On Causal Discovery and Inference from Observational Data

by

Fujin Zhu

A THESIS SUBMITTED
IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE OF

Doctor of Philosophy

Centre for Artificial Intelligence (CAI), School of Computer Science
Faculty of Engineering and Information Technology (FEIT)
University of Technology Sydney

August, 2019

CERTIFICATE OF AUTHORSHIP/ORIGINALITY

I certify that the work in this thesis has not previously been submitted for a degree nor has it been submitted as part of requirements for a degree except as fully acknowledged within the text.

I also certify that the thesis has been written by me. Any help that I have received in my research work and the preparation of the thesis itself has been acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This research is supported by the Australian Government Research Training Program

Signature of Candidate

To my parents, brother, and wife.

ACKNOWLEDGEMENTS

First and foremost, I would like to express my sincere gratitude to my supervisors Prof. Guangquan Zhang and Prof. Jie Lu for their continuous encouragement and guidance in my PhD career. I really appreciate them for offering me the opportunity to study and do my research in the Centre for Artificial Intelligence (CAI, former QCIS), Faculty of Engineering and Information Technology (FEIT), University of Technology Sydney (UTS).

As my principle supervisor, Prof. Zhang is more like a father to me. I still remember the meetings we had together. He always gives me sufficient freedom to learn new knowledge and explore unknowns. When my research got stuck and I started to vacillate and loss faith in myself, he believed in me and encouraged me to calm down, take a break and then carry on. He teaches me to look faraway and to become an independent researcher. He gives me so much patience even though I learned far too slow. I am always grateful to him.

I am very lucky to have Prof. Lu as my co-supervisor. Prof. Lu is an excellent and respectable researcher. She loves her students and influences us with her continual passion for almost everything. Prof. Lu shares with us her experience in study, research and life. She teaches me how to write professional papers, discusses and listens to my problems in research, and provides comments and suggestions which are really constructive and make this thesis possible. I learn

a lot from her.

I am also indebted to the rest of my committee – Dr. Haiyan Lu and Dr. Ling Chen – who have kindly contributed their time to make me a better researcher. I would like to thank all the other faculty, students, and staff at CAI and FEIT who have helped and influenced me. They are Camila Cremonese, Lily Qian, Janet Stack, Yi Zhang, Hongshu Chen, Junyu Xuan, Shirui Pan, Jia Wu, Mingmig Gong, Weiwei Liu, Anjin Liu, Fan Dong, Qian Zhang, Shan Xue, Wei Wang, Peng Hao, Guanjin Wang, Chenlian Hu, Yiliao Song, Feng Liu, Adi Lin, Dian Ouyang, Daokun Zhang, Yuangang Pan, Pingbo Pan, Ruiqi Hu, Di Wu, Zhibin Li and Xiaofeng Xu.

I would also like to thank my friends, Daniel Kim and Kai Austin Zhang, for their companion and the joyful days we spent together in a foreign country. The scholarship for my study and funding for my academic travels have come from a variety of sources. They are the China Scholarship Council, UTS, CAI, the School of Computer Science, and the FEIT.

Finally and above all, I would like to thank my family for their support all the way. They are my parents, brother and wife. During this long journey, they always believed and supported me unconditionally. I dedicate this thesis to them. I especially thank my wife, Mang Chen, who has been taking care of my everyday life and shared all my pain, sorrow and joy during my study and research. Her optimism and love makes all the dark days hopeful. No words could express my gratitude and love to her.

Fujin Zhu

Sydney, Australia, 2019

ABSTRACT

Causality is a fundamental component in all fields of science. In contrast to associational dependencies that are widely used in existing predictive machine learning and data-mining methods, causality implies the mechanism of how variables take their values and how the change of causes would lead to the change in the outcome. In the era of big data, for scientific discovery and rational decision-making, we fundamentally need methods for learning causal relationships between variables and estimating causal effects from observational data.

In this thesis, we aim to develop new models and algorithms for learning causal relationships and estimating causal effects using observational data. In particular, for the purpose of modelling and learning causal relationships from observational data, we study dynamic causal systems with feedbacks. To overcome the weakness of existing models that are unable to model both instantaneous and cross-temporal causal relations simultaneously, we propose a First-order Causal Process (FoCP) model and a causal structure learning algorithm to learn the causal graph of FoCPs from time series. For the purpose of estimating treatment effects, we investigate a range of existing methods for causal effect estimation, and propose three new methods using advanced machine learning techniques. First, to relieve the high-variance issue of the classic Inverse Propensity Weighting (IPW) estimator and thus to get more stable treatment effect estimates, we reframe it to the importance sampling framework and propose a novel Pareto-smoothing method using the generalized Pareto distribution

from the extreme value statistics. Second, for causal inference with unobserved confounders, we take advantage of proxy variables and use deep latent variable models to model the underlying data-generating process. Building on recent advances in Bayesian inference and deep generative models, we propose a Causal Effect Implicit Generative Model (CEIGM). Finally, with an observation that most of existing methods for causal inference are essentially indirect in that they estimate the target treatment effects by first estimating other auxiliary quantities, we propose the idea of direct treatment effect estimation. Based on this idea, we further propose two deep neural networks for direct treatment effect estimation.

We evaluate all the methods proposed in this thesis using simulated, semi-simulated or real-world data. Experiment results show that they perform generally better than their competitors. Given the key importance of learning causality and causal inference in both theory and real-world applications, we argue that our proposed models and algorithms are of both theoretical and practical significance.

Dissertation directed by

Prof. Guangquan Zhang, Dist. Prof. Jie Lu, and Prof. Donghua Zhu

Center for Artificial Intelligence, School of Computer Science, FEIT

TABLE OF CONTENT

ACKNOWLEDGEMENTS **iv**

ABSTRACT **vi**

LIST OF FIGURES **xii**

LIST OF TABLES **xiv**

Chapter 1 Introduction **1**
1.1 Background 1
1.2 Motivations 3
1.2.1 Motivation for Learning Causal Relationships 4
1.2.2 Motivation for Estimating Causal Effects 5
1.3 Research Problems 7
1.3.1 Causal Discovery: Learning Causal Relationship 7
1.3.2 Causal Inference: Estimating Causal Effects 8
1.4 Contributions and Significances 9
1.5 Thesis Organization 11
1.6 Publications 12

Chapter 2 Literature Review **14**
2.1 Mathematical Languages of Causality 14
2.1.1 Structural Causal Models and Causal Graph 15
2.1.2 The Potential Outcome Framework 17
2.2 Causal Discovery: Learning Causal Relationship 19
2.2.1 Assumptions 20
2.2.2 Constraint-based Causal Structure Learning 22
2.2.3 Modelling Dynamic Causal Systems with Feedbacks 24
2.3 Causal Inference: Estimating Causal Effects 27
2.3.1 Propensity Score Methods 29
2.3.2 Counterfactual Inference Methods 31
2.3.3 Doubly Robust Methods 32

2.3.4 Other Machine Learning Methods 33

Chapter 3 First-order Causal Process for Causal Modelling with Instantaneous and Cross-temporal Relations 35
3.1 A Motivating Example 36
3.2 First-order Causal Process 38
3.2.1 Two-stage State Evolution 38
3.2.2 The FoCP and Properties 40
3.3 Graphical Representations for FoCPs 42
3.3.1 The 2-time Variable Causal Graph 42
3.3.2 Feature Causal Graph 43
3.3.3 A Transformation Procedure 43
3.4 Structure Learning for FoCPs 44
3.4.1 Conditional Independence based Structure Learning 45
3.4.2 FoCP Learning 45
3.4.3 Computational Complexity 48
3.5 Experimental Analysis 49
3.5.1 Baselines and Evaluation Metrics 50
3.5.2 Simulated Data 52
3.5.3 Application to Climate Data 54
3.6 Summary 56

Chapter 4 A Pareto-smoothing Method for Causal Inference using Generalized Pareto Distribution 57
4.1 Problem Setup 58
4.2 Preliminaries 62
4.2.1 Estimating Expected Potential Outcomes 62
4.2.2 IPW Estimator 64
4.2.3 Truncated IPW Estimator 65
4.2.4 Self-normalized IPW Estimator 67
4.3 Methodology 67
4.3.1 GPD Fitting 69
4.3.2 Weight Smoothing 72
4.3.3 Estimators 72
4.3.4 Asymptotic Analysis 74
4.4 Simulation Studies 75
4.4.1 Simulated Data 77
4.4.2 Semi-simulated Data: IHDP 81
4.5 Application to the NHEFS Data 83
4.6 Summary 86

Chapter 5 Counterfactual Inference with Hidden Confounders Using Implicit Generative Models 88
5.1 Problem Setup . . . 89
5.2 Preliminaries . . . 90
5.2.1 Structural Causal Models . . . 91
5.2.2 Implicit Generative Models . . . 92
5.3 Counterfactual Inference Using IGMs . . . 94
5.3.1 Latent Variable Modelling for Causal Models . . . 94
5.3.2 Lower Bound Objective . . . 96
5.3.3 Inference . . . 97
5.4 Experiments . . . 98
5.4.1 Evaluation Metrics and Baselines . . . 98
5.4.2 Semi-simulated Data: IHDP . . . 100
5.4.3 Real World Data: Jobs . . . 102
5.5 Summary . . . 104

Chapter 6 Direct Treatment Effect Estimation using Deep Neural Networks 105
6.1 Problem Setup . . . 107
6.1.1 Treatment Effect Estimation: An Illustrative Example . . . 107
6.1.2 Definition and Assumptions . . . 110
6.2 Preliminaries . . . 112
6.2.1 Treatment Effect Estimation via Response Modelling . . . 112
6.2.2 DNNs for Treatment Effect Estimation . . . 114
6.3 Direct Treatment Effect Estimation Using DNNs . . . 116
6.3.1 Direct Treatment Effect Estimation . . . 116
6.3.2 CENet: Causal Effect Neural Network . . . 118
6.3.3 BCENet: CENet with Balanced Representation Layers . . . 122
6.4 Experimental Studies . . . 126
6.4.1 Baselines and Evaluation Metrics . . . 126
6.4.2 Semi-simulated Data . . . 127
6.4.3 Real World Data . . . 131
6.4.4 Experiment on Synthetic Data . . . 133
6.5 Summary . . . 137

Chapter 7 Conclusion and Future Directions 138
7.1 Conclusion . . . 138
7.2 Future Directions . . . 140
7.2.1 Learning Causality for General Entities . . . 140
7.2.2 Causal Inference with Continuous Treatment . . . 141
7.2.3 High-dimensional Causal Inference and Variable Selection . . . 141

7.2.4 Learning Treatment Policy from Observational Data . . . 142
7.2.5 Causality-based Machine Learning 143

Appendices **144**
A.1 Estimation of ATT 145
A.2 Identifiability of Counterfactuals and Treatment Effects 147
A.3 Representation Balancing Metrics 147
A.3.1 Calculating the Empirical MMD 148
A.3.2 Approximating the Wasserstein Distance 148
A.4 Hyperparameters 149
A.5 Additional Experimental Results 151

BIBLIOGRAPHY **154**

LIST OF FIGURES

1.1 Several possible causal graphs. 4
1.2 Data-generating processes for the example. 6
1.3 A diagram for causal structure learning. 8
1.4 Thesis structure. 12

2.1 Several typical DAGs for conditional independence. 16
2.2 Causal relationship between variables in the potential outcome framework. 17
2.3 An example first-order DBN. 25

3.1 Graphical representations for the SHO system. 44
3.2 Performance comparison in the Gaussian noises context. 53
3.3 Performance comparison in the non-Gaussian noises context. . . . 54
3.4 The learned causal graphs for the GSOD climate data. 55

4.1 ATE estimation bias and standard error in terms of sample size n for the simulated low-dimensional covariate data. 79
4.2 ATE estimation bias and standard error in terms of sample size n for the simulated high-dimensional covariate data. 80
4.3 Box plot of estimated ATE estimates by different estimators for the IHDP data in different settings. 83
4.4 The mean and standard deviation of weight gains for smoking non-quitters and quitters in the NHEFS data. 84

5.1 The underlying generative model and inference models. 92
5.2 Visualization of the IHDP dataset 100
5.3 Visualization of the Jobs dataset 102

6.1 Data for the illustrative example. 108
6.2 The joint neural networks for direct treatment effect estimation. . 116
6.3 Neural network architecture of CENet. 120
6.4 Neural network architecture of BCENet. 124
6.5 Visualization of the simulated data 134

6.6 Performance in terms of the imbalance parameter η when sample size $n = 1000$ for the simulated data. 136
6.7 Performance in terms of the sample size n when the imbalance parameter $\eta = 0.05$ for the simulated data. 136

A.5.1 ϵ_{ITE}^{out} in terms of the imbalance parameter η for different sample size n. 151
A.5.2 ϵ_{ITE}^{out} in terms of sample size n for different imbalance parameter η. 152
A.5.3 ϵ_{ATE}^{out} in terms of the imbalance parameter η for different sample size n. 152
A.5.4 ϵ_{ATE}^{out} in terms of sample size n for different imbalance parameter η. 153

LIST OF TABLES

4.1 Abbrivation (Abbr.) and description of ATE estimators 76
4.2 Performance comparisons for the simulated low-dimensional covariate data. 78
4.3 Performance comparisons for the simulated high-dimensional covariate data. 78
4.4 ATE estimation biases and standard errors (SE) for the IHDP dataset. 82
4.5 Estimation results for the NHEFS dataset. 85

5.1 Within-sample and out-of-sample results on the IHDP dataset . . 101
5.2 Within-sample and out-of-sample results on the Jobs dataset . . . 103

6.1 Within-sample and out-of-sample results on the Twins dataset . . 129
6.2 Within-sample and out-of-sample results on the IHDP dataset . . 131
6.3 Within-sample and out-of-sample results on the Jobs dataset . . . 133

A.3.1Hyperparameters and ranges 150
A.3.2Optimal hyper-parameters for CENet on each dataset 150
A.3.3Optimal hyper-parameters for BCENet on each dataset 150

Chapter 1

Introduction

1.1 Background

Recent advance in information collection and storage have made a huge amount of observational data available for researchers and policy makers. Empowered by the growing collection of big data and advances in computing power, during the last decade, predictive machine learning and data-mining algorithms [49, 177] have made spectacular progress, surpassing human performance in face recognition [208], natural language understanding [64], machine translation [14, 25], self-driving [17], etc. In general, these algorithms have mainly focused on eliciting *associational* knowledge and patterns from the collected data and making associational predictions on new data.

To make scientific conclusions and rational decisions, we, however, need to answer causal questions [9, 133], understand the causal relationships between variables or events [41, 214, 223], and estimate the possible change or difference in the outcome caused by a particular treatment or policy variable [61, 74]. For instance, in biology, scientists conduct experiments to discover genes for certain genotypes; in healthcare, patients need to known the effect of the medication

on their health to decide whether to take a particular medication or not; in economics, policy makers debate the possible effect of job training on employees' earning; and in marketing, what ad companies are really interested is the causal effect of an online advertisement on customers' purchasing habits.

As a core topic in science and philosophy, the research of causality (or causation) can be dated back to the early 1700s. The philosopher Aristotle wrote in his Physics that "we do not have knowledge of a thing until we have grasped its why, that is to say, its cause." In addition, David Hume defined causality as "'constant conjunction' in his book A Treaties of Human Nature [69]. Causality connects one phenomenon (the *cause*) with another (the *effect*) and establishes the relationship that the former is (partly) responsible for the latter to happen [221]. In the last three decades, the research on causality has been rejuvenated. Moreover, researchers from the Artificial Intelligence (AI) community argue that the ability to learn causality from data is a significant component of human-level intelligence [97, 101, 133, 138], and causal inference is a central topic for both scientific discovery and decision-making [9, 74, 105].

While causal claims have been used as common sense on a daily basis, making a causal conclusion from an observed dependence is not straight forward. A causal relation indicates the mechanism of the cause determines the effect, and the causal effect measures the strength of this determination process. In principle, scientists usually conduct randomized controlled trails (RCTs) to discover causal relationships and to make causal claims [44, 152]. In a RCT, subjects are randomly assigned to different treatment groups. Except for the treatment, all other conditions are considered identical for subjects. As a result, if we observe any statistically significant difference between these groups, we can then attribute it to the influence of the treatment variable and claim that the treatment is a cause for the difference.

Though RCTs are golden standard for learning causality and estimating causal effects, in many real-world applications, it will be expensive, unethical, or even impossible to conduct RCTs [61, 74]. As a result, we focus on learning causal relationships and estimating causal effects from purely observational data in this thesis.

1.2 Motivations

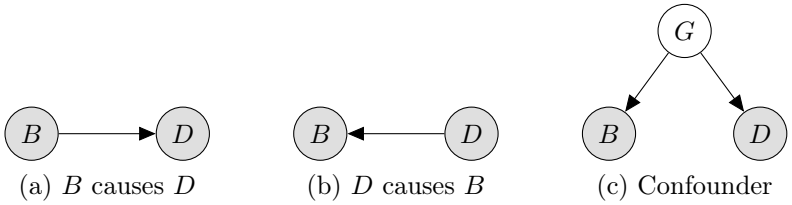
In his recent book *The Book of Why*, Judea Pearl, a Turing Award winner and pioneering researcher in AI, states that we are in an era of *Causal Revolution* and a human-like AI must master three levels of cognitive ability by climbing the *Ladder of Causation* [1] [135]: seeing, doing, and imagining.

While the ability of *seeing* and making observational predictions is based on associational analysis, we need to learn causality to acquire the ability of *doing* via intervention and *imagining* via introspection. To explain the motivation for studying causality in a high level, we cite the words of Pearl from [134]:

> *The answer to the question "why study causation?" is almost as immediate as the answer to "why study statistics." We study causation because we need to make sense of data, to guide actions and policies, and to learn from our success and failures ... causation is not merely an aspect of statistics; it is an addition to statistics, and enrichment that allows statistics to uncover workings of the world that traditional methods alone cannot.*

This thesis focus on learning causal relationships and estimating causal effects from observational data. Specifically, we explain the motivations of our research

[1] http://bayes.cs.ucla.edu/WHY/why-ch1.pdf



(a) B causes D (b) D causes B (c) Confounder

Figure 1.1: Several possible causal graphs that might explain the association between the biomarker B and the disease D. In (c), G stands for unobserved gene defect.

by showing that *association does not imply causation* and *associational prediction is different from causal prediction* in this section.

1.2.1 Motivation for Learning Causal Relationships

Without distinguishing association from causation, many observed associations have been over-interpreted as causal relationships in the literature. For example, in [115], Messerli observed a positive correlation between chocolate consumption and the number of Nobel laureates from a country, and claimed that chocolate consumption causes the sprouting of Nobel laureates. This is obviously a misinterpretation and the suggestion for winning more Nobel Prize by consuming more chocolate would be really farcical. To explain the difference between *association* and *causation*, and to emphasize the importance of understanding the causal relationship between variables in real-world applications, consider the simple example discussed in [84]:

In clinical diagnosis, suppose we have detected a positive correlation between the occurrence of a biomarker B and a disease D. To predict the occurrence of D, we can conduct a blood test and use the occurrence of B as a predictor for the disease D. After the diagnosis, doctors are often interested in a cure for this disease. For this purpose, if we could infer that the biomarker is a cause of the disease (i.e., B cause D Fig.1.1(a)), we may suggest that manipulating the

biomarker B in a proper way will cure the disease.

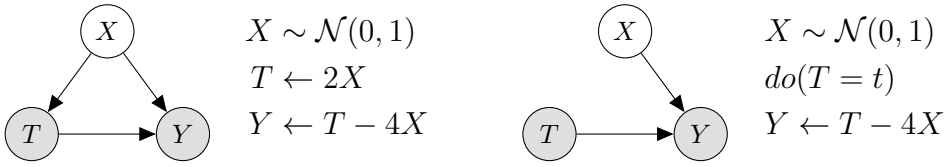
However, besides this causal relationship $B \rightarrow D$, it is easy to find at least two other scenarios that could explain the observed dependence, and the association between the occurrence of the biomarker and the disease may be explained by three causal relationships shown in Fig.1.1: (a) the causal relationship from the occurrence of the biomarker to the disease, (b) the causal relationship from the disease to the occurrence of the biomarker, and (c) there is no causal relationship between them but they both are caused by a common unobserved variable (e.g., gene defect).

As a result, we can conclude that while it is enough to make predictive conclusions using associational relationships, we need to understand the underlying causal relationship between variables to take effective actions. Since it is not trivial to infer causation from observed associations, we argue that the problem of how to learn causal relationships from observational data is of both theoretical and practical importance.

1.2.2 Motivation for Estimating Causal Effects

We have shown that *associational* relationships does not necessarily entail *causal* relationships. In this section, we introduce the motivation for estimating causal effects by showing the difference between *associational prediction* and *causal prediction*.

Given an actionable variable T that we can actively intervene to set its value and an outcome variable Y, associational prediction tries to estimate the conditional expectation of Y if we see that $T = t$, and causal prediction aims to estimate the conditional expectation Y if we set T to take the value t (denoted as $do(T = t)$). Obviously, the ability to make associational prediction of $\mathbb{E}[Y|T = t]$ reflects the cognitive ability of *seeing* and the ability to make causal prediction



(a) *seeing*:observational model (b) *doing*:interventional model

Figure 1.2: Data-generating processes for the example.

of $\mathbb{E}[Y|do(T = t)]$ reflects the cognitive ability of *doing* along Pearl's *Ladder of Causation* [135].

To explain their difference explicitly, consider the data-generating process as illustrated in Fig.1.2(a). With this model, suppose we see that $T = 1$, then according to the associational relationship between X and T, we get that $X = 0.5$. Consequently, we can easily make the associational prediction as

$$\mathbb{E}[Y|T = 1] = 1 - 4 \times 0.5 = -1$$

In fact, we can derive the close-form formula for associational prediction as $\mathbb{E}[Y|T = t] = -t$ for this example. For the purpose of causal prediction, we now obstruct the data-generating process and perform intervention $do(T = 1)$. The intervened data-generating model, then, becomes one shown in Fig.1.2(b) and we can make a causal prediction that

$$\mathbb{E}[Y|do(T = 1)] = 1 - 4 \times 0 = 1$$

As we can see from the example, the observational expectation $\mathbb{E}[Y|T = t]$ and the interventional expectation $\mathbb{E}[Y|do(T = t)]$ can be very different and even have different valences. This suggests that we need to be very cautious in making causal predictions using observational data and it is important to develop ad-hoc methods for causal inference in order to make rational decisions.

1.3 Research Problems

The objective of this thesis is to develop new models and algorithms for learning causality from observational data. This can be further divided into two problems: (1) How to learn causal relationships from observational data; and (2) How to estimate causal effect from observational data. The most significant point of this research is to use advanced machine learning techniques for learning causality and estimate causal effects. We introduce the above two research problems in this section and list four specific tasks for solving them as follows:

i. To conduct studies of causal modelling and learning for causal systems with feedback causal relationship by using time series data and incorporating temporal asymmetry into causal modelling. (Chapter 3)

ii. To investigate existing weighting estimators for causal inference and propose methods to improve their functionality. (Chapter 4)

iii. To conduct studies of causal inference with proxy variables and develop a latent-variable model-based causal inference method. (Chapter 5)

iv. To investigate recent advanced machine learning techniques so as to adapt them to develop direct treatment effect estimation algorithms. (Chapter 6)

1.3.1 Causal Discovery: Learning Causal Relationship

Causality connects one process (the *cause*) with another (the *effect*) and establishes the relationship that the former is (partly) responsible for the latter to happen [221]. Given its fundamental importance in science and the accumulation of observational data in the big data era, our first research problem is whether

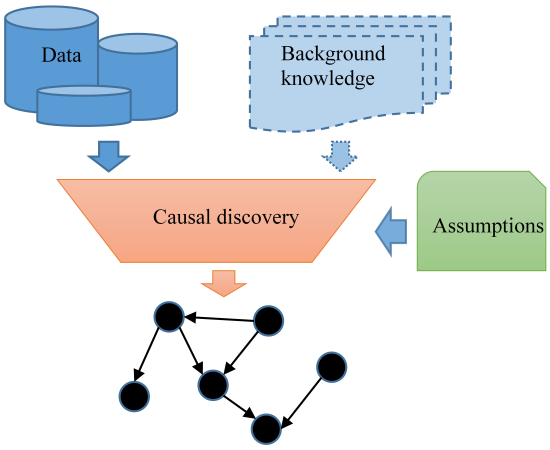


Figure 1.3: A diagram for causal structure learning.

we can learn the underlying causal relationships from observational data. This problem is also known as causal structure learning or causal discovery. We formally define the problem of learning causal relationships from observational data as follows:

Problem 1 (Causal Discovery, or Causal Structure Learning)**.** *Given a set of data and assumptions, how to infer the causal relationship between the variables and reveal the underlying causal graph $\mathcal{G}$ that have generated the observed data?*

The diagram for causal structure learning is illustrated in Fig.1.3. While most of existing work considers learning the causal structure of static systems, in this thesis, we study causal structure learning of dynamic systems from time series.

1.3.2 Causal Inference: Estimating Causal Effects

Causal structures discussed in the previous section indicates which variables have a direct effect on other variables. However, it does not indicate how strong this effect is (or whether it is positive or negative). In this section, we introduce the second research problem of estimating causal effects.

Causal effect estimation is also called causal inference [74] and is the problem of estimating the treatment effect of some intervention on a target outcome variable. We formally define it as follows:

Problem 2 (Causal Effect Estimation, or Causal Inference)**.** *For a target population, how can we consistently quantify and estimate the causal effect of the treatment T on an outcome variable Y?*

In this thesis, we investigate existing methods for causal inference and develop new treatment effect estimators based on advanced machine learning techniques.

1.4 Contributions and Significances

The main contributions of this thesis are summarized as follows:

- We propose a novel First-order Causal Process (FoCP) and a corresponding learning algorithm for modelling and learning the causal structure of dynamic causal systems with both instantaneous and cross-temporal causal relations.
- We reframe the classic IPW estimator for causal inference as an importance sampling method for estimating expectations and propose a Pareto-smoothing method for weight stabilization to achieve more stable propensity weighted causal estimates.
- We model the underlying data-generating process of a causal model with proxy variables using deep generative models and propose a causal effect implicit generative model (CEIGM) for treatment effect estimation.
- We propose direct treatment effect estimation as a complementary of existing indirect causal inference methods. In addition, we propose two neural network architectures for direct treatment effect estimation.

In this research, we investigate existing methods and develop new methods for learning causal relationships and estimating causal effects from observational data. In particular, we proposed models for representing and learning dynamic causal systems with feedbacks. New causal estimators are also proposed for treatment effect estimation. All the proposed methods are validated using simulated and real-world dataset.

In data science, most existing machine learning and data mining algorithms aim to achieve high predictive accuracy for a target outcome variable [177]. This goal is mainly realized by learning associational relationships between the observed variables from a training dataset and extend the elicited associational knowledge to make predictions on new data. In contrast, causality implies the mechanism of how variables take their values and how the change of causes would lead to the change in the outcome. In contrast to associational relationships, causality is universal and is a fundamental component in all fields of science [162]. Given the fundamental importance of causality in all these disciplines, our research is of theoretical significance.

Methods for learning causal relationships have been used in many real world applications. For example, in bio-informatics, learning Bayesian networks have been used for the interpretation and discovery of gene regulatory pathways [164] [94], analysing information flow in brain networks [37], and recovering causal relationships from neuro-imaging data [209]. In addition, causal inference methods have also been used in many real-world applications, including precise medicine [8], computational advertisement [18], social program evaluation [9, 62] and machine learning systems [191, 192]. As a result, the methods proposed in this thesis are of practical significance. We have also applied our proposed methods to real-world climate data analysis (Chapter 3), heath data analysis (Chapter 4), and job training program evaluation (Chapter 5 and Chapter 6).

1.5 Thesis Organization

Te remainder of this thesis is organised as follows:

- *Chapter 2*: This chapter presents a literature review of existing methods for causal discovery and causal inference.
- *Chapter 3*: This chapter studies causal structure learning from time series. We propose graphical representation models and a structure learning algorithm for dynamic systems with both instantaneous and cross-temporal relations.
- *Chapter 4*: This chapter investigates the classic IPW-based estimators for causal inference. To ease their high-variance and unstable issue, we develop a Pareto-smoothing method for stabilizing importance weights used in these estimators and propose two Pareto-smoothed causal estimators.
- *Chapter 5*: This chapter focuses on causal inference with proxy variables using counterfactual inference. We model the data-generating process using latent-variable models and then estimate treatment effects using fitted counterfactuals.
- *Chapter 6*: This chapter studies causal inference using deep neural networks. We propose a new idea for treatment effect estimation, i.e., modelling and learning the target treatment effect function directly. In addition, we extend on this idea by proposing two deep neural network architectures for direct treatment effect estimation.
- *Chapter 7*: We give a brief summary of the thesis contents and its contributions in this final chapter. Future research directions are also discussed.

We illustrate the structure of this thesis in Fig.1.4 .

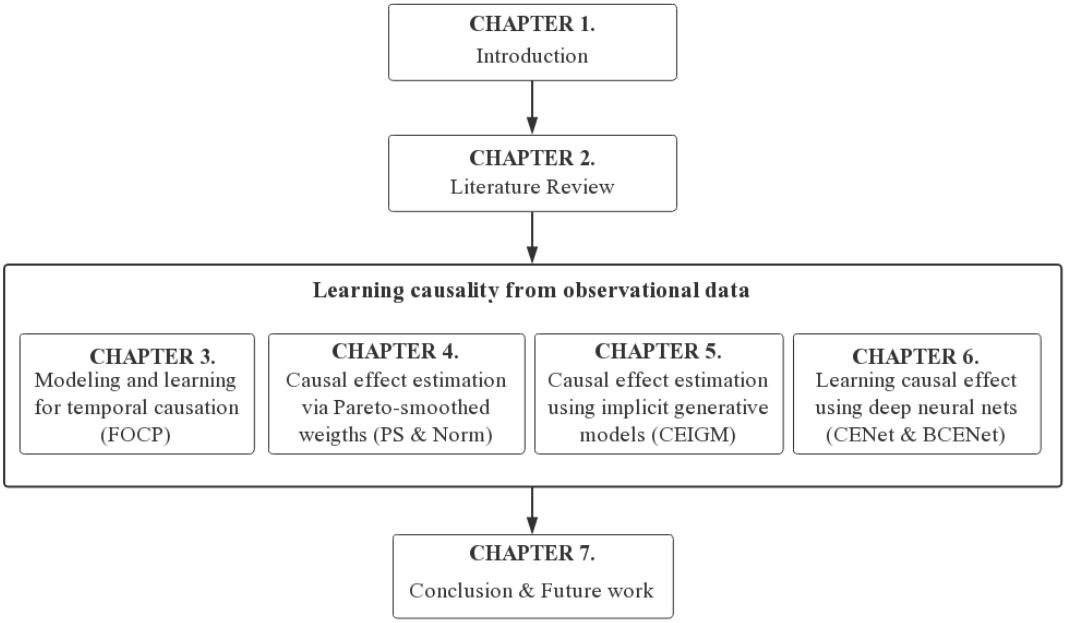


Figure 1.4: Thesis structure.

1.6 Publications

Conference Papers

C-1. **Fujin Zhu**, Adi Lin, Guangquan Zhang, and Jie Lu, "Counterfactual Inference with Hidden Confounders Using Implicit Generative Models," *Australasian Joint Conference on Artificial Intelligence*, pp.519-530, Dec. 11-14, 2018. Springer, Cham. **Best Student Paper Award.**

C-2. **Fujin Zhu**, Adi Lin, Guangquan Zhang, Jie Lu, and Donghua Zhu, "Pareto-smoothed inverse propensity weighing for causal inference," *Data Science and Knowledge Engineering for Sensing Decision Support: Proceedings of the 13th International FLINS Conference*, pp.413-420, Aug. 21-24, 2018.

C-3. **Fujin Zhu**, Guangquan Zhang, Jie Lu, and Donghua Zhu, "First-order Causal Process for Causal Modelling with Instantaneous and Cross-temporal Relations," *2017 International Joint Conference on Neural Networks (IJCNN)*, pp.380-387, May. 14-17, 2017. IEEE.

C-4. Adi Lin, Jie Lu, Junyu Xuan, **Fujin Zhu**, and Guangquan Zhang, "One-stage Deep Instrumental Variable Method for Causal Inference from Observational Data," *19th IEEE International Conference on Data Mining (ICDM)*, 2019. (To appear).

C-5. Ying Huang, **Fujin Zhu**, Ying Guo, Alan L. Porter, Yi Zhang, and Donghua Zhu, "Exploring technology evolution pathways to facilitate technology management: a study of Dye-sensitized solar cells (DSSCs)," *2016 Portland International Conference on Management of Engineering and Technology (PICMET)*, pp.764–776, May. 14-17, 2016. IEEE. **Brad W. Hosler Outstanding Student Paper Award.**

Journal Papers

J-1. **Fujin Zhu**, Jie Lu, Adi Lin, and Guangquan Zhang, "A Pareto-smoothing Method for Causal Inference using Generalized Pareto Distribution," DOI: https://doi.org/10.1016/j.neucom.2019.09.095, Neurocomputing, 2019.

J-2. **Fujin Zhu**, Jie Lu, Adi Lin, Donghua Zhu, and Guangquan Zhang, "Causal Effect Neural Networks for Direct Treatment Effect Estimation," In submission.

J-3. Adi Lin, Jie Lu, Junyu Xuan, **Fujin Zhu**, and Guangquan Zhang, "Causal DP: A New Bayesian Nonparametric Prior for Causal Inference," Transactions on Intelligent Systems and Technology, 2019. Under Review.

Chapter 2

Literature Review

In fields such as epidemiology [21, 149], economics [12, 73, 75], political science [7, 45, 54] and statistics [152, 161], an observational study draws inference from a sample to a population where the independent variable is not under the control of the researcher. One of the most important tasks is to understand the underlying causal mechanism and further estimate the causal effect of a treatment on subjects. In this chapter, we will give a brief review of two popular mathematical languages for communicating causality and existing methods for causal relationship learning and treatment effect estimation.

2.1 Mathematical Languages of Causality

There are mainly two languages for causal modelling and reasoning: Pearl's Structural Causal Models (SCMs) based on the *do*-calculus [131, 134, 136], and Rubin's potential outcomes framework [74, 160]. The two languages of causality are complementary to each other and have different strengths suitable for different tasks. For a more comprehensive understanding of both frameworks, we refer the readers to [55, 73].

2.1.1 Structural Causal Models and Causal Graph

From an interventional perspective, we say that X causes Y ($X \to Y$) if the conditional distribution of Y changes upon intervening on X. That is, $P(Y|do(X = x)) \neq P(Y|do(X = x'))$ where the *do*-operator $do(X = x)$ is proposed by Pearl [131, 132] and corresponds to forcing X to take the value x. This is the underlying idea of the structural causal model language for causality. In a SCM, to specify the quantitative characteristics of the causal variables, we specify a structural equation for each variable X_i as:

$$X_i = f_i\left(\boldsymbol{pa}(X_i), U_i\right), \quad i = 1, 2, \ldots, n \tag{2.1}$$

where $\boldsymbol{pa}(X_i)$ are the parents (or direct causes) of X_i, $\{U_1, U_2, , U_n\}$ are the error variables, and f_i is the causal mechanism for assigning the value of X_i from its parents and other disturbance errors. We can also represent these causal relationships with a causal graph defined as

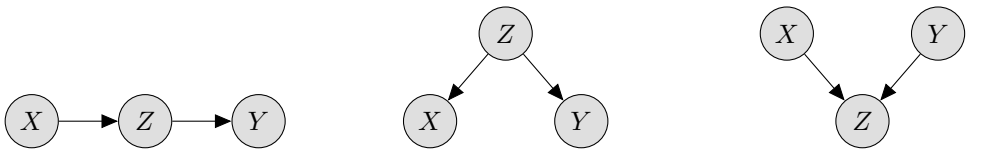
Definition 2.1 (Causal Graph [55])**.** *A Causal graph $\mathcal{G} = (\boldsymbol{V}, \boldsymbol{E})$ is a directed graph that describes the causal relationship between variables, where $\boldsymbol{V}$ is the set of nodes and $\boldsymbol{E}$ the set of all edges. In a causal graph, each node $X_i \in \boldsymbol{V}$ represents a causal variable; a direct edge $X \to Y$ denotes a causal relationship from X to Y.*

SCMs use causal graphs to represent causal relationships and are very popular for modelling and learning causal relationships. Usually, a causal graph $\mathcal{G}$ for $\boldsymbol{V} = \{X_1, X_1, \cdots, X_n\}$ is a directed acyclic graph (DAG) and is called a Causal Bayesian Network (CBN) [131, 183]. By taking advantage of their probabilistic reasoning ability and causal semantics, CBNs play a very important role for causal modelling and reasoning. The set of parents of a node X_i in the causal graph $\mathcal{G}$, denoted as $\boldsymbol{pa}(X_i)$, corresponds to the direct causes of the X_i. Every

f_i in Eq.(2.1) is a conditional distribution $P(X_i|\boldsymbol{pa}(X_i))$ that encodes an autonomous and modular causal mechanism for generating the value of X_i. With this representation, a CBN $\mathcal{G}$ implicitly encodes the joint distribution over $\boldsymbol{V}$ as a product of local conditional distributions:

$$P(\boldsymbol{V}) = P\left(X_1, \cdots, X_n\right) = \prod_{X_i \in \boldsymbol{V}} P\left(X_i|\boldsymbol{pa}\left(X_i\right)\right) \tag{2.2}$$

In a causal graph $\mathcal{G} = (\boldsymbol{V}, \boldsymbol{E})$, two nodes X and Y are *adjacent* if there is an edge between them. We denote the set of all adjacencies of any node X in $\mathcal{G}$ as $\texttt{Adjaceny}_{\mathcal{G}}(X)$. The *parents* of a node X are all nodes from which an arrow points to X. Correspondingly, all nodes to which an arrow points from X are called *children* of X. A *path* is a sequence of distinct, adjacent nodes. A path is directed if all edges on the path are directed in the same direction (e.g., Fig.2.1(a)). The *ancestors* of X are all nodes from which a directed path leads to X. Correspondingly, the *descendants* of X are all nodes to which a directed path leads from X. The *skeleton* of a causal graph is the undirected graph that is obtained when ignoring the directions of the arrows.



$X \not\perp\!\!\!\perp Y; X \perp\!\!\!\perp Y \mid Z$ (a) Directed path

$X \not\perp\!\!\!\perp Y; X \perp\!\!\!\perp Y \mid Z$ (b) Fork

$X \perp\!\!\!\perp Y; X \not\perp\!\!\!\perp Y \mid Z$ (c) Collider

Figure 2.1: Several typical DAGs for conditional independence.

Given an SCM, we can directly read out conditional independences by checking its causal graph represented as a CBN. Fig.2.1 shows several typical patterns and their conditional independence information. In Fig.2.1(a), X causally affects Y through its influence on Z. We call it a directed path or chain with Z

a blocking variable between X and Y. In Fig.2.1(b), X and Y are affected by a common cause Z and they are not directly affected by each other. This pattern is called a *fork* and the common cause Z is a *confounder*. In Fig.2.1(c), X and Y are independent to each other, but have a common effect variable Z. This is called a *collider* and acts as a key pattern used in constraint-based structure learning we will introduce later.

2.1.2 The Potential Outcome Framework

In statistics and economics [5, 67], many researchers focus on estimating the effects of some intervention or policy, without reference to a structural and graphical model. In these applications, the treatment (i.e., intervention or policy) and the outcome variables are generally clear by background knowledge. Originated from the literature on experimental and observational studies [44, 154, 155, 176], Rubin's potential outcome framework [74, 160] is the most popular language for defining treatment effects and conducting treatment effect estimation. To facilitate the introduction of different concepts and assumptions, we illustrate a general causal graph used in the potential outcome framework in Fig.2.2.

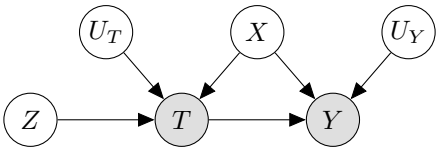


Figure 2.2: Causal relationship between variables in the potential outcome framework.

Our objective is to estimate the causal effect of the *treatment* T on some *outcome* Y, i.e., the strength of the causal relationship $T \to Y$. In observational studies, both T and Y are influenced by the confounders X. There are also independent errors U_T and U_Y influencing them respectively. In many cases, we

may also find some *instrumental variables* Z that influence the outcome Y only through the treatment T.

