achieve a good balance between prediction accuracy and fairness in complex real-world scenarios.

However, measuring unfairness as path-specific effects remains challenging due to the difficulty of estimation. For this reason, the FIO method [Nabi and Shpitser, 2018] uses the mean unfair effect (Definition 7), which can be estimated under reasonable assumptions. In particular, it relies only on the two standard assumtpions that are needed to marginal potential outcome probabilities $\mathrm{P}(Y_{A\Leftarrow 0} = 1)$ and $\mathrm{P}(Y_{A\Leftarrow 1\|\pi} = 1)$ in Eq. (5.7); we detail these assumptions later in Section 5.3.3 because our proposed method also makes the same assumptions. Unfortunately, forcing the mean unfair effect to be zero does not ensure that predictions are individually fair, as described in Section 5.2.4.

To make individually fair predictions, the PSCF method aims to remove the conditional mean unfair effect (Definition 8). To achieve this, it approximates the underlying data generating process (i.e., the SEM) by learning a deep generative model. However, as mentioned by [Kusner *et al.*, 2017], inferring such a data generating process needs an additional functional assumption that the value of each feature $V \in \boldsymbol{X}$ is determined by additive noise model $V = f_V(\boldsymbol{pa}(V)) + U_V$ [Hoyer *et al.*, 2009], where unobserved variable U_V is assumed to be an additive noise. However, as illustrated in Section 5.2.3, there are many examples of the data generating processes that do not satisfy such a functional assumption, and if the data do not satisfy this assumption, the PSCF method cannot guarantee individual-level fairness.

To resolve these issues, we propose a learning framework that guarantees fairness for each individual without restrictive functional assumptions.

5.3 Learning Individually Fair Classifier with Path-Specific Causal-Effect Constraints

5.3.1 Overview of Learning Framework

Given training data and a causal graph with unfair pathways π, by solving the penalized optimization problem in Eq. (5.1), we train a fair classifier that achieves the individual-level fairness condition (Definition 9).

To achieve this without demanding functional assumptions, we develop penalty function G_θ that forces an unfair effect to be (close to) zero for all individuals. To this end, our penalty function makes the potential outcomes take the same value (i.e., $Y_{A\Leftarrow 0} = Y_{A\Leftarrow 1\|\pi} = 0$ or $Y_{A\Leftarrow 0} = Y_{A\Leftarrow 1\|\pi} = 1$), regardless of the values of input features $\boldsymbol{X}$. Since this condition always implies the zero conditional mean unfair effect, it is sufficient to guarantee the individual-level fairness condition (Definition 9). It is also a more severe condition than Definition 9 since under the latter condition, potential outcomes $Y_{A\Leftarrow 0} = Y_{A\Leftarrow 1\|\pi}$ can take 0 or 1 depending on $\boldsymbol{X}$'s values. With such a severe fairness condition, prediction accuracy might decrease. Nevertheless, in Section 5.6.2, we experimentally show that our method can achieve comparable accuracy to the existing method for ensuring individual-level fairness (i.e., the PSCF method [Chiappa and Gillam, 2019]).

5.3.2 Achieving Individual-Level Fairness with PIU

To make potential outcomes take the same value for all individuals, we formulate penalty function G_θ in Eq. (5.1) based on the following quantity:

Definition 10. *Let π be the unfair pathways in a causal graph. For potential outcomes $Y_{A\Leftarrow 0}, Y_{A\Leftarrow 1\|\pi} \in \{0, 1\}$, we define the* ***probability of individual unfairness (PIU)*** *by* $\mathrm{P}(Y_{A\Leftarrow 0} \neq Y_{A\Leftarrow 1\|\pi})$.

PIU is the probability that potential outcomes $Y_{A\Leftarrow 0}$ and $Y_{A\Leftarrow 1\|\pi}$ take different values. Unlike the conditional mean unfair effect in Definition 9, PIU is not conditioned on features $\boldsymbol{X}$ of each individual. Nonetheless, PIU can be used to guarantee individual-level fairness. By constraining PIU to zero, we can guarantee that potential outcomes take the same value (i.e., $Y_{A\Leftarrow 0} = Y_{A\Leftarrow 1\|\pi} = 0$ or $Y_{A\Leftarrow 0} = Y_{A\Leftarrow 1\|\pi} = 1$)

with probability 1 regardless of the values of $\boldsymbol{X}$, and with such potential outcome values, we can ensure individual-level fairness.

Unfortunately, we cannot directly impose constraints on PIU. This is because estimating a PIU value requires the joint distribution of $Y_{A\Leftarrow 0}$ and $Y_{A\Leftarrow 1\|\pi}$, which is unavailable, as described in Section 5.2.4. To overcome this issue, instead of PIU, we utilize its upper bound that can be estimated from data. In particular, we formulate a penalty function that forces the upper bound on PIU to be nearly zero.

5.3.3 Penalty by Upper Bound on PIU

Deriving Upper Bound on PIU

To reduce the PIU value, we utilize the following upper bound on PIU:

Theorem 2 (Upper bound on PIU)**.** *Suppose that potential outcomes $Y_{A\Leftarrow 0}$ and $Y_{A\Leftarrow 1\|\pi}$ are binary (i.e., $Y_{A\Leftarrow 0}, Y_{A\Leftarrow 1\|\pi} \in \{0,1\}$). Then for any joint distribution of potential outcomes $\mathrm{P}(Y_{A\Leftarrow 0}, Y_{A\Leftarrow 1\|\pi})$, PIU is upper bounded as follows:*

$$\mathrm{P}(Y_{A\Leftarrow 0} \neq Y_{A\Leftarrow 1\|\pi}) \leq 2\,\mathrm{P}^I(Y_{A\Leftarrow 0} \neq Y_{A\Leftarrow 1\|\pi}), \tag{5.10}$$

where $\mathrm{P}^I(Y_{A\Leftarrow 0}, Y_{A\Leftarrow 1\|\pi}) = \mathrm{P}(Y_{A\Leftarrow 0})\,\mathrm{P}(Y_{A\Leftarrow 1\|\pi})$ is an independent joint distribution; for binary potential outcomes $Y_{A\Leftarrow 0}, Y_{A\Leftarrow 1\|\pi} \in \{0,1\}$, upper bound $2\,\mathrm{P}^I(Y_{A\Leftarrow 0} \neq Y_{A\Leftarrow 1\|\pi})$ is given as

$$\begin{aligned} &2\,\mathrm{P}^I(Y_{A\Leftarrow 0} \neq Y_{A\Leftarrow 1\|\pi}) \\ =&2(\mathrm{P}(Y_{A\Leftarrow 1\|\pi}=1)(1-\mathrm{P}(Y_{A\Leftarrow 0}=1)) + (1-\mathrm{P}(Y_{A\Leftarrow 1\|\pi}=1))\,\mathrm{P}(Y_{A\Leftarrow 0}=1)). \end{aligned} \tag{5.11}$$

The proof is described in Section 5.8.1. Theorem 2 states that whatever joint distribution potential outcomes $Y_{A\Leftarrow 0}$ and $Y_{A\Leftarrow 1\|\pi}$ follow, the resulting PIU value is at most twice the PIU value that is approximated with independent joint distribution P^I. The equality in (5.10) holds when the joint probability of satisfying $Y_{A\Leftarrow 0} \neq Y_{A\Leftarrow 1\|\pi}$ is zero; that is, when the marginal probabilities are $\mathrm{P}(Y_{A\Leftarrow 0}=1) = \mathrm{P}(Y_{A\Leftarrow 1\|\pi}=1) = 0$ or $\mathrm{P}(Y_{A\Leftarrow 0}=1) = \mathrm{P}(Y_{A\Leftarrow 1\|\pi}=1) = 1$.

Note that this upper bound can exceed 1; if so, the PIU value is not controlled because PIU is at most 1. However, by making the upper bound close to zero, we can guarantee that PIU is also close to zero.

Compared with the existing bounds discussed below, our upper bound has the following two advantages:

- It can deal with binary potential outcomes $Y_{A\Leftarrow 0}, Y_{A\Leftarrow 1\|\pi} \in \{0,1\}$. If we consider continuous potential outcomes, we can use several bounds on a functional of the joint distribution of potential outcomes [Fan *et al.*, 2017; Firpo and Ridder, 2019] because PIU is also such a functional. However, these existing bounds cannot be used for binary potential outcomes.

- It is much tighter than the existing result derived by Rubinstein and Singla [2017], which can also be used to derive an upper bound on PIU for binary potential outcomes. This existing result focuses on a function that takes as input random variables whose joint distribution is unavailable and obtains a *correlation gap* [Agrawal *et al.*, 2010], i.e., the worst-case ratio between two expected function values, each of whose expectation is taken with respect to an arbitrary joint distribution and an independent joint distribution that has the same marginal distributions with the former joint distribution. Since PIU can also be written as such an expected function value, i.e., $\mathbb{E}_{Y_{A\Leftarrow 0}, Y_{A\Leftarrow 1\|\pi}}[\mathbf{I}(Y_{A\Leftarrow 0} \neq Y_{A\Leftarrow 1\|\pi})]$, where $\mathbf{I}$ is an indicator function, we can apply this result to PIU to obtain its upper bound using independent joint distribution P^I in (5.10). However, the bound obtained with this existing result is much looser than ours. Although the multiplicative constant in (5.10) is 2, this value becomes 200 with the bound of Rubinstein and Singla [2017]. With such a loose upper bound, we need to impose an excessively severe penalty on it to ensure that PIU is close to zero.

Thanks to the first advantage, we can address the problem of learning a fair binary classifier where unfairness is measured using binary potential outcomes. Moreover, owing to the second one, we can avoid imposing an excessively severe penalty, thus preventing an unnecessary decrease in prediction accuracy.

Estimating Upper Bound

We estimate the upper bound on PIU in (5.10) (i.e., $2\,\mathrm{P}^I(Y_{A\Leftarrow 0} \neq Y_{A\Leftarrow 1\|\pi})$), which is twice the PIU value that is approximated using independent joint distribution P^I.

This approximated PIU value is given as follows:

$$\begin{aligned} &\mathrm{P}^{I}(Y_{A\Leftarrow 0} \neq Y_{A\Leftarrow 1\|\pi}) \\ =&\,\mathrm{P}(Y_{A\Leftarrow 1\|\pi}=1)(1-\mathrm{P}(Y_{A\Leftarrow 0}=1)) + (1-\mathrm{P}(Y_{A\Leftarrow 1\|\pi}=1))\,\mathrm{P}(Y_{A\Leftarrow 0}=1). \end{aligned} \tag{5.12}$$

To estimate marginal probabilities $\mathrm{P}(Y_{A\Leftarrow 0}=1)$ and $\mathrm{P}(Y_{A\Leftarrow 1\|\pi}=1)$ in Eq. (5.12), we make the two standard assumptions, both of which are widely used in the existing methods [Chiappa and Gillam, 2019; Nabi and Shpitser, 2018; Zhang and Wu, 2017; Zhang *et al.*, 2017].

One is an assumption on unfair pathways π, which is expressed using the following graphical condition called the *recanting witness criterion*:

Definition 11 (Recanting witness criterion [Avin *et al.*, 2005])**.** *Given pathways π, let Z be a node in a causal graph that satisfies the following:*

1. *There is a pathway from A to Z ($A \to \cdots \to Z$) in π.*
2. *There is a pathway from Z to Y ($Z \to \cdots \to Y$) in π.*
3. *There is another pathway from Z to Y ($Z \to \cdots \to Y$) that is in the causal graph but not in π.*

Then pathways π satisfy the recanting witness criterion with node Z, which is called a witness.

For example, consider the causal graph in Figure 5.3(a), where the unfair pathway is $\pi = \{A \to M_1 \to M_2 \to Y\}$. Clearly, pathway π satisfies the recanting witness criterion with witness M_1.

Avin *et al.* [2005] show that estimating marginal probabilities $\mathrm{P}(Y_{A\Leftarrow 0}=1)$ and $\mathrm{P}(Y_{A\Leftarrow 1\|\pi}=1)$ requires the assumption that there is no witness node; in other words,

Assumption 9. *Pathways π do* **not** *satisfy the recanting witness criterion.*

The other is a common assumption in causal inference called *sequential ignorability*, which requires conditional independence relations between variables.

We formulate this assumption based on the estimators in a previous work [Huber, 2014], which we use in our method, presented in Eq. (5.18). As an example, we show a formulation based on the causal graph in Figure 5.4 (see Huber [2014] for details).

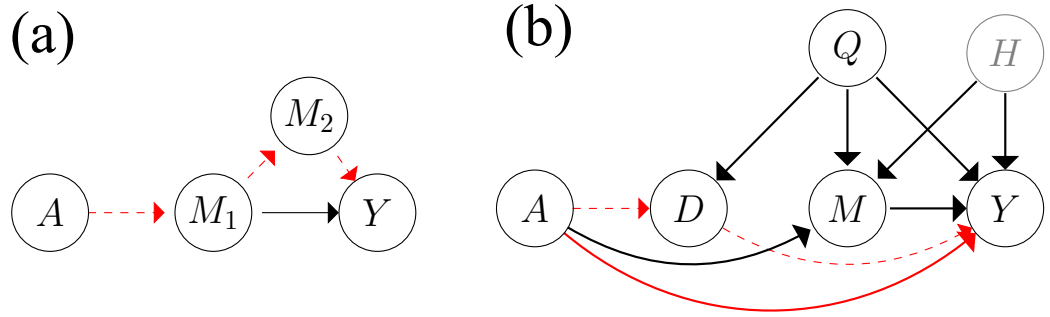

Figure 5.3: Two causal graphs that violate Assumptions 9 and 10: Unfair pathways (a): $\pi = \{A \to M_1 \to M_2 \to Y\}$ and (b): $\pi = \{A \to Y, A \to D \to Y\}$.

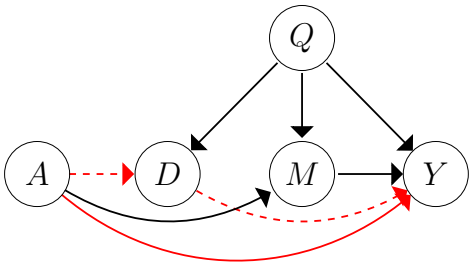

Figure 5.4: Causal graph example for illustration of our assumptions (Same graph structure with Figure 5.2(c))

In what follows, we use the same notations as those in the original paper [Huber, 2014]. Let potential outcomes $Y_{A \Leftarrow 0}$ and $Y_{A \Leftarrow 1 \| \pi}$ denote

$$Y_{A \Leftarrow 0} = Y(0, D(0), M(0)) \quad \text{and} \quad Y_{A \Leftarrow 1 \| \pi} = Y(1, D(1), M(0)), \tag{5.13}$$

respectively, where $D(0)$, $D(1)$, and $M(0)$ express counterfactual attributes formulated by Eq. (5.3). Then the sequential ignorability is expressed as follows:

Assumption 10 (Sequential ignorability)**.** *For all $a, a', a'' \in \{0, 1\}$ and d, m, q in the supports of D, M, and Q, the following four relations hold:*

$$\{Y(a, d, m), D(a'), M(a'')\} \perp\!\!\!\perp A \mid Q = q \tag{5.14}$$

$$Y(a', d, m) \perp\!\!\!\perp D \mid A = a, Q = q \tag{5.15}$$

$$Y(a', d, m) \perp\!\!\!\perp M \mid A = a, Q = q \tag{5.16}$$

$$\mathrm{P}(A = a \mid Q = q, D = d, M = m) > 0. \tag{5.17}$$

To express the potential outcome distributions using the observed data, three relations, (5.14), (5.15), and (5.16) in Assumption 10 are needed. In particular, relations (5.15) and (5.16), which are often called *cross-world independence assumptions* [Andrews and Didelez, 2020], require the exogenous variables in the structural equations to be mutually independent. As argued by Huber [2014], such independence relations between exogenous variables are violated if there is an unobserved variable called a latent confounder, which is an unobserved parent of the observed variables. For instance, when the causal graph in Figure 5.3(b) is given, the aforementioned relations do not hold due to a latent confounder, H (gray node).[3] However, even in the presence of latent confounders, in some cases, our method can achieve individual-level fairness using an extended penalty function, as described in Section 5.5.1. The last relation (5.17), which corresponds to the positivity assumption (Assumption 2) in Section 2.2.3, is used to avoid a division by zero.

