Causal Inference in the Presence of Heterogeneous Treatment Effects

by

Wei Liang

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Statistics

Waterloo, Ontario, Canada, 2025

© Wei Liang 2025

Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

External Examiner:	Erica Moodie Professor, Dept. of Epidemiology, Biostatistics, and Occupational Health, McGill University
Supervisor:	Changbao Wu Professor
Internal Members:	Pengfei Li Professor
	Glen McGee Assistant Professor
	Yeying Zhu Associate Professor
Internal-External Member:	Ashok Chaurasia Associate Professor, School of Public Heath Sciences

Author's Declaration

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Statement of Contributions

A manuscript based on the work in Chapter 3 of this thesis has been submitted to a journal and is currently under review. Chapter 4 is based on the preprint arXiv:2505.12209. The work presented in Chapter 5 is currently being prepared for submission to a statistical journal. All three papers are co-authored with my thesis supervisor, who contributed to discussions on the research problems and provided feedback on the initial drafts during the writing of these papers. I am responsible for theoretical developments, simulation studies, real world data applications, and the initial drafts.

Abstract

Causal inference has been widely accepted as a statistical tool in various areas for demystifying causality from data. Treatment effect heterogeneity is a common issue in causal inference which refers to variation in the causal effect of a treatment across different subgroups or individuals within a population. This thesis explores three topics in causal inference in the presence of heterogeneous treatment effects, aiming to provide some insights for this critical issue.

Chapter 2 introduces basic notation, frameworks, models, and parameters in causal inference, serving as preliminary material for the three topics studied in Chapters 3 - 5, with a focus on the Rubin causal model.

In Chapter 3, we discuss the first topic: causal inference with survey data. In the presence of heterogeneous treatment effects, a causal conclusion based on sample data may not generalize to a broader population if selection bias exists. We propose estimators for population average treatment effects by incorporating survey weights into the propensity score weighting approach to simultaneously mitigate confounding bias and selection bias. A robust sandwich variance estimator is developed to permit valid statistical inference for the population-level causal parameters under a proposed "two-phase randomization model" framework. The proposed estimators and associated inferential procedure are shown to be robust against model misspecifications. We further extend our results to observational non-probability survey samples and demonstrate how to combine auxiliary population information from multiple external reference probability samples for more reliable estimation. We illustrate our proposed methods through Monte Carlo simulation studies and the analysis of a real-world survey dataset.

Chapter 4 explores the second topic: estimation of treatment harm rate (THR), the proportion of individuals in a population who are negatively affected by a treatment. The THR is a measure of treatment risk and reveals the treatment effect heterogeneity within a subpopulation. However, the measure is generally non-identifiable even when the treatments are randomly assigned, and existing works focus primarily on the estimation of the THR under either untestable identification or ambiguous model assumptions. We develop a class of partitioning-based bounds for the THR with data from randomized controlled trials with two distinct features: Our proposed bounds effectively use available auxiliary covariates information and they can be consistently estimated without relying on any untestable or ambiguous model assumptions. Our methods are motivated from a key observation that the sharp bounds of the THR can be attained under a partition of the covariates space with at most four cells. Probabilistic classification algorithms are

employed to estimate nuisance parameters to realize the partitioning. The resulting interval estimators of the THR are model-assisted in the sense that they are highly efficient when the underlying models are well fitted, while their validity relies solely on the randomization of the trials. Finite sample performances of our proposed interval estimators along with a conservatively extended confidence interval for the THR are evaluated through Monte Carlo simulation studies. An application of the proposed methods to the ACTG 175 data is presented. A Python package named `partbte` for the partitioning-based algorithm has been developed and is available on https://github.com/w62liang/partition-te.

Chapter 5 investigates the third topic: causal mediation analysis in randomized controlled trials with noncompliance. The average causal mediation effect (ACME) and the natural direct effect (NDE) are two parameters of primary interest in causal mediation analysis. However, the two causal parameters are not identifiable in randomized controlled trials in the presence of mediator-outcome confounding and assignment-treatment noncompliance. In such scenarios, we explore partial identification of parameters and derive nonparametric bounds on the ACME and the NDE when the treatment assignment serves as an instrumental variable. The nonparametric sharp bounds for the local causal parameters defined on the subpopulation of treatment-assignment compliers are also provided. We demonstrate the practical application of the proposed bounds through an empirical analysis of a large-scale randomized online advertising dataset.

The thesis concludes in Chapter 6 with a brief summary and discussions of future work. Technical details, including the proofs of key propositions and theorems as well as additional simulation results, are provided at the end of each chapter.

Acknowledgements

I want to express my deepest gratitude to my supervisor, Dr. Changbao Wu, for his invaluable guidance and constant encouragement throughout the course of my research. His comments on my writing have significantly enhanced the quality of this thesis. He is remarkably generous in providing financial support, which allowed me to live and study in Canada without financial concerns. I have benefited greatly from our conversations over the past four years, which have imparted life lessons that extend far beyond the realm of academic research.

I would like to thank Dr. Erica Moodie, Dr. Pengfei Li, Dr. Glen McGee, Dr. Yeying Zhu, and Dr. Ashok Chaurasia for serving on my thesis examination committee. Pengfei Li and Yeying Zhu were instructors of courses I took in the first year of my PhD program. I want to thank them, as well as other faculty members, for providing me with invaluable instruction and knowledge in statistics. I extend my appreciation to my master's supervisor, Dr. Ying Yan, and my undergraduate supervisors, Dr. Lichun Wang and Dr. Liqun Wang, for their guidance that led me to the field of statistics and causal inference.

I am also grateful to Mary Lou Dufton, Greg Preston, Carla Daniels, Carlos Mendes, Shaleen Mathur, and many others for their administrative assistance. They are lovely people who have helped make the department a better place.

I feel fortunate to have some friends who have supported me in overcoming the challenges of living in a new country with different cultures. I want to express my heartfelt thanks to Chuanlong, who is incredibly selfless and always willing to lend a hand. I sincerely hope he will one day live the life he loves. I am also thankful to Xialing, Zhenggao, Xisi, Yuliang, Jingyue, Zhaoran, Liyuan, Chi, and many others for their help in daily life.

Finally, I wish to thank my parents for everything they have done for me.

Dedication

To my beloved family: Xiang, Qiutan, and Yucheng.

In memory of my grandparents.

"Life is like a boat."

— Rie fu

Table of Contents

Examining Committee Membership ii

Author's Declaration iii

Statement of Contributions iv

Abstract v

Acknowledgements vii

Dedication viii

List of Figures xiv

List of Tables xvi

List of Acronyms xviii

1 Introduction 1

2 Causal Inference: Frameworks, Models, and Parameters 6

2.1 The Rubin Causal Model and Potential Outcomes 6

2.2 Treatment Assignment Mechanisms 8

2.2.1 Randomized Experiments 9

2.2.2 Observational Studies . . . 9
2.2.3 Randomized Controlled Trials with Noncompliance: Instrumental Variables . . . 10
2.3 Causal Parameters and Identification . . . 11
2.3.1 Average Treatment Effects . . . 11
2.3.2 Treatment Harm Rate . . . 13
2.3.3 Natural Direct and Indirect Effects . . . 15
2.4 The Balancing Covariates Framework . . . 16

3 Propensity Score Weighting with Survey Data **18**
3.1 Introduction . . . 18
3.2 Problem Setup . . . 21
3.3 Survey-Weighted Propensity Score Weighting . . . 23
3.3.1 Two-Step Estimation . . . 23
3.3.2 SWPSW with Maximum Likelihood Estimator . . . 25
3.3.3 Balancing Covariates via SWPSW . . . 26
3.3.4 Variance Estimation . . . 29
3.3.5 Numerical Studies . . . 30
3.4 Survey-Weighted Propensity Score Weighting with Non-Probability Samples 36
3.4.1 Three-Step Estimation . . . 36
3.4.2 Identification with Partially Observed Covariates . . . 37
3.4.3 Estimation of Participation Probabilities Using Multiple Reference Samples . . . 39
3.4.4 Variance Estimation . . . 40
3.4.5 Numerical Studies with Non-Probability Survey Samples . . . 42
3.5 Real Data Analysis . . . 47
3.6 Some Additional Remarks . . . 53
3.7 Proofs and Technical Details . . . 54

3.7.1 Derivation of the Dual Problems . . . 54
3.7.2 Proof of Proposition 3.3.1 . . . 56
3.7.3 Proof of Proposition 3.4.1 . . . 58
3.7.4 Optimal Choice of the Tuning Parameters w_{kj} . . . 59

4 Estimation of Treatment Harm Rate via Partitioning 60
4.1 Introduction . . . 60
4.2 Partitioning-Based Bounds on Treatment Harm Rate . . . 63
4.3 Non-Parametric Estimation . . . 66
4.4 Space Partitioning . . . 68
4.4.1 Surrogate Partition . . . 68
4.4.2 Probabilistic Classifiers . . . 69
4.4.3 Cross-Fitting . . . 70
4.5 Assessment of Estimated Bounds . . . 71
4.6 Numerical Experiments . . . 74
4.7 Analysis of the AIDS Clinical Trials Group Study 175 Data . . . 80
4.8 Concluding Remarks . . . 82
4.9 Discussion and An Alternative Formulation . . . 83
4.10 Proofs and Technical Details . . . 85
4.10.1 Proofs of Main Propositions . . . 85
4.10.2 Model Calibration and Isotonic Regression . . . 89
4.10.3 A Brief Review of Probabilistic Classification Algorithms . . . 91
4.10.4 Additional Simulation Results . . . 95

5 Nonparametric Bounds in Mediation Analysis in the Presence of Unmeasured Confounding and Imperfect Compliance 105
5.1 Introduction . . . 105
5.2 Problem Setup . . . 108

5.2.1 Data Generating Mechanism . . . 108
5.2.2 Parameters of Interest . . . 110
5.2.3 Identification of Causal Parameters . . . 111
5.2.4 Additional Monotonicity Assumptions . . . 111
5.3 Bounds on Population Parameters . . . 112
5.4 Bounds on Local Parameters . . . 115
5.5 Analysis of the Criteo Uplift Dataset . . . 117
5.6 Concluding Remarks . . . 120
5.7 Proofs and Technical Details . . . 120
5.7.1 Derivation of the Bounds . . . 120
5.7.2 Bounds on Population Parameters under Monotonicity . . . 123
5.7.3 Bounds on Local Parameters under Monotonicity . . . 130

6 Future Work and Discussion **135**
6.1 Summary . . . 135
6.2 Future Work . . . 136
6.2.1 Variable Selection for Causal Inference with Nonprobability Samples 137
6.2.2 Partitioning-Based Bounds for General Causal Parameters on Binary Outcomes in RCTs . . . 138

References **141**

List of Figures

3.1 Boxplot of the health expenditures in the population and in the sub-population 50

3.2 Boxplot of the log health expenditures for different racial groups 50

3.3 Covariate-balance among groups using the absolute standardized mean difference. The weights are obtained for estimating the SATE. 51

3.4 Covariate-balance among groups using the absolute standardized mean difference. The weights are obtained for estimating the PATE with log-scaled outcomes. 51

4.1 Naive and Oracle bounds, estimated plug-in and partitioning-based bounds using the RF algorithm, and their corresponding 75% confidence bands. X-axis denotes the value of σ_ε. The plot for Scenario 1 is presented in the left panel, and the plot for Scenario 2 is presented in the right panel. 78

4.2 Naive and Oracle bounds, estimated plug-in and partitioning-based bounds using the RF algorithm combined with model calibration, and their corresponding 75% confidence bands. X-axis denotes the value of σ_ε. The plot for Scenario 1 is presented in the left panel, and the plot for Scenario 2 is presented in the right panel. 79

4.3 Naive and Oracle bounds, estimated plug-in and partitioning-based bounds using the KNN algorithm, and their corresponding 75% confidence bands. X-axis denotes the value of σ_ε. The plot for Scenario 1 is presented in the left panel, and the plot for Scenario 2 is presented in the right panel. . . . 103

4.4 Naive and Oracle bounds, estimated plug-in and partitioning-based bounds using the KNN algorithm combined with model calibration, and their corresponding 75% confidence bands. X-axis denotes the value of σ_ε. The plot for Scenario 1 is presented in the left panel, and the plot for Scenario 2 is presented in the right panel. 104

5.1 Feasible causal diagrams of the data generating mechanism under Assumptions 5.2.1 - 5.2.3. 110

List of Tables

2.1 Individualized treatment effect regarding different values of the potential outcomes . . . 13

2.2 Two hypothetical scenarios with different THR that yield the same observed RCT data and PATE. . . . 14

3.1 Simulation results for estimating the PATE with probability survey samples. 33

3.2 Simulation results for estimating the PATT with probability survey samples. 34

3.3 Simulation results for estimating the SATE with probability survey samples. 35

3.4 Designs for simulations with non-probability survey samples . . . 44

3.5 Simulation results with non-probability survey samples under Design I. . . 45

3.6 Simulation results with non-probability survey samples under Design II. . 46

3.7 Estimates and 95% confidence intervals of the racial disparity in health expenditures between Hispanics and non-Hispanic Whites. . . . 52

3.8 Estimates and 95% confidence intervals of the racial disparity in log health expenditures between Hispanics and non-Hispanic Whites. . . . 52

4.1 Monte Carlo simulation results for the interval estimators with $n = 500$. All values except the percentages in the table are presented in units scaled by 10^{-2}. . . . 77

4.2 Interval estimates of the treatment harm rate in ACTG 175 and 75% confidence intervals for the partitioning-based upper and lower bounds under two-fold cross-fitting. . . . 81

4.3 Interval estimates of the treatment harm rate in ACTG 175 and 75% confidence intervals for the partitioning-based upper and lower bounds under five-fold cross-fitting. . . . 82

4.4 Monte Carlo simulation results for the interval estimation under **Scenario 1**. All values except the percentages in the table are presented in units scaled by 10^{-2}. 97

4.5 Monte Carlo simulation results for the interval estimation under **Scenario 2**. All values except the percentages in the table are presented in units scaled by 10^{-2}. 98

4.6 Monte Carlo simulation results for the confidence intervals under **Scenario 1**. All values except the percentages in the table are presented in units scaled by 10^{-2}. 99

4.7 Monte Carlo simulation results for the confidence intervals under **Scenario 2**. All values except the percentages in the table are presented in units scaled by 10^{-2}. 99

4.8 Monte Carlo simulation results on the extended confidence intervals with $\alpha = 0.25$. All values except the percentages in the table are presented in units scaled by 10^{-2}. 100

4.9 Monte Carlo simulation results on the interval estimation combined with model calibration under **Scenario 1**. All values except the percentages in the table are presented in units scaled by 10^{-2}. 101

4.10 Monte Carlo simulation results on the interval estimation combined with model calibration under **Scenario 2**. All values except the percentages in the table are presented in units scaled by 10^{-2}. 102

5.1 Interpretation of the variables for the Criteo uplift data 118

5.2 Bounds on the population parameters under different assumptions. 119

5.3 Bounds on the local parameters under different assumptions. 120

5.4 The values of o_{yi} for $i = 0, \cdots, 15$. 121

5.5 The values of o_{mi} for $i = 0, 1, 2, 3$. 121

List of Acronyms

ACME	Average Causal Mediation Effect
CATE	Conditional Average Treatment Effect
CBPSW	Covariate Balancing Propensity Score Weighting
CMA	Causal Mediation Analysis
CRE	Completely Randomized Experiment
IV	Instrumental Variable
NDE	Natural Direct Effect
NIE	Natural Indirect Effect
PATE	Population Average Treatment Effect
PATT	Population Average Treatment Effect on the Treated
PBB	Partitioning-Based Bound
PS	Propensity Score
PSWMLE	Propensity Score Weighting with Maximum Likelihood Estimator
PWATE	Population Weighted Average Treatment Effect
RCM	Rubin Causal Model
RCT	Randomized Controlled Trial
SATE	Subpopulation Average Treatment Effect

SATT	Subpopulation Average Treatment Effect on the Treated
SWCBPSW	Survey-Weighted Covariate Balancing Propensity Score Weighting
SWPSW	Survey-Weighted Propensity Score Weighting
SWPSWMLE	Survey-Weighted Propensity Score Weighting with Maximum Likelihood Estimator
TAM	Treatment Assignment Mechanism
TEH	Treatment Effect Heterogeneity
THR	Treatment Harm Rate
TPR	Two-Phase Randomization

Chapter 1

Introduction

Causality is mysterious and important. It concerns causes and effects among events and variables, through the understanding of which we approach the essence of a system. Let us begin with a simple question: why does an apple fall from a tree? Owing to Issac Newton, we all know today that an apple falls due to gravity, the force exerted by the Earth, and the magnitude of the force relies on the Earth's mass and the radius. Why does iron rust when exposed to moist air? The causes of iron rusting are the presence of oxygen and water, which together initiate an oxidation reaction. The causal effect can be quantitatively described by the corresponding chemical equation. There is causality behind every phenomenon in our universe. Seeking causality is fundamentally a pursuit of truth.

Counterfactual thinking is a psychological approach to reasoning about causality. Consider the well-known causal statement from Imbens and Rubin (2015): "My headache went away because I took an aspirin". Counterfactual thinking involves the conceptual thought of "What if I had not taken the aspirin" or "What would my headache have been like had

I not taken the aspirin". The counterfactual outcome justifies whether "took aspirin" is the cause of "my headache went away" and what the causal effect is. However, counterfactual thinking remains a conceptual and psychological exercise. In practice, we can not verify a causal claim solely through counterfactual reasoning, as counterfactual outcomes are fundamentally unobservable.

Humanity's earliest practical attempts to verify causal hypotheses may be traced back to controlled experiments where certain variables that contribute to the dependent variable of interest are controlled experimentally to reduce experimental bias. Pioneered by Fisher's foundational book *The Design of Experiments* in the early 20th century, randomized experiments have been recognized as the golden rule for controlling confounding factors and drawing valid causal conclusions. Since then, a wide range of randomization techniques and corresponding theoretical frameworks have been extensively developed, including completely randomized designs, factorial experiments, and stratified randomization. In contrast to randomized experiments, observational studies lack scientific controls of confounding factors. This feature makes observational data much easier to collect but more difficult for inferring causality. However, observational studies play a central role in modern causal research due to the growing availability of observational data in the 21st century.

Causal inference, as a branch of statistics, focuses on demystifying causality from randomized experimental and observational data using statistical methodologies. The impact of causal inference has significantly expanded in the past decades across various disciplines, including the social sciences, psychology, epidemiology, political science, and technology. The Rubin causal model (RCM) serves as one of the most employed counterfactual frameworks for causal inference. This framework was originally formalized in the context of randomized experiments by the Polish mathematician Jerzy Neyman (Neyman, 1923a)

and was later substantially developed and extended to observational studies by the American statistician Donald Rubin (Rubin, 1974). In the RCM, the outcomes associated with different actions are called potential outcomes. Our understanding of causality arises from comparisons between these potential outcomes. This thesis adopts the RCM as the primary framework for addressing various causal problems.

Rubin holds the belief that "No causation without manipulation" (Imbens and Rubin, 2015). The RCM requires the treatment to be manipulable so that the potential outcomes can be well-defined. The above aspirin example is a well-defined causal statement in the RCM as "taking an aspirin" is a manipulable action — we could have chosen not to take the aspirin. In contrast, the statement "He was fired by his boss because he is Asian" illustrates an ill-defined causal claim, as race is an inherent and non-manipulable attribute of an individual. For this kind of non-manipulated treatments and ill-defined causal problems, this thesis introduces an alternative framework, called balancing covariates framework, for causal inference as described in Section 2.4.

Causal inference is inherently challenging and faces numerous difficulties. This thesis focuses on the challenge of treatment effect heterogeneity (TEH), which represents a major issue in causal inference. To the best of my knowledge, no universally accepted formal definition of TEH exists. Conceptually, TEH refers to variation in the causal effect of a treatment across different subgroups or individuals within a population. For instance, aspirin may be more effective in alleviating headaches among younger individuals compared to its average effect across the entire population. It is essential to note that discussions of TEH typically refer to variation with respect to observed auxiliary variables; that is, the subgroups in which TEH is present are defined by observed covariates. Without such covariates, it is difficult to justify the study of TEH, as any observed variation could be attributed merely to random noise. When TEH is present, researchers often shift their focus

from estimating average causal effects to exploring a broader class of causal parameters. This thesis covers three topics in causal inference motivated by the presence of TEH.

The first topic of this thesis is on the generalization of causal conclusions from a sample to a broader population. This topic is practically important because, in the presence of TEH, statistical analysis of causality based on the sample data may lead to biased and even misleading conclusions on the population of interest if the data at hand are not representative of the population. In history, there are tobacco companies who employed selective subgroup analyses to undermine the emerging scientific consensus on the harms of smoking. They focused on a specific subpopulation where the effect of smoking on disease outcomes appeared statistically insignificant to strategically fostered scientific doubt and delayed regulatory actions. The sample not representative of the population is known as a biased sample. In Chapter 3, we assume that the data for causal inference are probability survey data, enabling the use of survey weights to mitigate selection bias. Survey sampling is a standard statistical tool for data collection. A detailed discussion of survey sampling falls beyond the scope of this thesis; readers are referred to Wu and Thompson (2020) for a comprehensive introduction to this classical area of statistics. We develop in Chapter 3 a framework for statistical inference of a class of population average treatment effects based on observational survey data. This framework yields valid causal conclusions on the population by simultaneously mitigating the confounding bias and selection bias.

The second topic concerns the estimation of treatment harm rate (THR), a measure of TEH for binary outcomes. Subgroup analysis and the estimation of conditional average treatment effects are two commonly used strategies for evaluating TEH. However, these approaches primarily rely on difference-in-means metrics, which may fail to fully capture the extent of TEH by overlooking heterogeneity that exists in the subpopulation. For instance, while a medication may yield a positive average effect for a particular subpopulation, it

could still pose risks to certain individuals within that group — some may experience adverse effects despite the overall benefit. Investigating the treatment harm rate, along with related measures such as the treatment benefit rate, complements existing research on TEH and is increasingly important in real-world applications.

In the third topic of this thesis, we explore mediation analysis in randomized controlled trials with noncompliance. In causal mediation analysis, the researchers are interested in the causal effect of a treatment on outcomes through mediators. For example, we know that smoking is a leading cause of lung cancer. It is also reported that smoking may lead to high level tar deposits which increase the risk of lung cancer. In such case, the tar deposits serve as a mediator on the causal pathway between smoking and lung cancer. Chapter 5 focuses on evaluating causal mediation effects under a key setting in which treatment assignments are randomized, yet noncompliance arises due to subjects deviating from their assigned treatments.

The above three topics are studied in Chapters 3, 4 and 5. Preliminary material on causal inference are provided in Chapter 2, which covers overviews of the Rubin causal model, treatment assignment mechanisms, and causal parameters of primary interest in this thesis. A brief summary of this thesis and comments on potential extensions and future work are given in Chapter 6.

Chapter 2

Causal Inference: Frameworks, Models, and Parameters

The mathematical formulation of a causal inference problem typically involves three elements: a framework that describes causality, a statistical model for the data and particularly the treatments assignment mechanism, and causal parameters of interest. This chapter provides a brief overview of these foundational elements, serving as a preliminary introduction to the three topics in causal inference that will be explored in Chapters 3 – 5.

2.1 The Rubin Causal Model and Potential Outcomes

The Rubin causal model (RCM) invokes the notion of potential outcomes to quantify causality. Let $\mathcal{U} = \{1, \cdots, N\}$ be a finite population and $i \in \mathcal{U}$ a unit in the population. Denote $\mathbb{A}$ the value space of an action or a treatment of interest and $\mathbb{Y}$ the value space of an outcome of interest measured after the treatment. In the aspirin example mentioned

in Chapter 1, we note that $\mathbb{A} = \{\text{aspirin}, \text{no aspirin}\}$ and $\mathbb{Y} = \{\text{headache}, \text{no headache}\}$. A potential outcome for unit i, denoted by $Y_i(a) \in \mathbb{Y}$, is the outcome that would have been observed had the treatment $a \in \mathbb{A}$ been given to the unit. We assume throughout this thesis the stable unit treatment value assumption (SUTVA) introduced in Section 1.6 of Imbens and Rubin (2015) as follows.

Assumption 2.1.1 (SUTVA) *The potential outcomes for any unit do not vary with the treatments assigned to other units, and, for each unit, there are no different forms or versions of each treatment level, which lead to different potential outcomes.*

The SUTVA assumption rules out the possibility that potential outcomes may vary due to the treatment itself and implies a one-to-one correspondence between potential outcomes and treatment values in $\mathbb{A}$. In practice, however, for each unit we can observe only one of the potential outcomes because a unit can only be assigned one treatment; this is known as the fundamental difficulty of causal inference.

The practitioners may be concerned with the causal effect for either dichotomous, categorical, ordinal, or continuous treatments. This thesis focuses on binary treatments and assumes that $\mathbb{A} = \{0, 1\}$. Causal problems involving categorical or ordinal treatments can often be addressed by comparing specific pairs of treatment levels across different combinations. In regard to the outcome type, this thesis focuses on binary and continuous outcomes.

The notion of potential outcomes can be extended to more complex settings involving multiple factors of interest that influence the outcome beyond the assigned treatment. For instance, in randomized controlled trials with noncompliance, as considered in Angrist et al. (1996), the experimental units may not accept the treatments assigned to them. In this context, each unit is characterized by two treatment-related variables: the assigned

treatment z and the actual received treatment a, both taking values in $\mathbb{A}$. The corresponding potential outcome is denoted by $Y_i(z, a)$, representing the outcome that would have been observed had the assigned treatment been z and the actual received treatment been a. Another example is the mediation analysis, where there might be a mediator on the pathway between the assigned treatment and the outcome. The potential outcome can then be defined as $Y(a, m)$ where m is the value of the mediator.

2.2 Treatment Assignment Mechanisms

The fundamental issue of causal inference is that for each unit only the potential outcome given the assigned treatment can be observed and the others are missing. Therefore, how the treatments are practically assigned plays a crucial role in statistical analysis of the data. For unit i, suppose we observe the assigned treatment $A_i \in \{0, 1\}$, the outcome $Y_i = Y_i(A_i)$, and baseline pre-treatment covariates $\boldsymbol{X}_i$. The data at hand are $\{(Y_i, A_i, \boldsymbol{X}_i), i \in \mathcal{U}\}$ and the assigned treatments for the population are $\boldsymbol{A} = (A_1, \cdots, A_N)^\top$. Denote by $\mathbf{D} = \{(Y_i(1), Y_i(0), \boldsymbol{X}_i), i \in \mathcal{U}\}$ the collection of potential outcomes and covariates for the population. The $\boldsymbol{A}$ is a random vector under the RCM and the treatment assignment mechanism (TAM) is defined as the probability distribution of $\boldsymbol{A}$ conditional on $\mathbf{D}$, i.e., $\Pr(\boldsymbol{A} = \boldsymbol{a} \mid \mathbf{D})$ for $\boldsymbol{a} = (a_1, \cdots, a_N)^\top \in \{0, 1\}^N$, which is required to be strictly between 0 and 1 over all the possible values of $\mathbf{D}$. There are two primary sources of data for causal inference: randomized experiments and observational studies. The key distinction between them lies in whether the TAM is known to the researcher.

2.2.1 Randomized Experiments

In randomized experiments, the TAM is known and under experimental control. For example, in a completely randomized experiment (CRE) with the size of treated units set to be n, the TAM is known at the design stage and is equal to

$$\Pr(\boldsymbol{A} = \boldsymbol{a} \mid \mathbf{D}) = \begin{cases} \frac{1}{\binom{N}{n}} & \text{if } \sum_{i=1}^{N} a_i = n, \\ 0 & o.w. \end{cases}.$$

Randomized controlled trials (RCTs) are a special case of the randomized experiments, where the treatment assignments are randomly assigned and independent of the observed information, i.e., $\boldsymbol{A} \perp\!\!\!\perp \mathbf{D}$. Moreover, we require that the treatment rate $\pi = \mathrm{E}(A_i)$ is a fixed real number for $i = 1, \cdots, N$. It is clear that the CRE is a RCT. Another example of the RCT is the Bernoulli trials, in which $\Pr(\boldsymbol{A} = \boldsymbol{a} \mid \mathbf{D}) = \prod_{i=1}^{N} \Pr(A_i = a_i \mid \mathbf{D}) = \pi^{n_1}(1-\pi)^{N-n_1}$ where the treatment rate π is pre-specified and $n_1 = \sum_{i=1}^{N} a_i$. Randomization of treatment assignments yields consistent estimators of the average treatment effect.

2.2.2 Observational Studies

Observational studies are characterized by unknown treatment assignment mechanisms due to the lack of experimental control over treatment assignments. To enable statistical analysis of observational data, the RCM assumes that $A_1, \cdots, A_N$ are independent and the TAM is individualistic, i.e.,

$$\Pr(\boldsymbol{A} = \boldsymbol{a} \mid \mathbf{D}) = \prod_{i=1}^{N} \Pr(A_i = a_i \mid \mathbf{D}) = \prod_{i=1}^{N} \Pr(A_i = a_i \mid Y_i(1), Y_i(0), \boldsymbol{X}_i).$$

The researchers often further assume that the TAM is unconfounded provided the observed covariates, i.e., $\Pr(A_i = 1 \mid Y_i(1), Y_i(0), \boldsymbol{X}_i) = \Pr(A_i = 1 \mid \boldsymbol{X}_i)$, and that the TAM is strictly between 0 and 1 over the support of $\boldsymbol{X}_i$, which is known as the positivity or overlap assumption. The $e(\boldsymbol{X}_i) = \Pr(A_i = 1 \mid \boldsymbol{X}_i)$ is called propensity score (PS) or the PS function. The unconfoundedness assumption implies no unmeasured confounders, and that either the $\boldsymbol{X}_i$ are the pre-treatment covariates which are not affected by the treatment assignments or $\boldsymbol{X}_i$ are not affected by the potential outcomes, otherwise the assumption is false based on the d-separation criterion in Bayesian networks (Pearl, 2009). Weighted average treatment effects, a class of causal parameters as we will introduce later in Section 2.3.1, can be identified from the observed data under the unconfoundedness and overlap assumptions. Despite the unconfoundedness assumption is sometimes perceived untestable and controversial, it is plausible if we hold the prior belief that the covariates can completely explain the TAM or the outcome.

2.2.3 Randomized Controlled Trials with Noncompliance: Instrumental Variables

Noncompliance is a common issue in RCTs. It occurs in an experiment when the behaviors of some experimental units do not conform to the treatment assignments. The TAM of the actual received treatments in such scenario is unknown to and beyond the control of the researchers. Although we can still perform the "intent-to-treat" analysis to evaluate the causal effect of the assigned treatment, the conclusions based on which might be misleading if the treatment effectiveness rather than the assignment effectiveness is of primary interest.

Let Z_i denote the assigned treatment and A_i the actual received treatment for unit i. In a RCT with noncompliance, the TAM for $\boldsymbol{Z} = (Z_1, \cdots, Z_N)$ is known and completely

random, whereas the TAM for $\boldsymbol{A}$ is unknown. The individuals with $A_i \neq Z_i$ yield the noncompliance issue. Let $Y_i(a, z)$ denote the potential outcome under the actual treatment $A_i = a$ and the assigned treatment $Z_i = z$. The exclusion restriction assumption states that $Y_i(a, z) = Y_i(a, z^{'})$ for $z, z^{'}, a \in \mathbb{A}$, implying that Z_i affects the outcome solely through its effect on A_i. Under this assumption, the variable Z_i serves as an instrumental variable and the complier average treatment effect can be identified from the observed data (Angrist et al., 1996).

2.3 Causal Parameters and Identification

2.3.1 Average Treatment Effects

The RCM quantifies the causal effect of a treatment by comparing the resulting potential outcomes. The unit-level treatment effect for unit i is defined as $\tau_i = Y_i(1) - Y_i(0)$. The (finite) population average treatment effect is the population average of unit-level treatment effects, $\tau_{fp} = N^{-1} \sum_{i=1}^{N} \tau_i$. Statistical inference that views the potential outcomes as fixed and assumes that the randomness comes solely from the treatment assignments is called finite-population inference. In this thesis, we focus on model-based inference and further assume that the potential outcomes $Y_i(a), i \in \mathcal{U}$ are independent and identically distributed random variables from an infinite superpopulation, and defined the (super) population average treatment effect (PATE) as $\tau_{sp} = \mathrm{E}[\tau_i]$ where $\mathrm{E}[\cdot]$ denotes taking the expectation with respect to the probability distributions of the random variables. The τ_{sp} measures the overall causal effect of a treatment over the entire population. Let $\boldsymbol{X}_i$ denotes the random vector of baseline covariates for unit i and $\mathbb{X}$ the support of $\boldsymbol{X}_i$. When the primary interest is the average treatment effect over a subpopulation identified

by $\boldsymbol{X}_i$, we can use a pre-specified generic function of the covariates, namely $g(\boldsymbol{X}_i)$, to define the subpopulation, which yields a general class of average treatment effects over target populations as $\tau_g = \mathrm{E}[g(\boldsymbol{X}_i)\tau_i]/\mathrm{E}[g(\boldsymbol{X}_i)]$. With $g(\boldsymbol{X}_i) = 1$, τ_g is the PATE; when $g(\boldsymbol{X}_i) = e(\boldsymbol{X}_i) = \Pr(A_i = 1 \mid \boldsymbol{X}_i)$, τ_g is the population average treatment effect on the treated group (PATT); when $g(\boldsymbol{X}_i) = 1_{\mathbb{B}}(\boldsymbol{X}_i) = 1\{\boldsymbol{X}_i \in \mathbb{B}\}$ for some $\mathbb{B} \subset \mathbb{X}$, τ_g amounts to the average treatment effect for the subpopulation of individuals whose covariates fall into $\mathbb{B}$. For example, suppose the k-th coordinate of $\boldsymbol{X}_i$ represents the the unit's gender and $\mathbb{B} = \{\boldsymbol{x} \in \mathbb{X} : \text{the k-th coordinate of } \boldsymbol{x} \text{ is male}\}$, then τ_g represents the population average treatment effect for the males. A major topic of causal inference is to make inference about τ_g based on observed data from either randomized experiments or observational studies.

When $g(\boldsymbol{X}_i)$ is known and the TAM is unconfounded and strictly between 0 and 1, τ_g can be identified by

$$\tau_g = \mathrm{E}[g(\boldsymbol{X}_i)]^{-1}\mathrm{E}\left\{\left[\frac{A_i g(\boldsymbol{X}_i)}{e(\boldsymbol{X}_i)} - \frac{(1-A_i)g(\boldsymbol{X}_i)}{1-e(\boldsymbol{X}_i)}\right]Y_i\right\}.$$

With observational data, we can first estimate the PS function and then use it for estimation of τ_g. This yields the so-called inverse propensity score weighted estimator or the inverse probability weighted estimator. Since the estimated inverse weights $1/\hat{e}(\boldsymbol{X}_i)$ and $1/(1-\hat{e}(\boldsymbol{X}_i))$ may be very close to 0 and 1, the researchers often normalize the weights for more stable estimation, leading to the Hájek-type estimator

$$\tau_{IPW} = \frac{\sum_{i=1}^N A_i g(\boldsymbol{X}_i)Y_i/\hat{e}(\boldsymbol{X}_i)}{\sum_{i=1}^N A_i g(\boldsymbol{X}_i)/\hat{e}(\boldsymbol{X}_i)} - \frac{\sum_{i=1}^N (1-A_i) g(\boldsymbol{X}_i)Y_i/(1-\hat{e}(\boldsymbol{X}_i))}{\sum_{i=1}^N (1-A_i) g(\boldsymbol{X}_i)/(1-\hat{e}(\boldsymbol{X}_i))},$$

where $\hat{e}(\boldsymbol{X}_i)$ is an estimator of $e(\boldsymbol{X}_i)$. Under a parametric model $e_{\boldsymbol{\gamma}}(\boldsymbol{X}_i)$ for the PS function, the vector of nuisance parameters $\boldsymbol{\gamma}$ can be estimated by maximizing the Bernoulli

log-likelihood

$$\ell_N(\boldsymbol{\gamma}) = \frac{1}{N}\sum_{i=1}^{N} A_i \log e_{\boldsymbol{\gamma}}(\boldsymbol{X}_i) + \frac{1}{N}\sum_{i=1}^{N}(1-A_i)\log(1-e_{\boldsymbol{\gamma}}(\boldsymbol{X}_i)).$$

2.3.2 Treatment Harm Rate

The average treatment effects are measures of average uplift for the outcomes and do not fully reveal the treatment effect heterogeneity (TEH). Let us focus on binary outcomes and assume that $\mathbb{Y} = \{0, 1\}$ with $Y = 1$ being favorable. In Chapter 4, for simplicity of computation, we assume $\mathbb{Y} = \{-1, 1\}$. Under TEH, individuals in a population may respond differently to the treatment. The individualized treatment effect can be positive, ineffective, or negative, regarding different values of the potential outcome pair $(Y_i(1), Y_i(0))$ as shown in Table 2.1. To understand the full treatment effect heterogeneity within a target population, we need to know the distribution of the potential outcome pair over the population, i.e., $\Pr(Y_i(1) = s, Y_i(0) = t)$ for $s, t \in \mathbb{Y}$. In a RCT, it suffices to estimate the treatment harm rate (THR), $\theta = \Pr(Y_i(1) = 0, Y_i(0) = 1)$, as the other proportions can be identified from θ and $\Pr(Y_i(a) = t)$. The PATE on binary outcomes is given by $\tau_{sp} = \Pr(Y_i(1) = 1) - \Pr(Y_i(0) = 0) = \Pr(Y_i(1) = 1, Y_i(0) = 0) - \Pr(Y_i(1) = 0, Y_i(0) = 1)$, which is proportional to the difference between the treatment benefit rate and the THR. It is clear that the THR can not be inferred from the PATE. When $\tau_{sp} = 0.4$ for instance,

Table 2.1: Individualized treatment effect regarding different values of the potential outcomes

		$Y_i(1)$	
		1	0
$Y_i(0)$	1	ineffective	negative
	0	positive	ineffective

Table 2.2: Two hypothetical scenarios with different THR that yield the same observed RCT data and PATE.

	Observed data		Scenario 1		Scenario 2	
i	A_i	Y_i	$Y_i(1)$	$Y_i(0)$	$Y_i(1)$	$Y_i(0)$
1	1	1	1	0	1	0
2	1	1	1	0	1	0
3	1	1	1	0	1	1
4	1	0	0	1	0	0
5	1	0	0	0	0	0
6	0	1	0	1	1	1
7	0	0	0	0	0	0
8	0	0	1	0	0	0
9	0	0	1	0	1	0
10	0	0	1	0	1	0

it could be $0.4 = 0.6 - 0.2$ or $0.4 = 0.4 - 0$. The treatment risk is comparatively high in the former case compared to the latter. An example is illustrated in Table 2.2, where two hypothetical scenarios for the potential outcomes with respectively $\theta = 0.2$ and $\theta = 0$ yield the same observed RCT data and $\tau_{sp} = 0.4$.

Unfortunately, the THR is non-identifiable from the observed data. The Fréchet-Hoeffding bounds of θ provided RCT data without auxiliary information is given by

$$\max\{0, \Pr(Y_i(0) = 1) - \Pr(Y_i(1) = 1)\} \leq \theta \leq \min\{\Pr(Y_i(0) = 1), \Pr(Y_i(1) = 0)\}.$$

Regarding the RCT data in Table 2.2, one can verify that the estimated bounds for θ are $[0, 0.2]$. In the presence of covariates, the sharp bounds of the THR in observational studies given an unconfounded TAM has been derived by Zhang et al. (2013) and Kallus (2022) as $\theta \in [L(\theta), U(\theta)]$ where $L(\theta) = \mathrm{E}[\max\{0, \mu_0(\boldsymbol{X}_i) - \mu_1(\boldsymbol{X}_i)\}]$, $U(\theta) = \mathrm{E}[\min\{\mu_0(\boldsymbol{X}_i), 1 - \mu_1(\boldsymbol{X}_i)\}]$, $\mu_a(\boldsymbol{X}_i) = \Pr(Y_i(a) = 1|\boldsymbol{X}_i)$ for $a = 0, 1$. The $L(\theta)$ and $U(\theta)$, nonetheless, are

not easy to estimate. The plug-in estimators of the bounds by plugging in the estimators of $\mu_0(\boldsymbol{X}_i)$ and $\mu_1(\boldsymbol{X}_i)$ might be severely biased if the underlying models are misspecified. In Chapter 4, we constructed a class of partitioning-based bounds for θ, which are informative under a good partition of the covariates space. More importantly, the partitioning-based bounds can be consistently estimated given RCT data without requiring any ambiguous or untestable assumptions.

2.3.3 Natural Direct and Indirect Effects

The causal parameters we have introduced so far are concerned with the total effects of a treatment on specific outcomes within a target population. There is a practically important branch in causal inference called causal mediation analysis (CMA), which has been developed to understand how the treatment influences outcomes through mediators. A deeper understanding of mediation effects helps the practitioners to uncover the underlying causal mechanism of a treatment and develop more targeted interventions. For example, although it is widely known that smoking is a leading cause of lung cancer, this knowledge alone may be insufficient to deter smoking behavior. By identifying the mediators that lie on the causal pathway between smoking and lung cancer, it may be possible to mitigate the adverse health impacts of smoking by intervening on those intermediate factors.