In this framework, the treatment and outcome variables are clear from background knowledge. Potential outcomes are formally defined as

Definition 2.2 (Potential Outcome [55])**.** *For a population of n individuals $\{X_1, X_2, \ldots, X_n\}$, the potential outcome of individual X_i under treatment t is defined as the value the outcome variable Y would take if the treatment had been set to t and is denoted as $Y_i(t)$.*

In settings with binary treatment T, the treatment effect for individual i can be easily defined as $\tau_i = Y_i(1) - Y_i(0)$. Estimation of the individual treatment effect τ_i from observational data is fundamentally impossible since we can never observe both treatment outcomes $Y_i(0)$ and $Y_i(1)$ simultaneously. This is called *the fundamental problem of causal inference* [66, 155] in the causal inference literature. Alternatively, we can estimate the average treatment effect (ATE) for the population defined as

Definition 2.3 (Average Treatment Effect, ATE)**.** *In the setting with binary treatment T, the ATE is defined as*

$$\tau_{ATE} = \mathbb{E}[Y(1) - Y(0)]$$

Similarly, the average treatment effect for the treated subpopulation is defined as

Definition 2.4 (Average Treatment Effect for the Treated, ATT)**.** *In the setting with binary treatment T, the ATT is defined as*

$$\tau_{ATT} = \mathbb{E}[Y(1) - Y(0)|T = 1]$$

To incorporating the heterogeneity in treatment effects, the ATE for the subpopulation with features $X = x$ is defined as the conditional treatment effect (CATE) or heterogeneous treatment effect (HTE) as follows

Definition 2.5 (Conditional Average Treatment Effect, CATE)**.** *Given a subpopulation with features $X = x$ and binary treatment T, the CATE is a function of x and is defined as*

$$\tau_{CATE}(x) = \mathbb{E}[Y(1) - Y(0)|X = x]$$

To obtain consistent treatment effect from observational data, standard assumptions are essential and two basic assumptions are given below:

Assumption 2.1 (Consistency)**.** *For any individual i, $T_i = t$ implies $Y_i(t) = Y_i$.*

Assumption 2.2 (Stable Treatment Value Assumption, SUTVA)**.** *Each individual's potential outcomes are independent to the actual treatment assignment of other individuals.*

The consistency assumption [27, 146, 149, 201] indicates that the potential outcomes for an individual do not change no matter which treatment alternative is observed. The SUTVA [158, 159] indicates no interference among individuals, i.e., the potential outcomes of any individual are unrelated to other's treatment assignment.

2.2 Causal Discovery: Learning Causal Relationship

The problem of learning causal relationship from observational data is also known as causal structure learning or causal discovery. A broad range of methods

has been developed for this task [60, 165, 184, 224], including independence constraint-based structure learning [122, 184, 215], score and search-based structure learning [174, 185], and hybrid methods (e.g., the MMHC algorithm [195] and the Hybrid HPC algorithm [48]). Specifically, constraint-based methods learn causal Bayesian networks with conditional independence tests through analysing the probabilistic relations entailed by a set of assumptions. Score-based methods assign a goodness-of-fit score (e.g., the BIC score [173], the BDe score [59] and the K2 score [30]) to each candidate network for measuring how well the candidate network fits the dataset. Hybrid methods combine ideas from constraint-based methods and score-based methods.

To facilitate theoretical foundations for learning causal relationship, we firstly introduce several common assumptions for causal structure learning used in this thesis. Based on these assumptions, we provide a brief review of existing constraint-based structure learning methods, with an emphasis on the seminal PC algorithm [184]. In addition, since we focus on causal modelling and learning for dynamic causal systems, we also introduce existing models for representing and learning dynamic causal systems.

2.2.1 Assumptions

The first assumption is the causal sufficiency condition defined as:

Assumption 2.3 (Causal Sufficiency Condition, CSC [184])**.** *A set of observed variables $\boldsymbol{V}$ is causally sufficient for a causal system if and only if for all potential causal dependencies, all common causes are measured and included in $\boldsymbol{V}$.*

This condition indicates that all variables in the target causal system are measured and there exists no latent confounder or selection bias. The second assumption is the causal Markov condition (CMC) defined as

Assumption 2.4 (Causal Markov Condition, CMC [184]). *Let $\mathcal{G}$ be a causal graph with a set of variables $\boldsymbol{V}$ and P be a probability distribution over $\boldsymbol{V}$ generated by a causal model structured as $\mathcal{G}$. $\langle \mathcal{G}, P \rangle$ satisfies the Causal Markov Condition if and only if for every $X \in \boldsymbol{V}$, X is independent of its non-descendant variables given its parents $\boldsymbol{pa}(X)$.*

That is to say, every variable is independent of any subset of its non-descendant variables conditioned on its parents. By utilizing the CMC, the distribution of variables X in a Bayesian network $\langle \mathcal{G}, P \rangle$ can be factorized as Eq.(2.2). By virtue of CMC and the d-separation criterion [130] in DAGs and the m-separation criterion [145] in directed cyclic graphs, for distinct variables X , Y, and a disjoint set $\boldsymbol{S}$, we have:

$$X \perp\!\!\!\perp_{\mathcal{G}} Y \mid \boldsymbol{S} \Rightarrow X \perp\!\!\!\perp_{P} Y \mid \boldsymbol{S}$$

where $X \perp\!\!\!\perp_{\mathcal{G}} Y \mid \boldsymbol{S}$ is read as X and Y are d-separated [130] or m-separated [145] in $\mathcal{G}$ conditional on $\boldsymbol{S}$, and $X \perp\!\!\!\perp_{P} Y \mid \boldsymbol{S}$ indicates the probabilistic conditional independence of X and Y given $\boldsymbol{S}$. While the CMC specifies independence relationships among variables, the following causal faithfulness condition (CFC) specifies dependence relationships:

Assumption 2.5 (Causal Faithfulness Condition, CFC [184]). *A probability distribution P over a set of random variables $\boldsymbol{V}$ is called faithful relative to the graphical structure $\mathcal{G}$ on variables $\boldsymbol{V}$ if and only if every conditional independence relation of X and Y given $\boldsymbol{S}$ true in P is entailed by the CMC applied to $\mathcal{G}$, i.e.,*

$$X \perp\!\!\!\perp_{P} Y \mid \boldsymbol{S} \Rightarrow X \perp\!\!\!\perp_{\mathcal{G}} Y \mid \boldsymbol{S}$$

If $\mathcal{G}$ is a causal graph and $\langle \mathcal{G}, P \rangle$ satisfies the Faithfulness Condition, we say they are faithful to each other, and $\mathcal{G}$ is called a faithful causal graph. For a pair

of $\langle \mathcal{G}, P \rangle$ that satisfies both CMC and CFC, we can conclude that

$$X \perp\!\!\!\perp_{\mathcal{G}} Y \mid \boldsymbol{S} \Leftrightarrow X \perp\!\!\!\perp_{P} Y \mid \boldsymbol{S} \tag{2.3}$$

This indicates that we can discover causal independence relations in the underlying causal graph $\mathcal{G}$ by testing probabilistic conditional independences in P. And this is the key idea for the causal structure learning algorithms introduced in the next section.

2.2.2 Constraint-based Causal Structure Learning

Independence constraint-based structure learning algorithms learn causal relationships by testing whether certain conditional independences between causal variables hold [184]. We start with a fully connected graph. The conditional independence constraints are propagated throughout the graph and edges inconsistent with them are removed. A sound strategy for performing conditional independence tests ultimately retains only the statistically equivalent graphs consistent with the constraints. Typically, conditional independence tests are performed using statistical or information theoretic measures (e.g., G^2 test [123], χ^2 test [123], entropy-based test [39], kernel-based tests [119, 219] and non-parametric [189] conditional independence tests). Examples of independence constraint-based causal structure learning algorithms include the IC-algorithm [136], Grow-Shrink [110], SGS algorithm [184], PC algorithm [83, 184], and many others. Among them, the PC-algorithm is very popular and has been extended into additional algorithms to relax assumptions of the original algorithm to different data representations (e.g., the FCI algorithm [184], the CPC algorithm [143], the RPC algorithm [109] and the RFCI algorithm [29]).

As a core component used in the proposed structure learning method in this

Algorithm 2.1 PC Algorithm

Input: A data set $\mathcal{D}$ of the set of observational variables $\boldsymbol{V}$, and a conditional independence test method

Output: The causal graph $\mathcal{G}$

Begin with the fully connected undirected graph $\mathcal{G}$ on $\boldsymbol{V}$;
$n = 0$;
for Each adjacent pair X–Y with $|\texttt{Adjaceny}_{\mathcal{G}}(X)\backslash Y| \geq n$ or $|\texttt{Adjaceny}_{\mathcal{G}}(Y)\backslash X| \geq n$ **do**
  for Any $\boldsymbol{S} \subseteq \texttt{Adjaceny}_{\mathcal{G}}(X)\backslash Y \cup \texttt{Adjaceny}_{\mathcal{G}}(Y)\backslash X$ and $|\boldsymbol{S}| = n$ **do**
    Test whether X and Y are conditional independent given $\boldsymbol{S}$;
    if $X \perp\!\!\!\perp_P Y \mid \boldsymbol{S}$ **then**
      $\texttt{Sepset}(X, Y) \leftarrow \boldsymbol{S}$;
      Delete the edge X–Y in $\mathcal{G}$;
      $n \leftarrow n + 1$
    end if
  end for
end for
for Each triple of nodes X, Y, Z **do**
  if (X, Z) and (Y, Z) are adjacent and (X, Y) are not adjacent in $\mathcal{G}$ **then**
    Orient $X - Z - Y$ as $X \rightarrow Z \leftarrow Y$ if $Z \notin \texttt{Sepset}(X, Y)$ [Collider Detection];
  end if
end for
for Each triple of nodes X, Y, Z **do**
  if X, Y are not adjacent and $X \rightarrow Z - Y$ in $\mathcal{G}$ **then**
    Orient $Z - Y$ as $Z \rightarrow Y$ [Known Non-Collider Detection];
  end if
end for
for Each pair of nodes X, Y **do**
  if $X - Y$ and there is a directed path from X to Y in $\mathcal{G}$ **then**
    Orient $X - Y$ as $X \rightarrow Y$ [Cycle Avoidance];
  end if
end for

thesis, the PC algorithm is formalized in Algorithm 2.1 and can be outlined in three steps [84]. In the first step, the skeleton of the DAG is estimated. For doing this, the algorithm starts with a complete undirected graph. Then for each edge (say $X - Y$), we search and test whether there exists a separation set

$\boldsymbol{S}$ (denoted as $\texttt{Sepset}(X, Y)$) such that $X \perp\!\!\!\perp_P Y \mid \boldsymbol{S}$. If $\texttt{Sepset}(X, Y)$ is not null, then according Eq.(2.3), we get that $X \perp\!\!\!\perp_{\mathcal{G}} Y \mid \boldsymbol{S}$ and delete the edge $X - Y$ from the original graph. In the second step, the obtained separation sets are used to orientate unshielded colliders. Given an unshielded triplet $X - Z - Y$, if Z is not in $\texttt{Sepset}(X, Y)$, then Z is a collider and we can orientate them as $X \rightarrow Z \leftarrow Y$. In the third step, all undirected edges are checked and we orientate as many of them as possible to avoid new colliders and cycles. In this step, we assume all colliders have detected in the second step, and thus for an unshielded triplet $X \rightarrow Z - Y$, we orientate it as $X \rightarrow Z \rightarrow Y$ to avoid new colliders. Similarly, for a direct path $X \rightarrow Z \rightarrow Y$, if X and Y are adjacent, we orientate the edge $X \rightarrow Y$ to avoid cycles in the causal graph.

2.2.3 Modelling Dynamic Causal Systems with Feedbacks

The causal graphs we have introduced so far are acyclic, i.e., there is no feedback loops in the target causal system. However, causal feedbacks exist widely in real world systems in economics, biology, environmental sciences, and engineering. In general, these systems are dynamic in nature and data collected from them are multivariate time series [19]. In this setting, time provides an additional source of information as well as a new challenge for modelling and analysing these dynamic systems. Whether it is possible and how to discover the causal structure with feedbacks from time series is an essential problem for understanding the nature and measuring the effect of interventions in these dynamic systems [40, 42, 134].

Eichler provided a comprehensive discussion on how to define causality for time series in [42]. To model dynamic systems with feedbacks, we can assume the observational data is obtained from the equilibrium distribution [117] and represent the corresponding system by a static directed cyclic graph (DCG). Though being able to model feedbacks among different variables via directed

cycles, this category of models simply assume self-edges which are essential in any self-excitatory or self-inhibitory dynamic process do not exist [113]. For this reason, we give a brief review of two main approaches that are able to model self-edge feedbacks [228]: Dynamic Bayesian Networks (DBNs) [46, 120] and the Granger causality [52].

Dynamic Bayesian Networks

DBNs are dynamic versions of causal graphical models [93]. In a DBN, time is modelled as proceeding in discrete steps. A DBN contains a graph $\mathcal{G}$ over the set of random variables $\boldsymbol{V}$ at the current time-step t as well as nodes for $\boldsymbol{V}$ at each previous time-step in which there is a direct cause of the current values of $\boldsymbol{V}$. A simple example DBN for three random process is illustrated in Fig.2.3.

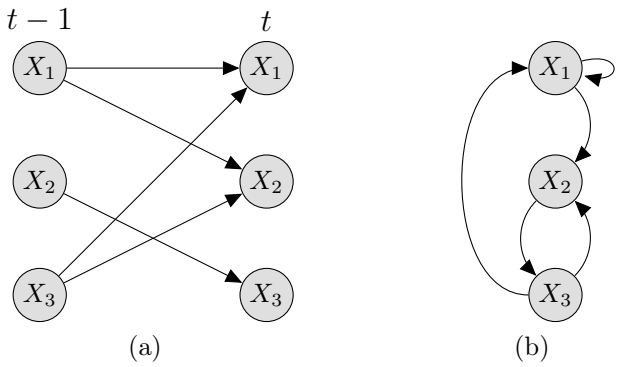


Figure 2.3: An example first-order DBN. (a) Unrolled graphical representation; (b) Compact graphical representation.

In a DBN, if the parents of variables at time t are those from the previous step $t-1$, we call it a *first-order* DBN. For a first-order DBN, we can easily represent the underlying causal dynamics using a Bayesian network of variables from two adjacent time slices, $\boldsymbol{V}(t-1)$ and $\boldsymbol{V}(t)$. Such a graphical representation is called a 2-time-slice Bayesian network (2-TBN) [93]. For more compact representation, we can also model a 2-TBN in a direct graph with possible self-loops as in

Fig.2.3(b). As can be seen from the figure, a key limitation of DBNs for modelling dynamic causal systems is that no instantaneous causal relations are permitted. That is, there are no direct causal relations between variables in the same time step.

Granger Causality

Granger causality has been widely used for modelling causal relationships from economic time series [52] and has a long history of applications across a wide range of domains [162]. A random process X is said to Granger cause another random process Y if and only if there is some unique information in X relevant for Y that is not contained in Y's past as well as the past of "all the information in the universe". Granger causal analysis usually assumes a linear model and can be written in the form of a Vector Auto-regressive (VAR) model [182] as:

$$\boldsymbol{V}(t) = \boldsymbol{A}\boldsymbol{V}(t-1) + \boldsymbol{\epsilon}(t)$$

where the coefficient matrix $\boldsymbol{A}$ contains temporal causal relations and is called the causal transition matrix.

Granger causality and VARs are limited to temporal lagged causal relationships. As an extension, the Structural Vector Auto-regressive (SVAR) model [118] considers both instantaneous and lagged relations between variables and can be formalized in a matrix form as

$$\boldsymbol{V}(t) = \boldsymbol{A}\boldsymbol{V}(t-1) + \boldsymbol{B}\boldsymbol{V}(t) + \boldsymbol{\epsilon}(t)$$

where the coefficient matrix $\boldsymbol{B}$ contains instantaneous causal relations.

Despite their popularity in prediction-oriented dynamic system modelling, all these models are essentially regression models and can merely discover con-

strained statistical associations. To make causal claims for the learnt relations, we need additional domain constraints or assumptions.

2.3 Causal Inference: Estimating Causal Effects

In the last section, we have introduced models for representing causal relationships and methods for learning causal structure from observational data. The obtained causal structure indicates whether a variable has a direct effect on other variables. However, it does not indicate how strong this effect is (or whether it is positive or negative). In this section, we will review methods for causal inference, i.e., estimating the causal effect from observational data.

Causal effects are usually measured as the difference between the treated and control groups in a randomized control trial (RCT). Take precise medicine [8] as an example, when doctors want to identify the efficiency of some medicine on a disease, the golden standard is to conduct a double-blind RCT where patients are randomly assigned into either the treated (taking medicine) or the control group (not taking medicine) and the treatment effect of the medicine is measured as the difference of recovery outcomes between the two groups. Through RCTs, the randomized treatment assignment mechanism ensures the observed outcomes will not be confounded by measured or unmeasured covariates, e.g., age, gender, and health status of the patients.

However in many real world applications, RCTs are expensive, unethical, or even impossible [61]. As a result, researchers have mainly focused on observational studies that conduct causal inference using purely observational data [155, 74]. Causal inference from observational data is challenging because we can only observe the outcome corresponding to the treatment received by an individual, while the outcome under the alternative treatment is unobserved.

This is called the fundamental problem of causal inference [155] and is known in the machine learning literature as counterfactual learning [18] and *learning from logged bandit feedback* [191, 192]. Moreover, the underlying treatment assignment mechanism that determines which outcome was observed is unknown and generally not random. This results in a covariate imbalance issue among the treatment groups and makes observational causal inference even more challenging. To tackle these problems, a range of methods using advanced machine learning techniques have been proposed in the literature [57, 172, 178, 206] and can be roughly grouped into three categories: Quasi-experiment methods, counterfactual inference methods, and doubly robust methods.

Quasi-experiment methods try to transform the collected observational data to mimic a balanced one as from RCTs. This is mainly realized via adjusting the biased treatment assignment mechanism. There are mainly two groups of Quasi-experiment methods for causal inference [88]: matching and weighting. Matching methods [153, 188] assume that similar individuals should have similar treatment outcomes, and estimate the unobserved counterfactual outcome for each individual by matching him or her with individuals in the counterpart group. Examples of matching methods include nearest neighborhood matching [157], propensity score matching [188], kernel matching [227], genetic matching [38] and optimal matching [85]. In general, matching methods are conceptually intuitive in that they maintain the units of analysis intact in an attempt to approximate an RCT [16]. One apparent drawback of matching methods is that they do not necessarily use all the observed data, in that some unmatched individuals are discarded and not used in the final treatment effect estimation [188].

In contrast, weighting methods use all observational data for causal inference and thus are more statistical efficient. There are two broad groups of approaches

for weighting [65]. The first group derives the weight for each individual by estimating the treatment model (e.g., inverse propensity weighting [63, 68]). The second group of weighting methods derives weights by directly reducing covariate imbalance across treatment groups. Examples include kernel balancing [58], entropy balancing [56, 222], adversarial balancing [129], etc.

In the following sections, we give a brief review of matching and weighting methods by estimating the treatment assignment model, counterfactual inference methods that models the potential outcomes and doubly robust methods that estimate both the treatment and the potential outcomes. In addition, we also give a brief review of recent causal inference methods using advanced machine learning techniques such as deep neural networks (DNNs) and latent variable modelling.

2.3.1 Propensity Score Methods

The propensity score defined as the conditional probability of the assignment to the treatment group given the observed covariates [155]: $e(x) = P(T = 1|X = x)$. Theoretically, we can account for the difference between the treatment and control groups by directly modelling the assignment mechanism with propensity scores, thus making populations from different groups more comparable [160]. In [155], the authors show that if the treatment assignment is ignorable given the observed covariates, the ATE can be consistently estimated by adjusting for the propensity scores alone. Given this balancing and de-biasing property, propensity score-based approaches have been widely used for causal inference from observational data [74], complete-case analysis for missing data [175], and survey sampling [166]. They have also recently been adopted by the data mining and machine learning communities for de-biasing in recommender systems [99, 167], information retrieval systems [207] and learning to rank systems [79].

There are many methods for causal inference using propensity scores, such as IPW [63, 68], stratification [155] and propensity score matching (PSM) [36, 111]. Propensity score methods for causal inference usually proceed in a two-step procedure. In the first step, the propensity scores are estimated using statistical methods (e.g., Logistic regression). Then the treatment effects are estimated by weighting the outcome using the inverse of the estimated propensity score for each individual or matching individuals from the control group to individuals from the treated group based on the similarity of the estimated propensity scores. These methods have been shown to be consistent and effective in estimating ATEs [63, 188].

Despite their popularity and theoretical appeal, a practical problem of propensity-based methods is that the true propensity scores are intrinsically unknown and must be estimated from finite observational data in pure observational studies. Research indicates that misspecification of the propensity score model can result in substantial bias in causal effect estimation. If the estimated propensity scores are close to one or zero for a substantial fraction of the population, the estimated causal effect may be of high variability and difficult to estimate precisely [11]. This is a particular concern in settings with many covariates, or simply when the assignment mechanism is highly skewed. When many of the estimated propensity scores are close to zero, the distribution of their reciprocals – the inverse propensity weights – are likely to have a heavy right tail, which leads to unstable estimates of treatment effects, sometimes with infinite variance.

To address the problem of variability, two methods for variance control in importance sampling, weight truncation and weight self-normalization, have been used to stabilize the importance weights-based estimators in the causal inference community [61, 74]. Researchers from the sampling and weighting community

have proposed a growing list of techniques for variance reduction [128]. We will discuss this problem in detail and propose a new method in Chapter 4.

2.3.2 Counterfactual Inference Methods

Instead of matching or weighting, Hill [62] observed that, if there is no unobserved confounders, we can estimate the treatment effect – the outcome difference between the treated group and the control group – by fitting the treatment outcome models $\mu(x,t) = \mathbb{E}[Y|x,t]$. By this observation, we are actually transforming the problem of causal inference into a predictive learning problem: learning the unknown counterfactual outcome models from bandit observational data. As a result, this method is also called counterfactual inference [62, 80, 169, 225].

Counterfactual inference for causal inference relies on a model of the potential outcome defined in Definition 2.2. Note that these potential outcomes are complementary in that, for any individual i, only one potential outcome is observable and the others are *counterfactuals* [35]. To complete the task of treatment effect estimation, counterfactual inference methods approximate the underlying treatment response functions using the conditional mean functions fitted via regression or other supervised machine learning algorithms. Given its strength in detecting interactions and discontinuities in modelling, the non-parametric Bayesian additive regression tree (BART) model [24] has been widely used for counterfactual inference. Recently, more advanced machine learning models such as Gaussian process [2, 144, 169] and DNNs [80, 178] have also been adapted for causal inference under the counterfactual inference framework.

Recently, Künzel et al. [95] categorize counterfactual inference methods for treatment effect estimation into S-learning that infers a single treatment response function for all treatments and T-learning that infers a treatment response function for each treatment. Although recent advances in non-parametric Bayesian

models [3, 62, 89] and DNNs permit us to learn very complex conditional mean functions and have achieved relatively good performance for treatment effect estimation [80, 178], counterfactual inference is originally designed for answering counterfactual questions such as *would this employee get higher salary had she received some job training?* For the task of treatment effect estimation, they solve it in an indirect way. As a result, treatment outcome functions fitted to minimize the prediction error for the observed outcomes are not guaranteed to produce accurate treatment effect estimation. Recently, experiment results in [225] indicate that counterfactual inference methods may learn the two treatment outcome surfaces quite well, but in opposite error directions, resulting unstable treatment effect estimations.

2.3.3 Doubly Robust Methods

While propensity score-based methods use information of the treatment assignment mechanism and rely on correct specification of the treatment propensity model $e(x) = P(T = 1|X = x)$, the counterfactual inference methods rely on correct specification of the treatment response model $\mu(x, t) = \mathbb{E}[Y|x, t]$. They are sensitive to the misspecification of the propensity score model or treatment response model [86]. To tackle this limitation, building on the basis of the semiparametric theory , doubly robust estimators [15, 47] model both the treatment and outcome models and, remarkably, give consistent treatment effect estimates as long as one of these two nuisance models is correctly specified.

Robins et al. [150] and Rotnitzky et al. [156] introduced the first doubly robust method called augmented inverse probability weighting (AIPW) to combine propensity score and treatment response modelling in estimating the ATE. Later, this idea is further extended in [15, 198]. Another well-known doubly robust technique is the targeted maximum likelihood estimation (TMLE, also

called targeted learning) [104, 197, 199, 200]. Take ATE as an example, TMLE requires initial estimates of the treatment propensity model $e(x)$ treatment response model $\mu(x,t)$, which can be flexibly estimated by ensemble and machine learning algorithms. Then a substitution "targeting" step that optimizes the bias-variance trade-off for the target ATE is conducted to get the final estimate.

Doubly robust estimators, in contrast to the causal estimators introduced in the previous two sections, provide us a double guarantee to make valid causal inference. Existing simulation-based evidence shows that they often perform better than their propensity score-based or counterfactual inference-based counterparts [104, 125, 199].

2.3.4 Other Machine Learning Methods

Besides the methods introduced in the above sections, researchers have also proposed various advanced methods that use machine learning for treatment effect estimation, including Lasso [142], support vector machines [72], trees and forests [10, 206], ensemble methods [141], meta learners [95], DNNs [23, 57, 80, 179] and Bayesian machine learning [2, 62, 102], seeking to provide more accurate treatment effect estimation under certain assumptions for different causal problems.

Recently, Schwab et al. [172] summarizes existing treatment effect estimation methods based on machine learning into the following five categories: (1) Matching-based methods that estimate the unobserved counterfactual outcome of an individual using the observed outcomes of other individuals by matching on some metric space; (2) Adjust regression methods that fit a single treatment outcome model with both covariates and the treatment indicator as inputs or multiple treatment outcome models, one for each treatment; (3) Tree-based methods that train many weak learners to build expressive ensemble models to non-parametrically learn the treatment outcome surfaces (e.g., BART [62])

or the CATE function (e.g., the Causal Tree [10] and Causal Forests [206]); (4) Representation-balancing methods that seek to learn balanced representations that are similar between the treated and control groups so as to alleviate the imbalanced treatment assignment in observational studies. Methods in this category include kernel balancing [58], entropy balancing [56, 222], balanced neural networks (BNN) [80], counterfactual regression (CFR) [178] and the local Similarity-preserved Individual Treatment Effect method (SITE) [212]; (5) Distribution-modelling methods that model the underlying treatment responses using probabilistic machine learning. Specifically, Causal Multi-task Gaussian Process [2] adapts multi-task Gaussian process to model the two treatment response surfaces. Causal Effect Variational Auto-encoder (CEVAE) [102] and Causal Effect Implicit Generative Model (CEIGM) [225] use deep latent-variable modelling approaches to model and infer the outcome generative probability. Generative Adversarial Nets for inference of Individual Treatment Effects (GANITE) [213] uses generative adversarial networks (GANs) to learn the counterfactual and ITE generators.

Among these methods, except for the Causal Tree and Causal Forests, other methods estimate the target treatment effect by modelling either the treatment responses or the treatment assignment mechanism using various machine learning techniques, and are thus indirect methods.

Chapter 3

First-order Causal Process for Causal Modelling with Instantaneous and Cross-temporal Relations

In this chapter, we propose a new causal model for representing and learning the causal relationship between variables in dynamic systems with both instantaneous and cross-temporal causal relations. Our proposed model is motivated by the *causal process* idea [40, 163] from the philosophy of science and some real physical systems. We introduce a novel 2-stage evolution semantic for dynamic systems with both instantaneous and cross-temporal causal relations. The first-order causal process (FoCP) is proposed to model these dynamic systems, compact graphical representations and a conditional independence test based structure learning algorithm are also proposed for FoCP.

The main content of this chapter has been published in [226], and the remainder of this chapter is organised as follows: in Section 3.1, we introduce a classic

dynamic system as a motivating example. Based on our observations about this motivating example, in Section 3.2, we conclude the 2-stage evolution semantic for dynamic systems and propose the FoCP model; properties of the new model are also discussed. In Section 3.3, we introduce graphical representations for the FoCP model. In Section 3.4, we propose an unified conditional independence based structure learning algorithm for the model by using a similar structure learning procedure of the classical PC algorithm [184] and the identified useful properties. Experiments on both simulated and real data are conducted in Section 3.5 to validate the model and algorithm. Finally, we summarize this chapter in Section 3.6.

3.1 A Motivating Example

Our illustrative example is adapted from [205]. Consider in a damped simple harmonic oscillator (SHO) system, a block of mass m is suspended from a spring in a special viscous fluid. There are several sources of forces acting on this block, including the gravity, the support, the spring force, and the fluid resistance. Denote the displacement and velocity of the block at time t as $x(t)$ and $v(t)$ respectively. In mechanics, the spring force and the fluid resistance at the moment t are determined by the displacement $x(t)$ and the velocity $v(t)$. Thus, we denote them as $F_x(t)$ and $F_v(t)$ respectively. In addition, denote the time interval as Δt, according to the momentum formula $m(v(t+\Delta t)-v(t)) = (F_x(t) + F_v(t))\,\Delta t$ and Newton's law of motion, we can easily conclude that $x(t + \Delta t) - x(t) = v(t)\Delta t$, $F_x(t) = -kx(t)$ and $F_v(t) = -\delta v(t)$, where k and δ are damping coefficients of the spring and the viscous fluid respectively. Obviously, since the coefficients k and δ are constant, the SHO system is a linear dynamic system.

In this SHO system, we observe that there are two kinds of causation be-

tween the system variables: the *cross-temporal causation* from an antecedent variable to a later variable, such as the influence from $F_x(t)$, $F_v(t)$ to $x(t+\Delta t)$, and the *instantaneous causation* between contemporary variables, such as the influence from $v(t)$ to $F_v(t)$. Moreover, we notice that a variable can be influenced by at most one kind of causation, either *instantaneous* or *cross-temporal*. Specifically, the displacement $x(t)$ and the velocity $v(t)$ are directly determined by *cross-temporal* causation, and the forces $F_x(t)$ and $F_v(t)$ are determined by instantaneous causation from the current system variables. This is a very important property of many physical dynamic systems and will be used to develop graphical models and structure learning algorithms in the following sections.

Denote the state vector of this linear dynamic system at time t as $s(t) = (F_x(t), F_v(t), v(t), x(t))^T$, the coefficients matrix for instantaneous causal transitions as $\boldsymbol{B}^{inst}$, and the coefficient matrix for temporal causal transitions as $\boldsymbol{B}^{temp}$, then the discrete approximation of the underlying causal mechanism (assume additive noises) can be formalized as the following structural equations:

$$\begin{aligned} F_x(t) &\leftarrow b_{14}^{inst} x(t) + e_1(t) \\ F_v(t) &\leftarrow b_{23}^{inst} v(t) + e_2(t) \\ v(t) &\leftarrow b_{33}^{temp} v(t-\Delta t) + b_{31}^{temp} F_x(t-\Delta t) + b_{32}^{temp} F_v(t-\Delta t) + \epsilon_3(t) \\ x(t) &\leftarrow b_{44}^{temp} x(t-\Delta t) + b_{43}^{temp} v(t-\Delta t) + \epsilon_4(t) \end{aligned} \tag{3.1}$$

where b_{ij}^{inst} and b_{ij}^{temp} are elements from the instantaneous coefficients matrix $\boldsymbol{B}^{inst}$ and the temporal coefficients matrix $\boldsymbol{B}^{temp}$ respectively. We can also write the structural equations in E.q(3.1) in a compact formalization as:

$$\begin{aligned} s(t) &= \widehat{\boldsymbol{B}}^{inst} \left(\boldsymbol{B}^{temp} s(t-\Delta t) + \epsilon(t) \right) + e(t) \\ &= \widehat{\boldsymbol{B}}^{inst} \boldsymbol{B}^{temp} s(t-\Delta t) + \widehat{\boldsymbol{B}}^{inst} \epsilon(t) + e(t) \end{aligned} \tag{3.2}$$

where $\widehat{\boldsymbol{B}}^{inst}$ is the augmented instantaneous transition matrix which augments the instantaneous transition matrix $\boldsymbol{B}^{inst}$ by setting the diagonal elements of zero rows to 1 to make sure the value of variables without contemporary parents are unchanged after this transformation; $\epsilon(t) = (0, 0, \epsilon_3(t), \epsilon_4(t))$ and $e(t) = (e_1(t), e_2(t), 0, 0)$ are additive temporal transition error and instantaneous error vectors. In this research, we assume causal sufficiency (Assumption 2.3) and all error terms are additive and mutually independent.

We can easily conclude from the above state evolutions in E.q(3.1) that, the displacement $x(t)$ is determined by the previous displacement $x(t-\Delta t)$ and velocity $v(t-\Delta t)$; similarly, the velocity $v(t)$ is determined by the previous spring force $F_x(t-\Delta t)$, viscosity $F_v(t-\Delta t)$ and velocity $v(t-\Delta t)$. While for $F_x(t)$ and $F_v(t)$, their values are directly determined by the contemporary variables $x(t)$ and $v(t)$ respectively, but not directly influenced by any variables from previous time slices. To obtain the state of the dynamic system at time $t+\Delta t$, $s(t+\Delta t)$, from $s(t)$, we can imaginably decompose the state evolutionary process into two stages: firstly, the velocity $v(t+\Delta t)$ and displacement $x(t+\Delta t)$ get their values from their directed parents at time t by the cross-temporal transition matrix $\boldsymbol{B}^{temp}$; then, $F_x(t+\Delta t)$ and $F_v(t+\Delta t)$ are instantaneously determined by the values of $v(t+\Delta t)$ and $x(t+\Delta t)$ through the instantaneous transition matrix $\boldsymbol{B}^{inst}$. This is directly entailed by the compact representation in E.q (3.2).

3.2 First-order Causal Process

3.2.1 Two-stage State Evolution

Motivated by the SHO example in the previous section, we postulate that a dynamic system is intrinsically a series of functional transformation of states. As a process, the system state $s(t)$ at time t is first transformed via a temporal causal

transformation $\mathcal{H}$ to the latent state $s^{in}(t+\Delta t)$, and then reaches its observed state $s^{out}(t+\Delta t)$ at time $t+\Delta t$ by the instantaneous causal transformation $\mathcal{F}$. Mathematically, we can generalize the underlying dynamics of the SHO system as the following state evolution process:

$$\begin{aligned} s^{in}(t+\Delta t) &= \mathcal{H}(s^{out}(t)) \\ s^{out}(t+\Delta t) &= \hat{\mathcal{F}}\left(s^{in}(t+\Delta t)\right) \end{aligned} \tag{3.3}$$

where $\hat{\mathcal{F}}$ is the augmented instantaneous transformation which augments the instantaneous causal transformation $\mathcal{F}$ by leaving variables without contemporary parents unchanged. We coin a causal dynamic system with such a 2-stage evolution semantic a *first-order causal process (FoCP)*. The term *first-order* means that we only consider *lag-one* cross-temporal causal relations for the temporal causal transformation $\mathcal{H}$. Apparently, the equivalent observational model of a FoCP is $s^{out}(t+\Delta t) = \hat{\mathcal{F}} \circ \mathcal{H}\left(s^{out}(t)\right)$.

In practice, most real-world dynamic systems are intrinsic stochastic, even in deterministic systems, noises coming from unobserved exogenous variables and measurements are often unavoidable. In this research, we assume all noises are additive and mutually independent. Thus, the 2-stage state evolution of a dynamic system can be generally formalized as:

$$\begin{aligned} s^{in}(t+\Delta t) &= \mathcal{H}\left(s^{out}(t)\right) + e^{in}(t+\Delta t) \\ s^{out}(t+\Delta t) &= \hat{\mathcal{F}}\left(s^{in}(t+\Delta t)\right) + e^{out}(t+\Delta t) \end{aligned} \tag{3.4}$$

For the simplest case, i.e., the additive noise linear dynamic system, the underlying causal dynamics is:

$$\begin{aligned} s^{in}(t+\Delta t) &= \boldsymbol{B}^{temp} s^{out}(t) + e^{in}(t+\Delta t) \\ s^{out}(t+\Delta t) &= \widehat{\boldsymbol{B}}^{inst} s^{in}(t+\Delta t) + e^{out}(t+\Delta t) \end{aligned} \tag{3.5}$$

Further define $u(t+\Delta t) = \widehat{\boldsymbol{B}}^{inst} e^{in}(t+\Delta t) + e^{out}(t+\Delta t)$, we have the observational version of the 2-stage state evolution as the following equation:

$$s(t+\Delta t) = \widehat{\boldsymbol{B}}^{inst} \boldsymbol{B}^{temp} s(t) + u(t+\Delta t) \tag{3.6}$$

By the observation that, in a dynamic system that falls into the scope of FoCPs, there may exist at most one kind of causal relation from X_j to X_i, either the *instantaneous causation* $X_j \to X_i$ or *cross-temporal causation* $X_j \dashrightarrow X_i$, or neither of them, it is easy to verify that $\widehat{\boldsymbol{B}}_{ij}^{\text{inst}} \boldsymbol{B}_{ij}^{\text{temp}} = 0$. Moreover, by the intuition *feedbacks need time to take effect*, existing static models that represent the causal system by a DAG or DCG can be regarded as special cases of the proposed model since they cannot model the self-loop feedbacks.

3.2.2 The FoCP and Properties

We now give a formal definition of the first-order causal process model:

Definition 3.1 (First-order causal process, FoCP)**.** *Denote the state vector of a discrete time-invariant dynamic system at time t as $s(t)$. A discrete time-invariant dynamic system is called a first-order causal process if (i) the value of any variable X_i is directly determined either by some variables at the previous time slice or some contemporary variables, or neither of them; (ii) feedbacks occur only through cross-temporal causal relations.*

Set the time interval of a discrete-time dynamic system to $\Delta t = 1$, a very similar model is the lag-one structural VAR (SVAR(1))[118], which models a linear dynamic system and is formalized in a compact matrix formation as:

$$s(t) = \boldsymbol{B} s(t) + \boldsymbol{A} s(t-1) + e(t) \tag{3.7}$$

where the matrix $\boldsymbol{B}$ illustrate the instantaneous causal relations between variables, and $\boldsymbol{A}$ contains their cross-temporal causal relations. However, this is actually a regression model where regression coefficients are interpreted into causality. By transforming the formula into the *reduced* form in E.q(3.8) which is similar with E.q(3.6), we claim that the SVAR(1) model can be treated as a special FoCP with additive errors.

$$(I - \boldsymbol{B})s(t) = \boldsymbol{A}s(t-1) + e(t)$$

$$s(t) = (I - \boldsymbol{B})^{-1}\boldsymbol{A}s(t-1) + (I - \boldsymbol{B})^{-1}e(t) \tag{3.8}$$

We now conclude several properties of the FoCP, which will be used to develop graphical representations in the next section and the structure learning algorithm in Section 3.4. Specifically, in a dynamic system whose dynamics can be modelled by a FoCP with T time slices, the following properties hold:

- **Property 1.** A direct cross-temporal causal relation is only possible from an antecedent variable to a later one;
- **Property 2.** At any time slice t, the contemporary causal structure over variables forms a DAG;
- **Property 3.** For any two variables X_i and X_j, $X_i(t) \to X_j(t) \Leftrightarrow X_i(t') \to X_j(t')$ where $t, t' \in \{1, 2, \ldots, T\}$;
- **Property 4.** For any two variables X_i and X_j, if $t_2 - t_1 = t_2' - t_1' > 0$, then $X_i(t_1) \to X_j(t_2) \Leftrightarrow X_i(t_1') \to X_j(t_2')$;
- **Property 5.** Any variable can at most be determined by either instantaneous causation from contemporary variables or cross-temporal causation from antecedent variables. This property can be further divided into: **Property (5a)**: If $X_i(t) \to X_i(t + \Delta t)$, then X_i has no contemporary

parents; **Property (5b)**: If $X_i(t) \to X_j(t + \Delta t)$, then X_j has no contemporary parents; and **Property (5c)**: If $X_i(t) \to X_j(t)$, then X_j has no antecedent parents.

In general, Property 1 is obvious by the arrow of time. Property 2 is entailed by the assumption that "*feedbacks take time to happen*". Property 3 and 4 represent the time-invariant property of a dynamic system we assumed earlier. Property 5 together with its three sub-properties is entailed by our assumption of the FoCP model: the value of a system variable X_i can only determined either by some antecedent variables or contemporary variables, or neither of them.

3.3 Graphical Representations for FoCPs

Denoting the ith variable at time t as $X_i(t)$ or X_i^t, We can define a *class* variable for the set of temporal variables $\{X_i(t), t = 1 : T\}$ as a feature X_i. Based on these notations, we introduce two equivalent graphical representations for the FoCP model in this section: the *2-time variable causal graph* (2TVCG), and the more compact *feature causal graph* (FCG).

3.3.1 The 2-time Variable Causal Graph

Similar to the compact graphical representation of 2-time-slice Bayesian networks (2TBN) in DBN introduced in Chapter 2.2.3, the 2TVCG of a FoCP is a DAG obtained by truncating the unrolled full causal graph and leaving only variables and edges in two adjacent time slices, let us say time slice t and $(t + 1)$. In a 2TVCG, instantaneous causal relations among variables in time slice $(t + 1)$ are explicitly illustrated while those among variables in time slice t are implicitly entailed by their latter copies; cross-temporal causal relations are illustrated

by arrows from temporal variables $X_i(t)$ at time slice t to temporal variables $X_j(t+1)$ at time slice $(t+1)$.

3.3.2 Feature Causal Graph

While the 2TVCG provides us an intuitive and compact tool to illustrate a first-order time-invariant dynamic system, an FCG is more compact by using the notation of features and different edge types to simultaneously encode instantaneous and cross-temporal causal relations. In an FCG $\mathcal{G}_f$ for an n-variate FoCP, there are n feature nodes and three types of edges:

- **Solid directed edge** ($\rightarrow$): A solid directed edge $X_i \rightarrow X_j$, means that there is a instantaneous causal relation from the variable $X_i(t)$ to the variable $X_j(t)$ for all $t = 1 : T$;
- **Dashed directed edge** ($\dashrightarrow$): A dashed directed edge $X_i \dashrightarrow X_j$ represents a time delayed cross-variable causal relation from variable $X_i(t)$ to variable $X_j(t+1)$;
- **Dotted Self-edge**: A dotted self-edge from feature X_i to itself represents a time delayed causal relation from $X_i(t)$ to $X_i(t+1)$.