There are various estimators that are founded on these assumptions. Among them, we utilize the computationally efficient estimator in Huber [2014], which can be computed in $O(n)$ time, where n is the number of training instances.

This estimator is formulated as a weighted average of conditional probabilities; this estimation technique is widely used and called *inverse probability weighting* (IPW). Let $c_\theta(\boldsymbol{X}) = \mathrm{P}(Y = 1|\boldsymbol{X})$ denote the conditional distribution provided by classifier h_θ; we let $c_\theta(\boldsymbol{X}) = h_\theta(\boldsymbol{X}) \in \{0, 1\}$ if h_θ is a deterministic classifier. Suppose that training data include the feature attributes of n individuals $\{\boldsymbol{x}_i\}_{i=1}^n$ ($i \in \{1, \ldots, n\}$), each of which contain sensitive attribute a_i. Then $\mathrm{P}(Y_{A \Leftarrow 0} = 1)$ and $\mathrm{P}(Y_{A \Leftarrow 1\|\pi} = 1)$ can be estimated as the following weighted averages of c_θ over individuals with $A = 0$ and $A = 1$, respectively:

$$\begin{aligned} \hat{p}_\theta^{A \Leftarrow 0} &= \frac{1}{n} \sum_{i=1}^{n} \mathbf{I}(a_i = 0) \hat{w}_i^{A \Leftarrow 0} c_\theta(\boldsymbol{x}_i) \quad \text{and} \\ \hat{p}_\theta^{A \Leftarrow 1\|\pi} &= \frac{1}{n} \sum_{i=1}^{n} \mathbf{I}(a_i = 1) \hat{w}_i^{A \Leftarrow 1\|\pi} c_\theta(\boldsymbol{x}_i), \end{aligned} \tag{5.18}$$

where $\mathbf{I}(\cdot)$ is an indicator function, and $\hat{w}_i^{A \Leftarrow 0}$ and $\hat{w}_i^{A \Leftarrow 1\|\pi}$ are non-negative weights. According to Huber [2014], weights $\hat{w}_i^{A \Leftarrow 0}$ and $\hat{w}_i^{A \Leftarrow 1\|\pi}$ are formulated in different

[3] Since variable H influences mediator M and outcome Y, it is also called a mediator-outcome confounder [VanderWeele, 2015, Section 5].

ways, depending on the causal graph structure and unfair pathways π. For instance, in case of **Example 2** in Section 5.2.2, letting the features of n individuals be $\{\boldsymbol{x}_i\}_{i=1}^n = \{a_i, q_i, d_i, m_i\}_{i=1}^n$, these weights are formulated as follows:

$$\begin{aligned} \hat{w}_i^{A\Leftarrow 0} &= \frac{1}{\hat{\mathrm{P}}(A=0|q_i)}, \\ \hat{w}_i^{A\Leftarrow 1\|\pi} &= \frac{\hat{\mathrm{P}}(A=1|q_i,d_i)\hat{\mathrm{P}}(A=0|q_i,d_i,m_i)}{\hat{\mathrm{P}}(A=1|q_i)\hat{\mathrm{P}}(A=0|q_i,d_i)\hat{\mathrm{P}}(A=1|q_i,d_i,m_i)}, \end{aligned} \tag{5.19}$$

where $\hat{\mathrm{P}}$ is a conditional distribution, which we infer in the same way as Zhang and Bareinboim [2018a], i.e., by learning a statistical model (e.g., a neural network) from the training data beforehand.[4] We derive the formulations of the weighted estimator (5.18) in Section 5.8.2.

Using the weighted estimators in Eq. (5.18), we formulate penalty function G_θ in our objective function in Eq. (5.1) as

$$G_\theta(\boldsymbol{x}_1, \dots \boldsymbol{x}_n) = \hat{p}_\theta^{A\Leftarrow 1\|\pi}(1 - \hat{p}_\theta^{A\Leftarrow 0}) + (1 - \hat{p}_\theta^{A\Leftarrow 1\|\pi})\hat{p}_\theta^{A\Leftarrow 0}. \tag{5.20}$$

In our experiments, we minimize the objective function using the stochastic gradient descent method [Sutskever *et al.*, 2013].

From the penalty function in Eq. (5.20), we can see why penalizing the upper bound on PIU guarantees individual-level fairness. As the penalty parameter value goes to infinity ($\lambda \to \infty$), the marginal probabilities $(\hat{p}_\theta^{A\Leftarrow 0}, \hat{p}_\theta^{A\Leftarrow 1\|\pi})$ approach $(0, 0)$ or $(1, 1)$. This guarantees that the potential outcomes take the same value with probability 1, which is sufficient to guarantee individual-level fairness.

In practice, we use a finite penalty parameter value $\lambda < \infty$. With such a penalty parameter value, our method reduces the value of penalty function G_θ to close to zero by controlling the predictions for individuals.

Controlling Predictions with Penalty Function

To reduce the value of penalty function G_θ, our method forces the marginal potential outcome probabilities to be $(\hat{p}_\theta^{A\Leftarrow 0}, \hat{p}_\theta^{A\Leftarrow 1\|\pi}) \approx (0, 0)$ or $(1, 1)$ by controlling the

[4]Note that the FIO method infers such conditional distributions not by learning statistical models beforehand but by simultaneously learning them with the predictive model of Y [Nabi and Shpitser, 2018]. This is because unlike our method, it addresses not only training a classifier but also learning a generative model of joint distribution $\mathrm{P}(\mathbf{X}, Y)$.

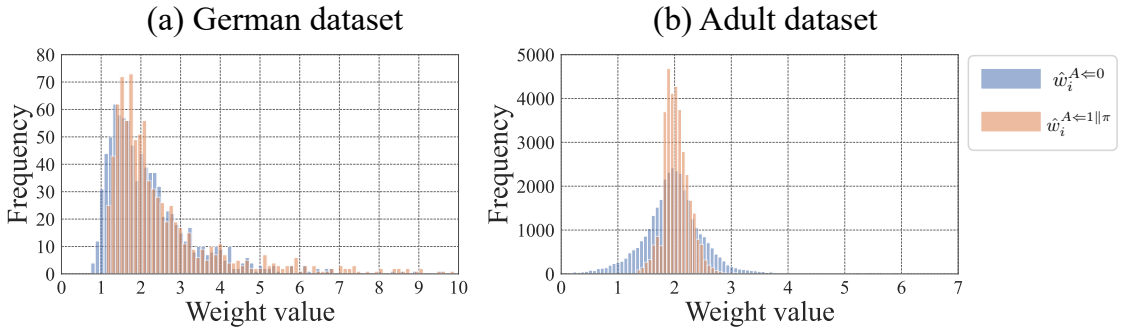

Figure 5.5: Histograms of weight values computed with two real-world datasets: (a) German dataset and (b) Adult dataset. See Section 5.6.1 for details of these datasets.

predictions for individuals.

In fact, our method can effectively adjust the control on the predictions by utilizing weighted estimators $\hat{p}_\theta^{A \Leftarrow 0}$ and $\hat{p}_\theta^{A \Leftarrow 1 \| \pi}$ in Eq. (5.18). With such weighted estimators, we do not have to push conditional probability $c_\theta(\boldsymbol{X})$ into a constant (i.e., $c_\theta(\boldsymbol{X}) = 0$ or $c_\theta(\boldsymbol{X}) = 1$); instead, we only have to impose strong penalties on the predictions for individual $i \in \{1, \ldots, n\}$ whose weight values $\hat{w}_i^{A \Leftarrow 0}$ and $\hat{w}_i^{A \Leftarrow 1 \| \pi}$ are large. These weight values take different values depending on individuals, as illustrated by the histograms computed with the two real-world datasets in Figure 5.5. The weight values depend on the (estimated) conditional distributions of sensitive feature A conditioned on a subset of features $\boldsymbol{X}$ (in case of (5.19), $\hat{\mathrm{P}}(A \mid q_i)$, $\hat{\mathrm{P}}(A \mid q_i, d_i)$, and $\hat{\mathrm{P}}(A \mid q_i, d_i, m_i)$). If such conditional distributions are extremely skewed (i.e., $A = 0$ or $A = 1$ is exceedingly concentrated), the weights take extreme values. Note that these weights are used to accurately infer the marginal potential outcome probabilities and that their values do not indicate how strongly each individual will suffer from discrimination.

5.4 Comparison with Existing Fairness Constraint

To show the effectiveness of our penalty function G_θ in Eq. (5.20), we compare it with the constraint of the FIO method [Nabi and Shpitser, 2018]. Both our penalty function and the FIO constraint are formulated using marginal potential outcome probabilities $\hat{p}_\theta^{A \Leftarrow 0}$ and $\hat{p}_\theta^{A \Leftarrow 1 \| \pi}$. However, our penalty function enables us to achieve individual-level fairness while the FIO constraint does not. We elucidate this difference from a viewpoint of feasible regions in optimization problems.

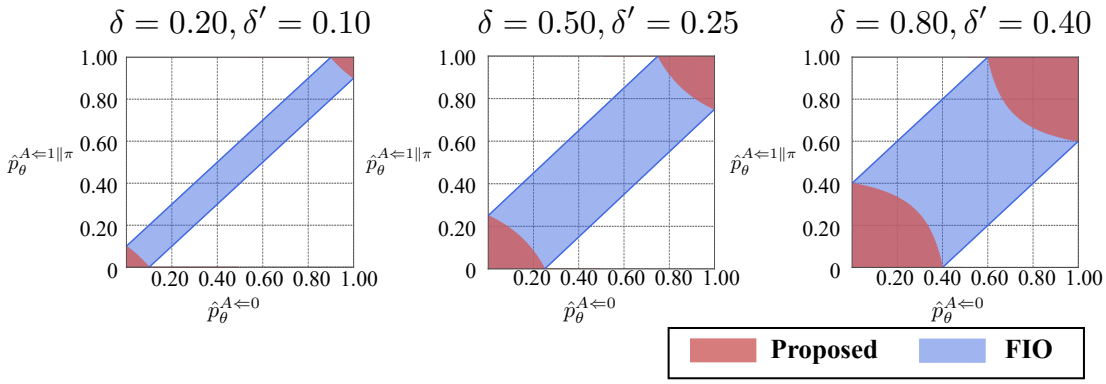

Figure 5.6: Feasible regions of our constraint (red) and FIO (blue) with $(\delta,\ \delta') = (0.20,\ 0.10),\ (0.50,\ 0.25),\ (0.80,\ 0.40)$

Suppose that our penalty function forces the upper bound on PIU to satisfy the following condition:

$$\hat{p}_\theta^{A\Leftarrow 1\|\pi}(1-\hat{p}_\theta^{A\Leftarrow 0}) + (1-\hat{p}_\theta^{A\Leftarrow 1\|\pi})\hat{p}_\theta^{A\Leftarrow 0} \leq \delta. \tag{5.21}$$

Here we let constant δ be $\delta \in [0,1]$ because otherwise we cannot force the PIU value to be less than 1.

The FIO constraint limits the mean unfair effect (5.7) to be in

$$-\delta' \leq \hat{p}_\theta^{A\Leftarrow 1\|\pi} - \hat{p}_\theta^{A\Leftarrow 0} \leq \delta', \tag{5.22}$$

where $\delta' \in [0,1]$ is a hyperparameter. If $\delta' = 0$, it ensures $\hat{p}_\theta^{A\Leftarrow 0} = \hat{p}_\theta^{A\Leftarrow 1\|\pi}$.

Figure 5.6 shows the feasible region of our fairness condition (red) and the FIO constraint (blue), respectively, obtained by graphing the hyperbolic inequality (5.21) and the linear inequality (5.22). Here, to clarify their difference, we consider the case where $\delta = 2\delta'$.

While our fairness condition with $\delta \approx 0$ only accepts region $(\hat{p}_\theta^{A\Leftarrow 0}, \hat{p}_\theta^{A\Leftarrow 1\|\pi}) \approx (0,0)$ or $(1,1)$,[5] FIO always accepts point $(\hat{p}_\theta^{A\Leftarrow 0}, \hat{p}_\theta^{A\Leftarrow 1\|\pi}) = (0.5, 0.5)$ with any δ'.

Due to the differences of these feasible regions, possible PIU values are largely different between two methods. To illustrate this, in what follows, we formulate the

[5]Obviously, this region is wider than red subregions $(\hat{p}_\theta^{A\Leftarrow 0}, \hat{p}_\theta^{A\Leftarrow 1\|\pi}) \approx (0,0)$ and $(\hat{p}_\theta^{A\Leftarrow 0}, \hat{p}_\theta^{A\Leftarrow 1\|\pi}) \approx (1,1)$. This indicates that compared with such naive constraints as $(\hat{p}_\theta^{A\Leftarrow 0}, \hat{p}_\theta^{A\Leftarrow 1\|\pi}) \approx (0,0)$ and $(\hat{p}_\theta^{A\Leftarrow 0}, \hat{p}_\theta^{A\Leftarrow 1\|\pi}) \approx (1,1)$, ours can accept more various values of classifier parameter θ to achieve high prediction accuracy.

lower and upper bounds on the PIU using marginal potential outcome probabilities. Let the (true) marginal probabilities of the potential outcomes be $\alpha = \mathrm{P}(Y_{A\Leftarrow 0} = 1)$ and $\beta = \mathrm{P}(Y_{A\Leftarrow 1\|\pi} = 1)$, and the (true) joint probabilities of $(Y_{A\Leftarrow 0}, Y_{A\Leftarrow 1\|\pi}) = (0,0)$, $(0,1)$, $(1,0)$, and $(1,1)$ be p_{00}, p_{01}, p_{10}, and p_{11}. With these notations, PIU can be formulated as $p_{01} + p_{10}$, and its lower and upper bounds can be expressed using marginal probabilities α and β as

$$|\alpha - \beta| \leq p_{01} + p_{10} \leq \min\{\beta, 1-\alpha\} + \min\{\alpha, 1-\beta\}, \tag{5.23}$$

which we prove in Section 5.8.3.

When the marginal probabilities are forced to be $(\alpha, \beta) \approx (0,0)$ or $(1,1)$, since the upper bound in Eq. (5.23) become close to zero, PIU is constrained to almost zero (i.e., $p_{01} + p_{10} \approx 0$). Hence, with our method, the unfair effect is likely to be zero, demonstrating how effectively our condition guarantees fairness for each individual.

By contrast, at point $(\alpha, \beta) = (0.5, 0.5)$, the lower and upper bounds in Eq. (5.23) become 0 and 1, respectively: $0 \leq p_{01} + p_{10} \leq 1$. This implies that with FIO, it is completely uncertain whether the PIU value is high since the joint probabilities are unknown in practice. Therefore, FIO cannot ensure that the unfair effect is zero for each individual, which is insufficient to guarantee individual-level fairness.

5.5 Extensions for Complex Real-World Scenarios

So far, we have made the following two assumptions:

- The marginal probabilities of potential outcomes can be estimated from data.
- The causal graph structure is given as input.

However, these assumptions might not hold in complex real-world scenarios. The former is not satisfied if there is a latent confounder [Pearl, 2009] (i.e., an unobserved variable that influences two (or more) observed variables and yields a spurious correlation), and the latter does not hold if domain experts do not know the true causal relationships between variables and if they cannot be inferred from data.