Let M_i be a mediator of interest and $\mathbb{M}$ the value space of M_i. The temporal sequence of the variables A_i, M_i, and Y_i is $A_i \to M_i \to Y_i$. Let $M_i(a)$ denote the potential outcome of the mediator under $A_i = a$ and $Y_i(m, a)$ the potential outcome of the outcome variable under $M_i = m$ and $A_i = a$ for $a \in \mathbb{A}$ and $m \in \mathbb{M}$. CMA examines the mediation mechanism through two central parameters — natural direct effect (NDE) and natural indirect effct (NIE). The former is defined by $\bar{\xi}(a) = \mathrm{E}[Y_i(1, M_i(a)) - Y_i(0, M_i(a))]$. The

latter is also referred to as the average causal mediation effect (ACME) and is defined by $\bar{\delta}(a) = \mathrm{E}[Y_i(a, M_i(1)) - Y_i(a, M_i(0))]$.

Imai et al. (2010) showed that $\mathrm{E}[Y_i(t, M_i(a))]$ can be nonparametrically identified by

$$\mathrm{E}[Y_i(t, M_i(a))] = \mathrm{E}_{\boldsymbol{X}_i}\left\{\mathrm{E}_{M_i|\boldsymbol{X}_i, A_i=a}\left[\mathrm{E}(Y_i \mid A_i = t, M_i, \boldsymbol{X}_i)\right]\right\}$$

under the sequential ignorability assumption which states that $\{Y_i(t, m), M_i(a)\} \perp\!\!\!\perp A_i \mid \boldsymbol{X}_i$ and $Y_i(t, m) \perp\!\!\!\perp M_i(a) \mid A_i = a, \boldsymbol{X}_i$ for $a, t \in \mathbb{A}$ and $m \in \mathbb{M}$. The sequential ignorability assumption is strong and untestable. It implicitly requires full observations of the treatment-mediator, mediator-outcome, and treatment-outcome confounding variables. However, the NDE and the ACME can still be partially identified without relying on this assumption. Sharp nonparametric bounds for the two parameters have been derived by Sjölander (2009).

2.4 The Balancing Covariates Framework

The RCM is a well-structured and self-consistent framework for studying the causal effect of manipulable treatments. However, when the treatment of interest is not manipulable, such as inherent human attributes like gender or race, the potential outcomes are not well-defined within the RCM. A motivating example arises in the study of racial disparity, a prevalent topic in social sciences such as education, health care, and criminal justice. The racial disparity is an unconfounded descriptive measure that evaluates the differences among racial groups due to socioeconomic status and racial discrimination rather than biological or behavioral differences. The treatment of interest in this context is "race" and the potential outcomes are ill-defined for this treatment. However, the study of the racial disparity can be still be approached from a causal inference perspective, as it in-

volves controlling for and balancing confounding variables across groups. Indeed, a central goal in observational studies given an unconfounded TAM is to balance confounding variables across treatment groups. This thesis therefore introduces a simple and alternative framework for causal inference: balancing covariates framework, which circumvents the need to explicitly define potential outcomes by focusing on the adjustment of confounding variables.

Let $A \in \{0, 1\}$ be a group indicator; e.g., $A = 1$ indicates Asian and $A = 0$ indicates non-Asian. Let Y denote the outcome of interest and $\boldsymbol{X}$ the covariates expected to be controlled in the study. We can define the PATE of A on Y as $\tau = \mathrm{E}[\tau(\boldsymbol{X})]$ where

$$\tau(\boldsymbol{X}) = \mathrm{E}[Y \mid \boldsymbol{X}, A = 1] - \mathrm{E}[Y \mid \boldsymbol{X}, A = 0].$$

The PATE defined in this context is a descriptive measure rather than a causal parameter, and is recognized by some researchers as a characterization of unconfounded association. However, when the TAM is unconfounded, one can verify under the RCM that

$$\mathrm{E}[Y(1) - Y(0) \mid \boldsymbol{X}] = \mathrm{E}[Y \mid \boldsymbol{X}, A = 1] - \mathrm{E}[Y \mid \boldsymbol{X}, A = 0],$$

implying that the PATE defined under the balancing covariates framework in terms of a manipulable treatment is a causal parameter and matches the one defined under the RCM. This framework will be employed in Chapter 3.

Chapter 3

Propensity Score Weighting with Survey Data

3.1 Introduction

As a main source of data for causal inference, observational studies place two major challenges on statistical analysis. The first challenge is the issue of confounding, resulted from distribution imbalance in confounders between the treated and the control groups. Second, the treatment effect may vary across subgroups, making it difficult to generalize the average treatment effect of the study sample to a broader target population.

The propensity score based weighting approach (Hirano and Imbens, 2001) has been adopted as a major tool for dealing with the first challenge under an unconfounded treatment assignment mechanism (TAM) over the past decades, which owns many advantages, including ease of implementation, robustness to misspecification in outcome regression models as compared to regression-based prediction methods, and the potential of doing

model checking by assessing the covariate-balance after weighting. We refer readers to Imbens and Rubin (2015); Austin and Stuart (2015); Li et al. (2018) for excellent reviews of the propensity score weighting approach and its appealing properties.

The second challenge, however, has received less attention in observational studies for causal inference. Lack of generalizability is considered to be a common issue for randomized controlled trials (RCTs) because of their strict inclusion criteria, which give rise to limited diversity of the candidates and the issue of extrapolation (Altman et al., 2001; Evans and Kalra, 2001; Rothwell, 2005). Observational studies, on the other hand, have the capabilities to encompass a broader and more diverse range of participants. However, this does not automatically lead to generalizability and external validity, as certain characteristics of an observational sample may significantly differ from those of the population of interest (Carlson and Morrison, 2009).

Typically, the data at hand are collected from a finite population, e.g., all adults in Canada. Probability sampling has been an established data collection tool in many fields with sound inferential procedures for the finite population. While many observational datasets are collected based on a probability sampling design, a large proportion of the existing research simply dismisses the sampling design information while drawing general conclusions on the target population, as shown in DuGoff et al. (2014). It is imperative to develop rigorous methods that simultaneously address internal validity by mitigating the confounding bias in the sample and external validity by incorporating the sampling design information to alleviate selection biases from the target population. In this chapter, we will focus on a special kind of observational data, post-treatment survey data, of the feature that the treatment stage occurs prior to data collection. There are many examples where the observational datasets are post-treatment survey data, especially when the treatment of interest involves self-selected health measures and behavioral changes. In contrary, most

experimental samples are pre-treatment where individuals are assigned to treatment and control groups after being included in the sample.

There have been a handful of references on combining the propensity score weighting approach with survey weights to generalize the results of an observational study to the target population (DuGoff et al., 2014; Ridgeway et al., 2015; Dong et al., 2020). A few numerical experiments in published papers suggest that the survey weights should be incorporated (either as weights or predictors in the propensity score model) when the parameters of interest are the population-level average treatment effects. However, rigorous theoretical justifications are not available in the existing literature with post-treatment survey data. The first major contribution of this chapter is to fill the aforementioned gap by developing a general framework for causal inference using post-treatment survey data for a general class of population parameters defined as the average treatment effects for either the overall target population or specific subpopulations of interest. In particular, we propose the "**survey-weighted covariate balancing propensity score weighting**" approach which achieves covariate balance not only between the treated and control groups but also between the treatment groups and the target population.This kind of three-way covariate balance yields estimators robust against model misspecifications. The proposed confidence intervals and other inferential procedures are rigorously justified under the "two-phase randomization" framework. Our methodological development involves both the technical details of the superpopulation causal inference and finite-population survey sampling.

The second major contribution of this chapter is the extension to cover scenarios where the observational datasets are a non-probability survey sample. Statistical analysis with non-probability samples is an emerging topic in survey sampling and official statistics. A popular setting for methodological development in recent years is to assume the availability of a single existing reference probability sample from the same target population

with the required auxiliary information (Rivers, 2007; Lee and Valliant, 2009; Elliott and Valliant, 2017; Chen et al., 2020). Similar data integration problems have also been studied for generalizing trial evidence. Existing methods include inverse probability weighting (Greenhouse et al., 2008; Cole and Stuart, 2010; Buchanan et al., 2018), the model-assisted approach of Dahabreh et al. (2019), and the non-parametric calibration weighting method of Lee et al. (2023). In this chapter, we extend our proposed methods for causal inference to scenarios where the observational dataset is a non-probability sample. In particular, we consider practically important scenarios where no single existing reference probability sample provides the required auxiliary information and develop valid inferential procedures for combining auxiliary information from two or multiple existing reference probability samples.

The rest of this chapter is organized as follows. Section 3.2 describes the setting, the inferential problems of interest and some notation. The proposed methodology for probability samples is presented in Section 3.3, with point estimation detailed in Sections 3.3.2 and 3.3.3, a robust sandwich variance estimator given in Section 3.3.4, and numerical studies with probability survey data placed in Section 3.3.5. Extensions of the methods to non-probability samples are described in Section 3.4. An example of the applications for the proposed methods to the real data are presented in Section 3.5. Some brief discussions are included in Section 3.6. Technical proof and details are provided in Section 3.7.

3.2 Problem Setup

Consider a superpopulation with a triplet of the attributes $(Y, A, \boldsymbol{X})$, where $Y \in \mathbb{R}$ denotes the outcome of interest, $A \in \{0, 1\}$ is a binary indicator of two groups, and $\boldsymbol{X}$ is a p-dimensional vector of covariates with the support $\mathbb{X}$. Denote by $e(\boldsymbol{X}) = \Pr(A = 1 \mid \boldsymbol{X})$ the

propensity score (PS). Let $m(\boldsymbol{X}) = \mathrm{E}(Y \mid \boldsymbol{X}, A = 0)$. We adopt the balancing covariates framework introduced in Section 2.4 and define the conditional average treatment effect (CATE) as $\tau(\boldsymbol{X}) = \mathrm{E}(Y \mid \boldsymbol{X}, A = 1) - \mathrm{E}(Y \mid \boldsymbol{X}, A = 0)$. We further assume that $0 < e(\boldsymbol{X}) < 1$ over the support of $\boldsymbol{X}$. We are interested in statistical inference of the population weighted average treatment effect (PWATE) defined as $\tau_g = \mathrm{E}[g(\boldsymbol{X})\tau(\boldsymbol{X})]/\mathrm{E}[g(\boldsymbol{X})]$ where $g : \mathbb{X} \to \mathbb{R}$ is a pre-specified function characterizing a target population of interest. The PWATE is a descriptive measure. However, it becomes a causal parameter under the RCM and a unconfounded TAM as in which scenario one can verify that $\tau(\boldsymbol{X}) = \mathrm{E}[Y(1) - Y(0) \mid \boldsymbol{X}]$ (Hirano et al., 2003; Li et al., 2018). In this chapter, we focus on two types of population parameters which are of common interest in practice: (i) the subpopulation average treatment effect (SATE), corresponding to $g(\boldsymbol{x}) = 1_{\mathbb{B}}(\boldsymbol{x})$, where $\mathbb{B}$ denotes a known subspace of $\mathbb{X}$ and $1_{\mathbb{B}}(\boldsymbol{x})$ is the indicator function taking the value 1 if $\boldsymbol{x} \in \mathbb{B}$ and 0 otherwise; and (ii) subpopulation average treatment effect on the treated (SATT), specified by $g(\boldsymbol{x}) = e(\boldsymbol{x})1_{\mathbb{B}}(\boldsymbol{x})$. When $\mathbb{B} = \mathbb{X}$ and $g(\boldsymbol{x}) = 1$ for all $\boldsymbol{x} \in \mathbb{X}$, τ_g refers to the population average treatment effect (PATE). When $g(\boldsymbol{x}) = e(\boldsymbol{x})$, the τ_g is called the population average treatment effect on the treated (PATT).

Consider a finite population with N labelled units, $\mathcal{U}_N = \{1, \cdots, N\}$. We assume that the finite-population data $\{(Y_i, A_i, \boldsymbol{X}_i), i \in \mathcal{U}_N\}$ are independently and identically distributed from $(Y, A, \boldsymbol{X})$. Suppose we have a survey sample $\mathcal{S}_A$ from the finite population $\mathcal{U}_N$ based on a probability sampling design, denoted by q. The observed data are $\{(Y_i, A_i, \boldsymbol{X}_i, d_i^A), i \in \mathcal{S}_A\}$ where $d_i^A = 1/\pi_i^A$ are the survey weights and $\pi_i^A = \Pr(i \in \mathcal{S}_A)$ are the first-order inclusion probabilities. Let n_A be size of $\mathcal{S}_A$ and $n_A^* = \mathrm{E}_q(n_A)$. We further assume that there exist constants $c_1, c_2 > 0$ which do not depend on N such that $c_1 n_A^*/N < \pi_i^A < c_2 n_A^*/N$ for $i = 1, \cdots, N$.

A two-phase randomization (TPR) framework assumes that the probability survey sam-

ple is selected from the finite population which is itself viewed as a random sample from the superpopulation, i.e.,

$$(Y, A, \boldsymbol{X}) \overset{s}{\Longrightarrow} \{(Y_i, A_i, \boldsymbol{X}_i)\}_{i=1}^{N} \overset{q}{\Longrightarrow} \{(Y_i, A_i, \boldsymbol{X}_i, d_i^A)\}_{i \in \mathcal{S}_A}. \tag{3.1}$$

Here, the notation s represents the superpopulation models for the random vector $(Y, A, \boldsymbol{X})$ and q, as aforementioned, represents the selection mechanism for the survey sample $\mathcal{S}_A$. The TPR model has been widely accepted in survey sampling as a framework for model-based and model-assisted inference (Wu and Thompson, 2020). The TPR model bridges superpopulation causal inference and finite-population survey sampling.

The asymptotic framework under the TPR model (3.1) is as follows. Let $\{\mathcal{U}_v, v = 1, \cdots\}$ be a sequence of finite populations with the population size $\{N_v, v = 1, \cdots\}$; let $\mathcal{S}_v$ be a sample of size n_v drawn from $\mathcal{U}_v$. As $v \to \infty$, $N_v \to \infty$ and $n_v \to \infty$. For ease of notation, we suppress the subscript v and use $N \to \infty$ to denote the limiting process.

3.3 Survey-Weighted Propensity Score Weighting

3.3.1 Two-Step Estimation

We propose a two-step estimation procedure for τ_g, termed as survey-weighted propensity score weighting (SWPSW). In the first step, we postulate a parametric model, $e(\boldsymbol{X}_i) = e_{\boldsymbol{\gamma}}(\boldsymbol{X}_i)$, for the PS and estimate the parameter $\boldsymbol{\gamma}$ by minimizing the survey-weighted loss function of the general form $S_n(\boldsymbol{\gamma}) = \sum_{i \in \mathcal{S}_A} d_i^A S(\boldsymbol{X}_i, A_i, \boldsymbol{\gamma})$ for some function S. A

common choice of the function $S_n(\boldsymbol{\gamma})$ is the negative survey-weighted log-likelihood

$$S_n(\boldsymbol{\gamma}) = -\sum_{i\in\mathcal{S}_A} d_i^A A_i \log e_{\boldsymbol{\gamma}}(\boldsymbol{X}_i) - \sum_{i\in\mathcal{S}_A} d_i^A (1-A_i)\log(1-e_{\boldsymbol{\gamma}}(\boldsymbol{X}_i)).$$

It amounts to solving the survey-weighted estimating equations $\sum_{i\in\mathcal{S}_A} d_i^A \boldsymbol{h}(\boldsymbol{X}_i, A_i, \boldsymbol{\gamma}) = 0$ where $\boldsymbol{h}(\boldsymbol{X}_i, A_i, \boldsymbol{\gamma}) = -\partial S(\boldsymbol{X}_i, A_i, \boldsymbol{\gamma})/\partial\boldsymbol{\gamma}$. The loss function based perspective is advantageous in dealing with high-dimensional data and carrying out adaptive algorithms such as recursive partitioning and boosting for estimation (Zhao, 2019). Let $\hat{\boldsymbol{\gamma}}$ denote the estimator of $\boldsymbol{\gamma}$ based on the first step. We re-weight the observations in the second step by the weights $\hat{\omega}_i = d_i^A A_i g(\boldsymbol{X}_i)/e_{\hat{\boldsymbol{\gamma}}}(\boldsymbol{X}_i) + d_i^A(1-A_i)g(\boldsymbol{X}_i)/(1-e_{\hat{\boldsymbol{\gamma}}}(\boldsymbol{X}_i))$ where $g(\boldsymbol{X}_i) = 1_{\mathbb{B}}(\boldsymbol{X}_i)$ for estimating the SATE and $g(\boldsymbol{X}_i) = 1_{\mathbb{B}}(\boldsymbol{X}_i)e_{\hat{\boldsymbol{\gamma}}}(\boldsymbol{X}_i)$ for estimating the SATT, and estimate τ_g by

$$\hat{\tau} = \frac{\sum_{i\in\mathcal{S}_A} A_i\hat{\omega}_i Y_i}{\sum_{i\in\mathcal{S}_A} A_i\hat{\omega}_i} - \frac{\sum_{i\in\mathcal{S}_A}(1-A_i)\hat{\omega}_i Y_i}{\sum_{i\in\mathcal{S}_A}(1-A_i)\hat{\omega}_i}.$$

The survey weights are essential in both steps. In the first step, the survey weights removed the biases for estimating the PS when the treatment indicator variable A_i and the inclusion probability π_i^A are both associated with some unobserved factors. In the second step, the survey weights removed the biases for estimating the population average treatment effects when $\tau(\boldsymbol{X})$ and the inclusion probability π_i^A are both associated with some observed or unobserved factors. The estimator of τ_g obtained from the above SWPSW approach is consistent if the PS model is correctly specified.

Let $\boldsymbol{\eta} = (\boldsymbol{\gamma}^\intercal, \tau_T, \tau)^\intercal$ and $\boldsymbol{O}_i = (Y_i, A_i, \boldsymbol{X}_i)$, where τ_T is a nuisance parameter used to define the main parameter of interest, τ. The SWPSW approach can be formulated as

solving the system of estimating equations

$$\mathbf{0} = \sum_{i \in \mathcal{S}_A} d_i^A \boldsymbol{\psi}_{g,\boldsymbol{h}}(\boldsymbol{O}_i, \boldsymbol{\eta}), \tag{3.2}$$

where

$$\boldsymbol{\psi}_{g,\boldsymbol{h}}(\boldsymbol{O}_i, \boldsymbol{\eta}) = \left\{ \boldsymbol{h}(\boldsymbol{X}_i, A_i, \boldsymbol{\gamma})^\intercal, \frac{A_i g(\boldsymbol{X}_i)}{e_{\boldsymbol{\gamma}}(\boldsymbol{X}_i)}(Y_i - \tau_T), \frac{(1 - A_i) g(\boldsymbol{X}_i)}{1 - e_{\boldsymbol{\gamma}}(\boldsymbol{X}_i)}(Y_i - \tau_T + \tau) \right\}^\intercal.$$

This estimating equation based formulation allows us to derive the variance for the estimator of $\boldsymbol{\eta}$, thus the variance for the estimator of τ_g, by a sandwich formula as shown in Section 3.3.4. The properties of our estimator for τ_g, denoted as $\hat{\tau}$, partially rely on the choice of the loss function (or the estimating function $\boldsymbol{h}$ as the first derivative of the loss function with respect to $\boldsymbol{\gamma}$) for the estimation of $\boldsymbol{\gamma}$. We give a discussion later to two popular options of the loss function-the negative log-likelihood function and the tailored loss function for covariate balancing (Zhao, 2019), with a focus on the latter one.

3.3.2 SWPSW with Maximum Likelihood Estimator

In regression problems with survey data, it is standard to maximize the survey-weighted Bernoulli log-likelihood function for the estimation of $\boldsymbol{\gamma}$. Denote by $\hat{\boldsymbol{\gamma}}_U$ and $\hat{\boldsymbol{\gamma}}_W$ respectively the estimator of $\boldsymbol{\gamma}$ by maximizing the survey-weighted log-likelihood function and the one by maximizing the unweighted sample log-likelihood function. Although $\hat{\boldsymbol{\gamma}}_U$ and $\hat{\boldsymbol{\gamma}}_W$ are both consistent for the same parameter when the sampling scheme is conditionally exogenous, i.e., $A_i \perp\!\!\!\perp \pi_i^A \mid \boldsymbol{X}_i$, the weighted estimator $\hat{\boldsymbol{\gamma}}_W$ is believed to be less efficient than $\hat{\boldsymbol{\gamma}}_U$ (Lumley and Scott, 2017). The issue of "to weight or not to weight" for the estimation of $\boldsymbol{\gamma}$ may be addressed through a preliminary test (DuMouchel and Duncan,

1983; Wu and Fuller, 2005). The ultimate goal of this chapter, however, is on the estimation of τ_g rather than $\boldsymbol{\gamma}$. The efficiency loss from estimating $\boldsymbol{\gamma}$ does not necessarily lead to loss of efficiency for estimating τ_g. While it is difficult in theory to compare the variance of the weighted estimators of τ_g using $\hat{\boldsymbol{\gamma}}_U$ or $\hat{\boldsymbol{\gamma}}_W$, Ridgeway et al. (2015) and Dong et al. (2020) have conducted extensive simulation studies and showed that, for estimating the PATE and the PATT, the survey weights should be incorporated in the estimation of the PS function in a variety of scenarios to improve the efficiency as well as the robustness of the estimator.

3.3.3 Balancing Covariates via SWPSW

We consider an alternative choice of the estimating functions for $\boldsymbol{\gamma}$ in this section, which leads to a robust estimator of τ_g, based on the covariate balancing propensity score weighting (CBPSW) method initially proposed by Imai and Ratkovic (2014) as solving over-identified estimating equations for $\boldsymbol{\gamma}$ involving both the score function (first derivative of the Bernoulli log-likelihood of $\boldsymbol{\gamma}$) and a covariate balancing condition. An optional form of the CBPSW method is to solve a set of just-identified equations that only contains the covariate balancing condition. We impose a logistic regression model for the PS, $e_{\boldsymbol{\gamma}}(\boldsymbol{x}) = \exp(\boldsymbol{x}^\top\boldsymbol{\gamma})/(1+\exp(\boldsymbol{x}^\top\boldsymbol{\gamma}))$, and define the covariate balancing function as

$$\boldsymbol{h}_{CB,g}(\boldsymbol{X}_i, A_i, \boldsymbol{\gamma}) = \left\{ \frac{A_i g(\boldsymbol{X}_i)}{e_{\boldsymbol{\gamma}}(\boldsymbol{X}_i)} - \frac{(1-A_i)g(\boldsymbol{X}_i)}{1-e_{\boldsymbol{\gamma}}(\boldsymbol{X}_i)} \right\} \boldsymbol{X}_i.$$

We propose to estimate $\boldsymbol{\gamma}$ by solving the survey-weighted (just-identified) covariate balancing estimating equations $\sum_{i\in\mathcal{S}_A} d_i^A \boldsymbol{h}_{CB,g}(\boldsymbol{X}_i, A_i, \boldsymbol{\gamma}) = \boldsymbol{0}$, which has a unique solution under mild conditions. The estimator of τ_g is obtained by setting $\boldsymbol{h} = \boldsymbol{h}_{CB,g}$ and then solving (3.2).

Duality. A dual optimization problem to the survey-weighted CBPSW approach can be formulated using the technique introduced in Zhao (2019). When $g(\boldsymbol{x}) = 1_{\mathbb{B}}(\boldsymbol{x})$ (for the estimation of the SATE), the weights $\hat{\omega}_i$ used to re-weight the observations defined in Section 3.3.1 are the solutions of the optimization problem:

$$\begin{aligned} \min_{\omega_i > d_i^A} & \sum_{i \in \mathcal{S}_A} (\omega_i - d_i^A) \log(\omega_i - d_i^A) - \omega_i (1 + \log(d_i^A)) \\ \text{s.t.} & \sum_{i \in \mathcal{S}_A, A_i = 1} \omega_i \boldsymbol{X}_i = \sum_{i \in \mathcal{S}_A, A_i = 0} \omega_i \boldsymbol{X}_i. \end{aligned} \tag{3.3}$$

For $i : \boldsymbol{X}_i \notin \mathbb{B}$, $\hat{\omega}_i$ are zeros because $g(\boldsymbol{X}_i) = 1_{\mathbb{B}}(\boldsymbol{X}_i) = 0$. The objective function of (3.3) reaches its global minimum at $\omega_i = d_i^A$ for $i \in \mathcal{S}_A, \boldsymbol{X}_i \in \mathbb{B}$. The dual optimization problem (3.3) shows that the estimated weights are forced towards the survey weights subject to the covariate balancing condition. It is worth noting that many popular non-parametric calibration methods in causal inference aim to solve an optimization problem with the objective function taking a variety of forms such as the empirical likelihood function (Qin and Zhang, 2007), entropy balancing (Hainmueller, 2012), and the general distance measure (Chan et al., 2016). Unfortunately, all these methods do not incorporate the survey weights into their frameworks and simply push the estimated weights to the uniform weights rather than the survey weights. Thus, these nonparametric methods only achieve covariate-balance between the treated and the control groups and do not provide results generalizable to the target population. Our proposed survey-weighted CBPSW method can simultaneously achieve covariate-balance among the treatment groups and the population of interest. Such three-way covariate-balance leads to robust estimation of the main parameters of interest. The dual problem formulation of (3.2) for estimating the SATT is given in Section 3.7.1. Let $\hat{\tau}$ denote the estimator of τ_g obtained by the survey-weighted approach. We have the following proposition.

Proposition 3.3.1 (Robustness) *The estimator $\hat{\tau}$ is consistent for τ_g under the TPR model (3.1) when the PS model is correctly specified. When the PS model is wrong, the estimator for SATE is still consistent if $m(\boldsymbol{x})$ is linear in $\boldsymbol{x}$ and $\tau(\boldsymbol{x})$ is a constant for $\boldsymbol{x} \in \mathbb{X}$. For estimating the SATT, the constant assumption on $\tau(\boldsymbol{x})$ can be removed and the robustness property remains.*

When the treatment effect is heterogeneous within the population, covariate mean balance after weighting among the treatment groups and the population ensures that the estimator of τ_g is consistent when the PS model is wrong but $m(\boldsymbol{x})$ is linear in $\boldsymbol{x}$ (Fan et al., 2021). However, this kind of three-way covariate-balance is not guaranteed under the weights for estimating the SATE. The weighted means of the predictors between the treated and the control groups are still balanced, but the weighted means of predictors for the two groups are not equal to the estimated mean of the target population. This is due to the fact that the three-way covariate-balance condition

$$\sum_{i\in\mathcal{S}_A,\boldsymbol{X}_i\in\mathbb{B}} \frac{d_i^A A_i}{e_{\boldsymbol{\gamma}}(\boldsymbol{X}_i)}\boldsymbol{X}_i = \sum_{i\in\mathcal{S}_A,\boldsymbol{X}_i\in\mathbb{B}} \frac{d_i^A (1-A_i)}{1-e_{\boldsymbol{\gamma}}(\boldsymbol{X}_i)}\boldsymbol{X}_i = \sum_{i\in\mathcal{S}_A,\boldsymbol{X}_i\in\mathbb{B}} d_i^A \boldsymbol{X}_i \tag{3.4}$$

is an over-identified equation system for $\boldsymbol{\gamma}$ under the parametric model and it typically has no solutions. This is also the reason why we require $\tau(\boldsymbol{x})$ to be a constant for theoretical analysis because when the treatment effect is homogeneous a two-way covariate-balance between the treated and the control groups is sufficient for the proof of consistency. When it comes to the estimation of the SATT, the constant CATE assumption is no longer necessary for the proof of consistency for $\hat{\tau}$ because the treated observations weighted by the survey weights compose the target population of interest for the SATT. In this scenario, the two-way covariate-balance implies the three-way covariate-balance. However, simulation results in Section 3.3.5 and a real data example in Section 3.5 show that, even

though the three-way covariate-balance is not exactly achieved with non-constant CATE, the results are substantially improved by the survey-weighted CBPSW approach. The resulting estimator for the SATE shows double robustness property in the presence of heterogeneous treatment effects.

3.3.4 Variance Estimation

The theoretical variance of $\hat{\tau}$ under the TPR model, known as the anticipated variance (Isaki and Fuller, 1982), involves the variance component due to the sampling design q and the component induced by the superpopulation model s, with the decomposition of the total variance as $\mathrm{Var}_{qs}(\hat{\tau}) = \mathrm{E}_s\{\mathrm{Var}_q(\hat{\tau})\} + \mathrm{Var}_s\{\mathrm{E}_q(\hat{\tau})\}$. We have $\mathrm{E}_s\{\mathrm{Var}_q(\hat{\tau})\} = O(n_A^{-1})$ and $\mathrm{Var}_s\{\mathrm{E}_q(\hat{\tau})\} = O(N^{-1})$ under some mild regularity conditions. When the sampling fraction n_A/N is small, the leading term of the total variance is the design component $\mathrm{E}_s\{\mathrm{Var}_q(\hat{\tau})\}$. Thus, estimation of $\mathrm{Var}_{qs}(\hat{\tau})$ amounts to estimating $\mathrm{Var}_q(\hat{\tau})$. The $\hat{\tau}$ as the solution of survey-weighted estimating equations is asymptotically normal around τ_g under the TPR model and certain regularity assumptions (Yuan and Jennrich, 1998; Rubin-Bleuer and Kratina, 2005). Therefore, once we have an estimator of $\mathrm{Var}_q(\hat{\tau})$, we can make inference about τ_g based on the normal approximation.

Based on the survey-weighted estimating equations (3.2), the robust design-based sandwich variance estimator of $\boldsymbol{\eta}$ is given by $\hat{\boldsymbol{A}}^{-1}\hat{\boldsymbol{\Omega}}(\hat{\boldsymbol{A}}^{\intercal})^{-1}$ where

$$\hat{\boldsymbol{A}} = \frac{1}{\hat{N}} \sum_{i \in \mathcal{S}_A} d_i^A \frac{\partial \boldsymbol{\psi}_{g,\boldsymbol{h}}(\boldsymbol{O}_i, \hat{\boldsymbol{\eta}})}{\partial \boldsymbol{\eta}^{\intercal}}$$

and

$$\hat{\boldsymbol{\Omega}} = \frac{n_A}{\hat{N}^2} \sum_{i \in \mathcal{S}_A} \sum_{j \in \mathcal{S}_A} \frac{\pi_{ij}^A - \pi_i^A \pi_j^A}{\pi_{ij}^A \pi_i^A \pi_j^A} \boldsymbol{\psi}_{g,\boldsymbol{h}}(\boldsymbol{O}_i, \hat{\boldsymbol{\eta}})^{\otimes 2}.$$

Here, $\pi_{ij}^A = \Pr(i, j \in \mathcal{S}_A)$ is the second-order inclusion probability for the pair (i, j), $\hat{N}$ is an estimate of N, and $\boldsymbol{H}^{\otimes 2}$ denotes $\boldsymbol{H}\boldsymbol{H}^\intercal$ for a vector or matrix $\boldsymbol{H}$. The π_{ij}^A's may be difficult to compute under complex survey designs, and practitioners often use approximate variance estimator not involving π_{ij}^A. For instance, one may use the variance formula under Poisson sampling:

$$\hat{\boldsymbol{\Omega}} = \frac{n_A}{\hat{N}^2} \sum_{i \in \mathcal{S}_A} \frac{1 - \pi_i^A}{{\pi_i^A}^2} \boldsymbol{\psi}_{g,\boldsymbol{h}}(\boldsymbol{O}_i, \hat{\boldsymbol{\eta}})^{\otimes 2}. \tag{3.5}$$

Another example is the Hájek variance estimator which performs well for the so-called high entropy survey designs; see Section 4.2.2 of Wu and Thompson (2020) for further detail. The derivation of the proposed variance estimator does rely on specific forms of the models for the PS function. Therefore, the resulting inferential procedure based on normal approximation is robust against model misspecifications.

3.3.5 Numerical Studies

In this subsection, we conduct Monte Carlo simulation studies to evaluate the finite-sample performance of the proposed methodologies on statistical inferences under probability sampling designs. The setup of the simulations is as follows. We first generate a group of covariates $\boldsymbol{X}_i = (X_{1i}, X_{2i}, X_{3i}, X_{4i}, X_{5i})^\intercal$ by $X_{1i} = Z_{1i}$, $X_{2i} = Z_{2i} + 0.3X_{1i}$, $X_{3i} = Z_{3i} + 0.2(X_{1i} + X_{2i})$, $X_{4i} = Z_{4i} + 0.1(X_{1i} + X_{2i} + X_{3i})$, $X_{5i} = Z_{5i}^2 + 0.2X_{3i}X_{4i}$ where $Z_{1i} \sim$ Bernoulli(0.5), $Z_{2i} \sim$ Unif$(0, 4)$, $Z_{3i} \sim$ Exp(1), $Z_{4i} \sim \chi(4)$, and $Z_{5i} \sim N(0, 2^2)$ so that the covariates are all positive (the auxiliary variables of a survey are typically positive) and mildly correlated with each other. We then generate the the treatment indicator

variable A_i from the binomial distribution with the PS specified as

$$e(\boldsymbol{X}_i) = \text{expit}(-X_{1i} + 0.6X_{2i} - 0.5X_{3i} + 0.1X_{4i} + \beta_0),$$

where β_0 is chosen so that $\sum_{i=1}^{N} e(\boldsymbol{X}_i) = 1/2$. The observed outcomes are generated from

$$Y_i = m(\boldsymbol{X}_i) + A_i\tau(\boldsymbol{X}_i) + N(0, 2^2)$$

where $m(\boldsymbol{X}_i) = X_{1i} + X_{2i} + X_{4i} + X_{5i} + 2$, $\tau(\boldsymbol{X}_i) = 0.5m(\boldsymbol{X}_i) + \beta_\tau$, and β_τ is chosen so that the PATE is equal to 3. Now, we have created the finite-population data $\{(Y_i, A_i, \boldsymbol{X}_i)\}_{i=1}^{N}$. We set $N = 2e4$. Every survey sample $\mathcal{S}_A$ from the finite population is selected by the randomized systematic PPS sampling method with the sample size $n_A = 500$ and the first inclusion probability π_i^A set to be proportional to $X_{4i} + c$ where c is a controlling factor so that $\max_i\{X_{4i} + c\}/\min_i\{X_{4i} + c\} = 50$ (Hartley and Rao, 1962). We replicate this procedure for $B = 1000$ times and the simulation results are based on such 1000 survey samples. The outcome Y_i and the treatment A_i are always observed for $\mathcal{S}_A$ while the covariates $\boldsymbol{X}_i$ may be not. Four different scenarios are studied: (1). TT: $X_{1i}, X_{2i}, X_{3i}, X_{4i}, X_{5i}$ are fully observed; (2). TF: $X_{1i}, X_{2i}, X_{4i}, X_{5i}$ are observed; (3). FT: $X_{1i}, X_{2i}, X_{3i}, X_{4i}$ are observed; (4). FF: X_{1i}, X_{2i}, X_{4i} are observed. We say the outcome model is correct if $m(\boldsymbol{X}_i)$ is linear in the observed covariates. TT yields correct outcome and PS models; TF yields correct outcome model but wrong PS model; FT yields correct PS model but wrong outcome model; FF yields two wrong models.

We apply five different approaches — Naive, propensity score weighting with maximum likelihood estimation (PSWMLE), CBPSW, survey-weighted PSWMLE (SWPSWMLE), and survey-weighted CBPSW (SWCBPSW) — to estimate the PATE, PATT, and SATE with the subpopulation $\mathbb{B} = \{2 < X_{4i} < 7\}$. The "Naive" corresponds to the difference in

means estimator; the PSWMLE and CBPSW correspond to the conventional propensity score weighting approaches; the SWPSWMLE and SWCBPSW correspond to the proposed survey-weighted two-step estimation approaches as described in Section 3.3. To evaluate these methods, we consider the following performance measures: (1) %RB: percentage relative bias for the point estimation, i.e., $\%\text{RB} = 1/B \sum_{b=1}^{B} (\hat{\tau}^{(b)} - \tau_g)/\tau_g \times 100$; (2) MCSD: Monte Carlo standard deviation; (3) AVESE: average of estimated standard errors; (4) RMSE: root mean squared error; (5) %CP: percentage empirical coverage probability. The simulation results for estimating the PATE, the PATT, and the SATE are presented in Tables 3.1 - 3.3, respectively. The findings based on these tables are summarized as follows.

First, the Naive method, PSWMLE and CBPSW are biased for all three estimands. The biases are substantially reduced after incorporating the survey weights as expected. In the scenario where the PS model is wrongly specified (TF), SWPSWMLE presents slight bias (around 3%), the relative bias of SWCBPSW is less than 1.5% for the PATT and SATE. This reveals that SWCBPSW is robust against model mis-specification for the PS as described in Proposition 3.3.1. When the outcome model is correct, the CBPSW and SWCBPSW estimators are significantly more efficient than the PSWMLE and SWP-SWMLE estimators. Although it is difficult to theoretically explain this phenomenon when the PS model is wrong, it is known that the CBPSW-based methods are closely related to semiparametric efficiency bound, and hence present the property of variance reduction; See Zhao and Percival (2017) and Zhao (2019) for further details. For estimating the PATE, the variances of the SWCBPSW estimators are under-estimated and the coverage rates fall below the nominal level but remain within an acceptable range.

Table 3.1: Simulation results for estimating the PATE with probability survey samples.

Scenario	Method	%RB	MCSD	AVESE	RMSE	%CP
TT	Naive	89.96	0.86	0.88	2.83	12.30
	PSWMLE	41.33	0.44	0.43	1.31	14.40
	CBPSW	40.19	0.30	0.28	1.24	0.60
	SWPSWMLE	2.02	0.48	0.47	0.48	95.40
	SWCBPSW	2.42	0.32	0.30	0.33	92.40
TF	Naive	89.96	0.86	0.88	2.83	12.30
	PSWMLE	44.11	0.39	0.40	1.38	4.30
	CBPSW	40.76	0.28	0.27	1.25	0.20
	SWPSWMLE	3.25	0.42	0.43	0.43	95.40
	SWCBPSW	2.51	0.30	0.29	0.31	92.80
FT	Naive	89.96	0.86	0.88	2.83	12.30
	PSWMLE	41.68	0.78	0.79	1.48	64.40
	CBPSW	40.84	0.75	0.75	1.44	64.20
	SWPSWMLE	0.14	0.92	0.90	0.92	94.40
	SWCBPSW	0.43	0.90	0.86	0.90	93.80
FF	Naive	89.96	0.86	0.88	2.83	12.30
	PSWMLE	17.12	0.74	0.75	0.90	90.80
	CBPSW	14.54	0.73	0.74	0.85	92.00
	SWPSWMLE	-18.84	0.85	0.85	1.02	85.90
	SWCBPSW	-19.34	0.85	0.84	1.03	85.10

Table 3.2: Simulation results for estimating the PATT with probability survey samples.

Scenario	Method	%RB	MCSD	AVESE	RMSE	%CP
TT	Naive	77.28	0.86	0.88	2.63	17.80
	PSWMLE	45.21	0.64	0.62	1.59	27.50
	CBPSW	41.02	0.33	0.33	1.36	1.70
	SWPSWMLE	-0.38	0.54	0.56	0.54	95.70
	SWCBPSW	1.90	0.35	0.34	0.35	93.80
TF	Naive	77.28	0.86	0.88	2.63	17.80
	PSWMLE	41.03	0.51	0.52	1.41	25.50
	CBPSW	40.28	0.32	0.32	1.33	1.50
	SWPSWMLE	-3.17	0.46	0.47	0.47	95.50
	SWCBPSW	1.44	0.34	0.34	0.34	94.20
FT	Naive	77.28	0.86	0.88	2.63	17.80
	PSWMLE	45.03	0.88	0.88	1.69	57.20
	CBPSW	42.90	0.79	0.78	1.59	56.10
	SWPSWMLE	-1.68	0.86	0.87	0.86	94.70
	SWCBPSW	1.02	0.84	0.83	0.84	94.60
FF	Naive	77.28	0.86	0.88	2.63	17.80
	PSWMLE	22.51	0.83	0.84	1.10	84.90
	CBPSW	18.62	0.78	0.77	0.98	88.60
	SWPSWMLE	-18.02	0.83	0.85	1.01	89.00
	SWCBPSW	-16.42	0.81	0.81	0.97	89.40

Table 3.3: Simulation results for estimating the SATE with probability survey samples.