3.3.3 A Transformation Procedure

We can easily derive a one-to-one mapping between the 2TVCG and the FCG representations of a FoCP. Denote the 2TVCG as $\mathcal{G}_v$ and its corresponding FCG as $\mathcal{G}_f$, we can conclude that: (1) the edge $X_i \rightarrow X_j$ exists in $\mathcal{G}_f$ if and only if $X_i(t) \rightarrow X_j(t)$ in $\mathcal{G}_v$ for every time slice; (2) the edge $X_i \dashrightarrow X_j$ exists in $\mathcal{G}_f$ if and only if $X_i(t) \rightarrow X_j(t+1)$ in $\mathcal{G}_v$; and (3) the dotted self-edge from X_i to itself exists in $\mathcal{G}_f$ if and only if $X_i(t) \rightarrow X_i(t+1)$ in $\mathcal{G}_v$. This mapping is formalized as the function `variableToFeatureGraph` defined in Algorithm 3.1

Procedure 3.1 `variableToFeatureGraph`($\mathcal{G}_v$)

For $i, j = 1 : n$ **do**
if $X_i^{t+1} \to X_j^{t+1}$ exists in $\mathcal{G}_v$, **then** draw $X_i \to X_j$ in $\mathcal{G}_f$
if $X_i^{t} \to X_j^{t+1}$ exists in $\mathcal{G}_v$, **then** draw $X_i \dashrightarrow X_j$ in $\mathcal{G}_f$
if $X_i^{t} \to X_i^{t+1}$ exists in $\mathcal{G}_v$, **then** draw $X_i \circlearrowleft X_i$ in $\mathcal{G}_f$

By this transformation procedure, the 2TVCG and its corresponding FCG of the SHO system is illustrated in Fig. 3.1.

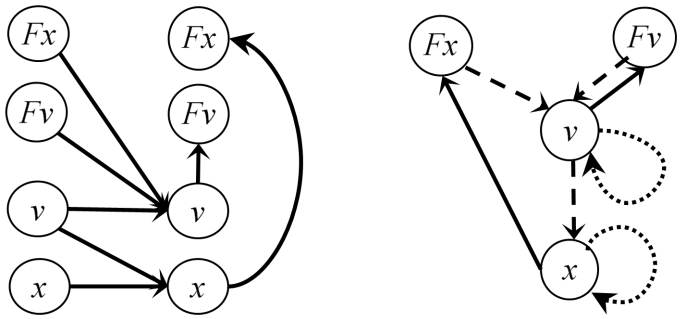


Figure 3.1: The 2TVCG (left) and FCG (right) representations for the SHO system.

3.4 Structure Learning for FoCPs

Based on the two graphical representations for FoCPs, the FoCP structure learning problem can be posed in the following way: given a set of multivariate time series over features $\boldsymbol{V}$, derive a 2TVCG over the set of variables $\boldsymbol{V}^t \cup \boldsymbol{V}^{t+\Delta t}$ from two adjacent time slices. Using the transformation procedure `variableToFeatureGraph`, we can then easily transform it to the corresponding FCG over $\boldsymbol{V}$.

3.4.1 Conditional Independence based Structure Learning

Our proposed FoCP learning algorithm belongs to the conditional independence constraint-based structure learning paradigm [184], which is mainly based on the following d-separation criteria:

Definition 3.2 (d-separation [130]). *In a DAG $\mathcal{G}$, a disjoint set of nodes Z is said to block a path p from X to Y if either (i) p contains at least one arrow-emitting vertex that is in Z, or (ii) p contains at least one collision vertex that is outside Z and has no descendant in Z.*

Specifically, under the causal sufficiency (Assumption 2.3), causal Markov (Assumption 2.4), and causal faithfulness (Assumption 2.5) conditions, the probabilistic conditional independence $X \perp\!\!\!\perp Y|Z$ is equivalent with graphical d-separation assertion $Dsep(X, Y|Z)$, i.e., $X \perp\!\!\!\perp Y|Z \Leftrightarrow Dsep(X, Y|Z)$. This indicates that we can discover causal independence relations in the underlying causal graph $\mathcal{G}$ by testing probabilistic conditional independences in the data distribution P. Typically, conditional independence tests between causal variables are performed using statistical or information theoretic measures. The algorithms start with a fully connected undirected graph. The conditional independence constraints are propagated throughout the graph, and edges inconsistent with them are removed. A sound strategy for performing conditional independence tests ultimately retains only the statistically equivalent graphs consistent with the constraints.

3.4.2 FoCP Learning

Following the above conditional independence constraint-based structure learning scheme, the procedure for learning the structure of a FoCP from a time series dataset is: firstly, initialize the graph $\mathcal{P}_0$ over $\boldsymbol{V} = \boldsymbol{V}^t \cup \boldsymbol{V}^{t+\Delta t}$ with directed edges

Algorithm 3.2 FoCP Learning Process

Input: A time series data set $\mathcal{D} = \{\mathrm{x}(t), t = 1 : T\}$, and an oracle for conditional independence between variables

Output: A FoCP feature causal graph $\mathcal{G}_f$

1: $\mathcal{P}_0 \leftarrow$ a graph over $\boldsymbol{V} = \boldsymbol{V}^t \cup \boldsymbol{V}^{t+\Delta t}$ with directed edges from every $X^t \in \boldsymbol{V}^t$ to every $Y^{t+\Delta t} \in \boldsymbol{V}^{t+\Delta t}$ and undirected edges between every pair of nodes from $\boldsymbol{V}^{t+\Delta t}$;
2: $\mathcal{P}_1 \leftarrow$ `instantaneousCausationSearch`$(\mathcal{P}_0, \mathcal{D})$;
3: $\mathcal{P}_2 \leftarrow$ `crossTemporalCausationPrune`$(\mathcal{P}_1, \mathcal{D})$;
4: $\mathcal{P}_3 \leftarrow$ `instantaneousCausationOrientation`$(\mathcal{P}_2, \mathcal{D})$;
5: $\mathcal{P} \leftarrow$ `variableToFeatureGraph`$(\mathcal{P}_3)$

from every antecedent variable $X^t \in \boldsymbol{V}^t$ to every latter variable $Y^{t+\Delta t} \in \boldsymbol{V}^{t+\Delta t}$ and undirected edges between every pair of variables in $\boldsymbol{V}^{t+\Delta t}$; secondly, search for adjacencies and orientations of instantaneous causal relations between variables in $\boldsymbol{V}^{t+\Delta t}$; thirdly, prune cross-temporal causal relations between variables from two adjacent time slices; and at last orientate more undirected instantaneous edges by constraints such that no cross-temporal and instantaneous causal relations can occur to a variable at the same time, and no cycles or new colliders can be introduced by orientations of undirected edges. The FoCP learning procedure is illustrated in Procedure 3.2 where we re-scale the time interval Δt of the target dynamic system to 1.

In the learning procedure, the procedure `instantaneousCausationSearch` inputs the initialized graph $\mathcal{P}_0$ (with fully connected contemporary and cross-temporal edges) and the time series data $\mathcal{D}$. Conditional independences between variables in $\boldsymbol{V}^{t+\Delta t}$ are tested to identify adjacencies among them. Since the contemporary causal structure over variables in $\boldsymbol{V}^{t+\Delta t}$ forms a DAG, the same orientation rules with the PC algorithm [184] are adopted to orientate as many instantaneous causal relations as possible.

To identify the correct set of cross-temporal causal relations, we only need to identify the adjacency relations between variables in $\boldsymbol{V}^t$ with variables in $\boldsymbol{V}^{t+\Delta t}$

Procedure 3.3 `instantaneousCausationSearch`$(\mathcal{P}_0, \mathcal{D})$

$\mathcal{P}_1 \leftarrow \mathcal{P}_0$
for $order = 0 : 2m - 2$ where m is the number of feature variables **do**
  for every ordered pair of variables $(X^{t+\Delta t}, Y^{t+\Delta t}) \in \boldsymbol{V}^{t+\Delta t}$ **do**
    $neighbors \leftarrow neighbors(X^{t+\Delta t}) \cup neighbors(Y^{t+\Delta t}) \setminus \{X^{t+\Delta t}, Y^{t+\Delta t}\}$
    if $|neighbors| \geq order$ **then**
      for $S \subseteq neighbors$ with $card(S) = order$ **do**
        if $X^{t+\Delta t} \perp\!\!\!\perp Y^{t+\Delta t} | S$ is entailed by the data $\mathcal{D}$ **then**
          Eliminate the edge $X^{t+\Delta t} — Y^{t+\Delta t}$
          $SepSet(X^{t+\Delta t}, Y^{t+\Delta t}) \leftarrow S$
        end if
      end for
    end if
  end for
end for
Let $\mathcal{G}$ be the sub-graph by deleting all variables $\boldsymbol{V}^t$ and edges collecting with them from $\mathcal{P}_1$. Use the same orientation rules with the PC algorithm (Algorithm 2.1) to orientate edges in $\mathcal{G}$;
Orient edges between variables $\boldsymbol{V}^{t+\Delta t}$ in $\mathcal{P}_1$ the same as edges in $\mathcal{G}$.
Output: $\mathcal{P}_1$

since their orientations are automatically entailed by their chronological order. The Property (5c) of FoCPs and conditional independence test are used in this stage. The resulting $\mathcal{P}_2$ theoretically contains all cross-temporal causal relations. The pruning procedure is formalized as `crossTemporalCausationPrune` in Procedure 3.4.

After the above identification procedures, the partially oriented causal graph $\mathcal{P}_2$ may contain some undirected instantaneous edges. We may further orientate some of them according to property (5) and by avoiding cycles and new colliders. Formally, this procedure is formalized as `instantaneousCausationOrientation` in Procedure 3.5.

In general, the proposed FoCP structure learning procedures can be modularized into the following stages: search preliminarily for instantaneous causal relations

Procedure 3.4 `crossTemporalCausationPrune`$(\mathcal{P}_1, \mathcal{D})$

$\mathcal{P}_2 \leftarrow \mathcal{P}_1$
Create an augmented dataset $\hat{\mathcal{D}} = \{(\mathrm{x}(t), \mathrm{x}(t+1), t = 1 : (T-1)\}$
for any $Y^{t+\Delta t} \in \boldsymbol{V}$ **do**
  if exist $X^{t+\Delta t} \in \boldsymbol{V} \cap X^{t+\Delta t} \to Y^{t+\Delta t}$ **then**
    Eliminate all cross-temporal arrows shooting from variables at time t to $Y^{t+\Delta t}$ from $\mathcal{P}_2$
  end if
end for
for $order = 0 : 2m-2$ where m is the number of feature variables **do**
  for every cross-temporal arrow $X^t \to Y^{t+\Delta t} \in \mathcal{P}_2$ **do**
    $neighbors \leftarrow neighbors(X^{t+\Delta t}) \cup neighbors(Y^{t+\Delta t}) \setminus \{X^{t+\Delta t}, Y^{t+\Delta t}\}$
    if $|neighbors| \geq order$ **then**
      for $S \subseteq neighbors$ with $card(S) = order$ **do**
        if $X^{t+\Delta t} \perp\!\!\!\perp Y^{t+\Delta t} | S$ is entailed by the data $\mathcal{D}$ **then**
          Eliminate the edge $X^{t+\Delta t} - Y^{t+\Delta t}$
          $SepSet(X^{t+\Delta t}, Y^{t+\Delta t}) \leftarrow S$
        end if
      end for
    end if
  end for
end for
Output: $\mathcal{P}_2$

$\rightarrow$(prune cross-temporal causal relations) $\rightarrow$(fine-tune instantaneous causal relations).

To obtain the final output FCG, the transformation procedure `variableToFeatureGraph` in Algorithm 3.1 is adopted to transform the causal graph $\mathcal{P}_3$ over $\boldsymbol{V}^t \cup \boldsymbol{V}^{t+\Delta t}$ to its equivalent FCG $\mathcal{G}_f$.

3.4.3 Computational Complexity

In the FoCP learning procedure, the computation is mainly used for testing conditional independences. For a m-variate FoCP, for searching instantaneous causal relations, the number of conditional independence tests is theoretically upper-bounded by $\binom{m}{2}\left[1 + \binom{2m-2}{1} + \cdots + \binom{2m-2}{2m-2}\right] = m(m-1) \cdot 2^{2m-3}$; for searching

Procedure 3.5 `instantaneousCausationOrientation`$(\mathcal{P}_2, \mathcal{D})$

1: $\mathcal{P}_3 \leftarrow \mathcal{P}_2$
2: Orientation undirected edges between variables $\boldsymbol{V}^{t+\Delta t}$ by the following rule while avoiding cycles and new colliders:
If $\exists X^t \to Y^{t+\Delta t}$ and $Y^{t+\Delta t}$—$Z^{t+\Delta t}$, **then** $Y^{t+\Delta t} \to Z^{t+\Delta t}$

Output: $\mathcal{P}_3$

cross-temporal causal relations, the number of conditional independence tests is upper-bounded by $\left[1 + \binom{2m-2}{1} + \cdots + \binom{2m-2}{2m-2}\right] = m^2 \cdot 2^{2m-2}$. The number of other rule checks is linear with the number of variables in the system. From this theoretical analysis, the computational complexity will become very high when the dimension of system variables increase. However, the above upper bounds will not practically be reached since the 2TVCG of a first order causal process is normally a sparse DAG given the Property (5) of FoCPs. As an initial work, we restrict ourselves to propose a reasonable model and correct learning algorithm for temporal causal modelling, but not focus on computational efficiency. However, for high dimensional dynamic systems, we recommend to manually set the maximum order of conditioning sets for conditional independence tests according to some prior knowledge for the sake of computational efficiency.

3.5 Experimental Analysis

To validate the FoCP model and the corresponding structure learning algorithm, we simulate data from the SHO system discussed in Section 3.1, and evaluate whether the proposed algorithm can discover the underlying causal structure from the observed data. In addition, we apply our method to a real-world climate dataset to demonstrate the viability of the FoCP model for modelling the causal structure of real dynamic system with possible instantaneous and cross-temporal causation. We implement the proposed algorithm using Matlab

and all experiments are run on a PC with four 2.30GHz i5 CPUs.

3.5.1 Baselines and Evaluation Metrics

To the best of our knowledge, there exists no temporal causal models [42] with such a 2-stage evolution semantic as the proposed FoCP model. As a result, no available algorithms are exactly suitable for comparison with our method. Even so, there are two similar works that are able to learn the structure of a dynamic system from time series when both instantaneous and cross-temporal causal relations may exist: Entner and Hoyer [43] extend the fast causal inference (FCI) algorithm [184] to time series data and propose the tsFCI algorithm; Hyvärinen et al. [71] take advantages of non-Gaussian noises to learn the structure of the SVAR model and their algorithm is named VARLiNGAM. Recently, Meek [113] proposed a similar *causal process* concept and the δ^*-separation for causal discovery from discrete-time continuous-valued time series. Unfortunately, the author only proposed a conceptual framework but did not provide any working algorithm. As a result, for experimental comparison, we choose the VARLiNGAM and tsFCI as our baseline methods.

Note that the two baseline algorithms are based on different assumptions. Specifically, VARLiNGAM makes the same CSC assumption as our FoCP learning algorithm, i.e., all system variables are observed; while tsFCI allows latent variables. As a consequence, the causal graph outputted by VARLiNGAM will only contain directed edges while tsFCI outputs a partial ancestral graph (PAG) [184] that may contain partially directed, undirected and bi-directed edges. For the sake of comparison, in this work, all cross-temporal edges in the learned PAG are replaced by directed edges from antecedent variables to latter variables, all bi-directed edges are replaced by undirected edges, and all partially directed edges are replaced by directed edges.

Following similar evaluation metrics in [26] and [185], we define the following edge error rates to evaluate the performance on edge identification of the aforementioned algorithms. The first evaluation metric is the *edge omission error rate* defined as:

$$E_{om} = \frac{\#\text{ edge omission errors}}{\#\text{ edges in the true 2TVCG}}$$

The second metric is the *edge commission error rate* defined as:

$$E_{com} = \frac{\#\text{ edge commission errors}}{\text{Maximum}\#\text{ of possible edge commission errors}}$$

In the above formulas, an edge omission error occurs when two variables are adjacent in the true 2TVCG but not in the learnt causal graph. An edge commission error occurs when two variables are adjacent in the learnt causal graph but not in the true 2TVCG. For a m-variate dynamic system with lag one cross-temporal causation, the maximum number of possible edge commission errors is equal to $\left(m^2 + \frac{m(m-1)}{2} - \#\text{ edges in the true } 2TVCG\right)$.

Similarly, we define the following arrowhead error rates to evaluate algorithm performances regarding edge orientations among contemporary variables. The arrowhead omission error rate is defined as

$$A_{om} = \frac{\#\text{ arrowhead omission errors}}{\#\text{ contemporary arrowheads in the learned graph}+1}$$

and the arrowhead commission error rate is defined as

$$A_{com} = \frac{\#\text{ arrowhead commission errors}}{\#\text{ contemporary arrowheads in the learned graph}+1}$$

We only consider instantaneous arrowheads because the orientations of cross-temporal edges are instinctive from antecedent variables to latter variables according to the arrow of time.

3.5.2 Simulated Data

We simulate the data from the toy SHO system with $\boldsymbol{B}^{temp} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0.1 & 0.1 & 1 & 0 \\ 0 & 0 & 0.1 & 1 \end{pmatrix}$, $\boldsymbol{B}^{inst} = \begin{pmatrix} 0 & 0 & 0 & -0.3 \\ 0 & 0 & -0.2 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$ and time interval $\Delta t = 0.1$. The initial state of the system is set to $\mathrm{x}(0) = (0, -1, 5, 0)$. Thus, the augmented instantaneous transition matrix $\widehat{\boldsymbol{B}}^{inst} = \begin{pmatrix} 0 & 0 & 0 & -0.3 \\ 0 & 0 & -0.2 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$ and the composited causal transition matrix for the system $\boldsymbol{B} = \widehat{\boldsymbol{B}}^{\mathrm{inst}}\boldsymbol{B}^{\mathrm{temp}} = \begin{pmatrix} 0 & 0 & -0.03 & -0.3 \\ -0.02 & -0.02 & -0.2 & 0 \\ 0.1 & 0.1 & 1 & 0 \\ 0 & 0 & 0.1 & 1 \end{pmatrix}$. Since the VARLiNGAM algorithm takes advantages of the non-Gaussianity of noises to learn causal structures, we simulate data with both additive Gaussian noises and non-Gaussian errors (specifically super-Gaussian noises are simulated). Furthermore, four time series data sets with sample size of 500, 1000, 1500 and 2000 are synthesised respectively in each noise context. Experimental results of the three algorithms in two noise contexts are illustrated in Fig. 3.2 and Fig. 3.3.

Generally, our FoCP learning algorithm has an overall lower error rates than the other two algorithms, except for the edge omission error rate compared with VARLiNGAM. Specifically, the VARLiNGAM algorithm tends to learn many suspicious cross-temporal relations from the antecedent time slice to the latter time slice (almost fully connected), and thus leading to zero edge omission errors while very high edge commission error rates in all experiments. This may be because VARLiNGAM are based on the VAR model, which is actually a multivariate regression model.

Basically both the tsFCI and our FoCP learning algorithm belong to the constraint-based structure learning paradigm. They both rely on some graphical separation criteria and conditional independence test, so they are very likely to output common edges. However, in our FoCP learning algorithm, the well designed cross-temporal causation identification and the additional instantaneous causation fine-tuning process theoretically guarantee us to make less edge com-

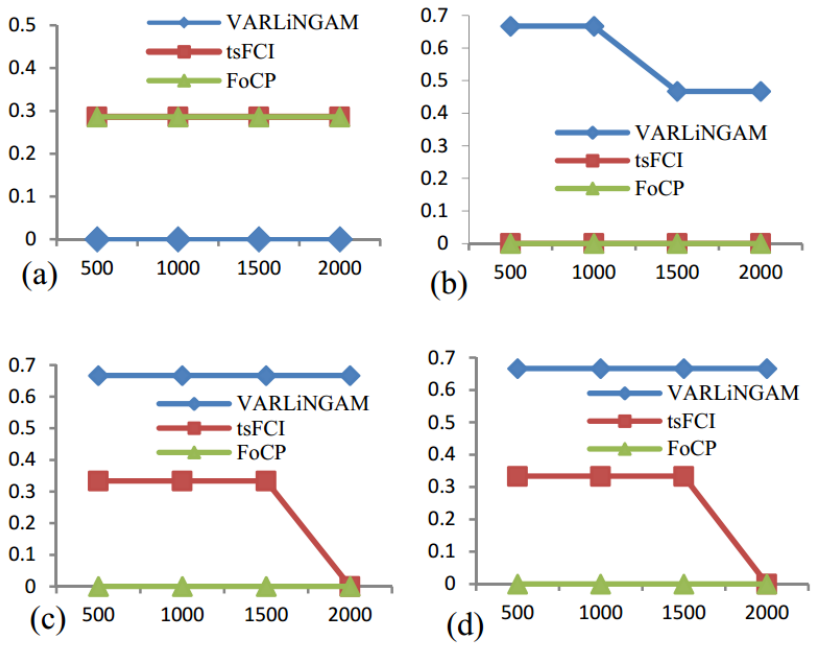


Figure 3.2: Comparisons of (a) edge omission error rates, (b) edge commission error rates, (c) arrowhead omission error rates, and (d) arrowhead commission error rates in the Gaussian noises context.

mission errors and arrowhead emission errors. This primary judgement is partially validated by Fig.3.2(c) and Fig.3.3(c).

Another finding from the results is that though not very significant, all algorithms tend to make less errors as the sample size gets larger. In addition, by comparing the results in Fig.3.2 with that in Fig.3.3, we find that Non-Gaussianity does benefit VARLiNGAM to identify the correct orientations of contemporary edges as theoretically expected.

3.5.3 Application to Climate Data

We also apply the FoCP model to the Global Summary of the Day (GSOD) climate data[1]. The collected data contains a number of 18159 daily records

[1]https://data.noaa.gov/dataset/global-surface-summary-of-the-day-gsod

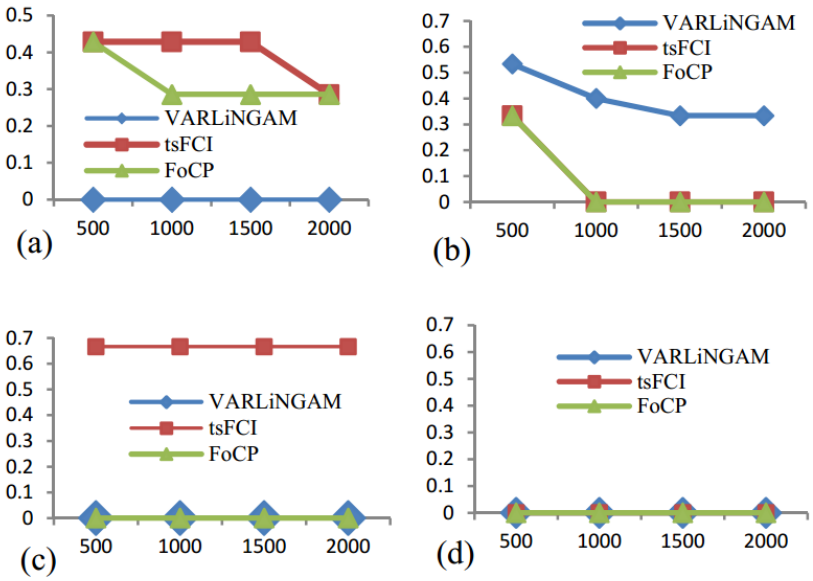


Figure 3.3: Comparisons of (a) edge omission error rates, (b) edge commission error rates, (c) arrowhead omission error rates, and (d) arrowhead commission error rates in the non-Gaussian noises context

of the temperature (*Temp*), sea level pressure (*SLP*), and average wind speed (*AWS*). The kernel based independence criteria, Hilbert Schmidt Independence Criteria (HSIC), is adopted for conditional independence test. The causal graphs learned by the VARLiNGAM algorithm, the tsFCI algorithm and our FoCP learning algorithm are illustrated in Fig.3.4. The original output PAG of the tsFCI algorithm is converted by the same transformational rules discussed in Section 3.5.1.

Without access to the ground truth, it is hard to say which learned causal graph is better than the others. However, we may still get some understanding of the underlying causal relationships among climate variables, and further validate our proposed method by the overlapping causal relations learned by our method and the other two methods.

In Fig.3.4, we may conclude some common causal relations. Specifically, all

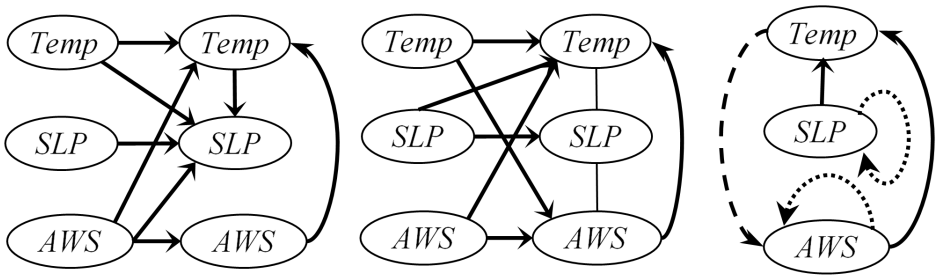


Figure 3.4: The learned causal graphs for the GSOD climate data learned by the VARLiNGAM algorithm (left), the tsFCI algorithm (middle), and our FoCP learning algorithm (right)

three causal graphs indicate that the *SLP* is directly influenced by the previous day's *SLP*; the *AWS* is also directly influenced by the previous day's *AWS*, while the daily temperature is directly influenced by the *AWS* in the same day. Both the FoCP and the tsFCI learning algorithm identify the feedback influence from temperature to the next day's *AWS*, while VARLiNGAM fails to learn this influence. Meanwhile, both causal graphs by VARLiNGAM and the FoCP learning algorithm indicate that no direct instantaneous causation exists between the *SLP* and the *AWS*.

As we have discussed in Section 3.3, the FCG representation of a FoCP encodes the instantaneous causation and cross-temporal causation (including self-edge feedbacks) by different kinds of edges. The 2-stage state evolution semantic guarantees the learned FCG to be generally more compact and simpler than its counterparts identified by the other two algorithms. Thus, the analysis on climate data indicates that the proposed FoCP model as well as the structure learning algorithm acts as a potential tool for temporal causal modelling to discover simple and meaningful causal structures.

3.6 Summary

In this chapter, we consider the issue of temporal causal modelling when both instantaneous and cross-temporal causal relations may exist. Although the concept of *causal process* has been discussed in previous work, there has not as yet a unified graphical representation or structure learning algorithms for it. As an initial work, we try to contribute this line of research by first drawing some insights from physical systems, and refine them into the 2-stage state evolution semantic further. Based on this new interpretation and its entailed properties, practical graphical representations and a structure learning algorithm are proposed to identify the underlying causal structure of a dynamic system. Experiments on simulated and real data validate the proposed method.

Chapter 4

A Pareto-smoothing Method for Causal Inference using Generalized Pareto Distribution

In this chapter, we consider the problem of causal inference via weighting. We reframe causal inference using the IPW estimator to the importance sampling framework and introduce a new smoothing method for importance weight stabilization using the smoothing property of the generalized Pareto distribution (GPD) from the extreme value statistics [28]. Based on the new interpretation of the IPW estimator and the proposed Pareto-smoothing method, we propose two IPW estimators for treatment effect estimation.

Our contributions are as follows: (1) We introduce the classic IPW causal estimator from the perspective of importance weighted estimation of expectations using data from a different proposal distribution. To the best of our knowledge, we are the first to formalize such an interpretation of the IPW estimator, which renders the high variability problem of importance weight-based estimators straightforward and easy to understand. (2) Building upon the above im-

portance sampling interpretation of the IPW estimator, we analyse the high variability problem of the IPW estimator with estimated propensity scores and conclude two existing stabilization methods for importance weight stabilization, i.e., weight truncation and self-normalization. (3) We propose a new Pareto-smoothing method for importance weight stabilization using GPDs and two Pareto-smoothed causal estimators based on the proposed method. We also discuss the selection of related parameters in the proposed method. Comprehensive experiments were conducted using both simulated and real data to demonstrate the practical validity of the proposed method.

The remainder of this chapter is organized as follows. In Section 4.1, we introduce notations, formalize the causal inference problem, and discuss the assumptions for identification. In Section 4.2, we reframe the classic IPW estimator for causal inference in the importance sampling framework, which leads to a straightforward understanding of its high variability problem in finite-sample settings. Within this framework, we briefly review two conventional methods for stabilizing the IPW estimator. In Section 4.3, we introduce the details of our new Pareto-smoothing method and the two proposed Pareto-smoothed causal estimators. Experiments on simulated data and an application on a real-world health dataset are conducted in Section 4.4 and Section 4.5. Section 4.6 summarizes this chapter.

4.1 Problem Setup

Consider a population of n individuals indexed by $i = 1, 2, \ldots, n$. Every individual i is characterized by a d-dimensional vector of features (also called pre-treatment covariates or attributes), $X_i \in \mathbb{R}^d$. Elements of these covariates might include age, gender, race, education, etc. For convenience, we use X_i and

i interchangeably to represent the ith individual, and X to represent a general individual from the population. Each individual makes a decision to choose an action or is assigned to a treatment T; for example, the treatment T could be whether to take a particular medicine or whether to receive a certain training program. We consider binary treatments and denote the treatment for an individual i as T_i, where $T_i = 0$ indicates that individual i received the control treatment and $T_i = 1$ indicates that individual i received the active treatment. Let Y be the outcome variable of interest. For any individual X, following Rubin's potential outcome framework [74], there is a pair of potential outcomes $Y_X(0)$ and $Y_X(1)$, denoting the outcome value of X if he or she had been in the control group or the treatment group respectively. By the principle of consistency, the observed outcome of individual X_i, denoted as $Y_{X_i}^{obs}$ or simply Y_i, is the potential outcome corresponding to the received treatment, i.e., $Y_{X_i}^{obs} = Y_i = Y_{X_i}(T_i) = Y_i(T_i)$

With these notations, the individual treatment effect for the ith individual is defined as the difference of the two potential outcomes $\tau_i = Y_i(1) - Y_i(0)$. The conditional average treatment effect is defined as $\tau(\mathrm{x}) = \mathbb{E}\left[\tau_i | X_i = \mathrm{x}\right]$ and the ATE of treatment T on the outcome Y is its expectation for this population,

$$\tau_{ATE} = \mathbb{E}\left[\tau(X)\right] = \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)] \tag{4.1}$$

Rather than the ATE for the whole population, sometimes we may only be concerned about the ATE for the treated individuals, i.e., the ATT defined as $\tau_{ATT} = \mathbb{E}\left[\tau(X) \mid T_i = 1\right]$. While we can analogously define the average treatment effect on the control (ATC), it is seldom of interest in practical applications. We formulate the ATE estimation problem for concreteness. The estimation of ATT is straightforward and is introduced in the Appendix.

Given the observational data $\mathcal{D} = \{(X_i, T_i, Y_i) : i = 1, 2, \ldots, n\}$, where n is the number of observations. Referring to the individuals with $T_i = 1$ as treated

individuals and the individuals $T_i = 0$ as control, we also denote the number of treated as $n_1 = \sum_{i=1}^{n} T_i$ and the number of controls as $n_0 = \sum_{i=1}^{n} (1 - T_i)$. For each $i = 1, 2, \ldots, n$, $Y_i(T_i) = Y_i$ is the observed factual outcome and $Y_i(1-T_i)$ is the counterfactual outcome, i.e., the outcome for individual i had she received the treatment $(1-T_i)$ instead of T_i. If we have access to both potential outcomes, ATE can be estimated by

$$\begin{aligned} \tau_{ATE} &= \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)] \\ &= \mathbb{E}_X \left[\mathbb{E}[Y_X(1)] - \mathbb{E}[Y_X(0)]\right] \\ &= \frac{1}{n} \sum_{i=1}^{n} Y_i(1) - Y_i(0) \end{aligned} \tag{4.2}$$

The ATE measures the average causal difference of a population if *all* individuals are treated versus *all* are untreated, which is generally different from the conditional difference between the outcomes of the treated group and the control group in the observational data. As a baseline, we denote the empirical conditional difference calculated in Eq.(4.3) as a naive ATE estimator,

$$\hat{\tau}_{ATE}^{\text{Naive}} = \frac{1}{n_1} \sum_{i=1}^{n} T_i Y_i - \frac{1}{n_0} \sum_{i=1}^{n} (1 - T_i) Y_i \tag{4.3}$$

Estimating the ATE from observational data is generally impossible because of *the fundamental problem of causal inference* [74]: for each individual, only one of the potential outcomes is observed. As a result, causal inference from observational data is by nature a missing data problem [100]. To ensure the identifiability [131], we assume *unconfoundedness* (or *conditional exchangeability*) defined in Assumption 4.1.

Assumption 4.1 (Unconfoundedness, or conditional exchangeability)**.** *Conditional on the observed pre-treatment covariates X, the potential outcomes $Y_X(0)$,$Y_X(1)$ are independent of the treatment T, i.e.,$\{Y_X(0), Y_X(1)\} \perp\!\!\!\perp T|X$*

This is to say that all confounders that affect both the treatment and outcome are observed. Under this assumption, the back-door adjustment criterion [131] suggests that we can identify the expected potential outcome $\mathbb{E}[Y_X(t)]$ by the conditional mean outcome via

$$\mathbb{E}[Y_X(t)] = \mathbb{E}[Y|X, T = t]$$

As a result, we can fit two conditional mean outcome models $\mathbb{E}[Y|X_i, T_i = 0]$ and $\mathbb{E}[Y|X_i, T_i = 1]$ from the observational data $\mathcal{D}$, and estimate the ATE in Eq.(4.2) by calculating the average of the covariate-stratified differences weighted by the probabilities of each stratum. Although feasible in principle, adjusting for all observed covariates to eliminate confounding bias may not be possible, especially when the covariates are continuous. So we need to find a lower-dimensional proxy for them that will suffice for removing the bias associated with imbalance in the pre-treatment covariates.

The propensity score in Definition 4.1 is such a low-dimensional proxy and plays a key role in many existing propensity score-based causal estimators.

Definition 4.1 (Propensity score [74]). *The* ***propensity score****, $e(X)$, of an individual X is its conditional probability to be assigned to the treatment group, i.e., $e(X) = p(T = 1 \mid X)$.*

For any individual, the treatment assignment T is independent of the pre-treatment covariates X conditional on the true propensity score $e(X)$. Moreover, the unconfoundedness assumption implies that $\{Y_X(0), Y_X(1)\} \perp\!\!\!\perp T|e(X)$. In practice, to guarantee enough randomness in the data-generating process so that unobserved counterfactuals can be estimated from the observed data, we also make the *Positivity* assumption.

Assumption 4.2 (Positivity, or overlap). $0 < e(X_i) < 1$ *for any* $i = 1, \ldots, n$

This assumption means that the treatment assignment is not deterministic. In words from the literature of observational studies, the observations are generated by a probabilistic assignment mechanism [74].

4.2 Preliminaries

As we can see from Eq.(4.1), a key task for treatment effect estimation is to estimate the expected potential outcomes of the population, $\mathbb{E}[Y(0)]$ and $\mathbb{E}[Y(1)]$. In this section, we introduce importance weighted expectation estimators from the importance sampling literature [128]. Within this importance weighting framework, we further introduce the IPW estimator, the truncated and the self-normalized estimators for causal inference.

4.2.1 Estimating Expected Potential Outcomes

To explain the deduction, let us first consider treatment effect estimation via RCTs and imagine there is a randomized control experiment in which the treatment propensity $p(T_i = 1|X_i)$ is constant for any $i = 1, 2, \ldots, n$. Using the Bayes rule, we can easily derive that the covariate distribution for the treated group, $p_X^{t=1} := p(X|T = 1)$, and the control group, $p_X^{t=0} := p(X|T = 1)$, all equals the population distribution, $p_X := p(X)$. Thus, we can identify both expected potential outcomes via

$$\mathbb{E}[Y(1)] = \mathbb{E}_X \mathbb{E}\left[Y|X, T = 1\right] = \mathbb{E}_{p_X^{t=1}} \mathbb{E}\left[Y|X, T = 1\right] \tag{4.4}$$

$$\mathbb{E}[Y(0)] = \mathbb{E}_X \mathbb{E}\left[Y|X, T = 0\right] = \mathbb{E}_{p_X^{t=0}} \mathbb{E}\left[Y|X, T = 0\right] \tag{4.5}$$

As a result, ATE can be directly identified from the experimental data via

$$\begin{aligned}\tau_{ATE} &= \mathbb{E}\left[Y(1)\right] - \mathbb{E}\left[Y(0)\right] \\ &= \mathbb{E}_{p_X^{t=1}}\mathbb{E}\left[Y|X, T=1\right] - \mathbb{E}_{p_X^{t=0}}\mathbb{E}\left[Y|X, T=0\right]\end{aligned}$$

However, in observational studies, the treatment assignment is generally not random, i.e., $p_X^{t=1} \neq p_X$ and $p_X^{t=0} \neq p_X$. Consequently, we cannot calculate the population expectations $\mathbb{E}[Y(1)]$ and $\mathbb{E}[Y(0)]$ from the observed data directly via Eq.(4.4) and Eq.(4.5). Importance sampling is one of the most generally applicable procedures for computing expectations when it is not possible to sample directly from the target distribution. Denote the target distribution as $\pi(\mathrm{x})$ and a proposal distribution $q(x)$. The expectation of any function $h(x)$ with respect to the target distribution $\pi(\mathrm{x})$ can be consistently estimated by the following importance weighting formula [128]

$$\mathbb{E}_\pi[h(\mathrm{x})] = \int h(\mathrm{x})\pi(\mathrm{x})d\mathrm{x} = \int h(\mathrm{x})q(\mathrm{x})\frac{\pi(\mathrm{x})}{q(\mathrm{x})}d\mathrm{x}$$

Denote $w(\mathrm{x}) = \pi(\mathrm{x})/q(\mathrm{x})$ and call $w(\mathrm{x}^\mathrm{s}) = \pi\left(\mathrm{x}^\mathrm{s}\right)/q\left(\mathrm{x}^\mathrm{s}\right)$ the *importance weight* for the sth sample. If we have S draws $\left\{\mathrm{x}^1, \mathrm{x}^2, \ldots, \mathrm{x}^S\right\}$ from $q(x)$, then we can approximate $\mathbb{E}_\pi[h(\mathrm{x})]$ using Monte Carlo by

$$\begin{aligned}\mathbb{E}_\pi[h(\mathrm{x})] &= \int q(\mathrm{x})h(\mathrm{x})\frac{\pi(\mathrm{x})}{q(\mathrm{x})}d\mathrm{x} \\ &= \mathbb{E}_q[w(\mathrm{x})h(\mathrm{x})] \\ &= \frac{1}{S}\sum_{s=1}^{S} w(\mathrm{x}^\mathrm{s})h\left(\mathrm{x}^\mathrm{s}\right)\end{aligned} \tag{4.6}$$

In our causal inference setting, the observational data $\mathcal{D} = \{(X_i, T_i, Y_i) : i = 1, 2, \ldots, n\}$ comes from the propensity model $p\left(T_i = 1|X_i\right) = e\left(X_i\right), p\left(T_i = 0|X_i\right) = 1 - e\left(X_i\right)$ and the outcome model $Y_i = Y_{X_i}(T_i)$. Knowing

that $p(T_i = 1) = \frac{n_1}{n}$, using the above importance weighting formula Eq.(4.6) and the Bayes rule, we can consistently estimate the expected treated outcome for the population by

$$\begin{aligned}\mathbb{E}[Y(1)] &= \frac{1}{n_1}\sum_{i:T_i=1}\frac{p(X_i)}{p(X_i|T_i=1)}Y_i \\ &= \frac{1}{n_1}\sum_{i:T_i=1}\frac{p\left(T_i=1\right)}{p\left(T_i=1|X_i\right)}Y_i \\ &= \frac{1}{n}\sum_{i=1}^{n}\frac{\mathbb{1}\left[T_i=1\right]}{p\left(T_i=1|X_i\right)}Y_i\end{aligned} \tag{4.7}$$

where $\mathbb{1}\left[T_i = t\right]$ is the indicator function. Similarly, the expected control outcome for the population can be estimated by

$$\mathbb{E}[Y(0)] = \frac{1}{n}\sum_{i=1}^{n}\frac{\mathbb{1}\left(T_i=0\right)}{p\left(T_i=0|X_i\right)}Y_i \tag{4.8}$$

4.2.2 IPW Estimator

Substituting Eq.(4.7) and Eq.(4.8) into the ATE definition in Eq.(4.2), we get the following ATE estimator

$$\begin{aligned}\hat{\tau}_{ATE} &= \mathbb{E}\left[Y(1)\right] - \mathbb{E}\left[Y(0)\right] \\ &= \frac{1}{n}\sum_{i=1}^{n}\frac{\mathbb{1}\left[T_i=1\right]}{p\left(T_i=1|X_i\right)}Y_i - \frac{1}{n}\sum_{i=1}^{n}\frac{\mathbb{1}\left[T_i=0\right]}{p\left(T_i=0|X_i\right)}Y_i\end{aligned} \tag{4.9}$$

Define the importance weight, W_i, for individual i in a general form as the reciprocal of its probability of receiving the observed treatment T_i. Formally,

$$W_i := \frac{1}{p(T_i|X_i)} = \frac{\mathbb{1}(T_i=1)}{e(X_i)} + \frac{\mathbb{1}(T_i=0)}{1-e(X_i)} \tag{4.10}$$

Then we can rewrite the estimator in Eq.(4.9) as the following importance

weighting estimator

$$\begin{aligned}\hat{\tau}_{ATE}^{\text{IPW}} &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}[T_i=1]W_iY_i - \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}[T_i=0]W_iY_i \\ &= \frac{1}{n}\sum_{i:T_i=1}W_iY_i - \frac{1}{n}\sum_{i:T_i=0}W_iY_i\end{aligned} \tag{4.11}$$

This is called the IPW estimator [63, 68] and is one of the most commonly used unbiased estimators for treatment effect estimation. In observational studies, the propensity score $e(X_i)$ for each individual is not available and need to be estimated from data by some statistical procedure (for example, Logistic regression). By using the estimated propensity scores $\hat{e}(X_i)$ directly, the finite-sample performance of the IPW estimator $\hat{\tau}_{ATE}^{\text{IPW}}$ could be poor. The reason is that the estimated propensity scores $\hat{e}(X_i)$ occur in the denominator in the definition of importance weight in Eq.(4.10), and small inaccuracies in $\hat{e}(X_i)$ can induce very high inaccuracies in the estimated ATE, especially when $\hat{e}(X_i)$ is close to zero or one. In this case, the importance weights W_i will be of high variability or even have unbounded variance, thus simple substitute estimators based on them may be unstable and misleading.