Although addressing such complications is extremely challenging, in some cases, we can achieve individual-level fairness by applying the extensions described below.

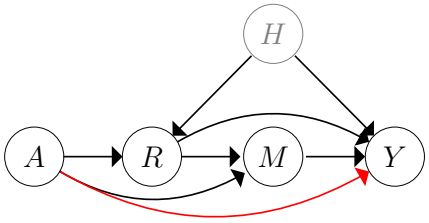

Figure 5.7: Example of causal graph containing latent confounder H (gray node), which affects R and prediction Y. Red pathway represents unfair pathway π.

5.5.1 Dealing with Latent Confounders

The presence of latent confounders yields a cumbersome spurious correlation between sensitive feature and decision outcome. For this reason, it becomes much more challenging to estimate marginal potential outcome probabilities.

However, even in such cases, our method can ensure individual-level fairness if the lower and upper bounds on marginal probabilities are available. Suppose that marginal probabilities $\mathrm{P}(Y_{A\Leftarrow 0} = 1)$ and $\mathrm{P}(Y_{A\Leftarrow 1\|\pi} = 1)$ are bounded by

$$\hat{l}_\theta^{A\Leftarrow 0} \leq \mathrm{P}(Y_{A\Leftarrow 0} = 1) \leq \hat{u}_\theta^{A\Leftarrow 0},$$
$$\hat{l}_\theta^{A\Leftarrow 1\|\pi} \leq \mathrm{P}(Y_{A\Leftarrow 1\|\pi} = 1) \leq \hat{u}_\theta^{A\Leftarrow 1\|\pi},$$

where $\hat{l}_\theta^{A\Leftarrow 0}$, $\hat{u}_\theta^{A\Leftarrow 0}$, $\hat{l}_\theta^{A\Leftarrow 1\|\pi}$, and $\hat{u}_\theta^{A\Leftarrow 1\|\pi}$ are the estimated lower and upper bounds. Then for any marginal probability values, the upper bound on PIU in Eq. (5.10) is always at most twice the value of

$$G_\theta(\boldsymbol{x}_1, \ldots \boldsymbol{x}_n) = \hat{u}_\theta^{A\Leftarrow 1\|\pi}(1 - \hat{l}_\theta^{A\Leftarrow 0}) + (1 - \hat{l}_\theta^{A\Leftarrow 1\|\pi})\hat{u}_\theta^{A\Leftarrow 0}. \tag{5.24}$$

Therefore, by making this penalty function value nearly zero, we can achieve individual-level fairness.

To formulate the penalty function in Eq. (5.24), several existing lower and upper bounds can be used [Robins and Richardson, 2010; Tchetgen and Phiri, 2014; Miles *et al.*, 2017]. For instance, the result in Miles *et al.* [2017] can be used when a feature affected by sensitive feature A, which is called a mediator, takes discrete values. An example causal graph structure is shown in Figure 5.7, where M is a mediator, R is an observed confounder,[6] and H is a latent confounder (a.k.a. a mediator-outcome

[6]More precisely, R is an *exposure-induced confounder* [VanderWeele, 2015, Chapter 5], i.e., an

confounder [VanderWeele, 2015, Section 5]). Under this causal graph, conditional ignorability in Assumption 10 does not hold due to the presence of latent confounder H; hence, we cannot represent the marginal potential outcome probabilities with the observed variable distributions. However, by applying the result of Miles *et al.* [2017], lower and upper bounds on $\mathrm{P}(Y_{A \Leftarrow 1 \| \pi} = 1)$ can be expressed as[7]

$$\hat{l}^{A \Leftarrow 1 \| \pi} = \sum_m \max\{0,\ \mathrm{P}(M = m | A = 0) - 1 + \sum_r \mathrm{P}(Y = 1 | A = 1, m, r)\, \mathrm{P}(R = r | A = 1)\},$$

$$\hat{u}^{A \Leftarrow 1 \| \pi} = \sum_m \min\{\mathrm{P}(M = m | A = 0),\ \sum_r \mathrm{P}(Y = 1 | A = 1, m, r)\, \mathrm{P}(R = r | A = 1)\}.$$

With conditional distribution $c_\theta(1, M, R) = \mathrm{P}(Y = 1 | A = 1, M, R)$ provided by classifier h_θ, these bounds can be expressed as the functions of parameter θ:

$$\hat{l}_\theta^{A \Leftarrow 1 \| \pi} = \sum_m \max\{0,\ \hat{\mathrm{P}}(M = m | A = 0) - 1 + \sum_r c_\theta(1, m, r) \hat{\mathrm{P}}(R = r | A = 1)\} \tag{5.25}$$

$$\hat{u}_\theta^{A \Leftarrow 1 \| \pi} = \sum_m \min\{\hat{\mathrm{P}}(M = m | A = 0),\ \sum_r c_\theta(1, m, r) \hat{\mathrm{P}}(R = r | A = 1)\}. \tag{5.26}$$

Here conditional distributions $\hat{\mathrm{P}}(M = m | A = 0)$ and $\hat{\mathrm{P}}(R = r | A = 1)$ can be estimated by learning statistical models (e.g., logistic regression or neural networks) from the training data beforehand. As with Eqs. (5.25) and (5.26), we can formulate the estimated lower and upper bounds on marginal probability $\mathrm{P}(Y_{A \Leftarrow 0} = 1)$ as follows:

$$\hat{l}_\theta^{A \Leftarrow 0} = \sum_m \max\{0,\ \hat{\mathrm{P}}(M = m | A = 0) - 1 + \sum_r c_\theta(0, m, r) \hat{\mathrm{P}}(R = r | A = 0)\} \tag{5.27}$$

$$\hat{u}_\theta^{A \Leftarrow 0} = \sum_m \min\{\hat{\mathrm{P}}(M = m | A = 0),\ \sum_r c_\theta(0, m, r) \hat{\mathrm{P}}(R = r | A = 0)\}. \tag{5.28}$$

By representing the lower and upper bounds with classifier parameter θ in this

observed variable that is affected by sensitive feature A and that influences multiple observed variables. An exposure-induced confounder is also a mediator. Unlike a mediator, however, it yields a spurious correlation among the observed variables.

[7] According to Miles *et al.* [2017], the lower and upper bounds coincide when the potential outcome and the potential mediator are degenerate.

way, we can formulate a penalty function in Eq. (5.24), which allows us to achieve fairness for each individual in the presence of latent confounders.

5.5.2 Addressing Uncertain Causal Graphs

Even without an unobserved confounder, it might be difficult in practice to depict the true causal graph structure. If domain experts cannot depict it, we need to infer this causal graph structure from the data. It is, however, a challenging task without making functional assumptions on the underlying SEM [Glymour *et al.*, 2019].

To alleviate this issue, we extend our learning framework so that it can take as input multiple candidates of causal graphs with unfair pathways. This idea is inspired by the Multi-World Fairness (MWF) algorithm in Russell *et al.* [2017], which achieves fairness based on multiple SEM candidates. MWF expresses an unfairness function based on each SEM and takes the sum of the unfairness functions to formulate a penalty function. Unfortunately, formulating an SEM candidate requires not only a causal graph candidate but also a formulation of data generating processes, which is unrealistic in practice. Therefore instead of using SEMs, we utilize multiple causal graph structures, each of which contains unfair pathways.

Let the number of causal graph candidates be $n_G > 1$. For j-th causal graph ($j \in \{1, \ldots, n_G\}$), suppose that marginal potential outcome probabilities are estimated as $\hat{p}_\theta^{A \Leftarrow 0(j)}$ and $\hat{p}_\theta^{A \Leftarrow 1 \| \pi(j)}$, which are differently formulated, depending on the causal graph structure (e.g., which variables are confounders, which pathways are unfair, etc). Using n_G marginal probability pairs, we formulate our penalty function as

$$G_\theta(\boldsymbol{x}_1, \ldots \boldsymbol{x}_n) = \frac{1}{n_G} \sum_{j=1}^{n_G} G_\theta^{(j)}(\boldsymbol{x}_1, \ldots \boldsymbol{x}_n), \tag{5.29}$$

where $G_\theta^{(j)}$ is the following function that measures unfairness based on j-th causal graph, expressed as:

$$G_\theta^{(j)}(\boldsymbol{x}_1, \ldots \boldsymbol{x}_n) = \hat{p}_\theta^{A \Leftarrow 1 \| \pi(j)}(1 - \hat{p}_\theta^{A \Leftarrow 0(j)}) + (1 - \hat{p}_\theta^{A \Leftarrow 1 \| \pi(j)})\hat{p}_\theta^{A \Leftarrow 0(j)}. \tag{5.30}$$

Penalty function G_θ depends on the weights in estimators $\hat{p}_\theta^{A \Leftarrow 0(j)}$ and $\hat{p}_\theta^{A \Leftarrow 1 \| \pi(j)}$ ($j \in \{1, \ldots, n_G\}$), whose weight values are assigned to each individual.

Lowering the value of G_θ in (5.29) implies imposing the average of penalty func-

tion $G_\theta^{(j)}$ over causal graph candidate $j = 1, \ldots, n_G$. In Section 5.6.3, we experimentally confirm that this average-based penalty function allows us to make fairer predictions than the original one in (5.20).

5.6 Experiments

Through synthetic and real-world data experiments, we evaluated the performance of our method. This section is organized as follows. In Section 5.6.1, we describe the experimental settings. In Section 5.6.2, we present the performance of our learning framework described in Section 5.3. Finally, in Section 5.6.3, we show the effectiveness of our extended framework presented in Section 5.5.

5.6.1 Experimental Settings

Baseline Methods

We compared the performance of our method with the following four baselines:

1. **FIO** [Nabi and Shpitser, 2018], which aims to reduce the mean unfair effect in Definition 7;

2. **PSCF** [Chiappa and Gillam, 2019], which can achieve individual-level fairness if the data are generated by additive noise models;

3. **Unconstrained**, which does not use any constraints or penalty terms related to fairness;

4. **Remove** [Kusner *et al.*, 2017, Section S4], which removes unfair effects simply by making predictions without input features that correspond to the nodes on unfair pathways π.

We show the performance of each method when using the two classifiers: a feed-forward neural network and logistic regression. The feed-forward neural network is composed of two linear layers with 100 and 50 hidden neurons, whose activation function, an output layer, and loss function are given by a sigmoid function, a log softmax function, and cross-entropy loss, respectively.

We train these classifiers using the stochastic gradient descent method [Sutskever *et al.*, 2013]. We set the minibatch size to 100 or 1,000; with the German dataset, since the number of training samples is less than 1,000, we set the minibatch size to 100, and with the other datasets, we set it to 1,000. We stopped the training after 1,000 epochs. The penalty parameter value is selected using a grid search with a 0.25 grid size from $[0.0, 5.0]$.

Data and Causal Graphs

For a performance evaluation, we used synthetic and real-world datasets, whose characteristics are summarized in Table 5.2.

We sampled four synthetic datasets (Synth1, Synth2, Synth3, and Synth4) from different SEMs.

Synth1 dataset is generated from the following SEM:

$$\begin{aligned} A &= U_A, \quad U_A \sim \text{Bernoulli}(0.6), \\ Q &= \lfloor U_Q \rfloor, \quad U_Q \sim \mathcal{N}(2, 5^2), \\ D &= A + \lfloor 0.5QU_D \rfloor, \quad U_D \sim \text{Tr}\mathcal{N}(1, 0.5^2, 0.1, 3.0), \\ M &= 3A + 0.4QU_M, \quad U_M \sim \text{Tr}\mathcal{N}(1.5, 0.5^2, 0.1, 3.0), \\ Y &= h(A, Q, D, M), \end{aligned} \tag{5.31}$$

where Bernoulli, $\mathcal{N}$, and $\text{Tr}\mathcal{N}$ represent the Bernoulli, Gaussian, and truncated Gaussian distributions, respectively, and $\lfloor \cdot \rfloor$ is a floor function that returns an integer by removing the decimal places. To output outcome Y, we used function h, which is a logistic regression model that provides the following conditional distribution:

$$\text{P}(Y = 1 \mid A, Q, D, M) = \text{Bernoulli}(\varsigma(-10 + 5A + Q + D + M)),$$

where $\varsigma(x) = 1/(1 + \exp(-x))$ is a standard sigmoid function. Note that this SEM does not satisfy the functional assumption of the PSCF method because the structural equations over D and M are not expressed by additive noise models [Hoyer *et al.*, 2009] due to multiplicative noises U_D and U_M.

Synth2 dataset is sampled from the following SEM, which follows the functional

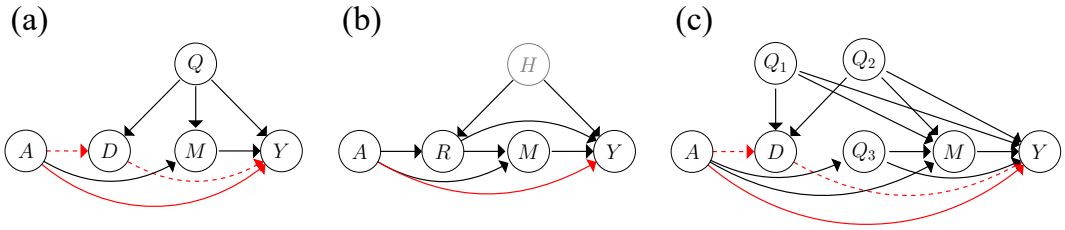

Figure 5.8: Causal graphs for four synthetic datasets: (a) Synth1 and Synth2 datasets, (b) Synth3 dataset, and (c) Synth4 dataset. Unfair pathways (a): $A \to Y$ (solid red edge) and $A \to D \to Y$ (dashed red pathway); (b): $A \to Y$ (solid red edge); (c): $A \to Y$ (solid red edge) and $A \to D \to Y$ (dashed red pathway).