Scenario	Method	%RB	MCSD	AVESE	RMSE	%CP
TT	Naive	96.83	0.86	0.88	3.03	7.80
	PSWMLE	8.46	0.80	0.76	0.84	93.00
	CBPSW	8.65	0.34	0.32	0.43	85.60
	SWPSWMLE	-0.21	0.84	0.81	0.84	94.30
	SWCBPSW	-0.12	0.35	0.33	0.35	93.80
TF	Naive	96.83	0.86	0.88	3.03	7.80
	PSWMLE	16.29	0.82	0.78	0.95	90.70
	CBPSW	9.53	0.32	0.31	0.43	84.70
	SWPSWMLE	3.07	0.83	0.81	0.84	94.80
	SWCBPSW	0.47	0.33	0.32	0.33	94.30
FT	Naive	96.83	0.86	0.88	3.03	7.80
	PSWMLE	9.12	1.02	0.99	1.05	94.20
	CBPSW	7.92	0.96	0.92	0.99	94.30
	SWPSWMLE	-2.27	1.09	1.05	1.10	94.80
	SWCBPSW	-2.11	1.02	0.97	1.03	94.40
FF	Naive	96.83	0.86	0.88	3.03	7.80
	PSWMLE	-9.13	0.98	0.96	1.01	93.10
	CBPSW	-10.54	0.93	0.91	0.98	92.20
	SWPSWMLE	-18.88	1.04	1.02	1.19	89.80
	SWCBPSW	-18.88	0.99	0.96	1.14	87.20

3.4 Survey-Weighted Propensity Score Weighting with Non-Probability Samples

In this section we extend our proposed methodologies to cover scenarios where the dataset is a non-probability survey sample from the finite population $\mathcal{U}_N$ with unknown sample selection mechanism. Let $\mathcal{S}_A$ be the non-probability survey sample and $R_i = 1_{\mathcal{S}_A}(i)$ be the sample inclusion indicator for unit i. We assume that the unknown sample selection mechanism is guided by an underlying probability model on R_i and, without loss of generality, we use the same notation q to denote the model.

3.4.1 Three-Step Estimation

One of the most crucial tasks in dealing with non-probability samples is the estimation of the so-called participation probabilities $\pi(\boldsymbol{X}_i) = \Pr_q(R_i = 1 \mid \boldsymbol{X}_i) = E_q(R_i \mid \boldsymbol{X}_i)$. Two commonly used assumptions are (i) $\mathrm{E}_s(Y_i \mid \boldsymbol{X}_i, A_i, R_i) = \mathrm{E}_s(Y_i \mid \boldsymbol{X}_i, A_i)$, where the inclusion of A_i is unique to our current setting; and (ii) there exists a reference probability sample, $\mathcal{S}_B$, with information on $\boldsymbol{X}$. It is also assumed that $R_i, i \in \mathcal{U}_N$ are independent of each other. Wu (2022) contains general discussions on statistical analysis with non-probability samples.

We consider practical scenarios where not all components of $\boldsymbol{X}$ are observed in the reference sample $\mathcal{S}_B$. Let $\mathcal{C}$ be the set of indices of observed components of $\boldsymbol{X}_i$ for $\mathcal{S}_B$. Let $\boldsymbol{X}_{\mathcal{C},i}$ be the corresponding observed covariates. The reference dataset is given by $\{(\boldsymbol{X}_{\mathcal{C},i}, d_i^B)\}_{i \in \mathcal{S}_B}$, where $d_i^B, i \in \mathcal{S}_B$ are the known survey weights for $\mathcal{S}_B$. Let $\pi(\boldsymbol{X}_{\mathcal{C},i}) = \mathrm{E}_q(R_i \mid \boldsymbol{X}_{\mathcal{C},i})$. We propose a three-step estimation procedure for τ_g as a natural extension of the two-step estimation procedure given in Section 3.3.1 for non-probability survey data.

In the first step, we estimate $\pi(\boldsymbol{X}_{\mathcal{C},i})$ using a calibration approach presented in Section 3.4.3, which is of some advantages over the pseudo-maximum-likelihood approach in Chen et al. (2020). In the last two steps, we follow the two-step estimation procedure described in Section 3.3.1 with the survey weights d_i^A replaced by the inverse of the estimated participation probabilities, $\hat{\pi}(\boldsymbol{X}_{\mathcal{C},i})^{-1}$. Essentially, our proposed estimator of τ_g can be obtanied by solving

$$\mathbf{0} = \sum_{i \in \mathcal{S}_A} \frac{\boldsymbol{\psi}_{g,\boldsymbol{h}}(\boldsymbol{O}_i, \boldsymbol{\eta})}{\hat{\pi}(\boldsymbol{X}_{\mathcal{C},i})}. \tag{3.6}$$

As compared to (3.2), the known survey weights d_i^A are replaced by the inverse of the estimated participation probabilities. Our focus for the rest of the section is to discuss theoretical properties of the resulting estimator of τ_g, and propose a new approach for the estimation of participation probabilities when information on certain components of $\boldsymbol{X}$ is scattered in two or multiple existing reference probability samples.

3.4.2 Identification with Partially Observed Covariates

We consider three additional assumptions: **B1**. $A_i \perp\!\!\!\perp R_i \mid \boldsymbol{X}_i$; **B2**. $g(\boldsymbol{X}_i)$ is a function of $\boldsymbol{X}_{\mathcal{C},i}$; and **B3**. $R_i \perp\!\!\!\perp \tau(\boldsymbol{X}_i) \mid \boldsymbol{X}_{\mathcal{C},i}$.

Proposition 3.4.1 *The estimating functions* $\boldsymbol{U}(\boldsymbol{\eta}) = \sum_{i \in \mathcal{S}_A} \boldsymbol{\psi}_{g,\boldsymbol{h}}(\boldsymbol{O}_i, \boldsymbol{\eta})/\pi(\boldsymbol{X}_{\mathcal{C},i})$ *are unbiased for* τ_g *under the TPR framework and the assumptions* ***B1***, ***B2*** *and* ***B3***.

B1 can be viewed as the ignorability of sample inclusion for $\mathcal{S}_A$ with respect to the treatment assignments. In fact, **B1** is not essential for estimating the PATE because we can switch our target from estimating the (population) PS to estimating the local PS defined

as $e_l(\boldsymbol{X}_i) = \Pr(A_i = 1 \mid R_i = 1, \boldsymbol{X}_i)$, and one can verify that the estimating equations $\boldsymbol{U}(\boldsymbol{\eta})$ remain unbiased for the PATE under **B3**. **B2** states that the target subpopulation can only be defined using the observed covariates in $\mathcal{S}_B$. For estimating the SATE, **B2** is easily verifiable because $g(\boldsymbol{X}_i)$ is known. **B3** implies that we do not have to observe all the confounders in $\mathcal{S}_B$ other than the ones related to both the heterogeneous treatment effect and the selection procedure. **B3** is not testable with the given sample unless we have information on $\tau(\boldsymbol{x})$.

The treatment indicator may also be observed in $\mathcal{S}_B$. When the confounders $\boldsymbol{X}_i$ are fully observed in $\mathcal{S}_B$, **B2** and **B3** must hold and by including A_i in the participation probability model, the condition **B1** can be removed for identification. However, if the confounders are only partially observed in $\mathcal{S}_B$, including A_i in the participation probability will ruin the identifiability provided by Proposition 3.4.1. That is, under **B1** - **B3**, the estimating functions $\sum_{i \in \mathcal{S}_A} \boldsymbol{\psi}_{g,\boldsymbol{h}}(\boldsymbol{O}_i, \boldsymbol{\eta})/\pi(\boldsymbol{X}_{\mathcal{C},i}, A_i)$ are biased unless $\boldsymbol{X}_{\mathcal{C},i} = \boldsymbol{X}_i$. Therefore, we do not suggest including A_i for the estimation of the participation probabilities even if the treatment indicator is observed in $\mathcal{S}_B$ unless the confounding variables are also fully observed in $\mathcal{S}_B$.

Although **B3** is difficult to verify, it becomes more plausible when we observe more relevant auxiliary variables in $\mathcal{S}_B$. In particular, **B3** holds when $\boldsymbol{X}_{\mathcal{C},i} = \boldsymbol{X}_i$. This motivates us to consider an important research problem of combining auxiliary information from two or multiple existing probability samples for more reliable estimation of participation probabilities.

3.4.3 Estimation of Participation Probabilities Using Multiple Reference Samples

Suppose that we have J independent reference probability samples $\mathcal{S}_{B_j}, j = 1, \cdots, J$, each of them may include a different set of covariates. Let $\mathcal{C}_j$ be the set of indices for the covariates observed in $\mathcal{S}_{B_j}$ with cardinality $p_j = |\mathcal{C}_j|$. We use q_j to denote the sampling design for $\mathcal{S}_{B_j}$. Let $\mathcal{C} = \cup_{j=1}^{J} \mathcal{C}_j$ be the combined set of fully observed covariates among all J reference samples with set size $p_c = |\mathcal{C}|$. Without loss of generality, we assume that $\mathcal{C} = \{1, \cdots, p_c\}$. For the j-th reference sample, we observe the data $\{\boldsymbol{X}_{\mathcal{C}_j,i}, d_i^{B_j}, i \in \mathcal{S}_{B_j}\}$ where $d_i^{B_j}, i \in \mathcal{S}_{B_j}$ are known survey weights for $\mathcal{S}_{B_j}$. We also assume that 1 is the first component of $\boldsymbol{X}_{\mathcal{C}_j}$.

We consider a parametric model for the participation probabilities with $\pi(\boldsymbol{X}_{\mathcal{C},i}) = \pi_{\boldsymbol{\theta}}(\boldsymbol{X}_{\mathcal{C},i}) = f(\boldsymbol{\theta}^{\intercal}\boldsymbol{X}_{\mathcal{C},i})$ where $f(\cdot)$ is an inverse link function. Note that the combined set of covariates $\boldsymbol{X}_{\mathcal{C},i}$ is available in the non-probability sample $\mathcal{S}_A$ but is not fully observed in any of the single reference sample $\mathcal{S}_{B_j}$. The likelihood function of $\boldsymbol{\theta}$ is given by

$$\ell_N(\boldsymbol{\theta}) = \sum_{i \in \mathcal{S}_A} \log\left\{\frac{\pi_{\boldsymbol{\theta}}(\boldsymbol{X}_{\mathcal{C},i})}{1 - \pi_{\boldsymbol{\theta}}(\boldsymbol{X}_{\mathcal{C},i})}\right\} + \sum_{i=1}^{N} \log\{1 - \pi_{\boldsymbol{\theta}}(\boldsymbol{X}_{\mathcal{C},i})\}.$$

Unfortunately, the pseudo-maximum-likelihood approach of Chen et al. (2020) is no longer feasible because the second term of $\ell_N(\boldsymbol{\theta})$ cannot be approximated using the J reference samples. To circumvent this difficulty, we propose to use the following set of unbiased estimating equations for $\boldsymbol{\theta}$,

$$\mathbf{0} = \sum_{i=1}^{N} \left\{\frac{R_i}{\pi_{\boldsymbol{\theta}}(\boldsymbol{X}_{\mathcal{C},i})} - 1\right\} \boldsymbol{X}_{\mathcal{C},i}, \tag{3.7}$$

which reduces to the conventional calibration equations in survey sampling in the form of $\sum_{i\in\mathcal{S}_A} \boldsymbol{X}_{\mathcal{C},i}/\pi_{\boldsymbol{\theta}}(\boldsymbol{X}_{\mathcal{C},i}) = \sum_{i=1}^{N} \boldsymbol{X}_{\mathcal{C},i}$. If we had access to the population totals $\boldsymbol{T}_{\mathcal{C}} = \sum_{i=1}^{N} \boldsymbol{X}_{\mathcal{C},i}$ from census data, an estimate of $\boldsymbol{\theta}$ can be obtained by solving (3.7). Under the current setting, each component of $\boldsymbol{T}_{\mathcal{C}} = (T_1, \cdots, T_{p_c})^\intercal$ can be estimated from at least one of the reference samples. More generally, for each $k \in \mathcal{C}$, let $\hat{T}_{kj} = \sum_{i\in\mathcal{S}_{B_j}} d_i^{B_j} X_{k,i}$ be the estimate for T_k if $X_{k,i}$ is available in $\mathcal{S}_{B_j}$. Then, a weighted average of $\hat{T}_{kj}$ for $j : k \in \mathcal{C}_j$ would be an unbiased and consistent estimator of the true population total T_k. The proposed estimator $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ is obtained by solving

$$0 = \sum_{i\in\mathcal{S}_A} \frac{X_{k,i}}{\pi_{\boldsymbol{\theta}}(\boldsymbol{X}_{\mathcal{C},i})} - \sum_{j:k\in\mathcal{C}_j} w_{kj}\hat{T}_{kj}, \qquad k = 1, \cdots, p_c, \tag{3.8}$$

where $w_{kj} \geq 0$ for $k \in \mathcal{C}_j$, $j = 1, \cdots, J$ satisfy $\sum_{j:k\in\mathcal{C}_j} w_{kj} = 1$. An optimal choice for w_{kj} in terms of estimating T_k using available data is to make it proportional to the inverse of the estimated design-based variance of $\hat{T}_{kj}$. Another simple choice in practice is $w_{kj} = n_{B_j}/\sum_{i:k\in\mathcal{C}_i} n_{B_i}$ where n_{B_j} is the sample size of $\mathcal{S}_{B_j}$. See Section 3.7.4 for further details.

3.4.4 Variance Estimation

The use of estimated participation probabilities in (3.6) with a consistent estimator $\hat{\boldsymbol{\theta}}$ leads to a consistent estimator of the main parameter of interest, τ_g. However, the estimator $\hat{\boldsymbol{\theta}}$ affects the variance of the estimator of τ_g. We provide a variance estimator for τ_g under the setting that the population size N is much bigger than n_A and $n_{B_j}, j = 1, \cdots, J$.

Let $\boldsymbol{W}_j = \{a_{kj}\}_{k=1}^{p_c}$ be the $p_c \times p_c$ diagonal matrix, $j = 1, \cdots, J$. We let the unobserved covariates in a particular reference sample be zero, i.e., $X_{k,j} = 0$ for $k \notin \mathcal{C}_j$. Let $\boldsymbol{\eta} =$

$(\boldsymbol{\theta}^\intercal, \boldsymbol{\gamma}^\intercal, \tau_T, \tau)^\intercal$. It follows that the SWPSW approach to estimating τ_g with non-probability samples can be formulated as part of the solution $\hat{\boldsymbol{\eta}}$ to the following system of equations:

$$\mathbf{0} = \boldsymbol{s}_n(\boldsymbol{\eta}) = \frac{1}{N}\sum_{i\in\mathcal{S}_A}\boldsymbol{\psi}_A(\boldsymbol{O}_i, \boldsymbol{\eta}) + \frac{1}{N}\sum_{j=1}^{J}\sum_{i\in\mathcal{S}_{B_j}}\boldsymbol{\psi}_{B_j}(\boldsymbol{X}_{\mathcal{C},i}), \tag{3.9}$$

where

$$\boldsymbol{\psi}_A(\boldsymbol{O}_i, \boldsymbol{\eta}) = \frac{1}{\pi_{\boldsymbol{\theta}}(\boldsymbol{X}_{\mathcal{C},i})}\left\{\boldsymbol{X}_i{}^\intercal, \boldsymbol{h}(\boldsymbol{X}_i, A_i, \boldsymbol{\gamma})^\intercal, \frac{A_i g(\boldsymbol{X}_i)(Y_i - \tau_T)}{e_{\boldsymbol{\gamma}}(\boldsymbol{X}_i)}, \frac{(1-A_i)g(\boldsymbol{X}_i)(Y_i - \tau_T + \tau)}{1 - e_{\boldsymbol{\gamma}}(\boldsymbol{X}_i)}\right\}^\intercal$$

and $\boldsymbol{\psi}_{B_j}(\boldsymbol{X}_{\mathcal{C},i}) = -d_i^{B_j}\left(\boldsymbol{X}_{\mathcal{C},i}^\intercal \boldsymbol{W}_j, \mathbf{0}, 0, 0\right)^\intercal$. Let Q denote the joint model of q for $\mathcal{S}_A$ and the survey design q_j for $\mathcal{S}_{B_j}$, $j = 1, \cdots, J$. We assume that the sizes of the J probability samples, n_{B_j}, are of the same order as n_A. Let $\boldsymbol{\eta}^*$ denote the probability limit of $\hat{\boldsymbol{\eta}}$ under the joint model Q. Under suitable regularity conditions, we have

$$\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}^* = \mathrm{E}_Q\left\{\frac{\partial \boldsymbol{s}_n(\boldsymbol{\eta}^*)}{\partial \boldsymbol{\eta}^\intercal}\right\}^{-1}\boldsymbol{s}_n(\boldsymbol{\eta}^*) + o_p\Big(n_A^{-1/2}\Big).$$

A variance estimator of $\hat{\boldsymbol{\eta}}$ is obtained by plugging in an estimator for $\mathrm{E}_Q\left[\partial \boldsymbol{s}_n(\boldsymbol{\eta}^*)/\partial \boldsymbol{\eta}^\intercal\right]$ and a design-based estimator for $\mathrm{Var}_Q\{\boldsymbol{s}_n(\boldsymbol{\eta}^*)\}$ in the sandwich form. The first term can be consistently estimated by $\hat{N}^{-1}\sum_{i\in\mathcal{S}_A}\partial\boldsymbol{\psi}_A(\boldsymbol{O}_i, \hat{\boldsymbol{\eta}})/\partial\boldsymbol{\eta}^\intercal$. For the second term, since the sampling designs $q, q_q, \cdots, q_J$ are independent of each other, we have $\mathrm{Var}_Q\{\boldsymbol{s}_n(\boldsymbol{\eta}^*)\} = \mathrm{Var}_q\{\boldsymbol{s}_n(\boldsymbol{\eta}^*)\} + \sum_{j=1}^{J}\mathrm{Var}_{q_j}\{\boldsymbol{s}_n(\boldsymbol{\eta}^*)\}$, which can be estimated by

$$\frac{1}{\hat{N}}\sum_{i\in\mathcal{S}_A}\frac{1-\hat{\pi}_i^A}{(\hat{\pi}_i^A)^2}\boldsymbol{\psi}_A(\boldsymbol{O}_i, \hat{\boldsymbol{\eta}})^{\otimes 2} + \frac{1}{\hat{N}}\sum_{j=1}^{J}\sum_{i\in\mathcal{S}_{B_j}}\sum_{k\in\mathcal{S}_{B_j}}\frac{\pi_{ik}^{B_j} - \pi_i^{B_j}\pi_k^{B_j}}{\pi_{ik}^{B_j}\pi_i^{B_j}\pi_k^{B_j}}\boldsymbol{\psi}_{B_j}(\boldsymbol{X}_{\mathcal{C},i})^{\otimes 2},$$

where $\hat{\pi}_i^A = \pi_{\hat{\boldsymbol{\theta}}}(\boldsymbol{X}_{\mathcal{C},i})$ and $\hat{N} = \sum_{i\in\mathcal{S}_A}(\hat{\pi}_i^A)^{-1} = \sum_{j=1}^{J} a_{1j}\sum_{i\in\mathcal{S}_j} d_i^{B_j}$. Similar to discussions

in Section 3.3.4, approximate variance estimators without involving second order inclusion probabilities can be used in practice. The variance estimator does not rely on superpopulation models, although it may still be sensitive to model misspecifications for participation probabilities.

3.4.5 Numerical Studies with Non-Probability Survey Samples

We slightly modify the setup of the simulations with probability survey samples to evaluate the performance of the proposed methods with non-probability samples. We redo the data generating process given in Section 3.3.5 for the finite population data excepting that $\tau(\boldsymbol{X}_i)$ is reset and will be identified later. The survey sample $\mathcal{S}_A$ is selected based on the Poisson sampling method with the first inclusion probability

$$\pi_i^A = \text{expit}(0.1X_{1i} + 0.1X_{2i} + 0.1X_{3i} + 0.1X_{4i} + T_i + \beta_A)$$

where β_A is a controlling parameter such that $\sum_{i=1}^N \pi_i^A = n_A$. We create three reference probability samples, $\mathcal{S}_{B_1}$, $\mathcal{S}_{B_2}$, and $\mathcal{S}_{B_3}$, with the sample sizes $n_{B_1} = 1000$, $n_{B_2} = 1000$, and $n_{B_3} = 1000$, based on the randomized systematic PPS sampling (without replacement) method with the first inclusion probabilities $\pi_i^{B_1}$, $\pi_i^{B_2}$, and π^{B_3}, respectively. Both $\pi_i^{B_1}$ and $\pi_i^{B_3}$ are set to be proportional to $X_{3i} + c$ where c is chosen so that $\max_i\{X_{3i} + c\}/\min_i\{X_{3i} + c\} = 50$. The $\pi_i^{B_2}$ is proportional to $X_{4i} + c$ where c is set to satisfy $\max_i\{X_{4i}+c\}/\min_i\{X_{4i}+c\} = 50$. However, the observed auxiliary variables are different for the three reference samples. We consider two different designs as summarized in Table 3.4. The population and sample size are $(N, n_A) = (2e4, 500)$.

The key to estimating the population average treatment effects with non-probability samples is to create pseudo-weights by estimating the participation probabilities $\pi_i^A =$

$\pi(\boldsymbol{X}_{\mathcal{C},i})$. Then, the estimator of τ_g is obtained by solving the survey-weighted estimating equations weighted by the pseudo-weights. We apply the following methods to generate the pseudo-weights. In the first, the weights are set to be uniform (e.g., 1), denoted by "Unif". In the second, we assume that π_i^A are known and set the weights as $(\pi_i^A)^{-1}$. This essentially reduces to the SWPSW with known survey weights and will be denoted by "Oracle". For the rest, we estimate the participation probabilities using the calibration approach presented in Section 3.4.3 and create the weights by $(\hat{\pi}_i^A)^{-1}$. The external-information of the reference probability samples are given as follows.

- SB1: $\mathcal{S}_{B_1}$ is available.
- SB2: $\mathcal{S}_{B_2}$ is available.
- SB3: $\mathcal{S}_{B_3}$ is available.
- SB12: $\mathcal{S}_{B_1}$ and $\mathcal{S}_{B_2}$ are both available.

SB3 contains the entire confounding variables. SB1 and SB2 represent the scenario where the confounding variables are only partially observed in the reference sample. The two designs as described in Table 3.4 are for different purposes of the evaluation. The main focus of Design I is to figure out whether the estimator of τ_g can be improved by borrowing information from multiple reference samples (SB12), while the main focus of Design II is to study which covariates play the key role in identification. We evaluate these methods under the scenarios TT, TF, and FT where at least one superpopulation model is correct. We focus on estimating the PATE and calculate the performance measures based on 1000 Monte Carlo replicated datasets. The simulation results for Design I and Design II are summarized respectively in Tables 3.5 and 3.6.

Table 3.4: Designs for simulations with non-probability survey samples

Design	Observed covariates			$\tau(\boldsymbol{X}_i)$
	SB1	SB2	SB3	
I	X_{1i}, X_{2i}	X_{1i}, X_{4i}	same as $\mathcal{S}_A$	$X_{1i} + X_{2i} + 0.2X_{4i} + \beta_\tau$
II	X_{2i}	X_{1i}, X_{3i}, X_{4i}		$X_{2i} + \beta_\tau$

From Table 3.5, it can be seen that the Unif estimators are biased. The biases are mildly corrected after involving either $\mathcal{S}_{B_1}$ or $\mathcal{S}_{B_2}$ for correction and substantially removed after integrating $\mathcal{S}_{B_1}$ and $\mathcal{S}_{B_2}$ together (SB12). SB12 performs closely to SB3 in terms of the bias, and outperforms SB3 in terms of the statistical efficiency. Especially, when the PS function is estimated by the CBPSW method, SB12 is as good as Oracle. This numerical experiment showcases the advantages of integrating multiple reference datasets for the estimation when a single reference sample cannot provide the required auxiliary information: in our setup, SB12 not only removes the selection bias, but also improves the estimator's efficiency in comparison to SB1, SB2 and SB3.

For Design II, the participation probabilities are supposed to be better estimated using $\mathcal{S}_{B_2}$ than $\mathcal{S}_{B_1}$ because $\mathcal{S}_{B_2}$ contains more important predictors in the true participation probabilities. However, remember that our ultimate goal is to estimate the population average treatment effects rather than the participation probabilities. As argued in Proposition 3.4.1, $\mathcal{S}_{B_1}$ is preferred for identification of τ_g because the $\tau(\boldsymbol{X}_i)$ merely relies on X_{2i}. Otherwise, we should combine $\mathcal{S}_{B_1}$ and $\mathcal{S}_{B_2}$ together to catch all the predictors for the participation probabilities. Table 3.6 shows that SB1 removes more biases than SB2 as expected in Scenarios TT, TF and FT. However, since $\pi_{\boldsymbol{\theta}}$ is wrongly specified by simply involving X_{2i}, SB1 is mildly biased. SB12 still achieves the best performance among all other methods in terms of the bias, the RMSEs, and the coverage rates.

Table 3.5: Simulation results with non-probability survey samples under Design I.

		PSWMLE					CBPSW				
Scenario	Method	%RB	MCSD	AVESE	RMSE	%CP	%RB	MCSD	AVESE	RMSE	%CP
TT	Naive	61.27	0.83	0.83	2.01	41.30	61.27	0.83	0.83	2.01	41.30
	Unif	23.06	0.67	0.61	0.96	59.00	20.86	0.28	0.25	0.68	33.60
	Oracle	1.94	0.39	0.40	0.39	94.60	2.39	0.26	0.25	0.27	93.60
	SB1	12.95	0.61	0.59	0.72	76.40	11.36	0.27	0.24	0.43	67.50
	SB2	9.77	0.49	0.50	0.57	83.40	11.58	0.27	0.25	0.44	68.80
	SB3	0.45	0.54	0.49	0.54	93.60	2.42	0.27	0.24	0.28	92.20
	SB12	0.33	0.46	0.48	0.46	94.20	2.41	0.26	0.24	0.27	92.60
TF	Naive	61.27	0.83	0.83	2.01	41.30	61.27	0.83	0.83	2.01	41.30
	Unif	19.98	0.48	0.46	0.76	63.00	20.27	0.26	0.24	0.66	31.60
	Oracle	2.26	0.33	0.35	0.34	95.30	1.63	0.25	0.24	0.25	93.90
	SB1	9.95	0.44	0.44	0.53	81.70	10.56	0.25	0.23	0.40	70.50
	SB2	7.06	0.39	0.39	0.45	86.90	11.08	0.25	0.24	0.42	69.20
	SB3	-2.10	0.44	0.39	0.45	95.60	1.67	0.25	0.23	0.26	93.30
	SB12	-2.33	0.37	0.38	0.38	95.90	1.69	0.25	0.23	0.25	93.20
FT	Naive	61.27	0.83	0.83	2.01	41.30	61.27	0.83	0.83	2.01	41.30
	Unif	21.25	0.90	0.84	1.10	79.70	20.39	0.75	0.71	0.97	79.30
	Oracle	0.39	0.74	0.74	0.74	94.70	1.07	0.71	0.70	0.71	93.90
	SB1	11.90	0.84	0.81	0.91	88.40	11.27	0.72	0.70	0.79	88.80
	SB2	7.64	0.78	0.77	0.81	91.80	9.94	0.73	0.71	0.79	89.40
	SB3	-0.70	0.79	0.75	0.79	94.70	1.34	0.72	0.70	0.72	94.40
	SB12	-1.03	0.75	0.75	0.75	94.70	1.22	0.71	0.70	0.71	94.10

Table 3.6: Simulation results with non-probability survey samples under Design II.

		PSWMLE					CBPSW				
Scenario	Method	%RB	MCSD	AVESE	RMSE	%CP	%RB	MCSD	AVESE	RMSE	%CP
TT	Naive	53.37	0.82	0.82	1.79	49.40	53.37	0.82	0.82	1.79	49.40
	Unif	15.17	0.67	0.61	0.81	72.40	13.08	0.27	0.25	0.48	59.90
	Oracle	1.73	0.38	0.40	0.39	94.20	2.20	0.26	0.24	0.27	93.20
	SB1	5.02	0.61	0.59	0.63	87.30	3.44	0.26	0.24	0.28	90.80
	SB2	9.17	0.50	0.51	0.57	84.80	10.85	0.27	0.25	0.42	71.20
	SB3	0.36	0.54	0.49	0.54	93.80	2.27	0.26	0.24	0.27	92.50
	SB12	0.27	0.46	0.49	0.46	94.20	2.30	0.26	0.24	0.27	93.30
TF	Naive	53.37	0.82	0.82	1.79	49.40	53.37	0.82	0.82	1.79	49.40
	Unif	12.09	0.48	0.46	0.60	78.30	12.52	0.26	0.24	0.45	61.20
	Oracle	2.12	0.33	0.34	0.33	95.60	1.52	0.25	0.24	0.25	93.80
	SB1	2.05	0.45	0.44	0.45	92.10	2.69	0.25	0.23	0.26	92.20
	SB2	6.96	0.39	0.39	0.44	87.00	10.96	0.25	0.24	0.41	68.90
	SB3	-2.18	0.42	0.39	0.42	96.00	1.58	0.25	0.23	0.25	92.60
	SB12	-2.24	0.39	0.38	0.40	95.70	1.60	0.24	0.24	0.25	93.90
FT	Naive	53.37	0.82	0.82	1.79	49.40	53.37	0.82	0.82	1.79	49.40
	Unif	13.36	0.90	0.84	0.98	86.10	12.60	0.75	0.71	0.84	86.30
	Oracle	0.21	0.74	0.74	0.74	94.70	0.90	0.71	0.70	0.71	94.00
	SB1	3.94	0.84	0.82	0.85	92.20	3.38	0.72	0.70	0.72	92.20
	SB2	7.28	0.79	0.77	0.82	91.70	9.40	0.74	0.71	0.79	89.60
	SB3	-0.88	0.77	0.75	0.77	94.40	1.12	0.71	0.70	0.71	94.10
	SB12	-1.08	0.76	0.76	0.76	94.50	1.08	0.71	0.70	0.71	94.50

3.5 Real Data Analysis

In this section, we apply the proposed SWPSW methods to analyze the Medical Expenditure Panel Survey (MEPS) data. The datasets were collected to obtain national estimates of health care expenditures in the US and contained rich individual-level information on demographic, medical insurance, and health status variables. Lê Cook et al. (2010) and Li et al. (2018) used the earlier datasets to estimate racial disparities in health care expenditures. Compared to the simple difference, the disparity is an unconfounded descriptive measure that evaluates group differences due to socioeconomic status and discrimination rather than the differences in clinical need and patient preferences (Lê Cook et al., 2010). Therefore, to infer the racial disparity, one needs to balance the distributions of certain confounding variables (e.g., the ones reflecting health status) among different racial groups. The MEPS dataset is a probability survey sample. However, survey weights were not used by Lê Cook et al. (2010) and Li et al. (2018), and the reported estimates may not be valid for the target population in the given year.

We illustrate our proposed methods for drawing population-level conclusions using the 2020 MEPS data and focusing on the comparison between Hispanics and non-Hispanic Whites racial groups. The 2020 MEPS data are linked to the National Health Interview Survey (NHIS) which has a multi-stage survey design and utilizes the systematic sampling method. The final personal level survey weights are further calibrated to the Current Population Survey (CPS) control totals based on six geographical and demographic variables: education, region, race, sex, metropolitan statistical area status, and age. Because of the potential conditional dependence of calibrated survey weights on the treatment variable "race" given the confounding variables, the survey weights are highly encouraged to be incorporated in estimating the PS for the target population. The potential inflated variance

of the estimators induced by the calibration step will be ignored in our analysis.

The survey sample has a total size of $26,847$ with $6,602$ Hispanics, $14,112$ (non-Hispanic) Whites, $3,788$ Blacks, $1,395$ Asians, and 950 other races. We focus on the comparison between Hispanics and non-Hispanic Whites and conduct our analysis using a sub-sample of size $20,714$. Based on the earlier analysis of Lê Cook et al. (2010), we consider 20 confounding variables (16 categorical, 4 continuous), including some demographic factors, chronic conditions, and health status measures, as predictors in the PS model. The health expenditures show a heavy tailed distribution, with some elder individuals having very large expenditures. These extreme values may be of concerns for asymptotic behaviours of our proposed estimators. We consider two analysis strategies for this application. The first is to consider a sub-population with individuals aged between 18 and 59. Figure 3.1 shows that this sub-population does not contain as many extreme outcomes as the initial one. Our goal is to make statistical inference on the SATE and the SATT over the subspace $\mathcal{A} = \{18 \leq \text{age} \leq 59\}$. The second analysis is to re-scale the outcomes by a log-transformation and focuses on estimating the PATE and the PATT. In this case, we define $\tau(\boldsymbol{X}_i) = \mathrm{E}[\log(Y_i) \mid A_i = 1, \boldsymbol{X}_i] - \mathrm{E}[\log(Y_i) \mid A_i = 0, \boldsymbol{X}_i]$. Figure 3.2 shows that the outcomes are much more stable in the log scale. A few missing values in the data are imputed with the median of the corresponding variable.

In an earlier analysis of the MEPS datasets, DuGoff et al. (2014) utilized the survey weights to re-weight the outcomes to estimate the average treatment effects, but they ignored the survey weights when estimating the PS function. Their methods may not be justifiable because the survey weights could rely on the treatment variable "race". In our current analyses, we incorporated the survey weights in the estimation of the PS function and applied the proposed SWPSW method for the estimation of the parameters of interest.

We use the absolute standardized mean difference (Austin, 2011) (ASMD) to com-

pare the performances of five different weighting methods on covariate-balancing: Naive, PSWMLE, CBPSW, SWPSWMLE, and SWCBPSW. Figures 3.3 and 3.4 show the ASMDs of some confounding variables after weighting using different methods among groups. The weights in Figure 3.3 are obtained for the subpopulation and the ones in Figure 3.4 are obtained for the entire population. It is generally believed that if the ASMD is greater than 0.1, there exists covariate im-balance between two groups, given by the red and dashed vertical lines. We find that the covariates are better balanced between the treated and the control groups after propensity score weighting. Especially for CBPSW, the covariate mean balance between the treated and control groups are perfectly reached. However, the ASMDs for some variables such as marriage and physical status are still greater than 0.1 between the treatment groups and the target population, implying that a three-way covariate mean balance has not been achieved. This issue is alleviated after incorporating survey weights. In particular, the SWCBPSW approach achieves excellent three-way covariate mean balance.

The point estimates and corresponding confidence intervals for the SATE and the SATT are given in Table 3.7. The naive PSW estimates in Table 3.7 show that the racial disparity is significantly bigger than 0 in the sample level. On the other hand, the survey-weighted estimates become smaller, and the corresponding CIs are wider and cover 0. This indicates that the racial disparity in health expenditures between Hispanics and non-Hispanic Whites is not significant in the sub-population $\{i : 18 \leq \text{age}_i \leq 59\}$. Table 3.8 summarizes the results for the estimation of the PATE and the PATT with log-transformed outcomes. From Table 3.8, we can conclude that the racial disparity in health expenditures under the log-scale between Hispanics and non-Hispanic Whites is significant over the entire US population.

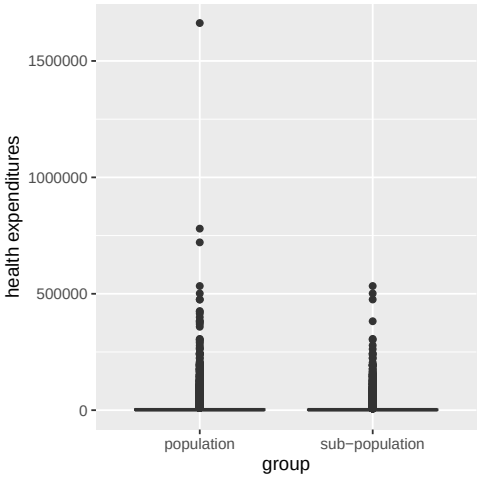


Figure 3.1: Boxplot of the health expenditures in the population and in the sub-population

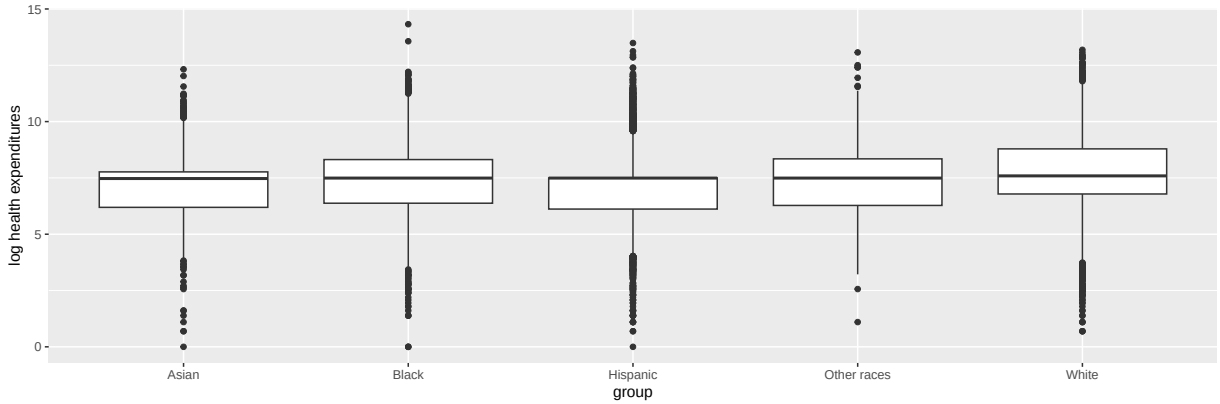


Figure 3.2: Boxplot of the log health expenditures for different racial groups

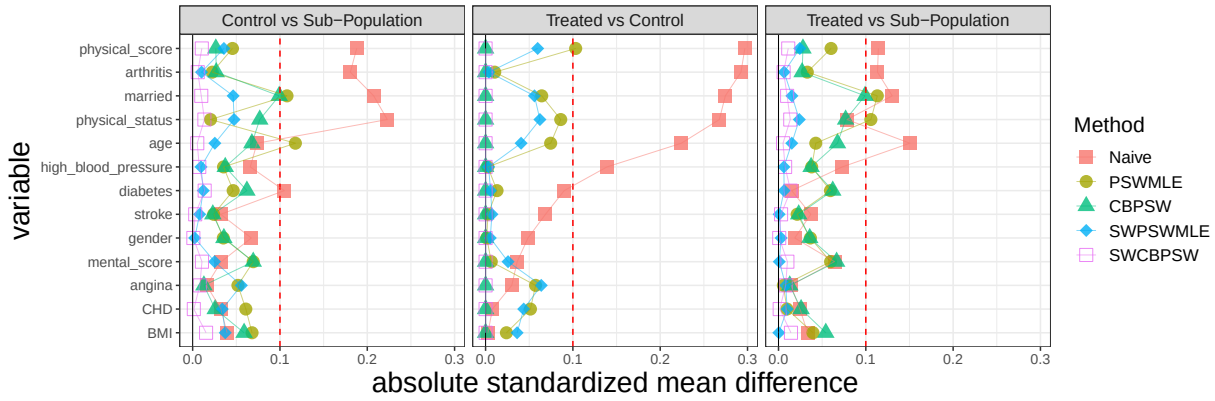


Figure 3.3: Covariate-balance among groups using the absolute standardized mean difference. The weights are obtained for estimating the SATE.

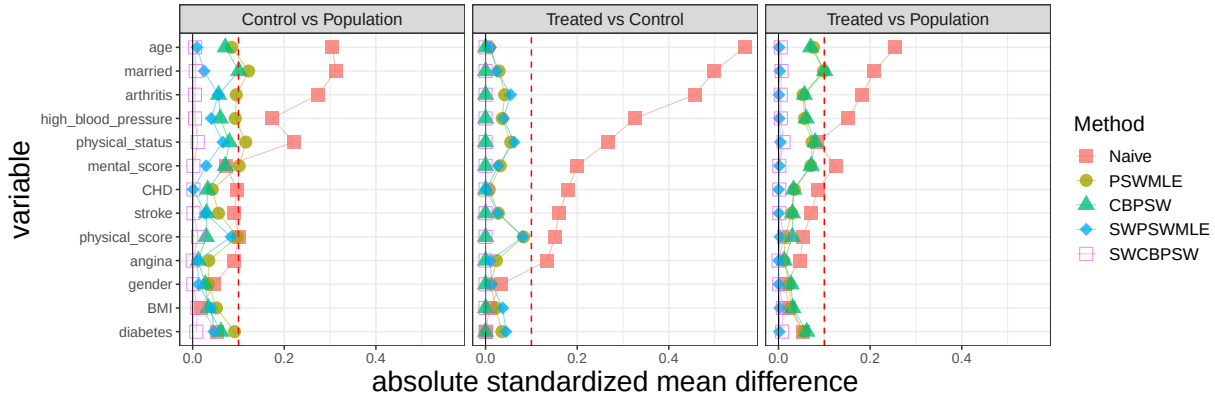


Figure 3.4: Covariate-balance among groups using the absolute standardized mean difference. The weights are obtained for estimating the PATE with log-scaled outcomes.

Table 3.7: Estimates and 95% confidence intervals of the racial disparity in health expenditures between Hispanics and non-Hispanic Whites.

	SATE		SATT	
Method	Estimate	95% CI	Estimate	95% CI
Naive	2101	$(1403, 2799)$	2101	$(1403, 2799)$
PSWMLE	1519	$(659, 2379)$	1238	$(275, 2200)$
CBPSW	1509	$(656, 2362)$	1301	$(374, 2227)$
SWPSWMLE	823	$(-477, 2124)$	606	$(-830, 2042)$
SWCBPSW	900	$(-342, 2143)$	746	$(-596, 2087)$

Table 3.8: Estimates and 95% confidence intervals of the racial disparity in log health expenditures between Hispanics and non-Hispanic Whites.