To remedy the high variability of the estimated importance weights, we introduce two existing methods for importance weighting estimator stabilization adopted from the importance sampling literature [61]: weight truncation and weight self-normalization.

4.2.3 Truncated IPW Estimator

Weight truncation is a common approach for variance reduction in the importance sampling literature [76, 128]. For the purpose of causal effect estimation, the truncated IPW estimator is defined as

$$\begin{aligned}\hat{\tau}_{ATE}^{\text{Trunc}} &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}[T_i=1]W_i^{\text{Trunc}}Y_i - \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}[T_i=0]W_i^{\text{Trunc}}Y_i \\ &= \frac{1}{n}\sum_{i:T_i=1}W_i^{\text{Trunc}}Y_i - \frac{1}{n}\sum_{i:T_i=0}W_i^{\text{Trunc}}Y_i\end{aligned} \tag{4.12}$$

where the truncated importance weight W_i^{Trunc} is derived by truncating the vanilla importance weight W_i by:

$$W_i^{\text{Trunc}} := \begin{cases} a, & \text{if } W_i < a \\ W_i, & \text{if } a \leq W_i \leq b \\ b, & \text{if } W_i > b \end{cases} \tag{4.13}$$

A consequence of weight truncation is the introduction of bias in the truncated importance weights, which in turn causes bias in the importance weight-based estimates. Moreover, the truncation thresholds are usually unknown and choosing them relies on experience or intuition. Crump et al. [31] proposed to keep individuals with estimated propensity score within the range $[0.1, 0.9]$. As a baseline, we follow this heuristic to truncate the importance weights in Eq.(4.13) by $a = \frac{10}{9}$ and $b = 10$. We denote the truncated IPW estimator with this truncation thresholds as *TruncCrump*. Recently, Yang and Ding [211] proposed to use a smooth weight function to approximate the existing sample truncation. Their method seems theoretically promising. However it requires us to tune the smooth weight function hyper-parameter and no open source code is available for comparison. In addition, Ju et al. [82] proposed a data-adaptive truncation algorithm which adaptively selects the optimal truncation threshold for the estimated propensity scores, but it is especially designed for target maximum likelihood estimators [171, 199]. In this work, we compare our proposed estima-

tors with the *TruncCrump* estimator and two other truncated IPW estimators in the experiment sections.

4.2.4 Self-normalized IPW Estimator

We can also apply the control variates technique [128] for variance reduction and divide the importance weights by their empirical mean in each treatment group. Denoting the average importance weight for the treated group as $\overline{W}_t := \frac{1}{n}\sum_{i:T_i=1} W_i$ and the average importance weight for the control group as $\overline{W}_c := \frac{1}{n}\sum_{i:T_i=0} W_i$, the self-normalized importance weight for each individual is then defined as

$$W_i^{\text{Norm}} := \mathbb{1}[T_i = 0]\frac{W_i}{\overline{W}_c} + \mathbb{1}[T_i = 1]\frac{W_i}{\overline{W}_t}$$

By replacing the importance weights W_i in Eq.(4.11) by the self-normalized importance weights W_i^{Norm}, we get the following self-normalized IPW estimator

$$\begin{aligned}\hat{\tau}_{ATE}^{\text{Norm}} &= \frac{1}{n}\sum_{i=1}^{n}\frac{\mathbb{1}[T_i = 1]W_i}{\overline{W}_t}Y_i - \sum_{i=1}^{n}\frac{\mathbb{1}[T_i = 0]W_i}{\overline{W}_c}Y_i \\ &= \frac{1}{n}\sum_{i:T_i=1} W_i^{\text{Norm}}Y_i - \frac{1}{n}\sum_{i:T_i=0} W_i^{\text{Norm}}Y_i\end{aligned} \tag{4.14}$$

In general, the self-normalized IPW estimator $\hat{\tau}_{ATE}^{\text{Norm}}$ has lower variance than the original IPW estimator $\hat{\tau}_{ATE}^{\text{IPW}}$. In the experimental study section, we evaluate the performance of the stabilized IPW estimator, which combines our proposed Pareto-smoothing method with the self-normalization method.

4.3 Methodology

In the previous section, we reframe importance weight-based causal estimators from the perspective of importance sampling estimation of expectations. A com-

mon phenomenon in the importance sampling literature is that the importance weighting estimator for expectations are subject to the instability problem in settings with finite samples. To cope with this problem so as to establish stable importance weight-based causal estimators, we also introduced weight truncation and self-normalization, leading to the truncated IPW estimator and the self-normalized estimator. As a complementary of these estimator stabilization methods, in this section, we introduce our Pareto-smoothing method for importance weight stabilization. Based on this method, we further propose two ATE estimators: the Pareto-smoothed IPW estimator and the Pareto-smoothed self-normalized IPW estimator.

Our proposed method builds upon results from the extreme value theory [28]. In extreme value statistics, if an unknown distribution function $F(w)$ lies in the "domain of attraction" of an extreme distribution function, then $F(w)$ has a generalized Pareto upper tail. As a result, we can approximate its upper tail by a GPD if the location μ of the tail can increase as the sample size increases.

Within the framework of importance weight-based causal inference, we estimate the importance weight for each individual and obtain the importance weights $\{W_1, W_2, \ldots, W_n\}$. To remedy the influence of extreme weights, rather than truncating the importance weights in a brute-force way, we fit a GPD over the upper tails of the estimated importance weights and smooth them by the fitted GPD. By this smoothing method, we try to stabilize the importance weights while retaining the information of their relative order.

4.3.1 GPD Fitting

The GPD probability density function for a scalar random variable W with parameter $\theta = (\mu, \sigma, \kappa)$ is defined as

$$f(w) = \begin{cases} 1/\sigma \left(1 + \frac{\kappa(w-\mu)}{\sigma}\right)^{-1/\kappa - 1}, & \kappa \neq 0 \\ 1/\sigma e^{-\frac{w-\mu}{\sigma}}, & \kappa = 0 \end{cases} \tag{4.15}$$

where μ is the location parameter, $\sigma > 0$ is the scale, and κ is the shape of the distribution. In addition, we will also use the cumulative density function $F(w; \mu, \sigma, \kappa)$ defined in E.q(4.16) to calculate its expected order statistics as the replacement of large importance weights

$$F(w) = \begin{cases} 1 - \left(1 + \frac{\kappa(w-\mu)}{\sigma}\right)^{-1/\kappa}, & \kappa \neq 0 \\ 1 - e^{-\frac{w-\mu}{\sigma}}, & \kappa = 0 \end{cases} \tag{4.16}$$

In this section, we describe the procedure for fitting a GPD over the upper tail of the estimated importance weights $\{W_1, W_2, \ldots, W_n\}$. This includes heuristics for choosing the location parameter μ, estimating the positive scale parameter σ and the shape parameter κ. In general, we only consider fitting the parameters with $\kappa \neq 0$.

Selecting μ

The location parameter μ of a GPD $F(w; \mu, \sigma, \kappa)$ determines the cut-point of the ordered importance weights and thus how many importance weights will be smoothed. In this section, we refer to existing literature and propose to choose it heuristically.

In order to obtain asymptotic consistency, Pickands [140] proposed that the lower bound parameter μ should be chosen so that the sample size M of to-be-

smoothed weights in the tail increases to infinity while M/n goes to zero. In addition, by extensive empirical comparisons, Vehtari et al. [202] recommended to choose μ so that the sample size M satisfies

$$M = \min(\lfloor 0.2n \rfloor, \lfloor 3\sqrt{n} \rfloor) \tag{4.17}$$

This is a reasonable heuristic for deciding the location parameter and the empirical study in [203] shows that the majority of results are not sensitive to the choice of M. In this work, following this routine, we first sort $W_1, W_2, \ldots, W_n$ in an ascending order and obtain the order statistics of these importance weights, $W_{[1]}, W_{[2]}, \ldots, W_{[n]}$ where $W_{[1]} \leq W_{[2]} \leq \cdots \leq W_{[n]}$. Then the location parameter μ is chosen by

$$\hat{\mu} = W_{[n-M]} \tag{4.18}$$

where M is derived according to Eq.(4.17).

Estimating k and σ

Having selected the location μ, we now estimate the scale σ and shape k of the GPD over the upper tail $\{W_{[n-M+1]}, W_{[n-M+2]}, \ldots, W_{[n]}\}$. In statistics, for a GPD $F(w; \mu, \sigma, \kappa)$ over $\{W_{[n-M+1]}, W_{[n-M+2]}, \ldots, W_{[n]}\}$, define

$$O_m = W_{[n-M+m]} - \mu, \quad m = 1, 2, \ldots, M$$

then $\{O_1, O_2, \ldots, O_M\}$ follow the GPD $F(w; 0, \sigma, \kappa)$.

There are many methods to estimate k and σ using the M residuals $\{O_1, O_2, \ldots, O_M\}$ in the literature [106]. Among these methods, Zhang and Stephens [216] reparametrized the GPD $F(w; 0, \sigma, \kappa)$ by two parameters (ρ, κ), where $\rho = \kappa/\sigma$. With this reparameterization, we can easily derive the log-

likelihood for the samples $\{O_1, O_2, \ldots, O_M\}$ as

$$\ell(\rho, \kappa) = M \log \frac{\rho}{\kappa} - \frac{\kappa + 1}{\kappa} \sum_{i=1}^{M} \log(1 + \rho O_i) \tag{4.19}$$

Set the gradient over κ to 0,

$$\nabla_\kappa \ell = -\frac{M}{\kappa} + \frac{\sum_{i=1}^{M} \log(1 + \rho O_i)}{\kappa^2} = 0$$

We get

$$\kappa = \frac{1}{M} \sum_{i=1}^{M} \log(1 + \rho O_i) \tag{4.20}$$

Substituting Eq.(4.20) into Eq.(4.19), we get the following *profile log-likelihood* function for ρ

$$\ell(\rho) = M \log \frac{\rho}{\kappa} - M(\kappa + 1) \tag{4.21}$$

where κ is a function of ρ as indicated in Eq.(4.20). Thus, the key is to get an estimate of ρ. Zhang and Stephens [216] proposed to estimate it using the Bayes-flavoured estimation method as

$$\hat{\rho} = \int \rho \cdot \pi(\rho) L(\rho) d\rho \Big/ \int \pi(\rho) L(\rho) d\rho \tag{4.22}$$

where $L(\rho) = e^{\ell(\rho)}$ is the *profile likelihood function* and the prior $\pi(\rho)$ is specified in a way such that the estimates always exist and can be expressed as explicit functions of the observations. For more details of its derivation, we refer the readers to [216].

The estimate has a small bias, is highly efficient, and is simple and fast to compute. With an esitmation $\hat{\rho}$ from Eq.(4.22), the final estimates for κ and σ

are given by

$$\hat{\kappa} = \frac{1}{M}\sum_{i=1}^{M}\log\left(1+\hat{\rho}O_i\right), \qquad \hat{\sigma} = \frac{\hat{\kappa}}{\hat{\rho}} \tag{4.23}$$

4.3.2 Weight Smoothing

The original importance weights, $\{W_i, i = 1, \ldots, n\}$, are smoothed by replacing the M largest weights with the expected values of the order statistics of the fitted GPD $F(w; \hat{\mu}, \hat{\sigma}, \hat{\kappa})$, i.e.,

$$W_{[n-M+m]} = F^{-1}\left(\frac{m-1/2}{M}\right), \quad m = 1, \ldots, M \tag{4.24}$$

where $F^{-1}(\cdot)$ is the inverse cumulative distribution of the fitted $F(w; \hat{\mu}, \hat{\sigma}, \hat{\kappa})$.

Denote the resulting Pareto-smoothed importance weight for X_i as W_i^{PS}, the above weight replacement procedure is equivalent with

$$W_i^{PS} := \begin{cases} W_i & \text{if } W_i \leq \hat{\mu} \\ F^{-1}\left(\frac{m_i-n+M-0.5}{M}\right), & \text{otherwise} \end{cases} \tag{4.25}$$

where m_i is the order number of W_i in the ascendingly sorted importance weights used in the previous section.

By this procedure, we obtain the Pareto-smoothed importance weights $\{W_1^{\mathrm{PS}}, W_2^{\mathrm{PS}}, \ldots, W_n^{\mathrm{PS}}\}$, which are the basis of the Pareto-smoothed IPW estimator ($\hat{\tau}_{ATE}^{\mathrm{PS}}$) and the Pareto-smoothed self-normalized IPW estimator ($\hat{\tau}_{ATE}^{\mathrm{PSNorm}}$) introduced in the following section.

4.3.3 Estimators

Given a set of n observations $\mathcal{D} = \{(X_i, T_i, Y_i), \ldots, (X_n, T_n, Y_n)\}$, we fit the importance weights by Logistic regression, obtain the Pareto-smoothed impor-

Algorithm 4.1 Pareto-smoothed IPW ATE Estimator

Input: Observation data $\mathcal{D} = \{(X_i, T_i, Y_i), \ldots, (X_n, T_n, Y_n)\}$
Output: The estimated $\hat{\tau}_{ATE}$

1: Fit the Logistic regression propensity model $e(X) = p(T = 1|X)$ from $\mathcal{D}$;
2: Calculate the importance weights for each individual $\{W_i, i = 1, \ldots, n\}$ via E.q(4.10);
3: Sort the importance weights $\{W_i, i = 1, \ldots, n\}$ ascendingly to obtain the sorted importance weights $\{W_{[1]}, W_{[2]}, \ldots, W_{[n]}\}$
4: Choose the location parameter $\hat{\mu}$ by E.q(4.18)
5: Estimate the parameters σ and k by E.q(4.23)
6: Smooth the importance weights $\{W_1, W_2, \ldots, W_n\}$ by E.q(4.25) to obtain the Pareto-smoothed importance weights $\{W_1^{\text{PS}}, W_2^{\text{PS}}, \ldots, W_n^{\text{PS}}\}$
7: Estimate the ATE $\hat{\tau}_{ATE}$ via E.q(4.26)

tance weights $\{W_1^{\text{PS}}, W_2^{\text{PS}}, \ldots, W_n^{\text{PS}}\}$ using the above procedures, and estimate the ATE by

$$\begin{aligned}\hat{\tau}_{ATE}^{\text{PS}} &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}[T_i = 1]W_i^{\text{PS}}Y_i - \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}[T_i = 0]W_i^{\text{PS}}Y_i \\ &= \frac{1}{n}\sum_{i:T_i=1}W_i^{\text{PS}}Y_i - \frac{1}{n}\sum_{i:T_i=0}W_i^{\text{PS}}Y_i\end{aligned} \tag{4.26}$$

The process of ATE estimation by our proposed estimator is summarized in Algorithm 4.1. The corresponding process for ATT estimation is described in the Appendix. In addition, we can also make use of the self-normalization trick after weight smoothing and estimate the ATE by

$$\hat{\tau}_{ATE}^{\text{PSNorm}} = \frac{1}{n}\left(\sum_{i:T_i=1}\frac{W_i^{\text{PS}}}{\overline{W}_t^{\text{PS}}}Y_i - \sum_{i:T_i=0}\frac{W_i^{\text{PS}}}{\overline{W}_c^{\text{PS}}}Y_i\right) \tag{4.27}$$

where $\overline{W}_t^{PS} = \frac{1}{n}\sum_{i:T_i=1}W_i^{PS}$ and $\overline{W}_c^{PS} = \frac{1}{n}\sum_{i:T_i=0}W_i^{PS}$. This Pareto-smoothed self-normalized IPW estimator proceeds as Algorithm 4.1 by estimating the ATE using E.q(4.27) in the last step.

In general, the Pareto-smoothed IPW Estimator $\hat{\tau}_{ATE}^{\text{PS}}$ stabilizes the IPW es-

timator with a novel weight smoothing trick. The self-normalized IPW estimator tries to stabilize the estimate by standardizing the importance weights by the average weight in each group. The Pareto-smoothed self-normalized IPW estimator ($\hat{\tau}_{ATE}^{\text{PSNorm}}$) takes advantage of the self-normalization trick used by $\hat{\tau}_{ATE}^{\text{Norm}}$ and further stabilizes $\hat{\tau}_{ATE}^{\text{PS}}$ by standardizing the smoothed importance weights.

4.3.4 Asymptotic Analysis

To analyse the asymptotic property of the propoed estimators using results from existing literature, define the weight function

$$\lambda(X_i) := \frac{\mathbb{1}[T_i = 1]n_1}{n} \frac{W_i^{\text{PS}}}{\overline{W}_t^{\text{PS}}} + \frac{\mathbb{1}[T_i = 0]n_0}{n} \frac{W_i^{\text{PS}}}{\overline{W}_c^{\text{PS}}} \tag{4.28}$$

We can conclude according to Eq.(4.25) that $\lambda_i = \lambda(X_i)$ is a function of the covariates X_i parameterized by the propensity model parameters and the fitted GPD parameters. With this notation, we can rewrite $\hat{\tau}_{ATE}^{\text{PSNorm}}$ in Eq.(4.27) as a standard weighting estimator

$$\hat{\tau}_{ATE}^{\text{PSNorm}} = \frac{1}{n_1} \sum_{i:T_i=1} \lambda_i Y_i - \frac{1}{n_0} \sum_{i:T_i=0} \lambda_i Y_i \tag{4.29}$$

where the weights λ_i satisfy the following two summation restrictions:

$$\frac{1}{n_1} \sum_{i:T_i=1} \lambda_i = \frac{1}{n} \sum_{i:T_i=1} \frac{W_i^{\text{PS}}}{\overline{W}_t^{\text{PS}}} = 1$$

and

$$\frac{1}{n_0} \sum_{i:T_i=0} \lambda_i = \frac{1}{n} \sum_{i:T_i=0} \frac{W_i^{\text{PS}}}{\overline{W}_c^{\text{PS}}} = 1$$

Define the two conditional variance functions $\sigma_0^2(\text{x}) := \mathbb{V}(Y(0)|X = \text{x})$ and $\sigma_1^2(\text{x}) := \mathbb{V}(Y(1)|X = \text{x})$. According to the results in [31, 211] and [74, Chap. 19],

if the weighting function $\lambda(X_i)$ is continuous and differentiable, the estimator $\hat{\tau}_{ATE}^{\text{PSNorm}}$ is asymptotic linear and its asymptotic variance can be approximated by

$$\mathbb{V}(\hat{\tau}_{ATE}^{\text{PSNorm}}) = \frac{1}{n_1^2} \sum_{i:T_i=1} \lambda_i^2 \cdot \sigma_1^2(X_i) + \frac{1}{n_0^2} \sum_{i:T_i=0} \lambda_i^2 \cdot \sigma_0^2(X_i)$$

However, it is easy to verify that the weighting function $\lambda_i = \lambda(X_i)$ in Eq.(4.28) is not smooth nor differentiable. In this case, we cannot guarantee the consistency of the proposed estimator $\hat{\tau}_{ATE}^{\text{PSNorm}}$. Moreover, the inference of its asymptotic variance is an open problem in the causal inference literature and existing methods are unable to conduct inference to the population [31, 211]. Analogously, the Pareto-smoothed IPW estimator $\hat{\tau}_{ATE}^{\text{PS}}$ is also inconsistent. To quantify the estimation uncertainty of the causal estimators, in the simulation and experiment sections, we replicate the experiments multiple times and report the empirical standard error of each estimator.

We summarize this section by comparing the proposed Pareto-smoothing method with the weight truncation method for causal estimator stabilization. Both methods are biased. while the weight truncation method truncate the extreme weights by fixed values, our proposed method tries to smooth them and keep their relative order. As a result, the proposed Pareto-smoothed estimators are expected to be less biased than truncated estimators. This is validated by the empirical results in the simulation experiments.

4.4 Simulation Studies

Since the ground truth counterfactual outcomes are not available in real-world observational datasets, evaluating causal inference algorithms is not straightforward. In this section, we validate our proposed method using simulated and semi-simulated data, where the ground truth is available to us such that we

Table 4.1: Abbrivation (Abbr.) and description of ATE estimators

Estimator	Abbr.	Description
$\hat{\tau}_{ATE}^{\text{Naive}}$	Naive	Naive estimator for ATE as in E.q(4.3)
$\hat{\tau}_{ATE}^{\text{IPW}}$	IPW	IPW estimator for ATE as in E.q(4.11)
$\hat{\tau}_{ATE}^{\text{Trunc}}$	Trunc	Truncated IPW estimator for ATE as in E.q(4.12) with the truncation thresholds $a = 1$ and b specified by E.q(4.17)
	TruncNorm	Truncated IPW estimator for ATE by normalizing the truncated importance weights used in the *Trunc* estimator
	TruncCrump	Truncated IPW estimator for ATE with weight truncation thresholds $a = \frac{10}{9}$ and $b = 10$ in (4.13)
$\hat{\tau}_{ATE}^{\text{Norm}}$	Norm	IPW estimator for ATE with weight self-normalization for ATE as in E.q(4.14)
$\hat{\tau}_{ATE}^{\text{PS}}$ (Ours)	PS	Pareto-smoothed IPW estimator for ATE as in E.q(4.26)
$\hat{\tau}_{ATE}^{\text{PSNorm}}$ (Ours)	PSNorm	Pareto-smoothed self-normalized IPW estimator for ATE as in E.q(4.27)

can evaluate the performance of different methods. Descriptions of all ATE estimators used in the work are listed in Table 4.1. Specifically, according to the criterion used for choosing the truncation thresholds, we specify three variants of the truncated IPW estimator. The first truncation estimator, *Trunc*, uses the same criterion, E.q(4.17), to specify the truncation thresholds as our Pareto-smoothed estimators. The second truncation estimator, *TruncCrump* uses the truncation threshold in [31] (discussed in Section 4.2.3). In addition, we also use the self-normalization trick used in the proposed Pareto-smoothed self-normalized IPW estimator to the *Trunc* estimator and denote the resulting estimator the *TruncNorm* estimator. In all the experiments, we follow most of the literature on propensity score estimation and use Logistic regression to fit the propensity scores. In all simulations, the underlying potential outcomes $Y_i(0)$

and $Y_i(1)$ for each individual are known, so we can calculate the true sample ATE empirically by $\tau_{ATE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i(1) - Y_i(0))$. For an estimator $\hat{\tau}_{ATE}$, its estimation bias is calculated by

$$Bias_{\text{ATE}} = |\hat{\tau}_{ATE} - \tau_{ATE}| = \left|\hat{\tau}_{ATE} - \frac{1}{n}\sum_{i=1}^{n}(Y_i(1) - Y_i(0))\right|$$

4.4.1 Simulated Data

We simulated data in the context of both low dimensional and relatively high dimensional covariates.

Low dimensional covariates

There are two pre-treatment covariates in the first simulation: one binary $X_{i1}|T_i = 0 \sim \text{Bernoulli}\ (0.4)$,$X_{i1}|T_i = 1 \sim \text{Bernoulli}\ (0.5)$ and one continuous $X_{i2}|T_i = 0 \sim \mathcal{N}(-1.0, 1)$, $X_{i2}|T_i = 1 \sim \mathcal{N}(1.0, 1)$. We simulated data with sample size $n = 100, 200, 300, 500, 1000, 1500, 2000$. For each sample size, we assigned exactly half of the subjects to the treatment group $T = 1$, and the other half to the control group $T = 0$. The potential outcomes for each subject i are adapted from [124] with $Y_i(0) = 0.85X_{i2} + 0.05X_{i2}^2 + 2$ and $Y_i(1) = 0.25X_{i1} + (1 + \exp(1 - 0.85X_{i2}))^{-1}$. With this data generating process, the true ATE is approximately -1.52. For each sample size n, we replicated the experiment 1000 times. The result on ATE estimation biases and standard errors is listed in Table 4.2. To clearly compare the estimation performance, we also illustrate the estimation bias in terms of the sample size in Fig. 4.1.

As we can see from the result, as the sample size increases, all importance weight-based estimators obtain better estimates. In general, estimators based on our proposed Pareto-smoothing method, i.e., the *PS* estimator and the *PSNorm* estimator, achieve the best performance in all sample sizes. As two unbiased

Table 4.2: Comparison of ATE estimation bias and standard error (SE) averaged over 1000 replicates on the simulated low-dimensional covariate data for different estimators list in Table 4.1

n	**Naive**	**IPW**	**Trunc**	**TruncNorm**	**TruncCrump**	**Norm**	**PS**	**PSNorm**
100	$1.194 \pm .004$	$0.761 \pm .042$	$1.173 \pm .005$	$1.494 \pm .006$	$0.651 \pm .008$	$0.769 \pm .041$	**$0.608 \pm .013$**	$0.647 \pm .013$
200	$1.194 \pm .003$	$0.668 \pm .033$	$1.173 \pm .003$	$1.491 \pm .004$	$0.624 \pm .006$	$0.656 \pm .033$	**$0.515 \pm .012$**	$0.526 \pm .012$
300	$1.198 \pm .002$	$0.551 \pm .030$	$1.142 \pm .003$	$1.439 \pm .004$	$0.616 \pm .005$	$0.548 \pm .029$	**$0.449 \pm .011$**	$0.461 \pm .010$
500	$1.197 \pm .002$	$0.460 \pm .023$	$1.085 \pm .002$	$1.354 \pm .003$	$0.612 \pm .004$	$0.458 \pm .023$	**$0.390 \pm .010$**	$0.396 \pm .010$
1000	$1.197 \pm .001$	$0.430 \pm .035$	$0.995 \pm .002$	$1.223 \pm .002$	$0.609 \pm .002$	$0.425 \pm .035$	$0.334 \pm .009$	**$0.333 \pm .008$**
1500	$1.198 \pm .002$	$0.358 \pm .022$	$0.946 \pm .001$	$1.154 \pm .002$	$0.608 \pm .002$	$0.358 \pm .022$	**$0.268 \pm .006$**	$0.271 \pm .006$
2000	$1.196 \pm .001$	$0.291 \pm .012$	$0.907 \pm .001$	$1.100 \pm .002$	$0.604 \pm .002$	$0.287 \pm .012$	$0.247 \pm .006$	**$0.246 \pm .006$**

Table 4.3: Comparison of ATE estimation bias and standard error (SE) averaged over 1000 replicates on the simulated high-dimensional covariate data for different estimators list in Table 4.1

n	**Naive**	**IPW**	**Trunc**	**TruncNorm**	**TruncCrump**	**Norm**	**PS**	**PSNorm**
500	$1.876 \pm .014$	$1.657 \pm .106$	$3.869 \pm .015$	$2.100 \pm .012$	$1.535 \pm .020$	$1.774 \pm .107$	**$1.300 \pm .035$**	$1.388 \pm .038$
1000	$1.892 \pm .010$	$1.172 \pm .047$	$3.353 \pm .012$	$1.908 \pm .010$	$1.478 \pm .015$	$1.270 \pm .048$	**$0.990 \pm .027$**	$1.064 \pm .029$
1500	$1.895 \pm .008$	$0.968 \pm .036$	$3.064 \pm .010$	$1.781 \pm .008$	$1.435 \pm .012$	$1.057 \pm .038$	**$0.839 \pm .026$**	$0.907 \pm .027$
2000	$1.898 \pm .007$	$0.887 \pm .034$	$2.887 \pm .008$	$1.705 \pm .007$	$1.439 \pm .010$	$0.965 \pm .035$	**$0.737 \pm .018$**	$0.799 \pm .020$
2500	$1.910 \pm .006$	$0.784 \pm .026$	$2.765 \pm .007$	$1.653 \pm .006$	$1.448 \pm .009$	$0.860 \pm .027$	**$0.697 \pm .018$**	$0.758 \pm .019$
3000	$1.896 \pm .006$	$0.764 \pm .029$	$2.641 \pm .007$	$1.589 \pm .006$	$1.431 \pm .008$	$0.829 \pm .030$	**$0.660 \pm .017$**	$0.713 \pm .018$

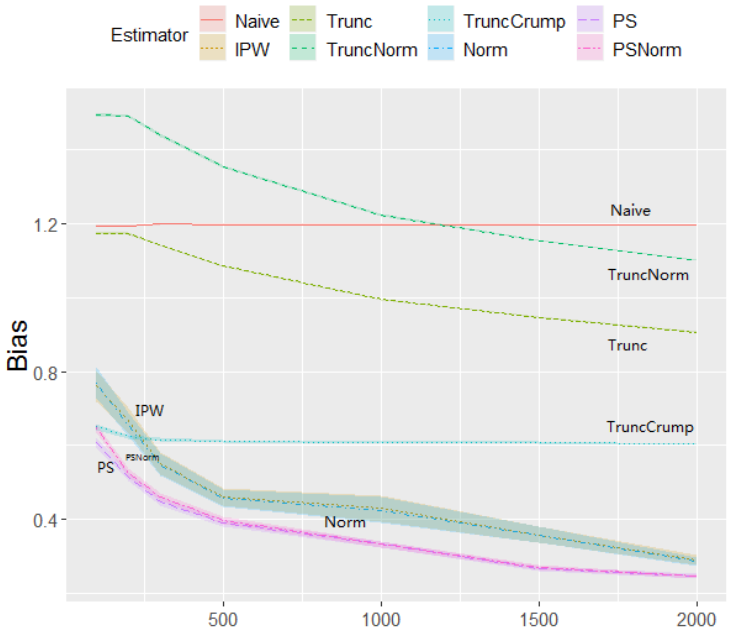


Figure 4.1: ATE estimation bias and standard error in terms of sample size n over 1000 replicates for the simulated low-dimensional covariate data.

estimators, the *IPW* estimator and the *Norm* estimator achieve similar performance. Among the three weight stabilization methods, weight truncation, self-normalization and Pareto-smoothing, our proposed Pareto-smoothing method is the least biased and is more stable than the self-normalization method. By further comparing *TruncNorm* and *Trunc* as well as *PSNorm* and *PS*, we find that self-normalization is likely to worsen the estimation when the sample size is small and one weight stabilization strategy has already been used, either truncation or Pareto-smoothing.

High dimensional covariates

With finite data, the estimated importance weights are more likely to be highly variable in settings with high dimensional covariates. To investigate the performance of the proposed Pareto-smoothing method in this setting, we

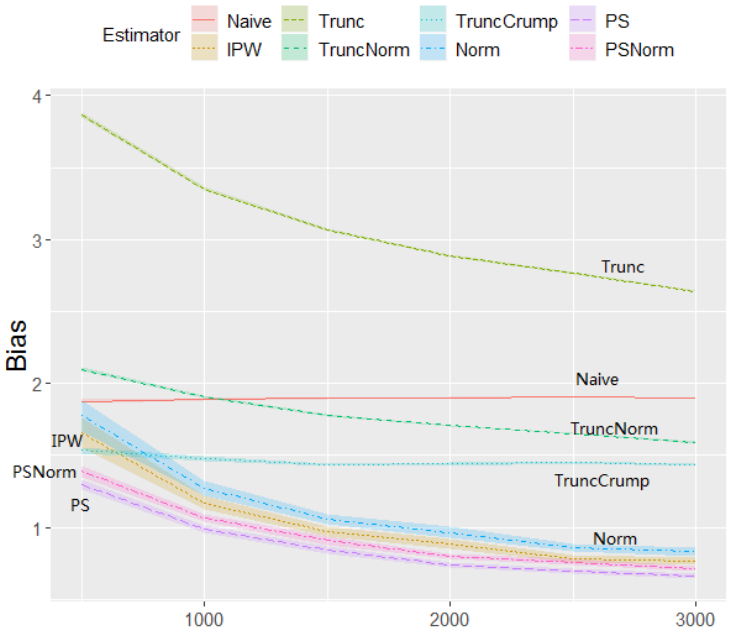


Figure 4.2: ATE estimation bias and standard error in terms of sample size n over 1000 replicates for the simulated high-dimensional covariate data.

adapt the simulation in [124] and generate data by assigning half of the samples to the treatment group $T = 1$, and the other half to the control group $T = 0$. In this simulation, there are 10 confounders, 5 binary and 5 continuous. The values of the binary confounders are generated by $X_i|T = 0 \sim \text{Bernoulli}(0.4), X_i|T = 1 \sim \text{Bernoulli}(0.45), i \in \{1, 2, 3, 4, 5\}$ and that of the continuous confounders was generated by $X_i|T = 0 \sim \mathcal{N}(-1, 3^2), X_i|T = 1 \sim \mathcal{N}(1.25, 3^2), i \in \{6, 7, 8, 9, 10\}$. The potential outcomes were generated so that they exhibit non-linear trends in the estimated propensity scores. For each individual i, the two underlying potential outcomes are generated by $Y_i(0) = 5 + 0.2\,(X_{i1} + X_{i2} + X_{i3} + X_{i4} + X_{i5}) + (1 + \exp(1 - 8X_{i5}))^{-1} + X_{i7} + X_{i8} + X_{i9} + X_{i10}$ and $Y_i(1) = -5 + 0.2(X_{i1} + X_{i2} + X_{i3} + X_{i4} + X_{i5}) - 0.5(X_{i6} + X_{i7} + X_{i8} + X_{i9} + X_{i10})$

We simulate data with sample size $n = 500, 1000, 1500, 2000, 2500, 3000$. The treatment propensity scores are unknown and are estimated via simple Logistic

regression, which is clearly misspecified. Estimation biases of different estimators are listed in Table 4.3. The estimation biases and corresponding standard errors in terms of the sample size are illustrated in Fig.4.2. The result is similar with that in the low-dimensional covariate setting, the proposed *PS* estimator obtains the lowest estimation biases in all sample sizes. The proposed *PSNorm* estimator performs slightly worse than *IPW* but better than the other estimators. Results for this high-dimensional covariate setting further validate the superiority of our proposed Pareto-smoothing method.

4.4.2 Semi-simulated Data: IHDP

In this section, we evaluate the performance of our algorithm through the semi-simulated dataset based on the Infant Health and Development Program (IHDP) which was introduced in [62] and used as a benchmark dataset in the causal inference literature [102, 178, 225]. The IHDP is a real randomized experiment to enhance the cognitive and health status of low birth weight, premature infants through paediatric follow-ups and parent support groups. The observed covariates and treatments in the semi-simulated data are from the IHDP program, while all outcomes (response surfaces) are simulated so that the true treatment effects are known. In total, the IHDP dataset consists of 747 individuals (139 treated, 608 control), and 25 covariates measuring the properties of children and their mothers. The binary treatment T indicates whether the child was assigned into a program where both intensive high-quality childcare and home visits from a trained provider are provided. Examples of covariates include the sex and birth weights of the child, and the age and education attainment level of the mother.

We conduct experiments on all three response simulation settings proposed in [62]. In setting A, the response surfaces are linear and parallel across the two treatment groups and there is no treatment effect heterogeneity. The response

surfaces for settings B and C are nonlinear and not parallel across treatment conditions. The outcomes are simulated so that the underlying treatment effects are 4.0. For more details of the three simulation settings, refer to [62]. We simulated the outcomes using the NPCI package[1] and ran the experiment 1000 times. The boxplot of the ATE estimates of different estimators in three settings is illustrated in Fig.4.3. Results of estimation biases are listed in Table 4.4.

Table 4.4: Results for the IHDP dataset. A, B and C stand for three outcome simulation settings. Estimation biases and standard errors (SE) are computed by replicating the experiment 1000 times.

	A		**B**		**C**	
	Bias	**SE**	**Bias**	**SE**	**Bias**	**SE**
Naive	0.949	(0.048)	0.711	(0.016)	0.491	(0.014)
IPW	0.744	(0.029)	0.652	(0.015)	0.461	(0.014)
Trunc	6.854	(0.178)	4.989	(0.029)	2.744	(0.030)
TruncNorm	0.733	(0.037)	0.542	(0.013)	0.371	(0.010)
TruncCrump	2.104	(0.059)	1.447	(0.019)	0.519	(0.015)
Norm	0.696	(0.036)	0.539	(0.013)	0.432	(0.013)
PS	0.779	(0.030)	0.678	(0.015)	0.456	(0.013)
PSNorm	0.697	(0.036)	0.538	(0.013)	0.428	(0.012)

As we can see from Fig.4.3, the *Norm* and *PSNorm* estimators perform similarly in all three settings, with estimated ATEs around the true ATE. The estimators *IPW*, *Trunc*, *TruncCrump* and *PS* tend to under-estimate the ATE in setting A and B. Furthermore, the result in Table 4.4 indicates that the *Norm* estimator and the *PSNorm* estimator achieve the lowest bias in settings A and B respectively. While in setting C, the *TruncNorm* estimator performs the best. In addition, we find that weight truncation or Pareto-smoothing alone deteriorates the ATE estimation performance in settings A and B. The reason may be that since the treatment assignments in the IHDP data are random, the negative influence of weight truncation and Pareto-smoothing proposed for handling

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

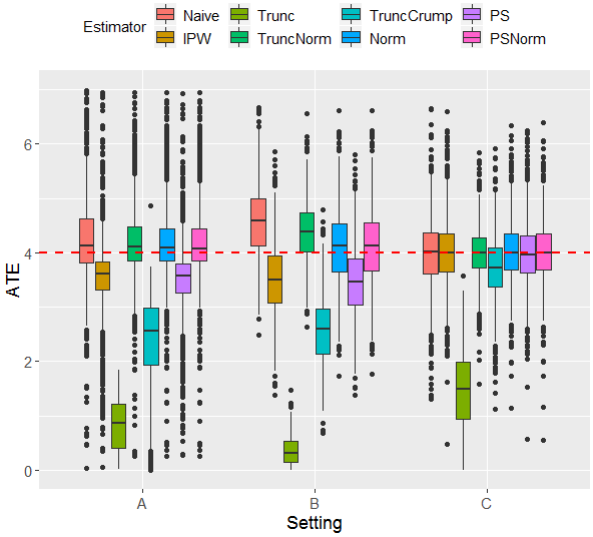


Figure 4.3: Box plot of ATEs estimated by different estimators for the IHDP data in different settings. The underlying true ATE is 4 (the red dashed line) in all settings.

extreme importance weights surpasses the benefit they bring for this balanced dataset. Fortunately, by combining them with weight self-normalization, the resulting *TruncNorm* and *PSNorm* estimators acheive better estimation than the naive *IPW* estimator.

4.5 Application to the NHEFS Data

The National Health and Nutrition Examination Survey (NHANES) is a population survey designed to assess the heath and nutritional status of adults and children in the United States. It was jointly initiated by the National Center for Health Statistics (NCHS), which is part of the Centers for Disease Control and Prevention, and the National Institute on Aging in collaboration with other agencies of the US Public Health Service. The datasets, with a detailed description and documentation, are publicly available online[2]. Under the NHANES,

[2] https://wwwn.cdc.gov/nchs/nhanes/nhefs/default.aspx/

the NHANES I Epidemiologic Follow-up Study (NHEFS) was designed to investigate the relationships between clinical, nutritional and behavioural factors assessed in NHANES I.

We use the subset of the NHEFS dataset used in [61] to estimate the ATE of smoking cessation on weight gain. There are 1746 cigarette smokers in the original data with a baseline visit in the year of 1971-1975. After removing missing and censored records, there are 1566 individuals left, aged 25-74 years old and with a follow-up visit in 1982. Individuals who reported having quit smoking before the follow-up visit are classified as treated $T = 1$, and as untreated $T = 0$ otherwise. The outcome variable – weight gain Y – of each individual is the body weight at the follow-up visit minus the body weight at the baseline visit, measured in kg. Examples of pre-treatment covariates X include the age, sex, race, baseline weight, and smoking intensity of each individual.

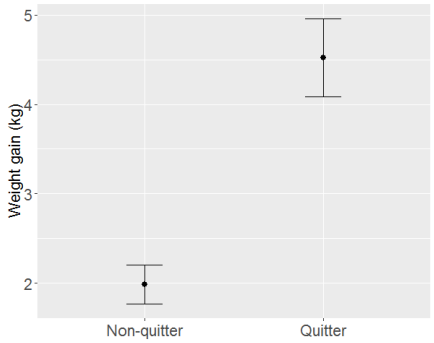


Figure 4.4: The mean and standard deviation of weight gains for smoking non-quitters and quitters in the NHEFS data.

Of the selected 1566 individuals, 1163 are non-quitters and the other 403 are quitters. The mean weight gain of non-quitters and quitters is $1.98kg$ and $4.53kg$ respectively (see Fig.4.4), which means that for the studied individuals, quitters experience approximately $2.55kg$ more weight gain than non-quitters on average. However, as we have discussed, this associational mean difference $\mathbb{E}[Y|T = 1] - \mathbb{E}[Y|T = 0]$ is not the causal effect $\mathbb{E}[Y(1)] - \mathbb{E}[Y(0)]$ of quitting smoking because

of the existence of confounding bias. For example, in the observational data, older people are more likely to quit smoking and gain less weight than younger people regardless of whether they quit smoking. Moreover, more males quit smoking than females. As a result, naive estimation from the observational data may underestimate the true treatment effects and we need to adjust for these possible confounders for the purpose of treatment effect estimation. Following [61], we assume unconfoundedness conditional on the observed covariates.