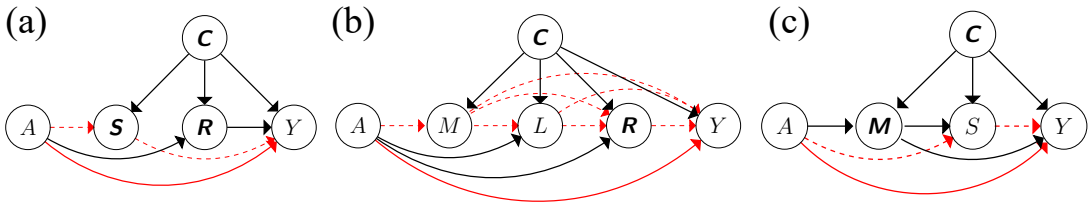

Figure 5.9: Causal graphs for real-world datasets: (a) German credit dataset, (b) Adult dataset, and (c) COMPAS dataset. Unfair pathways (a): direct pathway $A \to Y$ (solid red edge; A: gender; Y: predicted credit risk) and pathway via financial information $\boldsymbol{S}$, i.e., $A \to \boldsymbol{S} \to Y$ (dashed red pathway); (b): direct pathway $A \to Y$ (solid red edge; A: gender; Y: predicted income) and those that go via marital status M, i.e., $A \to M \to \cdots \to Y$ (dashed red pathways); (c) direct pathway $A \to Y$ (solid red edge; A: race; Y: predicted recidivism) and pathway via COMPAS score S, i.e., $A \to S \to Y$ (dashed red pathway).

assumption of the PSCF method:

$$\begin{aligned} A &= U_A, \quad U_A \sim \text{Bernoulli}(0.6), \\ Q &= \lfloor U_Q \rfloor, \quad U_Q \sim \mathcal{N}(5, 2.5^2), \\ D &= A + \lfloor 0.1Q + U_D \rfloor, \quad U_D \sim \lfloor \mathcal{N}(1, 0.5^2) \rfloor, \\ M &= 3A + 0.4Q + U_M, \quad U_M \sim \lfloor \mathcal{N}(1, 0.5^2) \rfloor, \\ Y &= h(A, Q, D, M), \end{aligned} \tag{5.32}$$

where function h is given by the following conditional distribution:

$$\text{P}(Y = 1 \mid A, Q, D, M) = \text{Bernoulli}(\varsigma(-10 + 2A + 2Q + 2D + 2M)).$$

Synth3 dataset is drawn from the following SEM, which contains latent con-

founder H:

$$\begin{aligned} &A = U_A, \quad U_A \sim \text{Bernoulli}(0.6), \\ &R = 3A + \lfloor 10H \rfloor + \lfloor U_R \rfloor, \quad U_R \sim \mathcal{N}(1, 0.5^2), \\ &M = A + R + \lfloor U_M \rfloor, \quad U_M \sim \mathcal{N}(1, 0.5^2), \\ &Y = h(A, R, M, H), \end{aligned} \tag{5.33}$$

where H denotes a latent confounder, which is sampled by $H \sim \mathcal{N}(1, 0.5^2)$, and function h is expressed by the following conditional distribution:

$$\text{P}(Y = 1 \mid A, R, M, H) = \text{Bernoulli}(\varsigma(-15 + 3A + R + M + 5H)).$$

Synth4 dataset is prepared using the following SEM:

$$\begin{aligned} &A = U_A, \quad U_A \sim \text{Bernoulli}(0.6), \\ &Q_1 = \lfloor U_{Q_1} \rfloor, \quad U_{Q_1} \sim \mathcal{N}(2, 1^2), \\ &Q_2 = \lfloor U_{Q_2} \rfloor, \quad U_{Q_2} \sim \mathcal{N}(2, 1^2), \\ &Q_3 = A + \lfloor U_{Q_3} \rfloor, \quad U_{Q_3} \sim \mathcal{N}(0, 1^2), \\ &D = A + \lfloor 0.1(Q_1 + Q_2)U_D \rfloor, \quad U_D \sim \mathcal{N}(1, 0.5^2) \\ &M = 2A + \lfloor 0.01\text{exp}(Q_1) + 0.2 \cdot (Q_2 + Q_3)) \rfloor + \lfloor U_M \rfloor, \quad U_M \sim \mathcal{N}(1, 1^2), \\ &Y = h(A, Q_1, Q_2, Q_3, D, M), \end{aligned} \tag{5.34}$$

where function h is given by the following conditional distribution:

$$\begin{aligned} &\text{P}(Y = 1 | A, Q_1, Q_2, Q_3, D, M) \\ &= \text{Bernoulli}(\varsigma(-5 + 2A + 0.5(Q_1 + Q_2 + Q_3) + 0.5D + 2M)). \end{aligned}$$

The SEMs in Eqs. (5.31), (5.32), (5.33), and (5.34) are associated with the causal graphs presented in Figure 5.8. Through all synthetic data experiments, we used $5,000$ samples to train the classifier and $1,000$ samples to test the performance.

In real-world data experiments, we used three datasets: the German credit dataset, the Adult dataset [Bache and Lichman, 2013], and the dataset about a risk assessment tool called Correctional Offender Management Profiling for Alterna-

Table 5.2: Dataset characteristics: number of training instances n_{tr}, number of test instances n_{te}, number of input features s, ratio $A = 1{:}A = 0$, and ratio $Y = 1{:}Y = 0$.

Data	n_{tr}	n_{te}	s	Ratio of A (%)	Ratio of Y (%)
Synth1	5000	1000	4	60:40	47:53
Synth2	5000	1000	4	60:40	64:36
Synth3	5000	1000	3	60:40	76:24
Synth4	5000	1000	6	60:40	55:45
German	900	100	9	69:31	70:30
Adult	34001	10870	9	67:33	25:75
COMPAS	4278	1000	7	40:60	47:53

tive Sanctions (COMPAS).[8] The German credit dataset consists of the records of loan applicants that contain gender A, financial information $\boldsymbol{S}$ (i.e., savings amount, checking account balance, and housing (rent or own)), information about debts $\boldsymbol{R}$ (i.e., amount of credit debt and repayment duration), and other attributes $\boldsymbol{C}$ (i.e., age and loan's purpose). With this dataset, we predicted the risk of each loan applicant (Y), where 900 and 100 samples were used as training and test data. On the other hand, the Adult dataset is comprised of US census data that contain such features of as marital status M, education L, occupation information $\boldsymbol{R}$ (e.g., weekly working hours), age and nationality $\boldsymbol{C}$. Using this dataset, we predicted whether annual income exceeds \$50,000 (Y), where $34,001$ and $10,870$ samples were employed as training and test data. To measure the unfairness of the predictions, following Chiappa and Gillam [2019], we used the causal graphs in Figure 5.9(a) and (b). Regarding the COMPAS dataset, we provide its detail in Section 5.6.3, including the causal graph in Figure 5.9(c).

5.6.2 Evaluation of Proposed Framework

In this section, we present the performance of our proposed learning framework described in Section 5.3.

Accuracy and Fairness

We tested the performance of each method using Synth1 dataset, generated by an SEM that does not satisfy the functional assumption of the **PSCF** method, and two real-world datasets: the German credit dataset and the Adult dataset.

[8]We used the modified COMPAS dataset included in R package "fairness" [Kozodoi and V. Varga, 2021].

Table 5.3: Test accuracy (%) on each dataset when using feed-forward neural network (DNN) and logistic regression (LR). **PSCF** is not shown with LR because it is a neural network-based method.

Method	Synth1		German		Adult	
	DNN	LR	DNN	LR	DNN	LR
Proposed	80.1 ± 0.6	78.0 ± 1.3	74.0 ± 2.4	72.3 ± 1.4	75.4 ± 0.4	75.2 ± 0.8
FIO	84.5 ± 0.5	83.5 ± 1.3	71.9 ± 3.1	69.4 ± 1.5	80.3 ± 0.5	76.3 ± 1.6
PSCF	75.3 ± 1.2	-	75.1 ± 1.3	-	74.2 ± 0.9	-
Unconstrained	88.0 ± 0.9	87.4 ± 0.4	77.1 ± 1.2	73.7 ± 0.8	83.2 ± 0.3	78.8 ± 1.2
Remove	76.6 ± 1.3	75.9 ± 0.9	71.1 ± 1.7	67.0 ± 1.6	74.8 ± 0.1	72.5 ± 1.9

Table 5.4: AUC on each dataset when using feed-forward neural network (DNN) and logistic regression (LR). **PSCF** is not shown with LR because it is a neural network-based method.

Method	Synth1		German		Adult	
	DNN	LR	DNN	LR	DNN	LR
Proposed	0.786 ± 0.003	0.790 ± 0.004	0.638 ± 0.007	0.632 ± 0.009	0.618 ± 0.003	0.644 ± 0.008
FIO	0.845 ± 0.011	0.842 ± 0.009	0.673 ± 0.008	0.705 ± 0.007	0.702 ± 0.006	0.723 ± 0.007
PSCF	0.720 ± 0.015	-	0.592 ± 0.012	-	0.642 ± 0.013	-
Unconstrained	0.874 ± 0.006	0.872 ± 0.003	0.722 ± 0.012	0.735 ± 0.004	0.801 ± 0.005	0.773 ± 0.006
Remove	0.760 ± 0.005	0.756 ± 0.007	0.595 ± 0.008	0.622 ± 0.012	0.604 ± 0.009	0.585 ± 0.001

We evaluated the test accuracy, the area under the curve (AUC), and four statistics of unfair effects: (i) the mean unfair effect (Definition 7), (ii) the standard deviation in the conditional mean unfair effects conditioned on the features of each individual (Definition 8), (iii) the upper bound on PIU (Theorem 2), and (iv) the PIU (Definition 10).

Tables 5.3 and 5.4 present the test accuracy and the AUC measures when using the neural network and logistic regression, and Figures 5.10 and 5.11 show the four statistics of the unfair effects that are obtained with the neural network and logistic regression, respectively. We performed 20 experiments by randomly splitting each dataset into training and test data and evaluated the means and the standard deviations. In Figures 5.10 and 5.11, statistics (ii) and (iv) are not displayed for the German and Adult datasets because computing them requires true SEMs, which are unavailable for these real-world datasets. Regarding **PSCF**, statistics (i) and (i ii) are not shown in Figure 5.10 because they are not well-defined for this method, and the performance when using logistic regression is not presented in Table 5.3 and Figure 5.11 since it is a neural network-based method.

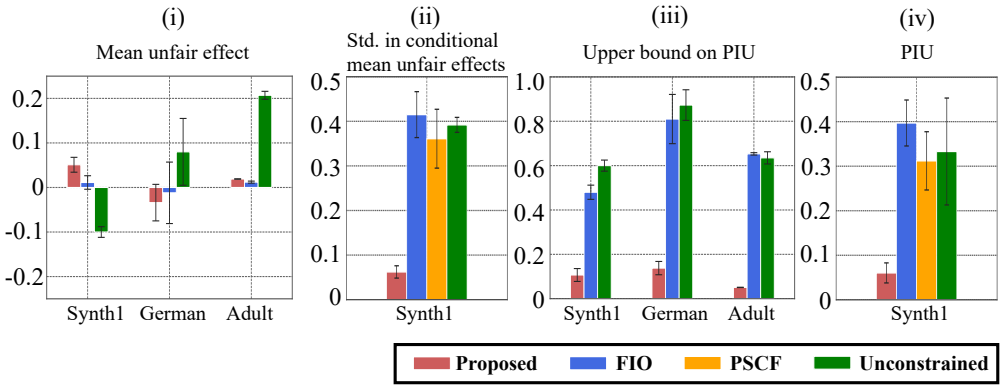

Figure 5.10: Four statistics of unfair effects on test data when using feed-forward neural network: The closer they are to zero, the fairer predictions are. With **Remove**, all statistics are zero (not shown). With **PSCF**, (i) and (iii) are not well-defined.

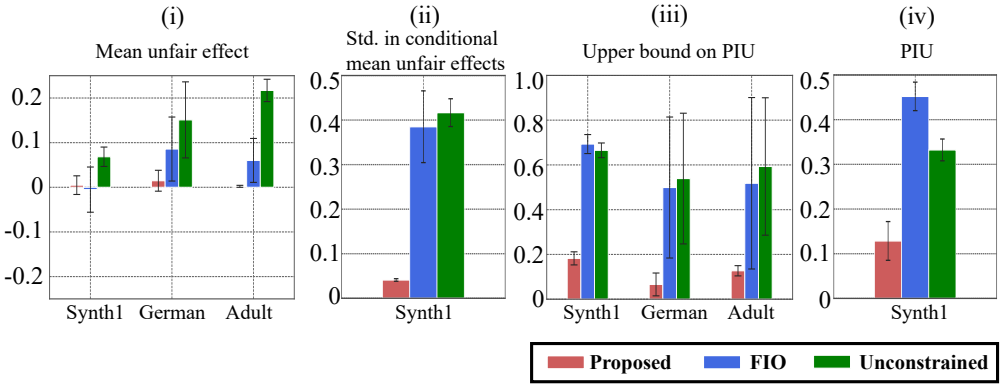

Figure 5.11: Four statistics of unfair effects on test data when using logistic regression model: The closer they are to zero, the fairer predictions are. With **Remove**, all statistics are zero (not shown). **PSCF** is not shown because it is a neural network-based method.

With **Proposed** using the neural network and logistic regression, all the statistics of the unfair effects were sufficiently close to zero, demonstrating that it made fair predictions for all individuals. This is because by imposing a penalty on the upper bound on PIU, **Proposed** forced unfair effect values to be close to zero for all individuals, guaranteeing that the other statistics were close to zero.

By contrast, regarding **FIO** and **PSCF**, the unfair effect values were much larger. With **FIO**, although the mean unfair effect (i.e., (i)) was close to zero, the other statistics deviated from zero, indicating that constraining the mean unfair effect did not ensure individual-level fairness. **PSCF** failed to reduce the value of the standard deviation in the conditional mean unfair effects (i.e., (ii)). This is because

the data are not generated from additive noise models (see Section 5.6.1 for the data), violating the functional assumption of **PSCF**. Since the large values of (ii) imply that unfair effects are greatly affected by the attributes of input features $\boldsymbol{X}$, these results indicate that **FIO** and **PSCF** made unfair predictions based on these attributes. With real-world datasets, **FIO** provided large upper bound values of PIU (i.e., (iii)). Since these values represent the upper bound, they do not necessarily imply that the predictions are unfair for some individuals. However, we cannot state with certainty that the predictions are individually fair, which might be problematic in practice.

The test accuracy and the AUC of **Proposed** were lower than **FIO**, higher than **Remove**, and comparable to **PSCF**. This result is reasonable because **FIO** imposes a much weaker fairness constraint than **Proposed**, **Remove**, and **PSCF**, all of which aim to guarantee individual-level fairness unlike **FIO**. By contrast, since **Remove** removes all informative input features that are affected by the sensitive feature to guarantee individual-level fairness, it provided the lowest accuracy. A comparison of **Proposed** and **PSCF** indicates that although our method employs a more severe fairness condition than **PSCF**, it barely sacrifices accuracy, demonstrating that it strikes a better balance between individual-level fairness and accuracy.

Since most methods achieved better performance with the neural network than with logistic regression, in the rest of this section, we present the experimental results when using the neural network.

Performance on Synthetic Data that Satisfy Functional Assumptions

We further compared the performance of **Proposed** with **PSCF** using the Synth2 dataset, which satisfies the functional assumptions of **PSCF**. With such data, we expect that both methods can learn an individually fair classifier.

Table 5.5 shows the test accuracy and the AUC, and Figure 5.12 presents the standard deviation in the conditional mean unfair effects and the PIU values on the Synth2 dataset. The test accuracy and the AUC of **Proposed** and **PSCF** were almost the same, which were lower than **Unconstrained**. Their PIU values were much close to zero than **Unconstrained**. These results demonstrate that if the data generating processes satisfy the functional assumptions of **PSCF**, **Proposed** and **PSCF** can achieve almost the same performance.

Table 5.5: Test accuracy (%) and AUC on Synth2 dataset

Method	Test accuracy (%)	AUC
Proposed	72.3 ± 0.9	0.760 ± 0.007
PSCF	72.5 ± 0.5	0.765 ± 0.004
Unconstrained	79.4 ± 1.1	0.816 ± 0.003

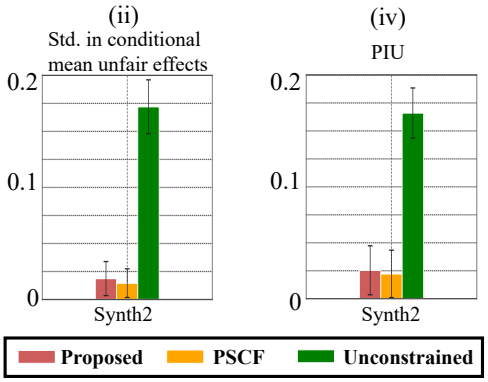

Figure 5.12: Two statistics of unfair effects on test data in Synth2 dataset: The closer they are to zero, the fairer predictions are.

Effectiveness of Proposed Upper Bound on PIU

As described in Section 5.3.3, our proposed learning framework utilizes an upper bound on PIU, which is much tighter than the existing bounds. To demonstrate its tightness, we compared the performance of **Proposed** with **Oracle**, which uses true PIU values as penalties during the training phase with the same penalty parameter value. Using a two-layered neural network as a classifier, we evaluated the test accuracy, the AUC, and the PIU value on the Synth1 dataset.