	PATE		PATT	
Method	Estimate	95% CI	Estimate	95% CI
Naive	0.395	$(0.097, 0.693)$	0.395	$(0.097, 0.693)$
PSWMLE	0.297	$(0.236, 0.358)$	0.273	$(0.197, 0.350)$
CBPSW	0.345	$(0.293, 0.396)$	0.343	$(0.283, 0.403)$
SWPSWMLE	0.277	$(0.207, 0.347)$	0.259	$(0.179, 0.340)$
SWCBPSW	0.325	$(0.265, 0.386)$	0.322	$(0.256, 0.388)$

3.6 Some Additional Remarks

We proposed a survey-weighted propensity score weighting approach to causal inference with observational survey data in this chapter, and establish a theoretical framework on the consistency of the point estimators and valid variance estimation for general causal parameters for the target population or sub-population. It is known in the literature, however, that survey weighted approach can lead to less efficient estimators in terms of inflated variance if the weighting is not necessary.

Under our current setting, analysis involving survey weights may not be required for the following scenarios: (1) The objective is to estimate the sample average treatment effects; (2) The treatment effect is homogeneous within the target population; (3) The CATE function and the selection procedure have no factors in common. In the absence of clear identification of these scenarios, survey weighted methods are strongly recommended. The real data example shows that the proposed survey weighted CBPSW approach is generally favored as it achieves excellent three-way covariate-balance.

Causal inference with observational non-probability survey samples is an emerging research topic that has not been studied in the existing literature. Valid statistical methods require auxiliary population information from other sources such as a reference probability survey sample. We derive a set of identification assumptions for the population quantities of interest when only a partial set of confounding variables is observed in the reference sample. However, these assumptions are difficult to verify in practice. Our recommendation is to search from multiple external sources to include as many relevant covariates as possible in the analysis using estimation strategies we proposed in Section 3.4.3.

3.7 Proofs and Technical Details

3.7.1 Derivation of the Dual Problems

Recall and follow the notation in Chapter 3, the survey-weighted covariate balancing propensity score weighting approach is formulated as solving the system of estimating equations

$$\sum_{i\in\mathcal{S}_A} d_i^A \boldsymbol{\psi}_{g,\boldsymbol{h}}(\boldsymbol{O}_i, \boldsymbol{\eta}) = \mathbf{0} \tag{3.10}$$

where $\boldsymbol{h} = \boldsymbol{h}_{CB,g}$ are a set of just-identified covariate balancing conditions. When $g(\boldsymbol{x}) = 1_{\mathbb{B}}(\boldsymbol{x})$ (SATE), the estimator of τ_g by solving (3.10) has the form

$$\hat{\tau} = \frac{\sum_{i\in\mathcal{S}_A} A_i\hat{\omega}_i Y_i}{\sum_{i\in\mathcal{S}_A} A_i\hat{\omega}_i} - \frac{\sum_{i\in\mathcal{S}_A}(1-A_i)\hat{\omega}_i Y_i}{\sum_{i\in\mathcal{S}_A}(1-A_i)\hat{\omega}_i}$$

where $\hat{\omega}_i, i \in \mathcal{S}_A$ are the solutions of the optimization problem:

$$\begin{aligned} \min_{\omega_i \geq \tilde{d}_i^A} & \sum_{i\in\mathcal{S}_A} (\omega_i - \tilde{d}_i^A)\log(\omega_i - \tilde{d}_i^A) - \omega_i(1+\log(\tilde{d}_i^A)) \\ \text{s.t.} & \sum_{i\in\mathcal{S}_A, A_i=1} \omega_i \boldsymbol{\phi}(\boldsymbol{X}_i) = \sum_{i\in\mathcal{S}_A, A_i=0} \omega_i \boldsymbol{\phi}(\boldsymbol{X}_i). \end{aligned} \tag{3.11}$$

Here, we denote $\tilde{d}_i^A = d_i^A 1_{\mathbb{B}}(\boldsymbol{X}_i)$ for simplicity. The objective function of (3.11) reaches its minimums at $\omega_i = \tilde{d}_i^A$ for $i \in \mathcal{S}_A$. When $\boldsymbol{h} = \boldsymbol{h}_{CB,g}$ and $g(\boldsymbol{x}) = e_{\boldsymbol{\gamma}}(\boldsymbol{x})1_{\mathbb{B}}(\boldsymbol{x})$ (SATT), the estimator of τ_g has the form

$$\hat{\tau} = \frac{\sum_{i\in\mathcal{S}_A} A_i\tilde{d}_i^A Y_i}{\sum_{i\in\mathcal{S}_A} A_i\tilde{d}_i^A} - \frac{\sum_{i\in\mathcal{S}_A}(1-A_i)\hat{\omega}_i Y_i}{\sum_{i\in\mathcal{S}_A}(1-A_i)\hat{\omega}_i}$$

where $\hat{\omega}_i, i \in \mathcal{S}_A, A_i = 0$ are the solutions of

$$\begin{aligned} &\min_{\omega_i > 0} \sum_{i \in \mathcal{S}_A, A_i = 0} \omega_i \log(\omega_i / \tilde{d}_i^A) - \omega_i + \tilde{d}_i^A \\ &\text{s.t.} \sum_{i \in \mathcal{S}_A, A_i = 0} \omega_i \boldsymbol{\phi}(\boldsymbol{X}_i) = \sum_{i \in \mathcal{S}_A, A_i = 1} \tilde{d}_i^A \boldsymbol{\phi}(\boldsymbol{X}_i) \end{aligned} \tag{3.12}$$

The objective function of (3.12) is exactly the Kullback–Leibler divergence which also reaches its minimums at $\omega_i = \tilde{d}_i^A$ for $i \in \mathcal{S}_A, A_i = 0$.

Derivation of the above dual problems takes advantage of the technique described in Zhao (2019). In the following, we show how to obtain the dual problem for estimating the SATE. The dual problem for estimating the SATT can be derived in a similar manner. First, when $g(\boldsymbol{x}) = 1_{\mathbb{B}}(\boldsymbol{x})$, solving for (3.10) can be formulated equivalently as

$$\max_{\boldsymbol{\gamma}, \zeta} \sum_{i \in \mathcal{S}_A} \tilde{d}_i^A S(f(\zeta_i), A_i), \tag{3.13}$$

$$\text{s.t. } \zeta_i = \boldsymbol{\gamma}^\intercal \boldsymbol{\phi}(\boldsymbol{X}_i), i \in \mathcal{S}_A, \tag{3.14}$$

where $\tilde{d}_i^A = d_i^A 1_{\mathbb{B}}(\boldsymbol{X}_i)$, $f(x) = \text{expit}(x) = \exp(x)/[1+\exp(x)]$ and $S(p,t) = \int (t-p)p^{-2}(1-p)^{-2}dp$. Define the Lagrange function as

$$L(\boldsymbol{\gamma}, \boldsymbol{\zeta}, \lambda) = \sum_{i \in \mathcal{S}_A} \tilde{d}_i^A S(f(\zeta_i), A_i) + \sum_{i \in \mathcal{S}_A} \lambda_i (\boldsymbol{\gamma}^\intercal \boldsymbol{\phi}(\boldsymbol{X}_i) - \zeta_i)$$

where $\boldsymbol{\zeta} = (\zeta_1, \cdots, \zeta_N)^\intercal$. The Lagrangian dual problem of (3.10) is given by

$$\min_{\lambda} \max_{\boldsymbol{\gamma}, \boldsymbol{\zeta}} L(\boldsymbol{\gamma}, \boldsymbol{\zeta}, \lambda)$$

with the duality gap equal to 0 due to Slater's condition (Boyd et al., 2004). Taking the

first derivative of L with respect to $\boldsymbol{\gamma}$ and ζ we obtain

$$0 = \frac{\partial L}{\partial \zeta_i} = \tilde{d}_i^A (A_i - f(\zeta_i)) f(\zeta_i)^{-1} (1 - f(\zeta_i))^{-1} - (2A_i - 1)\omega_i, i \in \mathcal{S}_A, \tag{3.15}$$

$$0 = \frac{\partial L}{\partial \boldsymbol{\gamma}} = \sum_{i \in \mathcal{S}_A} (2A_i - 1)\omega_i \boldsymbol{\phi}(\boldsymbol{X}_i) = 0, \tag{3.16}$$

where $\omega_i = (2A_i - 1)^{-1}\lambda_i$. Recall that $e_{\boldsymbol{\gamma}}(\boldsymbol{x})$ is postulated as a logistic model, then we have $f(\zeta_i) = \text{expit}(\boldsymbol{\gamma}^\intercal \boldsymbol{\phi}(\boldsymbol{X}_i)) = e_{\boldsymbol{\gamma}}(\boldsymbol{X}_i)$. Thus, it follows from (3.15) that $\omega_i = \tilde{d}_i^A \{A_i / e_{\boldsymbol{\gamma}}(\boldsymbol{X}_i) + (1 - A_i)/(1 - e_{\boldsymbol{\gamma}}(\boldsymbol{X}_i))\}$ for $i \in \mathcal{S}_A$ which are exactly the (non-normalized) weights for the observations. Let ω_i be the dual variable of interest and $\tilde{L}(\omega, \boldsymbol{\gamma}) = \max_\zeta L(\boldsymbol{\gamma}, \zeta, \lambda)$. By solving (3.15) we get $\zeta_i = A_i f^{-1}(\tilde{d}_i^A/\omega_i) + (1 - A_i) f^{-1}(1 - \tilde{d}_i^A/\omega_i)$ and $\omega_i \geq \tilde{d}_i^A$ for $i \in \mathcal{S}_A$. By substituting ζ_i, simple algebra gives

$$\begin{aligned}
\tilde{L}(\omega, \boldsymbol{\gamma}) &= \sum_{i \in \mathcal{S}_A} A_i \int \frac{-\omega_i}{\omega_i - \tilde{d}_i^A} d\omega_i - A_i \log\left(\frac{\tilde{d}_i^A}{\omega_i - \tilde{d}_i^A}\right) + (1 - A_i) \log\left(\frac{\omega_i - \tilde{d}_i^A}{\tilde{d}_i^A}\right) \\
&\quad + \sum_{i \in \mathcal{S}_A} (2A_i - 1)\omega_i \boldsymbol{\gamma}^\intercal \boldsymbol{\phi}(\boldsymbol{X}_i) \\
&= \sum_{i \in \mathcal{S}_A} (\omega_i - \tilde{d}_i^A) \log(\omega_i - \tilde{d}_i^A) - \omega_i (1 + \log(\tilde{d}_i^A)) + \sum_{i \in \mathcal{S}_A} (2A_i - 1)\omega_i \boldsymbol{\gamma}^\intercal \boldsymbol{\phi}(\boldsymbol{X}_i), \omega_i \geq \tilde{d}_i^A.
\end{aligned}$$

By noting the Lagrangian dual problem is equivalent to $\min_\omega \max_{\boldsymbol{\gamma}} \tilde{L}(\omega, \boldsymbol{\gamma})$, we obtain (3.11).

3.7.2 Proof of Proposition 3.3.1

Let $\tau_{g,T} = \mathrm{E}\{g(\boldsymbol{X})\mathrm{E}(Y \mid \boldsymbol{X}, A = 1)\}/\mathrm{E}[g(\boldsymbol{X})]$, $\boldsymbol{\gamma}_0$ be the true value of $\boldsymbol{\gamma}$ under correct model specification, and $\boldsymbol{\eta}_0 = (\boldsymbol{\gamma}_0^\intercal, \tau_{g,T}, \tau_g)^\intercal$. Under correct PS model, we have $\mathrm{E}\left\{\sum_{i \in \mathcal{S}_A} d_i^A \boldsymbol{\psi}_{g,\boldsymbol{h}}(\boldsymbol{O}_i, \boldsymbol{\eta}_0)\right\} = 0$ under the TPR framework which has unique solutions.

Thus, the $\hat{\boldsymbol{\eta}}$ is a standard M-estimator which is consistent under certain regularity conditions.

To prove the consistency of the proposed estimators under wrong PS model, we denote by

$$\hat{\omega}_i = d_i^A \left\{ \frac{A_i g(\boldsymbol{X}_i)}{e_{\hat{\boldsymbol{\gamma}}}(\boldsymbol{X}_i)} + \frac{(1-A_i) g(\boldsymbol{X}_i)}{1 - e_{\hat{\boldsymbol{\gamma}}}(\boldsymbol{X}_i)} \right\}$$

the estimated weights and $\hat{\omega}_i^* = \hat{\omega}_i / 2 \sum_{i \in \mathcal{S}_A} \hat{\omega}_i$ the normalized weights so that $\sum_{i \in \mathcal{S}_A} A_i \hat{\omega}_i^* = \sum_{i \in \mathcal{S}_A} (1 - A_i) \hat{\omega}_i^* = 1$. Then, simple algebra shows

$$\begin{aligned} \hat{\tau} - \tau_g = & \sum_{i \in \mathcal{S}_A} (2A_i - 1) \hat{\omega}_i^* (Y_i - m(\boldsymbol{X}_i) - A_i \tau(\boldsymbol{X}_i)) + \sum_{i \in \mathcal{S}_A} (2A_i - 1) \hat{\omega}_i^* m(\boldsymbol{X}_i) \\ & + \left[\sum_{i \in \mathcal{S}_A} A_i \hat{\omega}_i^* \tau(\boldsymbol{X}_i) - \tau_g \right]. \end{aligned} \tag{3.17}$$

The first term at the right side of the equation (3.17) vanishes because

$$\mathrm{E}\left(Y_i - m(\boldsymbol{X}_i) - A_i \tau(\boldsymbol{X}_i) \mid \boldsymbol{X}_i, A_i\right) = 0.$$

The second term is exactly zero when $m(\boldsymbol{X}_i)$ is linear in $\boldsymbol{X}_i$ due to the covariate-balancing property of the weights. For the estimation of the SATT, we have $\hat{\omega}_i = d_i^A A_i 1_{\mathbb{B}}(\boldsymbol{X}_i)$ for $A_i = 1$ and thus the third term converges in probability to zero because $\tau_g = \mathrm{E}[\hat{\omega}_i A_i \tau(\boldsymbol{X}_i)] / \mathrm{E}[\hat{\omega}_i A_i]$. For the estimation of the SATE, the third term can not be ignored unless $\tau(\boldsymbol{X}_i)$ is a constant.

3.7.3 Proof of Proposition 3.4.1

Simple algebra gives that

$$\mathrm{E}\left\{\sum_{i=1}^{N}\frac{R_i}{\pi(\boldsymbol{X}_{\mathcal{C},i})}\boldsymbol{h}(\boldsymbol{X}_i,A_i,\boldsymbol{\gamma}_0)\mid\boldsymbol{X}_i\right\}=\sum_{i=1}^{N}\frac{\mathrm{E}_q(R_i|\boldsymbol{X}_i)}{\pi(\boldsymbol{X}_{\mathcal{C},i})}\mathrm{E}_s\{\boldsymbol{h}(\boldsymbol{X}_i,A_i,\boldsymbol{\gamma}_0)\mid\boldsymbol{X}_i\}=0.$$

The first equation is because $A_i \perp\!\!\!\perp R_i \mid \boldsymbol{X}_i$. The last equation holds due to

$$\mathrm{E}_s\left\{\boldsymbol{h}(\boldsymbol{X}_i,A_i,\boldsymbol{\gamma}_0)\mid\boldsymbol{X}_i\right\}=0.$$

Simple algebra gives that

$$\begin{aligned}&\mathrm{E}\left\{\begin{array}{c}\sum_{i=1}^{N}\frac{R_iA_i}{\pi(\boldsymbol{X}_{\mathcal{C},i})e_{\boldsymbol{\gamma}_0}(\boldsymbol{X}_i)}g(\boldsymbol{X}_i)(Y_i-\tau_T)\\ \sum_{i=1}^{N}\frac{R_i(1-A_i)}{\pi(\boldsymbol{X}_{\mathcal{C},i})(1-e_{\boldsymbol{\gamma}_0}(\boldsymbol{X}_i))}g(\boldsymbol{X}_i)(Y_i-\tau_T+\tau_g)\end{array}\right\}\\ =&\mathrm{E}\left\{\begin{array}{c}\sum_{i=1}^{N}\frac{R_i}{\pi(\boldsymbol{X}_{\mathcal{C},i})}g(\boldsymbol{X}_{\mathcal{C},i})\left[m(\boldsymbol{X}_i)+\tau(\boldsymbol{X}_i)-\tau_T\right]\\ \sum_{i=1}^{N}\frac{R_i}{\pi(\boldsymbol{X}_{\mathcal{C},i})}g(\boldsymbol{X}_{\mathcal{C},i})\left[m(\boldsymbol{X}_i)-\tau_T+\tau_g\right]\end{array}\right\}.\end{aligned}$$

Since τ_T is only a nuisance parameter for estimating τ_g, it suffices to verify that

$$\mathrm{E}\left\{\frac{R_i}{\pi(\boldsymbol{X}_{\mathcal{C},i})}g(\boldsymbol{X}_{\mathcal{C},i})[\tau(\boldsymbol{X}_i)-\tau_g]\right\}=0.$$

It follows that

$$\begin{aligned}&\mathrm{E}\left\{\frac{R_i}{\pi(\boldsymbol{X}_{\mathcal{C},i})}g(\boldsymbol{X}_{\mathcal{C},i})[\tau(\boldsymbol{X}_i)-\tau_g]\right\}\\ =&\mathrm{E}_s\left\{g(\boldsymbol{X}_{\mathcal{C},i})[\tau(\boldsymbol{X}_i)-\tau_g]\right\}=0.\end{aligned}$$

The first equation holds due to $\tau(\boldsymbol{X}_i) \perp\!\!\!\perp R_i \mid \boldsymbol{X}_{\mathcal{C},i}$.

3.7.4 Optimal Choice of the Tuning Parameters w_{kj}

The term $\widehat{T}_k^w = \sum_{j:k\in\mathcal{C}_j} w_{kj}\hat{T}_{kj}$ with $\sum_{j:k\in\mathcal{C}_j} w_{kj} = 1$ is always a consistent estimator of the population total $T_k = \sum_{i=1}^N X_{k,i}$. One of the criteria of setting w_{kj} is trying to maximize the asymptotic efficiency of $\widehat{T}_k^w$. By the Lagrange multiplier method, one can verify that letting $w_{kj} = V_{kj}^{-1}/\sum_{j:k\in\mathcal{C}_j} V_{kj}^{-1}$ for $j : k \in \mathcal{C}_j$ leads to the smallest design-based variance of $\widehat{T}_k^w$ where V_{kj} is the design-based variance of $\hat{T}_{kj}$. Hence, given k, an optimal choice of w_{kj} under the above criterion is

$$w_{kj} = \frac{\hat{V}_{kj}^{-1}}{\sum_{j:k\in\mathcal{C}_j} \hat{V}_{kj}^{-1}} \text{ for } j : k \in \mathcal{C}_j$$

where

$$\hat{V}_{kj} = \sum_{i\in\mathcal{S}_{B_j}} \sum_{l\in\mathcal{S}_{B_j}} \frac{\pi_{il}^{B_j} - \pi_i^{B_j}\pi_l^{B_j}}{\pi_{il}^{B_j}} \frac{X_{k,i}}{\pi_i^{B_j}} \frac{X_{k,l}}{\pi_l^{B_j}}$$

is a design-consistent estimate of V_{kj}, and $\pi_i^{B_j}$ and $\pi_{il}^{B_j}$ are respectively the first and second-order inclusion probabilities for $\mathcal{S}_{B_j}$. Note that for many sampling designs, V_{kj} is of order $O(N^2/n_{B_j}^*)$ under suitable regularities where $n_{B_j}^* = \mathrm{E}_{q_j}(n_{B_j})$. Then, $V_{kj}^{-1}/\sum_{j:k\in\mathcal{C}_j} V_{kj}^{-1}$ would be of the order $O(n_{B_j}/\sum_{k\in\mathcal{C}_i} n_{B_i})$.

Chapter 4

Estimation of Treatment Harm Rate via Partitioning

4.1 Introduction

We have studied weighted average treatment effects in Chapter 3. These quantities defined as the difference of two means in a population or in a subgroup do not fully reveal treatment effect heterogeneity (TEH). To obtain a more comprehensive understanding of TEH, it is necessary to consider the joint distribution of counterfactual outcomes, which is encapsulated by the treatment harm rate (THR) in a randomized controlled trial (RCT) when the counterfactual outcomes are binary. The THR is defined as the proportion of individuals negatively affected by a treatment in a given population and reflects the average risk in a population associated with the treatment. Exploring the THR complements existing research on TEH and is increasingly important in real-world applications.

The THR is generally nonidentifiable even in a RCT, unless we impose some strong

conditions on the data structure such as conditional independence (Zhang et al., 2013; Shen et al., 2013) or joint distribution restrictions (Huang et al., 2019). Gadbury et al. (2004) studied the partial identifiability and derived the sharp bounds (Fréchet-Hoeffding bounds) of the THR in RCTs without covariates. The Fréchet-Hoeffding bounds of the THR were later extended to observational studies in the presence of covariates (Zhang et al., 2013; Kallus, 2022). Wu et al. (2024) further improved the bounds assuming positive conditional correlation between the two potential outcomes and developed a sensitivity analysis framework for the assumption.

In RCTs, it is common to report the estimated average treatment effect of the clinical subjects, which has been justified at the design stage by randomized treatment assignments. However, there are two important considerations on the estimators of the THR that should also be reported for RCTs. **First, auxiliary information should be effectively incorporated in the estimation of the THR to reduce the conservativeness of the estimators.** For example, the interval estimator provided by Gadbury et al. (2004) might be too wide to be practically informative. The width of the interval estimator can be substantially reduced by effective use of the auxiliary covariates information. **Second, justification of the estimators of the THR after covariate adjustment should not involve any untestable identification or ambiguous model assumptions other than treatment randomization**. For instances, Huang et al. (2019) required the users to impose zero proportion restrictions on some potential outcome pairs in order to identify the THR; Zhang et al. (2013) and Shen et al. (2013) assumed that the two potential outcomes are conditionally independent given the covariates. These identification assumptions on the joint distribution of the potential outcomes are un-testable and conclusions based on which are thus unreliable. One might argue that sensitivity analyses can be incorporated to address issues of assumption violations, but choosing a reasonable range

of the sensitivity parameter brings in another challenging issue in the absence of sufficient background knowledge. The recent work of Kallus (2022) tried to estimate the Fréchet-Hoeffding bounds of the THR under minimal identification assumptions. However, the bounds estimators proposed in the paper are valid only when the estimators of some nuisance parameters converge fast enough, say in the rate of $o_p(n^{-\gamma})$ for some $0 \leq \gamma < 1$, to the true value. These model assumptions are difficult to verify and often not satisfied in practice.

In this chapter, we develop a class of partitioning-based interval estimators for the THR by taking into account of the two aforementioned considerations. Our estimators are motivated from a key observation that the sharp bounds of the THR can be attained under a partition of the covariates space with at most four cells and the quantity of interest in each partition can be consistently estimated under the randomization. We show that the optimal partitioning rule is the Bayes classifier for a weighted multiclass classification problem. Probabilistic classification algorithms are employed to estimate the nuisance parameters and to form the partitioning. The proposed interval estimators are valid under the randomization regardless of the accuracy of the estimators for the nuisance parameters. However, the latter estimators can be highly informative if the underlying models capture the true data patterns. The proposed interval estimators are in a spirit similar to the model-assisted estimators in survey sampling and causal inference (Wu and Sitter, 2001; Wu, 2003; Wu and Thompson, 2020; Ye et al., 2023).

The rest of this chapter is organized as follows. In Section 4.2, we derive the partitioning-based bounds for the THR and show that a partition of four cells can yield the sharp bounds of the THR under standard assumptions used in causal inference literature. In Section 4.3, we show how to estimate the partitioning-based bounds given a partition under the randomization of treatment assignments. We discuss in Section 4.4 on how to

create a partition by plugging in probabilistic classifiers and how to create pseudo external data by cross-fitting. Confidence intervals for the partitioning-based bounds, which can be computed numerically based on the asymptotic distributions, are provided in Section 4.5 for the evaluation of the estimated bounds. In Section 4.6, we report results from Monte Carlo simulation studies to assess the finite-sample performance of the proposed interval estimators. An application of the proposed estimators to analyzing the AIDS Clinical Trials Group Study 175 is presented in Section 4.7, and some concluding remarks are given in Section 4.8. Proofs and technical details as well as additional simulation results are relegated to Section 4.10.

4.2 Partitioning-Based Bounds on Treatment Harm Rate

Suppose we are interested in the causal effect of a treatment of two levels, $a \in \{0, 1\}$, on individuals from an infinite superpopulation. We assume that the potential outcomes $Y(1)$ and $Y(0)$ take values in $\mathbb{Y} = \{-1, 1\}$, with $Y(a) = 1$ being favorable. Let $A \in \mathbb{A} = \{0, 1\}$ be the treatment indicator, $Y = Y(A) = AY(1) + (1 - A)Y(0)$ the observed outcome, $\boldsymbol{X}$ a vector of pre-treatment baseline covariates, and $\mathbb{X}$ the support of $\boldsymbol{X}$. We assume the unconfoundedness and the overlap assumptions for the TAM. We consider randomized designs in following sections for estimation and thus the unconfoundedness assumptions is only the focus of the current section on more general results. The parameter of interest is the THR, defined as $\theta = \Pr(Y(0) = 1, Y(1) = -1)$.

The parameter θ is generally non-identifiable even in RCTs unless $\text{Var}(Y \mid A = 1, \boldsymbol{X}) = 0$ or $\text{Var}(Y \mid A = 0, \boldsymbol{X}) = 0$ almost surely (Kallus, 2022). However, the θ can be partially

identified under the above assumptions with the sharp bounds given by $\theta \in [L(\theta), U(\theta)]$ where $L(\theta) = \mathrm{E}[\max\{0, \mu_0(\boldsymbol{X}) - \mu_1(\boldsymbol{X})\}]$, $U(\theta) = \mathrm{E}[\min\{\mu_0(\boldsymbol{X}), 1 - \mu_1(\boldsymbol{X})\}]$, $\mu_a(\boldsymbol{X}) = \Pr(Y(a) = 1|\boldsymbol{X})$ for $a = 0, 1$ (Zhang et al., 2013; Kallus, 2022; Wu et al., 2024). The sharp bounds are not easy to estimate because they involve the nondifferential functionals.

We consider a class of partitioning-based bounds (PBBs) for θ. Given an event $\mathbb{X}_j$, we define $\mu_a(\mathbb{X}_j) = \Pr(Y(a) = 1 \mid \boldsymbol{X} \in \mathbb{X}_j)$ for $a = 0, 1$. Consider a partition of the support of $\boldsymbol{X}$ as $\Delta = \{\mathbb{X}_1, \cdots, \mathbb{X}_J\}$ so that $\mathbb{X} = \bigcup_{j=1}^{J} \mathbb{X}_j$ and $\mathbb{X}_j$ are disjoint of each other (we assume $\mathbb{X}_j$ are all measurable). Based on the partition Δ, we can decompose θ into $\theta = \sum_{j=1}^{J} \rho_j \theta_j$ where $\theta_j = \Pr(Y(0) = 1, Y(1) = -1 \mid \boldsymbol{X} \in \mathbb{X}_j)$ and $\rho_j = \Pr(\boldsymbol{X} \in \mathbb{X}_j)$. Each θ_j is bounded by $\max\{0, \mu_0(\mathbb{X}_j) - \mu_1(\mathbb{X}_j)\} \leq \theta_j \leq \min\{\mu_0(\mathbb{X}_j), 1 - \mu_1(\mathbb{X}_j)\}$ due to the Fréchet-Hoeffding inequality (Fréchet, 1935). Then, the partitioning-based bounds of θ under the partition Δ are given by $B_\theta(\Delta) = [L(\theta, \Delta), U(\theta, \Delta)]$ where $L(\theta, \Delta) = \sum_{j=1}^{J} \rho_j \max\{0, \mu_0(\mathbb{X}_j) - \mu_1(\mathbb{X}_j)\}$ and $U(\theta, \Delta) = \sum_{j=1}^{J} \rho_j \min\{\mu_0(\mathbb{X}_j), 1 - \mu_1(\mathbb{X}_j)\}$.

Two trivial partitions are $\Delta_0 = \{\mathbb{X}\}$ and $\Delta_\infty = \bigcup_{\boldsymbol{x} \in \mathbb{X}} \{\{\boldsymbol{x}\}\}$, the latter of which is a collection of all singleton sets. Let us define $\sum_{x \in \mathbb{X}} \Pr(\boldsymbol{X} = \boldsymbol{x}) h(\boldsymbol{x}) = \int_{\mathbb{X}} h(\boldsymbol{x}) f_{\boldsymbol{X}}(\boldsymbol{x}) \nu(d\boldsymbol{x})$ for any function $h : \mathbb{X} \to \mathbb{R}$ where $f_{\boldsymbol{X}}(\boldsymbol{x})$ is the density of $\boldsymbol{X}$ with respect to the Lebesgue measure ν. Then, $B_\theta(\Delta_0)$ are the sharp bounds of θ in RCTs without covariates, initially derived by Gadbury et al. (2004), and $B_\theta(\Delta_\infty) = [L(\theta), U(\theta)]$ refer to the sharp bounds of θ in observational studies in the presence of covariates under the unconfoundedness and overlap assumptions. Nonetheless, it may not be surprising that, $B_\theta(\Delta_\infty)$ can be attained given a partition of $\mathbb{X}$ with four cells. Furthermore, to attain either the sharp upper bound or the sharp lower bound of θ, a partition with two cells is sufficient.

Proposition 4.2.1 *Let $\mathbb{L} = \{0, 1\} \times \{-1, 1\}$ denote the value space of (A, Y). The partition $\Delta_{full} = \{\mathbb{X}_{a,t}, (a, t) \in \mathbb{L}\}$ yields the sharp bounds of θ, i.e., $B_\theta(\Delta_{full}) = B_\theta(\Delta_\infty)$, where*

$\mathbb{X}_{a,t} = \{\boldsymbol{x} \in \mathbb{X} : (1-t)/2 + t\mu_a(\boldsymbol{x}) \geq \mu_{max}(\boldsymbol{x})\}$, $\mu_{max}(\boldsymbol{x}) = \max_{(a,t)\in\mathbb{L}}\{(1-t)/2 + t\mu_a(\boldsymbol{x})\}$. *Ties can be broken randomly. Because* $\mathbb{X}_{a,t}$ *could be an empty set, the size of* Δ_{full} *ranges from 1 to 4. Note that* $(1-t)/2 + t\mu_a(\boldsymbol{x}) = \Pr(Y(a) = t \mid \boldsymbol{X} = \boldsymbol{x})$.

Proposition 4.2.2 *Let* $\mathbb{X}_{L,t} = \{\boldsymbol{x} \in \mathbb{X} : t\mu_0(\boldsymbol{x}) \leq t\mu_1(\boldsymbol{x})\}$ *and* $\mathbb{X}_{U,t} = \{\boldsymbol{x} \in \mathbb{X} : t\mu_0(\boldsymbol{x}) + t\mu_1(\boldsymbol{x}) \leq t\}$ *for* $t \in \{-1, 1\}$. *Consider the partitions* $\Delta_{L,full} = \{\mathbb{X}_{L,1}, \mathbb{X}_{L,-1}\}$ *and* $\Delta_{U,full} = \{\mathbb{X}_{U,1}, \mathbb{X}_{U,-1}\}$. *Then, we have* $L(\theta, \Delta_{L,full}) = L(\theta)$ *and* $U(\theta, \Delta_{U,full}) = U(\theta)$.

Let $g^*(\boldsymbol{x}) = \arg\max_{(a,t)\in\mathbb{L}}\{(1-t)/2 + t\mu_a(\boldsymbol{x})\}$ be the optimal partitioning rule which yields Δ_{full}. Then, the $g^*(\boldsymbol{x})$ is essentially a Bayes classifier for a weighted multiclass classification problem.

Proposition 4.2.3 *Let us take the partitioning as a multiclass classification problem where the label for each observation is* $(A, Y) \in \{(1,1), (1,-1), (0,1), (0,-1)\}$, *or equivalently,* $(Y(1), Y(0)) \in \{(1,?), (-1,?), (?,1), (?,-1)\}$. *Consider the classifier* $g : \mathbb{X} \to \mathbb{L}$. *Let* $m_a(g) = \Pr(g(\boldsymbol{X}) \neq (a, Y(a)))$ *be the misclassification rate for individuals whose first label is equal to* a *and* $\pi_a = \Pr(A = a)$. *Then,* $g^*(\boldsymbol{x})$ *minimizes* $m_1(g) + m_0(g)$, *while the ordinary multiclass Bayes classifier,* $g^B(\boldsymbol{x}) = \arg\max_{(a,t)\in\mathbb{L}} \Pr(A = a, Y = t \mid \boldsymbol{X} = \boldsymbol{x})$, *minimizes* $\Pr(g(\boldsymbol{X}) \neq (A, Y)) = \pi_1 m_1(g) + \pi_0 m_0(g)$.

The PBBs of θ possess a very attractive property that they can always be consistently estimated under the randomization of treatment assignments as long as the size of the partition does not grow too large, as will be discussed in the next section. Whether the bounds are informative relies on the provided partition. Proposition 4.2.1 and 4.2.3 provide a guideline on how to find the optimal partition that yields the sharp bounds. The second key observation in this chapter is that finer partitioning cannot hurt the sharpness of the bounds.

Proposition 4.2.4 *A partition Δ_1 is said to be finer than a partition Δ_2 if and only if for every cell $\omega_2 \in \Delta_2$, there exists a cell $\omega_1 \in \Delta_1$ such that $\omega_2 \subseteq \omega_1$. Then for two partitions Δ_1 and Δ_2, we have $B_\theta(\Delta_1) \subseteq B_\theta(\Delta_2)$ if Δ_1 is finer than Δ_2.*

Remark. Due to Proposition 4.2.2 and 4.2.4, the partition, denoted by $\Delta_{full,L} \times \Delta_{full,U} = \{\mathbb{X}_{L,s} \cap \mathbb{X}_{U,t}, s, t \in \{1, -1\}\}$, as pairwise intersections of $\Delta_{L,full}$ and $\Delta_{U,full}$ must yield sharp bounds of θ. Indeed, it can be shown that $\Delta_{full,L} \times \Delta_{full,U} = \Delta_{full}$.

Some bounds are informative and some are not. For example, the bounds $[0, 1]$ are totally non-informative for θ. We define the information (for θ) carried by Δ, or equivalently the information of $B_\theta(\Delta)$, as one minus the width of $B_\theta(\Delta)$, i.e.,

$$I_\theta(\Delta) = 1 + L(\theta, \Delta) - U(\theta, \Delta) = \sum_{j=1}^{J} \rho_j \max\{\mu_1(\mathbb{X}_j), 1 - \mu_1(\mathbb{X}_j), \mu_0(\mathbb{X}_j), 1 - \mu_0(\mathbb{X}_j)\}.$$

The sharper the $B_\theta(\Delta)$ is, the more information the partition Δ carries. Because $B_\theta(\Delta_\infty)$ is sharp in the presence of covariates, the information upper bound for θ is $I_\theta(\Delta_\infty) = I_\theta(\Delta_{full}) = \mathrm{E}[\max\{\mu_0(\boldsymbol{X}), 1 - \mu_0(\boldsymbol{X}), \mu_1(\boldsymbol{X}), 1 - \mu_1(\boldsymbol{X})\}]$. Proposition 4.2.4 implies that we will not lose information by finer partitioning, which is an important property for the technique we will discuss in Section 4.4.3.

4.3 Non-Parametric Estimation

Consider a sample of individuals indexed in $\mathcal{S} = \{1, \cdots, n\}$. Let $\{(Y_i(1), Y_i(0), \boldsymbol{X}_i), i \in \mathcal{S}\}$ be the attributes of the sample units as n i.i.d. copies of $(Y(1), Y(0), \boldsymbol{X})$. Let A_i be the treatment assigned to the i-th unit. We assume that the observed data are RCT data, identified by the following two assumptions:

(1) $\{A_1, \cdots, A_n\} \perp\!\!\!\perp \{(Y_i(1), Y_i(0), \boldsymbol{X}_i)\}_{i=1}^n$.

(2) $\mathrm{E}(A_i) = \pi_1 \in (0,1)$ for $i = 1, \cdots, n$.

Let $Y_i = Y_i(A_i)$ denote the observed outcome and $\mathbf{S} = \{(Y_i, A_i, \boldsymbol{X}_i), i \in \mathcal{S}\}$ the observed sample data. The goal of this section is to estimate $B_\theta(\Delta)$ provided a partition Δ and the sample data $\mathbf{S}$ under simple randomization. We propose to estimate $\mu_a(\mathbb{X}_j)$ nonparametrically by

$$\hat{\mu}_a(\mathbb{X}_j) = \frac{\#\{i \in \mathcal{S} : Y_i = 1, A_i = a, \boldsymbol{X}_i \in \mathbb{X}_j\}}{\#\{i \in \mathcal{S} : A_i = a, \boldsymbol{X}_i \in \mathbb{X}_j\}}.$$

Because of randomization, $\hat{\mu}_a(\mathbb{X}_j)$ is an unbiased and consistent estimator of $\mu_a(\mathbb{X}_j)$. Let $n_j = \#\{i \in \mathcal{S} : \boldsymbol{X}_i \in \mathbb{X}_j\}$ be the size of the sample units in $\mathbb{X}_j$. Then, the $L(\theta, \Delta)$ can be estimated as $\widehat{L}(\theta, \Delta) = \sum_{j=1}^J n_j/n \max\{0, \hat{\mu}_0(\mathbb{X}_j) - \hat{\mu}_1(\mathbb{X}_j)\}$, and $U(\theta, \Delta)$ can be estimated as $\widehat{U}(\theta, \Delta) = \sum_{j=1}^J n_j/n \min\{\hat{\mu}_0(\mathbb{X}_j), 1 - \hat{\mu}_1(\mathbb{X}_j)\}$. The partitioning-based interval estimator of θ (under the partition Δ) is given by $\hat{B}_\theta(\Delta) = [\widehat{L}(\theta, \Delta), \widehat{U}(\theta, \Delta)]$.

Proposition 4.3.1 *Let* $\eta_n = \sum_{j=1}^J \sum_{a=0}^1 \sigma_{Y(a),j} \sqrt{\rho_j / n\pi_a} + 1/\sqrt{n}$ *where* $\sigma^2_{Y(a),j} = \mathrm{Var}(Y(a) \mid \boldsymbol{X} \in \mathbb{X}_j)$ *and* $\pi_a = \Pr(A = a)$. *Denote* $\hat{I}_\theta(\Delta) = 1 + \widehat{L}(\theta, \Delta) - \widehat{U}(\theta, \Delta)$. *Provided RCT data and a partition* Δ, *we have* $\widehat{L}(\theta, \Delta) = L(\theta, \Delta) + O_p(\eta_n)$, $\widehat{U}(\theta, \Delta) = U(\theta, \Delta) + O_p(\eta_n)$, *and* $\mathrm{E}[\hat{I}_\theta(\Delta)] \geq I_\theta(\Delta)$.

Justification of the bounds estimators is solely randomization-based. It does not require any assumptions on the partition Δ except finite J as long as Δ is fixed. Typically, we do not hope J gets too large because the bias term η_n converges more slowly as J increases. Proposition 4.3.1 also indicates that the non-parametric estimation procedure will not lead to a loss of information carried by Δ.

4.4 Space Partitioning

In this section, we present a strategy for partitioning. To validate the non-parametric estimators of the bounds proposed in Section 4.3, the partition should be fixed regarding the sample data $\mathbf{S}$. Therefore, we assume there is an external dataset $\mathbf{G} = \{(Y_i, A_i, \boldsymbol{X}_i), i \in \mathcal{G}\}$, independent of $\mathbf{S}$, so that any partitions created using $\mathbf{G}$ can be viewed as fixed conditional on $\mathbf{G}$. The data $\mathbf{G}$ can be from observational studies where treatment assignments are non-randomized. Later in this section, we will show how to create pseudo external datasets by cross-fitting using the RCT data at hand when the reference data are not available from external sources.

4.4.1 Surrogate Partition

Let $N = |\mathbf{G}|$ be the size of the external data $\mathbf{G}$ and $\hat{\mu}_{a,N}(\boldsymbol{x})$ be an estimator of $\mu_a(\boldsymbol{x})$ trained on $\mathbf{G}$ for $a = 0, 1$. The $\hat{\mu}_{a,N}(\boldsymbol{x})$ should be independent of the sample $\mathbf{S}$. Motivated by Proposition 4.2.1, we create a surrogate partition as $\Delta_{plug} = \Delta_{plug}(\hat{\mu}_{0,N}(\boldsymbol{x}), \hat{\mu}_{1,N}(\boldsymbol{x})) = \{\hat{\mathbb{X}}_{a,t}, (a,t) \in \mathbb{L}\}$, where $\hat{\mathbb{X}}_{a,t} = \{\boldsymbol{x} \in \mathbb{X} : (1-t)/2 + t\hat{\mu}_{a,N}(\boldsymbol{x}) \geq \hat{\mu}_{max,N}(\boldsymbol{x})\}$ as a guess of $\mathbb{X}_{a,t}$ by plugging in the estimators $\hat{\mu}_{1,N}(\boldsymbol{x})$, $\hat{\mu}_{0,N}(\boldsymbol{x})$, and $\hat{\mu}_{max,N}(\boldsymbol{x}) = \max_{(a,t)\in\mathbb{L}}\{(1-t)/2 + t\hat{\mu}_{a,N}(\boldsymbol{x})\}$. Ties will be broken randomly. Here, the functions $\mu_a(\boldsymbol{x})$ are referred to as nuisance parameters. They are not the parameters of primary interest. However, the information carried by Δ_{plug} does rely on whether we can accurately estimate these nuisance parameters.