Table 4.5: Estimation results for the NHEFS dataset. The ATE and standard errors (SE) are computed from 1000 replications. The reference estimation in [61] is $3.4kg$ with a 95% confidence interval of $2.4 \sim 4.5kg$.

	ATE	**SE**
Naive	2.534	(0.010)
IPW	3.412	(0.010)
Trunc	2.466	(0.009)
TruncNorm	3.245	(0.010)
TruncCrump	3.354	(0.010)
Norm	3.432	(0.010)
PS	3.412	(0.010)
PSNorm	3.438	(0.010)

We use the same Logistic regression model used in [61] to fit the treatment propensities and replicate the estimation 1000 times by randomly sampling 70% individuals from each group (quitter and non-quitter) in each replication. The estimated ATE and corresponding empirical standard errors are listed in Table 4.5. As we can see from the result, except for the *Naive* estimator, the *Trunc* estimator and the *TruncNorm* estimator, all the other estimators have a similar estimated ATE of about $3.4kg$, i.e., quitting smoking increases weight by about $3.4kg$ for the investigated population. This estimate is quite close to that in [61], which is $3.4kg$ with a 95% confidence interval of $2.4 \sim 4.5kg$. Though the ground truth is unknown, we can conclude from the result that the estimates

of the proposed Pareto-smoothed estimators match existing unbiased estimators (the IPW estimator and the self-normalized IPW estimator). In addition, by comparing the estimates of different truncated IPW estimators, we find that the estimate of truncated estimators is sensitive to the selected truncation thresholds.

4.6 Summary

In this chapter, we reframe the classic IPW estimator for causal inference into the framework of expectation estimation using importance sampling. To handle extreme importance weights commonly existed in importance weight-based estimators using finite samples, we take advantage of the smoothing property of the GPD from the extreme value statistics and propose a new Pareto-smoothing method to stabilize the IPW causal estimator. Based on this method, we further propose two Pareto-smoothed causal estimators, the Pareto-smoothed IPW estimator and the Pareto-smoothed self-normalized IPW estimator. Comprehensive experiments using both simulated and semi-simulated data indicate that, for causal inference from finite observational data, the proposed Pareto-smoothed estimators generally achieve lower bias than estimators using weight truncation or weight self-normalization. Moreover, they are more stable than the vanilla IPW estimator and the self-normalized IPW estimator. We also validate the proposed method with a real-world health dataset.

Note that although we focus on IPW-based estimation of the ATE in this work, the key component of the proposed method is in principle to stabilize the estimated importance weights by fitting a GPD over the tail to smooth the extreme weights. This is quite general and can be easily adapted for the estimation of other causal estimands (e.g., ATT and ATC) with any other propensity score based causal estimators. As a result, one of our future research undertakings

will be to investigate the application of the proposed method in other causal estimators such as propensity score matching and weighted outcome regression.

In addition, we assume unconfoundedness and estimate the treatment propensities with all the observed pre-treatment covariates for simplicity in this work. Many researchers have recently noticed that variable selection in propensity score estimation using the outcome adaptive LASSO [181] or the highly adaptive LASSO [81] can also stabilize the resulting propensity score-based estimators. We believe this will also be beneficial for our proposed estimators and leave that item for future study.

Chapter 5

Counterfactual Inference with Hidden Confounders Using Implicit Generative Models

As we have introduced in Chapter 2, counterfactual inference tries to fulfil the problem of causal inference by learning the treatment exposure surfaces. One of the biggest challenges in counterfactual inference is the existence of unobserved confounders, which are latent variables that affect both the treatment and outcome variables. Building on recent advances in latent variable modelling and efficient Bayesian inference techniques, deep latent variable models, such as variational auto-encoders (VAEs) [91], have been used to ease the challenge by learning the latent confounders from the observations [102].

However, for the sake of tractability, the posterior of latent variables used in existing methods is assumed to be Gaussian with diagonal covariance matrix. This specification is quite restrictive and even contradictory with the underlying truth, limiting the quality of the resulting generative models and the causal effect estimation. In this chapter, we propose to take advantage of implicit

generative models to detour this limitation by using black-box inference models. In addition, to make inference for the implicit generative model with intractable likelihood, we take advantage of recent advances in implicit variational inference based on adversary training to obtain a close approximation to the true posterior.

This work has been published in [225]. The remainder of this chapter is organized as follows: in Section 5.2, we firstly introduce preliminary knowledge on causal models and implicit models; details of the proposed method are presented in Section 5.3; Section 5.4 illustrates our experiments on two benchmark datasets; we conclude this chapter in Section 5.5.

5.1 Problem Setup

Denote the treatment space by $\mathcal{T}$, the set of contexts by $\mathcal{X}$, and the set of possible outcomes by $\mathcal{Y}$. For example, for an employee with covariates $x \in \mathcal{X}$, the set of treatments $\mathcal{T}$ might be whether he or she joined a specific training program and the set of outcomes might be $Y = [0, 10K]$ indicating his/her monthly salary in dollars. For an individual x (e.g., an employee), let $Y_t(x) \in \mathcal{Y}$ be the potential outcome of x under the treatment $t \in \mathcal{T}$. The fundamental problem of causal inference is that only one of potential outcomes $Y_t(x), t \in \mathcal{T}$ is observed for a given individual x. In the machine learning literature, this kind of partial feedback is often called *bandit feedback*[191, 192].

Without loss of generality, we consider the case of a binary treatment set, i.e., $\mathcal{T} = \{0, 1\}$, where $t = 1$ indicates the individual is allocated into the *treated* group and $t = 0$ the *control* group. In this setting, the individual treatment effect $ITE(x) = Y_1(x) - Y_0(x)$ for individual x is of high interest. Knowing this quantity enables us to choose the best treatment options and to give personalized recommendations. Based on ITE, the average treatment effect,

$ATE = \mathbb{E}_{x\sim p(x)}[ITE(x)]$, for a population with distribution $p(x)$ quantifies the average treatment effect difference between the two actions. Sometimes, we are only interested in the ATE for the treated group, i.e., the average treatment effect on the treated, $ATT = \mathbb{E}_{x\sim p(x)}[ITE(x)|t = 1]$.

The problem of causal effect estimation from observational data has been studied extensively in the literature [18, 192, 80, 178, 102]. One of the most widely used approaches is counterfactual inference, also known as potential outcome modelling. The main idea is: given n samples $\mathcal{D} = \{(x_1, t_1, y_1), \ldots, (x_n, t_n, y_n)\}$, where the observed *factual* outcome $y_i = t_i Y_1(x_i) + (1 - t_i)Y_0(x_i)$, if we can unbiasedly learn the potential outcome model $Y_t(x) = h(x, t)$ using the observed data, the estimated ITE is then

$$\widehat{ITE}(x_i) = \begin{cases} y_i - h(x_i, 0), & t_i = 1 \\ h(x_i, 1) - y_i, & t_i = 0 \end{cases} \tag{5.1}$$

Therefore, the key is to learn the potential outcome function $h(x, t)$. In the literature, $Y_0(x) = h(x, 0)$ and $Y_1(x) = h(x, 1)$ are also called the *response surfaces*. As a learning problem, this is different from classic learning in that we never see the individual-level treatment effect in the observations. Because of the existence of unobserved confounders that affect both the treatment assignment and the outcome, naïvely fitting the outcome model from observational data is subject to confounding bias [74, 61].

5.2 Preliminaries

In this section, we introduce two basic components of our proposed method introduced in the next section: the structural causal models and IGMs.

5.2.1 Structural Causal Models

Structural causal models [131], or functional causal models, defined in Definition 5.1, represent variables as deterministic functions of their parents and exogenous noises. They take advantages of the functional causal semantics of structural equation models (SEMs) [196] and the representa-tion and reasoning power of Bayesian networks [130].

Definition 5.1 (Structural causal model, SCM)**.** *A structural causal model $\mathcal{M}$ is a tuple $(\boldsymbol{V}, U, F, P(\mathrm{u}))$ that consists of (i) a set of observed endogenous variables $\boldsymbol{V} = \{V_1, \cdots, V_n\}$; (ii) a set of unobserved background (or exogenous) variables U; (iii) a set of causal mechanisms $F = \{f_1, \cdots, f_n\}$ that determines the endogenous variables $\boldsymbol{V}$; and (iv) the joint distribution $P(\mathrm{u})$ over the background variables U. Each causal mechanism f_i tells us the value of $V_i \in V$ given the value of all other variables, i.e., $V_i \leftarrow f_i\left(PA_i, U\right), U \sim P(\mathrm{u})$, where $PA_i \subseteq \boldsymbol{V} \backslash V_i$ is called the parents of V_i.*

In this definition, the endogenous variables $\boldsymbol{V}$ are regarded as deterministic functions of other variables and randomness comes from unobserved exogenous variables U. Together with Pearl's *do*-calculus and counterfactual notations [131], it permits us to answer intervention and counterfactual questions. In this chapter, we consider causal models with the observed set $\boldsymbol{V}$ including a treatment variable t, an outcome variable y, and some evidence variables x that act as proxies of the unobserved confounders z. The corresponding causal graph (or data-generating process) is illustrated as in Fig.5.1. In this setting, the following Theorem 5.1 gives the identifiability condition of causal effect.

Theorem 5.1. *[102] If we can recover the joint distribution $P(x, z, t, y)$, then we can identify the ITE under the causal model represented in Fig.5.1.*

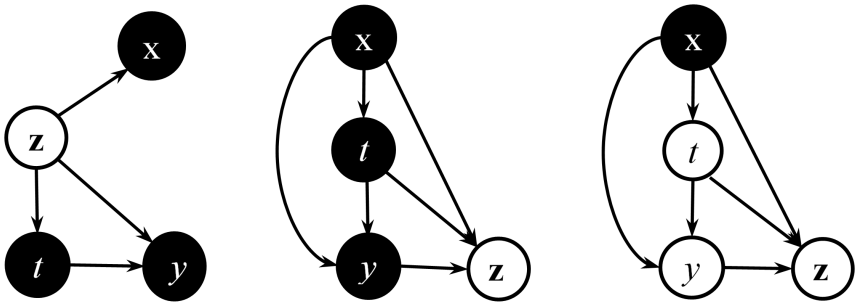


Figure 5.1: Left: The underlying causal model (the generative model). Middle: The inference model for with-in-sample observations. Right: The inference model for out-of-sample data. Solid nodes are observed and hollow nodes are unobserved.

5.2.2 Implicit Generative Models

In probabilistic machine learning, IGMs [116, 194] capture an unknown distribution by hypothesizing about its data-generating process. For a distribution $p(\mathrm{x})$ of observations x, we define a function g that takes in noise $\epsilon \sim p(\epsilon)$ and output x given parameters θ (including possible null set of parents),

$$\mathrm{x} = g(\epsilon|\theta), \qquad \epsilon \sim p(\epsilon)$$

The induced implicit density of $\mathrm{x} \in S$ given θ is derived via

$$p(\mathrm{x} \in S) = \int_{\{g(\epsilon|\theta)=\mathrm{x}\in S\}} p(\epsilon)d\epsilon \tag{5.2}$$

In an IGM, the function g is usually a deep neural network that is a universal approximator to any continuous function. By separating randomness (noise ϵ) from the transformation (function g), IGMs imitate the structural invariance of causal models [193]. A weakness of IGMs is that the integral in Eq.(5.2) is typically intractable and does not admit a tractable likelihood, making the inference of the parameters very difficult.

In its general form, an SCM $\mathcal{M}$ defined in Definition 5.1 is a non-parametric causal model, and each structural equation in F is a nonlinear, nonparametric generalization of linear SEMs [134]. SCMs work regardless of the type of equations, linear or nonlinear, parametric or non-parametric. That means, SCMs provide us a framework to conduct causal modeling and reasoning. Existing simple parametric models apply simple nonlinearities such as polynomials, hand-engineered low order interactions between variables, and assume additive interactions with Gaussian noise. Deep neural networks provide us rich models to encode the causal mechanisms in high-dimensional complex causal systems. Recently, Tran and Blei [193] proposed to use implicit causal models (ICMs). Analogous to the well known approximator theorem of feedforward neural networks [34], they present a similar universal approximation theorem for using implicit models to approximate causal models, as formally described in Theorem 5.2.

Theorem 5.2. *In an SCM $\mathcal{M} = (V, U, F, s(\mathrm{u}))$, assume each causal mechanism is a continuous function on the n-dimensional unit cube $f \in \mathcal{C}([0,1]^n)$. Let σ be a non-constant, bounded, and monotonically increasing continuous function. For each causal mechanism f and any error $\sigma > 0$, there exist parameters $\theta = (\alpha, \beta, b)$ for a H layer neural network, where $\alpha_h, b_h \in \mathbb{R}$ and $\beta_h \in \mathbb{R}^n, h = 1, 2, , H$, such that the following function approximates f:*

$$\forall v \in [0,1]^n \qquad g(v|\theta) = \sum_{h=1}^{H} \alpha_h \sigma\left(\beta_h^T v + b_h\right), \quad |g(v|\theta) - f(v)| < \delta$$

Besides the universal approximation property of deep implicit models for causal mechanisms, recent advances in the machine learning community, for example, approximate Bayesian computation [77], adversarial training [50, 78], and probabilistic programming [194], permit us to use fast algorithms for their Bayesian inference of the parameters.

5.3 Counterfactual Inference Using IGMs

In this section, we firstly introduce our proposed counterfactual inference method using implicit models. The lower bound objective and implicit variational inference method based on adversary training are then presented.

5.3.1 Latent Variable Modelling for Causal Models

As discussed in Section 5.1, we need to learn the potential outcome function $Y_t(\mathrm{x}) = h(\mathrm{x}, t)$. If the latent confounders are available, we can estimate the potential outcome by the following adjustment formula [131, 61]:

$$Y_t(\mathrm{x}) = h(\mathrm{x}, t) = \mathbb{E}\left[Y|\mathrm{x}, t\right] \tag{5.3}$$

However, when the underlying confounders are unobserved and the data-generating process is illustrated as the generative model in Fig.5.1, we need to uncover the posterior of the unobserved confounders z and then estimate the potential outcomes via:

$$Y_t(\mathrm{x}) = h(\mathrm{x}, t) = \mathbb{E}\left[Y_t|\mathrm{x}\right] = \int_{\mathrm{z}} \mathbb{E}[Y|\mathrm{z}, t]p(\mathrm{z}|\mathrm{x})d(\mathrm{z}) \tag{5.4}$$

Learning confounders for causal inference has its root from the *abduction-action-prediction* procedure for counterfactual inference [134, Chap. 4]. Instead of such a multi-stage induction process, in this work, we propose to jointly learn the response surfaces and latent confounder space. This is analogous to deep generative models which learn the *generative* and *inference* models jointly. The generative and inference models for our proposed method are illustrated in Fig.5.1. For an observed tuple (x_i, t_i, y_i), the log-likelihood is

$$\log p\left(\mathrm{x}_i, t_i, y_i\right) = \int \log p_\theta\left(\mathrm{x}_i, t_i, y_i|\mathrm{z}_i\right) p\left(\mathrm{z}_i\right) d\mathrm{z}_i \tag{5.5}$$

where θ denotes the generative parameters. The generative model for each component in the tuple (x_i, t_i, y_i) is

$$\begin{aligned} \mathrm{x}_i &\sim p_\theta\left(\mathrm{x}|\mathrm{z}_i\right) \\ t_i &\sim p_\theta\left(t|\mathrm{z}_i\right) \\ y_i &\sim p_\theta\left(y|t_i, \mathrm{z}_i\right) \end{aligned} \tag{5.6}$$

We put Gaussian priors on the latent confounders z_i, i.e., $\mathrm{z}_i \sim \mathcal{N}\left(z|0, I_M\right)$

$$\mathrm{z}_i = \mathrm{z}_\phi\left(\mathrm{x}_i, t_i, y_i, \epsilon\right), \quad \epsilon \sim s(\cdot) \tag{5.7}$$

where ϕ denotes the variational parameters. Based on Eq.(5.2), the induced implicit density is denoted as $q_\phi(\mathrm{z}|x, t, y)$. According to the generative models in E.q(5.6), we can obtain the decoder from latent variables z_i to the observed tuple (x_i, t_i, y_i) as

$$\log p_\theta\left(x_i, t_i, y_i|\mathrm{z}_i\right) = \log p_\theta\left(y_i|t_i, \mathrm{z}_i\right) + \log p_\theta\left(t_i|\mathrm{z}_i\right) + \log p_\theta\left(x_i|\mathrm{z}_i\right) \tag{5.8}$$

How can this joint learning framework account for the confounding bias? This can be realized because the posterior of the latent confounders z, $q_\phi(\mathrm{z}|x, t, y)$, depends on both the outcome y and the treatment t. Moreover, the learning of latent confounders z are tailored to good generative models for the outcome y and the treatment t. This joint learning process will hopefully extract information from the observations to learn a good representation of the latent confounder that will account for the confounding bias. Such a philosophy is also discussed in [102] and [193].

5.3.2 Lower Bound Objective

To maximize the log-likelihood of the observed data

$$\ell = \sum_{(x,t,y)\in\mathcal{D}_{obs}} \mathbb{E}[\log p(x,t,y)] \tag{5.9}$$

variational inference minimizes the Kullback-Leibler (KL) divergence from the variational approximation $q_\phi(z|x,t,y)$ to the posterior $p_\theta(z|x,t,y)$, denoted as $KL\left[q_\phi(z|x,t,y)||p_\theta(z|x,t,y)\right]$. This is equivalent to maximizing the *evidence lower bound* (ELBO)

$$\text{ELBO} = \sum_{(x,t,y)\in\mathcal{D}_{obs}} \mathbb{E}_{q_\phi(z|x,t,y)}\left[\log p_\theta(z,x,t,y) - \log q_\phi(z|x,t,y)\right] \tag{5.10}$$

Note that in observational causal effect estimation, the treatment assignment t and corresponding outcome y required for inferring $q_\phi(z|x,t,y)$ are not observed for new test samples. For this reason, we need to take two auxiliary approximation models into consideration in our variational lower bound.

$$t_i \sim q_\phi\left(t|x_i\right), \quad y_i \sim q_\phi\left(y|x_i,t_i\right) \tag{5.11}$$

This is first recognized in [102] and formalized as the following *causal effect lower bound*

$$\mathcal{L}^{\text{CE}} = \text{ELBO} + \sum_{i=1}^{n}\left(\log q_\phi\left(t_i^*|x_i\right) + \log q_\phi\left(y_i^*|x_i,t_i^*\right)\right) \tag{5.12}$$

where $(\mathrm{x}_i, t_i^*, y_i^*)$ are the observed values in the training set. We try to maximize $\mathcal{L}^{\text{CE}}$ to learn the generative parameters θ and the variational parameters ϕ for counterfactual inference via Eq.(5.4).

5.3.3 Inference

Notice that the ELBO in Eq.(5.10) can be written as

$$\text{ELBO} = \sum_{(x,t,y)\in\mathcal{D}_{obs}} \mathbb{E}_{q_\phi(z|x,t,y)} \left[\log p_\theta(x,t,y|z) - \log \frac{q_\phi(z|x,t,y)}{p(z)}\right] \tag{5.13}$$

When we have an explicit representation $q_\phi(z|x,t,y)$ such as the neural network parameterized Gaussian distribution used in VAE [91] and the CEVAE, the ELBO $\mathcal{L}$ can be maximized using the reparameterization trick [91] and stochastic gradient descent. Unfortunately, when we use black-box approximation families, the implicit density $q_\phi(z|x,t,y)$ becomes intractable. In this work, we follow [114] and define the log density ratio (also called *prior contrastive*) $r(z,x,t,y,\phi) = \log \frac{q_\phi(z|x,t,y)}{p(z)}$. Then we have

$$\text{ELBO} = \mathbb{E}_{q_\phi(z|x,t,y)} \left[\log p_\theta(x,t,y|z) - r(z,x,t,y,\phi)\right] \tag{5.14}$$

By introducing the following objective for the discriminator $D(z,x,t,y;\psi)$

$$\max_\psi \mathbb{E}_{q_\phi(z|x,t,y)}[\log \sigma(D(z,x,t,y;\psi))] + \mathbb{E}_{p(z)}[\log(1-\sigma(D(z,x,t,y;\psi)))]$$

where $\sigma(\cdot)$ is the sigmoid activation function, the following proposition indicates that we can obtain the value of the prior contrastive via optimizing the discriminator.

Proposition 5.1. *For fixed generative model $p_\theta(x,t,y|z)$ and inference model $q_\phi(z|x,t,y)$, the optimal discriminator parameter ϕ^* is given by*

$$D\left(z,x,t,y;\psi^*\right) = r(z,x,t,y,\phi) = \log q_\phi(z|x,t,y) - \log p(z) \tag{5.15}$$

Proof. The proof is analogous to the proof of Proposition 1 in [50]. □

As we get the optimal discriminator $D\left(z, x, t, y; \psi^*\right)$, Proposition 5.1 allows us to use it as a proxy of the log density ratio $r(z, x, t, y, \phi)$ and the ELBO can be calculated by

$$\text{ELBO} = \mathbb{E}_{q_\phi(z|x,t,y)}\left[\log p_\theta(x, t, y|z) - D\left(z, x, t, y; \psi^*\right)\right] \tag{5.16}$$

Substitute Eq.(5.16) into Eq.(5.12), we get the causal effect lower bound objective as

$$\begin{aligned}\mathcal{L}^{\text{CE}} = \sum_{i=1}^{n} \{\mathbb{E}_{q_\phi(z_i|x_i,t_i,y_i)}\left[\log p_\theta(x_i, t_i, y_i|z_i) - D(z_i, x_i, t_i, y_i; \psi^*)\right] + \\ \log q_\phi(t_i^*|x_i) \log q_\phi(y_i^*|x_i, t_i^*)\}\end{aligned} \tag{5.17}$$

5.4 Experiments

Evaluating causal inference methods using observational data is always challenging because we do not have access to the ground-truth for the target causal effects. Common evaluation approaches include creating synthetic or semi-synthetic datasets, where real data is modified in a way that allows us to know the true causal effect. In this section, we firstly introduce several metrics and baseline methods used for comparison. Experiment performances on two existing benchmark datasets, IHDP (continuous outcomes) and Jobs (binary outcomes), are then discussed to validate the proposed method. Our experiments are conducted using the TensorFlow [1] platform. The noise distributions $s(\epsilon)$ used in implicit inference networks are standard multivariate Gaussians.

5.4.1 Evaluation Metrics and Baselines

For causal inference evaluation, the absolute error of the ATE estimator, ϵ_{ATE}, is defined as

$$\begin{aligned}\epsilon_{ATE} &= |\widehat{ATE} - ATE| \\ &= \left| \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_1(x_i) - \hat{y}_0(x_i)) - \frac{1}{n}\sum_{i=1}^{n}(Y_1(x_i) - Y_0(x_i)) \right|\end{aligned}$$

where $\hat{y}_t(x_i) = h(x_i, t), t = 0, 1$. Analogously, the absolute error of the ATT estimator, ϵ_{ATT}, is defined as

$$\begin{aligned}\epsilon_{ATT} &= |\widehat{ATT} - ATT| \\ &= \left| \frac{1}{n_1}\sum_{t_i=1}(\hat{y}_1(x_i) - \hat{y}_0(x_i)) - \frac{1}{n_1}\sum_{t_i=1}(Y_1(x_i) - Y_0(x_i)) \right|\end{aligned}$$

where n_1 is the number of units that are in the treatment group.

To evaluate the estimation of ITE, when the underlying ground truth are known, the metric *precision in estimation of heterogeneous effect* (PEHE) [62] is defined in Eq.(5.18). We will report its square root.

$$PEHE = \frac{1}{n}\sum_{i=1}^{n}\left[(\hat{y}_1(x_i) - \hat{y}_0(x_i)) - (Y_1(x_i) - Y_0(x_i))\right]^2 \tag{5.18}$$

When the true ITEs are unknown, we can not calculate PEHE. Alternatively, the policy risk defined in Eq. (5.19) can be used as a proxy to the ITE performance

$$\begin{aligned}R_{pol}(\pi_{\hat{\tau}}) = 1 - \{&p(\pi_{\hat{\tau}}(\mathrm{x}) = 1) \cdot \mathbb{E}\left[Y_1 \mid \pi_{\hat{\tau}}(\mathrm{x}) = 1\right] + \\ &(1 - p(\pi_{\hat{\tau}}(\mathrm{x}) = 1)) \cdot \mathbb{E}\left[Y_0 \mid \pi_{\hat{\tau}}(\mathrm{x}) = 0\right]\}\end{aligned} \tag{5.19}$$

where $\pi_{\hat{\tau}} : \mathcal{X} \to \{0, 1\}$ is an policy induced from an ITE estimator $\hat{\tau}(\cdot)$ with $\pi_{\hat{\tau}}(\mathrm{x}) = 1$ if $\hat{\tau}(\mathrm{x}) > 0$, and $\hat{\tau}(\mathrm{x}) = 0$ otherwise.

Since our method is based on implicit generative models, we call it CEIGM. Baseline methods used for comparison include Ordinary Least Squares (OLS-1,

for continuous outcomes) / Logistic Regression (LR1 for binary outcomes) with treatment as feature, Ordinary Least Squares (OLS-2, for continuous outcomes) / Logistic Regression (LR2 for binary outcomes) with separate regressors for each treatment, k-nearest neighor (k-NN), the double robust method Targeted Maximum Likelihood Estimation (TMLE) [199], Bayesian Additive Regression Trees (BART) estimator [24, 62], Random Forest (Rand. For.) [20, 13], Causal Forest (Caus. For.) [206], Balancing Linear Regression (BLR) and Balancing Neural Network (BNN) by [80], and CEVAE [102]. Following [80] and [102], we report both the within-sample and out-of-sample results.

5.4.2 Semi-simulated Data: IHDP

The benchmark dataset IHDP was first compiled by Hill [62] based in the Infant Health and Development Program (IHDP), which aims at studying the effect of high-quality child care and home visits on future cognitive test scores. The dataset consists of 747 subjects (139 treated and 608 control), each represented by 25 covariates measuring aspects of children and their mothers. We illustrated the 2D projection of covariates using the UMAP projection algorithm [112] and histogram of samples sizes in the treatment and control groups in Fig. 5.2.

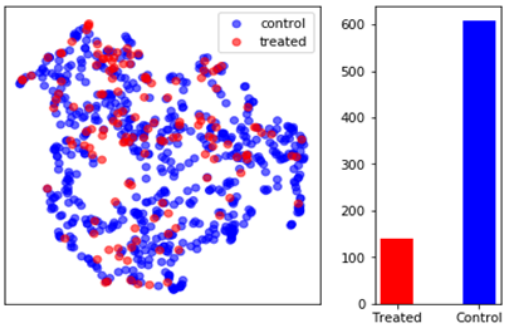


Figure 5.2: Visualization of the 2D projection of the covariates and a histogram of sample sizes by treatment groups for the IHDP dataset.

Table 5.1: Within-sample and out-of-sample results on the IHDP dataset. $\sqrt{\epsilon^{in}_{PEHE}}$ and $\sqrt{\epsilon^{out}_{PEHE}}$ are the in-sample and out-sample squared PEHE errors respectively. ϵ^{in}_{ATE} and ϵ^{out}_{ATE} are the in-sample and out-sample ATE estimation errors respectively. Estimation errors and corresponding empirical standard errors are calculated by replicating the experiments 10 times.

	$\sqrt{\epsilon^{in}_{PEHE}}$	ϵ^{in}_{ATE}	$\sqrt{\epsilon^{out}_{PEHE}}$	ϵ^{out}_{ATE}
OLS1	$5.8 \pm .3$	$.73 \pm .04$	$5.8 \pm .3$	$.94 \pm .06$
OLS2	$2.4 \pm .1$	$.14 \pm .01$	$2.5 \pm .1$	$.31 \pm .02$
BLR	$5.8 \pm .3$	$.72 \pm .04$	$5.8 \pm .3$	$.93 \pm .05$
k-NN	$2.1 \pm .1$	$.14 \pm .01$	$4.1 \pm .2$	$.79 \pm .05$
TMLE	$5.0 \pm .2$	$.30 \pm .01$	–	–
BART	$2.1 \pm .1$	$.23 \pm .01$	$2.3 \pm .1$	$.34 \pm .02$
Rand.For	$4.2 \pm .2$	$.73 \pm .05$	$6.6 \pm .3$	$.96 \pm .06$
Caus.For	$3.8 \pm .2$	$.18 \pm .01$	$3.8 \pm .2$	$.40 \pm .03$
BNN	$2.2 \pm .1$	$.37 \pm .03$	$2.1 \pm .1$	$.42 \pm .03$
CEVAE	$2.7 \pm .1$	$.34 \pm .01$	$2.6 \pm .1$	$.46 \pm .02$
CEIGM	$2.0 \pm .1$	$1.1 \pm .2$	$2.0 \pm .2$	$1.2 \pm .2$

For the sake of comparison, we follow [102] and use the noiseless outcome to compute the true effects. The results are presented in Table 5.1. The results shows that the proposed CEIGM method gets the lowest within-sample and out-of-sample PEHE errors. This indicates CEIGM fits both response surfaces $\mathbb{E}[Y_0|x]$ and $\mathbb{E}[Y_1|x]$ quite well. Unfortunately, CEIGM gets the highest errors for estimating the ATE. This is beyond our expectation. One possible reason is that, though the two response surfaces are well fitted, they differ from the underlying true response surfaces in opposite directions. For example, the fitted potential outcomes for the control $\mathbb{E}[Y_0|x]$ tend to be smaller than the true control outcomes, while the fitted potential outcomes for the treated $\mathbb{E}[Y_1|x]$ tend to be larger than the true treated outcome. As a result, even though both of them have small errors, the average of their difference may induce a relatively large error.

5.4.3 Real World Data: Jobs

We also validate the proposed CEIGM method using the real-world Jobs dataset, which combines a randomized study $\mathcal{R}$ based on the National Supported Work (NSW) program with observational data $\mathcal{O}$ to form a larger dataset. For more details of the data, refer [1]. The 2D projection of the features and histogram of the samples sizes for two groups are illustrated in Fig.5.3.

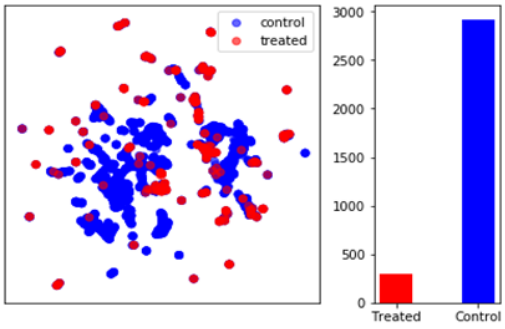


Figure 5.3: Visualization of the 2D projection of the covariates and a histogram of sample sizes by treatment groups for the Jobs dataset.

Instead of the ATE, the NSW program aims at estimating the effect of job training on employment after training, i.e., the true average treatment effect on the treated (ATT). Since all the treated individuals come from the randomized study $\mathcal{R}$, we can easily estimate ATT by

$$\begin{aligned} ATT &:= \frac{1}{|\mathcal{T}_1|} \sum_{i \in \mathcal{T}_1} (Y_1(x_i) - Y_0(x_i)) \\ &= \frac{1}{|\mathcal{T}_1|} \sum_{i \in \mathcal{T}_1} y_i - \frac{1}{|\mathcal{T}_0 \cap \mathcal{R}|} \sum_{i \in \mathcal{T}_0 \cap \mathcal{R}} y_i \end{aligned}$$

where $\mathcal{T}_1$ and $\mathcal{T}_0$ are the treated and control group in the full dataset. Following [178] and [102], we use the NSW experimental sample (297 treated and 425 control) and the PSID comparison group (2490 control) and report the $\epsilon_{ATT} =$

[1] http://users.nber.org/ rdehejia/data/nswdata2.html

$|\widehat{ATT} - ATT|$. For evaluating ITE estimation, we use the policy risk R_{pol}. The results are list in Table 5.2.

From the result in Table 5.2, we can conclude that our proposed CEIGM method achieves lower out-of-sample ATT error and policy risk than most of the benchmarks. Specifically, CEIGM gets the second smallest values for both the out-of-sample policy risk and ATT error. Compared with CEVAE, our proposed CEIGM method has lower out-of-sample policy risk but higher ATT error. The result validate again that the proposed CEIGM method is able to learn better potential outcome functions because the implicit posteriors are theoretically able to approximate arbitrarily complex distributions.

Table 5.2: Within-sample and out-of-sample results on the IHDP dataset. R_{pol}^{in} and R_{pol}^{out} are the in-sample and out-sample policy risk respectively. ϵ_{ATT}^{in} and ϵ_{ATT}^{out} are the in-sample and out-sample ATT estimation errors respectively. Estimation errors and corresponding empirical standard errors are calculated by replicating the experiments 10 times.

	R_{pol}^{in}	R_{pol}^{out}	ϵ_{ATT}^{in}	ϵ_{ATT}^{out}
LR1	$.22 \pm .0$	$.01 \pm .00$	$.23 \pm .0$	$.08 \pm .04$
LR2	$.21 \pm .0$	$.01 \pm .01$	$.24 \pm .0$	$.08 \pm .03$
BLR	$.22 \pm .0$	$.01 \pm .01$	$.25 \pm .0$	$.08 \pm .03$
k-NN	$.02 \pm .0$	$.21 \pm .01$	$.26 \pm .0$	$.13 \pm .05$
TMLE	$.22 \pm .0$	$.02 \pm .01$	–	–
BART	$.23 \pm .0$	$.02 \pm .00$	$.25 \pm .0$	$.08 \pm .03$
Rand.For	$.23 \pm .0$	$.03 \pm .01$	$.28 \pm .0$	$.09 \pm .04$
Caus.For	$.19 \pm .0$	$.03 \pm .01$	$.20 \pm .0$	$.07 \pm .03$
BNN	$.20 \pm .0$	$.04 \pm .01$	$.24 \pm .0$	$.09 \pm .04$
CEVAE	$.15 \pm .0$	$.02 \pm .01$	$.26 \pm .0$	$.03 \pm .01$
CEIGM	$.22 \pm .0$	$.02 \pm .00$	$.23 \pm .0$	$.05 \pm .01$

5.5 Summary

In this chapter, we model the causal mechanisms in a causal model by IGMs, which are proved universal approximators for the underlying causal mechanisms. The proposed CEIGM method is a generalization of the CEVAE method proposed in [102]. Specifically, we generalize the Gaussian inference model of latent confounders used in CEVAE to general black box inference models parametrized by deep neural networks. To tackle the intractability of implicit inference model, we adopt an adversary training scheme using a discriminator to learn the parameters.

We validate the proposed method via experiments on two benchmark datasets. Results of both experiments indicate that the proposed method tend to learn better potential outcome functions with opposite error directions, leading to better ITE estimation but worse ATE/ATT estimation. This issue is out of our expectation and we leave it as future investigation. We also notice that recent research [70, 180] on implicit model inference indicate that discriminator-based adversary training may lead to noisy gradients and thus unstable results. In future work, more implicit variational inference algorithms will be investigated to realize methods that are more robust.

Chapter 6

Direct Treatment Effect Estimation using Deep Neural Networks

In this chapter, we propose the idea of direct treatment effect estimation, which parametrizes the target treatment effect function with DNNs and learn it via gradient-based optimization directly without a detour of learning the treatment response function or the treatment assignment mechanism. This idea is pretty intuitive and motivated by policy gradient methods for policy optimization from the reinforcement learning literature [190]. Unlike other value-based algorithms, e.g., Q-learning, that learn an optimal policy indirectly by estimating the state-action function (i.e., the Q function) first, policy gradient methods parametrize the target policy directly and learn it using gradient-based optimization. We also note that several algorithms have recently been proposed to model the individual treatment effect (ITE) function directly. Specifically, Wager et al. [206] proposed to directly estimate the ITE non-parametrically using random forests with an ad-hoc leaf splitting criteria. By this causal forest method, individuals in each leaf can be regarded as randomly assigned as an RCT. However, tree-based methods need manual feature engineering which are not as automatic

as our DNN-based method. We will also compare it with our proposed models in the experiment section empirically. Nie et al. [126] also parametrize the target treatment effect function directly and propose to learn it using their proposed R-loss objective function. However, the proposed R-learner still learns the parametrized treatment effect function in a two-step manner by first estimating two auxiliary functions, the propensity score function and the mean outcome function.

Our main contributions are: (1) We introduce the idea of direct treatment effect estimation that learns the target treatment effect function directly from observational data; (2) We propose a novel Causal Effect Network (CENet) model for direct treatment effect estimation using DNNs. The proposed CENet learns the target treatment effect function and two auxiliary treatment response functions jointly. (3) We further combine the idea of direct treatment effect estimation and balanced representation learning to propose the Balanced Causal Effect Neural Network (BCENet) model; (4) We validate the proposed methods with comprehensive experiments on synthetic, semi-simulated and real world datasets. Experiment results suggest that our proposed models generally have better or competitive performance than existing state-of-art models. Moreover, estimations of our direct estimation models are generally more stable than the other models since they are estimated directly in an end-to-end manner rather than indirectly by a two-stage process.

The remainder of this chapter is organized as follows: In Section 6.1, we introduce definitions, notations and formalize the causal inference problem. In Section 6.2, we give a brief review of related work. As the core section, Section 6.3 introduces the direct treatment effect estimation idea and two neural network architectures for direct treatment effect estimation. Experiments on simulated data and an application on real health data are conducted in Section 6.4. Section

6.5 summarizes this chapter.

6.1 Problem Setup

The causal inference literature stresses the importance of defining the causal estimands of interest (or the *target parameter*) first and thinking carefully about necessary assumptions for identification. In this section, we demonstrate the problem of causal inference from observational data via an illustrative example, introduce preliminary definitions and assumptions for identifiability.

6.1.1 Treatment Effect Estimation: An Illustrative Example

Consider a number of $n = 400$ individuals with a scaler covariate $\mathrm{x}_i \sim U(-2, 2)$ for $i = 1, \ldots, n$. For each individual x, suppose there is a binary treatment with $t = 1$ indicating treated and $t = 0$ not treated (i.e., control). The target outcome after treated/ control is denoted as y. For an individual with covariate value x, denote the underlying treatment response functions if she is assigned into the treated group and the control group as $\mu_1(\mathrm{x})$ and $\mu_0(\mathrm{x})$ respectively. In causal inference, we are interested in the treatment effect of the treatment t on the outcome, which is defined as the expected difference between the two potential treatment responses, i.e., $\tau(\mathrm{x}) = \mu_1(\mathrm{x}) - \mu_0(\mathrm{x})$.

This is called the ITE or CATE in the causal inference literature [74], and is intrinsically important in settings where we want to evaluate the efficiency of some policy and make personalized recommendations. Suppose in the observational data, we observed $n_1 = 150$ individuals got treated and the other $n_0 = 250$ individuals not treated. The treatment assignment mechanism that allocated the observed treatment to each individual is not random and depends

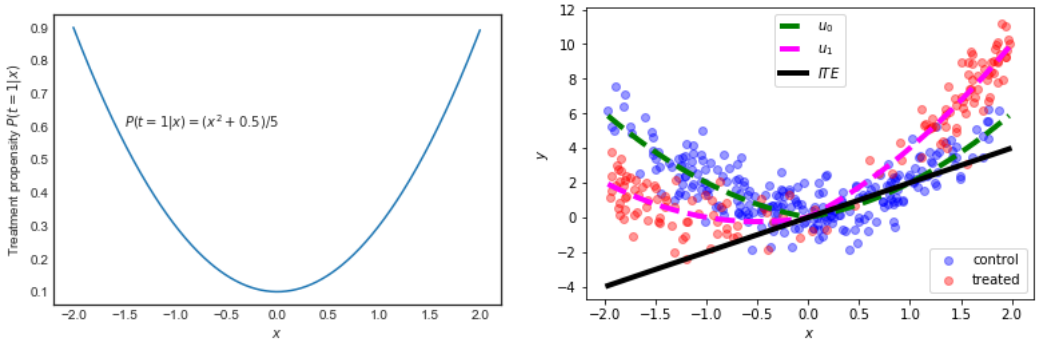


Figure 6.1: Illustration of (a) the treatment propensity function and (b) the observational data. There are $n = 400$ individuals in (b), $n_1 = 150$ of them are treated and the other $n_0 = 250$ are assigned into the control group. The x-axis is the feature x and y-axis is the observed outcome y. The dashed green and magenta curves are respectively the treatment responses under control and treated. The black line is the target treatment effect function.

on the covariate value x via

$$t \sim \text{Bern}\left(\frac{x^2 + 0.5}{5}\right)$$

We illustrate the underlying treatment propensity function, i.e., the probability of an individual is assigned into the treatment group $P(t = 1 \mid \text{x})$ in Fig.6.1(a). Obviously, individuals with value x far away from 0 is more likely to be treated, and individuals with value x in a *normal* range near 0 are more likely be assigned into the control group. This kind of imbalanced treatment selection preference is called *selection bias* in the causal inference community [74].