Table 5.6 presents the result. None of the test accuracy, the AUC, and the PIU value greatly differ, even if we have an oracle access to the true PIU values. This demonstrates that our upper bound is an effective alternative to the true PIU, whose value is unavailable in real-world scenarios.

Table 5.6: Test accuracy, AUC, and PIU value on Synth1 dataset

Method	Test accuracy (%)	AUC	PIU ($\times 10^{-2}$)
Proposed	80.1 ± 0.6	0.786 ± 0.003	6.04 ± 2.21
Oracle	79.2 ± 0.7	0.774 ± 0.011	3.15 ± 1.12

Table 5.7: Test accuracy and AUC on Synth3 dataset

Method	Test accuracy (%)	AUC
Proposed$_{lc}$	85.1 ± 0.4	0.822 ± 0.008
Proposed	86.0 ± 0.9	0.856 ± 0.011
FIO	87.5 ± 0.8	0.872 ± 0.005
PSCF	84.2 ± 1.2	0.809 ± 0.012
Unconstrained	89.1 ± 1.0	0.886 ± 0.003
Remove	83.8 ± 0.9	0.782 ± 0.010

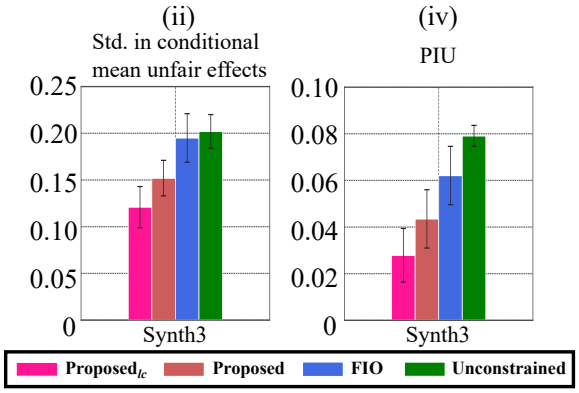

Figure 5.13: Two statistics of unfair effects on test data: The closer they are to zero, the fairer predictions are. With **PSCF** and **Remove**, both statistics are zero.

5.6.3 Testing Extended Frameworks

Here we show the performance of our extended learning framework described in Section 5.5. Section 5.6.3 presents the experimental results on synthetic data that are affected by latent confounders, and Section 5.6.3 illustrate the performance when using multiple candidates of causal graph structures.

Performance in Presence of Latent Confounders

We tested our extended framework (**Proposed**$_{lc}$), which addresses cases with latent confounders. We evaluated the performance with the Synth3 dataset, generated based on the causal graph in Figure 5.8(b) that contains a latent confounder.

To evaluate the unfairness of the predictions, we computed (ii) the standard deviation in the conditional mean unfair effects and (iv) the PIU. For a fair comparison, we did not evaluate the other two statistics (i.e., (i) and (iii)) because they depend on marginal potential outcome probabilities, whose estimators are formulated in different ways between **Proposed**$_{lc}$ and the other methods.

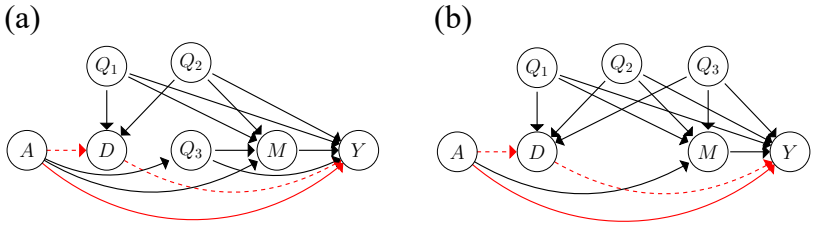

Figure 5.14: Two causal graph candidates for Synth4 dataset: (a): True causal graph; (b) Misspecified causal graph.

We present the test accuracy and the AUC in Table 5.7 and the two statistics of unfair effects in Figure 5.13. Note that not only **Remove** but also **PSCF** reduce the unfair effects to exactly zero. This is because **PSCF** makes predictions using the same A's value for all individuals, which completely removes unfair effects if the unfair pathways are only direct pathway, $\pi = \{A \to Y\}$, as in the case of the causal graph in Figure 5.8(b).

With **Proposed**$_{lc}$, both unfair-effect statistics were closer to zero than **Proposed** and **FIO** because it uses more reliable estimators of marginal potential outcome probabilities, which are designed for dealing with latent confounders. These results demonstrate that in the presence of latent confounders, our proposed extension makes fairer predictions than those methods.

The test accuracy and the AUC of **Proposed**$_{lc}$ exceeded **PSCF** and **Remove**, both of which completely eliminate unfair effects, indicating that our **Proposed**$_{lc}$ can strike a better balance between prediction accuracy and fairness.

Achieving a good balance between accuracy and fairness in the presence of latent confounders remains an open problem. Nevertheless, these experimental results suggest that if reliable estimators of lower and upper bounds on marginal probabilities are available, our proposed extension can strike a good balance between individual-level fairness and prediction accuracy.

Synthetic Data Experiments under Uncertain Causal Graph Structure

Using synthetic data, we tested our extension, which aims to achieve individual-level fairness when multiple causal graphs are given as input. We compared its performance with the original one that has oracle access to the true causal graph.

We used Synth4 dataset, whose true causal graph is shown in Figure 5.14(a). As

Table 5.8: Test accuracy and AUC on Synth4 dataset

Method	Test accuracy (%)	AUC
Proposed$_{ab}$	70.5 ± 2.5	0.708 ± 0.019
Proposed$_{a}$	71.4 ± 1.4	0.741 ± 0.013
Proposed$_{b}$	71.2 ± 1.1	0.733 ± 0.015
Unconstrained	86.3 ± 1.1	0.868 ± 0.012
Remove	68.3 ± 1.2	0.650 ± 0.015

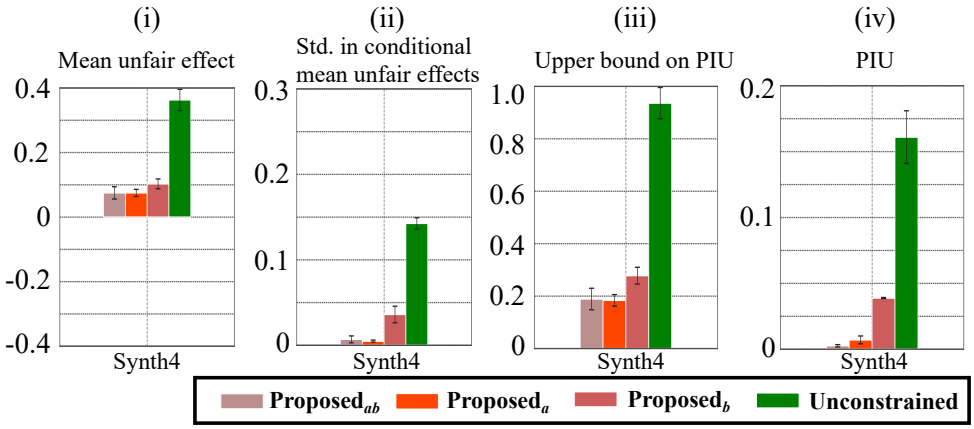

Figure 5.15: Four statistics of unfair effects on test data in Synth4 dataset: The closer they are to zero, the fairer predictions are. **Proposed**$_a$ uses true causal graph, **Proposed**$_{ab}$ takes two causal graphs, and **Proposed**$_b$ employs misspecified causal graph. With **Remove**, all statistics are zero (not shown).

with Russell *et al.* [2017], for simplicity, we focus on the case where there are two causal graph candidates. Our extended framework (**Proposed**$_{ab}$) takes as input the two causal graphs in Figure 5.14(a) and (b): the true causal graph for Synth4 dataset and the misspecified causal graph whose mediators and confounders are different from those of the true causal graph. To test **Proposed**$_{ab}$, we add the following baselines:

1. **Proposed**$_a$: Our method employing the true causal graph (Figure 5.14(a)).
2. **Proposed**$_b$: Our method using the misspecified causal graph (Figure 5.14(b)).

Note that in this experiment, we did not compare our method with **FIO** [Nabi and Shpitser, 2018] and **PSCF** [Chiappa and Gillam, 2019] since these methods are not designed for multiple causal graph candidates.

We show the test accuracy and the AUC in Table 5.8 and display the four statistics of the unfair effects in Figure 5.15.

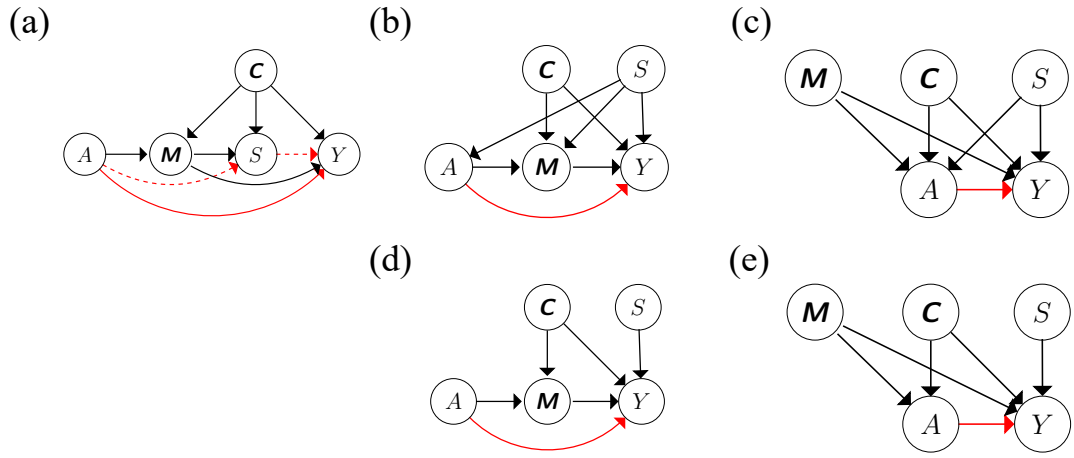

Figure 5.16: Causal graphs for COMPAS dataset: (a) true graph structure and (b)-(e) incorrect graph structures. A, $\boldsymbol{M}$, $\boldsymbol{C}$, S, and Y denote race, prior conviction, age and gender, COMPAS score, and predicted recidivism.

As expected, our extended framework (**Proposed**$_{ab}$) made fairer predictions than the original one with the misspecified causal graph (**Proposed**$_{b}$) and maintained similar test accuracy. With **Proposed**$_{ab}$, all the statistics of the unfair effects were comparable to those of **Proposed**$_{a}$, and three statistics, (ii), (iii), and (iv), were much closer to zero than **Unconstrained**.

Real-World Data Experiments under Uncertain Causal Graph Structure

To further evaluate the performance of our proposed extension, we performed real-world data experiments using the dataset about a risk assessment tool for prisoners called COMPAS [Angwin *et al.*, 2016].

This dataset contains the records of prisoners including race A, prior conviction $\boldsymbol{M}$, age and gender $\boldsymbol{C}$, COMPAS score S, and recidivism Y. The news media called Propublica [Angwin *et al.*, 2016] provided a report on COMPAS score S that the scoring by the COMPAS, which evaluates the probability of recidivism Y from the attributes of each prisoner, was discriminatory with respect to race A. Based on this report, we tested each method by evaluating the performance of predicting recidivism Y from features $\boldsymbol{X} = \{A, \boldsymbol{M}, \boldsymbol{C}, S\}$. Following Russell *et al.* [2017], we regarded the causal graph in Figure 5.16(a) as the true causal graph structure and tested our method.

To test our proposed extension, we used five causal graph candidates presented

Table 5.9: Test accuracy and AUC on COMPAS dataset

Method	Test accuracy (%)	AUC
Proposed$_{abcde}$	63.4 ± 1.1	0.614 ± 0.009
Proposed$_{a}$	63.1 ± 0.9	0.615 ± 0.011
Proposed$_{b}$	62.5 ± 1.3	0.614 ± 0.013
Proposed$_{c}$	62.0 ± 0.9	0.604 ± 0.009
Proposed$_{d}$	61.2 ± 1.3	0.600 ± 0.012
Proposed$_{e}$	62.5 ± 1.0	0.610 ± 0.004
Unconstrained	67.6 ± 1.3	0.674 ± 0.013
Remove	58.8 ± 0.9	0.581 ± 0.013

in Figure 5.16. Here the causal graphs in Figure 5.16(b)-(e) are incorrect: all these causal graphs illustrate that COMPAS score S is not affected by the attributes of each prisoner; however, this contradicts the truth because COMPAS score S is computed from these attributes. Using such causal graph candidates, we compared the performance of our proposed extension (**Proposed**$_{abcde}$) with the following baselines:

1. **Proposed**$_{a}$: Our method using the true causal graph structure in Figure 5.16(a).

2. **Proposed**$_{b}$, **Proposed**$_{c}$, **Proposed**$_{d}$, and **Proposed**$_{e}$: Our method employing the incorrect causal graph structures Figure 5.16(b), (c), (d), and (e), respectively.

We display the test accuracy and the AUC of each method in Table 5.9 and show the unfairness of the predictions in Figure 5.17.

All variants of our proposed method achieved higher test accuracy and AUC than **Remove** and made fairer predictions than **Uncosntrained**, whose mean unfair effect and the upper bound on PIU of **Unconstrained** were -0.195 ± 0.006 and 0.886 ± 0.003, respectively (not shown in Figure 5.17).

Compared with other variants, the test accuracy and AUC of our proposed extension **Proposed**$_{abcde}$ did not greatly differ. However, we observed that the unfair effects of **Proposed**$_{abcde}$ were closer to the case of using the true causal graph structure (**Proposed**$_{a}$) than the misspecified cases (i.e., **Proposed**$_{b}$, **Proposed**$_{c}$, **Proposed**$_{d}$, and **Proposed**$_{e}$). These results demonstrate that our proposed extension achieves a good tradeoff between accuracy and fairness.

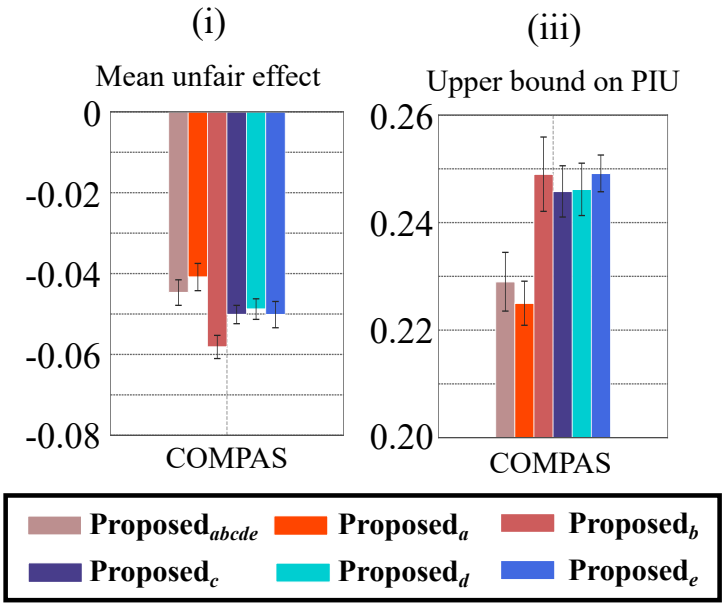

Figure 5.17: Two statistics of unfair effects on test data in COMPAS dataset: The close they are to zero, the fairer predictions are.

5.7 Conclusion

We proposed a learning framework for guaranteeing individual-level fairness without impractical functional assumptions. Based on a concept called path-specific effects, we defined a quantity called PIU and derived its upper bound that can be estimated from data without making restrictive functional assumptions. By forcing this upper bound value to be nearly zero, our framework trains an individually fair classifier. Furthermore, we show that this framework can be extended to deal with challenging real-world scenarios where there are unobserved variables called latent confounders and where the causal graph is uncertain. From a viewpoint of the feasible regions of optimization problems, we illustrate why making the upper bound value (close to) zero guarantees individual-level fairness. We experimentally show that our method makes individually fairer predictions than the existing methods at a slight cost of accuracy, indicating that it strikes a better balance between fairness and accuracy.