Proposition 4.4.1 *The information carried by Δ_{plug} satisfies*

$$I_\theta(\Delta_{plug}) \geq I_\theta(\Delta_\infty) - \mathrm{E}\left[\mu_{max}(\boldsymbol{X}) - \hat{\mu}_{max,N}(\boldsymbol{X}) + \max_{a=0,1}|\hat{\mu}_{a,N}(\boldsymbol{X}) - \mu_a(\boldsymbol{X})| \mid \mathbf{G}\right],$$

where $\mathrm{E}[\cdot \mid \mathbf{G}]$ *denotes taking expectation conditional on the external data* $\mathbf{G}$.

Corollary 4.4.2 *If* $\mathrm{E}[|\hat{\mu}_{a,N}(\boldsymbol{X}) - \mu_a(\boldsymbol{X})| \mid \mathbf{G}] = o_p(1)$ *for* $a = 0, 1$, *then there must be* $I_\theta(\Delta_{plug}) \geq I_\theta(\Delta_\infty) - o_p(1)$. *The proof can be easily obtained by noting that* $|\mu_{max}(\boldsymbol{X}) - \hat{\mu}_{max,N}(\boldsymbol{X})| \leq |\hat{\mu}_{1,N}(\boldsymbol{X}) - \mu_1(\boldsymbol{X})| + |\hat{\mu}_{0,N}(\boldsymbol{X}) - \mu_0(\boldsymbol{X})|$.

The information upper bound is asymptotically reached if the estimator $\hat{\mu}_{a,N}$ is consistent in the sense that $\mathrm{E}[|\hat{\mu}_{a,N}(\boldsymbol{X}) - \mu_a(\boldsymbol{X})| \mid \mathbf{G}] = o_p(1)$ for $a = 0, 1$. We consider popular machine learning classification algorithms for the estimation of the nuisance parameters in the next section.

Remark. An alternative way to partitioning is to solve the weighted multiclass classification problem as formulated in Proposition 4.2.3, which can be realized by minimizing a surrogate loss function of the weighted misclassification risk. The topic is for future work and will not be discussed in the thesis.

4.4.2 Probabilistic Classifiers

In machine learning, a classifier is a function that takes the covariates as input and the predicted label of the categorical outcome as output. A probabilistic classifier is a classifier that also outputs the estimated probability of the outcome belonging to a certain class given the covariates. Many machine learning models are probabilistic, and their predicted labels of the outcomes are essentially produced by applying a threshold to the estimated probabilities. Some models are margin-based and not inherently probabilistic such as support vector machines. With certain techniques such as Platt scaling, however, the margins to the decision boundaries can be transformed into probabilistic estimates (Platt et al., 1999). In Section 4.10.3, we give a brief review of six classification algorithms on binary

outcomes — logistic regression, Gaussian naive Bayes, K-nearest-neighbors, support vector machines, random forests, and gradient boosting — and how they do probabilistic estimation. These algorithms will be employed in our numerical experiments for the estimation of nuisance parameters. For more details of these algorithms, we refer readers to Hastie et al. (2009). While highly flexible and capable of being fast implemented in software such as Python, these algorithms have a potential drawback: some are not specifically designed for probabilistic estimation and could produce inaccurate probabilistic estimates. These probabilistic estimates, however, can be more reliable after model calibration; see Section 4.10.2 for further details.

4.4.3 Cross-Fitting

Cross-fitting is one of the basic tools in machine learning for model evaluation and parameter selection. It has also received more and more attention in the field of statistics as a standard approach to creating valid confidence intervals. See, e.g., Wager et al. (2016); Athey and Imbens (2016); Chernozhukov et al. (2018). In this chapter, cross-fitting will serve to create pseudo external datasets when no external data sources are available.

In K-fold cross-fitting, the sample data are randomly split into K subsets with equal sample size. In each loop, $(K-1)$ out of the K data subsets are used as a pseudo external data for the estimation of the nuisance parameters μ_a, and the remaining one is used for the interval estimation of θ based on the nuisance estimators. This fitting-estimation procedure repeats K times, going through all leave-one-out subsets, and the final interval estimator of θ is the average of the K interval estimators. The fitting-estimation procedure creates independency between the data for model training and the data for interval estimation, and thus justifies the non-parametric bounds estimators provided in Section 4.3.

Cross-fitting itself can be taken as an action of partitioning. However, because the data splitting procedure is completely random, every split dataset is representative of the population. Therefore, the partitioning-based bounds under cross-fitting should not differ from the ordinary partitioning-based bounds. Even if certain covariate distribution shift occurs after data splitting due to the randomness involved in the splitting procedure, there is no need for concern as finer partitioning does not cause loss of information, as stated in Proposition 4.2.4.

Although cross-fitting does not result in loss of information, it does affect the variances and convergence rates of the nonparametric bounds estimators provided in Section 4.3. The number of folds, K, essentially yields a trade-off between the information (sharpness) of the bounds and the precisions of the bounds estimators. With more folds, on the one hand, we have more data in each loop for model training which leads to better estimators of the nuisance parameters and hence a more informative partition. On the other hand, the variances of the bounds estimators may get bigger as K increases because of less data for bounds estimation in each loop. The K can be determined according to the sample size. We suggest setting $K = 2$ or $K = 5$ in practice, and adopt $K = 2$ in the numerical studies.

4.5 Assessment of Estimated Bounds

One of the major issues for practitioners, after obtaining an interval estimator of θ, is to assess the precision of the interval estimator. Let us first focus on the case that external data are available and thus the partition Δ can be viewed as fixed (conditional on the external data). In this case, our estimated bounds $[\widehat{L}(\theta, \Delta), \widehat{U}(\theta, \Delta)]$ converge to $[L(\theta, \Delta), U(\theta, \Delta)]$ under simple randomization. The information on the precisions of the estimated bounds

is included in the distribution of $\widehat{L}(\theta,\Delta) - L(\theta,\Delta)$ and $\widehat{U}(\theta,\Delta) - U(\theta,\Delta)$, of which the quantile estimators can be used to construct an extended confidence interval of θ.

We first give a discussion on $\widehat{L}(\theta,\Delta) - L(\theta,\Delta)$. Provided a partition $\Delta = \{\mathbb{X}_1, \cdots, \mathbb{X}_J\}$, recall that

$$\widehat{L}(\theta,\Delta) = \sum_{j=1}^{J} \frac{n_j}{n} \max\{0, \hat{\mu}_0(\mathbb{X}_j) - \hat{\mu}_1(\mathbb{X}_j)\}.$$

Let us proceed conditional on $\boldsymbol{n} = (n_1, \cdots, n_J)$. Then, the $\hat{\mu}_a(\mathbb{X}_j) \mid n_j$ is asymptotically normally distributed as $n_j \to \infty$ with the mean $\tau_{a,j} = \mu_a(\mathbb{X}_j)$ and the variance $\sigma^2_{Y(a),j}/(n_j\pi_a)$ for $a = 0, 1$ under certain regularity conditions, and the variance of $\hat{\mu}_0(\mathbb{X}_j) - \hat{\mu}_1(\mathbb{X}_j)$ is given by $\sigma^2_{Y(0),j}/(n_j\pi_0) + \sigma^2_{Y(1),j}/(n_j\pi_1)$ (Li and Ding, 2017; Ye et al., 2023). Note that the mean $\tau_{a,j}$ does not rely on n_j. To be clear, we denote by $\boldsymbol{n}^* = (n_1^*, \cdots, n_J^*)$ a random vector so that $\sum_{j=1}^J n_j^* = n$ and $\boldsymbol{n}$ is a realization of $\boldsymbol{n}^*$. Then, the distribution of $\widehat{L}(\theta,\Delta) - L(\theta,\Delta)$ is approximately $\sum_{j=1}^J n_j^*/n \max\{0, N(\tau_{0,j} - \tau_{1,j}, \sigma^2_{1,j} + \sigma^2_{0,j})\} - L(\theta,\Delta)$ where $\sigma^2_{a,j} = \sigma^2_{Y(a),j}/(n_j^*\pi_a)$ for $a = 0, 1$. Conditional on $\boldsymbol{n}^*$, this distribution can be estimated by

$$\sum_{j=1}^{J} \frac{n_j^*}{n} \max\{0, N(\hat{\mu}_0(\mathbb{X}_j) - \hat{\mu}_1(\mathbb{X}_j), \hat{\sigma}^2_{1,j} + \hat{\sigma}^2_{0,j})\} - \widehat{L}(\theta,\Delta) \mid \mathbf{S}, \boldsymbol{n}^* \tag{4.1}$$

where

$$\hat{\sigma}^2_{a,j} = \frac{\sum_{i:\boldsymbol{X}_i \in \mathbb{X}_j} 1\{A_i = a\}(Y_i - \bar{Y}_{a,j})^2}{n_j^* n_{a,j}/n_j (n_j^* n_{a,j}/n_j - 1)},$$

$\bar{Y}_{a,j} = n_{a,j}^{-1} \sum_{\boldsymbol{X}_i \in \mathbb{X}_j} 1\{A_i = a\} Y_i$, $n_{a,j} = \#\{i : A_i = a, \boldsymbol{X}_i \in \mathbb{X}_j\}$ for $a = 0, 1$. Since Y_i is binary, the sampling variance has a simple form: $\hat{\sigma}^2_{a,j} = n_{a,j}\bar{Y}_{a,j}(1-\bar{Y}_{a,j})/\{n_j^* n_{a,j}/n_j (n_j^* n_{a,j}/n_j - 1)\}$. The distribution (4.1) can be simulated by Monte Carlo. Similarly, the distribution

of $\widehat{U}(\theta,\Delta) - U(\theta,\Delta) \mid \boldsymbol{n}$ can be approximated by

$$\sum_{j=1}^{J} \frac{n_j^*}{n} \min\{N(\hat{\mu}_0(\mathbb{X}_j), \hat{\sigma}_{0,j}^2), 1 - N(\hat{\mu}_1(\mathbb{X}_j), \hat{\sigma}_{1,j}^2)\} - \widehat{U}(\theta,\Delta) \mid \mathbf{S}, \boldsymbol{n}^*. \tag{4.2}$$

The distribution of $\boldsymbol{n}^*$ can be approximately by a multivariate hypergeometric distribution with sufficiently large M (so that $n_j M/n$ are integers) as

$$\Pr(n_1^* = m_1, \cdots, n_J^* = m_J) = \begin{cases} \frac{\prod_{j=1}^{J} \binom{n_j M/n}{m_j}}{\binom{M}{n}} & \text{if } \sum_{j=1}^{J} m_j = n, \\ 0 & o.w. \end{cases}.$$

Based on the above results, we can numerically approximate the distribution

$$\sum_{j=1}^{J} \frac{n_j^*}{n} \max\{0, N(\hat{\mu}_0(\mathbb{X}_j) - \hat{\mu}_1(\mathbb{X}_j), \hat{\sigma}_{1,j}^2 + \hat{\sigma}_{0,j}^2)\} \mid \mathbf{S},$$

which is an approximation of the distribution of $\widehat{L}(\theta,\Delta)$. Let $\hat{\xi}_{\alpha,L}$ be its α-upper-quantile. Together with the constraint $L(\theta,\Delta) \in [0,1]$, the proposed two-sided $(1-\alpha)$ confidence interval for $L(\theta,\Delta)$ is

$$C_{\alpha,L}(\Delta) = [\max\{2\widehat{L}(\theta,\Delta) - \hat{\xi}_{\alpha/2,L}, 0\}, \min\{2\widehat{L}(\theta,\Delta) - \hat{\xi}_{1-\alpha/2,L}, 1\}].$$

Similarly, we can get a $(1-\alpha)$ confidence interval for $U(\theta,\Delta)$ as

$$C_{\alpha,U}(\Delta) = [\max\{2\widehat{U}(\theta,\Delta) - \hat{\xi}_{\alpha/2,U}, 0\}, \min\{2\widehat{U}(\theta,\Delta) - \hat{\xi}_{1-\alpha/2,U}, 1\}]$$

where $\hat{\xi}_{\alpha,U}$ is the estimated α-upper-quantile of the distribution

$$\sum_{j=1}^{J} \frac{n_j^*}{n} \min\{N(\hat{\mu}_0(\mathbb{X}_j), \hat{\sigma}_{0,j}^2), 1 - N(\hat{\mu}_1(\mathbb{X}_j), \hat{\sigma}_{1,j}^2)\} \mid \mathbf{S}.$$

However, since the proposed confidence intervals are constructed based on the assumption that $\hat{\mu}_a(\mathbb{X}_j) \mid n_j$ is approximately normal, they are valid only when $n_j, j = 1, \cdots, J$ are large enough, and may fail if n is small and $\rho_j = \Pr(\boldsymbol{X} \in \mathbb{X}_j) \approx 0$ for some j. After Bonferroni correction, an extended $(1-\alpha)$ confidence interval for $[L(\theta, \Delta), U(\theta, \Delta)]$ is given as

$$C_\alpha(\theta, \Delta) = [\max\{2\widehat{L}(\theta, \Delta) - \hat{\xi}_{\alpha/2,L}, 0\}, \min\{2\widehat{U}(\theta, \Delta) - \hat{\xi}_{1-\alpha/2,U}, 1\}],$$

which can be used as a $(1-\alpha)$ confidence interval for θ. This extended confidence interval is, of course, too conservative for θ in general cases. It might over-cover θ even in the extreme case that $L(\theta, \Delta) = L(\theta) = \theta$ or $U(\theta, \Delta) = U(\theta) = \theta$ because of the Bonferroni correction. In the context of cross-fitting, the confidence intervals obtained from the split samples are aggregated by taking their average.

4.6 Numerical Experiments

In this section, we evaluate the proposed partitioning-based interval estimators through Monte Carlo numerical experiments. We consider two partitions, Δ_0 and Δ_{full}, with the resulting interval estimators of θ respectively referred to as "Naive" and "Oracle". Also, we consider surrogate partitions induced by six probabilistic classification algorithms: Logistic regression (Logit), Naive Bayes (NBayes), Support Vector Machines (SVM, Cortes and

Vapnik (1995)) with Platt scaling (Platt et al., 1999), K-Nearest Neighbors (KNN), Random Forests (RF, Breiman (2001)), and eXtreme Gradient Boosting (XGBoost, Chen and Guestrin (2016)). See Section 4.10.3 for a brief description and Hastie et al. (2009) for further details of these classification algorithms. We adopt the Python package `scikit-learn` in our coding to realize the classification algorithms, and the folds of cross-fitting is set to $K = 2$.

We can also estimate the sharp bounds of θ by directly plugging in the nuisance estimators in the bounds as $\widehat{L}(\theta) = \mathrm{E}[\max\{0, \hat{\mu}_{0,N}(\boldsymbol{X}) - \hat{\mu}_{1,N}(\boldsymbol{X})\} \mid \mathbf{G}]$ and $\widehat{U}(\theta) = \mathrm{E}[\min\{\hat{\mu}_{0,N}(\boldsymbol{X}), 1 - \hat{\mu}_{1,N}(\boldsymbol{X})\} \mid \mathbf{G}]$. Similarly, the cross-fitting strategy will be utilized in the estimation. We label the resulting estimators by "Plug-in" and include them in the simulation results for comparison. Furthermore, we calculate and present the cross-validation misclassification rates (MCR) for the classification algorithms applied to the Monte Carlo data.

The data generating procedure is as follows. The observed covariates are generated by $\boldsymbol{X} = (X_1, \cdots, X_{10})^\intercal \sim N(0, \boldsymbol{I}_{10})$. Let $\boldsymbol{Z} = (3X_1^2, X_2X_3, X_3, X_4, X_5)^\intercal$. The potential outcome is generated as $Y(a) = 1\{\boldsymbol{\beta}_a^\intercal \boldsymbol{Z} + \beta_{a,0} + \varepsilon\}$ where $\varepsilon \sim N(0, \sigma_\varepsilon^2)$ and $\beta_{a,0}$ is chosen so that $\Pr(Y(0) = 1) = 0.2$ and $\Pr(Y(1) = 1) = 0.4$. The σ_ε controls the magnitude of the noise factor and thus the information upper bound of the partitions. The treatment assignments are generated based on the completely randomized design under the treated ratio equal to $1/2$. We consider two scenarios for $\boldsymbol{\beta}_a$:

- **Scenario 1** (Homogeneous treatment effect). $\boldsymbol{\beta}_a = (1, 1, 1, 1, 1)^\intercal$ for $a = 0, 1$.
- **Scenario 2** (Heterogeneous treatment effect). $\boldsymbol{\beta}_0 = (1, 1, 1, 1, 1)^\intercal$ and

$$\boldsymbol{\beta}_1 = (-1.2, 1, -0.8, 0.5, -0.3)^\intercal.$$

For each setup, we generate $M = 1000$ Monte Carlo datasets. Let $\widehat{B}(\theta) = [\widehat{L}(\theta), \widehat{U}(\theta)]$ denote an interval estimator of θ and $\widehat{B}_m(\theta) = [\widehat{L}_m(\theta), \widehat{U}_m(\theta)]$ an interval estimate based on the m-th Monte Carlo dataset. We adopt the following three metrics to evaluate $\widehat{B}(\theta)$ based on the M Monte Carlo replications:

- **Bias**: $\text{Bias}(\widehat{B}(\theta)) = M^{-1}\sum_{m=1}^{M} \max\{\widehat{L}_m(\theta) - \theta, \theta - \widehat{U}_m(\theta), 0\}$.
- **Width**: $\text{Width}(\widehat{B}(\theta)) = M^{-1}\sum_{m=1}^{M}(\widehat{U}_m(\theta) - \widehat{L}_m(\theta))$.
- **Coverage rate (CovR)**: $\text{CovR}(\widehat{B}(\theta)) = M^{-1}\sum_{m=1}^{M} 1\{\widehat{L}_m(\theta) \leq \theta \leq \widehat{U}_m(\theta)\}$.

Let n denote the size of the sample data in each Monte Carlo replication. We consider combinations of the parameters $\sigma_\varepsilon \in \{1, 2\}$ and $n \in \{500, 2000\}$. Simulation results for KNN and RF with $n = 500$ are summarized in Table 4.1. Additional results for the six aforementioned classification algorithms, with model-calibration, and for $n = 2000$, are given in Section 4.10. The metrics TMCR and CMCR in the table denotes respectively the empirical cross-validation misclassification rate of the corresponding classification algorithm on the treated and the control group data. Results presented in Table 4.1 can be summarized as follows: (1) The proposed partitioning-based interval estimators have much smaller biases as compared to the plug-in based interval estimators; (2) under the RF algorithm, the partitioning-based interval estimators are more informative than the plug-in based estimators with smaller widths; under the KNN algorithm, the plug-in based estimators are narrower, but they are severely biased and fail to cover the true value. For illustrations of a more complete picture, we vary the value of σ_ε from 0 to 5 and conduct simulations for the interval estimators under the RF algorithm with $n = 2000$. The estimated upper and lower bounds and 75% confidence bands are depicted in Figure 4.1. We observe that the partitioning-based estimated bounds are much more informative than

Table 4.1: Monte Carlo simulation results for the interval estimators with $n = 500$. All values except the percentages in the table are presented in units scaled by 10^{-2}.

		Plug-in				Partitioning					
Method	θ	Estimate	Bias	Width	CovR	Estimate	Bias	Width	CovR	TMCR	CMCR
Scenario 1, $\sigma_\varepsilon = 1$											
Naive	0.1	$[0.0, 20.5]$	0.0	20.5	100.0%	$[0.0, 20.5]$	0.0	20.5	100.0%	NaN	NaN
KNN	0.1	$[2.5, 11.7]$	2.4	9.2	0.0%	$[0.1, 15.2]$	0.1	15.1	76.6%	31.3%	18.8%
RF	0.1	$[0.8, 15.2]$	0.7	14.4	0.5%	$[0.1, 5.3]$	0.1	5.2	76.5%	17.7%	10.6%
Oracle	0.1	NaN	NaN	NaN	NaN	$[0.0, 0.6]$	0.0	0.6	73.5%	NaN	NaN
Scenario 1, $\sigma_\varepsilon = 2$											
Naive	1.9	$[0.0, 22.7]$	0.0	22.7	100.0%	$[0.0, 22.7]$	0.0	22.7	100.0%	NaN	NaN
KNN	1.9	$[2.7, 13.8]$	0.9	11.1	21.2%	$[0.2, 17.7]$	0.0	17.5	99.2%	35.1%	21.1%
RF	1.9	$[0.9, 17.7]$	0.0	16.8	95.8%	$[0.2, 8.6]$	0.0	8.4	100.0%	23.8%	14.1%
Oracle	1.9	NaN	NaN	NaN	NaN	$[0.0, 4.3]$	0.0	4.3	98.5%	NaN	NaN
Scenario 2, $\sigma_\varepsilon = 1$											
Naive	19.7	$[0.0, 20.6]$	0.6	20.6	61.6%	$[0.0, 20.6]$	0.6	20.6	61.6%	NaN	NaN
KNN	19.7	$[3.0, 12.4]$	7.2	9.5	0.1%	$[7.3, 20.4]$	0.6	13.2	62.3%	35.9%	18.7%
RF	19.7	$[7.6, 20.5]$	0.5	12.9	64.2%	$[16.7, 20.4]$	0.7	3.7	49.5%	21.7%	10.6%
Oracle	19.7	NaN	NaN	NaN	NaN	$[19.2, 19.8]$	1.4	0.6	11.4%	NaN	NaN
Scenario 2, $\sigma_\varepsilon = 2$											
Naive	20.6	$[0.0, 22.6]$	0.3	22.6	75.9%	$[0.0, 22.6]$	0.3	22.6	75.9%	NaN	NaN
KNN	20.6	$[3.5, 14.6]$	6.1	11.1	1.3%	$[6.2, 22.4]$	0.3	16.2	74.7%	38.5%	21.0%
RF	20.6	$[7.4, 22.5]$	0.3	15.1	78.5%	$[15.6, 22.3]$	0.3	6.7	73.7%	28.6%	14.0%
Oracle	20.6	NaN	NaN	NaN	NaN	$[18.6, 21.8]$	0.6	3.3	51.9%	NaN	NaN

the plug-in based estimated bounds and close to the Oracle bounds. The results under model calibration are given in Figure 4.2, where the plug-in based estimated bounds perform much better in terms of bias because of the probabilities correction, despite they are still wider than the partitioning-based interval estimators. However, when σ_ε is small the plug-in based estimated lower bound is biased and greater than the true value of interest.

Additional tables for other simulation settings presented in Section 4.10.4 demonstrate similar patterns. A brief summary on extended confidence intervals: a significance level of $\alpha = 0.25$ leads to around 98% coverage rates of the extended confidence intervals on θ under all the classification algorithms in our setups. Additional figures on the results under the KNN algorithm and the tables on evaluation of the confidence intervals developed in

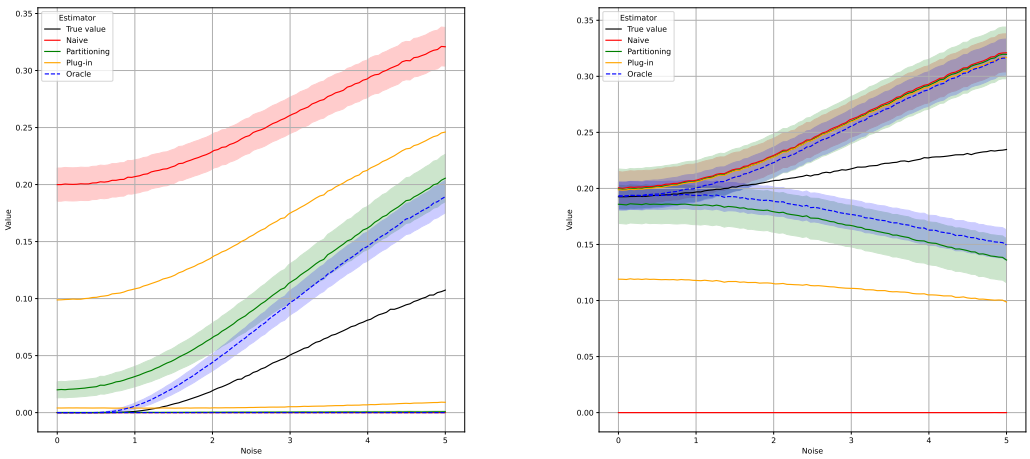


Figure 4.1: Naive and Oracle bounds, estimated plug-in and partitioning-based bounds using the RF algorithm, and their corresponding 75% confidence bands. X-axis denotes the value of σ_ε. The plot for Scenario 1 is presented in the left panel, and the plot for Scenario 2 is presented in the right panel.

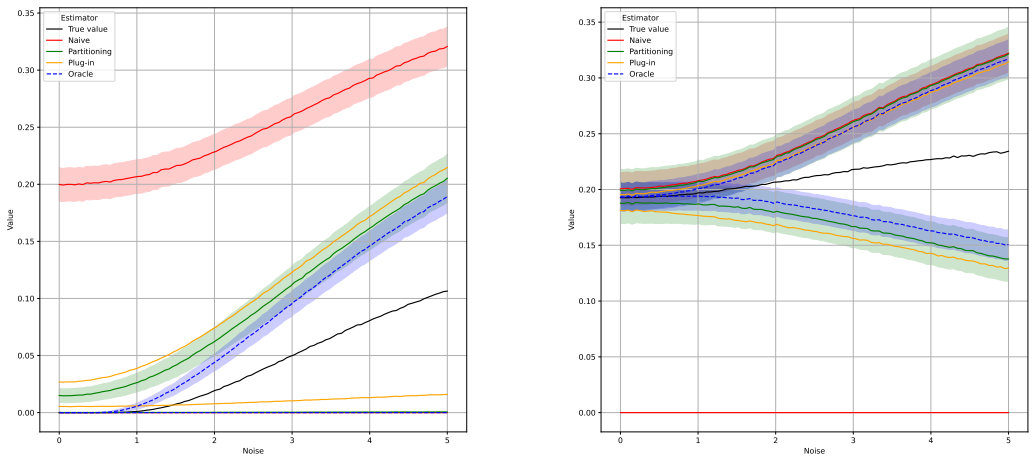


Figure 4.2: Naive and Oracle bounds, estimated plug-in and partitioning-based bounds using the RF algorithm combined with model calibration, and their corresponding 75% confidence bands. X-axis denotes the value of σ_ε. The plot for Scenario 1 is presented in the left panel, and the plot for Scenario 2 is presented in the right panel.

Section 4.5 are also provided in Section 4.10.4. In summary, the empirical coverage rates of the proposed confidence intervals are very close to the nominal levels, except some extreme cases as explained in Section 4.10.

4.7 Analysis of the AIDS Clinical Trials Group Study 175 Data

AIDS Clinical Trials Group Study 175 (ACTG 175) is a RCT that aimed to compare the treatment effects of different therapies on patients infected with human immunodeficiency virus type 1 (HIV-1) whose CD4 cell counts were between 200 and 500 per cubic millimeter (Hammer et al., 1996). Four therapy treatments to the clinical subjects were considered and randomized in the study: monotherapy with zidovudine alone, monotherapy with didanosine alone, combination therapy with zidovudine and didanosine, and combination therapy with zidovudine and zalcitabine. In the public ACTG 175 dataset that is available online, the CD4 cell counts of all the patients were recorded at baseline prior to the treatments and 20+/-5 weeks after the treatments. The CD4 cell counts reveal the damage of HIV-1 infection to the immune system, as which progresses the CD4 cell counts decrease. We define a new binary outcome as the indicator of whether the CD4 cell counts had increased at 20+/-5 weeks since the baseline, denoted by Y, to assess the treatment effects of different therapies, with $Y = 1$ being favorable. We focus on the comparison between the combination therapy with zidovudine and zalcitabine, termed as the treated group ($A = 1$), and monotherapy with zidovudine alone, termed as the control group ($A = 0$). The average treatment effect is estimated as 0.119 with the 95% confidence interval as $[0.059, 0.179]$, implying the combination therapy is, on average, beneficial to the clinical

Table 4.2: Interval estimates of the treatment harm rate in ACTG 175 and 75% confidence intervals for the partitioning-based upper and lower bounds under two-fold cross-fitting.

Method	Estimate	Upper-CI	Lower-CI	Extended-CI
Naive	$[0, 0.436]$	$[0.424, 0.465]$	$[0, 0]$	$[0, 0.465]$
Logit	$[0.003, 0.34]$	$[0.309, 0.375]$	$[0, 0.006]$	$[0, 0.375]$
NBayes	$[0.002, 0.355]$	$[0.326, 0.391]$	$[0, 0.004]$	$[0, 0.391]$
SVM	$[0, 0.329]$	$[0.297, 0.364]$	$[0, 0]$	$[0, 0.364]$
KNN	$[0, 0.356]$	$[0.323, 0.391]$	$[0, 0]$	$[0, 0.391]$
RF	$[0.004, 0.318]$	$[0.285, 0.353]$	$[0, 0.009]$	$[0, 0.353]$
XGBoost	$[0.001, 0.358]$	$[0.329, 0.395]$	$[0, 0.001]$	$[0, 0.395]$

subjects during the first 20 weeks.

We are interested in the question "What is the potential proportion of patients getting physically worse after the combination therapy"? To answer this question, we apply our proposed partitioning-based interval estimation approach to estimate the treatment harm rate. The interval estimates and the 75% confidence intervals for the upper and lower bounds as well as the extended confidence intervals for the THR are given in Tables 4.2 and 4.3. The naive Fréchet-Hoeffding bounds are $[0, 0.436]$, implying no more than 43.6% patients might be negatively affected by the combination therapy. The partitioning-based interval estimators are more precise after covariate adjustment and substantially sharpen the upper bound. For example, the partitioning-based interval estimate of the THR under the RF algorithm and two-folds cross-fitting is $[0.004, 0.318]$, indicating that a potential of 0.4% treatment harm rate but no greater than 31.8%. If we take the possible variability into consideration, a 75% extended confidence interval of the THR under the RF is given as $[0, 0.353]$, indicating that there is no evidence that the combination therapy does harm to the patients.

Table 4.3: Interval estimates of the treatment harm rate in ACTG 175 and 75% confidence intervals for the partitioning-based upper and lower bounds under five-fold cross-fitting.

Method	Estimate	Upper-CI	Lower-CI	Extended-CI
Naive	$[0, 0.436]$	$[0.424, 0.465]$	$[0, 0]$	$[0, 0.465]$
Logit	$[0.008, 0.333]$	$[0.296, 0.397]$	$[0, 0.015]$	$[0, 0.397]$
NBayes	$[0.006, 0.348]$	$[0.309, 0.419]$	$[0, 0.009]$	$[0, 0.419]$
SVM	$[0.015, 0.317]$	$[0.271, 0.375]$	$[0, 0.03]$	$[0, 0.375]$
KNN	$[0.011, 0.346]$	$[0.308, 0.409]$	$[0, 0.018]$	$[0, 0.409]$
RF	$[0.011, 0.344]$	$[0.304, 0.406]$	$[0, 0.019]$	$[0, 0.406]$
XGBoost	$[0.016, 0.344]$	$[0.303, 0.408]$	$[0, 0.027]$	$[0, 0.408]$

4.8 Concluding Remarks

The focus of this chapter is on how to construct a robust and informative interval estimator for the THR with RCT data. We developed a class of partitioning-based interval estimators motivated from a class of partitioning-based bounds of the THR. Our proposed methodologies do not impose any untestable identification or ambiguous model assumptions for the estimation procedures other than the randomization which is part of the design. Our interval estimators not only achieve information gains from the baseline auxiliary covariates but also remain valid under model-misspecification, which is a highly desirable model-assisted property for applications such as clinical trials.

The idea of partitioning can be viewed as a generalization of stratification, where some covariates are discrete or categorical, and the covariates space can be naturally partitioned based on the levels of these covariates. In the context of estimating the average treatment effect, stratification serves as a covariate adjustment method to improve the estimator's efficiency. In the context of estimating the THR, the idea of using stratification to sharpen the bounds has been previously mentioned by, for example, Zhang et al. (2013) and Huang et al. (2017). However, there are no guarantees in adequate information gains after strati-

fication relying on the given levels of the categorical covariates. If there are a large number of levels within the covariates, the size of the strata may become excessively large. The proposed partitioning-based method provides a systematic approach to achieving an informative partition with a maximum of four cells and is applicable to data containing any types of covariates. The modern machine learning methods can be easily incorporated into our framework to further improve the interval estimators as discussed in this chapter.

One possible limitation of our proposed methods is that they are specifically designed for binary outcomes. They cannot be directly applied to continuous or ordinal outcomes. Also, we only consider simple randomization designs such as completely randomized trials and Bernoulli trials. It might be of interest to practitioners to consider blocking randomization and covariate-adaptive randomization which are often used in practice. These extensions will be pursued in future research.

4.9 Discussion and An Alternative Formulation

It turns out that there is an alternative formulation of the PBBs which can be motivated from the estimation procedure considered in Kallus (2022). The connection was not realized until after the main body of the research presented in this chapter had been completed. Without loss of generality, we focus on the lower bounds of θ and show this connection below.

Given a partition $\Delta = \{\mathbb{X}_1, \cdots, \mathbb{X}_J\}$, the partitioning-based lower bound of θ has the

expression:

$$
\begin{aligned}
L(\theta, \Delta) &= \sum_{j=1}^{J} \Pr(\boldsymbol{X} \in \mathbb{X}_j) \max\{0, \mu_0(\mathbb{X}_j) - \mu_1(\mathbb{X}_j)\} \\
&= \sum_{j=1}^{J} \max\{0, \Pr(Y(0) = 1, \boldsymbol{X} \in \mathbb{X}_j) - \Pr(Y(1) = 1, \boldsymbol{X} \in \mathbb{X}_j)\} \\
&= \sum_{j=1}^{J} \max\{0, \mathrm{E}[1\{\boldsymbol{X} \in \mathbb{X}_j\}(Y(0) - Y(1))]\}
\end{aligned}
$$

The sharp lower bound of θ can be formulated as

$$
\begin{aligned}
L(\theta) &= \mathrm{E}[\max\{0, \mu_0(\boldsymbol{X}) - \mu_1(\boldsymbol{X})\}] && (4.3) \\
&= \mathrm{E}[1\{\mu_0(\boldsymbol{X}) \geq \mu_1(\boldsymbol{X})\}(Y(0) - Y(1))]. && (4.4)
\end{aligned}
$$

We can see that the two bounds match with each other when we take $\Delta = \{\mathbb{X}_1, \mathbb{X}_2\}$ with $\mathbb{X}_1 = \{\boldsymbol{x} : \mu_0(\boldsymbol{X}) \geq \mu_1(\boldsymbol{X})\}$ and $\mathbb{X}_2 = \{\boldsymbol{x} : \mu_0(\boldsymbol{X}) < \mu_1(\boldsymbol{X})\}$. This is consistent with our observations in Proposition 4.2.2. Define $L(\theta, f_0, f_1) = \mathrm{E}[1\{f_0(\boldsymbol{X}) \geq f_1(\boldsymbol{X})\}(Y(0) - Y(1))]$ where f_0 and f_1 may be estimators of μ_0 and μ_1 based on external datasets, we can verify that $L(\theta, f_0, f_1) \leq L(\theta, \mu_0, \mu_1) \leq \theta$ for any f_1 and f_0; the choices of f_0 and f_1 correspond to the partitioning rules and $L(\theta, f_0, f_1)$ is maximized at $f_0 = \mu_0$, $f_1 = \mu_1$. In RCTs, because of the randomized treatment assignments, we can identify $L(\theta, f_0, f_1)$ by

$$
L(\theta, f_0, f_1) = \mathrm{E}[1\{f_0(\boldsymbol{X}) \geq f_1(\boldsymbol{X})\}Y \mid A = 0] - \mathrm{E}[1\{f_0(\boldsymbol{X}) \geq f_1(\boldsymbol{X})\}Y \mid A = 1],
$$

which yield the proposed partitioning-based method. In an observational study with an unconfounded TAM, the $L(\theta, f_0, f_1)$ can be identified by incorporating the doubly robust

identification of $\mathrm{E}[Y(a)]$, leading to

$$
\begin{aligned}
L(\theta, f_0, f_1) =& \mathrm{E}\left\{1\{f_0(\boldsymbol{X}) \geq f_1(\boldsymbol{X})\}\left[\frac{A}{e(\boldsymbol{X})}(Y-\mu_1(\boldsymbol{X}))+\mu_1(\boldsymbol{X})\right]\right\} \\
& -\mathrm{E}\left\{1\{f_0(\boldsymbol{X}) \geq f_1(\boldsymbol{X})\}\left[\frac{1-A}{1-e(\boldsymbol{X})}(Y-\mu_0(\boldsymbol{X}))+\mu_0(\boldsymbol{X})\right]\right\}.
\end{aligned}
$$

The above identification formula is exactly the one used in Kallus (2022).

4.10 Proofs and Technical Details

4.10.1 Proofs of Main Propositions

Proof of Proposition 4.2.1

Let $\rho_{a,t} = \Pr(\boldsymbol{X} \in \mathbb{X}_{a,t})$ for $a = 0, 1$, $t = -1, 1$. It follows that

$$
\begin{aligned}
I_\theta(\Delta_{full}) &= \sum_{(a,t)\in\mathbb{L}} \rho_{a,t} \max\{\mu_1(\mathbb{X}_{a,t}), 1-\mu_1(\mathbb{X}_{a,t}), \mu_0(\mathbb{X}_{a,t}), 1-\mu_0(\mathbb{X}_{a,t})\} \\
&\geq \sum_{(a,t)\in\mathbb{L}} \rho_{a,t}\{(1-t)/2 + t\mu_a(\mathbb{X}_{a,t})\} \\
&\geq \sum_{(a,t)\in\mathbb{L}} \int_{\mathbb{X}_{a,t}} \mu_{max}(\boldsymbol{x}) f(\boldsymbol{x}) v(d\boldsymbol{x}) \\
&\geq \int_{\mathbb{X}} \mu_{max}(\boldsymbol{x}) f(\boldsymbol{x}) v(d\boldsymbol{x}) = I_\theta(\Delta_\infty).
\end{aligned}
$$

Since $I_\theta(\Delta_{full})$ reaches the information upper bound, we must have $B_\theta(\Delta_{full}) = B_\theta(\Delta_\infty)$.

Proof of Proposition 4.2.3

Some algebra shows that

$$\begin{aligned} m_1(g) + m_0(g) &= \Pr(g(\boldsymbol{X}) \neq (1, Y(1))) + \Pr(g(\boldsymbol{X}) \neq (0, Y(0))) \\ &= \mathrm{E}\left\{ \sum_{(a,t)\in\mathbb{L}} \Pr(Y(a) = t \mid \boldsymbol{X}) 1\{g(\boldsymbol{X}) \neq (a,t)\} \right\} \\ &= C - \mathrm{E}\left\{ \sum_{(a,t)\in\mathbb{L}} \Pr(Y(a) = t \mid \boldsymbol{X}) 1\{g(\boldsymbol{X}) = (a,t)\} \right\} \end{aligned}$$

for some constant C. This loss is minimized at $g^*(\boldsymbol{x}) = \underset{(a,t)\in\mathbb{L}}{\mathrm{argmax}} \Pr(Y(a) = t \mid \boldsymbol{X} = \boldsymbol{x})$.