Moreover, the two underlying treatment response functions are $\mu_0(\text{x}) = \text{x}^2 + |\text{x}|$ and $\mu_1(\text{x}) = \text{x}^2 + |\text{x}| + 2\text{x}$. Then the target treatment effect function is

$$\tau(\text{x}) = \mu_1(\text{x}) - \mu_0(\text{x}) = 2\text{x}$$

However, in real world observational data, we do not have access to the treat-

ment assignment mechanism or the underlying treatment response functions. For an individual x_i, we only know her assigned treatment t_i and the corresponding observed outcome $y_i = y_i(t_i)$, while the potential outcome for the alternative treatment option $y_i(1 - t_i)$ can never been observed in principle. The observed data for this illustrative example is depicted in Fig.6.1(b). We also draw the unobserved treatment response curves, $\mu_0(\mathrm{x})$ and $\mu_1(\mathrm{x})$, as well as the treatment effect function $\tau(\mathrm{x})$ in the figure.

With the $n = 400$ observed examples, $\mathcal{D} = \{(\mathrm{x}_i, t_i, y_i), i = 1, \ldots, n\}$, we want to learn the treatment effect function from the observational data, which then can be utilized to answer the following questions:

- **Question (1)**: What is the overall effect of the treatment t over the whole population?
- **Question (2)**: How efficient is the treatment t over the $n_1 = 150$ treated individuals?
- **Question (3)**: For a new individual with covariate value x, should we assign her into the treated or the control group?

Note that, since real world treatment assignment mechanisms are usually not randomized and the probability of an individual to be treated depends on her own covariate values, the collected data in real world observational studies are imbalanced (as illustrated in Fig.6.1(b)). Moreover, the fundamental issue that we can only observed the outcome corresponding to the assigned treatment renders treatment effect estimation from observational data generally impossible. In the next section, we formally introduce definitions and assumptions needed for solving this problem.

6.1.2 Definition and Assumptions

Consider an observational study consisting of n observations $\mathcal{D} = \{(\mathrm{x}_i, t_i, y_i), i = 1, \ldots, n\}$ of the variables (X, T, Y) drawn i.i.d. from some underlying distribution such that for each i, $\mathrm{x}_i \in \mathcal{X}$ denotes the baseline pre-treatment covariates, $t_i \in \mathcal{T}$ the assigned treatment, and $y_i \in \mathcal{Y}$ the observed outcome. Take job training as an example, for an employee with covariate $\mathrm{x} \in \mathcal{X}$, the set of treatments $\mathcal{T}$ might be whether she joined a specific job training program, and the set of outcomes might be $\mathcal{Y} = [0, 10K]$ indicating her monthly salary in dollars. In this work, we only consider the binary treatment case, i.e., $\mathcal{T} = \{0, 1\}$. Denote the treated group as $\mathcal{T}_1 = \{i : t_i = 1\}$ and the control group as $\mathcal{T}_0 = \{i : t_i = 0\}$. For an individual i, let $Y_i(t) \in \mathcal{Y}$ be her potential outcome under the treatment option t. The fundamental problem of causal inference is that only one of the two potential outcomes, $Y_i(0)$ and $Y_i(1)$, can be observed for a given individual, i.e., $y_i = t_i Y_i(1) + (1 - t_i) Y_i(0)$. In the machine learning literature, this kind of partial feedback is called *bandit feedback* [192, 191].

Two key functions mentioned in the last section are the treatment response functions $\mu_0(\mathrm{x})$ and $\mu_1(\mathrm{x})$. In the language of Pearl's do-calculus [131], they are defined as

$$\begin{aligned} \mu_t(\mathrm{x}) &= \mathbb{E}[Y_i | X_i = \mathrm{x}, do(T_i = t)] \\ &= \mathbb{E}\left[Y_i(t) | X_i = \mathrm{x}\right], \qquad t = 0, 1 \end{aligned} \tag{6.1}$$

where $do\left(T_i = t\right)$ is the *do*-operator meaning to "*set*" the treatment as t rather than "*seeing*" the treatment t. By this definition, the difference between the two treatment responses is defined as the CATE, which measures the covariate-specific treatment effect

$$\begin{aligned} \tau(\mathrm{x}) &= \mu_1(\mathrm{x}) - \mu_0(\mathrm{x}) \\ &= \mathbb{E}\left[Y_i(1) | X_i = \mathrm{x}\right] - \mathbb{E}\left[Y_i(0) | X_i = \mathrm{x}\right] \end{aligned} \tag{6.2}$$

This is a fundamental component in the causal inference literature. We can use it to answer Question (1) in the last section by estimating the ATE via $ATE = \mathbb{E}\left[\tau(\mathrm{x}_i)\right]$ and Question (2) by the ATT via $ATT = \mathbb{E}\left[\tau\left(\mathrm{x}_i\right)|t_i = 1\right]$. Question (3) can be answered by denoting a treatment policy that depends on the pre-treatment covariates x as $\pi\left(\mathrm{x}_i\right) = P\left(t_i = 1|\mathrm{x}_i\right)$, then $\tau(\mathrm{x})$ is a sufficient statistics for evaluating an existing policy $\pi : \mathcal{X} \to \mathcal{T}$ and for optimizing any treatment policy since we can re-write the policy optimization objective as the difference to the always-control policy $\pi_0(\mathrm{x}) \equiv 0$, i.e.[127],

$$\begin{aligned}\pi^* &\in \underset{\pi \in \mathcal{X} \to \mathcal{T}}{\operatorname{argmin}} \mathbb{E}[(\mu_0(\mathrm{x}_i) + \pi(\mathrm{x}_i)\tau(\mathrm{x}_i)) - \mu_0\left(\mathrm{x}_i\right)] \\ &= \underset{\pi \in \mathcal{X} \to \mathcal{T}}{\operatorname{argmin}} \mathbb{E}\left[\pi(\mathrm{x}_i)\tau(\mathrm{x}_i)\right]\end{aligned}$$

Despite of its importance, treatment effect estimation from observational data is fundamental impossible without causal assumptions since we can never observe both treatment responses for any individual. For identifiability, besides Consistency (Assumption 2.1) and SUTVA (Assumption 2.2), we also make the following *Ignorability* assumption common in the causal inference literature [74].

Assumption 6.1 (Ignorability)**.** *For each individual X_i, the potential outcome variables $Y_i(t), t \in \mathcal{T}$ are statistically independent of the treatment actually taken. That is, $Y_i(t) \perp\!\!\!\perp T_i \mid X_i$ for all $i = 1, 2, \ldots, n$.*

This assumption means that there exist no unobserved confounders. It is generally uncheckable from data only and must be determined by domain knowledge. Under these assumptions, we provide a self-contained proof of the identifiability of various treatment effects in the Appendix A.2. In practice, we also make the following *Positivity* assumption to guarantee enough randomness in the data-generating process so that unobserved counterfactuals can be estimated from the observed data.

Assumption 6.2 (Positivity, or Common Support)**.** *The treatment propensity is positive for any covariate* $\mathrm{x} \in \mathcal{X}$ *i.e.,* $0 < P(t = 1 \mid \mathrm{x}) < 1$.

6.2 Preliminaries

In this section, we introduce two groups of methods that are related to our proposed method: treatment response modelling for treatment effect estimation and DNNs for treatment effect estimation.

6.2.1 Treatment Effect Estimation via Response Modelling

Note that the ITE of an individual $X_i = \mathrm{x}$ is defined as the difference of the outcome if treated versus that if untreated. An intuitive method would be to use any supervised learning method (e.g., linear regression, random forest, neural networks) to fit the two treatment response functions $\mu_0(\mathrm{x})$ and $\mu_1(\mathrm{x})$. Then for an individual x_i in the training data (also called in-sample individual), the CATE is estimated by

$$\hat{\tau}\left(\mathrm{x}_i\right) = \begin{cases} y_i - \mu_0\left(\mathrm{x}_i\right), & t_i = 1 \\ \mu_1\left(\mathrm{x}_i\right) - y_i, & t_i = 0 \end{cases} \tag{6.3}$$

For an out-sample individual with covariates x, the CATE is estimated via

$$\hat{\tau}(\mathrm{x}) = \mu_1(\mathrm{x}) - \mu_0(\mathrm{x}) \tag{6.4}$$

Therefore, $\tau(\mathrm{x})$ is estimated indirectly by a two-stage procedure, and the key is to obtain a good estimate of the underlying conditional mean functions $\mu_0(\mathrm{x})$ and $\mu_1(\mathrm{x})$. This is called simulated twins, G-computation [146, 147, 148], outcome regression or counterfactual inference in the literature [61, 225, 170].

In practice, this is realized by regressing the observed outcomes on the co-

variates in the control group $\mathcal{T}_0$ to fit the control response function μ_0 and in the treated group $\mathcal{T}_1$ to fit μ_1. Künzel et al. [95] call this approach T-learning (T for "two models" or "twins"). In T-learning, we treat the treatment indicator $t \in \{0, 1\}$ as a function indicator and learn separate treatment response models for each treatment. This is favourable when the two treatment response functions $\mu_0(\mathrm{x})$ and $\mu_1(\mathrm{x})$ have drastically different properties (e.g., different explanatory covariates and different interactions among these covariates). Alternatively, we can regard t as just another covariate and define treatment responses under different treatments as a single conditional mean function, $\mu(\mathrm{x}, t) = t \cdot \mu_1(\mathrm{x}) + (1 - t) \cdot \mu_0(\mathrm{x})$. CATE is then derived by replacing $\mu_0(\mathrm{x})$ and $\mu_1(\mathrm{x})$ with $\mu(\mathrm{x}, 0)$ and $\mu(\mathrm{x}, 1)$ respectively in the above estimators (6.3) and (6.4). This is called S-learning (S for "single model") in [95]. Besides T-learning and S-learning, Künzel et al. [95] also proposed X-learning that estimates the two treatment response functions $\mu_0(\mathrm{x})$ and $\mu_1(\mathrm{x})$ as in T-learning first, and then impute the estimated ITEs for individuals in the treated group $\mathcal{T}_1$ using $\hat{\tau}(\mathrm{x}_i) = y_i - \mu_0(\mathrm{x}_i)$ and individuals in $\mathcal{T}_0$ using $\hat{\tau}(\mathrm{x}_i) = \mu_1(\mathrm{x}_i) - y_i$. Lastly, X-learning learns the target treatment effect function $\tau(\mathrm{x})$ with the imputed ITEs in a supervised manner.

As a learning problem, treatment effect estimation via treatment response modelling is different from classic learning in that the explicit label – the true ITEs – can never been observed. For each individual, we can only observe her response to the treatment she actually received. This kind of *bandit feedback* is called *the fundamental problem of causal inference* [155]. As we have discussed in Chapter 2.3.2, treatment response modelling is originally designed for answering counterfactual questions and fulfils the task of treatment effect estimation in an indirect manner. As a result, treatment response functions fitted to minimize the prediction error for the observed outcomes are not guaranteed to produce

accurate treatment effect estimation.

6.2.2 DNNs for Treatment Effect Estimation

DNNs [49] have proved to be successful in a plenty of machine learning applications due to their automatic feature engineering ability and universal approximation property. Compared with other machine learning techniques, DNNs are able to extract complex nonlinear features from raw inputs so that minimizing the need for manual feature engineering, and can be trained continually in an end-to-end manner. This allows us to train a single neural network targeting several distinct objectives and allows multiple networks to be co-trained on the same set of data while keep them coupled with shared layers. By virtue of the powerful representation learning ability of DNNs, Johansson et al. [80] proposed the BNN network, which learns a single treatment response function $\mu(\mathrm{x}, t) = h(\phi(\mathrm{x}), t)$ for treatment effect estimation. The covariate imbalance between the two treatment groups are handled by learning a shared representation $\phi(\mathrm{x})$ of the pre-treatment covariates. Later in [178], they argued that such a S-learning method may lose the influence of the scaler treatment indicator t on the shared high-dimensional representation during training. To avoid this issue, they proposed two neural networks, TARNet and CFR, to learn two separate outcome models $h_0(\phi(\mathrm{x})$ and $h_1(\phi(\mathrm{x})$ on top of the shared representation layers $\phi(\mathrm{x})$. In these models, the treated and control groups are able to share information in the process of learning the two treatment response functions. Moreover, to guide the learning of the shared representation layers so as to realize the goal of covariate balance in the representation space, BNN and CFR add a balancing constraint on the shared representation using integral probability metrics (IPMs) [186].

Besides deep representation learning for balancing, researchers have also

adapted recent advances in deep generative models and GAN training for the task of treatment effect estimation. Specifically, The CEVAE method [102] uses variational auto-encoders to model and infer the outcome generative probability in a latent-variable modelling framework. The CEIGM model [225] uses more general implicit generative models to model the underlying data-generating process. The GANITE method [213] uses GANs to learn the counterfactual and ITE generators. Other researchers have adopted DNNs into the instrumental variable framework [57] and the multi-task learning framework [2] for treatment effect estimation.

In general, balanced representation learning methods realize the goal of treatment effect estimation via treatment response modelling. Methods based on deep generative models such as CEVAE and CEIGM are from a probabilistic machine learning perspective and rely on different assumptions of the underlying data-generating mechanism to ours. Building on the idea of X-learning, Stadie et al. [187] proposed a DNN architecture called Y-learner for treatment effect estimation. The Y-learner is also a direct treatment effect estimation method which learns the treatment response functions and treatment effect function jointly. However, as we will describe in the next section, our method goes further than the Y-learner by incorporating the idea of balanced representation learning in the learning of the auxiliary treatment response functions. We argue that this should be beneficial for the target treatment effect estimation. However, since the source code of the Y-learner is not available yet, we are unable to compare it with our proposed methods.

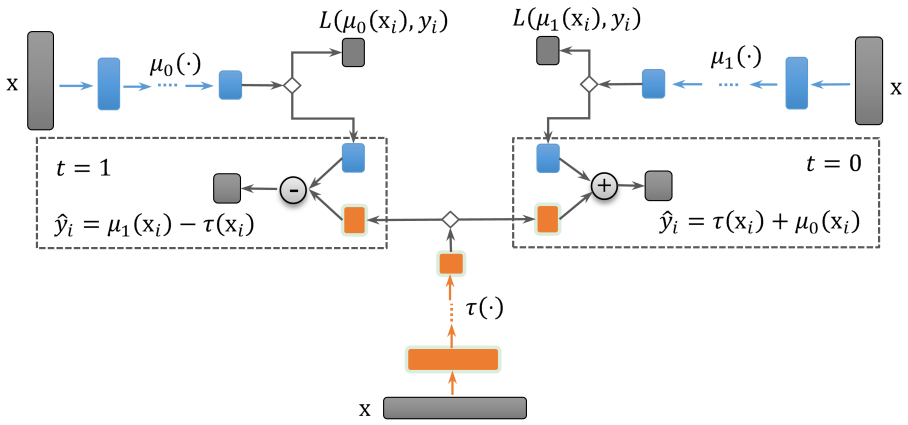


Figure 6.2: The joint neural networks for direct treatment effect estimation. Blue and orange nodes are hidden layers or estimated values. Gray nodes corresponds to observed values. Hollow diamonds are decision nodes which decide the input is a *treated* sample or a *control* sample according to the corresponding t.

6.3 Direct Treatment Effect Estimation Using DNNs

In this section, we first introduce the idea of direct treatment effect estimation by learning the treatment effect function directly. Two DNN-based models, CENet and BCENet, for direct treatment effect estimation are then described in details.

6.3.1 Direct Treatment Effect Estimation

Since our target of interest is the CATE, we may be better off modelling it directly without a first stage estimation of other functions. To motivate and have a better understanding of such a direct estimation process, suppose we have access to an oracle of the true ITE $\tau_i^* = \tau^*(\mathrm{x}_i)$ for each individual x_i, then we can cast the problem of treatment effect estimation into a supervised learning problem: Given a class $\mathcal{C}$ (usually interpretable like linear models or tree-based models) and a set of observational data $\mathcal{D}^* = \{(\mathrm{x}_i, \tau_i^*), i = 1, \ldots, n\}$,

learn the best-in-class CATE function with the following minimum mean square error (MSE)

$$\min_{\tau \in \mathcal{C}} \frac{1}{n} \sum_{i=1}^{n} (\tau(\mathrm{x}_i) - \tau^*(\mathrm{x}_i))^2$$

Moreover, if we have some prior knowledge of the treatment effect function class $\mathcal{C}$, we can learn it by the following regularized objective

$$\min_{\tau \in \mathcal{C}} \frac{1}{n} \sum_{i=1}^{n} (\tau(x_i) - \tau^*(\mathrm{x}_i))^2 + \Omega(\tau)$$

where $\Omega(\tau)$ is the regularization term represents possible prior knowledge of $\tau(\cdot)$. Though the idea to modelling the target CATE function directly via machine learning is appealing, the problem is that we never observe τ_i^* directly but only the bandit feedback data $\mathcal{D} = \{(\mathrm{x}_i, t_i, y_i), i = 1, \ldots, n\}$. To detour this problem and still to learn the CATE function directly, we need to find conditions that relates $\tau(\mathrm{x})$ with other quantities that can be estimated from the observational data.

Based on the analogy between the *covariates-treatment-outcome* data structure (x_i, t_i, y_i) in observational causal inference with the *state-action-reward* bandit feedback data structure (s, a, r) in reinforcement learning, the idea of direct treatment effect estimation, which parametrizes and learns the target CATE function $\tau(\mathrm{x})$ directly, is motivated by deep policy gradient methods that parametrize the target policy by DNNs and optimize it using gradient-based optimization in the reinforcement learning literature [190]. In direct treatment effect estimation, we model the treatment response functions and the target CATE function with DNNs and couple them via their relationship formulated in (6.2). The computation process of the joint neural networks is depicted in Fig.6.2.

The joint neural network architecture comprises the target CATE network

and two auxiliary outcome prediction networks. In this architecture, our target is the CATE function $\tau(\cdot;\theta)$. In order to update its parameters, we bridge it with the observed outcome via two treatment outcome networks $\mu_0(\mathrm{x};\beta_0)$ and $\mu_1(\mathrm{x};\beta_1)$. On one hand, $\mu_0(\mathrm{x};\beta_0)$ and $\mu_1(\mathrm{x};\beta_1)$ are parametrized by separate neural networks, which are able to model complex treatment responses as T-learning. On the other hand, they are coupled via the target CATE network $\tau(\cdot;\theta)$ and can be trained efficiently using the whole observational dataset. To optimize the neural network parameters, we use the standard back propagation based training paradigm. Note that in the joint neural network architecture, the two auxiliary treatment response functions $\tau_0(\mathrm{x})$ and $\tau_1(\mathrm{x})$ still need to be learned in the training stage, but they are only intermediate estimands used to bridge the unknown ITE with the observed outcome for each individual so as to provide guidance for optimizing the target treatment effect function in the training stage. We do not need them for out-sample predictions. In the following sections, we introduce two practical models for direct treatment effect estimation.

6.3.2 CENet: Causal Effect Neural Network

Our first direct treatment effect estimation model, CENet, consists of two components: the outcome prediction component and the CATE component.

The Outcome Prediction Component

Note that the observational dataset $\mathcal{D} = \{(\mathrm{x}_i, t_i, y_i), i = 1, \ldots, n\}$ can be divided into the treated subset $\mathcal{D}_1 = \{(\mathrm{x}_i, y_i), i \in \mathcal{T}_1\}$ and the control subset $\mathcal{D}_0 = \{(\mathrm{x}_i, y_i), i \in \mathcal{T}_0\}$. A naive method for learning the two treatment response functions $\mu_0(\mathrm{x};\beta_0)$ and $\mu_1(\mathrm{x};\beta_1)$ is to minimize the following loss functions over

$\mathcal{D}_0$ and $\mathcal{D}_1$ respectively,

$$\begin{aligned}\mathcal{L}^{\mu_0} &= \frac{1}{n_0}\sum_{i\in\mathcal{T}_0} L\left(\mu_0\left(\mathrm{x}_i\right), y_i\right)\\ \mathcal{L}^{\mu_1} &= \frac{1}{n_1}\sum_{i\in\mathcal{T}_1} L\left(\mu_1\left(\mathrm{x}_i\right), y_i\right)\end{aligned} \tag{6.5}$$

Possible loss functions are the L_2 loss $L(\mu(\mathrm{x}_i), y_i) = (\mu(\mathrm{x}_i) - y_i)^2$ for continuous outcomes and the log-loss $L\left(\mu\left(\mathrm{x}_i\right), y_i\right) = -y_i \log \mu\left(\mathrm{x}_i\right) - (1-y_i)\log\left(1-\mu\left(\mathrm{x}_i\right)\right)$ for binary outcomes. This naive method is actually the main idea of T-learning [95], which is statistical inefficiency and biased. That is, the estimated control outcome function $\mu_0(\mathrm{x};\beta_0)$ trained solely over the control subset $\mathcal{D}_0$ will not generalize well to the treated subset $\mathcal{D}_1$ and the treated outcome function $\mu_1(\mathrm{x};\beta_1)$ trained over $\mathcal{D}_1$ will not generalize well to $\mathcal{D}_0$.

In Section 6.3.2, we will introduce the joint learning process that, besides factual outcome prediction errors, the counterfactual prediction error for individuals in subset $\mathcal{D}_1$ and counterfactual predictions for individuals in subset $\mathcal{D}_0$ also guide the learning of $\mu_0(\mathrm{x};\beta_0)$ and $\mu_1(\mathrm{x};\beta_1)$. By this training scheme, we can use all observations in $\mathcal{D} = \{(\mathrm{x}_i, t_i, y_i), i = 1, \ldots, n\}$ to train both $\mu_0(\mathrm{x};\beta_0)$ and $\mu_1(\mathrm{x};\beta_1)$ simultaneously and the training process is targeted to the CATE function. Obviously, this is more flexible than S-learning and more data efficient than T-learning.

The CATE Component

The target CATE function, $\tau : \mathcal{X} \to \mathbb{R}$, is a mapping from an observation $\mathrm{x} \in \mathcal{X}$ to a real-valued treatment effect. Suppose we parameterize this function with a neural network $\tau(\cdot;\theta)$, to optimize the neural network parameters θ using the observed data $\mathcal{D}$, we need to bridge $\tau(\mathrm{x}_i;\theta)$ with the observed outcome y_i for each individual x_i. To understand our derivation, first let us assume we have

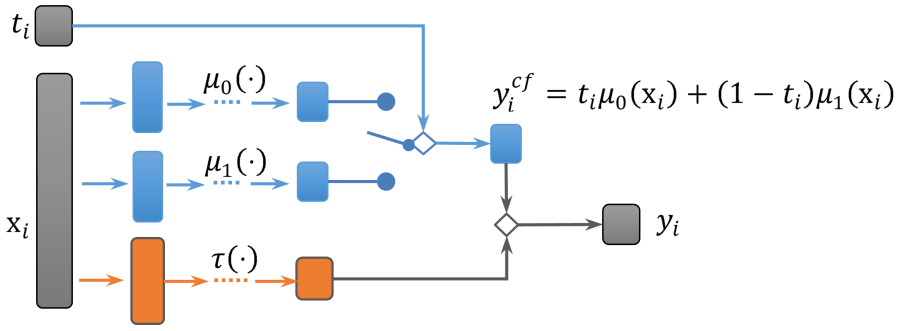


Figure 6.3: Neural network architecture of CENet. The diamond toggle is switched according to the observed treatment for each individual to obtain the unobserved counterfactual outcome.

an oracle the counterfactual outcome y_i^{cf} for each individual x_i, several lines of algebra then imply that

$$\mathbb{E}\left[y_i\right] = \left(2t_i - 1\right)\tau(\mathrm{x}_i) + \mathbb{E}\left[y_i^{cf}\right] \tag{6.6}$$

This then permits us to learn the target CATE function $\tau(\mathrm{x};\theta)$ in a supervised learning manner. However, we do not have access to such an oracle. We replace it using the following formula

$$\mathbb{E}\left[y_i^{cf}\right] = t_i\mu_0(\mathrm{x}_i) + (1 - t_i)\mu_1(\mathrm{x}_i) \tag{6.7}$$

where $\mu_0(\mathrm{x}_i)$ and $\mu_1(\mathrm{x}_i)$ are the expected treatment responses for x_i predicted using the outcome prediction networks introduced in the last section. Overall, this supervised learning process is formulated as

$$\mathbb{E}\left[y_i\right] = \left(2t_i - 1\right)\tau\left(\mathrm{x}_i\right) + t_i\mu_0\left(\mathrm{x}_i\right) + \left(1 - t_i\right)\mu_1\left(\mathrm{x}_i\right) \tag{6.8}$$

Thus we can obtain the loss function for learning the CATE function as

Algorithm 6.1 Learning Process for CENet

Input: Observation data $\mathcal{D} = \{(\mathrm{x}_i, t_i, y_i), \ldots, (\mathrm{x}_n, t_n, y_n)\}$, hyper-parameters $\gamma, \lambda > 0$, training batch size B, number of epochs K, and learning rate η

Output: The learned parameters $(\theta, \beta_0, \beta_1)$

1: Initialize parameters $(\theta, \beta_0, \beta_1)$ for the CATE network $\tau(\cdot;\theta)$ and outcome prediction networks $\mu_0(\cdot;\beta_0)$, $\mu_1(\cdot;\beta_1)$;
2: Split $\mathcal{D}$ into training and validation sets $\mathcal{D}_{train}$ and $\mathcal{D}_{valid}$;
3: **for** $k = 1, 2, \ldots, K$ **do**
4: Sample $\mathcal{D}_{batch} = \left\{\left(\mathrm{x}_1^k, t_1^k, y_1^k\right), \ldots, \left(\mathrm{x}_B^k, t_B^k, y_B^k\right)\right\}$ from $\mathcal{D}_{train}$
5: Update $(\theta, \beta_0, \beta_1)$ to minimize Eq.(6.10) on $\mathcal{D}_{batch}$ via

$$\begin{aligned}\theta &\leftarrow \theta - \eta \nabla_\theta \mathcal{L}_B^{CENet} \\ \beta_0 &\leftarrow \beta_0 - \eta \nabla_{\beta_0} \mathcal{L}_B^{CENet} \\ \beta_1 &\leftarrow \beta_1 - \eta \nabla_{\beta_1} \mathcal{L}_B^{CENet}\end{aligned}$$

6: Test convergence using $\mathcal{D}_{valid}$, **if** converge
7: **break**
8: **end for**

$$\mathcal{L}^\tau = \frac{1}{n} \sum_{i=1}^{n} \left\{(2t_i - 1)\tau(\mathrm{x}_i) + t_i \mu_0(\mathrm{x}_i) + (1 - t_i)\mu_1(\mathrm{x}_i) - y_i\right\}^2 \tag{6.9}$$

The Objective Function

By combining the learning objectives in (6.5) and (6.9) together and adding a model complexity regularization term, we obtain the following joint loss function for the CENet model:

$$\mathcal{L}_n^{CENet} = \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}_i^\tau + \gamma_0 \mathcal{L}^{\mu_0} + \gamma_1 \mathcal{L}^{\mu_1} + \lambda \Omega(\tau) \tag{6.10}$$

where $\gamma_0, \gamma_1, \lambda > 0$ are hyper-parameters. Normally, the two treatment groups have different sample sizes, we use the following parameter configuration to com-

pensate for this difference

$$\gamma_1 = \gamma, \quad \gamma_0 = \frac{p}{1-p}\gamma$$

where $p \triangleq p(t=1) = \frac{1}{n}\sum_{i=1}^{n} t_i$ is simply the treatment proportion in the training dataset. As a result, we get the following objective function

$$\mathcal{L}_n^{CENet} = \frac{1}{n}\sum_{i=1}^{n}\mathcal{L}_i^{\tau} + \gamma\left(\mathcal{L}^{\mu_1} + \frac{p}{1-p}\mathcal{L}^{\mu_0}\right) + \lambda\Omega(\tau) \tag{6.11}$$

The neural network architecture for the CENet model is illustrated in Fig.6.3. We use the stochastic optimization method Adam [90] to train the model. The pseudocode for the joint learning process is summarized in Algorithm 6.1.

6.3.3 BCENet: CENet with Balanced Representation Layers

In observational causal inference, it is well-acknowledged that covariate imbalance between the treated and control groups biases treatment effect estimation. Theorem 1 in [178] derives an upper bound of the treatment effect estimation error using the factual outcome prediction error and the distance between the covariate distributions of the treated and control groups. This suggests that treatment response modelling under a constraint that encourages better distributional balance between the two treatment groups will theoretically benefit treatment effect estimation. In this section, we adopt this idea and propose to add shared representation layers into the above CENet model. The resulting neural network architecture is illustrated in Fig.6.4. Since we are in principle introducing a balancing constraint into CENet, we call the resulting model Balanced Causal Effect Neural Network (BCENet).

Balanced Representation Layers

Denote the shared representation layers in the BCENet by $\phi : \mathcal{X} \to \Phi$ and parametrize the corresponding transformation function by $\phi(\cdot; W)$. For each individual x_i, the observed covariate will first be transformed into $\phi_W(\mathrm{x}_i) \in \Phi$. By this transformation function, we obtain two samples, $\Phi_0 = \{\phi_W(\mathrm{x}_i) : i \in \mathcal{T}_0\}$ and $\Phi_1 = \{\phi_W(\mathrm{x}_i) : i \in \mathcal{T}_1\}$. Denote the distribution density of Φ_0 and Φ_1 as $p^{t=0}$ and $p^{t=1}$ respectively. We follow [178] and quantify the distributional distance between the two treatment groups by the IPM between $p^{t=0}$ and $p^{t=1}$, i.e.,

$$\mathrm{Disc}(p^{t=0}, p^{t=1}) = IPM_{\mathcal{F}}(p^{t=0}, p^{t=1}) \tag{6.12}$$

IPMs are a class of metric for measuring the distance between distributions [186]. For two probability distributions p and q defined on $\mathcal{X}$, the IPM between them with related to a function family $\mathcal{F} : \mathcal{X} \to \mathbb{R}$ is defined as

$$IPM_{\mathcal{F}}(p, q) = \sup_{f \in \mathcal{F}} \left| \int f(\mathrm{x})(p(\mathrm{x}) - q(\mathrm{x})) \right|$$

With this definition, one goal of BCENet is to predict the observed outcome while minimizing $IPM_{\mathcal{F}}(p^{t=0}, p^{t=1})$. Empirically, given the two transformed samples Φ_0 and Φ_1, it is calculated by $IPM_{\mathcal{F}}(\Phi_0, \Phi_1)$ via

$$IPM_{\mathcal{F}}(\Phi_0, \Phi_1) = \sup_{f \in \mathcal{F}} \left| \frac{1}{n_0} \sum_{i \in \mathcal{T}_0} f(\phi_W(\mathrm{x}_i)) - \frac{1}{n_1} \sum_{j \in \mathcal{T}_1} f(\phi_W(\mathrm{x}_j)) \right| \tag{6.13}$$

According to the choice of the function family $\mathcal{F}$, popular examples of IPMs include the maximum mean discrepancy (MMD) [53] with $\mathcal{F} = \{f : \|f\|_{\mathcal{H}_k} \leq 1\}$ where $\mathcal{H}$ is the Hilbert space induced by a kernel function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$; and the Wasserstein distance [186] with $\mathcal{F} = \{f : f \text{ is 1-Lipschitz}\}$. The calculation

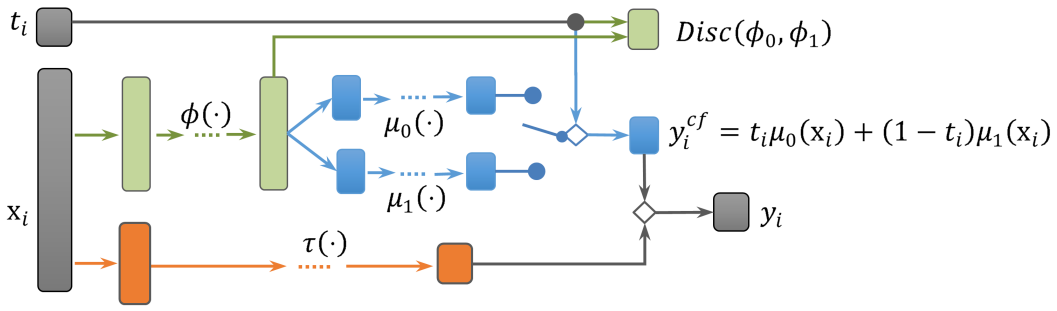


Figure 6.4: Neural network architecture of BCENet. The diamond *toggle* is switched according to the observed treatment for each individual to obtain the unobserved counterfactual outcome.

of the MMD with linear kernel and the Wasserstein distance are provided in the Appendix A.3.

The Objective Function

As illustrated in Fig.6.4, the two outcome prediction networks build on top of the shared representation layers. As a result, the outcome prediction risks in (6.5) becomes

$$\begin{aligned} \mathcal{L}^{\mu_0} &= \frac{1}{n_0} \sum_{i \in \mathcal{T}_0} L\left(\mu_0\left(\phi_W\left(\mathrm{x}_i\right); \beta_0\right), y_i\right) \\ \mathcal{L}^{\mu_1} &= \frac{1}{n_1} \sum_{i \in \mathcal{T}_1} L\left(\mu_1\left(\phi_W\left(\mathrm{x}_i\right); \beta_1\right), y_i\right) \end{aligned} \tag{6.14}$$

Similarly, the loss function for CATE learning will also indirectly depends on the transformation function $\phi(\cdot; W)$ as

$$\mathcal{L}^{\tau} = \frac{1}{n} \sum_{i=1}^{n} \{(2t_i - 1)\tau(\mathrm{x}_i) + t_i\mu_0(\phi_W(\mathrm{x}_i); \beta_0) + (1-t_i)\mu_1(\phi_W(\mathrm{x}_i); \beta_1) - y_i\}^2 \tag{6.15}$$

Combining the outcome prediction risks in (6.14), the CATE risk in (6.15),

Algorithm 6.2 Learning Process for BCENet

Input: Observation data $\mathcal{D} = \{(\mathrm{x}_i, t_i, y_i), \ldots, (\mathrm{x}_n, t_n, y_n)\}$, hyper-parameters $\gamma, \alpha, \lambda > 0$, training batch size B, number of epochs K, and learning rate η

Output: The learned parameters $(\theta, W, \beta_0, \beta_1)$

1: Initialize parameters $(\theta, W, \beta_0, \beta_1)$ for the CATE network $\tau(\cdot;\theta)$, the balanced representation network $\phi(\cdot; W)$ and the outcome prediction networks $\mu_0(\cdot;\beta_0)$, $\mu_1(\cdot;\beta_1)$;

2: Split $\mathcal{D}$ into training and validation sets $\mathcal{D}_{train}$ and $\mathcal{D}_{valid}$;

3: **for** $k = 1, 2, \ldots, K$ **do**

4: Sample $\mathcal{D}_{batch} = \left\{\left(\mathrm{x}_1^k, t_1^k, y_1^k\right), \ldots, \left(\mathrm{x}_B^k, t_B^k, y_B^k\right)\right\}$ from $\mathcal{D}_{train}$

5: Update $(\theta, W, \beta_0, \beta_1)$ to minimize Eq.(6.16) on $\mathcal{D}_{batch}$

$$\begin{aligned}\theta &\leftarrow \theta - \eta \nabla_\theta L_B^{BCENet}\\ W &\leftarrow W - \eta \nabla_W L_B^{BCENet}\\ \beta_0 &\leftarrow \beta_0 - \eta \nabla_{\beta_0} L_B^{BCENet}\\ \beta_1 &\leftarrow \beta_1 - \eta \nabla_{\beta_1} \mathcal{L}_B^{BCENet}\end{aligned}$$

6: Test convergence using $\mathcal{D}_{valid}$, **if** converge

7: **break**

8: **end for**

and the representation balancing regularizer in (6.13) as well as the model complexity regularization term, we get the joint loss function for the BCENet model as

$$\mathcal{L}_n^{BCENet} = \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}_i^{\tau} + \gamma(\mathcal{L}^{\mu_1} + \frac{p}{1-p}\mathcal{L}^{\mu_0}) + \alpha \cdot \mathrm{Disc}(\Phi_0, \Phi_1) + \lambda\Omega(\tau) \quad (6.16)$$

where $\gamma, \alpha, \lambda \geq 0$ are hyper-parameters. According to the IPM used in (6.16), we get two BCENet models: BCENet with the MMD metric (BCENet-MMD), and BCENet with the Wasserstein distance (BCENet-Wass). The pseudocode for BCENet learning is summarized in Algorithm 6.2.

6.4 Experimental Studies

In this section, we evaluate our proposed direct treatment effect estimation models: CENet, BCENet-MMD, and BCENet-Wass. In general, it is hard to validate treatment effect estimation models on observational datasets since we have no access to the underlying ITEs. To cope with this problem, we used semi-simulated data, experimental data from real world applications and simulated data. Details on hyper-parameter configurations are described in the Appendix A.4.

6.4.1 Baselines and Evaluation Metrics

Baseline methods used for comparison include: Least square regression with the treatment as a covariate (OLS/LR-1); separate least square regressions for different groups (OLS/LR-2); BART [62]; S-learning (S-Learner), T-learning (T-Learner), and X-learning (X-Learner) with gradient boosting as the base outcome model and random forest classifier as the base propensity model; domain adaptation learner (DA-Learner), which is a variant of the X-Learner that uses domain adaptation weighting techniques to learn the outcome models via cost-sensitive learning; double robust learner (DR-Learner) that estimates the potential outcomes with double robust statistical techniques [86]; PSM [188], CF [206]; BNN [80], TARNet [178], CFR with MMD discrepancy (CFR-MMD), and CFR with Wasserstein discrepancy (CFR-Wass) [178]. They are grouped into four categories: regression-based methods (OLS/LR and BART); meta-learners (S-Leaner, T-Learner, X-Learner, DA-Learner and DR-Learner), non-parametric methods (PSM, CF), and the other methods are DNN-based methods.

To evaluate the estimation performance, the following metrics are used for ITE, ATE and ATT estimation respectively when the ground truth are known

$$\epsilon_{ITE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\tau_i - \hat{\tau}_i\right)^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\left(\mu_i(1) - \mu_i(0)\right) - \hat{\tau}_i\right)^2}$$

$$\epsilon_{ATE} = |\widehat{ATE} - ATE| = \frac{1}{n}\left|\sum_{i=1}^{n}\left(\left(\mu_i(1) - \mu_i(0)\right) - \hat{\tau}_i\right)\right|$$

$$\epsilon_{ATT} = |\widehat{ATT} - ATT| = \frac{1}{n_1}\left|\sum_{i\in\mathcal{T}_1}\left(\left(\mu_i(1) - \mu_i(0)\right) - \hat{\tau}_i\right)\right|$$

For all evaluation metrics, we report both the within-sample error and the out-of-sample error, where the former is computed over the training and validation sets, and the later is computed over the test set. Standard deviations for multiple replications are also reported.

6.4.2 Semi-simulated Data

In this section, we evaluate the proposed methods using semi-simulated data. The covariates are collected from real world applications, but the treatment responses or the treatment assignments are simulated in order to form an observational dataset. We validate the methods in settings with binary and continuous outcomes respectively.

Binary Outcome: Twins Dataset

The Twins dataset [4] comes from the all twins birth in the USA between 1989-1991 and first used as a benchmark for evaluating causal inference algorithms in [102]. In Twins, the treatment $t = 1$ is being born the heavier twin, and the binary outcome y corresponds to the mortality of each of the twins in their first year of life. Since we have records for both twins (the heavier twin $t = 1$ and the lighter twin $t = 0$), their outcomes could be considered as the two potential

outcomes with respect to the treatment of being born heavier. We focus on the same sex twin-pair whose weights are less than $2kg$ and follow the same pre-processing procedure in [212]. There are in total 5409 records in the final processed data, each record contains 40 pre-treatment covariates related to the parents, the pregnancy and the birth. In order to simulate an observational study with selection bias, one of the twins in each record is selectively chosen as the observation and the other is unobserved. The biased treatment assignment is simulated by

$$t_i|\mathrm{x}_i \sim \mathrm{Bern}\left(\sigma\left(w^T\mathrm{x}_i + \epsilon_i\right)\right)$$

where $\sigma(z) = \frac{1}{1+\exp(-z)}$ is the sigmoid function, $w \sim \mathcal{U}\left((-0.1, 0.1)^{40\times 1}\right)$ and $\epsilon_i \sim \mathcal{N}(0, 0.1)$.

We run the experiments 10 times with a 63/27/10 train/validation/test split ratio. The mean and standard deviation of estimation errors are listed in Table 6.1. In general, all estimators except for BART and PSM obtained pretty similar performance while our proposed models performs generally better than other baselines. Specifically, our BCENet-MMD model obtained the lowest out-of-sample errors in both ITE and ATE estimations. Moreover, for this dataset, balanced representation learning benefits all estimations slightly except for the in-sample ATE estimation. Among the baselines, on one hand, CF obtains similar performance with our proposed models but with a higher out-of-sample ATE error; on the other hand, BART and PSM get relatively much higher estimation errors than the other methods.