Our proposed learning framework indicates that even if estimating the causality-based unfairness measure is difficult, by utilizing its bounds, we can learn fair predictive models. Although deriving such bounds is often challenging, once we obtain them, they allow us to achieve fairness in complex real-world scenarios.

5.8 Proofs

We derive the upper bound on PIU (Theorem 2), the estimators of marginal potential outcome probabilities in Eq. (5.18), and the lower bound on PIU in Eq. (5.23).

5.8.1 Upper Bound on PIU (Theorem 2)

Proof. Let the marginal potential outcome probabilities be $\alpha = \mathrm{P}(Y_{A\Leftarrow 0} = 1)$ and $\beta = \mathrm{P}(Y_{A\Leftarrow 1\|\pi} = 1)$, and let their joint probabilities of $(Y_{A\Leftarrow 0}, Y_{A\Leftarrow 1\|\pi}) = (0,0)$, $(0,1)$, $(1,0)$, and $(1,1)$ be p_{00}, p_{01}, p_{10}, and p_{11}, respectively. Then we have

$$\begin{aligned} &p_{10} + p_{11} = \alpha, \quad p_{00} + p_{01} = 1 - \alpha, \\ &p_{01} + p_{11} = \beta, \quad \text{and} \quad p_{10} + p_{00} = 1 - \beta. \end{aligned} \tag{5.35}$$

Using marginal probabilities α and β, joint probability $\mathrm{P}^I(Y_{A\Leftarrow 0} \neq Y_{A\Leftarrow 1\|\pi})$ can be represented as

$$\mathrm{P}^I(Y_{A\Leftarrow 0} \neq Y_{A\Leftarrow 1\|\pi}) = \beta(1-\alpha) + \alpha(1-\beta).$$

As a result, the right-hand side in Eq. (5.10) can be written as $2(\beta(1-\alpha)+\alpha(1-\beta))$.

Therefore, our goal is to prove

$$p_{01} + p_{10} \leq 2(\beta(1-\alpha) + \alpha(1-\beta)).$$

Since all the joint probabilities in Eq. (5.35) are non-negative, p_{01} and p_{10} become at most $\min\{\beta,\, 1-\alpha\}$ and $\min\{\alpha,\, 1-\beta\}$, respectively; this implies

$$p_{01} + p_{10} \leq \min\{\beta, 1-\alpha\} + \min\{\alpha, 1-\beta\}. \tag{5.36}$$

Hence, it suffices to prove

$$\min\{\beta, 1-\alpha\} + \min\{\alpha, 1-\beta\} \leq 2\beta(1-\alpha) + 2\alpha(1-\beta). \tag{5.37}$$

Since both sides in inequality (5.37) are symmetric with respect to lines $\beta = \alpha$ and $\beta = 1-\alpha$, it is sufficient to consider the case when $\alpha \leq \beta \leq 1-\alpha$, which is illustrated in Figure 5.18 as the red triangle. In this case, since $\min\{\beta,\, 1-\alpha\} = \beta$

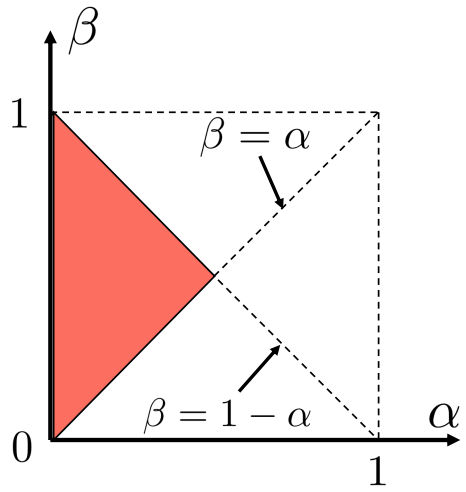

Figure 5.18: Red triangle represents region where marginal probability values α and β satisfy $\alpha \leq \beta \leq 1 - \alpha$.

and $\min\{\alpha, 1-\beta\} = \alpha$, inequality (5.37) can be written as

$$\beta + \alpha \leq 2\beta(1-\alpha) + 2\alpha(1-\beta) \Leftrightarrow \alpha + \beta - 4\alpha\beta \geq 0. \tag{5.38}$$

Since $\alpha + \beta \leq 1$ holds in this case, we have inequality $\alpha + \beta - (\alpha + \beta)^2 \geq 0$. Using this inequality, inequality (5.38) can be proven as follows:

$$\alpha + \beta - 4\alpha\beta = \alpha + \beta - (\alpha + \beta)^2 + (\alpha - \beta)^2 \geq 0. \tag{5.39}$$

Thus, we obtain Theorem 2. □

5.8.2 Marginal Potential Outcome Probabilities in Eq. (5.18)

Following the original paper [Huber, 2014], we derived the following formulation of the existing estimators of the marginal potential outcome probabilities:

$$\begin{aligned} \hat{p}_\theta^{A \Leftarrow 0} &= \frac{1}{n} \sum_{i=1}^{n} \mathbf{I}(a_i = 0) \hat{w}_i^{A \Leftarrow 0} c_\theta(a_i, q_i, d_i, m_i), \\ \hat{p}_\theta^{A \Leftarrow 1 \| \pi} &= \frac{1}{n} \sum_{i=1}^{n} \mathbf{I}(a_i = 1) \hat{w}_i^{A \Leftarrow 1 \| \pi} c_\theta(a_i, q_i, d_i, m_i), \end{aligned} \tag{5.18}$$

where $c_\theta(\boldsymbol{X}) = \mathrm{P}(Y = 1 \mid \boldsymbol{X})$ is the conditional distribution given by classifier h_θ, $\mathbf{I}(\cdot)$ is an indicator function, and $\hat{w}_i^{A \Leftarrow 0}$ and $\hat{w}_i^{A \Leftarrow 1 \| \pi}$ are the following weights:

$$\begin{aligned} \hat{w}_i^{A \Leftarrow 0} &= \frac{1}{\hat{\mathrm{P}}(A = 0 \mid q_i)}, \\ \hat{w}_i^{A \Leftarrow 1 \| \pi} &= \frac{\hat{\mathrm{P}}(A = 1 | q_i, d_i)\hat{\mathrm{P}}(A = 0 \mid q_i, d_i, m_i)}{\hat{\mathrm{P}}(A = 1 \mid q_i)\hat{\mathrm{P}}(A = 0 \mid q_i, d_i)\hat{\mathrm{P}}(A = 1 \mid q_i, d_i, m_i)}, \end{aligned} \tag{5.19}$$

where $\hat{\mathrm{P}}$ is the conditional distribution that is estimated by learning the statistical models (e.g., neural networks) to the training data beforehand.

Following the notations in the original paper [Huber, 2014], let the potential outcomes denote $Y_{A \Leftarrow 0} = Y(0, D(0), M(0))$ and $Y_{A \Leftarrow 1 \| \pi} = Y(1, D(1), M(0))$.

Then with the causal graph in Figure 5.2(c), marginal probability $\mathrm{P}(Y_{A \Leftarrow 1 \| \pi} = 1)$ can be written as

$$\begin{aligned} &\mathrm{P}(Y_{A \Leftarrow 1 \| \pi} = 1) \\ =& \mathrm{P}(Y(1, D(1), M(0)) = 1) \\ =& \mathbb{E}_Q[\mathbb{E}_{D(1)|Q}[\mathbb{E}_{M(0)|Q,D(1)}[\mathrm{P}(Y(1, d, m) = 1 \mid A = 1, Q = q, D(1) = d, M(0) = m)]]]. \end{aligned}$$

Using Assumption 10, this can be rewritten as

$$\begin{aligned} &\mathrm{P}(Y_{A \Leftarrow 1 \| \pi} = 1) \\ =& \mathbb{E}_Q[\mathbb{E}_{D|A=1,Q}[\mathbb{E}_{M|A=0,Q,D}[\mathrm{P}(Y(1, d, m) = 1 \mid A = 1, Q = q, D = d, M = m)]]]. \end{aligned}$$

With Bayes' theorem, this can be expressed as

$$\begin{aligned} &\mathrm{P}(Y_{A \Leftarrow 1 \| \pi} = 1) \\ =& \mathbb{E}_Q[\mathbb{E}_{D|Q}[\mathbb{E}_{M|Q,D}[\omega^{A \Leftarrow 1 \| \pi} \mathrm{P}(Y = 1 \mid dA = 1, Q = q, D = d, M = m)]]], \end{aligned}$$

where $\omega^{A \Leftarrow 1 \| \pi}$ is expressed as follows:

$$\omega^{A \Leftarrow 1 \| \pi} = \frac{\mathrm{P}(A = 1 \mid Q = q, D = d)\, \mathrm{P}(A = 0 \mid Q = q, D = d, M = m)}{\mathrm{P}(A = 1 \mid Q = q)\, \mathrm{P}(A = 0 \mid Q = q, D = d)}.$$

With indicator function $\mathbf{I}(\cdot)$, this can be formulated as

$$\mathrm{P}(Y_{A\Leftarrow 1\|\pi}=1)=\mathbb{E}[\mathbf{I}(A=1)w^{A\Leftarrow 1\|\pi}\,\mathrm{P}(Y=1\mid A=1,q,d,m)], \tag{5.40}$$

where weight w' is expressed as

$$w^{A\Leftarrow 1\|\pi}=\frac{1}{\mathrm{P}(A=1\mid Q=q,D=d,M=m)}\omega^{A\Leftarrow 1\|\pi}.$$

In a similar manner, marginal probability $\mathrm{P}(Y_{A\Leftarrow 0}=1)$ can be represented as

$$\mathrm{P}(Y_{A\Leftarrow 0}=1)=\mathbb{E}[\mathbf{I}(A=0)w^{A\Leftarrow 0}\,\mathrm{P}(Y=1\mid A=0,q,d,m)], \tag{5.41}$$

where weight $w^{A\Leftarrow 0}$ is formulated as

$$w^{A\Leftarrow 0}=\frac{1}{\mathrm{P}(A=0\mid Q=q)}.$$

Given empirical distribution, by plugging conditional distribution c_θ into $\mathrm{P}(Y=1\mid A=1,Q=q,D=d,M=m)$, we can estimate (5.41) and (5.40) as (5.18).

5.8.3 Lower Bound on PIU in Eq. (5.23)

Since we already proved the upper bound in (5.36), below we derive the lower bound. Since α and β are marginal probabilities, we have

$$p_{10}+p_{11}=\alpha,\quad p_{01}+p_{11}=\beta,$$

which are equivalent to

$$p_{10}=\alpha-p_{11},\quad p_{01}=\beta-p_{11},$$

respectively. By summing up both, we have

$$p_{01}+p_{10}=\alpha+\beta-2p_{11}.$$

Since joint probability p_{11} is less than marginal probabilities α and β, we have $p_{11} \leq \min\{\alpha, \beta\}$. Therefore,

$$p_{01} + p_{10} \geq \alpha + \beta - 2\min\{\alpha, \beta\} = |\alpha - \beta|. \tag{5.42}$$

Combined with the upper bound on $p_{01} + p_{10}$ in (5.36), we obtain (5.23).

Chapter 6

Conclusion

6.1 Contribution Summary

In this dissertation, we have established the three causal inference frameworks for accelerating scientific discoveries and improving the reliability of machine learning predictions. Below we discuss the contribution of each framework.

Causal Discovery from Time Series Data (Chapter 3)

Complex nonlinear time series are common in various scientific fields, such as bioinformatics, neuroscience, and meteorology. Inferring the causal relationships in them is challenging, especially when the data are scarce. To tackle this challenge, we have proposed a supervised learning approach that can improve the inference accuracy using training data, i.e., the time series data whose causal relationships are obvious.

To further ameliorate the inference accuracy, we would need to overcome the three limitations of proposed method:

1. As can be seen from the definition of Granger causality, our supervised learning framework cannot detect an *instantaneous causal relationship* (a.k.a. *contemporaneous causal relationship*), where one variable influences another at the same time t (i.e., $X_t \to Y_t$). This limitation is crucial especially when the time series data are infrequently sampled.

2. Our framework cannot correctly infer the causal relationship between time-dependent variables X and Y if more than one variable acts as their common

cause (i.e., confounders $Z_1, Z_2, \ldots$ such that $X \leftarrow Z_1, Z_2, \cdots \rightarrow Y$) because our extended feature representation is designed for trivariate time series data.

3. Related to the above, our framework wrongly outputs the results if there are unobserved common cause variables (a.k.a. unobserved confounders).

The first two limitations have already been resolved by the context-aware Dependency to Causality (caD2C) algorithm [Bontempi, 2020], which is a recently proposed supervised learning approach that is motivated by ours. However, it cannot overcome limitation 3; indeed, correctly inferring the causal relationships in presence of unobserved confounders is extremely challenging in causal discovery.

Hence, it would be interesting to investigate how to develop a supervised learning approach that avoids outputting incorrect causal relationships by inferring the maximal ancestral graph (MAG) [Richardson and Spirtes, 2002], which displays the possibility of the presence of unobserved confounders by bi-directed edge $\leftrightarrow$.

Interpretable Treatment Effect Estimation (Chapter 4)

When the treatment effects are different across individuals, elucidating why such heterogeneity exists is a common interest in many scientific fields, such as medicine and economics. To deepen the understanding of the causal mechanisms that yield heterogeneous treatment effects, we have developed a feature selection framework for discovering distributional treatment effect modifiers. By utilizing the distributional information, our method can find a wider variety of important features related to treatment effect heterogeneity, compared with the existing mean-based methods. This advantage leads to better understanding of the underlying causal mechanisms and thus is helpful for making scientific discoveries.

As future work, it would be interesting to tackle *large p and small n problems*, i.e., the cases where there are much more features than the number of observed data points. Such cases are common in various tasks related to bioinformatics, such as biomarker selection and toxicogenomics selection. However, as can be seen from the experimental results in Section 4.5.2, the performance of proposed method is not sufficiently good under a low-sample setting. Hence, establishing a statistical framework for dealing with large p and small n problems is left as our future work.

Making Accurate and Fair Predictions based on Causality (Chapter 5)

Causality-based fairness criteria have got increasing attention in the field of machine learning and fairness. However, learning fair predictive models based on these criteria is challenging due to the difficulty of estimating the causality-based unfairness measure. To overcome this difficulty, we have proposed a learning framework that effectively utilizes the upper bound on the unfairness measure. We show that such an idea of bounding the unfairness measure can be extended to complex settings, such as the presence of unobserved confounders and the cases where the causal graph structure is uncertain. Experimental results show that our framework makes much fairer predictions for each individual than the existing methods at a slight expense of prediction accuracy.

Thus, our learning framework indicates that we can overcome the difficulty of estimating unfair causal effects by imposing a constraint on the bound on the functional of the joint distribution of potential outcomes. Indeed, deriving such a bound has been actively studied recently [Fan *et al.*, 2017; Firpo and Ridder, 2019; Shingaki and Kuroki, 2021]. An interesting future work direction is to utilize these bounds for achieving a better balance between fairness and accuracy.

6.2 Conclusion and Future Directions

Throughout this dissertation, we have discussed the two ultimate goals of causal inference: scientific discoveries and fairness-aware machine learning.

To accomplish these ultimate goals, taking an interdisciplinary approach that combines various tools and concepts in statistics, machine learning, and causal inference is crucially important. All the three causal inference frameworks presented in this dissertation indicate the importance of such an interdisciplinary viewpoint. In Chapter 3, we have solved the traditional statistical problem of Granger causality identification via supervised learning. In Chapter 4, we have improved the selection accuracy of treatment effect modifiers by utilizing the kernel MMD, which is a common distributional discrepancy measure in machine learning. And in Chapter 5, we have shown that we can make much better balance between fairness and accuracy by utilizing the causality concepts.