Proof of Proposition 4.3.1

Let $\delta_{i,j} = 1\{\boldsymbol{X}_i \in \mathbb{X}_j\}$, $\mathcal{F}_j = \{\delta_{i,j}, i = 1, \cdots, n\}$, and $\mathcal{A} = \{A_1, \cdots, A_n\}$. It is easy to see that $\mathrm{E}[\hat{\mu}_a(\mathbb{X}_j)|\mathcal{A}, \mathcal{F}_j] = \mu_a(\mathbb{X}_j)$ for $a = 0, 1$. For any $\varepsilon > 0$, the Chebyshev's inequality gives

$$\Pr(|\hat{\mu}_a(\mathbb{X}_j) - \mu_a(\mathbb{X}_j)| \geq \varepsilon|\mathcal{A}, \mathcal{F}_j) \leq \varepsilon^{-2} \operatorname{Var}(\hat{\mu}_a(\mathbb{X}_j)|\mathcal{A}, \mathcal{F}_j) = \frac{\sigma^2_{Y(a),j}}{\varepsilon^2 \sum_{i=1}^n 1\{A_i = a\}\delta_{i,j}}$$

where $\sigma^2_{Y(a),j} = \operatorname{Var}(Y(a)|\boldsymbol{X} \in \mathbb{X}_j)$. Then, we have

$$\Pr(|\hat{\mu}_a(\mathbb{X}_j) - \mu_a(\mathbb{X}_j)| \geq \varepsilon) = \mathrm{E}_{\mathcal{A},\mathcal{F}_j} \Pr(|\hat{\mu}_a(\mathbb{X}_j) - \mu_a(\mathbb{X}_j)| \geq \varepsilon|\mathcal{A}, \mathcal{F}_j) \leq \frac{\sigma^2_{Y(a),j}}{\varepsilon^2 n \pi_a \rho_j}$$

and thus $\hat{\mu}_a(\mathbb{X}_j) = \mu_a(\mathbb{X}_j) + O_p(\sigma_{Y(a),j}/\sqrt{n\pi_a\rho_j})$. Some algebra shows

$$\begin{aligned}\widehat{L}(\theta,\Delta) &= \sum_{j=1}^{J} \frac{n_j}{n} \max\{0, \hat{\mu}_0(\mathbb{X}_j) - \hat{\mu}_1(\mathbb{X}_j)\} \\ &= \sum_{j=1}^{J} (\frac{n_j}{n} - \rho_j) \max\{0, \hat{\mu}_0(\mathbb{X}_j) - \hat{\mu}_1(\mathbb{X}_j)\} + \sum_{j=1}^{J} \rho_j \max\{0, \hat{\mu}_0(\mathbb{X}_j) - \hat{\mu}_1(\mathbb{X}_j)\}.\end{aligned}$$

The first term at the right side of the above equation is $O_p(1/\sqrt{n})$ because $n_j/n = \rho_j + O_p(1/\sqrt{n})$ by the Markov inequality and $\max\{0, \hat{\mu}_0(\mathbb{X}_j) - \hat{\mu}_1(\mathbb{X}_j)\} \leq 1$. For the second term,

$$\begin{aligned}&\sum_{j=1}^{J} \rho_j \max\{0, \hat{\mu}_0(\mathbb{X}_j) - \hat{\mu}_1(\mathbb{X}_j)\} \\ &\leq \sum_{j=1}^{J} \rho_j \max\{0, \hat{\mu}_0(\mathbb{X}_j) - \mu_0(\mathbb{X}_j) - \hat{\mu}_1(\mathbb{X}_j) + \mu_1(\mathbb{X}_j)\} + L(\theta,\Delta) \\ &= L(\theta,\Delta) + O_p\left(\sum_{j=1}^{J}\sum_{a=0}^{1} \sigma_{Y(a),j}\sqrt{\rho_j/n\pi_a}\right).\end{aligned}$$

Similarly, for the estimated upper bound, we have

$$\begin{aligned}
\widehat{U}(\theta,\Delta) =& \sum_{j=1}^{J} \frac{n_j}{n} \min\{\hat{\mu}_0(\mathbb{X}_j), 1-\hat{\mu}_1(\mathbb{X}_j)\} \\
=& \sum_{j=1}^{J} (\frac{n_j}{n} - \rho_j) \min\{\hat{\mu}_0(\mathbb{X}_j), 1-\hat{\mu}_1(\mathbb{X}_j)\} + \sum_{j=1}^{J} \rho_j \max\{\hat{\mu}_0(\mathbb{X}_j), 1-\hat{\mu}_1(\mathbb{X}_j)\} \\
\leq& \sum_{j=1}^{J} \rho_j \max\{\mu_0(\mathbb{X}_j), 1-\mu_1(\mathbb{X}_j)\} + |\hat{\mu}_0(\mathbb{X}_j) - \mu_0(\mathbb{X}_j)| + |\mu_1(\mathbb{X}_j) - \hat{\mu}_1(\mathbb{X}_j)| \\
& + O_p(1/\sqrt{n}) \\
=& U(\theta,\Delta) + O_p\left(\sum_{j=1}^{J} \sum_{a=0}^{1} \sigma_{Y(a),j} \sqrt{\rho_j / n\pi_a} + 1/\sqrt{n} \right).
\end{aligned}$$

For the expected information of the estimated bounds, some algebra shows

$$\begin{aligned}
& \mathrm{E}[\hat{I}_\theta(\Delta) | \boldsymbol{n} = \{n_1, \cdots, n_J\}] \\
=& \sum_{j=1}^{J} \frac{n_j}{n} \mathrm{E}[\max\{\hat{\mu}_0(\mathbb{X}_j), 1-\hat{\mu}_0(\mathbb{X}_j), \hat{\mu}_1(\mathbb{X}_j), 1-\hat{\mu}_1(\mathbb{X}_j)\} | \boldsymbol{n}] \\
\geq& \sum_{j=1}^{J} \frac{n_j}{n} \max\{\mathrm{E}[\hat{\mu}_0(\mathbb{X}_j)|\boldsymbol{n}], 1-\mathrm{E}[\hat{\mu}_0(\mathbb{X}_j)|\boldsymbol{n}], \mathrm{E}[\hat{\mu}_1(\mathbb{X}_j)|\boldsymbol{n}], 1-\mathrm{E}[\hat{\mu}_0(\mathbb{X}_j)|\boldsymbol{n}]\} \\
\geq& \sum_{j=1}^{J} \frac{n_j}{n} \max\{\mu_0(\mathbb{X}_j), 1-\mu_0(\mathbb{X}_j), \mu_1(\mathbb{X}_j), 1-\mu_1(\mathbb{X}_j)\}.
\end{aligned}$$

The second inequality is because of the Jensen's inequality. It follows that $\mathrm{E}[\hat{I}_\theta(\Delta)] = \mathrm{EE}[\hat{I}_\theta(\Delta)|\boldsymbol{n}] \geq \sum_{j=1}^{J} \rho_j \max\{\mu_0(\mathbb{X}_j), 1-\mu_0(\mathbb{X}_j), \mu_1(\mathbb{X}_j), 1-\mu_1(\mathbb{X}_j)\} = I_\theta(\Delta)$.

Proof of Proposition 4.4.1

Consider the information carried by Δ_{plug}, we have

$$
\begin{aligned}
I_\theta(\Delta_{plug}) &= \sum_{(a,t)\in\mathbb{L}} \rho_{a,t} \max\{\mu_1(\hat{\mathbb{X}}_{a,t}), 1-\mu_1(\hat{\mathbb{X}}_{a,t}), \mu_0(\hat{\mathbb{X}}_{a,t}), 1-\mu_0(\hat{\mathbb{X}}_{a,t})\} \\
&\geq \sum_{(a,t)\in\mathbb{L}} \rho_{a,t}\{(1-t)/2 + t\mu_a(\hat{\mathbb{X}}_{a,t})\} \\
&\geq \sum_{(a,t)\in\mathbb{L}} \int_{\hat{\mathbb{X}}_{a,t}} \{(1-t)/2 + t\mu_a(\boldsymbol{x})\} f_{\boldsymbol{X}}(\boldsymbol{x})\nu(d\boldsymbol{x}) \\
&\geq \int_{\mathbb{X}} \mu_{max}(\boldsymbol{x}) f_{\boldsymbol{X}}(\boldsymbol{x})\nu(d\boldsymbol{x}) + \sum_{(a,t)\in\mathbb{L}} \int_{\hat{\mathbb{X}}_{a,t}} \{(1-t)/2 + t\mu_a(\boldsymbol{x}) - \mu_{max}(\boldsymbol{x})\} f_{\boldsymbol{X}}(\boldsymbol{x})\nu(d\boldsymbol{x}) \\
&\geq I_\theta(\Delta_\infty) + \sum_{(a,t)\in\mathbb{L}} \int_{\hat{\mathbb{X}}_{a,t}} \{t(\mu_a(\boldsymbol{x}) - \hat{\mu}_a(\boldsymbol{x})) + \hat{\mu}_{max,N}(\boldsymbol{x}) - \mu_{max}(\boldsymbol{x})\} f_{\boldsymbol{X}}(\boldsymbol{x})\nu(d\boldsymbol{x}) \\
&\geq I_\theta(\Delta_\infty) - \int_{\mathbb{X}} \{(\mu_{max}(\boldsymbol{x}) - \hat{\mu}_{max,N}(\boldsymbol{x})) + \max_{a=0,1} |\mu_a(\boldsymbol{x}) - \hat{\mu}_a(\boldsymbol{x})|\} f_{\boldsymbol{X}}(\boldsymbol{x})\nu(d\boldsymbol{x}).
\end{aligned}
$$

4.10.2 Model Calibration and Isotonic Regression

Model calibration in classification problems is a separate topic of long history and great importance in machine learning (Guo et al., 2017; Vaicenavicius et al., 2019; Minderer et al., 2021). It aims to provide more trustworthy probabilistic estimates while not changing the ranking performance of the original classification algorithms. Suppose we want to estimate $\mu(\boldsymbol{x}) = \Pr(Y = 1 \mid \boldsymbol{X} = \boldsymbol{x})$ with the sample data $\{(Y_i, \boldsymbol{X}_i), i = 1, \cdots, n\}$ as n i.i.d. copies of $(Y, \boldsymbol{X})$ where Y is a binary response variable and $\boldsymbol{X}$ is a vector of covariates. In our context of causal inference, the sample data will be the treated observations $\{(Y_i, \boldsymbol{X}_i), A_i = 1\}$ if we want to estimate $\mu_1(\boldsymbol{x}) = \Pr(Y = 1 \mid A = 1, \boldsymbol{X} = \boldsymbol{x})$ and the control observations $\{(Y_i, \boldsymbol{X}_i), A_i = 0\}$ if we want to estimate $\mu_0(\boldsymbol{x}) = \Pr(Y = 1 \mid A = 0, \boldsymbol{X} = \boldsymbol{x})$. Let $\hat{\mu}$ be an estimator of μ at hand, probably learned from the sample data. We focus in this subsection on why the model calibration can bring us better estimators of μ and how to

construct a well-calibrated estimator based on $\hat{\mu}$.

Definition A.1 (Well-calibrated estimator) *An estimator $\hat{\mu}$ is said well-calibrated if* $\Pr(Y(a) = 1 \mid \hat{\mu}(\boldsymbol{X})) = \hat{\mu}(\boldsymbol{X})$ *over the support of* $\boldsymbol{X}$.

Proposition A.1 *The true model μ is well-calibrated.*

It is often difficult and sometimes even impossible to get a perfect estimator of μ. However, the true model is necessary to be well-calibrated. We can therefore reasonably speculate that an uncalibrated estimator will be made closer to the truth after calibration, and shift our attention to seeking a better-calibrated probabilistic classifier. Indeed, in our numerical experiments, the interval estimators of θ perform better after we calibrate the nuisance estimators.

Proposition A.2 *Let* $g(z) = \Pr(Y(a) = 1 \mid \hat{\mu}(\boldsymbol{X}) = z)$ *for* $z \in [0, 1]$. *Then,* $g \circ \hat{\mu}$ *is well-calibrated.*

Proposition A.2 provides a way to calibrate an estimator: the key is to find the conditional probability $g(z) = \Pr(Y(a) = 1 \mid \hat{\mu}(\boldsymbol{X}) = z)$, which refers to a calibrator. This may look similar to an ordinary binary regression problem, but they are different in the following three aspects. First, the input of the calibrator are probabilistic estimates in the range $[0, 1]$, thus the regression is expected to be done in a unit square. Second, in order to retain the properties of the algorithms on classification, we would not like to change the rankings of the estimated probabilities after calibration, which implies that $g(z)$ should be a monotonically increasing function. Third, if $\hat{\mu}$ is obtained based on the sample data, there might be over-fitting issues if we use the sample data to estimate the calibrator as well. The last concern, luckily, can be easily addressed by a popular technique in machine learning called cross-fitting.

Platt scaling (Platt et al., 1999) and isotonic regression (Ayer et al., 1955) are two commonly used approaches in model calibration for the estimation of a calibrator (Niculescu-Mizil and Caruana, 2005). Platt scaling is a parametric method that assumes a logistic model for $g(z)$. The isotonic regression, also known as monotone regression, is non-parametric. We focus on isotonic regression because it is more flexible and generally performs better than Platt scaling in practice.

Without loss of generality, we assume that $\hat{\mu}(\boldsymbol{X}_1) \leq \hat{\mu}(\boldsymbol{X}_2) \leq \cdots \leq \hat{\mu}(\boldsymbol{X}_n)$. The isotonic regression assumes $g(z)$ is a step function as $g(z) = n^{-1} \sum_{i=1}^{n} 1\{z \leq w_i\}$ for some $0 \leq w_1, \cdots, w_n \leq 1$ and estimates $w_1, \cdots, w_n$ by maximizing $\sum_{i=1}^{n}(Y_i - w_i)^2$ with respect to the constraints $0 \leq w_1 \leq w_2 \leq \cdots \leq w_n \leq 1$. It is clear that the sequence $\hat{\mu}(\boldsymbol{X}_i), i = 1, \cdots, n$ maintains the order after isotonic regression because $g(z)$ is monotonically increasing. This quadratic optimization problem can be solved by an algorithm called pool-adjacent-violators (PAV) algorithm (Ayer et al., 1955; Busing, 2022). The cross-fitting isotonic regression algorithm can be rapidly implemented by a function named `CalibratedClassifierCV` in the Python library `scikit-learn`.

4.10.3 A Brief Review of Probabilistic Classification Algorithms

We give a brief review of the six classification algorithms adopted in our numerical experiments and on how they produce the probabilistic estimates. Similarly, we assume that we want to estimate $\mu(\boldsymbol{x}) = \Pr(Y = 1 | \boldsymbol{X} = \boldsymbol{x})$ with the sample data $\{(Y_i, \boldsymbol{X}_i), i = 1, \cdots, n\}$ as n i.i.d. copies of $(Y, \boldsymbol{X})$ where Y is a binary response variable and $\boldsymbol{X}$ is a vector of covariates.

Logistic Regression. *In logistic regression, a parametric model* $f(\boldsymbol{X}^\intercal \boldsymbol{\beta}) = 1/(1 + \exp(-\boldsymbol{X}^\intercal \boldsymbol{\beta}))$ *is postulated for* $\mu(\boldsymbol{X})$ *and an estimator* $\hat{\boldsymbol{\beta}}$ *of* $\boldsymbol{\beta}$ *is computed by maximiz-*

ing the regularized log-likelihood function

$$\frac{1}{n}\sum_{i=1}^{n}\{Y_i \log f(\boldsymbol{X}_i^\intercal\boldsymbol{\beta}) + (1-Y_i)\log(1-f(\boldsymbol{X}_i^\intercal\boldsymbol{\beta}))\} + \lambda_n||\boldsymbol{\beta}||.$$

The conditional probability is estimated as $\hat{\mu}(\boldsymbol{X}) = f(\boldsymbol{X}^\intercal\hat{\boldsymbol{\beta}})$.

Gaussian Naive Bayes. *In the perspective of Bayesian statistics, the conditional probability* $\mu(\boldsymbol{x})$ *is* $\mu(\boldsymbol{x}) = P(Y=1)f_1(\boldsymbol{x})/(P(Y=1)f_1(\boldsymbol{x}) + P(Y=-1)f_0(\boldsymbol{x}))$ *where* $f_t(\boldsymbol{x})$ *is the density of* $\boldsymbol{X}$ *given* $Y=t$ *for* $t=-1,1$. *For the* i*-th individual, let* $\boldsymbol{X}_i = (X_{i,1},\cdots,X_{i,p})^\intercal$. *In Gaussian Naive Bayes, it is assumed that the covariates are independent of each other so that* $f_t(\boldsymbol{X}_i) = \prod_{k=1}^{p} f_{t,k}(\boldsymbol{X}_{i,k})$ *where* $f_{t,k}(\boldsymbol{X}_{i,k})$ *is the density of* $\boldsymbol{X}_{i,k}$ *given* $Y=t$, *and* $f_{t,k}(\boldsymbol{X}_{i,k})$ *is modeled by the standard normal density function. For more general cases, one can consider one-dimensional kernel density estimates for the margins.*

K-Nearest-Neighbors. *The K-nearest-neighbors (KNN) algorithm is very simple and does not require any parametric assumptions on the form of* $\mu(\boldsymbol{X})$. *The simplest KNN estimates* $\mu(\boldsymbol{x})$ *by* $\hat{\mu}(\boldsymbol{x}) = \sum_{d(\boldsymbol{X}_i,\boldsymbol{x})\le q_K} Y_i/K$ *where* $d(\cdot,\cdot)$ *is a distance metric and* q_K *is the* K*-th smallest values of* $\{d(\boldsymbol{X}_i,\boldsymbol{x}), i=1,\cdots,n\}$. *Ties can be broken randomly.*

Support Vector Machines. *The support vector machines (SVM) aims to separate the individuals from different classes with a hyperplane identified by a vector* $(\omega_0,\boldsymbol{\omega}^\intercal) = (\omega_0,\omega_1,\cdots,\omega_p)$ *so that the margin between the two classes can be maximized. Since the margin is proportional to* $1/||\boldsymbol{\omega}||$, *the naive SVM solves the optimization problem*

$$\min_{\boldsymbol{\omega},\omega_0} ||\boldsymbol{\omega}|| \quad s.t. \quad Y_i(\omega^\intercal\boldsymbol{X}_i + \omega_0) \ge 1, i=1,\cdots,n.$$

Samples on the boundary of the margin, i.e., those satisfy $Y_i(\omega^\intercal \boldsymbol{X}_i + \omega_0) = 1$, *are called support vectors. When the two classes are not linearly separable, the SVM solves the following optimization problem*

$$\begin{aligned} &\min_{\boldsymbol{\omega},\omega_0} ||\boldsymbol{\omega}||^2/2 + C\sum_{i=1}^{n}\xi_i \\ &s.t. \quad Y_i(\omega^\intercal \boldsymbol{X}_i + \omega_0) \geq 1 - \xi_i, \xi_i \geq 0, i = 1, \cdots, n, \end{aligned} \tag{4.5}$$

where $\xi_1, \cdots, \xi_n$ *are slack variables that allow the samples to be on the wrong side of the margin by the distance* ξ_i, *and* C *is a cost parameter that controls the magnitude of* $\sum_{i=1}^{n}\xi_i$. *A dual problem of (4.5) can be derived as*

$$\begin{aligned} &\max_{\alpha_1,\cdots,\alpha_n} \sum_{i=1}^{n}\alpha_i + \sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j Y_i Y_j \boldsymbol{X}_i^\intercal \boldsymbol{X}_j \\ &s.t. 0 \leq \alpha_i \leq C, i = 1, \cdots, n, \sum_{i=1}^{n}\alpha_i Y_i = 0. \end{aligned} \tag{4.6}$$

A kernel trick is suggested by replacing $\boldsymbol{X}_i^\intercal \boldsymbol{X}_i$ *with* $K(\boldsymbol{X}_i, \boldsymbol{X}_j)$ *where a Gaussian radial basis kernel corresponds to* $K(\boldsymbol{X}_i, \boldsymbol{X}_j) = \exp(-\gamma||\boldsymbol{X}_i - \boldsymbol{X}_j||)$ *so that the hyperplane will be built in a transformed space with high-dimensional features. The inverse distance of any observation with covariates* $\boldsymbol{X}_i$ *to the decision hyperplane is proportional to* $D_i = \sum_{j=1}^{n}\hat{\alpha}_j K(\boldsymbol{X}_i, \boldsymbol{X}_j)$. *The Platt scaling transformes these values into probabilities by doing logistic regression with the data* $(D_i, Y_i), i = 1, \cdots, n$.

Random Forests. *A classification tree is a recursive binary tree that divides the covariates space into a bunch of partitions with its leaf nodes. The tree is usually constructed through two steps: (1) splitting the tree via splitting covariates by cutoff values based on certain rule such as Gini index or cross-entropy; the splitting will be stopped once the tree reaches*

maximum depth of the tree, maximum samples in each node, maximum number of nodes, or minimum information gains. (2) Pruning the tree based on errors of the tree on validation data and complexity of the tree. The estimated probability of a point will be the proportion of $Y = 1$ *in the node where the point is located.*

Random forests for classification are an ensemble of classification trees. However, each tree in random forests is constructed based on a bootstrap sample and $\sqrt{p}$ *of the original* p *features. The estimated probability is the average of the estimated probabilities of the trees.*

Gradient Boosting. *The gradient boosting for classification assumes that* $\mu(\boldsymbol{x}) = logit^{-1}(f(\boldsymbol{x}))$ *for a function* f *where* $logit(x) = \log(x/(1-x))$*. Here, we will assume that* $Y_i \in \{0, 1\}$*. The model for* f*, denoted by* $f_M(\boldsymbol{x}) = \sum_{i=1}^{M} \eta h_m(\boldsymbol{x})$*, is a weighted aggregation of a sequence of weak classifiers (e.g., trees). The algorithm learns the weak classifiers by sequentially minimizing the loss* $L_n(f) = n^{-1} \sum_{i=1}^{n} Y_i \log(\mu(\boldsymbol{X}_i)) + (1 - Y_i) \log(1 - \mu(\boldsymbol{X}_i)) = n^{-1} \sum_{i=1}^{n} Y_i f(\boldsymbol{X}_i) - \log\{1 + \exp(f(\boldsymbol{X}_i))\}$*. Let us begin by initializing the model* f_M *with a constant prediction as*

$$f_0(x) = c = \log(\bar{Y}/(1 - \bar{Y}))$$

where $\bar{Y} = n^{-1} \sum_{i=1}^{n} Y_i$*. At each iteration* m*, the model* f *is updated by adding the new weak learner* h_m*:*

$$f_m(x) = f_{m-1}(x) + \eta \cdot h_m(x),$$

where η *is the learning rate, a small constant that controls the contribution of each weak learner. For original gradient boosting, the weak learner* h_m *is trained to predict the negative gradient of the loss function with respect to* f *at* f_{m-1}*, i.e.,* $-\partial L_n(f_{m-1})/\partial f =$

$\bar{Y} - n^{-1}\sum_{i=1}^{n} logit^{-1}(f_{m-1}(\boldsymbol{X}_i))$. *The XGBoost trains* h_m *to predict*

$$-(\partial^2 L_n(f_{m-1})/\partial f \partial f^{\intercal})\partial L_n(f_{m-1})/\partial f,$$

which matches the Newton's method. The algorithm continues iterating until a stopping criterion is met, such as (1) a predefined number of iterations is reached, or (2) the loss function no longer improves significantly. The final probability estimator will be $logit^{-1}(\sum_{i=1}^{M} \eta \hat{h}_m(\boldsymbol{x}))$.

4.10.4 Additional Simulation Results

This section provides additional tables of the simulation results on (1) interval estimators of the six classification algorithms (Tables 4.4 and 4.5); (2) confidence intervals and their coverage rates under the nominal levels $\alpha \in \{0.05, 0.1, 0.25, 0.5\}$ for the Naive and Oracle bounds (Tables 4.6 and 4.7); (3) 75% extended confidence intervals and their coverage rates for θ (Tables 4.8); (4) interval estimators of the six classification algorithms combined with model calibration (Tables 4.9 and 4.10). This section also provides additional figures on the estimated lower and upper bounds under the KNN algorithm with σ_ε varying from 0 to 5 (Figures 4.3 and 4.4).

We briefly explain the results for the confidence intervals given in Tables 4.6 and 4.7. Denote by $C_{\alpha,L}(\Delta)$ the confidence interval for the lower bound and $C_{\alpha,U}(\Delta)$ the one for the upper bound. We focus on the two partitions Δ_0 and Δ_{full} because the true lower and upper bounds under them are known and computable. We consider the significance levels $\alpha \in \{0.05, 0.1, 0.25, 0.5\}$ and the simulation results under the Scenario 1 and 2 are respectively summarized in Tables 4.6 and 4.7. The LCovR and UCovR in the tables

represent respectively the empirical coverage rate of $C_{\alpha,L}(\Delta)$ on $L(\theta,\Delta)$ and $C_{\alpha,U}(\Delta)$ on $U(\theta,\Delta)$. In general cases where $L(\theta,\Delta)$ or $U(\theta,\Delta)$ is away from zero, the empirical coverage rates of the proposed confidence intervals are close to the nominal levels. However, there are two cases to be noted and explained. First, under Scenario 1 with $\sigma_\varepsilon = 1$ and $n = 500$, the coverage rates of $C_{\alpha,U}(\Delta_{full})$ are away from the true nominal levels because there are too few samples in some cells of Δ_{full} to validate the confidence intervals constructed based on the asymptotic distributions of the estimated bounds in those cells. As n increases to 2000, the coverage rates are much better. Second, we can see that $C_{\alpha,L}(\Delta_0)$ over-covers $L(\theta,\Delta_0)$ with the coverage rate 100% under all the setups. This is because the probability of $\widehat{L}(\theta,\Delta_0)$ is massed at zero under all the setups.

Table 4.4: Monte Carlo simulation results for the interval estimation under **Scenario 1**. All values except the percentages in the table are presented in units scaled by 10^{-2}.

		Plug-in				Partitioning					
Method	θ	Estimate	Bias	Width	CovR	Estimate	Bias	Width	CovR	TMCR	CMCR
$\sigma_\varepsilon = 1 \quad n = 500$											
Naive	0.1	$[0.0, 20.5]$	0.0	20.5	100.0%	$[0.0, 20.5]$	0.0	20.5	100.0%	NaN	NaN
Logit	0.1	$[1.5, 17.9]$	1.4	16.3	0.4%	$[0.3, 18.5]$	0.2	18.2	59.4%	35.1%	21.3%
NBayes	0.1	$[1.2, 12.2]$	1.1	11.0	0.3%	$[0.1, 5.5]$	0.1	5.4	76.6%	18.7%	12.8%
SVM	0.1	$[2.2, 15.9]$	2.1	13.7	0.0%	$[0.2, 10.5]$	0.2	10.3	58.2%	26.7%	18.6%
KNN	0.1	$[2.5, 11.7]$	2.4	9.2	0.0%	$[0.1, 15.2]$	0.1	15.1	76.6%	31.3%	18.8%
RF	0.1	$[0.8, 15.2]$	0.7	14.4	0.5%	$[0.1, 5.3]$	0.1	5.2	76.5%	17.7%	10.6%
XGBoost	0.1	$[1.3, 4.1]$	1.2	2.8	0.0%	$[0.1, 4.5]$	0.1	4.4	74.5%	17.5%	10.1%
Oracle	0.1	NaN	NaN	NaN	NaN	$[0.0, 0.6]$	0.0	0.6	73.5%	NaN	NaN
$\sigma_\varepsilon = 1 \quad n = 2000$											
Naive	0.1	$[0.0, 20.7]$	0.0	20.7	100.0%	$[0.0, 20.7]$	0.0	20.7	100.0%	NaN	NaN
Logit	0.1	$[0.2, 18.9]$	0.1	18.8	43.2%	$[0.0, 18.4]$	0.0	18.3	84.2%	33.0%	20.4%
NBayes	0.1	$[0.1, 12.8]$	0.0	12.6	47.1%	$[0.0, 3.4]$	0.0	3.4	96.7%	13.9%	10.4%
SVM	0.1	$[1.3, 7.5]$	1.2	6.2	0.0%	$[0.0, 3.8]$	0.0	3.8	97.0%	17.4%	11.9%
KNN	0.1	$[2.1, 10.7]$	2.0	8.6	0.0%	$[0.0, 11.5]$	0.0	11.5	98.5%	26.1%	16.9%
RF	0.1	$[0.4, 10.9]$	0.3	10.5	0.0%	$[0.0, 3.2]$	0.0	3.2	94.9%	13.5%	7.6%
XGBoost	0.1	$[0.3, 1.3]$	0.2	1.0	1.1%	$[0.0, 2.3]$	0.0	2.3	95.0%	13.1%	7.0%
Oracle	0.1	NaN	NaN	NaN	NaN	$[0.0, 0.6]$	0.0	0.6	98.1%	NaN	NaN
$\sigma_\varepsilon = 2 \quad n = 500$											
Naive	1.9	$[0.0, 22.7]$	0.0	22.7	100.0%	$[0.0, 22.7]$	0.0	22.7	100.0%	NaN	NaN
Logit	1.9	$[1.5, 19.6]$	0.2	18.1	69.8%	$[0.3, 20.7]$	0.0	20.4	99.5%	38.0%	23.7%
NBayes	1.9	$[1.4, 14.0]$	0.1	12.6	79.0%	$[0.2, 8.8]$	0.0	8.6	99.6%	24.1%	15.1%
SVM	1.9	$[2.0, 19.0]$	0.5	16.9	49.0%	$[0.3, 14.0]$	0.0	13.7	98.7%	31.1%	20.8%
KNN	1.9	$[2.7, 13.8]$	0.9	11.1	21.2%	$[0.2, 17.7]$	0.0	17.5	99.2%	35.1%	21.1%
RF	1.9	$[0.9, 17.7]$	0.0	16.8	95.8%	$[0.2, 8.6]$	0.0	8.4	100.0%	23.8%	14.1%
XGBoost	1.9	$[2.2, 5.9]$	0.5	3.7	43.2%	$[0.2, 8.5]$	0.0	8.3	99.7%	24.5%	14.2%
Oracle	1.9	NaN	NaN	NaN	NaN	$[0.0, 4.3]$	0.0	4.3	98.5%	NaN	NaN
$\sigma_\varepsilon = 2 \quad n = 2000$											
Naive	1.9	$[0.0, 22.9]$	0.0	22.9	100.0%	$[0.0, 22.9]$	0.0	22.9	100.0%	NaN	NaN
Logit	1.9	$[0.1, 20.7]$	0.0	20.6	100.0%	$[0.0, 20.6]$	0.0	20.6	100.0%	35.7%	22.6%
NBayes	1.9	$[0.1, 14.6]$	0.0	14.4	100.0%	$[0.0, 6.8]$	0.0	6.8	100.0%	19.8%	12.6%
SVM	1.9	$[1.5, 11.4]$	0.0	9.9	86.6%	$[0.0, 7.6]$	0.0	7.6	100.0%	23.1%	14.6%
KNN	1.9	$[2.4, 12.9]$	0.5	10.5	9.9%	$[0.0, 14.3]$	0.0	14.3	100.0%	30.6%	19.2%
RF	1.9	$[0.4, 13.6]$	0.0	13.2	100.0%	$[0.0, 6.6]$	0.0	6.6	100.0%	19.7%	11.0%
XGBoost	1.9	$[0.8, 2.7]$	0.0	1.9	98.3%	$[0.0, 6.3]$	0.0	6.3	100.0%	20.6%	11.2%
Oracle	1.9	NaN	NaN	NaN	NaN	$[0.0, 4.4]$	0.0	4.4	100.0%	NaN	NaN

Table 4.5: Monte Carlo simulation results for the interval estimation under **Scenario 2**. All values except the percentages in the table are presented in units scaled by 10^{-2}.

		Plug-in				Partitioning					
Method	θ	Estimate	Bias	Width	CovR	Estimate	Bias	Width	CovR	TMCR	CMCR
$\sigma_\varepsilon = 1 \quad n = 500$											
Naive	19.7	[0.0, 20.6]	0.6	20.6	61.6%	[0.0, 20.6]	0.6	20.6	61.6%	NaN	NaN
Logit	19.7	[3.7, 19.9]	0.9	16.2	51.7%	[1.9, 20.3]	0.7	18.4	58.4%	39.3%	21.4%
NBayes	19.7	[13.2, 21.3]	0.5	8.0	70.6%	[17.0, 20.5]	0.7	3.5	50.7%	23.1%	12.8%
SVM	19.7	[6.2, 20.4]	0.6	14.1	60.7%	[10.5, 20.4]	0.6	9.9	61.4%	32.4%	18.6%
KNN	19.7	[3.0, 12.4]	7.2	9.5	0.1%	[7.3, 20.4]	0.6	13.2	62.3%	35.9%	18.7%
RF	19.7	[7.6, 20.5]	0.5	12.9	64.2%	[16.7, 20.4]	0.7	3.7	49.5%	21.7%	10.6%
XGBoost	19.7	[16.2, 17.9]	2.2	1.7	13.6%	[17.6, 20.4]	0.8	2.9	43.8%	20.0%	10.0%
Oracle	19.7	NaN	NaN	NaN	NaN	[19.2, 19.8]	1.4	0.6	11.4%	NaN	NaN
$\sigma_\varepsilon = 1 \quad n = 2000$											
Naive	19.7	[0.0, 20.7]	0.2	20.7	78.1%	[0.0, 20.7]	0.2	20.7	78.1%	NaN	NaN
Logit	19.7	[2.2, 20.5]	0.2	18.3	73.9%	[1.7, 20.6]	0.2	18.8	75.9%	37.1%	20.4%
NBayes	19.7	[13.3, 22.3]	0.0	8.9	96.5%	[18.7, 20.6]	0.2	1.9	62.4%	19.1%	10.4%
SVM	19.7	[13.5, 19.9]	0.3	6.4	58.1%	[16.3, 20.5]	0.1	4.2	78.7%	24.5%	11.9%
KNN	19.7	[4.8, 12.8]	6.9	8.1	0.0%	[11.3, 20.6]	0.2	9.3	79.2%	30.6%	16.9%
RF	19.7	[11.7, 20.5]	0.1	8.8	80.0%	[18.4, 20.6]	0.2	2.1	66.9%	16.5%	7.6%
XGBoost	19.7	[18.7, 19.2]	0.8	0.5	14.9%	[18.7, 20.5]	0.3	1.8	58.1%	16.3%	7.1%
Oracle	19.7	NaN	NaN	NaN	NaN	[19.3, 19.9]	0.6	0.7	24.8%	NaN	NaN
$\sigma_\varepsilon = 2 \quad n = 500$											
Naive	20.6	[0.0, 22.6]	0.3	22.6	75.9%	[0.0, 22.6]	0.3	22.6	75.9%	NaN	NaN
Logit	20.6	[4.1, 21.8]	0.5	17.6	65.5%	[2.0, 22.2]	0.4	20.2	70.5%	40.2%	23.6%
NBayes	20.6	[13.2, 23.0]	0.3	9.8	79.1%	[15.9, 22.4]	0.3	6.5	75.5%	28.4%	15.2%
SVM	20.6	[5.2, 22.2]	0.4	17.0	74.5%	[9.1, 22.3]	0.4	13.2	74.5%	35.2%	20.6%
KNN	20.6	[3.5, 14.6]	6.1	11.1	1.3%	[6.2, 22.4]	0.3	16.2	74.7%	38.5%	21.0%
RF	20.6	[7.4, 22.5]	0.3	15.1	78.5%	[15.6, 22.3]	0.3	6.7	73.7%	28.6%	14.0%
XGBoost	20.6	[16.2, 18.8]	2.3	2.6	19.6%	[16.0, 22.4]	0.3	6.4	73.5%	27.7%	14.3%
Oracle	20.6	NaN	NaN	NaN	NaN	[18.6, 21.8]	0.6	3.3	51.9%	NaN	NaN
$\sigma_\varepsilon = 2 \quad n = 2000$											
Naive	20.6	[0.0, 23.0]	0.0	23.0	95.4%	[0.0, 23.0]	0.0	23.0	95.4%	NaN	NaN
Logit	20.6	[2.5, 22.8]	0.0	20.2	94.9%	[1.8, 22.8]	0.0	21.0	94.5%	37.7%	22.6%
NBayes	20.6	[13.2, 24.3]	0.0	11.1	99.5%	[18.1, 22.8]	0.0	4.7	95.4%	25.2%	12.6%
SVM	20.6	[11.9, 22.0]	0.1	10.1	86.8%	[15.3, 22.8]	0.0	7.5	96.5%	29.6%	14.6%
KNN	20.6	[5.2, 15.2]	5.5	10.0	0.0%	[10.0, 22.9]	0.0	12.9	95.8%	34.8%	19.2%
RF	20.6	[11.5, 22.8]	0.0	11.3	96.1%	[17.9, 22.8]	0.0	4.9	95.8%	24.2%	11.1%
XGBoost	20.6	[19.2, 20.4]	0.7	1.2	28.3%	[17.7, 22.8]	0.0	5.1	95.9%	24.8%	11.3%
Oracle	20.6	NaN	NaN	NaN	NaN	[18.8, 22.3]	0.1	3.5	88.0%	NaN	NaN

Table 4.6: Monte Carlo simulation results for the confidence intervals under **Scenario 1**. All values except the percentages in the table are presented in units scaled by 10^{-2}.

		Naive					Oracle					
α	$L(\theta,\Delta_0)$	$C_{\alpha,L}(\Delta_0)$	LCovR	$U(\theta,\Delta_0)$	$C_{\alpha,U}(\Delta_0)$	UCovR	$L(\theta,\Delta_{full})$	$C_{\alpha,L}(\Delta_{full})$	LCovR	$U(\theta,\Delta_{full})$	$C_{\alpha,U}(\Delta_{full})$	UCovR
$\sigma_\varepsilon=1$	$n=500$											
0.05	0.0	$[0.0,0.0]$	100.0%	20.7	$[15.7,25.8]$	94.1%	0.0	$[0.0,0.0]$	100.0%	0.6	$[0.0,1.4]$	74.0%
0.10	0.0	$[0.0,0.0]$	100.0%	20.7	$[16.5,25.0]$	89.8%	0.0	$[0.0,0.0]$	100.0%	0.6	$[0.0,1.2]$	73.6%
0.25	0.0	$[0.0,0.0]$	100.0%	20.7	$[17.8,23.7]$	76.1%	0.0	$[0.0,0.0]$	100.0%	0.6	$[0.1,1.0]$	67.3%
0.50	0.0	$[0.0,0.0]$	100.0%	20.7	$[19.0,22.5]$	51.3%	0.0	$[0.0,0.0]$	100.0%	0.6	$[0.3,0.8]$	55.3%
$\sigma_\varepsilon=1$	$n=2000$											
0.05	0.0	$[0.0,0.0]$	100.0%	20.7	$[18.2,23.2]$	95.1%	0.0	$[0.0,0.0]$	100.0%	0.6	$[0.1,1.0]$	90.7%
0.10	0.0	$[0.0,0.0]$	100.0%	20.7	$[18.6,22.8]$	90.8%	0.0	$[0.0,0.0]$	100.0%	0.6	$[0.2,0.9]$	88.5%
0.25	0.0	$[0.0,0.0]$	100.0%	20.7	$[19.2,22.2]$	75.2%	0.0	$[0.0,0.0]$	100.0%	0.6	$[0.3,0.8]$	73.4%
0.50	0.0	$[0.0,0.0]$	100.0%	20.7	$[19.8,21.5]$	50.3%	0.0	$[0.0,0.0]$	100.0%	0.6	$[0.4,0.7]$	43.6%
$\sigma_\varepsilon=2$	$n=500$											
0.05	0.0	$[0.0,0.0]$	100.0%	22.8	$[17.5,27.9]$	94.3%	0.0	$[0.0,0.0]$	100.0%	4.3	$[1.9,6.9]$	92.5%
0.10	0.0	$[0.0,0.0]$	100.0%	22.8	$[18.4,27.1]$	91.6%	0.0	$[0.0,0.0]$	100.0%	4.3	$[2.3,6.5]$	89.5%
0.25	0.0	$[0.0,0.0]$	100.0%	22.8	$[19.7,25.8]$	76.0%	0.0	$[0.0,0.0]$	100.0%	4.3	$[2.9,5.9]$	74.8%
0.50	0.0	$[0.0,0.0]$	100.0%	22.8	$[20.9,24.5]$	51.3%	0.0	$[0.0,0.0]$	100.0%	4.3	$[3.5,5.2]$	48.4%
$\sigma_\varepsilon=2$	$n=2000$											
0.05	0.0	$[0.0,0.0]$	100.0%	23.0	$[20.3,25.5]$	95.2%	0.0	$[0.0,0.0]$	100.0%	4.5	$[3.2,5.7]$	94.1%
0.10	0.0	$[0.0,0.0]$	100.0%	23.0	$[20.7,25.0]$	90.2%	0.0	$[0.0,0.0]$	100.0%	4.5	$[3.4,5.5]$	90.7%
0.25	0.0	$[0.0,0.0]$	100.0%	23.0	$[21.3,24.4]$	77.8%	0.0	$[0.0,0.0]$	100.0%	4.5	$[3.7,5.2]$	76.9%
0.50	0.0	$[0.0,0.0]$	100.0%	23.0	$[22.0,23.8]$	53.7%	0.0	$[0.0,0.0]$	100.0%	4.5	$[4.0,4.9]$	50.0%

Table 4.7: Monte Carlo simulation results for the confidence intervals under **Scenario 2**. All values except the percentages in the table are presented in units scaled by 10^{-2}.