Continuous Outcome: IHDP Dataset

The continuous outcome dataset IHDP was first compiled by[62] based on the Infant Health and Development Program (IHDP), which is a randomized experiment to enhance the cognitive and health status of low birth weight, premature

Table 6.1: Within-sample and out-of-sample results on the Twins dataset

	ϵ_{ITE}^{in}	ϵ_{ITE}^{out}	ϵ_{ATE}^{in}	ϵ_{ATE}^{out}
LR1	0.366 ± 0.020	0.366 ± 0.025	0.018 ± 0.011	0.015 ± 0.013
LR2	0.403 ± 0.015	0.411 ± 0.024	0.034 ± 0.035	0.039 ± 0.030
BART	0.465 ± 0.047	0.478 ± 0.061	0.186 ± 0.062	0.182 ± 0.065
T-Learner	0.383 ± 0.007	0.437 ± 0.024	0.012 ± 0.008	0.020 ± 0.015
S-Learner	0.309 ± 0.007	0.346 ± 0.018	0.012 ± 0.007	0.025 ± 0.015
X-Learner	0.334 ± 0.012	0.361 ± 0.022	0.016 ± 0.012	0.027 ± 0.015
DA-Learner	0.376 ± 0.005	0.391 ± 0.018	0.011 ± 0.009	0.025 ± 0.020
DR-Learner	0.302 ± 0.003	0.317 ± 0.014	0.008 ± 0.006	0.012 ± 0.009
PSM	0.357 ± 0.044	0.358 ± 0.049	0.140 ± 0.092	0.140 ± 0.097
CF	0.303 ± 0.002	0.307 ± 0.015	0.008 ± 0.007	0.012 ± 0.009
TARNet	0.329 ± 0.005	0.333 ± 0.019	0.014 ± 0.012	0.015 ± 0.008
BLR	0.306 ± 0.002	0.308 ± 0.014	0.014 ± 0.006	0.015 ± 0.010
BNN	0.307 ± 0.002	0.309 ± 0.015	0.009 ± 0.006	0.013 ± 0.006
CFR-MMD	0.329 ± 0.005	0.333 ± 0.019	0.015 ± 0.008	0.014 ± 0.008
CFR-Wass	0.329 ± 0.005	0.335 ± 0.017	0.014 ± 0.013	0.017 ± 0.015
CENet	0.306 ± 0.002	0.307 ± 0.014	0.005 ± 0.004	0.009 ± 0.006
BCENet-MMD	0.305 ± 0.002	0.306 ± 0.014	0.007 ± 0.004	0.008 ± 0.006
BCENet-Wass	0.305 ± 0.002	0.307 ± 0.014	0.007 ± 0.005	0.008 ± 0.007

infants through paediatric follow-ups and parent support groups. To create an observational study dataset, records with non-white mothers in the treatment group were omitted to make the treatment and control groups unbalanced. The resulting IHDP dataset consists of 747 individuals (139 treated, 608 control), and 25 covariates (6 continuous and 19 binary) measuring properties of children and their mothers. The binary treatment t indicate whether the child was assigned into a program where both intensive high-quality childcare and home visits from a trained provider are provided. Examples of covariates include sex and birth weights of the child, and age, education attainment of the mother.

The observed covariates and treatments in the semi-simulated data are from the IHDP program while all treatment responses are simulated so that the true treatment effects are known. In order to simulate heterogeneous treatment effects

for different children, we used the nonlinear response surface B setting in [62], where the two treatment response functions are simulated as follows:

$$\mu_0\left(\mathrm{x}_i\right)=\exp\left(\left(\mathrm{x}_i+0.5\mathrm{I}\right)^T\beta\right) \quad \mu_1\left(\mathrm{x}_i\right)=\mathrm{x}_i^T\beta-\omega$$

where the 25-dimensional vector of regression coefficients was randomly sampled from $[0, 0.1, 0.2, 0.3, 0.4]$ with probabilities $[0.6, 0.1, 0.1, 0.1, 0.1]$, and the offset ω was chosen such that the true ATE equals 4. The noisy observational outcome is simulated as

$$y_i=t_i\mu_1\left(\mathrm{x}_i\right)+\left(1-t_i\right)\mu_0\left(\mathrm{x}_i\right)+\mathcal{N}(0,I)$$

The simulated noiseless outcomes are used to compute the true effects and results for 10 experiments with a 63/27/10 train/validation/test split ratio are demonstrated in Table 6.2.

As we can see from Table 6.2, the BCENet-MMD model obtained the best out-of-sample performance in both ITE and ATE estimation. Comparing CENet with its two BCENet variants, we found that balancing constraints with the MMD and Wasserstein distance metrics both improve ITE estimation. However, balancing with MMD is preferred than Wasserstein distance for this dataset. This is also indicated by comparing the performance between CFR MMD and CFR Wass. In regard to estimation deviations, estimates of CENet and BCENets are likely to be more stable in the sense of lower standard deviations than that of other estimators. We argue that this is because we estimate the target individual treatment effect in a direct end-to-end manner while other methods do this indirectly. In addition, since the two treatment response function $\mu_0(\mathrm{x})$ and $\mu_1(\mathrm{x})$ are very different to each other, we find that baseline methods using a single outcome model $\mu(\mathrm{x}, t)$, including OLS1, S-Learner, DR-Learner, BLR and BNN all get very high estimation errors and standard deviations. This further

Table 6.2: Within-sample and out-of-sample results on the IHDP dataset

	ϵ^{in}_{ITE}	ϵ^{out}_{ITE}	ϵ^{in}_{ATE}	ϵ^{out}_{ATE}
OLS1	7.218 ± 6.052	7.190 ± 6.595	0.812 ± 0.936	0.810 ± 0.828
OLS2	2.749 ± 2.378	2.791 ± 2.657	0.135 ± 0.113	0.314 ± 0.375
BART	1.137 ± 0.605	1.837 ± 1.512	0.087 ± 0.073	0.204 ± 0.105
T-Learner	2.121 ± 1.512	2.835 ± 2.287	0.078 ± 0.072	0.388 ± 0.617
S-Learner	4.607 ± 4.422	4.544 ± 4.630	0.550 ± 0.780	0.932 ± 1.538
X-Learner	3.239 ± 2.776	3.357 ± 2.943	0.252 ± 0.323	0.291 ± 0.344
DA-Learner	2.032 ± 1.457	2.804 ± 2.414	0.089 ± 0.075	0.319 ± 0.557
DR-Learner	3.766 ± 3.573	3.906 ± 3.878	0.473 ± 0.686	0.780 ± 1.243
PSM	7.860 ± 6.490	7.758 ± 6.816	2.707 ± 3.135	2.624 ± 2.477
CF	5.916 ± 5.088	5.922 ± 5.529	0.355 ± 0.261	0.864 ± 0.961
TARNet	2.035 ± 1.172	1.986 ± 1.000	0.433 ± 0.391	0.458 ± 0.407
BLR	2.276 ± 1.048	2.223 ± 1.013	0.247 ± 0.185	0.247 ± 0.124
BNN	2.734 ± 1.216	2.721 ± 1.205	0.415 ± 0.221	0.379 ± 0.251
CFR-MMD	1.790 ± 0.599	1.868 ± 0.924	0.501 ± 0.355	0.468 ± 0.466
CFR-Wass	2.858 ± 1.786	2.680 ± 1.735	0.792 ± 0.591	0.890 ± 0.566
CENet	1.186 ± 0.123	1.202 ± 0.157	0.158 ± 0.137	0.159 ± 0.117
BCENet-MMD	1.147 ± 0.279	1.086 ± 0.210	0.205 ± 0.111	0.134 ± 0.130
BCENet-Wass	1.150 ± 0.206	1.103 ± 0.183	0.300 ± 0.190	0.300 ± 0.188

validates the superiority of our direct estimation models to classical treatment response modelling models.

6.4.3 Real World Data

We also validate the proposed method using the real Jobs dataset, which combines a randomized study $\mathcal{R}$ based on the National Supported Work program with a larger observational dataset $\mathcal{O}$. This dataset was collected to evaluate the effect of job training programs on the employment status. In the original LaLonde randomized sample $\mathcal{R}$ by [98], there are 722 employees (297 treated and 425 control) with 8 covariates such as age, education, and previous earnings. The binary treatment is whether an employee was enrolled in the job

training program. For more details of the randomized study and data, refer[1]. To evaluate causal inference algorithms, Shalit et al.[178] constructed the Job dataset by combining the LaLonde randomized sample $\mathcal{R}$ with the observational PSID comparison sample $\mathcal{O}$ (2490 control) to predict unemployment after job training. In the Jobs dataset, the original 8 covariates are transformed into a 17 dimension feature set. As a result, we obtain a real world binary-treatment binary-outcome dataset with 3212 examples and 17 dimensional features.

For the Jobs dataset, since the true ITEs are unknown, we are unable to calculate the RMSE ϵ_{ITE}. Following [178] and [102], we use the policy risk estimated for the randomized subset $\mathcal{R}$ as a proxy to the ITE performance

$$R_{pol}(\pi_{\hat{\tau}}) = 1 - \{p(\pi_{\hat{\tau}}(\mathrm{x}) = 1) \cdot \mathbb{E}\left[Y_1 \mid \pi_{\hat{\tau}}(\mathrm{x}) = 1\right] + (1 - p(\pi_{\hat{\tau}}(\mathrm{x}) = 1)) \cdot \mathbb{E}\left[Y_0 \mid \pi_{\hat{\tau}}(\mathrm{x}) = 0\right]\}$$

where $\pi_{\hat{\tau}} : \mathcal{X} \to \{0, 1\}$ is an policy induced from an ITE estimator $\hat{\tau}(\cdot)$ with $\pi_{\hat{\tau}}(\mathrm{x}) = 1$ if $\hat{\tau}(\mathrm{x}) > 0$, and $\hat{\tau}(\mathrm{x}) = 0$ otherwise. This measures the average regret when treating with the induced policy $\pi_{\hat{\tau}}$ and thus can serve as a proxy of the ITE estimation error. Instead of ATE, the NSW program aims at estimating the effect of job training on employment after training for employees enrolled in the training program, i.e., the ATT. Since all the treated individuals came from the randomized study $\mathcal{R}$, we can easily estimate ATT by

$$\begin{aligned} ATT &:= \frac{1}{|\mathcal{T}_1|} \sum_{i \in \mathcal{T}_1} (Y_1(\mathrm{x}_i) - Y_0(\mathrm{x}_i)) \\ &= \frac{1}{|\mathcal{T}_1|} \sum_{i \in \mathcal{T}_1} y_i - \frac{1}{|T_0 \cap \mathcal{R}|} \sum_{i \in T_0 \cap \mathcal{R}} y_i = ATE \end{aligned}$$

where $\mathcal{T}_1$ and $\mathcal{T}_0$ are the treated and control group in the full dataset. We

[1] http://users.nber.org/ rdehejia/data/nswdata2.html

Table 6.3: Within-sample and out-of-sample results on the Jobs dataset

	R_{pol}^{in}	R_{pol}^{out}	ϵ_{ATT}^{in}	ϵ_{ATT}^{out}
LR1	0.268 ± 0.000	0.334 ± 0.000	0.008 ± 0.000	0.045 ± 0.000
LR2	0.274 ± 0.000	0.275 ± 0.000	0.132 ± 0.000	0.156 ± 0.000
BART	0.204 ± 0.009	0.346 ± 0.011	0.123 ± 0.008	0.048 ± 0.007
T-Learner	0.071 ± 0.003	0.338 ± 0.010	0.012 ± 0.003	0.045 ± 0.003
S-Learner	0.140 ± 0.009	0.274 ± 0.016	0.021 ± 0.003	0.044 ± 0.005
X-Learner	0.100 ± 0.006	0.320 ± 0.017	0.006 ± 0.003	0.046 ± 0.007
DA-Learner	0.067 ± 0.007	0.364 ± 0.026	0.011 ± 0.002	0.042 ± 0.011
DR-Learner	0.066 ± 0.004	0.410 ± 0.026	0.007 ± 0.003	0.051 ± 0.023
PSM	0.288 ± 0.000	0.321 ± 0.000	0.350 ± 0.000	0.300 ± 0.000
CF	0.156 ± 0.008	0.291 ± 0.016	0.013 ± 0.004	0.032 ± 0.005
TARNet	0.199 ± 0.016	0.333 ± 0.028	0.041 ± 0.034	0.042 ± 0.046
BLR	0.219 ± 0.000	0.296 ± 0.000	0.047 ± 0.023	0.051 ± 0.023
BNN	0.219 ± 0.000	0.296 ± 0.000	0.049 ± 0.030	0.051 ± 0.034
CFR-MMD	0.197 ± 0.015	0.345 ± 0.020	0.025 ± 0.021	0.060 ± 0.025
CFR-Wass	0.200 ± 0.008	0.350 ± 0.014	0.028 ± 0.022	0.057 ± 0.041
CENet	0.167 ± 0.016	0.265 ± 0.021	0.025 ± 0.015	0.058 ± 0.028
BCENet-MMD	0.183 ± 0.014	0.285 ± 0.039	0.020 ± 0.015	0.016 ± 0.015
BCENet-Wass	0.186 ± 0.017	0.292 ± 0.043	0.027 ± 0.009	0.035 ± 0.019

replicated the experiment 10 times with a 56/24/20 train/validation/test ratio. Average performances and their standard deviations are list in Table 6.3.

As indicated in Table 6.3, CENet obtained the best out-of-sample performance in ITE estimation and BCENet-MMD obtained the best out-of-sample performance in ATE estimation. For this extremely imbalanced dataset, balanced representation learning does not helps ITE estimation but does benefit ATE estimation. Meta-learners based on ensemble algorithms tended to overfit the training data while performed relatively worse for out-of-sample estimation.

6.4.4 Experiment on Synthetic Data

To further check the robustness of the proposed models and their performance in different sample size and imbalance settings, we adapted the data simulation

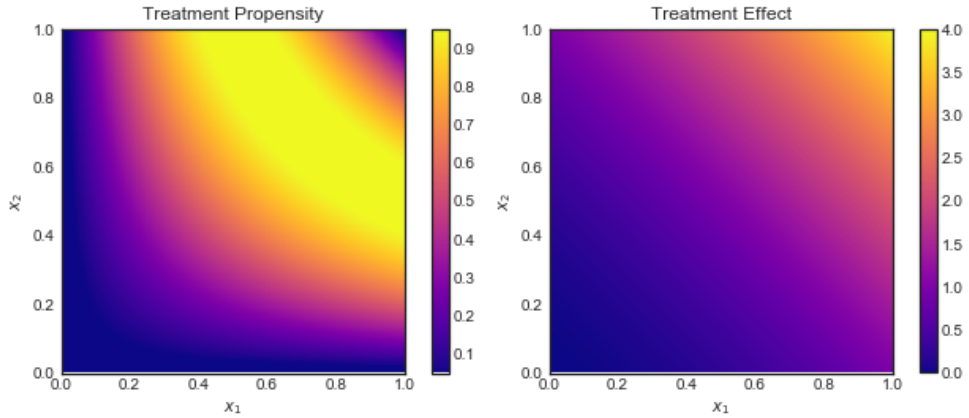


Figure 6.5: Visualization of (a): the treatment propensity function $\text{trim}_\eta\left(\sin\left(\pi \text{x}_{i1}\text{x}_{i2}\right)\right)$ when $\eta = 0.05$, and (b): the underlying heterogeneous treatment effect function in terms of the first two dimensions.

setup A in [126] and simulate data by the following data-generating process:

$$\text{x}_i \sim U(0,1)^5, \quad t_i|\text{x}_i \sim \text{Bern}\left(\text{trim}_\eta\left(\sin\left(\pi \text{x}_{i1}\text{x}_{i2}\right)\right)\right)$$

where $\text{trim}_\eta(z) = \max\{\eta, \min(z, 1-\eta)\}$ and $\eta \in (0,1)$ is the imbalance parameter. To illustrate the selection bias in the underlying treatment assignment mechanism, we plot the treatment propensities in terms of the first two dimensions for $\eta = 0.05$ in Fig.6.5(a). The treatment response functions and the observed outcome for each individual are respectively

$$\mu_0(\text{x}_i) = \sin(\pi \text{x}_{i1}\text{x}_{i2}) + 2(\text{x}_{i3} - 0.5)^2 + \text{x}_{i4} + 0.5\text{x}_{i5}$$

$$\mu_1(\text{x}_i) = \mu_0(\text{x}_i) + (\text{x}_{i1} + \text{x}_{i2})^2$$

$$y_i = t_i\mu_1(\text{x}_i) + (1 - t_i)\mu_0(\text{x}_i)$$

As a result, the underlying CATE function is

$$\tau(x_i) = \mu_1(x_i) - \mu_0(x_i) = (x_{i1} + x_{i2})^2$$

and is illustrated in Fig.6.5(b). We simulated data with sample size $n = 500, 1K, 3K, 5K, 7K, 10K$ and $\eta = 0.005, 0.01, 0.05, 0.1, 0.5$. For each simulation setting, we split the data into train/validation/test sets with a ratio of 56/24/20 and replicated the experiments 10 times. We compared CENet and BCENet with other DNN-based baselines, BLR, TARNet, BNN, CFR-MMD and CFR-WASS. All neural networks have similar configurations, with 2 hidden layers for each component and 50 neurons each layer. Hyper-parameters are set as $\gamma = 1$ and $\alpha = \lambda = 0.0001$. The training batch size was 200.

We first evaluated the performance of the proposed models in different imbalance settings. Error bar plots for out-of-sample ITE and ATE estimation errors in terms of the imbalance parameter η in the setting of sample size $n = 1000$ are illustrated in Fig.6.5. More estimation performances for other sample sizes are provided in Appendix A.4. As we can see from the results, as the imbalance parameter η increases, estimation errors of all methods generally decreased. For ITE estimaiton, BLR and BNN have higher errors than other estimators in each η level, and our BCENet models obtained the lowest errors and have smaller standard deviations. As for ATE estimation, BCENet models also perform the best in most imbalance settings except when $\eta = 0.005$. We also investigated the influence of sample size n on estimation performances. Plots of out-of-sample estimation errors in terms of sample sizes when $\eta = 0.05$ are illustrated in Fig.6.7 (plots for other settings are provided in Appendix A.5).

In general, as sample size gets larger, the performances did not change very much. Both BCENet models obtained the lowest error in ITE estimation but obtained slightly larger errors than TARNet and CFR in ATE estimation. How-

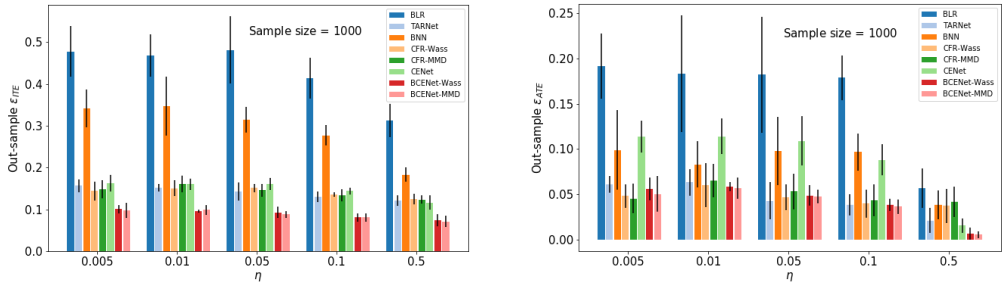


Figure 6.6: Comparisons of out-of-sample errors (left: ITE error ϵ_{ITE}, right: ATE error ϵ_{ATE}) and corresponding standard deviations in terms of the imbalance parameter η when sample size $n = 1000$.

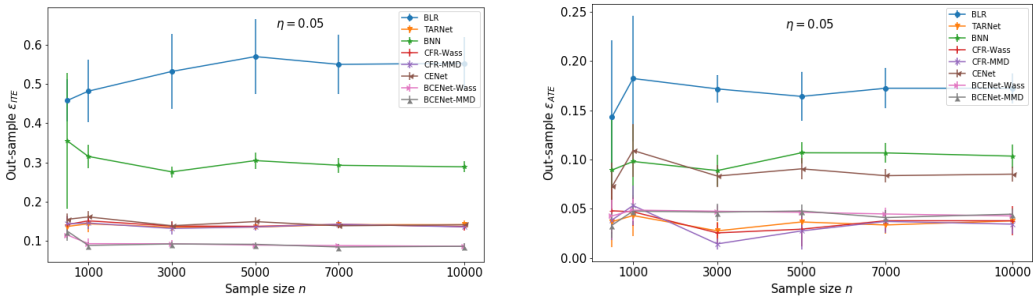


Figure 6.7: Comparisons of out-of-sample errors (left: ITE error ϵ_{ITE}, right: ATE error ϵ_{ATE}) and corresponding standard deviations in terms of the sample size n when the imbalance parameter $\eta = 0.05$.

ever as the sample size increase, this gap is getting smaller. Similar with that in ITE estimation, our proposed models generally obtained estimations with smaller standard deviations.

From these simulations, we can conclude that our BCENet models outperform other competitors in ITE estimation and match them in ATE estimation in different imbalance and sample size settings, moreover, they are generally more stable. Comparing them with the naive CENet, we further know that adding balanced representation learning layers improves both ITE and ATE estimations.

6.5 Summary

In this chapter, we proposed the idea of direct treatment effect estimation and related models for causal inference from observational data. Compared with existing methods that fulfil the task in an indirect way by first learning some auxiliary functions, direct treatment effect estimation parametrizes and learns the target treatment effect function directly. With this idea, we proposed a causal effect neural network, CENet, which parametrizes the treatment effect function with DNNs and learns it directly. In addition, we adopted the idea of learning balanced representations with IPMs to further reduce the impact of selection bias in observational data. This results in our second direct effect estimation model, BCENet. An obvious strength of the proposed models is that the target treatment effect function is learned directly in an end-to-end manner, which is very simple and intuitive. We conducted comprehensive experiments and compared the proposed models with a range of baselines. Experiment results showed that the proposed models performed generally better than existing baselines and tended to obtain more stable estimates. Note that we focused on treatment effect estimation with binary treatments and a possible future research question is how to extend the direct treatment effect estimation to settings with multivariate treatment and even continuous treatments.

Chapter 7

Conclusion and Future Directions

In this chapter, we summarize the thesis and discuss several future research directions.

7.1 Conclusion

In this thesis, we study models and algorithms for learning causal relationships and estimating causal effects from observational data.

For the purpose of learning causal relationships, most of existing models are unable to model feedbacks or instantaneous and cross-temporal causal relations commonly existed in real world causal systems. To handle this limitation, in Chapter 3, we propose the FoCP for modelling dynamic causal systems with both instantaneous and cross-temporal causal relations. In addition, an algorithm based on the classic PC algorithm is also proposed to learn the causal structure of FoCPs from time series data. The proposed FoCP and corresponding structure learning algorithm are validated using simulated data and are also applied for modelling real world climate data.

For the purpose of estimating treatment effects, we propose three methods

within different frameworks. In Chapter 4, we consider weighting estimators for causal inference. To stabilize the traditional IPW estimator, we proposed a pareto-smoothing method and two pareto-smoothed IPW estimators for ATE and ATT estimation. Experiments on simulated and semi-simulated data indicate that the proposed method performs generally better than other methods. In addition, an application of the propose method to a health survey dataset confirms its efficacy.

In Chapter 5, we consider the counterfactual inference framework for treatment effect estimation. Specifically, we consider causal inference with unobserved confounders and observed proxies and model the underlying data-generating process using IGMs. Based on recent advances in Bayesian inference and deep generative models, we proposed the CEIGM for treatment effect estimation. We evaluate the proposed CEIGM using both semi-simulated and real-world experimental datasets, and experiment results indicate that CEIGM can learn better treatment outcome functions.

In Chapter 6, we propose direct treatment effect estimation that models the target CATE function with deep neural networks and learns it directly from observational data. By considering different constraints, we further proposed two neural network architectures for direct treatment effect estimation, the CENet and BCENet. The proposed direct treatment effect estimation idea is simple and intuitive. Results from comprehensive experiments show that our proposed direct estimators perform generally better and tend to obtain stable causal estimates in various settings.

7.2 Future Directions

Causality is a very broad and emerging topic in the field of AI. In recent years, a range of interesting and promising problems and methods have sprung up. In this section, I summarize several research topics that are unexplored in this thesis as future directions.

7.2.1 Learning Causality for General Entities

Existing studies on learning causality are mostly for observations of random variables or random processes. In these settings, the definition of cause and effect variables is clear and observations can be organized in a data frame or matrix. However, when the objects to be analysed are general entities with complex inner structures(e.g., images and documents), the concept of random variables is not as obvious as what we have considered in this thesis. For example, a theorem consists of words and equations; an image is made up by a set of pixels representing different objects. How to learn causal relationships between these general entities and estimate treatment effects for them is an important while almost untouched problem.

As a first attempt, Carulla et al. [151] proposed to extract random proxy variables using a proxy projections for causal discovery from static entities such as images and documents. For causal inference, Veitch et al. [204] use text embedding to conduct causal inference from text documents. In general, research for learning causality for general entities is limited in the literature. Challenges in this setting include that the data is usually unstructured and of high-dimension. In addition, how to define the treatment and outcome variables in these setting remains a problem. We leave it as a future research topic.

7.2.2 Causal Inference with Continuous Treatment

Most methods for causal inference consider the treatment to be binary. The justification is that either the individual received the treatment or not. In reality, many treatments are continuous and can take a range of values. The response to these continuous treatments is usually dose-dependent and non-linear [137]. For instance, drugs often have a recommended dosage. An insufficient does will fail to be effective and an excessive dose can have a negative effete.

Recently, causal inference with continuous treatments (e.g., dose, duration, and frequency) has received many attentions [87, 45, 137, 210]. Importantly, such treatments lead to effects that are naturally described by curves (e.g., dose-response curves) and extending causal inference methods for binary treatments to continuous settings is non-trivial [87]. As a result, research on either extending existing methods for binary treatments or developing ad-hoc methods for causal inference with continuous treatments will be a theoretical and practical promising direction.

7.2.3 High-dimensional Causal Inference and Variable Selection

Causal inference from observational data requires adjustment for all confounding variables. In this thesis, we generally assume no unobserved confounders (Assumption 4.1) and estimate treatment effects by controlling for all observed covariates. However, in high-dimensional settings, the set of candidate control variables is often quite large relative to the available sample size, and controlling for all observed covariates becomes infeasible in practice. Researchers have recently noticed the importance of variable selection for causal inference with high-dimensional covariates [81, 181].

There are many methods for variable selection in high-dimensional predictive machine learning. However, as we have discussed in Chapter 1, causal prediction is different from associational prediction and the set of covariates useful for prediction may be very different from that for causal estimation. In the age of big data, with the use of high-dimensional electronic health records, administrative databases, and large-scale genomic and imaging datasets getting increasingly common [6]. The problem of high-dimensional causal inference and variable selection for causal inference is getting even more urgent.

7.2.4 Learning Treatment Policy from Observational Data

Existing literature on causality has mainly focused on learning causal relationships and estimating causal effects from observational data. Apart from causal relationships and causal effects, a very important problem for learning with observational data is how to get an optimal treatment policy for future planning. Given a set of treatment options, policy optimization or learning is the problem of choosing the best option for each individual. To learn such policies, it is essential for us to be able to evaluate the efficacy of an existing policy using observational data. This is essentially a causal inference problem – estimating the treatment effect of each treatment option induced by the target policy – and models and techniques for causal inference can be adopted for this problem. Since the data used for policy evaluation and optimization is off-line observations. This problem is also called as off-policy optimization.

Recently, Lu et al. [103] proposed to combine RL with causal inference for learning good polices from historical data with confounding bias. Their research indicates that causality and RL are complementary and can be integrated from the causal perspective to enhance both. Given the growing observational data in healthcare [51] and online advertising [18], how to adapt techniques for ad-

justing confounding bias from causal inference to develop effective algorithms for learning treatment policies from observational data is an interesting direction for future research.

7.2.5 Causality-based Machine Learning

We have shown that the research of learning causality from observational data, including learning causal relationships and estimating causal effects, benefits greatly from recent advances in machine learning. As indicated in [168] and [220], causal information about the underlying data-generating process can in turn help to understand the behaviour of the target system under changing, unseen environments, and thus is able to facilitate solving a number of machine learning problems such as semi-supervised learning, covariate shift, and transfer learning [108, 151, 168, 218].

There is a growing interest in causality-based machine learning. Schölkopf et al. [168] classify existing predictive machine learning tasks into *causal learning* (predicting effect from cause) and *anti-causal learning* (predicting cause from effect). More recently, causal models and causal inference techniques have been adapted for learning generative models that we can intervene [92], for quantifying algorithmic fairness [96, 107, 217] and for interpreting machine learning algorithms [22, 121].

Causality focus on the underlying data-generating process of a system and the causal relationships between variables. Causal information should be more stable and robust to possible environment changing than the purely associational information widely used in existing machine learning algorithms. As a result, developing new machine learning algorithms and applications with causality in mind is another promising research direction.

Appendices

A.1 Estimation of ATT

In the main text, we demonstrate the proposed method by focusing on the estimation of ATE. If the estimand of interest is the ATE for the treated, $\tau_{ATT} = \mathbb{E}[Y(1)-Y(0)|T=1]$, the covariate distribution for our target population is then $p_X^{t=1} := p(X|T=1)$. The expected potential outcomes $\mathbb{E}[Y(0)|T=1]$ and $\mathbb{E}[Y(1)|T=1]$ are estimated using importance sampling via

$$\mathbb{E}[Y(1)|T=1] = \mathbb{E}_{p_X^{t=1}}\mathbb{E}[Y|X] = \frac{1}{n_1}\sum_{i:T_i=1} Y_i$$

$$\begin{aligned}
\mathbb{E}[Y(0)|T=1] &= \mathbb{E}_{p_X^{t=1}}\mathbb{E}[Y(0)|X] \\
&= \mathbb{E}_{p_X^{t=1}}\mathbb{E}[Y|X,T=0] \\
&= \frac{1}{n_0}\sum_{i:T_i=0}\frac{p(X_i|T_i=1)}{p(X_i|T_i=0)}Y_i \\
&= \frac{1}{n_0}\sum_{i:T_i=0}\frac{e(X_i)}{1-e(X_i)}Y_i
\end{aligned}$$

Apparently, we only need to weight the individuals in the control group to match the treatment group. Define the importance weight for X_i as

$$W_i = \begin{cases} 1, & \text{if } T_i = 1 \\ \frac{e(X_i)}{1-e(X_i)}, & \text{if } T_i = 0 \end{cases} \tag{A.1.1}$$

After Pareto-smoothing W_i for individuals in the control group, we obtain the Pareto-smoothed importance weights $\{W_1^{PS}, W_2^{PS}, \ldots, W_n^{PS}\}$, and the Pareto-smoothed IPW estimator for ATT is defined as

$$\begin{aligned}\hat{\tau}_{ATT}^{\mathrm{PS}} &= \frac{1}{n_1}\sum_{i:T_i=1} W_i^{\mathrm{PS}} Y_i - \frac{1}{n_0}\sum_{i:T_i=0} W_i^{\mathrm{PS}} Y_i \\ &= \frac{1}{n_1}\sum_{i:T_i=1} Y_i - \frac{1}{n_0}\sum_{i:T_i=0} W_i^{\mathrm{PS}} Y_i\end{aligned} \tag{A.1.2}$$

The implementation procedure is summarized in Algorithm A.1.1

Algorithm A.1.1 Pareto-smoothed IPW ATT Estimator

Input: Observation data $\mathcal{D} = \{(X_i, T_i, Y_i), \ldots, (X_n, T_n, Y_n)\}$
Output: The estimated $\hat{\tau}_{ATT}^{\mathrm{PS}}$

1: Fit a treatment propensity model $e(X)$ form $\mathcal{D}$;
2: Calculate the importance weights for each unit by E.q(A.1.1) to obtain $\{W_i, i = 1, \ldots, n\}$
3: Sort the importance weights $\{W_i, i = 1, \ldots, n\}$ ascendingly to obtain the sorted importance weights $W_{[1]}, W_{[2]}, \ldots, W_{[n]}$
4: Choose the location parameter $\hat{\mu}$ by E.q(4.18)
5: Estimate the parameters σ and k by E.q(4.23)
6: Smooth the importance weights $\{W_1, W_2, \ldots, W_n\}$ by E.q(4.25) to obtain the Pareto-smoothed importance weights $\{W_1^{\mathrm{PS}}, W_2^{\mathrm{PS}}, \ldots, W_n^{\mathrm{PS}}\}$
7: Estimate the ATE $\hat{\tau}_{ATT}^{\mathrm{PS}}$ via E.q(A.1.2)

Moreover, in the case of estimating the ATT, the corresponding estimation bias is defined as

$$Bias_{\mathrm{ATT}} = |\hat{\tau}_{ATT} - \tau_{ATT}| = \left|\hat{\tau}_{ATT} - \frac{1}{n_1}\sum_{t_i=1}(Y_i(1) - Y_i(0))\right|$$

A.2 Identifiability of Counterfactuals and Treatment Effects

Since the training process of the CATE funciton $\tau(\mathrm{x})$ in our proposed models needs an estimation of the counterfactual outcome as an intermediate estimand as in (6.7), we give a brief proof of the identifiability of the treatment response function $\mu_t(\mathrm{x}) = \mathbb{E}\left[Y_i(t)|X_i = \mathrm{x}\right]$ from a set of observational data $\mathcal{D} = \{(\mathrm{x}_i, t_i, y_i), i = 1, \ldots, n\}$. This is realized by

$$\begin{aligned}\mathbb{E}\left[Y_i(t)|X_i = \mathrm{x}\right] &= \mathbb{E}\left[Y_i(t)|X_i = \mathrm{x}, T_i = t\right] \text{ (by Assumption 6.1)} \\ &= \mathbb{E}\left[Y_i|X_i = \mathrm{x}, T_i = t\right] \quad \text{(by Assumption 2.1)}\end{aligned}$$

As a result, under Assumption 2.1 and Assumption 6.1, we can estimate $\mu_t(\mathrm{x})$ by fitting a prediction model $\mathbb{E}\left[Y_i|X_i = \mathrm{x}, T_i = t\right]$ from $\mathcal{D}$. In practice, this is realizable when the observational data satisfies the positivity assumption (Assumption 6.2).

Furthermore, by the definition $\tau(\mathrm{x}) = \mu_1(\mathrm{x}) - \mu_0(\mathrm{x}) = \mathbb{E}\left[Y_i(1)|X_i = \mathrm{x}\right] - \mathbb{E}\left[Y_i(0)|X_i = \mathrm{x}\right]$, it is obvious that $\tau(\mathrm{x})$ is identifiable given that we can estimate $\mu_t(\mathrm{x})$ from observational data. Since $ATE = \mathbb{E}\left[\tau\left(\mathrm{x}_i\right)\right]$ and $ATT = \mathbb{E}\left[\tau\left(\mathrm{x}_i\right)|t_i = 1\right]$. They are also identifiable.

A.3 Representation Balancing Metrics

In this section, we introduce the calculation of the IPM between two samples when the function family $\mathcal{F}$ is $\mathcal{F} = \{f : \|f\|_{\mathcal{H}_k} \leq 1\}$ (MMD) where $\mathcal{H}_k$ is a universal reproducing Hilbert kernel space and $\mathcal{F} = \{f : f \text{ is 1-Lipschitz}\}$ (Wasserstein distance) used in our experiments.

A.3.1 Calculating the Empirical MMD

Denote the size of treated and control groups in a batch as B_0 and B_1. Samples in the two treatment groups are then transformed into $S_0 = \left\{\phi_W(\mathrm{x}_1^0), \ldots, \phi_W(\mathrm{x}_{B_0}^0)\right\}$ and $S_1 = \left\{\phi_W(\mathrm{x}_1^1), \ldots, \phi_W(\mathrm{x}_{B_1}^1)\right\}$ by the representation learning network $\phi_W(\cdot)$. For a kernel $k(\cdot,\cdot)$ in the feature space Φ, the MMD of the two samples can be written as

$$\begin{aligned} MMD_k\left(S_0, S_1\right) = & \frac{1}{B_0\left(B_0-1\right)} \sum_{i=1}^{B_0} \sum_{v=1, v \neq i}^{B_0} k\left(\phi_w\left(\mathrm{x}_i^0\right), \phi_w\left(\mathrm{x}_v^0\right)\right) \\ & + \frac{1}{B_1\left(B_1-1\right)} \sum_{j=1}^{B_1} \sum_{v=1, v \neq j}^{B_1} k\left(\phi_w\left(\mathrm{x}_j^1\right), \phi_w\left(\mathrm{x}_v^1\right)\right) \\ & + \frac{2}{B_0 B_1} \sum_{i=1}^{B_0} \sum_{j=1}^{B_1} k\left(\phi_w\left(\mathrm{x}_i^0\right), \phi_W\left(\mathrm{x}_j^1\right)\right) \end{aligned}$$

In the experiment, we use the linear kernel $k(\mathrm{x}, \mathrm{x}') = |\mathrm{x} - \mathrm{x}'|$. The corresponding linear MMD is the distance between the embedded means of the two groups, i.e.,

$$MMD_k\left(S_0, S_1\right) = 2\left\|\frac{1}{B_0} \sum_{i=1}^{B_0} \phi_W\left(\mathrm{x}_i^0\right) - \frac{1}{B_1} \sum_{j=1}^{B_1} \phi_W\left(\mathrm{x}_j^1\right)\right\|_2$$

A.3.2 Approximating the Wasserstein Distance

In the main text, the Wasserstein distance between two samples is defined as an IPM with the function family $\mathcal{F} = \{f : f \text{ is 1-Lipschitz}\}$. However, it is not clear how to compute it or its gradients with such a definition. In this section, we introduce its original definition as a distance induced by solving an optimal transport (OT) problem [139]

$$\begin{aligned} \text{Wass}_C\left(S_0, S_1\right) &:= \min_{P \in \Gamma_{n_0,n_1}} \langle C, P \rangle \\ &= \min_{P \in \Gamma_{n_0,n_1}} \sum_{i=1}^{n_0} \sum_{j=1}^{n_1} C_{i,j} P_{i,j} \end{aligned} \tag{A.3.1}$$

where $\Gamma_{n_0,n_1} = \left\{P \in \mathbb{R}_+^{n_0 \times n_1} : P\mathbb{1}_{n_1} = n_0^{-1}, P^T\mathbb{1}_{n_0} = n_1^{-1}\right\}$ is the set of $n_0 \times n_1$ matrices with non-negative entries and is called the set of couplings between S_0 and S_1; C is the transportation cost matrix with $C_{(i,j)} := c(a_i, b_j)$ the cost of transporting a unit mass from $a_i \in S_0$ to $b_j \in S_1$. By this definition, computing the Wasserstein distance involves solving a linear program and the solution $P^* \in \Gamma_{n_0,n_1}$ is called the optimal transport matrix. Historically, the IPM definition of the Wasserstein distance is derived by recasting the OT problem in (A.3.1) into its Lagrangian dual problem and several lines of algebraic manipulations.

Solving the linear program in (A.3.1) is prohibitively expensive for high-dimensional covariates. In our experiment, we followed [178] and approximated the empirical Wasserstein distance between the two transformed samples $S_0 = \left\{\phi_W(\mathrm{x}_1^0), \ldots, \phi_W(\mathrm{x}_{B_0}^0)\right\}$ and $S_1 = \left\{\phi_W(\mathrm{x}_1^1), \ldots, \phi_W(\mathrm{x}_{B_1}^1)\right\}$ through the Sinkhorn-Knopp matrix scaling algorithm from [32]. In the implementation, the $B_0 \times B_1$ transportation cost matrix C in for each batch is calculated by

$$C_{i,j} = c\left(\mathrm{x}_i, \mathrm{x}_j\right) = \left\|\phi_W\left(\mathrm{x}_i^0\right) - \phi_W\left(\mathrm{x}_j^1\right)\right\|_2$$

With this transportation cost matrix, the optimal transport matrix P^* is approximated using Algorithm 3 of [33].

A.4 Hyperparameters

In all experiments, we applied the Xavier initialization for weight matrices, bias vectors are initialized by 0, and scalar biases are initialized by 0.1, dropout

Table A.3.1: Hyperparameters and ranges

	IHDP	**Twins**	**Jobs**
Outcome prediction parameter, γ	$\{10^{-k}\}_{k=0}^{2}$	$\{10^{-k}\}_{k=0}^{2}$	$\{10^{-k}\}_{k=0}^{2}$
Representation balancing parameter, α	$\{10^{-k}\}_{k=2}^{4}$	$\{10^{-k}\}_{k=2}^{4}$	$\{10^{-k}\}_{k=2}^{4}$
Weight-decay parameter, λ	$\{10^{-k}\}_{k=2}^{4}$	$\{10^{-k}\}_{k=2}^{4}$	$\{10^{-k}\}_{k=2}^{4}$
Num. of hidden layers	2,3	2,3	2,3
Dim. of hidden layers	50,100,200	50,100,200	50,100,200
Batch size	200,500	200,500	200,500

Table A.3.2: Optimal hyper-parameters for CENet on each dataset

	IHDP	**Twins**	**Jobs**
Outcome prediction parameter, γ	1	1	1
Weight-decay parameter, λ	0.01	0.01	0.0001
Num. of outcome prediction layers	3	3	3
Num. of CATE layers	2	3	3
Dim. of outcome prediction layers	200	200	200
Dim. of CATE layers	200	50	50
Batch size	500	200	500

Table A.3.3: Optimal hyper-parameters for BCENet on each dataset

	IHDP	**Twins**	**Jobs**
Outcome prediction parameter, γ	1	1	1
Representation balancing parameter, α	0.0001	0.001	0.0001
Weight-decay parameter, λ	0.01	0.01	0.0001
Num. of representation layers	2	2	3
Num. of outcome prediction layers	3	3	3
Num. of CATE layers	2	3	3
Dim. of representation layers	200	200	200
Dim. of outcome prediction layers	200	200	200
Dim. of CATE layers	200	50	50
Batch size	500	200	500

rates in all hidden layers are set to 0.1. The search range of neural network hyperparameters are listed in Table A.3.1. Hyperparameter configurations of the proposed CENet and BCENet models for the three benchmark datasets are listed in Table A.3.2 and Table A.3.3 respectively.

A.5 Additional Experimental Results

More results for simulation studies are illustrated in Fig.A.5.1-A.5.4.