From a broader perspective, we would need to focus our attention on the task dependency among causal discovery, treatment effect estimation, and fairness-aware machine learning. That is, causal discovery is helpful for treatment effect estimation (because the correct identification of confounders is a prerequisite for the estimation), and treatment effect estimation contributes to making fair predictions (since an accurate causal-effect estimator allows us to precisely measure the unfairness of predictions). This task dependency tells us that taking a higher viewpoint might be important in achieving the ultimate goals. For instance, to create an ideal future where machine learning predictions are used with no concern for fairness, it might be essential to tackle fundamental causal inference challenges: improving the inference accuracy of causal discovery and developing an accurate treatment effect estimator.

We believe that the lens of causality has infinite potential toward making scientific discoveries and achieving trustworthy machine learning. To unfold this potential, however, we would still need to overcome enormous challenges and resolve many methodological limitations. Taking these challenges and limitations from a broad and interdisciplinary perspective will lead to the future success in causal inference.

Bibliography

Shipra Agrawal, Yichuan Ding, Amin Saberi, and Yinyu Ye. Correlation robust stochastic optimization. In *SODA*, pages 1087–1096, 2010. ↰ p.89

Ahmed M. Alaa and Mihaela van der Schaar. Bayesian inference of individualized treatment effects using multi-task Gaussian processes. In *NeurIPS*, pages 3427–3435, 2017. ↰ p.2

Pierre-Olivier Amblard and Olivier J.J. Michel. The relation between Granger causality and directed information theory: A review. *Entropy*, 15(1):113–143, 2013. ↰ p.9

Ryan M Andrews and Vanessa Didelez. Insights into the cross-world independence assumption of causal mediation analysis. *Epidemiology*, 32(2):209–219, 2020. ↰ p.92

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine Bias, 2016. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. ↰ p.2, ↰ p.73, ↰ p.112

Susan Athey and Guido Imbens. Recursive partitioning for heterogeneous causal effects. *Proceedings of National Academy of Sciences*, 113(27):7353–7360, 2016. ↰ p.2

Chen Avin, Ilya Shpitser, and Judea Pearl. Identifiability of path-specific effects. In *IJCAI*, pages 357–363, 2005. ↰ p.19, ↰ p.75, ↰ p.79, ↰ p.90

K. Bache and M. Lichman. UCI machine learning repository: Datasets. http://archive.ics.uci.edu/ml/datasets, 2013. ↰ p.103

Lionel Barnett, Adam B. Barrett, and Anil K. Seth. Granger causality and transfer entropy are equivalent for Gaussian variables. *Physical Review Letters*, 103(23):238701, 2009. ↰ p.9

David Bell, Jim Kay, and Jim Malley. A non-parametric approach to non-linear causality testing. *Economics Letters*, 51(1):7–18, 1996. ↰ p.22, ↰ p.24, ↰ p.26, ↰ p.33

Alexis Bellot and Mihaela van der Schaar. A kernel two-sample test with selection bias. In *UAI*, 2021. ↰ p.49, ↰ p.57

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995. ↰ p.55

Til Ole Bergmann and Gesa Hartwigsen. Inferring causality from noninvasive brain stimulation in cognitive neuroscience. *Journal of Cognitive Neuroscience*, 33(2):195–225, 2021. ↰ p.1

Gianluca Bontempi and Maxime Flauder. From dependency to causality: A machine learning approach. *JMLR*, 16:2437–2457, 2015. ↰ p.22, ↰ p.25

Gianluca Bontempi. Learning causal dependencies in large-variate time series. In *IJCNN*, pages 1–7, 2020. ↰ p.122

Emmanuel Candes, Yingying Fan, Lucas Janson, and Jinchi Lv. Panning for gold:'Model-X'knockoffs for high dimensional controlled variable selection. *Journal of Royal Statistical Society: Series B (Statistical Methodology)*, 80(3):551–577, 2018. ↰ p.54, ↰ p.55

Yale Chang and Jennifer Dy. Informative subspace learning for counterfactual inference. In *AAAI*, pages 1770–1776, 2017. ↰ p.46

Min Chen and Hong Zhi An. A kolmogorov-smirnov type test for conditional heteroskedasticity in time series. *Statistics and Probability Letters*, 33(3):321–331, 1997. ↰ p.28

Heng Chen, Yanqin Fan, and Ruixuan Liu. Inference for the correlation coefficient between potential outcomes in the Gaussian switching regime model. *Journal of Econometrics*, 195(2):255–270, 2016. ↰ p.48, ↰ p.67

Dehua Cheng, Mohammad Taha Bahadori, and Yan Liu. FBLG: A simple and effective approach for temporal dependence discovery from time series data. In *KDD*, pages 382–391, 2014. ↰ p.22, ↰ p.26

Silvia Chiappa and Thomas P.S. Gillam. Path-specific counterfactual fairness. In *AAAI*, pages 7801–7808, 2019. ↰ p.20, ↰ p.74, ↰ p.75, ↰ p.76, ↰ p.77, ↰ p.83, ↰ p.85, ↰ p.87, ↰ p.90, ↰ p.100, ↰ p.104, ↰ p.111

Yoichi Chikahara and Akinori Fujino. Causal inference in time series via supervised learning. In *IJCAI*, pages 2042–2048, 2018. ↰ p.vii, ↰ p.85

Yoichi Chikahara and Akinori Fujino. A supervised learning approach to granger causality inference. *Information Processing Society of Japan (IPSJ) Transactions on Mathematical Modeling and its Applications (TOM)*, 11(3):58–73, 2018. ↰ p.vii

Yoichi Chikahara, Shinsaku Sakaue, Akinori Fujino, and Hisashi Kashima. Learning individually fair classifier with path-specific causal-effect constraint. In *AISTATS*, pages 145–153, 2021. ↰ p.vii, ↰ p.48

Yoichi Chikahara, Makoto Yamada, and Hisashi Kashima. Feature selection for discovering distributional treatment effect modifiers. In *UAI*, pages 400–410, 2022. ↰ p.vii

Alexandra Chouldechova, Diana Benavides-Prado, Oleksandr Fialko, and Rhema Vaithianathan. A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In *FAT*, pages 134–148, 2018. ↰ p.2, ↰ p.73

Thomas M. Cover. *Elements of Information Theory*. 1999. ↰ p.9

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *ITCS*, pages 214–226, 2012. ↰ p.3, ↰ p.74, ↰ p.86

Michael Eichler. Causal inference with multiple time series: Principles and problems. *Philosophical Transactions of Royal Society A: Mathematical, Physical and Engineering Sciences*, 371(1997), 2013. ↰ p.5, ↰ p.8

Felix Elwert and Christopher Winship. Endogenous selection bias: The problem of conditioning on a collider variable. *Annual Review of Sociology*, 40:31–53, 2014. ↰ p.45, ↰ p.51

Yanqin Fan, Emmanuel Guerre, and Dongming Zhu. Partial identification of functionals of the joint distribution of "potential outcomes". *Journal of Econometrics*, 197(1):42–59, 2017. ↰ p.89, ↰ p.123

Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *KDD*, pages 259–268, 2015. ↰ p.74, ↰ p.86

Sergio Firpo and Geert Ridder. Partial identification of the treatment effect distribution and its functionals. *Journal of Econometrics*, 213(1):210–234, 2019. ↰ p.89, ↰ p.123

Jean-Pierre Florens. Some technical issues in defining causality. *Journal of Econometrics*, 112(1):127–127, 2003. ↰ p.8

John Geweke. Measurement of linear dependence and feedback between multiple time series. *Journal of American Statistical Association*, 77(378):304–313, 1982. ↰ p.9

John F. Geweke. Measures of conditional linear dependence and feedback between time series. *Journal of American Statistical Association*, 79(388):907–915, 1984. ↰ p.31

Amir Gilad, Harsh Parikh, Sudeepa Roy, and Babak Salimi. Heterogeneous treatment effects in social networks. *arXiv preprint arXiv:2105.10591*, 2021. ↰ p.56

Clark Glymour, Kun Zhang, and Peter Spirtes. Review of causal discovery methods based on graphical models. *Frontiers in Genetics*, 10:524, 2019. ↰ p.77, ↰ p.85, ↰ p.99

Clive W.J. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of Econometric Society*, pages 424–438, 1969. ↰ p.5, ↰ p.6, ↰ p.7, ↰ p.22, ↰ p.24, ↰ p.26, ↰ p.33

Clive W.J. Granger. Testing for causality: A personal viewpoint. *Journal of Economic Dynamics and control*, 2:329–352, 1980. ↰ p.1, ↰ p.5, ↰ p.6, ↰ p.7

Arthur Gretton, Karsten M. Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex J. Smola. A kernel method for the two-sample-problem. In *NIPS*, pages 513–520, 2007. ↰ p.23, ↰ p.27

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *JMLR*, 13(1):723–773, 2012. ↰ p.44, ↰ p.49, ↰ p.66

Isabelle Guyon. ChaLearn cause-effect pair challenge. https://www.kaggle.com/c/cause-effect-pairs/, 2013. ↰ p.22, ↰ p.24

P. Richard Hahn, Jared S. Murray, and Carlos M. Carvalho. Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects. *Bayesian Analysis*, 15(3):965–1056, 2020. ↰ p.14, ↰ p.56

Moritz Hardt, Eric Price, Nati Srebro, et al. Equality of opportunity in supervised learning. In *NeurIPS*, pages 3315–3323, 2016. ↰ p.3, ↰ p.74, ↰ p.86

Negar Hassanpour and Russell Greiner. Counterfactual regression with importance sampling weights. In *IJCAI*, pages 5880–5887, 2019. ↰ p.2, ↰ p.14, ↰ p.46

James J. Heckman and Rodrigo Pinto. Econometric mediation analyses: Identifying the sources of treatment effects from experimentally estimated production technologies with unmeasured and mismeasured inputs. *Econometric Reviews*, 34(1-2):6–31, 2015. ↰ p.20

Miguel A. Hernán and James M. Robins. *Causal Inference: What if. Boca Raton: Chapman & Hill/CRC*. 2020. ↰ p.45

Tom Heskes, Evi Sijben, Ioan Gabriel Bucur, and Tom Claassen. Causal Shapley values: Exploiting causal knowledge to explain individual predictions of complex models. In *NeurIPS*, pages 4778–4789, 2020. ↰ p.2, ↰ p.20

Jennifer L. Hill. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011. ↰ p.2, ↰ p.14, ↰ p.46, ↰ p.56

Wassily Hoeffding. The strong law of large numbers for u-statistics. Technical report, 1961. ↰ p.71

Kimberly A. Houser. Can AI solve the diversity problem in the tech industry: Mitigating noise and bias in employment decision-making. *Stanford Technology Law Review*, 22:290, 2019. ↰ p.2, ↰ p.73

Patrik O. Hoyer, Dominik Janzing, Joris M. Mooij, Jonas Peters, and Bernhard Schölkopf. Nonlinear causal discovery with additive noise models. In *NeurIPS*, pages 689–696, 2009. ↰ p.75, ↰ p.84, ↰ p.86, ↰ p.101

Martin Huber. Identifying causal mechanisms (primarily) based on inverse probability weighting. *Journal of Applied Econometrics*, 29(6):920–943, 2014. ↰ p.90, ↰ p.91, ↰ p.92, ↰ p.116, ↰ p.117

Kosuke Imai and Marc Ratkovic. Estimating treatment effect heterogeneity in randomized program evaluation. *Annals of Applied Statistics*, 7(1):443–470, 2013. ↰ p.44, ↰ p.45, ↰ p.46

Kamal Abu Jabal, Hila Ben-Amram, Karine Beiruti, Yunis Batheesh, Christian Sussan, Salman Zarka, and Michael Edelstein. Impact of age, ethnicity, sex and prior infection status on immunogenicity following a single dose of the BNT162b2 mRNA COVID-19 vaccine: Real-world evidence from healthcare workers, Israel, December 2020 to January 2021. *Eurosurveillance*, 26(6), 2021. ↰ p.2, ↰ p.44

Dominik Janzing and Bernhard Scholkopf. Causal inference using the algorithmic markov condition. *IEEE Transactions on Information Theory*, 56(10):5168–5194, 2010. ↰ p.25

Fredrik Johansson, Uri Shalit, and David Sontag. Learning representations for counterfactual inference. In *ICML*, pages 3020–3029, 2016. ↰ p.2, ↰ p.14, ↰ p.56

Muhsin Kar, Şaban Nazlıoğlu, and Hüseyin Ağır. Financial development and economic growth nexus in the MENA countries: Bootstrap panel Granger causality analysis. *Economic modelling*, 28(1):685–693, 2011. ↰ p.21

Amir E. Khandani, Adlar J. Kim, and Andrew W. Lo. Consumer credit-risk models via machine-learning algorithms. *Journal of Banking and Finance*, 34(11):2767–2787, 2010. ↰ p.2, ↰ p.73

Niki Kilbertus, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. Avoiding discrimination through causal reasoning. In *NeurIPS*, pages 656–666, 2017. ↰ p.85

Samantha Kleinberg and George Hripcsak. A review of causal inference for biomedical informatics. *Journal of Biomedical Informatics*, 44(6):1102–1112, 2011. ↰ p.1

Tuen Kloek and Herman K. Van Dijk. Bayesian estimates of equation system parameters: An application of integration by monte carlo. *Econometrica: Journal of Econometric Society*, pages 1–19, 1978. ↰ p.12

Michael R. Kosorok and Eric B. Laber. Precision medicine. *Annual review of statistics and its application*, 6:263–286, 2019. ↰ p.10

Nikita Kozodoi and Tibor V. Varga. Fairness: Algorithmic fairness metrics, 2021. R package version 1.2.1; https://CRAN.R-project.org/package=fairness. ↰ p.104

Marlene Kretschmer, Samantha V. Adams, Alberto Arribas, Rachel Prudden, Niall Robinson, Elena Saggioro, and Theodore G Shepherd. Quantifying causal pathways of teleconnections. *Bulletin of American Meteorological Society*, 102(12):2247–2263, 2021. ↰ p.1

G.M. Kuersteiner. Granger-Sims causality. In *Macroeconometrics and Time Series Analysis*, pages 119–134. 2010. ↰ p.8

Solomon Kullback and Richard A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22(1):79–86, 1951. ↰ p.28

Sören R. Künzel, Jasjeet S. Sekhon, Peter J. Bickel, and Bin Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of National Academy of Sciences*, 116(10):4156–4165, 2019. ↰ p.2, ↰ p.14, ↰ p.46, ↰ p.56

Matt J. Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *NeurIPS*, pages 4066–4076, 2017. ↰ p.74, ↰ p.77, ↰ p.78, ↰ p.85, ↰ p.86, ↰ p.100

Matt Kusner, Chris Russell, Joshua Loftus, and Ricardo Silva. Making decisions that reduce discriminatory impacts. In *ICML*, pages 3591–3600, 2019. ↰ p.86

Rasmus Larsen. Reinforcement learning of causal variables using mediation analysis. In *AAAI*, 2022. ↰ p.2, ↰ p.20

Kwonsang Lee, Dylan S. Small, Jesse Y. Hsu, Jeffrey H. Silber, and Paul R. Rosenbaum. Discovering effect modification in an observational study of surgical mortality at hospitals with superior nursing. *Journal of Royal Statistical Society: Series A (Statistics in Society)*, 181(2):535–546, 2018. ↰ p.43, ↰ p.45

Kwonsang Lee, Falco J. Bargagli-Stoffi, and Francesca Dominici. Causal rule ensemble: Interpretable inference of heterogeneous treatment effects. *arXiv preprint arXiv:2009.09036*, 2020. ↰ p.56