		Naive					Oracle					
α	$L(\theta,\Delta_0)$	$C_{\alpha,L}(\Delta_0)$	LCovR	$U(\theta,\Delta_0)$	$C_{\alpha,U}(\Delta_0)$	UCovR	$L(\theta,\Delta_{full})$	$C_{\alpha,L}(\Delta_{full})$	LCovR	$U(\theta,\Delta_{full})$	$C_{\alpha,U}(\Delta_{full})$	UCovR
$\sigma_\varepsilon=1$	$n=500$											
0.05	0.0	$[0.0,0.0]$	100.0%	20.7	$[15.5,25.5]$	94.4%	19.3	$[14.2,23.3]$	94.1%	20.0	$[15.4,24.6]$	95.2%
0.10	0.0	$[0.0,0.0]$	100.0%	20.7	$[16.3,24.7]$	89.6%	19.3	$[15.2,22.6]$	89.9%	20.0	$[16.1,23.6]$	89.5%
0.25	0.0	$[0.0,0.0]$	100.0%	20.7	$[17.6,23.4]$	78.0%	19.3	$[16.4,21.5]$	73.5%	20.0	$[17.2,22.4]$	77.1%
0.50	0.0	$[0.0,0.0]$	100.0%	20.7	$[18.8,22.2]$	52.6%	19.3	$[17.5,20.5]$	49.7%	20.0	$[18.3,21.3]$	52.9%
$\sigma_\varepsilon=1$	$n=2000$											
0.05	0.0	$[0.0,0.0]$	100.0%	20.7	$[18.2,23.3]$	94.7%	19.3	$[17.2,21.5]$	94.3%	20.0	$[17.9,22.2]$	94.7%
0.10	0.0	$[0.0,0.0]$	100.0%	20.7	$[18.6,22.9]$	89.0%	19.3	$[17.6,21.2]$	89.8%	20.0	$[18.3,21.9]$	88.9%
0.25	0.0	$[0.0,0.0]$	100.0%	20.7	$[19.3,22.2]$	74.8%	19.3	$[18.1,20.6]$	76.1%	20.0	$[18.8,21.3]$	75.1%
0.50	0.0	$[0.0,0.0]$	100.0%	20.7	$[19.9,21.6]$	50.6%	19.3	$[18.6,20.1]$	49.4%	20.0	$[19.3,20.8]$	49.7%
$\sigma_\varepsilon=2$	$n=500$											
0.05	0.0	$[0.0,0.0]$	100.0%	22.8	$[17.8,28.2]$	95.4%	18.7	$[13.6,23.1]$	96.8%	22.2	$[17.7,27.7]$	96.4%
0.10	0.0	$[0.0,0.0]$	100.0%	22.8	$[18.6,27.4]$	90.7%	18.7	$[14.7,22.4]$	92.1%	22.2	$[18.4,26.5]$	90.7%
0.25	0.0	$[0.0,0.0]$	100.0%	22.8	$[19.9,26.1]$	75.0%	18.7	$[16.1,21.3]$	77.2%	22.2	$[19.6,25.1]$	75.3%
0.50	0.0	$[0.0,0.0]$	100.0%	22.8	$[21.2,24.8]$	47.6%	18.7	$[17.2,20.3]$	50.0%	22.2	$[20.7,23.9]$	50.0%
$\sigma_\varepsilon=2$	$n=2000$											
0.05	0.0	$[0.0,0.0]$	100.0%	22.8	$[20.3,25.6]$	94.8%	18.8	$[16.6,21.0]$	95.3%	22.2	$[20.0,24.6]$	94.9%
0.10	0.0	$[0.0,0.0]$	100.0%	22.8	$[20.8,25.1]$	90.0%	18.8	$[17.0,20.7]$	90.1%	22.2	$[20.4,24.2]$	90.7%
0.25	0.0	$[0.0,0.0]$	100.0%	22.8	$[21.4,24.5]$	75.0%	18.8	$[17.5,20.1]$	74.2%	22.2	$[21.0,23.7]$	75.1%
0.50	0.0	$[0.0,0.0]$	100.0%	22.8	$[22.1,23.8]$	49.4%	18.8	$[18.1,19.6]$	50.4%	22.2	$[21.5,23.1]$	49.7%

Table 4.8: Monte Carlo simulation results on the extended confidence intervals with $\alpha = 0.25$. All values except the percentages in the table are presented in units scaled by 10^{-2}.

		Scenario 1				Scenario 2			
Method	θ	Estimate	Bias	Width	CovR	Estimate	Bias	Width	CovR
$\sigma_\varepsilon = 1 \quad n = 500$									
Naive	19.7	[0.0, 23.5]	0.0	23.5	100.00%	[0.0, 23.6]	0.1	23.6	93.20%
Logit	19.7	[0.0, 23.0]	0.0	23.0	95.40%	[0.3, 25.0]	0.0	24.6	97.30%
NBayes	19.7	[0.0, 7.9]	0.0	7.9	98.60%	[13.0, 24.5]	0.0	11.5	98.30%
SVM	19.7	[0.0, 13.8]	0.0	13.8	97.40%	[7.0, 24.6]	0.0	17.7	98.40%
KNN	19.7	[0.0, 19.0]	0.0	19.0	98.70%	[4.0, 24.6]	0.0	20.7	97.20%
RF	19.7	[0.0, 7.7]	0.0	7.7	97.60%	[13.2, 24.2]	0.0	11.0	96.80%
XGBoost	19.7	[0.0, 6.6]	0.0	6.6	98.40%	[13.8, 24.4]	0.0	10.5	97.00%
Oracle	19.7	[0.0, 1.0]	0.0	1.0	73.50%	[16.6, 22.6]	0.2	6.0	82.80%
$\sigma_\varepsilon = 1 \quad n = 2000$									
Naive	19.7	[0.0, 22.2]	0.0	22.2	100.00%	[0.0, 22.2]	0.0	22.2	96.90%
Logit	19.7	[0.0, 20.4]	0.0	20.4	98.70%	[0.6, 22.8]	0.0	22.2	98.90%
NBayes	19.7	[0.0, 4.3]	0.0	4.3	99.90%	[16.8, 22.5]	0.0	5.7	98.90%
SVM	19.7	[0.0, 4.7]	0.0	4.7	100.00%	[14.6, 22.4]	0.0	7.9	99.00%
KNN	19.7	[0.0, 13.2]	0.0	13.2	99.90%	[9.6, 22.6]	0.0	13.0	99.00%
RF	19.7	[0.0, 4.1]	0.0	4.1	99.80%	[16.7, 22.4]	0.0	5.7	98.90%
XGBoost	19.7	[0.0, 3.1]	0.0	3.1	100.00%	[16.9, 22.4]	0.0	5.4	98.80%
Oracle	19.7	[0.0, 0.8]	0.0	0.8	99.60%	[18.0, 21.2]	0.1	3.2	85.10%
$\sigma_\varepsilon = 2 \quad n = 500$									
Naive	20.6	[0.0, 25.8]	0.0	25.8	100.00%	[0.0, 25.6]	0.0	25.6	97.60%
Logit	20.6	[0.0, 25.5]	0.0	25.5	100.00%	[0.3, 27.0]	0.0	26.7	99.40%
NBayes	20.6	[0.0, 11.9]	0.0	11.9	100.00%	[11.7, 26.8]	0.0	15.0	99.40%
SVM	20.6	[0.0, 17.9]	0.0	17.8	100.00%	[5.7, 26.7]	0.0	21.0	99.50%
KNN	20.6	[0.0, 21.8]	0.0	21.8	100.00%	[2.8, 26.8]	0.0	24.0	99.30%
RF	20.6	[0.0, 11.6]	0.0	11.6	100.00%	[11.9, 26.5]	0.0	14.6	99.20%
XGBoost	20.6	[0.0, 11.4]	0.0	11.4	100.00%	[11.9, 26.7]	0.0	14.8	99.50%
Oracle	20.6	[0.0, 5.8]	0.0	5.8	99.80%	[15.8, 24.8]	0.0	9.0	95.10%
$\sigma_\varepsilon = 2 \quad n = 2000$									
Naive	20.6	[0.0, 24.4]	0.0	24.4	100.00%	[0.0, 24.5]	0.0	24.5	99.70%
Logit	20.6	[0.0, 22.8]	0.0	22.8	100.00%	[0.6, 25.1]	0.0	24.5	99.80%
NBayes	20.6	[0.0, 8.1]	0.0	8.1	100.00%	[16.1, 24.9]	0.0	8.7	100.00%
SVM	20.6	[0.0, 9.0]	0.0	9.0	100.00%	[13.5, 24.9]	0.0	11.3	100.00%
KNN	20.6	[0.0, 16.1]	0.0	16.1	100.00%	[8.2, 25.0]	0.0	16.8	99.60%
RF	20.6	[0.0, 7.9]	0.0	7.9	100.00%	[16.1, 24.8]	0.0	8.8	100.00%
XGBoost	20.6	[0.0, 7.6]	0.0	7.6	100.00%	[15.9, 24.8]	0.0	9.0	100.00%
Oracle	20.6	[0.0, 5.2]	0.0	5.2	100.00%	[17.6, 23.7]	0.0	6.1	99.00%

Table 4.9: Monte Carlo simulation results on the interval estimation combined with model calibration under **Scenario 1**. All values except the percentages in the table are presented in units scaled by 10^{-2}.

		Plug-in				Partitioning					
Method	θ	Estimate	Bias	Width	CovR	Estimate	Bias	Width	CovR	TMCR	CMCR
$\sigma_\varepsilon = 1 \quad n = 500$											
Logit	0.1	[1.6, 18.0]	1.5	16.4	0.10%	[0.3, 18.4]	0.2	18.2	57.60%	35.10%	21.30%
NBayes	0.1	[2.3, 9.2]	2.2	6.9	0.00%	[0.1, 5.4]	0.1	5.3	76.70%	18.70%	12.80%
SVM	0.1	[3.0, 13.4]	2.9	10.4	0.00%	[0.2, 10.4]	0.2	10.2	60.60%	26.70%	18.60%
KNN	0.1	[2.0, 18.1]	1.9	16.0	0.00%	[0.2, 14.7]	0.2	14.5	57.70%	31.30%	18.80%
RF	0.1	[2.0, 7.0]	1.9	5.0	0.00%	[0.1, 4.3]	0.1	4.2	77.00%	17.70%	10.60%
XGBoost	0.1	[1.7, 7.1]	1.6	5.4	0.00%	[0.1, 4.5]	0.1	4.4	74.90%	17.50%	10.10%
$\sigma_\varepsilon = 1 \quad n = 2000$											
Logit	0.1	[0.4, 17.6]	0.3	17.3	10.90%	[0.1, 18.2]	0.0	18.2	79.40%	33.00%	20.40%
NBayes	0.1	[0.5, 4.5]	0.4	3.9	0.20%	[0.0, 2.9]	0.0	2.9	96.00%	13.90%	10.40%
SVM	0.1	[1.3, 7.1]	1.2	5.8	0.00%	[0.0, 4.0]	0.0	4.0	97.20%	17.40%	11.90%
KNN	0.1	[2.0, 16.1]	1.9	14.1	0.00%	[0.0, 10.1]	0.0	10.1	95.20%	26.10%	16.90%
RF	0.1	[0.6, 3.9]	0.5	3.3	0.00%	[0.0, 2.7]	0.0	2.7	95.50%	13.50%	7.60%
XGBoost	0.1	[0.5, 3.7]	0.4	3.2	0.00%	[0.0, 2.3]	0.0	2.3	96.00%	13.10%	7.00%
$\sigma_\varepsilon = 2 \quad n = 500$											
Logit	1.9	[1.6, 19.8]	0.3	18.2	66.10%	[0.3, 20.7]	0.0	20.4	98.70%	38.00%	23.70%
NBayes	1.9	[2.3, 11.6]	0.6	9.3	36.90%	[0.2, 8.8]	0.0	8.6	99.80%	24.10%	15.10%
SVM	1.9	[3.0, 16.0]	1.2	13.0	15.30%	[0.3, 13.7]	0.0	13.4	99.20%	31.10%	20.80%
KNN	1.9	[1.8, 20.5]	0.3	18.7	58.30%	[0.3, 17.7]	0.0	17.4	98.60%	35.10%	21.10%
RF	1.9	[2.5, 10.8]	0.7	8.3	27.50%	[0.2, 8.0]	0.0	7.8	99.80%	23.80%	14.10%
XGBoost	1.9	[2.3, 11.4]	0.6	9.1	34.90%	[0.2, 8.5]	0.0	8.3	99.70%	24.50%	14.20%
$\sigma_\varepsilon = 2 \quad n = 2000$											
Logit	1.9	[0.3, 19.6]	0.0	19.3	100.00%	[0.1, 20.5]	0.0	20.5	100.00%	35.70%	22.60%
NBayes	1.9	[0.5, 7.5]	0.0	6.9	100.00%	[0.0, 6.5]	0.0	6.5	100.00%	19.80%	12.60%
SVM	1.9	[1.5, 10.7]	0.0	9.2	84.80%	[0.0, 7.8]	0.0	7.8	100.00%	23.10%	14.60%
KNN	1.9	[1.7, 19.1]	0.1	17.5	71.90%	[0.0, 13.8]	0.0	13.8	100.00%	30.60%	19.20%
RF	1.9	[0.8, 7.4]	0.0	6.7	100.00%	[0.0, 6.3]	0.0	6.3	100.00%	19.70%	11.00%
XGBoost	1.9	[0.9, 8.1]	0.0	7.2	100.00%	[0.0, 6.2]	0.0	6.2	100.00%	20.60%	11.20%

Table 4.10: Monte Carlo simulation results on the interval estimation combined with model calibration under **Scenario 2**. All values except the percentages in the table are presented in units scaled by 10^{-2}.

		Plug-in				Partitioning					
Method	θ	Estimate	Bias	Width	CovR	Estimate	Bias	Width	CovR	TMCR	CMCR
$\sigma_\varepsilon = 1 \quad n = 500$											
Logit	19.7	[3.6, 20.1]	0.8	16.5	56.10%	[1.9, 20.4]	0.7	18.5	59.40%	39.30%	21.40%
NBayes	19.7	[14.1, 19.8]	0.9	5.7	51.10%	[17.2, 20.5]	0.7	3.3	49.00%	23.10%	12.80%
SVM	19.7	[8.7, 19.3]	1.1	10.6	42.30%	[10.6, 20.3]	0.6	9.8	61.50%	32.40%	18.60%
KNN	19.7	[3.8, 19.9]	0.9	16.1	52.30%	[7.3, 20.4]	0.6	13.0	60.90%	35.90%	18.70%
RF	19.7	[14.7, 19.4]	1.0	4.7	43.50%	[17.0, 20.5]	0.7	3.5	51.20%	21.70%	10.60%
XGBoost	19.7	[15.7, 19.9]	0.9	4.2	48.40%	[17.6, 20.5]	0.8	3.0	46.60%	20.00%	10.00%
$\sigma_\varepsilon = 1 \quad n = 2000$											
Logit	19.7	[3.1, 20.4]	0.2	17.3	70.70%	[2.0, 20.5]	0.2	18.6	76.00%	37.10%	20.40%
NBayes	19.7	[17.7, 20.4]	0.2	2.7	68.20%	[18.7, 20.6]	0.2	1.9	62.00%	19.10%	10.40%
SVM	19.7	[13.7, 19.5]	0.6	5.8	41.20%	[16.1, 20.5]	0.1	4.4	77.70%	24.50%	11.90%
KNN	19.7	[6.2, 20.1]	0.3	13.9	61.30%	[12.0, 20.6]	0.1	8.6	79.00%	30.60%	16.90%
RF	19.7	[17.6, 20.2]	0.2	2.7	67.70%	[18.6, 20.6]	0.2	2.0	65.00%	16.50%	7.60%
XGBoost	19.7	[18.2, 20.3]	0.2	2.1	65.60%	[18.8, 20.5]	0.3	1.8	58.60%	16.30%	7.10%
$\sigma_\varepsilon = 2 \quad n = 500$											
Logit	20.6	[3.9, 22.0]	0.5	18.1	69.30%	[2.0, 22.2]	0.4	20.3	71.80%	40.20%	23.60%
NBayes	20.6	[13.7, 21.7]	0.6	8.1	65.60%	[16.1, 22.4]	0.3	6.3	74.50%	28.40%	15.20%
SVM	20.6	[8.3, 21.2]	0.7	13.0	58.30%	[9.5, 22.3]	0.3	12.8	75.20%	35.20%	20.60%
KNN	20.6	[3.6, 21.9]	0.5	18.3	66.20%	[6.3, 22.3]	0.4	16.0	74.10%	38.50%	21.00%
RF	20.6	[13.3, 21.2]	0.7	7.9	57.10%	[15.8, 22.4]	0.3	6.7	76.30%	28.60%	14.00%
XGBoost	20.6	[13.9, 21.8]	0.5	7.9	66.50%	[16.1, 22.4]	0.3	6.3	73.00%	27.70%	14.30%
$\sigma_\varepsilon = 2 \quad n = 2000$											
Logit	20.6	[3.4, 22.7]	0.0	19.3	92.60%	[2.0, 22.8]	0.0	20.8	94.70%	37.70%	22.60%
NBayes	20.6	[17.2, 22.6]	0.0	5.4	94.00%	[18.1, 22.8]	0.0	4.8	95.10%	25.20%	12.60%
SVM	20.6	[12.6, 21.7]	0.1	9.1	78.20%	[15.1, 22.8]	0.0	7.7	96.40%	29.60%	14.60%
KNN	20.6	[5.3, 22.4]	0.1	17.1	91.10%	[10.6, 22.9]	0.0	12.2	95.70%	34.80%	19.20%
RF	20.6	[16.8, 22.4]	0.0	5.6	92.30%	[18.0, 22.8]	0.0	4.8	95.40%	24.20%	11.10%
XGBoost	20.6	[16.7, 22.4]	0.0	5.7	93.00%	[18.0, 22.8]	0.0	4.8	94.90%	24.80%	11.30%

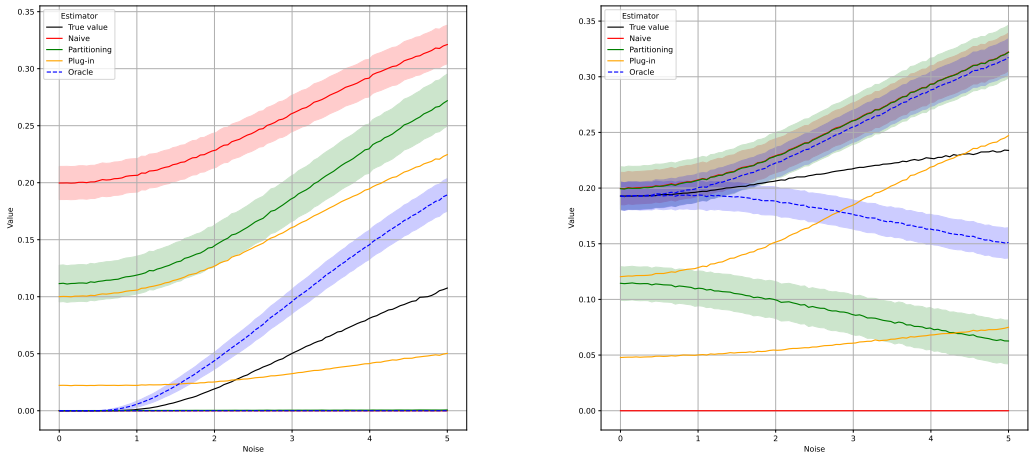


Figure 4.3: Naive and Oracle bounds, estimated plug-in and partitioning-based bounds using the KNN algorithm, and their corresponding 75% confidence bands. X-axis denotes the value of σ_ε. The plot for Scenario 1 is presented in the left panel, and the plot for Scenario 2 is presented in the right panel.

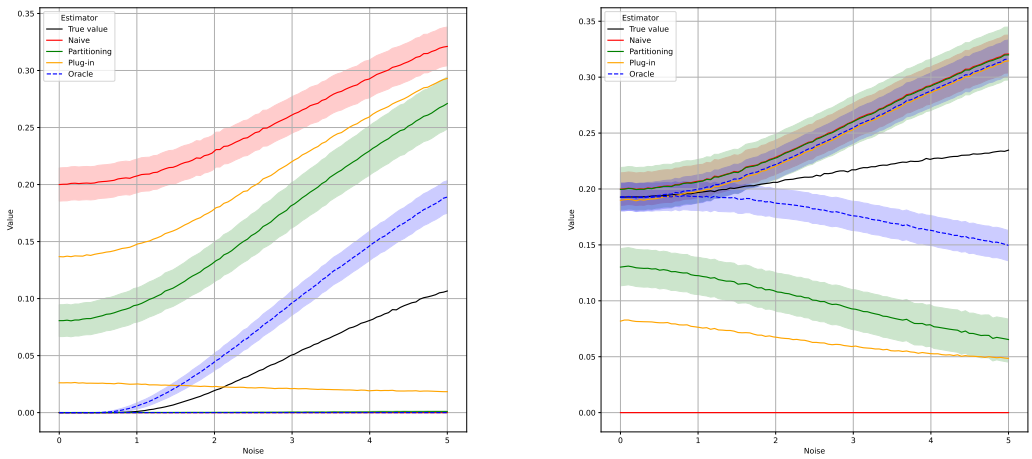


Figure 4.4: Naive and Oracle bounds, estimated plug-in and partitioning-based bounds using the KNN algorithm combined with model calibration, and their corresponding 75% confidence bands. X-axis denotes the value of σ_ε. The plot for Scenario 1 is presented in the left panel, and the plot for Scenario 2 is presented in the right panel.

Chapter 5

Nonparametric Bounds in Mediation Analysis in the Presence of Unmeasured Confounding and Imperfect Compliance

5.1 Introduction

Causal mediation anlysis (CMA) plays a crucial role in various fields, including social sciences, psychological research, economics, and beyond. Unlike traditional causal inference, which primarily examines the total effects of a treatment on specific outcomes, CMA digs into the underlying causal mechanisms by exploring how the treatment influences outcomes through mediators. This approach shifts the focus from merely identifying treatment effects to understanding the causal pathways that explain how the treatment produces its

effects.

We focus on CMA in randomized controlled trials (RCTs). Specifically, we consider the setting described in Angrist et al. (1996) and Balke and Pearl (1997) where treatment assignments are randomized but imperfect compliance to the assignments introduces confounding bias. The treatment assignment serves as an instrumental variable in this setting assuming the exclusion-restriction assumption. Such scenario is very common in RCTs (Cuzick et al., 1997; Imbens and Rubin, 1997), where the practitioners typically perform an "intent-to-treat" analysis to evaluate the effect of the treatment assignment. However, the naive "intent-to-treat" analysis can result in misleading conclusions if the treatment effectiveness rather than the assignment effectiveness is of primary interest (Pearl, 2009).

In CMA, the average causal mediation effect (ACME), which operates through the mediator, and the natural direct effect (NDE), which captures the effect through other pathways, are two central parameters (Pearl, 2001; Imai et al., 2010). In RCTs with imperfect compliance, however, the ACME and the NDE can not be identified from the data unless we have full observation of the treatment-mediator, treatment-outcome, and mediator-outcome confounding factors. Imai et al. (2010) demonstrated that these effects can be identified under the sequential ignorability assumption, which, however, is very strong, untestable, and often fragile in practice. In many cases, imposing such stringent identification assumptions may lead to biased results, making it impractical to rely solely on point estimates. To address this challenge, two strategies are commonly employed in existing literature. The first is to adopt sensitivity analysis, which assesses how robust parameter estimates are to potential violations of identification assumptions (VanderWeele and Chiba, 2014; Ding and Vanderweele, 2016). The second approach, known as partial identification, derives bounds on the parameters under minimal assumptions. In this chapter, we focus on the latter approach.

Extensive research has been conducted on partial identification in causal inference. For example, Manski (1990) showed nonparametric bounds of the average treatment effect in two-arm RCT with imperfect compliance, and its sharp bounds when the outcomes are binary are given by Balke and Pearl (1997). Cheng and Small (2006) extended their results to average treatment effects within principal strata in three-arm randomized experiments. For CMA in randomized trials, Sjölander (2009) derived the bounds on the NDE. The bounds on other causal parameters such as the controlled direct and indirect effects also have been studied in either randomized experiments or observational studies (Cai et al., 2008; VanderWeele, 2011). For a comprehensive review of the partial identification results on average treatment effects using instrumental variables, readers can refer to Swanson et al. (2018).

To the best of our knowledge, however, no prior research has established the bounds on the ACME and the NDE in RCTs with noncompliance, which is essentially a pivotal scenario in practice. We fill in the gap in this chapter by providing nonparametric bounds on the ACME and the NDE under possibly minimal assumptions that enable the treatment assignment variable to be a valid instrument. Furthermore, we consider additional monotonicity assumptions which lead to tighter bounds of the causal parameters if the imposed monotonicity assumptions are true. Those monotonicity assumptions can be guaranteed practically by design. We also derive the sharp (tightest) nonparametric bounds for a set of local causal parameters defined on the subpopulation of treatment-assignment compliers. These local parameters may also be concerned in practice.

The rest of this chapter is organized as follows. Section 5.2 formulates the problem of interest and explains the assumptions underlying our model. Sections 5.3 and 5.4 derive nonparametric bounds on the causal parameters of interest, with the former addressing the bounds on population causal parameters and the latter focusing on the bounds on

local causal parameters. In Section 5.5, we apply these bounds to a large-scale online A/B testing dataset to illustrate their practical utility. We conclude this chapter in Section 5.6. Technical proofs and additional results are provided in Section 5.7.

5.2 Problem Setup

5.2.1 Data Generating Mechanism

We consider CMA with RCT data. The units in the study are randomly assigned to two treatment groups, indicated by a random variable $Z \in \{0, 1\}$. However, there are possibly assignment noncompliance and the actual treatment state is given by the variable $A \in \{0, 1\}$. We define $A(1)$ and $A(0)$ as two potential outcomes of A, i.e., the values of A would have been, under respectively $Z = 1$ and $Z = 0$. Then, the pair $(A(1), A(0))$ has four potential values, with $(1, 0)$ corresponding to the (treatment-assignment) compliers, $(1, 1)$ corresponding to the always-takers, $(0, 0)$ corresponding to the never-takers, and $(0, 1)$ corresponding to the defiers. Let Y denote the outcome and M a mediator of interest, both measured after the treatment. We further assume Y and M are also dichotomous, taking value either 0 or 1. The temporal sequence of the variables is $Z \to A \to M \to Y$.

Because of the temporal relationship, we can also define potential outcomes for Y with respect to Z, A and M, and the ones for M with respect to Z and A. For $z, a, m \in \{0, 1\}$, we assume $Y(z, a, m), Y(a, m)$ are the potential outcomes of Y under respectively $(Z, A, M) = (z, a, m)$ and $(A, M) = (a, m)$, and $M(z, a)$ and $M(a)$ are the potential outcomes of M under respectively $(Z, A) = (z, a)$ and $A = a$. The following three assumptions are considered throughout this chapter.

Assumption 5.2.1 (Exclusion restriction) $\forall(z,a,m) \in \{0,1\}^3$, *we have* $Y(z,a,m) = Y(a,m)$ *and* $M(z,a) = M(a)$.

Assumption 5.2.2 (Relevance) $\mathrm{E}[A(1) - A(0)] \neq 0$.

Assumption 5.2.3 (Randomized assignment)

$$Z \perp\!\!\!\perp \{Y(1,1), Y(1,0), Y(0,1), Y(0,0), M(1), M(0), A(1), A(0)\}.$$

The above three assumptions, together with the Assumption 5.2.4 in Section 5.2.4, ensure that Z is an instrumental variable (IV) in the sense of Angrist et al. (1996) for the causal effects of A on M and Y. In the sense of Chapter 7 in Pearl (2009) under the structural causal models, however, the Assumption 5.2.4 seems not essential to define an instrument. To make it clear, we refer to the model under Assumptions 5.2.1 - 5.2.3 as the IV setting/setup with noncompliance (or imperfect compliance), and the model under Assumptions 5.2.1 - 5.2.4 as the IV setting/setup without defiers. Assumption 5.2.2 says that the assignment Z has an overall impact on the treatment A, which is very weak and often desirable in practice. Assumption 5.2.3 is a naturally satisfied in RCTs due to the design. The most controversial part is Assumption 5.2.1, which states that Z does not have a direct causal path to M and Y; The causal path from Z to M must be through A, and the one from Z to Y must be through A or M. This assumption is untestable and assessing it requires background knowledge and arguments based on the real situations. It sounds plausible if we believe the assignment status does not affect the mediator and the outcome given the actual treatment status. Angrist et al. (1996) includes comprehensive discussions on this assumption. The Assumptions 5.2.1 - 5.2.3 are the basis of our setting for RCTs with noncompliance and admit a wide range of causal diagrams for the data, with three

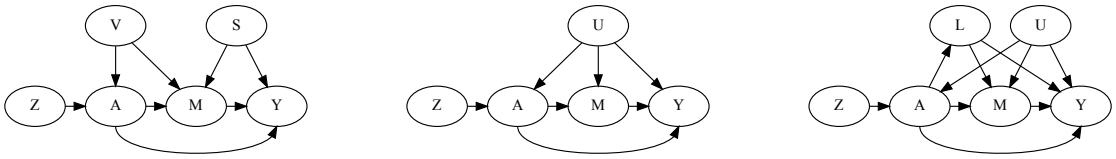


Figure 5.1: Feasible causal diagrams of the data generating mechanism under Assumptions 5.2.1 - 5.2.3.

instances given in Figure 5.1. We allow for the presence of unmeasured treatment-mediator confounding, treatment-outcome confounding, mediator-outcome confounding, as well as treatment-mediator-outcome confounding. Also, our model allows for the existence of unmeasured post-treatment mediator-outcome confounding, corresponding to the rightmost plot of Figure 5.1. The key feature of these causal diagrams is that there are no (even observed) confounding factors between Z and (A, M, Y) due to the randomization of Z, and Z has no direct causal paths to M and Y because of the exclusion restriction assumption.

5.2.2 Parameters of Interest

Two key components in CMA are $\delta(a) = Y(a, M(1)) - Y(a, M(0))$ and $\xi(a) = Y(1, M(a)) - Y(0, M(a))$. The former component quantifies the effect of the treatment A on the outcome Y through the mediator M by hypothetically fixing the treatment to $A = a$ but reversing the mediator from $M(0)$ to $M(1)$. The latter quantifies the direct treatment effect had the mediator been set to what it would have been under $A = a$. We focus on two population quantities: the average causal mediation effect (ACME; Imai et al., 2010), also called the natural indirect effect (Pearl, 2001), $\bar{\delta}(a) = \mathrm{E}[\delta(a)]$, and the natural direct effect (NDE), given by $\bar{\xi}(a) = \mathrm{E}[\xi(a)]$. The average treatment effect, $\tau = \mathrm{E}[Y(1, M(1)) - Y(0, M(0))]$ can be decomposed as $\tau = \bar{\delta}(0) + \bar{\xi}(1) = \bar{\delta}(1) + \bar{\xi}(0)$.

5.2.3 Identification of Causal Parameters

As we have shown in Section 2.3.3, the causal parameters $\bar{\delta}(a)$ and $\bar{\xi}(a)$ are identifiable under the sequential ignorability assumption introduced by Imai et al. (2010), which implicitly requires the treatment-mediator, treatment-outcome, and mediator-outcome confounding factors to be fully observed. This assumption is untestable and obviously very strong. It also excludes some causal diagrams, e.g., the third causal diagram in Figure 5.1. With the help of instrumental variables, however, it is possible to identify a set of local parameters for the treatment-assignment compliers under a weaker assumption that only requires fully observed mediator-outcome confounding (Rudolph et al., 2021). We will discuss this further in Section 5.4. Rather than relying on strong identification assumptions such as the sequential ignorability, we address the issue of partial identification and, in Section 5.3, establish nonparametric bounds for $\bar{\delta}(a)$ and $\bar{\xi}(a)$ under the IV setup with noncompliance.

5.2.4 Additional Monotonicity Assumptions

We consider a few additional monotonicity assumptions that are not essential but are plausible and yield tighter bounds on the parameters of interest in certain circumstances.

Assumption 5.2.4 (Treatment-assignment monotonicity) $A(1) \geq A(0)$ *almost surely.*

Assumption 5.2.5 (Mediator-treatment monotonicity) $M(1) \geq M(0)$ *almost surely.*

Assumption 5.2.6 (Outcome-mediator monotonicity) $Y(a,1) \geq Y(a,0)$ *almost surely for* $a = 0, 1$.

Assumption 5.2.7 (Outcome-treatment monotonicity) $Y(1,m) \geq Y(0,m)$ *almost surely for* $m = 0, 1$.

The monotonicity assumptions imply the absence of defiers. Although these assumptions are defined in terms of the joint distribution of potential outcomes and are therefore not empirically testable, they can sometimes be justified by design. For example, Assumption 5.2.4 can be validated by ensuring that units assigned to the control group are effectively prevented from receiving the treatment. In our real data application in Section 5.5, individuals assigned to the control group are deliberately blocked from accessing the treatment. In such a scenario, Assumptions 5.2.4 is desirable. It is worth noting that the tightest bounds under the additional monotonicity assumptions are not necessarily tighter than the tightest bounds under Assumptions 5.2.1 - 5.2.3 unless the imposed monotonicity assumptions are true.

5.3 Bounds on Population Parameters

We focus in this section on nonparametric bounds of $\bar{\delta}(a)$ and $\bar{\xi}(a)$ under Assumptions 5.2.1 - 5.2.3 for conciseness of the main body. The readers are referred to Section 5.7 for the bounds on the population causal parameters under additional monotonicity assumptions. Let $p_{yma.z} = \Pr(Y = y, M = m, A = a \mid Z = z)$, the conditional probabilities for observed data. The bounds of $\bar{\delta}(a)$ and $\bar{\xi}(a)$ under Assumptions 5.2.1 - 5.2.3 are established as follows.

$$\bar{\delta}(1) \geq \max \left\{ \begin{array}{c} -p_{000.1} - p_{001.0} - p_{010.0} - p_{011.1} - p_{100.1} - p_{110.0} \\ -p_{000.0} - p_{001.1} - p_{010.1} - p_{011.0} - p_{100.0} - p_{110.1} \\ p_{101.1} + p_{111.1} - 1 \end{array} \right\},$$

$$\bar{\delta}(1) \leq \min \left\{ \begin{array}{c} p_{000.0} + p_{010.1} + p_{100.0} + p_{101.1} + p_{110.1} + p_{111.0} \\ p_{000.1} + p_{010.0} + p_{100.1} + p_{101.0} + p_{110.0} + p_{111.1} \\ -p_{001.1} - p_{011.1} + 1 \end{array} \right\},$$

$$\bar{\delta}(0) \geq \max \left\{ \begin{array}{c} -p_{001.0} - p_{011.1} - p_{100.1} - p_{101.0} - p_{110.0} + p_{111.0} - p_{111.1} \\ -p_{001.1} - p_{011.0} - p_{100.0} - p_{101.1} - p_{110.1} \\ p_{000.0} + p_{010.0} + p_{111.0} - 1 \end{array} \right\},$$

$$\bar{\delta}(0) \leq \min \left\{ \begin{array}{c} p_{000.1} + p_{001.0} + p_{010.0} - p_{011.0} + p_{011.1} + p_{101.0} + p_{111.1} \\ p_{000.0} + p_{001.1} + p_{010.1} + p_{101.1} + p_{111.0} \\ -p_{011.0} - p_{100.0} - p_{110.0} + 1 \end{array} \right\}.$$

$$\bar{\xi}(1) \geq \max \left\{ \begin{array}{c} -p_{000.0} - p_{001.1} - p_{010.1} - p_{100.0} - p_{110.0} - p_{111.0} + p_{111.1} \\ p_{000.0} + p_{111.0} + 2p_{011.0} + p_{101.1} - p_{000.1} - p_{011.1} - 1 \\ p_{011.0} + p_{101.1} + p_{111.1} - 1 \end{array} \right\},$$

$$\bar{\xi}(1) \leq \min \left\{ \begin{array}{c} p_{000.0} + p_{100.0} + p_{010.0} + p_{011.0} - p_{011.1} + p_{101.1} + p_{110.1} \\ -p_{001.1} - p_{011.0} - p_{100.0} + p_{100.1} - 2p_{111.0} + p_{111.1} + 1 \\ -p_{001.1} - p_{011.1} - p_{111.0} + 1 \end{array} \right\},$$

$$\bar{\xi}(0) \geq \max \left\{ \begin{array}{c} p_{000.1} + p_{100.1} + p_{111.1} + p_{010.0} - p_{100.0} - p_{111.0} - 1 \\ p_{000.0} + p_{010.1} - p_{101.0} + p_{101.1} - p_{110.0} + p_{110.1} - 1 \\ p_{000.0} + p_{010.0} - 1 \end{array} \right\},$$

$$\bar{\xi}(0) \leq \min \left\{ \begin{array}{c} p_{000.0} - p_{000.1} + p_{011.0} - p_{011.1} - p_{100.1} - p_{110.0} + 1 \\ p_{001.0} - p_{001.1} + p_{010.0} - p_{010.1} - p_{100.0} - p_{110.1} + 1 \\ -p_{100.0} - p_{110.0} + 1 \end{array} \right\}.$$

Proofs of these bounds are presented in Section 5.7.1. The provided lower/upper bounds are in a general form of taking the maximum/minimum value from a set of observed probability entries. Section 5.7.1 shows that these symbolic entries correspond to the vertices of (symbolic) polyhedral.

The above lower and upper bounds of $\bar{\delta}(a)$ and $\bar{\xi}(a)$ are not sharp. The tightest bounds of $\bar{\delta}(a)$ and $\bar{\xi}(a)$ under Assumptions 5.2.1 - 5.2.3 involve over 30 entries and we will not show those lengthy expressions in this thesis. The above bounds are in fact the tightest bounds of $\bar{\delta}(a)$ and $\bar{\xi}(a)$ under Assumptions 5.2.1 - 5.2.4. This implies that for bounding $\bar{\delta}(a)$ and $\bar{\xi}(a)$ no information is gained by assuming treatment-assignment monotonicity. This is because among the the entries for the bounds under Assumptions 5.2.1 - 5.2.3, only those for the bounds under Assumptions 5.2.1 - 5.2.4 are active when the treatment-assignment monotonicity is true. The similar phenomenon has been observed when bounding the average treatment effect in the IV setting with noncompliance (Balke and Pearl, 1997; Gabriel et al., 2023). We also note that the upper bounds of $\bar{\delta}(a)$ under Assumptions 5.2.1 - 5.2.5, 5.2.7 are same as the ones under Assumptions 5.2.1 - 5.2.7.

5.4 Bounds on Local Parameters

The IV setting without defiers specified by Assumptions 5.2.1 - 5.2.4 allows us to identify the local average treatment effect (Angrist et al., 1996), $\tau^l = \mathrm{E}[Y(1, M(1)) - Y(0, M(0)) \mid A(1) > A(0)]$, by

$$\tau^l = \frac{\mathrm{E}[Y(1, A(1), M(A(1))) - Y(0, A(0), M(A(0)))]}{\mathrm{E}[A(1) - A(0)]} = \frac{\Pr(Y = 1|Z = 1) - \Pr(Y = 1|Z = 0)}{\Pr(A = 1|Z = 1) - \Pr(A = 1|Z = 0)}.$$

This is the average treatment effect for the subpopulation of treatment-assignment compliers. We analogously define the local average causal mediation effect as $\bar{\delta}^l(a) = \mathrm{E}[\delta(a) \mid A(1) > A(0)]$ and the local natural indirect effect as $\bar{\xi}^l(a) = \mathrm{E}[\xi(a) \mid A(1) > A(0)]$. These local causal parameters defined over a specific subpopulation might be the focus in real-world problems. Especially when the compliers occupy only a very small fraction of the entire population, even the tightest bounds on the population parameters can be noninformative and useless because they need to consider the best/worst situations for the noncompliance which forms the majority of the population.

Identification of these local parameters under Assumptions 5.2.1 - 5.2.4 is free from treatment-mediator and treatment-outcome confounding factors, despite it still requires fully observed mediator-outcome confounding. In this section, we derive the bounds of $\bar{\delta}^l(a)$ under the IV setting without defiers which allows for the presence of unmeasured mediator-outcome confounding. The bounds for $\bar{\xi}^l(a)$ can be easily obtained from the bounds of $\bar{\delta}^l(1-a)$ by noting that $\tau^l = \bar{\delta}^l(1-a) + \bar{\xi}^l(a)$. Let $\bar{\delta}^*(a) = \mathrm{E}[\delta(a)A(1)(1-A(0))]$. Under Assumptions 5.2.1 - 5.2.4, we can show that $\bar{\delta}^l(a) = \bar{\delta}^*(a)/\Pr(A(1) = 1, A(0) = 0)$ where $\Pr(A(1) = 0, A(0) = 1) = \Pr(A = 1|Z = 1) - \Pr(A = 1|Z = 0)$ is the proportion of compliers, which is identifiable from the observed data. It is therefore equivalent to study the bounds of $\bar{\delta}^*(a)$. In the following, we give the bounds of $\bar{\delta}^*(a)$ under Assumptions

5.2.1 - 5.2.4. The bounds of $\bar{\delta}^*(a)$ under additional monotonicity assumptions are shown in Section 5.7.

Under Assumptions 5.2.1 - 5.2.4, the sharp bounds of $\bar{\delta}^*(a)$ are given by

$$\bar{\delta}^*(1) \geq \max \left\{ \begin{array}{c} -p_{010.0} + p_{010.1} + p_{011.0} - p_{011.1} - p_{110.0} + p_{110.1} \\ -p_{000.0} + p_{000.1} + p_{001.0} - p_{001.1} - p_{100.0} + p_{100.1} \\ p_{001.0} - p_{001.1} + p_{011.0} - p_{011.1} \end{array} \right\},$$

$$\bar{\delta}^*(1) \leq \min \left\{ \begin{array}{c} p_{000.0} - p_{000.1} + p_{100.0} - p_{100.1} - p_{101.0} + p_{101.1} \\ p_{010.0} - p_{010.1} + p_{110.0} - p_{110.1} - p_{111.0} + p_{111.1} \\ -p_{101.0} + p_{101.1} - p_{111.0} + p_{111.1} \end{array} \right\},$$

$$\bar{\delta}^*(0) \geq \max \left\{ \begin{array}{c} p_{011.0} - p_{011.1} - p_{110.0} + p_{110.1} + p_{111.0} - p_{111.1} \\ p_{001.0} - p_{001.1} - p_{100.0} + p_{100.1} + p_{101.0} - p_{101.1} \\ -p_{100.0} + p_{100.1} - p_{110.0} + p_{110.1} \end{array} \right\},$$

$$\bar{\delta}^*(0) \leq \min \left\{ \begin{array}{c} p_{000.0} - p_{000.1} - p_{001.0} + p_{001.1} - p_{101.0} + p_{101.1} \\ p_{010.0} - p_{010.1} - p_{011.0} + p_{011.1} - p_{111.0} + p_{111.1} \\ p_{000.0} - p_{000.1} + p_{010.0} - p_{010.1} \end{array} \right\}.$$

We note that the probability entries in the above bounds of the local causal parameters

show up in pairs of $p_{yma.0} - p_{yma.1}$. We can verify that

$$\frac{(1-2a)(p_{yma.0} - p_{yma.1})}{\Pr(A(1)=1, A(0)=0)} = p^c_{yma.a}$$

where $p^c_{yma.a} = \Pr(Y=y, M=m, A=a \mid Z=a, A(1)=1, A(0)=0)$ are the observed conditional probabilities over the subpopulation of treatment-assignment compliers. Therefore, by replacing $(p_{yma.0} - p_{yma.1})/\Pr(A(1)=1, A(0)=0)$ with $(1-2a)p^c_{yma.a}$, one can obtain reduced expressions of the bounds for the local causal parameters.