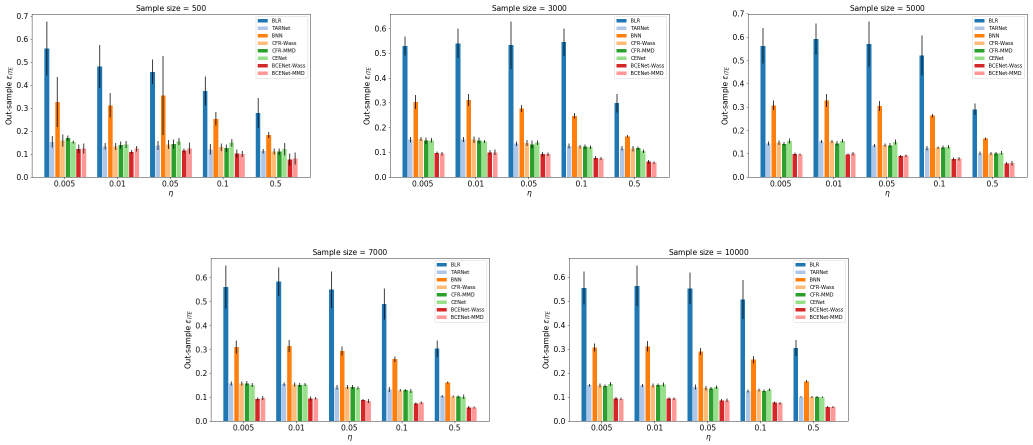


Figure A.5.1: ϵ_{ITE}^{out} in terms of the imbalance parameter η for different sample size n.

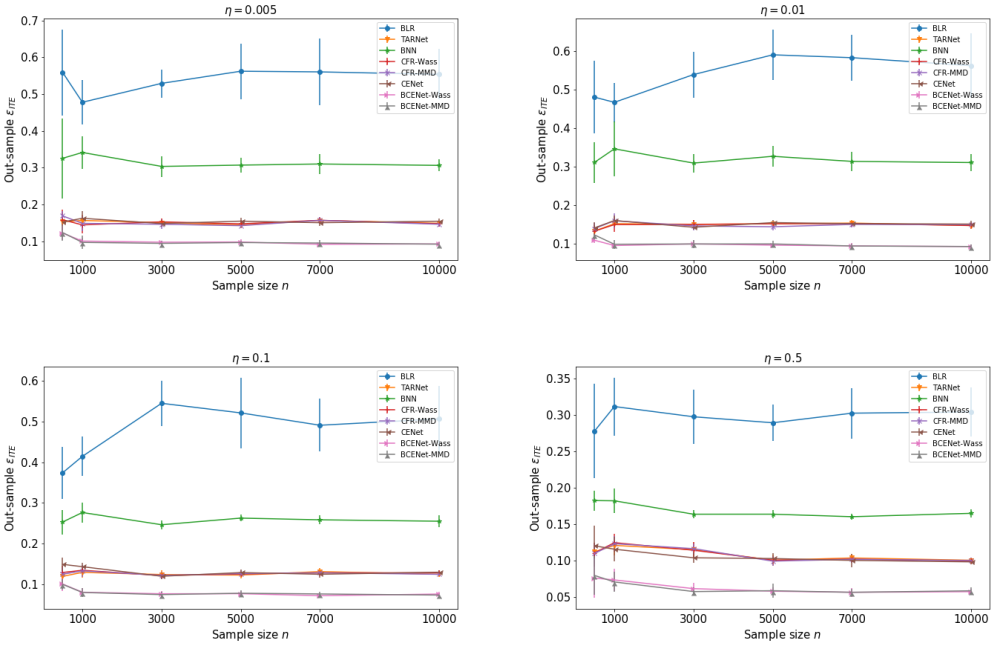


Figure A.5.2: ϵ_{ITE}^{out} in terms of sample size n for different imbalance parameter η.

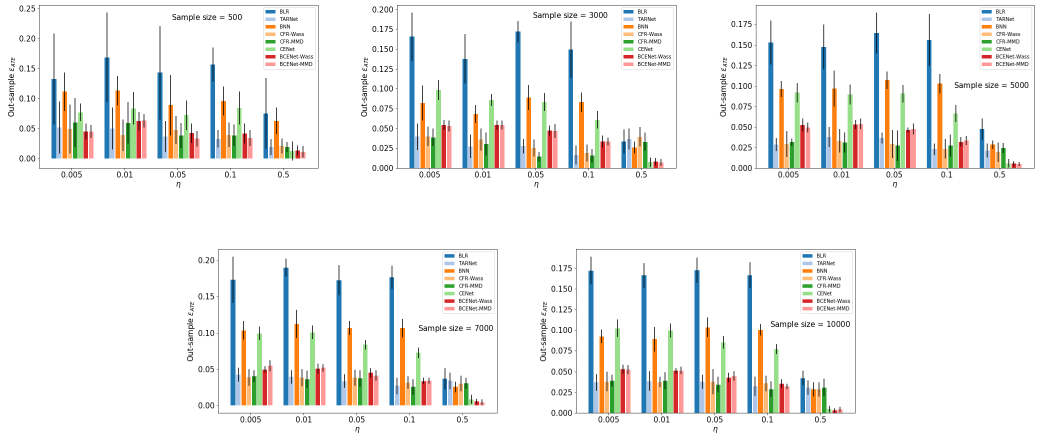


Figure A.5.3: ϵ_{ATE}^{out} in terms of the imbalance parameter η for different sample size n.

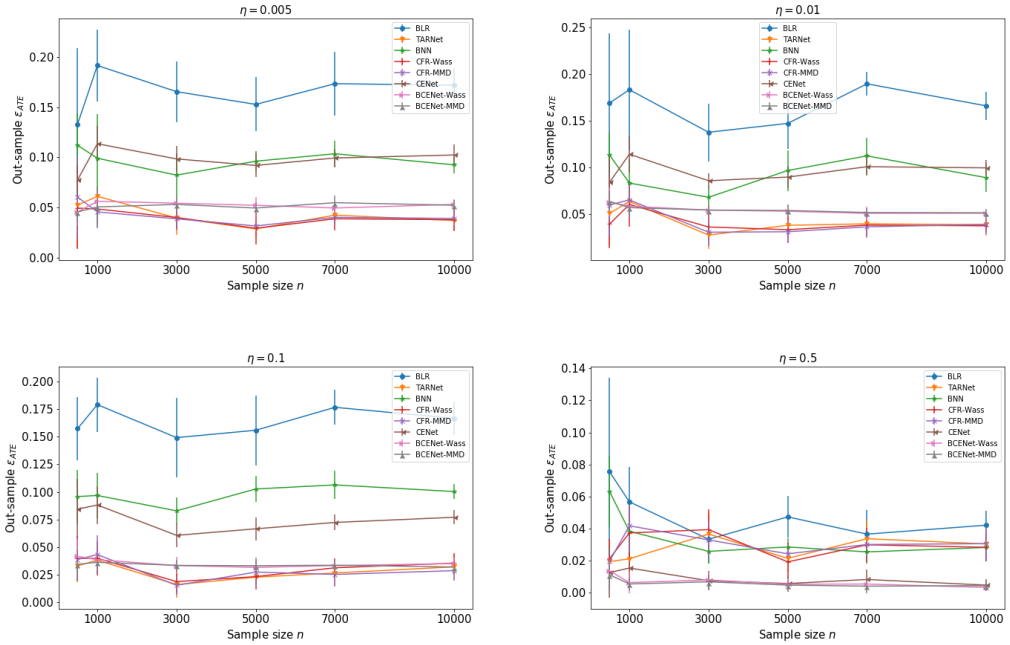


Figure A.5.4: ϵ_{ATE}^{out} in terms of sample size n for different imbalance parameter η.

BIBLIOGRAPHY

[1] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: A system for large-scale machine learning. In *12th Symposium on Operating Systems Design and Implementation*, pages 265–283, 2016.

[2] Ahmed M Alaa and Mihaela van der Schaar. Bayesian inference of individualized treatment effects using multi-task gaussian processes. In *Advances in Neural Information Processing Systems*, pages 3424–3432, 2017.

[3] Ahmed M Alaa and Mihaela van der Schaar. Bayesian nonparametric causal inference: Information rates and learning algorithms. *IEEE Journal of Selected Topics in Signal Processing*, 12(5):1031–1046, 2018.

[4] Douglas Almond, Kenneth Y Chay, and David S Lee. The costs of low birth weight. *The Quarterly Journal of Economics*, 120(3):1031–1083, 2005.

[5] Joshua D Angrist and Jörn-Steffen Pischke. *Mostly harmless econometrics: An empiricist's companion*. Princeton university press, 2008.

[6] Joseph Antonelli, Matthew Cefalu, Nathan Palmer, and Denis Agniel. Doubly robust matching estimators for high dimensional confounding adjustment. *Biometrics*, 74(4):1171–1179, 2018.

[7] Kevin Arceneaux, Alan S Gerber, and Donald P Green. Comparing experimental and matching methods using a large-scale voter mobilization experiment. *Political Analysis*, 14(1):37–62, 2006.

[8] Onur Atan, James Jordon, and Mihaela van der Schaar. Deep-treat: Learning optimal personalized treatments from observational data using neural networks. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[9] Susan Athey. Beyond prediction: Using big data for policy problems. *Science*, 355(6324):483–485, 2017.

[10] Susan Athey and Guido Imbens. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, 2016.

[11] Susan Athey, Guido Imbens, Thai Pham, and Stefan Wager. Estimating average treatment effects: Supplementary analyses and remaining challenges. *American Economic Review*, 107(5):278–81, 2017.

[12] Susan Athey and Guido W Imbens. The state of applied econometrics: Causality and policy evaluation. *Journal of Economic Perspectives*, 31(2):3–32, 2017.

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

[14] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations (ICLR)*, 2014.

[15] Heejung Bang and James M Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005.

[16] Magdalena Bennett, Juan Pablo Vielma, and Jose R Zubizarreta. Building representative matched samples with multi-valued treatments in large observational studies: Analysis of the impact of an earthquake on educational attainment. *arXiv preprint arXiv:1810.06707*, 2018.

[17] Mariusz Bojarski, Davide Del Testa, Daniel Dworakowski, Bernhard Firner, Beat Flepp, Prasoon Goyal, Lawrence D Jackel, Mathew Monfort, Urs Muller, Jiakai Zhang, Xin Zhang, Jake Zhao, and Karol Zieba. End to end learning for self-driving cars. *arXiv preprint arXiv:1604.07316*, 2016.

[18] Léon Bottou, Jonas Peters, Joaquin Quiñonero-Candela, Denis X Charles, D Max Chickering, Elon Portugaly, Dipankar Ray, Patrice Simard, and Ed Snelson. Counterfactual reasoning and learning systems: The example of computational advertising. *The Journal of Machine Learning Research*, 14(1):3207–3260, 2013.

[19] George EP Box, Gwilym M Jenkins, Gregory C Reinsel, and Greta M Ljung. *Time series analysis: forecasting and control*. John Wiley & Sons, 2015.

[20] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[21] M Alan Brookhart, Sebastian Schneeweiss, Kenneth J Rothman, Robert J Glynn, Jerry Avorn, and Til Stürmer. Variable selection for propensity score models. *American journal of epidemiology*, 163(12):1149–1156, 2006.

[22] Chun-Hao Chang, Elliot Creager, Anna Goldenberg, and David Kristjanson Duvenaud. Explaining image classifiers by counterfactual generation. In *International Conference on Learning Representations (ICLR)*, 2018.

[23] Ran Chen and Hanzhong Liu. Heterogeneous treatment effect estimation through deep learning. *arXiv preprint arXiv:1810.11010*, 2018.

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

[25] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

[26] Tianjiao Chu and Clark Glymour. Search for additive nonlinear time series causal models. *Journal of Machine Learning Research*, 9(May):967–991, 2008.

[27] Stephen R Cole and Constantine E Frangakis. The consistency statement in causal inference: a definition or an assumption? *Epidemiology*, 20(1):3–5, 2009.

[28] Stuart Coles, Joanna Bawa, Lesley Trenner, and Pat Dorazio. *An introduction to statistical modeling of extreme values*, volume 208. Springer, 2001.

[29] Diego Colombo, Marloes H Maathuis, Markus Kalisch, and Thomas S Richardson. Learning high-dimensional directed acyclic graphs with latent and selection variables. *The Annals of Statistics*, pages 294–321, 2012.

[30] Gregory F Cooper and Edward Herskovits. A bayesian method for the induction of probabilistic networks from data. *Machine learning*, 9(4):309–347, 1992.

[31] Richard K Crump, V Joseph Hotz, Guido W Imbens, and Oscar A Mitnik. Dealing with limited overlap in estimation of average treatment effects. *Biometrika*, 96(1):187–199, 2009.

[32] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in neural information processing systems*, pages 2292–2300, 2013.

[33] Marco Cuturi and Arnaud Doucet. Fast computation of wasserstein barycenters. In *International Conference on Machine Learning*, pages 685–693, 2014.

[34] George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314, 1989.

[35] A Philip Dawid. Causal inference without counterfactuals. *Journal of the American Statistical Association*, 95(450):407–424, 2000.

[36] Rajeev H Dehejia and Sadek Wahba. Propensity score-matching methods for nonexperimental causal studies. *Review of Economics and statistics*, 84(1):151–161, 2002.

[37] Mukeshwar Dhamala, Govindan Rangarajan, and Mingzhou Ding. Analyzing information flow in brain networks with nonparametric granger causality. *Neuroimage*, 41(2):354–362, 2008.

[38] Alexis Diamond and Jasjeet S Sekhon. Genetic matching for estimating causal effects: A general multivariate matching method for achiev-

ing balance in observational studies. *Review of Economics and Statistics*, 95(3):932–945, 2013.

[39] Andreia Dionísio, Rui Menezes, and Diana A Mendes. Entropy-based independence test. *Nonlinear Dynamics*, 44(1-4):351–357, 2006.

[40] Phil Dowe. *Physical causation*. Cambridge University Press, 2007.

[41] Sizhen Du, Guojie Song, and Haikun Hong. Collective causal inference with lag estimation. *Neurocomputing*, 323:299–310, 2019.

[42] Michael Eichler. *Causal inference in time series analysis*. na, 2012.

[43] Doris Entner and Patrik O Hoyer. On causal discovery from time series data using fci. *Probabilistic graphical models*, pages 121–128, 2010.

[44] RA Fisher. *The design of experiments*. Oliver & Boyd, 1935.

[45] Christian Fong, Chad Hazlett, Kosuke Imai, et al. Covariate balancing propensity score for a continuous treatment: application to the efficacy of political advertisements. *The Annals of Applied Statistics*, 12(1):156–177, 2018.

[46] Nir Friedman, Kevin Murphy, and Stuart Russell. Learning the structure of dynamic probabilistic networks. In *Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence*, pages 139–147. Morgan Kaufmann Publishers Inc., 1998.

[47] Michele Jonsson Funk, Daniel Westreich, Chris Wiesen, Til Stürmer, M Alan Brookhart, and Marie Davidian. Doubly robust estimation of causal effects. *American journal of epidemiology*, 173(7):761–767, 2011.

[48] Maxime Gasse, Alex Aussem, and Haytham Elghazel. An experimental comparison of hybrid algorithms for bayesian network structure learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 58–73. Springer, 2012.

[49] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[50] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[51] Omer Gottesman, Fredrik Johansson, Joshua Meier, Jack Dent, Donghun Lee, Srivatsan Srinivasan, Linying Zhang, Yi Ding, David Wihl, Xuefeng Peng, Jiayu Yao, Isaac Lage, Christopher Mosch, Li wei H. Lehman, Matthieu Komorowski, Matthieu Komorowski, Aldo Faisal, Leo Anthony Celi, David Sontag, and Finale Doshi-Velez. Evaluating reinforcement learning algorithms in observational health settings. *arXiv preprint arXiv:1805.12298*, 2018.

[52] Clive WJ Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society*, pages 424–438, 1969.

[53] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(Mar):723–773, 2012.

[54] Justin Grimmer, Solomon Messing, and Sean J Westwood. Estimating het-

erogeneous treatment effects and the effects of heterogeneous treatments with ensemble methods. *Political Analysis*, 25(4):413–434, 2017.

[55] Ruocheng Guo, Lu Cheng, Jundong Li, P Richard Hahn, and Huan Liu. A survey of learning causality with data: Problems and methods. *arXiv preprint arXiv:1809.09337*, 2018.

[56] Jens Hainmueller. Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis*, 20(1):25–46, 2012.

[57] Jason Hartford, Greg Lewis, Kevin Leyton-Brown, and Matt Taddy. Deep iv: A flexible approach for counterfactual prediction. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1414–1423. JMLR. org, 2017.

[58] Chad Hazlett. Kernel balancing: A flexible non-parametric weighting procedure for estimating causal effects. *Available at SSRN 2746753*, 2018.

[59] David Heckerman, Dan Geiger, and David M Chickering. Learning bayesian networks: The combination of knowledge and statistical data. *Machine learning*, 20(3):197–243, 1995.

[60] Christina Heinze-Deml, Marloes H Maathuis, and Nicolai Meinshausen. Causal structure learning. *Annual Review of Statistics and Its Application*, 5:371–391, 2018.

[61] Miguel A Hernan and James M Robins. *Causal inference*. Boca Raton: Chapman & HallCRC, forthcoming, 2019.

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

[63] Keisuke Hirano and Guido W Imbens. Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization. *Health Services and Outcomes research methodology*, 2(3-4):259–278, 2001.

[64] Julia Hirschberg and Christopher D Manning. Advances in natural language processing. *Science*, 349(6245):261–266, 2015.

[65] David A Hirshberg and José R Zubizarreta. On two approaches to weighting in causal inference. *Epidemiology*, 28(6):812–816, 2017.

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

[67] Kevin D Hoover. Causality in economics and econometrics. *The New Palgrave Dictionary of Economics: Volume 1–8*, pages 719–728, 2008.

[68] Daniel G Horvitz and Donovan J Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, 47(260):663–685, 1952.

[69] David Hume. *A treatise of human nature*. London: John Noon, 1739.

[70] Ferenc Huszár. Variational inference using implicit distributions. *arXiv preprint arXiv:1702.08235*, 2017.

[71] Aapo Hyvärinen, Kun Zhang, Shohei Shimizu, and Patrik O Hoyer. Estimation of a structural vector autoregression model using non-gaussianity. *Journal of Machine Learning Research*, 11(May):1709–1731, 2010.

[72] Kosuke Imai, Marc Ratkovic, et al. Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics*, 7(1):443–470, 2013.

[73] Guido W Imbens. Potential outcome and directed acyclic graph approaches to causality: Relevance for empirical practice in economics. *arXiv preprint arXiv:1907.07271*, 2019.

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

[75] Guido W Imbens and Jeffrey M Wooldridge. Recent developments in the econometrics of program evaluation. *Journal of economic literature*, 47(1):5–86, 2009.

[76] Edward L Ionides. Truncated importance sampling. *Journal of Computational and Graphical Statistics*, 17(2):295–311, 2008.

[77] Rafael Izbicki, Ann B Lee, and Taylor Pospisil. Abc–cde: Toward approximate bayesian computation with complex high-dimensional data and limited simulations. *Journal of Computational and Graphical Statistics*, pages 1–20, 2019.

[78] Vinay Jethava and Devdatt Dubhashi. Gans for life: generative adversarial networks for likelihood free inference. *arXiv preprint arXiv:1711.11139*, 2017.

[79] Thorsten Joachims, Adith Swaminathan, and Tobias Schnabel. Unbiased learning-to-rank with biased feedback. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, pages 781–789. ACM, 2017.

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

[81] Cheng Ju, David Benkeser, and Mark J van der Laan. Flexible collaborative estimation of the average causal effect of a treatment using the outcome-highly-adaptive lasso. *arXiv preprint arXiv:1806.06784*, 2018.

[82] Cheng Ju, Joshua Schwab, and Mark J van der Laan. On adaptive propensity score truncation in causal inference. *Statistical methods in medical research*, page 0962280218774817, 2018.

[83] Markus Kalisch and Peter Bühlmann. Estimating high-dimensional directed acyclic graphs with the pc-algorithm. *Journal of Machine Learning Research*, 8(Mar):613–636, 2007.

[84] Markus Kalisch and Peter Bühlmann. Causal structure learning and inference: a selective review. *Quality Technology & Quantitative Management*, 11(1):3–21, 2014.

[85] Nathan Kallus. A framework for optimal matching for causal inference. In *Artificial Intelligence and Statistics*, pages 372–381, 2017.

[86] Joseph DY Kang, Joseph L Schafer, et al. Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical science*, 22(4):523–539, 2007.

[87] Edward H Kennedy, Zongming Ma, Matthew D McHugh, and Dylan S Small. Non-parametric methods for doubly robust estimation of continuous treatment effects. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(4):1229–1245, 2017.

[88] Ronald C Kessler, Robert M Bossarte, Alex Luedtke, Alan M Zaslavsky, and Jose R Zubizarreta. Machine learning methods for developing precision treatment rules with observational data. *Behaviour research and therapy*, 120:103412, 2019.

[89] Chanmin Kim, Michael J Daniels, Bess H Marcus, and Jason A Roy. A framework for bayesian nonparametric inference for causal effects of mediation. *Biometrics*, 73(2):401–409, 2017.

[90] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

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

[92] Murat Kocaoglu, Christopher Snyder, Alexandros G Dimakis, and Sriram Vishwanath. Causalgan: Learning causal implicit generative models with adversarial training. In *International Conference on Learning Representations (ICLR)*, 2018.

[93] Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[94] Andreas Krämer, Jeff Green, Jack Pollard Jr, and Stuart Tugendreich. Causal analysis approaches in ingenuity pathway analysis. *Bioinformatics*, 30(4):523–530, 2013.

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

[96] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, pages 4066–4076, 2017.

[97] Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J

Gershman. Building machines that learn and think like people. *Behavioral and brain sciences*, 40, 2017.

[98] Robert J LaLonde. Evaluating the econometric evaluations of training programs with experimental data. *The American economic review*, pages 604–620, 1986.

[99] Dawen Liang, Laurent Charlin, and David M Blei. Causal inference for recommendation. In *Causation: Foundation to Application, Workshop at UAI*. AUAI, 2016.

[100] Roderick JA Little and Donald B Rubin. *Statistical analysis with missing data*, volume 793. Wiley, 2019.

[101] David Lopez-Paz. *From Dependence to Causation*. PhD thesis, Cambridge University, 2016.

[102] Christos Louizos, Uri Shalit, Joris M Mooij, David Sontag, Richard Zemel, and Max Welling. Causal effect inference with deep latent-variable models. In *Advances in Neural Information Processing Systems*, pages 6446–6456, 2017.

[103] Chaochao Lu, Bernhard Schölkopf, and José Miguel Hernández-Lobato. Deconfounding reinforcement learning in observational settings. *arXiv preprint arXiv:1812.10576*, 2018.

[104] Miguel Angel Luque-Fernandez, Michael Schomaker, Bernard Rachet, and Mireille E Schnitzer. Targeted maximum likelihood estimation for a binary treatment: A tutorial. *Statistics in medicine*, 37(16):2530–2546, 2018.

[105] Liang Ma, Jie Dong, and Kaixiang Peng. Root cause diagnosis of quality-

related faults in industrial multimode processes using robust gaussian mixture model and transfer entropy. *Neurocomputing*, 285:60–73, 2018.

[106] Edward BL Mackay, Peter G Challenor, and AbuBakr S Bahaj. A comparison of estimators for the generalised pareto distribution. *Ocean Engineering*, 38(11-12):1338–1346, 2011.

[107] David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Fairness through causal awareness: Learning causal latent-variable models for biased data. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 349–358. ACM, 2019.

[108] Sara Magliacane, Thijs van Ommen, Tom Claassen, Stephan Bongers, Philip Versteeg, and Joris M Mooij. Domain adaptation by using causal inference to predict invariant conditional distributions. In *Advances in Neural Information Processing Systems*, pages 10846–10856, 2018.

[109] Marc Maier, Katerina Marazopoulou, and David Jensen. Reasoning about independence in probabilistic models of relational data. *arXiv preprint arXiv:1302.4381*, 2013.

[110] Dimitris Margaritis and Sebastian Thrun. Bayesian network induction via local neighborhoods. *Advances in Neural Information Processing Systems*, 12:505–511, 1999.

[111] Daniel F McCaffrey, Greg Ridgeway, and Andrew R Morral. Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological methods*, 9(4):403, 2004.

[112] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[113] Christopher Meek. Toward learning graphical and causal process models. In *Proceedings of the UAI 2014 Conference on Causal Inference: Learning and Prediction-Volume 1274*, pages 43–48. CEUR-WS. org, 2014.

[114] Lars Mescheder, Sebastian Nowozin, and Andreas Geiger. Adversarial variational bayes: Unifying variational autoencoders and generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2391–2400. JMLR. org, 2017.

[115] FH Messerli. Chocolate consumption, cognitive function, and nobel laureates. *The New England journal of medicine*, 367(16):1562–1564, 2012.

[116] Shakir Mohamed and Balaji Lakshminarayanan. Learning in implicit generative models. *arXiv preprint arXiv:1610.03483*, 2016.

[117] Joris M Mooij, Dominik Janzing, Tom Heskes, and Bernhard Schölkopf. On causal discovery with cyclic additive noise models. In *Advances in neural information processing systems*, pages 639–647, 2011.

[118] Mary S Morgan, Tarja Knuuttila, et al. Models and modelling in economics. *Handbook of the Philosophy of Economics*, pages 49–87, 2012.

[119] 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.

[120] Kevin Patrick Murphy. *Dynamic Bayesian Networks: Representation, Inference and Learning*. PhD thesis, University of California, Berkeley, 2002.

[121] Tanmayee Narendra, Anush Sankaran, Deepak Vijaykeerthy, and Senthil

Mani. Explaining deep learning models using causal inference. *arXiv preprint arXiv:1811.04376*, 2018.

[122] Kazuki Natori, Masaki Uto, and Maomi Ueno. Consistent learning bayesian networks with thousands of variables. In *Advanced Methodologies for Bayesian Networks*, pages 57–68, 2017.

[123] Richard E Neapolitan et al. *Learning bayesian networks*, volume 38. Pearson Prentice Hall Upper Saddle River, NJ, 2004.

[124] Rachel C Nethery, Fabrizia Mealli, and Francesca Dominici. Estimating population average causal effects in the presence of non-overlap: A bayesian approach. *arXiv preprint arXiv:1805.09736*, 2018.

[125] Romain Neugebauer and Mark van der Laan. Why prefer double robust estimators in causal inference? *Journal of statistical planning and inference*, 129(1-2):405–426, 2005.

[126] Xinkun Nie and Stefan Wager. Quasi-oracle estimation of heterogeneous treatment effects. *arXiv preprint arXiv:1712.04912*, 2017.

[127] Miruna Oprescu, Vasilis Syrgkanis, and Zhiwei Steven Wu. Orthogonal random forest for heterogeneous treatment effect estimation. *arXiv preprint arXiv:1806.03467*, 2018.

[128] Art B Owen. *Monte Carlo theory, methods and examples (book draft)*. 2014.

[129] Michal Ozery-Flato, Pierre Thodoroff, and Tal El-Hay. Adversarial balancing for causal inference. *arXiv preprint arXiv:1810.07406*, 2018.

[130] Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, 1988.

[131] Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge university press, 2009.

[132] Judea Pearl. The do-calculus revisited. In *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence*, pages 3–11. AUAI Press, 2012.

[133] Judea Pearl. Theoretical impediments to machine learning with seven sparks from the causal revolution. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pages 3–3. ACM, 2018.

[134] Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. *Causal inference in statistics: A primer*. John Wiley & Sons, 2016.

[135] Judea Pearl and Dana Mackenzie. *The book of why: the new science of cause and effect*. Basic Books, 2018.

[136] Judea Pearl and Thomas S Verma. A theory of inferred causation. In *Studies in Logic and the Foundations of Mathematics*, volume 134, pages 789–811. Elsevier, 1995.

[137] Joel Persson. *Counterfactual Prediction Methods for Causal Inference in Observational Studies with Continuous Treatments*. PhD thesis, Lund University, 2019.

[138] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. Cambridge, MA: The MIT Press, 2017.

[139] Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

[140] James Pickands III et al. Statistical inference using extreme order statistics. *the Annals of Statistics*, 3(1):119–131, 1975.

[141] Scott Powers, Junyang Qian, Kenneth Jung, Alejandro Schuler, Nigam Haresh Shah, Trevor J. Hastie, and Robert Tibshirani. Some methods for heterogeneous treatment effect estimation in high dimensions. *Statistics in medicine*, 37 11:1767–1787, 2018.

[142] Min Qian and Susan A Murphy. Performance guarantees for individualized treatment rules. *Annals of statistics*, 39(2):1180, 2011.

[143] Joseph Ramsey, Jiji Zhang, and Peter L Spirtes. Adjacency-faithfulness and conservative causal inference. *arXiv preprint arXiv:1206.6843*, 2012.

[144] Carl Edward Rasmussen. Gaussian processes in machine learning. In *Summer School on Machine Learning*, pages 63–71. Springer, 2003.

[145] Thomas Richardson, Peter Spirtes, et al. Ancestral graph markov models. *The Annals of Statistics*, 30(4):962–1030, 2002.

[146] James Robins. A new approach to causal inference in mortality studies with a sustained exposure period – application to control of the healthy worker survivor effect. *Mathematical modelling*, 7(9-12):1393–1512, 1986.

[147] James M Robins. The analysis of randomized and non-randomized aids treatment trials using a new approach to causal inference in longitudinal studies. *Health service research methodology: a focus on AIDS*, pages 113–159, 1989.

[148] James M Robins. Causal inference from complex longitudinal data. In *Latent variable modeling and applications to causality*, pages 69–117. Springer, 1997.

[149] James M Robins, Miguel Angel Hernán, and Babette Brumback. Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5):551, 2000.

[150] James M Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, 89(427):846–866, 1994.

[151] M Rojas-Carulla, M Baroni, and D Lopez-Paz. Causal discovery using proxy variables. In *International Conference on Learning Representations (ICLR)*. OpenReview. net, 2018.

[152] Paul R Rosenbaum. Observational study. *Encyclopedia of statistics in behavioral science*, 2005.

[153] Paul R Rosenbaum. Modern algorithms for matching in observational studies. *Annual Review of Statistics and Its Application*, 7, 2019.

[154] Paul R Rosenbaum et al. *Design of observational studies*, volume 10. Springer, 2010.

[155] 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.

[156] Andrea Rotnitzky, James M Robins, and Daniel O Scharfstein. Semiparametric regression for repeated outcomes with nonignorable nonresponse. *Journal of the american statistical association*, 93(444):1321–1339, 1998.

[157] Donald B Rubin. Matching to remove bias in observational studies. *Biometrics*, pages 159–183, 1973.

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

[159] Donald B Rubin. Comment: Which ifs have causal answers. *Journal of the American Statistical Association*, 81(396):961–962, 1986.

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

[161] Donald B Rubin. *Matched sampling for causal effects*. Cambridge University Press, 2006.

[162] J Runge. Causal network reconstruction from time series: From theoretical assumptions to practical estimation. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 28(7):075310, 2018.

[163] Bertrand Russell. *Human knowledge, its scope and limits.* Simon & Schuster, 1948.

[164] Karen Sachs, Omar Perez, Dana Pe'er, Douglas A Lauffenburger, and Garry P Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721):523–529, 2005.

[165] Ramon Sangüesa and Ulises Cortés. Learning causal networks from data: a survey and a new algorithm for recovering possibilistic causal networks. *AI Communications*, 10(1):31–61, 1997.

[166] Richard L Scheaffer, William Mendenhall III, R Lyman Ott, and Kenneth G Gerow. *Elementary survey sampling*. Cengage Learning, 2011.

[167] Tobias Schnabel, Adith Swaminathan, Ashudeep Singh, Navin Chandak, and Thorsten Joachims. Recommendations as treatments: Debiasing learning and evaluation. In *International Conference on Machine Learning*, pages 1670–1679, 2016.

[168] B Schölkopf, D Janzing, J Peters, E Sgouritsa, K Zhang, and J Mooij. On causal and anticausal learning. In *29th International Conference on Machine Learning (ICML 2012)*, pages 1255–1262. International Machine Learning Society, 2012.

[169] Peter Schulam and Suchi Saria. Reliable decision support using counterfactual models. In *Advances in Neural Information Processing Systems*, pages 1697–1708, 2017.

[170] Alejandro Schuler, Michael Baiocchi, Robert Tibshirani, and Nigam Shah. A comparison of methods for model selection when estimating individual treatment effects. *arXiv preprint arXiv:1804.05146*, 2018.

[171] Megan S Schuler and Sherri Rose. Targeted maximum likelihood estimation for causal inference in observational studies. *American journal of epidemiology*, 185(1):65–73, 2017.

[172] Patrick Schwab, Lorenz Linhardt, Stefan Bauer, Joachim M Buhmann, and Walter Karlen. Learning counterfactual representations for estimating individual dose-response curves. *arXiv preprint arXiv:1902.00981*, 2019.

[173] Gideon Schwarz et al. Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464, 1978.

[174] Marco Scutari, Claudia Vitolo, and Allan Tucker. Learning bayesian networks from big data with greedy search: computational complexity and efficient implementation. *Statistics and Computing*, pages 1–14, 2019.

[175] Shaun R Seaman and Ian R White. Review of inverse probability weighting for dealing with missing data. *Statistical methods in medical research*, 22(3):278–295, 2013.

[176] William R Shadish, Thomas D Cook, and Donald T Campbell. *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin, 2002.

[177] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[178] Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3076–3085. JMLR. org, 2017.

[179] Claudia Shi, David M Blei, and Victor Veitch. Adapting neural networks for the estimation of treatment effects. *arXiv preprint arXiv:1906.02120*, 2019.

[180] Jiaxin Shi, Shengyang Sun, and Jun Zhu. Kernel implicit variational inference. *arXiv preprint arXiv:1705.10119*, 2017.

[181] Susan M Shortreed and Ashkan Ertefaie. Outcome-adaptive lasso: Variable selection for causal inference. *Biometrics*, 73(4):1111–1122, 2017.

[182] Christopher A Sims. Macroeconomics and reality. *Econometrica: journal of the Econometric Society*, pages 1–48, 1980.

[183] Peter Spirtes. Introduction to causal inference. *Journal of Machine Learning Research*, 11(May):1643–1662, 2010.

[184] Peter Spirtes, Clark N Glymour, Richard Scheines, David Heckerman, Christopher Meek, Gregory Cooper, and Thomas Richardson. *Causation, prediction, and search*. MIT press, 2000.

[185] Peter Spirtes and Christopher Meek. Learning bayesian networks with discrete variables from data. In *KDD*, volume 1, pages 294–299, 1995.

[186] Bharath K Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Bernhard Schölkopf, Gert RG Lanckriet, et al. On the empirical estimation of integral probability metrics. *Electronic Journal of Statistics*, 6:1550–1599, 2012.

[187] Bradly C Stadie, Sören R Künzel, Nikita Vemuri, and Jasjeet S Sekhon. Estimating heterogeneous treatment effects using neural networks with the y-learner. 2018.

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

[189] Liangjun Su and Xun Lu. Nonparametric dynamic panel data models: kernel estimation and specification testing. *Journal of Econometrics*, 176(2):112–133, 2013.

[190] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[191] Adith Swaminathan and Thorsten Joachims. Batch learning from logged bandit feedback through counterfactual risk minimization. *Journal of Machine Learning Research*, 16(1):1731–1755, 2015.

[192] Adith Swaminathan and Thorsten Joachims. Counterfactual risk mini-

mization: Learning from logged bandit feedback. In *International Conference on Machine Learning*, pages 814–823, 2015.

[193] Dustin Tran and David M. Blei. Implicit causal models for genome-wide association studies. *CoRR*, abs/1710.10742, 2018.

[194] Dustin Tran, Rajesh Ranganath, and David Blei. Hierarchical implicit models and likelihood-free variational inference. In *Advances in Neural Information Processing Systems*, pages 5523–5533, 2017.

[195] Ioannis Tsamardinos, Laura E Brown, and Constantin F Aliferis. The max-min hill-climbing bayesian network structure learning algorithm. *Machine learning*, 65(1):31–78, 2006.

[196] Jodie B Ullman and Peter M Bentler. Structural equation modeling. *Handbook of Psychology, Second Edition*, 2, 2012.

[197] Mark J van der Laan. Targeted estimation of nuisance parameters to obtain valid statistical inference. *The international journal of biostatistics*, 10(1):29–57, 2014.

[198] Mark J Van der Laan, MJ Laan, and James M Robins. *Unified methods for censored longitudinal data and causality*. Springer Science & Business Media, 2003.

[199] Mark J Van der Laan and Sherri Rose. *Targeted learning: causal inference for observational and experimental data*. Springer Science & Business Media, 2011.

[200] Mark J Van der Laan and Sherri Rose. *Targeted Learning in Data Science*. Springer, 2018.

[201] Tyler J VanderWeele. Concerning the consistency assumption in causal inference. *Epidemiology*, 20(6):880–883, 2009.

[202] Aki Vehtari, Andrew Gelman, and Jonah Gabry. Pareto smoothed importance sampling. *arXiv preprint arXiv:1507.02646*, 2015.

[203] Aki Vehtari, Andrew Gelman, and Jonah Gabry. Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and Computing*, 27(5):1413–1432, 2017.

[204] Victor Veitch, Dhanya Sridhar, and David M Blei. Using text embeddings for causal inference. *arXiv preprint arXiv:1905.12741*, 2019.

[205] Mark Voortman, Denver Dash, and Marek J Druzdzel. Learning why things change: the difference-based causality learner. In *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence*, pages 641–650. AUAI Press, 2010.

[206] Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.

[207] Xuanhui Wang, Michael Bendersky, Donald Metzler, and Marc Najork. Learning to rank with selection bias in personal search. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 115–124. ACM, 2016.

[208] Harry Wechsler, Jonathon P Phillips, Vicki Bruce, Françoise Fogelman Soulié, and Thomas S Huang. *Face recognition: From theory to applications*, volume 163. Springer Science & Business Media, 2012.

[209] Sebastian Weichwald, Arthur Gretton, Bernhard Scholkopf, and Moritz Grosse-Wentrup. Recovery of non-linear cause-effect relationships from linearly mixed neuroimaging data. In *2016 International Workshop on Pattern Recognition in Neuroimaging (PRNI)*, pages 1–4. IEEE, 2016.

[210] Xiao Wu, Fabrizia Mealli, Marianthi-Anna Kioumourtzoglou, Francesca Dominici, and Danielle Braun. Matching on generalized propensity scores with continuous exposures. *arXiv preprint arXiv:1812.06575*, 2018.

[211] S Yang and P Ding. Asymptotic inference of causal effects with observational studies trimmed by the estimated propensity scores. *Biometrika*, 105(2):487–493, 2018.

[212] Liuyi Yao, Sheng Li, Yaliang Li, Mengdi Huai, Jing Gao, and Aidong Zhang. Representation learning for treatment effect estimation from observational data. In *Advances in Neural Information Processing Systems*, pages 2633–2643, 2018.

[213] Jinsung Yoon, James Jordon, and Mihaela van der Schaar. Ganite: Estimation of individualized treatment effects using generative adversarial nets. In *International Conference on Learning Representations (ICLR)*, 2018.

[214] Hector Zenil, Narsis A Kiani, Allan A Zea, and Jesper Tegnér. Causal deconvolution by algorithmic generative models. *Nature Machine Intelligence*, 1(1):58, 2019.

[215] Jiji Zhang. *Causal inference and reasoning in causally insufficient systems*. PhD thesis, Carnegie Mellon University, 2006.

[216] Jin Zhang and Michael A Stephens. A new and efficient estimation method

for the generalized pareto distribution. *Technometrics*, 51(3):316–325, 2009.

[217] Junzhe Zhang and Elias Bareinboim. Fairness in decision-making: the causal explanation formula. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[218] Kun Zhang, Mingming Gong, and Bernhard Schölkopf. Multi-source domain adaptation: A causal view. In *Twenty-ninth AAAI conference on artificial intelligence*, 2015.

[219] Kun Zhang, Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Kernel-based conditional independence test and application in causal discovery. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, pages 804–813. AUAI Press, 2011.

[220] Kun Zhang, Bernhard Schölkopf, Peter Spirtes, and Clark Glymour. Learning causality and causality-related learning: some recent progress. *National science review*, 5(1):26–29, 2017.

[221] Qingyuan Zhao. *Topics in Causal and High Dimensional Inference*. PhD thesis, Stanford University, 2016.

[222] Qingyuan Zhao and Daniel Percival. Entropy balancing is doubly robust. *Journal of Causal Inference*, 5(1), 2017.

[223] Sendong Zhao, Ting Liu, Sicheng Zhao, Yiheng Chen, and Jian-Yun Nie. Event causality extraction based on connectives analysis. *Neurocomputing*, 173:1943–1950, 2016.

[224] Yang Zhou. Structure learning of probabilistic graphical models: a comprehensive survey. *arXiv preprint arXiv:1111.6925*, 2011.

[225] Fujin Zhu, Adi Lin, Guangquan Zhang, and Jie Lu. Counterfactual inference with hidden confounders using implicit generative models. In *Australasian Joint Conference on Artificial Intelligence*, pages 519–530. Springer, 2018.

[226] Fujin Zhu, Guangquan Zhang, Jie Lu, and Donghua Zhu. First-order causal process for causal modelling with instantaneous and cross-temporal relations. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 380–387. IEEE, 2017.

[227] Yeying Zhu, Jennifer S Savage, and Debashis Ghosh. A kernel-based metric for balance assessment. *Journal of causal inference*, 6(2), 2018.

[228] Cunlu Zou and Jianfeng Feng. Granger causality vs. dynamic bayesian network inference: a comparative study. *BMC bioinformatics*, 10(1):122, 2009.