Molei Liu, Eugene Katsevich, Lucas Janson, and Aaditya Ramdas. Fast and powerful conditional randomization testing via distillation. *Biometrika*, 2021. ↰ p.56

Joseph T. Lizier, Mikhail Prokopenko, and Albert Y. Zomaya. Local information transfer as a spatiotemporal filter for complex systems. *Physical Review E*, 77(2):026110, 2008. ↰ p.9

Joseph T. Lizier, Jakob Heinzle, Annette Horstmann, John-Dylan Haynes, and Mikhail Prokopenko. Multivariate information-theoretic measures reveal directed information structure and task relevant changes in fMRI connectivity. *Journal of computational neuroscience*, 30(1):85–107, 2011. ↰ p.9

David Lopez-Paz, Krikamol Muandet, Bernhard Schölkopf, and Ilya Tolstikhin. Towards a learning theory of cause-effect inference. In *ICML*, pages 1452–1461, 2015. ↰ p.22, ↰ p.25, ↰ p.33, ↰ p.34

David Lopez-Paz, Robert Nishihara, Soumith Chintala, Bernhard Schölkopf, and Léon Bottou. Discovering causal signals in images. In *CVPR*, pages 6979–6987, 2017. ↰ p.22, ↰ p.25

Karima Makhlouf, Sami Zhioua, and Catuscia Palamidessi. Survey on causal-based machine learning fairness notions. *arXiv*, 2020. ↰ p.3, ↰ p.86

Daniele Marinazzo, Mario Pellicoro, and Sebastiano Stramaglia. Kernel-Granger causality and the analysis of dynamical networks. *Physical Review E*, 77(5):056215, 2008. ↰ p.22, ↰ p.24, ↰ p.26, ↰ p.34

Caleb H. Miles, Phyllis Kanki, Seema Meloni, and Eric J. Tchetgen Tchetgen. On partial identification of the natural indirect effect. *Journal of Causal Inference*, 5(2):1–12, 2017. ↰ p.97, ↰ p.98

Christoph Molnar. *Interpretable Machine Learning*. Lulu. com, 2020. ↰ p.2

Joris M. Mooij, Jonas Peters, Dominik Janzing, Jakob Zscheischler, and Bernhard Schölkopf. Distinguishing cause from effect using observational data: Methods and benchmarks. *JMLR*, 17(1):1103–1204, 2016. ↰ p.38

Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, Bernhard Schölkopf, et al. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends® in Machine Learning*, 10(1-2):1–141, 2017. ↰ p.31

Krikamol Muandet, Motonobu Kanagawa, Sorawit Saengkyongam, and Sanparith Marukatat. Counterfactual mean embeddings. *JMLR*, 22(162):1–71, 2021. ↰ p.49, ↰ p.57

Razieh Nabi and Ilya Shpitser. Fair inference on outcomes. In *AAAI*, pages 1931–1940, 2018. https://github.com/raziehna/fair-inference-on-outcomes. ↰ p.20, ↰ p.74, ↰ p.75, ↰ p.76, ↰ p.77, ↰ p.83, ↰ p.86, ↰ p.90, ↰ p.93, ↰ p.94, ↰ p.100, ↰ p.111

Razieh Nabi, Daniel Malinsky, and Ilya Shpitser. Learning optimal fair policies. In *ICML*, pages 4674–4682, 2019. ↰ p.20, ↰ p.86

Elizbar A. Nadaraya. On estimating regression. *Theory of Probability & Its Applications*, 9(1):141–142, 1964. ↰ p.52

Xinkun Nie and Stefan Wager. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, 108(2):299–319, 2021. ↰ p.2, ↰ p.14, ↰ p.46, ↰ p.56

Inga Sif Ólafsdóttir, Thorarinn Gislason, B. Thjodleifsson, I. Olafsson, D. Gislason, Rain Jögi, and Christer Janson. C reactive protein levels are increased in non-allergic but not allergic asthma: A multicentre epidemiological study. *Thorax*, 60(6):451–454, 2005. ↰ p.63

Edward Ott. *Chaos in Dynamical Systems*. Cambridge University Press, 2002. ↰ p.40

Junhyung Park, Uri Shalit, Bernhard Schölkopf, and Krikamol Muandet. Conditional distributional treatment effect with kernel conditional mean embeddings and U-statistic regression. In *ICML*, pages 8401–8412, 2021. ↰ p.49, ↰ p.57

Judea Pearl. A probabilistic calculus of actions. In *UAI*, pages 454–462, 1994. ↰ p.16

Judea Pearl. Direct and indirect effects. In *UAI*, pages 411–420, 2001. ↰ p.17, ↰ p.82

Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2009. ↰ p.1, ↰ p.5, ↰ p.14, ↰ p.15, ↰ p.76, ↰ p.96

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *NeurIPS*, pages 1177–1184, 2007. ↰ p.31, ↰ p.45, ↰ p.53

Hans Reichenbach. *The Direction of Time*, volume 65. Univ of California Press, 1956. ↰ p.1

Thomas Richardson and Peter Spirtes. Ancestral graph Markov models. *Annals of Statistics*, 30(4):962–1030, 2002. ↰ p.122

Lorenzo Richiardi, Rino Bellocco, and Daniela Zugna. Mediation analysis in epidemiology: Methods, interpretation and bias. *International Journal of Epidemiology*, 42(5):1511–1519, 2013. ↰ p.20

James M. Robins and Thomas S. Richardson. Alternative graphical causal models and the identification of direct effects. *Causality and Psychopathology*, pages 103–158, 2010. ↰ p.97

Peter M. Robinson. Root-n-consistent semiparametric regression. *Econometrica: Journal of Econometric Society*, pages 931–954, 1988. ↰ p.2

Heriberto Rodríguez-Hernández, Luis E. Simental-Mendía, Gabriela Rodríguez-Ramírez, and Miguel A. Reyes-Romero. Obesity and inflammation: Epidemiology, risk factors, and markers of inflammation. *International journal of endocrinology*, 2013. ↰ p.62

Paul R. Rosenbaum and Donald B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983. ↰ p.12, ↰ p.50

Kenneth J. Rothman, Sander Greenland, Timothy L. Lash, et al. *Modern Epidemiology*, volume 3. Wolters Kluwer Health/Lippincott Williams & Wilkins Philadelphia, 2008. ↰ p.46

Donald B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688, 1974. ↰ p.2, ↰ p.5, ↰ p.10, ↰ p.14, ↰ p.45

Aviad Rubinstein and Sahil Singla. Combinatorial prophet inequalities. In *SODA*, pages 1671–1687, 2017. ↰ p.89

Chris Russell, Matt J Kusner, Joshua Loftus, and Ricardo Silva. When worlds collide: Integrating different counterfactual assumptions in fairness. In *NeurIPS*, pages 6414–6423, 2017. ↰ p.86, ↰ p.99, ↰ p.111, ↰ p.112

Thomas M. Russell. Sharp bounds on functionals of the joint distribution in the analysis of treatment effects. *Journal of Business & Economic Statistics*, 39(2):532–546, 2021. ↰ p.47, ↰ p.48, ↰ p.67

Piotr Rzepakowski and Szymon Jaroszewicz. Uplift modeling in direct marketing. *Journal of Telecommunications and Information Technology*, pages 43–50, 2012. ↰ p.10

Babak Salimi, Luke Rodriguez, Bill Howe, and Dan Suciu. Interventional fairness: Causal database repair for algorithmic fairness. In *SIGMOD*, pages 793–810, 2019. ↰ p.86

Peter Z. Schochet, Mike Puma, and John Deke. Understanding variation in treatment effects in education impact evaluations: An overview of quantitative methods.

National Center for Education Evaluation and Regional Assistance, 2014. ↰ p.43, ↰ p.45

Bernhard Scholkopf and Alexander J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT press, 2001. ↰ p.34

Bernhard Schölkopf, Alexander J. Smola, Francis Bach, et al. *Learning with kernels: Support vector machines, regularization, optimization, and beyond*. MIT press, 2002. ↰ p.58

Thomas Schreiber. Measuring information transfer. *Physical Review Letters*, 85(2):461, 2000. ↰ p.5, ↰ p.7, ↰ p.9, ↰ p.34

Konstantinos Sechidis, Matthias Kormaksson, and David Ohlssen. Using knockoffs for controlled predictive biomarker identification. *Statistics in Medicine*, 40(25):5453–5473, 2021. ↰ p.46

Robert J. Serfling. *Approximation theorems of mathematical statistics*, volume 162. John Wiley & Sons, 2009. ↰ p.71

Uri Shalit, Fredrik D. Johansson, and David Sontag. Estimating individual treatment effect: Generalization bounds and algorithms. In *ICML*, pages 3076–3085, 2017. ↰ p.2, ↰ p.14, ↰ p.46, ↰ p.56

Claude Elwood Shannon. A mathematical theory of communication. *Bell system technical journal*, 27(3):379–423, 1948. ↰ p.9

Ryusei Shingaki and Manabu Kuroki. Identification and estimation of joint probabilities of potential outcomes in observational studies with covariate information. In *NeurIPS*, 2021. ↰ p.48, ↰ p.67, ↰ p.123

Christopher A. Sims. Money, income, and causality. *American Economic Review*, 62(4):540–552, 1972. ↰ p.5, ↰ p.7, ↰ p.8

Stephen M. Smith. The future of fMRI connectivity. *Neuroimage*, 62(2):1257–1266, 2012. ↰ p.21

Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. In *NeurIPS*, pages 3483–3491, 2015. ↰ p.54

Paul T. Spellman, Gavin Sherlock, Michael Q. Zhang, Vishwanath R. Iyer, Kirk Anders, Michael B. Eisen, Patrick O. Brown, David Botstein, and Bruce Futcher. Comprehensive identification of cell cycle-regulated genes of the yeast *saccharomyces cerevisiae* by microarray hybridization. *Molecular Biology of Cell*, 9(12):3273–3297, 1998. ↰ p.41

Bharath K. Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Bernhard Schölkopf, and Gert R.G. Lanckriet. Hilbert space embeddings and metrics on probability measures. *JMLR*, 11:1517–1561, 2010. ↰ p.27

Xiaohai Sun. Assessing nonlinear Granger causality from multivariate time series. In *ECML*, pages 440–455, 2008. ↰ p.22, ↰ p.26

Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In *ICML*, pages 1139–1147, 2013. ↰ p.93, ↰ p.101

Matt Taddy, Matt Gardner, Liyun Chen, and David Draper. A nonparametric bayesian analysis of heterogenous treatment effects in digital experimentation. *Journal of Business & Economic Statistics*, 34(4):661–672, 2016. ↰ p.2, ↰ p.44, ↰ p.45

Eric J. Tchetgen Tchetgen and Kelesitse Phiri. Bounds for pure direct effect. *Epidemiology*, 25(5):775, 2014. ↰ p.97

Lu Tian, Ash A. Alizadeh, Andrew J. Gentles, and Robert Tibshirani. A simple method for estimating interactions between a treatment and a large number of covariates. *Journal of American Statistical Association*, 109(508):1517–1532, 2014. ↰ p.44, ↰ p.46

Jonathan Y.C. Ting and Amanda S. Barnard. Data-driven causal inference of process-structure relationships in nanocatalysis. *Current Opinion in Chemical Engineering*, 36:100818, 2022. ↰ p.1

Elizabeth Tipton and Robert B. Olsen. A review of statistical methods for generalizing from evaluations of educational interventions. *Educational Researcher*, 47(8):516–524, 2018. ↰ p.10

Tyler J. VanderWeele. On the distinction between interaction and effect modification. *Epidemiology*, 20(6):863–871, 2009. ↰ p.46

Tyler VanderWeele. *Explanation in Causal Inference: Methods for Mediation and Interaction*. Oxford University Press, 2015. ↰ p.92, ↰ p.97, ↰ p.98

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. In *NeurIPS*, pages 12388–12401, 2020. ↰ p.20

Marjolein Visser, Lex M. Bouter, Geraldine M. McQuillan, Mark H. Wener, and Tamara B. Harris. Elevated C-reactive protein levels in overweight and obese adults. *Journal of American Medical Association*, 282(22):2131–2135, 1999. ↰ p.63

Geoffrey S. Watson. Smooth regression analysis. *Sankhyā: Indian Journal of Statistics, Series A*, pages 359–372, 1964. ↰ p.52

Dominik Wied and Rafael Weißbach. Consistency of the kernel density estimator: A survey. *Statistical Papers*, 53(1):1–21, 2012. ↰ p.72

Yongkai Wu, Lu Zhang, and Xintao Wu. On discrimination discovery and removal in ranked data using causal graph. In *KDD*, pages 2536–2544, 2018. ↰ p.20, ↰ p.86

Yongkai Wu, Lu Zhang, and Xintao Wu. Counterfactual fairness: Unidentification, bound and algorithm. In *IJCAI*, pages 1438–1444, 2019. ↰ p.86

Yongkai Wu, Lu Zhang, Xintao Wu, and Hanghang Tong. PC-fairness: A unified framework for measuring causality-based fairness. In *NeurIPS*, pages 3399–3409, 2019. ↰ p.83, ↰ p.84, ↰ p.85

Depeng Xu, Yongkai Wu, Shuhan Yuan, Lu Zhang, and Xintao Wu. Achieving causal fairness through generative adversarial networks. In *IJCAI*, pages 1452–1458, 2019. ↰ p.20, ↰ p.86

Shun Yao, Shinjae Yoo, and Dantong Yu. Prior knowledge driven Granger causality analysis on gene regulatory network discovery. *BMC Bioinformatics*, 16(1):1–18, 2015. ↰ p.21

Xuan Yin and Liangjie Hong. The identification and estimation of direct and indirect effects in A/B tests through causal mediation analysis. In *KDD*, pages 2989–2999, 2019. ↰ p.20

Jinsung Yoon, James Jordon, and Mihaela Van Der Schaar. GANITE: Estimation of individualized treatment effects using generative adversarial nets. In *ICLR*, 2018. ↰ p.2, ↰ p.46, ↰ p.56

Junzhe Zhang and Elias Bareinboim. Equality of opportunity in classification: A causal approach. In *NeurIPS*, pages 3675–3685, 2018. ↰ p.86, ↰ p.93

Junzhe Zhang and Elias Bareinboim. Fairness in decision-making–the causal explanation formula. In *AAAI*, pages 2037–2045, 2018. ↰ p.86

Lu Zhang and Xintao Wu. Anti-discrimination learning: A causal modeling-based framework. *International Journal of Data Science and Analytics*, 4(1):1–16, 2017. ↰ p.20, ↰ p.77, ↰ p.86, ↰ p.90

Lu Zhang, Yongkai Wu, and Xintao Wu. A causal framework for discovering and removing direct and indirect discrimination. In *IJCAI*, pages 3929–3935, 2017. ↰ p.20, ↰ p.74, ↰ p.83, ↰ p.86, ↰ p.90

Lu Zhang, Yongkai Wu, and Xintao Wu. Causal modeling-based discrimination discovery and removal: Criteria, bounds, and algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 31(11):2035–2050, 2018. ↰ p.20, ↰ p.86

Qingyuan Zhao, Dylan S. Small, and Ashkan Ertefaie. Selective inference for effect modification via the lasso. *Journal of Royal Statistical Society: Series B (Statistical Methodology)*, 84(2):382–413, 2022. ↰ p.44, ↰ p.45, ↰ p.46, ↰ p.57, ↰ p.62

Pingping Zhu, Badong Chen, and Jose C. Principe. Learning nonlinear generative models of time series with a kalman filter in RKHS. *IEEE Transactions on Signal Processing*, 62(1):141–155, 2014. ↰ p.28, ↰ p.29, ↰ p.30