5.5 Analysis of the Criteo Uplift Dataset

The Criteo uplift data are large-scale online A/B testing data collected by the online advertising company Criteo, with the details of the data discussed in Diemert et al. (2018) and Diemert et al. (2021). The dataset is originally released for uplift modeling to identify the individuals most likely to be converted by online advertising (conversion means to make a purchase). The data are used for inllustration of the proposed nonparametric bounds on the causal mediation effects.

The dataset contains over 13M observations, with 85% of them randomly assigned to the treatment group. The treatment indicates whether a company participates in the real-time bidding (RTB) auction for a user, while the user is not effectively exposed unless the company wins in the RTB auction for the user. For each observation in the dataset, four binary indicators are recorded: (1) Whether the user was assigned to the treatment group (Z); (2) whether the user was effectively exposed (A); (3) whether a visit occurred to the user (M); and (4) whether a conversion occurred to the user (Y). See Table 5.1 for the specific meanings of the four variables. The setup is a RCT with noncompliance since the

Table 5.1: Interpretation of the variables for the Criteo uplift data

Z	whether a company participates in RTB auction for a user
A	whether the company wins RTB auction for the user
M	whether the user visits the website advertised
Y	whether the user achieves conversion (purchase)

effect of A on Y is assumed to be of primary concern. The Z serves as an IV if we assume the exclusion-restriction assumption which is indeed plausible in this problem. To figure out the causal mechanisms of effective exposition on the conversion rate and the role of the website-visit behavior in this procedure, we focus on evaluating the ACME and the NDE.

We can reasonably make Assumption 5.2.4 because at the design stage, the individuals in the control group ($Z = 0$) were intentionally blocked from being exposed ($A = 1$), i.e., $\Pr(A = 1 \mid Z = 0) = 0$. This is actually a stronger result compared to Assumption 5.2.4. It is also warranted by design that no conversion can be achieved without a visit, i.e., $Y(a, 0) = 0$ almost surely for $a = 0, 1$. However, this structural restriction would lead to a zero NDE, thereby rendering the mediation analysis meaningless. To remedy this issue, we hypothetically allow for the possibility that users may achieve conversion through alternative means, such as visiting offline (brick-and-mortar) stores, and emphasize that the data are used here solely for illustrative purposes in evaluating the proposed method.

The bounds on the population causal parameters $\bar{\delta}(a)$ and $\bar{\xi}(a)$ for the data under different assumptions are given in Table 5.2. Because Assumption 5.2.4 is true in this example, we see that the bounds for all the causal parameters under Assumptions 5.2.1 - 5.2.3 are the same as the ones under Assumptions 5.2.1 - 5.2.4, verifying our arguments in Section 5.3. The intervals formed by the bounds are very wide because the noncompliance rate is greater than $\Pr(A = 0 \mid Z = 1) \approx 96.4\%$, yielding a great amount of uncertainty. However, we are still able to acquire information from the bounds. For example, under

Table 5.2: Bounds on the population parameters under different assumptions.

	Assumptions				
	5.2.1 - 5.2.3	5.2.1 - 5.2.4	5.2.1 - 5.2.4, 5.2.6	5.2.1 - 5.2.6	5.2.1 - 5.2.7
$\bar{\delta}(1)$	$[-0.982, 0.966]$	$[-0.982, 0.966]$	$[-0.038, 0.932]$	$[0, 0.932]$	$[0, 0.932]$
$\bar{\delta}(0)$	$[-0.002, 0.982]$	$[-0.002, 0.982]$	$[-0.002, 0.945]$	$[0, 0.941]$	$[0, 0.932]$
$\bar{\xi}(1)$	$[-0.982, 0.966]$	$[-0.982, 0.966]$	$[-0.945, 0.966]$	$[-0.941, 0.964]$	$[0, 0.964]$
$\bar{\xi}(0)$	$[-0.002, 0.982]$	$[-0.002, 0.982]$	$[-0.002, 0.969]$	$[-0.002, 0.964]$	$[0, 0.964]$

Assumptions 5.2.1 - 5.2.4, 5.2.6, the $\bar{\delta}(0)$ cannot be smaller than -0.002.

Facing such a huge proportion of noncompliance, it might be better to shift our attention from the entire population to the subpopulation of treatment-assignment compliers. The bounds on the local causal parameters are shown in Table 5.3, which are more informative compared to the ones on the population causal parameters. Under Assumptions 5.2.1 - 5.2.4, the bounds $[-0.488, 0.054]$ on $\bar{\delta}^l(1)$ imply a possible strong negative mediation effect of advertising through website-visit for the compliers effectively exposed. One can compute the average treatment effect of A on M over the compliers as

$$\tau_m = \frac{\Pr(M = 1 \mid Z = 1) - \Pr(M = 1 \mid Z = 0)}{\Pr(A = 1 \mid Z = 1) - \Pr(A = 1 \mid Z = 0)} = 0.287,$$

indicating that if $\bar{\delta}^l(1)$ is negative, then visiting the website has a reversed effect on conversion for the compliers effectively exposed. If $\bar{\delta}^l(1)$ is truly negative, the practitioners may consider improving the quality of the website or encouraging the effectively exposed users to go to offline stores for purchase. The bounds of $\bar{\delta}^l(a)$ and $\bar{\xi}^l(1-a)$ are complementary to each other, i.e., crosswise addition of them always leads to the value 0.032 which is exactly the IV estimate of the local average treatment effect because $\tau^l = \bar{\delta}^l(a) + \bar{\xi}^l(1-a)$.

Table 5.3: Bounds on the local parameters under different assumptions.

	Assumptions			
	5.2.1 - 5.2.4	5.2.1 - 5.2.4, 5.2.6	5.2.1 - 5.2.6	5.2.1 - 5.2.7
$\bar{\delta}^l(1)$	$[-0.488, 0.054]$	$[-0.128, 0.054]$	$[0.0, 0.054]$	$[0.0, 0.032]$
$\bar{\delta}^l(0)$	$[-0.022, 0.520]$	$[-0.022, 0.415]$	$[0.0, 0.287]$	$[0.0, 0.032]$
$\bar{\xi}^l(1)$	$[-0.488, 0.054]$	$[-0.383, 0.054]$	$[-0.255, 0.032]$	$[0.0, 0.032]$
$\bar{\xi}^l(0)$	$[-0.022, 0.520]$	$[-0.022, 0.160]$	$[-0.022, 0.032]$	$[0.0, 0.032]$

5.6 Concluding Remarks

We derived the nonparametric bounds of the ACME and the NDE in the IV setting with imperfect compliance and extended our results to the local causal parameters in the IV setting without defiers. The bounds provide partial information on the parameters of interest under minimal assumptions on the data. However, the readers should be aware of that the derived bounds can barely be used for inferences of whether the causal effects are positive or negative, as in most cases the intervals formed by the bounds include zero. They are expected to complement the information offered by point estimates (Robins and Greenland, 1996; Richardson et al., 2014).

5.7 Proofs and Technical Details

5.7.1 Derivation of the Bounds

The bounds in the main body as well as the ones we will present in Sections 5.7.2 and 5.7.3 are derived based on the method outlined in Balke and Pearl (1994) and Balke (1996). We briefly review the method in this section. Let $O_y = (Y(1,1), Y(1,0), Y(0,1), Y(0,0))$,

Table 5.4: The values of o_{yi} for $i = 0, \cdots, 15$.

i	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0
	1	1	1	1	0	0	0	0	1	1	1	1	0	0	0	0
o_{yi}	1	1	0	0	1	1	0	0	1	1	0	0	1	1	0	0
	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0

$O_m = (M(1), M(0))$, and $O_a = (A(1), A(0))$. The values of O_y and O_m are encoded as given respectively in Table 5.4 and Table 5.5. Define $q_{ijk} = \Pr(O_y = o_{yi}, O_m = o_{mj}, O_a = o_{mk})$ for $i = 0, \cdots, 15$, $j, k = 0, 1, 2, 3$ (See Tables 5.4 and 5.5 for the definitions of o_{yi} and o_{mj}). Under Assumptions 5.2.1 and 5.2.2, the parameters of interest are linear in q_{ijk}. For example, simple algebra shows that

$$
\begin{aligned}
\bar{\delta}(1) &= \mathrm{E}[Y(1, M(1)) - Y(1, M(0))] \\
&= \Pr(Y(1,1) = 1, M(1) = 1) + \Pr(Y(1,0) = 1, M(1) = 0) \\
&\quad - \Pr(Y(1,1) = 1, M(0) = 1) - \Pr(Y(1,0) = 1, M(0) = 0) \\
&= \sum_{i=4,5,6,7} \sum_{j=1} \sum_{k=0,1,2,3} q_{ijk} + \sum_{i=8,9,10,11} \sum_{j=2} \sum_{k=0,1,2,3} q_{ijk} \\
&\quad - \sum_{i=4,5,6,7} \sum_{j=2} \sum_{k=0,1,2,3} q_{ijk} - \sum_{i=8,9,10,11} \sum_{j=1} \sum_{k=0,1,2,3} q_{ijk}.
\end{aligned}
$$

Because $\sum_{ijk} q_{ijk} = 1$, the degree of freedom for q_{ijk} is 255. The density of the observed data (Y, M, A, Z) is given by $\Pr(Y = y, M = m, A = a, Z = z) = p(y, m, a, z) =$

Table 5.5: The values of o_{mi} for $i = 0, 1, 2, 3$.

i	0	1	2	3
	1	1	0	0
o_{mi}	1	0	1	0

$p(y, m, a|z)p(z)$ which is fully characterized by $p(y, m, a|z) = \Pr(Y = y, M = m, A = a \mid Z = z)$ and $p(z) = \Pr(Z = z)$ for $y, m, a, z \in \{0, 1\}$. Denote $p_{yma.z} = p(y, m, a|z)$ as in the main body. Under Assumption 5.2.3, we can show that $p_{yma.z} = \Pr(Y(a, m) = y, M(a) = m, A(z) = a)$ which impose equality constraints on and further reduce the degree of freedom of q_{ijk}. For instance, one can easily show that

$$p_{111.1} = \sum_{i=0,1,\cdots,7} \sum_{j=0,1} \sum_{k=0,1} q_{ijk}.$$

Note that $p(z)$ do not impose any constraints on q_{ijk}. Under Assumptions 5.2.1 - 5.2.3, there will be $2^4 + 1 = 17$ equality constraints for q_{ijk}. The monotonicity assumptions introduce more equality constraints on q_{ijk}. For example, Assumption 5.2.4 which states that $A(1) \geq A(0)$ enforces $q_{ijk} = 0$ when $k = 2$; Assumption 5.2.6 which states that $Y(a, 1) \geq Y(a, 0)$ enforces $q_{ijk} = 0$ when $i = 2, 6, 8, 9, 10, 11, 14$. Let $q = (q_{ijk})$ be a 256-dimensional vector of q_{ijk}. Computing, e.g., the lower bound of $\bar{\delta}(1)$ under Assumptions 5.2.1 - 5.2.3 and possibly others corresponds to solving the symbolic linear programming problem

$$\begin{aligned} &\min_q c^\top q \\ \text{s.t. } &Aq = b, \\ &q \geq 0, \end{aligned} \tag{5.1}$$

where $\bar{\delta}(1) = c^\top q$ and $Aq = b$ are the equality constraints for q_{ijk}. To solve (5.1), we consider its dual problem

$$\begin{aligned} &\max_w b^\top w \\ \text{s.t. } &A^\top w \leq c. \end{aligned} \tag{5.2}$$

The duality gap is zero because the constraints in (5.1) are all affine (linear). The $P = \{w \mid$

$A^\top w \leq c\}$ is a polyhedron. The maximum value of (5.2) is attained at either the vertices or the rays of the polyhedron P. Due to the constraints $q \geq 0$ and $1^\top q = 1$, the solution to (5.1) must be bounded and thus the problem (5.2) must also be bounded. Because of the boundedness, there is no need to consider the rays of P and the solution to (5.2) will be

$$\max_{w \in V} b^\top w \tag{5.3}$$

where V is a collection of the vertices of the polyhedron P. To solve (5.2), therefore, we can enumerate all the vertices of the polyhedron P and then substitute into (5.3). For the extreme vertex enumeration algorithms, we refer readers to Mattheiss (1973) and Matheiss and Rubin (1980).

In the presence of covariates $\boldsymbol{X}$, we let $q_{ijk} = \Pr(O_y = o_{yi}, O_m = o_{mj}, O_a = o_{mk} \mid \boldsymbol{X})$ and $p_{yma.z} = p(y, m, a|z, \boldsymbol{x})$. The density of the observed data $(Y, M, A, Z, \boldsymbol{X})$ is $p(y, m, a, z, \boldsymbol{x}) = p(y, m, a|z, \boldsymbol{x})p(z)p(\boldsymbol{x})$ where $p_{yma.z}$ impose the same constraints on q_{ijk}, while $p(z)$ and $p(\boldsymbol{x})$ do not impose any constraints. Then, we solve the same symbolic linear programming problem. Since the causal parameters in this scenario are in the form of $\mathrm{E}(c^\top q)$, the results are followed by noting that $\mathrm{E}(\min_q c^\top q) \leq \min_q \mathrm{E}(c^\top q) \leq \max_q \mathrm{E}(c^\top q) \leq \mathrm{E}(\max_q c^\top q)$ due to Jensen's inequality.

5.7.2 Bounds on Population Parameters under Monotonicity

In Section 5.3, we have presented the bounds on $\bar{\delta}(a)$ and $\bar{\xi}(a)$ under Assumptions 5.2.1 - 5.2.3 which are also the sharp bounds of $\bar{\delta}(a)$ and $\bar{\xi}(a)$ under Assumptions 5.2.1 - 5.2.4. With additional monotonicity assumptions given in Section 5.2.4, these bounds can be tightened (if the imposed assumptions are true). The results are given in the following two

subsections, with the first subsection presenting the sharp bounds for $\bar{\delta}(a)$ and the second presenting the sharp bounds for $\bar{\xi}(a)$. Proofs of the above bounds have been outlined in Section 5.7.1, using the Pearl-Balke method.

Bounds on the ACME

The sharp bounds for $\bar{\delta}(a)$ under different assumptions are established as follows. Under Assumptions 5.2.1 - 5.2.5, we have

$$\bar{\delta}(1) \geq \max \left\{ \begin{array}{c} p_{001.1} + p_{010.0} + p_{101.1} + p_{110.0} + p_{111.0} - 1 \\ -p_{000.1} - p_{011.1} - p_{100.1} \end{array} \right\},$$

$$\bar{\delta}(1) \leq \min \left\{ \begin{array}{c} -p_{001.1} - p_{010.0} - p_{011.0} - p_{101.1} - p_{110.0} + 1 \\ p_{000.1} + p_{100.1} + p_{111.1} \end{array} \right\},$$

$$\bar{\delta}(0) \geq \max \left\{ \begin{array}{c} p_{000.1} + p_{001.1} + p_{010.0} + p_{101.1} + p_{110.0} + p_{111.0} - 1 \\ -p_{011.0} - p_{100.0} \end{array} \right\},$$

$$\bar{\delta}(0) \leq \min \left\{ \begin{array}{c} -p_{001.1} - p_{010.0} - p_{011.0} - p_{100.1} - p_{101.1} - p_{110.0} + 1 \\ p_{000.0} + p_{111.0} \end{array} \right\}.$$

Under Assumptions 5.2.1 - 5.2.4, 5.2.6, we have

$$\max \left\{ \begin{array}{c} -p_{001.1} - p_{010.1} - p_{110.1} \\ -p_{001.0} - p_{010.0} - p_{110.0} \end{array} \right\} \leq \bar{\delta}(1) \leq \min \left\{ \begin{array}{c} p_{000.0} + p_{100.0} + p_{111.0} \\ p_{000.1} + p_{100.1} + p_{111.1} \end{array} \right\},$$

$$\max \left\{ \begin{array}{c} -p_{001.1} - p_{101.1} - p_{110.1} \\ -p_{001.0} - p_{101.0} - p_{110.0} \end{array} \right\} \leq \bar{\delta}(0) \leq \min \left\{ \begin{array}{c} p_{000.0} + p_{111.0} \\ p_{000.1} - p_{011.0} + p_{011.1} + p_{111.1} \end{array} \right\}.$$

Under Assumptions 5.2.1 - 5.2.4, 5.2.7, we have

$$\bar{\delta}(1) \geq \max \left\{ \begin{array}{c} -p_{000.0} - p_{001.1} - p_{010.1} - p_{011.0} - p_{100.0} - p_{110.1} \\ -p_{000.1} - p_{001.0} - p_{010.0} - p_{011.1} - p_{100.1} - p_{110.0} \\ p_{101.1} + p_{111.1} - 1 \end{array} \right\},$$

$$\bar{\delta}(1) \leq \min \left\{ \begin{array}{c} p_{000.1} + p_{010.0} + p_{101.0} + p_{111.1} \\ p_{000.0} + p_{010.1} + p_{101.1} + p_{111.0} \\ -p_{001.1} - p_{011.1} - p_{100.0} - p_{110.0} + 1 \end{array} \right\},$$

$$\bar{\delta}(0) \geq \max \left\{ \begin{array}{c} -p_{001.0} - p_{011.1} - p_{100.1} - p_{101.0} - p_{110.0} + p_{111.0} - p_{111.1} \\ -p_{001.1} - p_{011.0} - p_{100.0} - p_{101.1} - p_{110.1} \\ p_{000.0} + p_{010.0} + p_{111.0} - 1 \end{array} \right\},$$

$$\bar{\delta}(0) \leq \min \left\{ \begin{array}{c} p_{000.0} + p_{010.1} + p_{101.1} + p_{111.0} \\ p_{000.1} + p_{010.0} + p_{101.0} + p_{111.1} \\ -p_{001.1} - p_{011.1} - p_{100.0} - p_{110.0} + 1 \end{array} \right\}.$$

Under Assumptions 5.2.1 - 5.2.6, we have

$$0 \leq \bar{\delta}(1) \leq \min \left\{ \begin{array}{c} -p_{001.1} - p_{010.0} - p_{011.0} - p_{101.1} - p_{110.0} + 1 \\ p_{000.1} + p_{100.1} + p_{111.1} \end{array} \right\},$$

$$0 \leq \bar{\delta}(0) \leq \min \left\{ \begin{array}{c} -p_{001.1} - p_{010.0} - p_{011.0} - p_{100.1} - p_{101.1} - p_{110.0} + 1 \\ p_{000.0} + p_{111.0} \end{array} \right\}.$$

Under Assumptions 5.2.1 - 5.2.5, 5.2.7, we have

$$\bar{\delta}(1) \geq \max \left\{ \begin{array}{c} p_{001.1} + p_{010.0} + p_{101.1} + p_{110.0} + p_{111.0} - 1 \\ -p_{000.1} - p_{011.1} - p_{100.1} \end{array} \right\},$$

$$\bar{\delta}(1) \leq \min \left\{ \begin{array}{c} -p_{001.1} - p_{010.0} - p_{011.0} - p_{100.1} - p_{101.1} - p_{110.0} + 1 \\ -p_{001.1} - p_{010.1} - p_{011.1} - p_{100.0} - p_{101.0} - p_{110.0} + 1 \\ p_{000.0} + p_{001.0} - p_{001.1} + p_{111.0} \\ p_{000.1} - p_{110.0} + p_{110.1} + p_{111.1} \end{array} \right\},$$

$$\bar{\delta}(0) \geq \max \left\{ \begin{array}{c} p_{000.1} + p_{001.1} + p_{010.0} + p_{101.1} + p_{110.0} + p_{111.0} - 1 \\ -p_{011.0} - p_{100.0} \end{array} \right\},$$

$$\bar{\delta}(0) \leq \min \left\{ \begin{array}{c} -p_{001.1} - p_{010.0} - p_{011.0} - p_{100.1} - p_{101.1} - p_{110.0} + 1 \\ -p_{001.1} - p_{010.1} - p_{011.1} - p_{100.0} - p_{101.0} - p_{110.0} + 1 \\ p_{000.0} + p_{001.0} - p_{001.1} + p_{111.0} \\ p_{000.1} - p_{110.0} + p_{110.1} + p_{111.1} \end{array} \right\}.$$

Under Assumptions 5.2.1 - 5.2.7, we have

$$0 \leq \bar{\delta}(1) \leq \min \left\{ \begin{array}{c} -p_{001.1} - p_{010.0} - p_{011.0} - p_{100.1} - p_{101.1} - p_{110.0} + 1 \\ -p_{001.1} - p_{010.1} - p_{011.1} - p_{100.0} - p_{101.0} - p_{110.0} + 1 \\ p_{000.1} - p_{110.0} + p_{110.1} + p_{111.1} \\ p_{000.0} + p_{001.0} - p_{001.1} + p_{111.0} \end{array} \right\},$$

$$0 \leq \bar{\delta}(0) \leq \min \left\{ \begin{array}{c} -p_{001.1} - p_{010.1} - p_{011.1} - p_{100.0} - p_{101.0} - p_{110.0} + 1 \\ -p_{001.1} - p_{010.0} - p_{011.0} - p_{100.1} - p_{101.1} - p_{110.0} + 1 \\ p_{000.0} + p_{001.0} - p_{001.1} + p_{111.0} \\ p_{000.1} - p_{110.0} + p_{110.1} + p_{111.1} \end{array} \right\}.$$

As we impose more assumptions, the derived bounds are expected to be tighter because the parameter space is reduced. Even if not guaranteed by design, in many practical cases the monotonicity assumptions are considered mild and plausible. If the imposed assumptions are false, however, the derived bounds will be invalid and not necessarily tighter, while they can still be useful. The procedure of imposing more and more assumptions is similar to the procedure of changing sensitivity parameters in sensitivity analysis.

Bounds on the NDE

The sharp bounds for $\bar{\xi}(a)$ under different assumptions are established as follows. Under Assumptions 5.2.1 - 5.2.5, we have

$$\bar{\xi}(1) \geq \max \left\{ \begin{array}{c} -p_{001.0} + p_{001.1} + p_{010.0} + p_{011.0} - p_{100.0} + p_{100.1} - p_{101.0} + 2p_{101.1} + p_{111.1} - 1 \\ p_{010.0} + p_{011.0} + p_{101.1} + p_{111.1} - 1 \end{array} \right\},$$

$$\bar{\xi}(1) \leq \min \left\{ \begin{array}{c} p_{000.0} - p_{000.1} + p_{001.0} - 2p_{001.1} - p_{011.1} + p_{101.0} - p_{101.1} - p_{110.0} - p_{111.0} + 1 \\ -p_{001.1} - p_{011.1} - p_{110.0} - p_{111.0} + 1 \end{array} \right\},$$

$$\bar{\xi}(0) \geq \max \left\{ \begin{array}{c} p_{000.0} + 2p_{010.0} - p_{010.1} + p_{011.0} - p_{011.1} + p_{101.1} + p_{110.0} - p_{110.1} - 1 \\ p_{000.0} + p_{010.0} + p_{101.1} - 1 \end{array} \right\},$$

$$\bar{\xi}(0) \leq \min \left\{ \begin{array}{c} -p_{001.1} - p_{010.0} + p_{010.1} - p_{100.0} - 2p_{110.0} + p_{110.1} - p_{111.0} + p_{111.1} + 1 \\ -p_{001.1} - p_{100.0} - p_{110.0} + 1 \end{array} \right\}.$$

Under Assumptions 5.2.1 - 5.2.4, 5.2.6, we have

$$\bar{\xi}(1) \geq \max \left\{ \begin{array}{c} -p_{000.1} - p_{001.0} + p_{011.0} - p_{011.1} - p_{100.0} - p_{101.0} + p_{101.1} - p_{110.0} \\ p_{010.0} + p_{011.0} + p_{101.1} + p_{111.1} - 1 \end{array} \right\},$$

$$\bar{\xi}(1) \leq \min \left\{ \begin{array}{c} p_{000.0} + p_{010.0} + p_{011.0} - p_{011.1} + p_{101.1} + p_{110.1} \\ -p_{001.1} - p_{011.1} - p_{100.0} - p_{111.0} + 1 \end{array} \right\},$$

$$\bar{\xi}(0) \geq \max \left\{ \begin{array}{c} -p_{001.1} + p_{010.0} - p_{010.1} + p_{011.0} - p_{011.1} - p_{100.0} - p_{110.1} - p_{111.0} \\ p_{000.0} + p_{010.0} + p_{011.0} + p_{101.1} - 1 \end{array} \right\},$$

$$\bar{\xi}(0) \leq \min \left\{ \begin{array}{c} p_{000.1} + p_{001.0} + p_{010.0} - p_{100.0} + p_{100.1} + p_{101.1} + p_{111.1} \\ -p_{011.1} - p_{100.0} - p_{110.0} + 1 \end{array} \right\}.$$

Under Assumptions 5.2.1 - 5.2.4, 5.2.7, we have

$$\bar{\xi}(1) \geq \max \left\{ \begin{array}{c} -p_{000.0} + p_{000.1} - p_{100.0} + p_{100.1} - p_{110.0} + p_{110.1} - p_{111.0} + p_{111.1} \\ -p_{010.0} + p_{010.1} - p_{100.0} + p_{100.1} - p_{101.0} + p_{101.1} - p_{110.0} + p_{110.1} \\ 0 \end{array} \right\},$$

$$\bar{\xi}(1) \leq \min \left\{ \begin{array}{c} -p_{001.0} - p_{011.1} - p_{101.0} + p_{101.1} - p_{110.0} + p_{110.1} - p_{111.0} + 1 \\ -p_{001.1} - p_{011.0} - p_{100.0} + p_{100.1} - 2p_{111.0} + p_{111.1} + 1 \\ -p_{001.1} - p_{011.1} - p_{111.0} + 1 \end{array} \right\},$$

$$\bar{\xi}(0) \geq \max \left\{ \begin{array}{c} -p_{010.0} + p_{010.1} - p_{100.0} + p_{100.1} - p_{101.0} + p_{101.1} - p_{110.0} + p_{110.1} \\ p_{001.0} - p_{001.1} + p_{010.0} - p_{010.1} + p_{011.0} - p_{011.1} + p_{101.0} - p_{101.1} \\ 0 \end{array} \right\},$$

$$\bar{\xi}(0) \leq \min \left\{ \begin{array}{c} p_{001.0} - p_{001.1} + p_{010.0} - p_{010.1} - p_{100.0} - p_{110.1} + 1 \\ p_{000.0} - p_{000.1} + p_{011.0} - p_{011.1} - p_{100.1} - p_{110.0} + 1 \\ -p_{100.0} - p_{110.0} + 1 \end{array} \right\}.$$

Under Assumptions 5.2.1 - 5.2.6, we have

$$\bar{\xi}(1) \geq \max \left\{ \begin{array}{c} -p_{001.0} + p_{001.1} + p_{010.0} + p_{011.0} - p_{100.0} + p_{100.1} - p_{101.0} + 2p_{101.1} + p_{111.1} - 1 \\ p_{010.0} + p_{011.0} + p_{101.1} + p_{111.1} - 1 \end{array} \right\},$$

$$\bar{\xi}(1) \leq -p_{001.1} - p_{011.1} - p_{100.0} - p_{110.0} - p_{111.0} + 1,$$

$$\bar{\xi}(0) \geq \max \left\{ \begin{array}{c} p_{000.0} + 2p_{010.0} - p_{010.1} + 2p_{011.0} - p_{011.1} + p_{101.1} + p_{110.0} - p_{110.1} - 1 \\ p_{000.0} + p_{010.0} + p_{011.0} + p_{101.1} - 1 \end{array} \right\},$$

$$\bar{\xi}(0) \leq -p_{001.1} - p_{011.1} - p_{100.0} - p_{110.0} + 1.$$

Under Assumptions 5.2.1 - 5.2.5, 5.2.7, we have

$$\bar{\xi}(1) \geq \max \left\{ \begin{array}{c} p_{010.0} - p_{010.1} + p_{011.0} - p_{011.1} - p_{100.0} + p_{100.1} - p_{101.0} + p_{101.1} \\ -p_{100.0} + p_{100.1} - p_{101.0} + p_{101.1} \\ p_{010.0} - p_{010.1} + p_{011.0} - p_{011.1} \\ 0 \end{array} \right\},$$

$$\bar{\xi}(1) \leq \min \left\{ \begin{array}{c} p_{000.0} - p_{000.1} + p_{001.0} - 2p_{001.1} - p_{011.1} + p_{101.0} - p_{101.1} - p_{110.0} - p_{111.0} + 1 \\ -p_{001.1} - p_{011.1} - p_{110.0} - p_{111.0} + 1 \end{array} \right\},$$

$$\bar{\xi}(0) \geq \max \left\{ \begin{array}{c} p_{010.0} - p_{010.1} + p_{011.0} - p_{011.1} - p_{100.0} + p_{100.1} - p_{101.0} + p_{101.1} \\ -p_{100.0} + p_{100.1} - p_{101.0} + p_{101.1} \\ p_{010.0} - p_{010.1} + p_{011.0} - p_{011.1} \\ 0 \end{array} \right\},$$

$$\bar{\xi}(0) \leq \min \left\{ \begin{array}{c} -p_{001.1} - p_{010.0} + p_{010.1} - p_{100.0} - 2p_{110.0} + p_{110.1} - p_{111.0} + p_{111.1} + 1 \\ -p_{001.1} - p_{100.0} - p_{110.0} + 1 \end{array} \right\}.$$

Under Assumptions 5.2.1 - 5.2.7, we have

$$\bar{\xi}(1) \geq \max \left\{ \begin{array}{c} p_{010.0} - p_{010.1} + p_{011.0} - p_{011.1} - p_{100.0} + p_{100.1} - p_{101.0} + p_{101.1} \\ -p_{100.0} + p_{100.1} - p_{101.0} + p_{101.1} \\ p_{010.0} - p_{010.1} + p_{011.0} - p_{011.1} \\ 0 \end{array} \right\},$$

$$\bar{\xi}(1) \leq -p_{001.1} - p_{011.1} - p_{100.0} - p_{110.0} - p_{111.0} + 1$$

$$\bar{\xi}(0) \geq \max \left\{ \begin{array}{c} p_{010.0} - p_{010.1} + p_{011.0} - p_{011.1} - p_{100.0} + p_{100.1} - p_{101.0} + p_{101.1} \\ -p_{100.0} + p_{100.1} - p_{101.0} + p_{101.1} \\ p_{010.0} - p_{010.1} + p_{011.0} - p_{011.1} \\ 0 \end{array} \right\},$$

$$\bar{\xi}(0) \leq -p_{001.1} - p_{011.1} - p_{100.0} - p_{110.0} + 1.$$

5.7.3 Bounds on Local Parameters under Monotonicity

For local causal parameters, we have shown the sharp bounds on $\bar{\delta}^*(a)$ under Assumptions 5.2.1 - 5.2.4 in Section 5.4 of the main body. In this section, we present the sharp bounds of $\bar{\delta}^*(a)$ under additional monotonicity assumptions. Under Assumptions 5.2.1 - 5.2.5, we

have

$$\bar{\delta}^*(1) \geq \max \left\{ \begin{array}{c} -p_{000.0} + p_{000.1} - p_{001.0} + p_{001.1} - p_{100.0} + p_{100.1} - p_{101.0} + p_{101.1} \\ p_{011.0} - p_{011.1} \end{array} \right\},$$

$$\bar{\delta}^*(1) \leq \min \left\{ \begin{array}{c} p_{000.0} - p_{000.1} + p_{001.0} - p_{001.1} + p_{100.0} - p_{100.1} + p_{101.0} - p_{101.1} \\ -p_{111.0} + p_{111.1} \end{array} \right\},$$

$$\bar{\delta}^*(0) \geq \max \left\{ \begin{array}{c} -p_{000.0} + p_{000.1} - p_{001.0} + p_{001.1} - p_{100.0} + p_{100.1} - p_{101.0} + p_{101.1} \\ -p_{100.0} + p_{100.1} \end{array} \right\},$$

$$\bar{\delta}^*(0) \leq \min \left\{ \begin{array}{c} p_{000.0} - p_{000.1} + p_{001.0} - p_{001.1} + p_{100.0} - p_{100.1} + p_{101.0} - p_{101.1} \\ p_{000.0} - p_{000.1} \end{array} \right\}.$$

Under Assumptions 5.2.1 - 5.2.4, 5.2.6, we have

$$\bar{\delta}^*(1) \geq \max \left\{ \begin{array}{c} -p_{010.0} + p_{010.1} - p_{110.0} + p_{110.1} \\ p_{001.0} - p_{001.1} \end{array} \right\},$$

$$\bar{\delta}^*(1) \leq \min \left\{ \begin{array}{c} p_{000.0} - p_{000.1} + p_{100.0} - p_{100.1} \\ -p_{111.0} + p_{111.1} \end{array} \right\},$$

$$\bar{\delta}^*(0) \geq \max \left\{ \begin{array}{c} p_{001.0} - p_{001.1} + p_{101.0} - p_{101.1} \\ -p_{110.0} + p_{110.1} \end{array} \right\},$$

$$\bar{\delta}^*(0) \leq \min \left\{ \begin{array}{c} -p_{011.0} + p_{011.1} - p_{111.0} + p_{111.1} \\ p_{000.0} - p_{000.1} \end{array} \right\}.$$

Under Assumptions 5.2.1 - 5.2.4, 5.2.7, we have

$$\bar{\delta}^*(1) \geq \max \left\{ \begin{array}{c} -p_{010.0} + p_{010.1} + p_{011.0} - p_{011.1} - p_{110.0} + p_{110.1} \\ -p_{000.0} + p_{000.1} + p_{001.0} - p_{001.1} - p_{100.0} + p_{100.1} \\ p_{001.0} - p_{001.1} + p_{011.0} - p_{011.1} \end{array} \right\},$$

$$\bar{\delta}^*(1) \leq \min \left\{ \begin{array}{c} p_{000.0} - p_{000.1} + p_{001.0} - p_{001.1} + p_{010.0} - p_{010.1} + p_{011.0} - p_{011.1} \\ p_{000.0} - p_{000.1} - p_{101.0} + p_{101.1} \\ p_{010.0} - p_{010.1} - p_{111.0} + p_{111.1} \end{array} \right\},$$

$$\bar{\delta}^*(0) \geq \max \left\{ \begin{array}{c} p_{011.0} - p_{011.1} - p_{110.0} + p_{110.1} + p_{111.0} - p_{111.1} \\ p_{001.0} - p_{001.1} - p_{100.0} + p_{100.1} + p_{101.0} - p_{101.1} \\ -p_{100.0} + p_{100.1} - p_{110.0} + p_{110.1} \end{array} \right\},$$

$$\bar{\delta}^*(0) \leq \min \left\{ \begin{array}{c} p_{000.0} - p_{000.1} + p_{001.0} - p_{001.1} + p_{010.0} - p_{010.1} + p_{011.0} - p_{011.1} \\ p_{000.0} - p_{000.1} - p_{101.0} + p_{101.1} \\ p_{010.0} - p_{010.1} - p_{111.0} + p_{111.1} \end{array} \right\}.$$

Under Assumptions 5.2.1 - 5.2.6, we have

$$0 \leq \bar{\delta}^*(1) \leq \min \left\{ \begin{array}{c} p_{000.0} - p_{000.1} + p_{001.0} - p_{001.1} + p_{100.0} - p_{100.1} + p_{101.0} - p_{101.1} \\ -p_{111.0} + p_{111.1} \end{array} \right\},$$

$$0 \leq \bar{\delta}^*(0) \leq \min \left\{ \begin{array}{c} p_{000.0} - p_{000.1} + p_{001.0} - p_{001.1} + p_{100.0} - p_{100.1} + p_{101.0} - p_{101.1} \\ p_{000.0} - p_{000.1} \end{array} \right\}.$$

Under Assumptions 5.2.1 - 5.2.5, 5.2.7, we have

$$\bar{\delta}^*(1) \geq \max \left\{ \begin{array}{c} -p_{000.0} + p_{000.1} - p_{001.0} + p_{001.1} - p_{100.0} + p_{100.1} - p_{101.0} + p_{101.1} \\ p_{011.0} - p_{011.1} \end{array} \right\},$$

$$\bar{\delta}^*(1) \leq \min \left\{ \begin{array}{c} p_{000.0} - p_{000.1} + p_{001.0} - p_{001.1} + p_{100.0} - p_{100.1} + p_{101.0} - p_{101.1} \\ p_{000.0} - p_{000.1} + p_{001.0} - p_{001.1} + p_{010.0} - p_{010.1} + p_{011.0} - p_{011.1} \\ p_{000.0} - p_{000.1} + p_{001.0} - p_{001.1} \\ -p_{110.0} + p_{110.1} - p_{111.0} + p_{111.1} \end{array} \right\},$$

$$\bar{\delta}^*(0) \geq \max \left\{ \begin{array}{c} -p_{000.0} + p_{000.1} - p_{001.0} + p_{001.1} - p_{100.0} + p_{100.1} - p_{101.0} + p_{101.1} \\ -p_{100.0} + p_{100.1} \end{array} \right\},$$

$$\bar{\delta}^*(0) \leq \min \left\{ \begin{array}{c} p_{000.0} - p_{000.1} + p_{001.0} - p_{001.1} + p_{100.0} - p_{100.1} + p_{101.0} - p_{101.1} \\ p_{000.0} - p_{000.1} + p_{001.0} - p_{001.1} + p_{010.0} - p_{010.1} + p_{011.0} - p_{011.1} \\ -p_{110.0} + p_{110.1} - p_{111.0} + p_{111.1} \\ p_{000.0} - p_{000.1} + p_{001.0} - p_{001.1} \end{array} \right\}.$$

Under Assumptions 5.2.1 - 5.2.7, we have

$$0 \leq \bar{\delta}^*(1) \leq \min \left\{ \begin{array}{c} p_{000.0} - p_{000.1} + p_{001.0} - p_{001.1} + p_{010.0} - p_{010.1} + p_{011.0} - p_{011.1} \\ p_{000.0} - p_{000.1} + p_{001.0} - p_{001.1} + p_{100.0} - p_{100.1} + p_{101.0} - p_{101.1} \\ p_{000.0} - p_{000.1} + p_{001.0} - p_{001.1} \\ -p_{110.0} + p_{110.1} - p_{111.0} + p_{111.1} \end{array} \right\},$$

$$0 \leq \bar{\delta}^*(0) \leq \min \left\{ \begin{array}{c} p_{000.0} - p_{000.1} + p_{001.0} - p_{001.1} + p_{010.0} - p_{010.1} + p_{011.0} - p_{011.1} \\ p_{000.0} - p_{000.1} + p_{001.0} - p_{001.1} + p_{100.0} - p_{100.1} + p_{101.0} - p_{101.1} \\ p_{000.0} - p_{000.1} + p_{001.0} - p_{001.1} \\ -p_{110.0} + p_{110.1} - p_{111.0} + p_{111.1} \end{array} \right\}.$$

The above bounds are the tightest bounds under corresponding assumptions. Similarly, they would be tighter if we impose more assumptions. It is worth noting that the probability entry $p_{000.0} - p_{000.1} + p_{001.0} - p_{001.1} + p_{010.0} - p_{010.1} + p_{011.0} - p_{011.1}$ frequently appear in the probability entries of the upper bounds of $\bar{\delta}^*(1)$ and $\bar{\delta}^*(0)$. We observe that

$$\frac{p_{000.0} - p_{000.1} + p_{001.0} - p_{001.1} + p_{010.0} - p_{010.1} + p_{011.0} - p_{011.1}}{\Pr(A(1) = 1, A(0) = 0)} = \tau^l.$$

Then, the $\bar{\delta}^l(a)$ cannot be greater than τ^l under certain monotonicity assumptions ensuring that $\bar{\xi}^l(1-a)$ is non-negative.

Chapter 6

Future Work and Discussion

A brief summary and discussions on potential extensions and future work are provided in this chapter to conclude this thesis.

6.1 Summary

In Chapter 2, we introduced the frameworks, models, and parameters of interest for causal inference as preliminaries for the three topics we later explored in Chapters 3 - 5.

The first topic is presented in Chapter 3, which focused on statistical inference of population weighted average treatment effects. We developed in this chapter a framework that incorporates the survey weights in the propensity score weighting procedure to simultaneously mitigate the confounding bias as well as the selection bias. The key of our method is to utilize the survey weights not only for re-weighting the outcomes, but also for the estimation of the propensity score. We later extended the proposed method to address

the scenario with non-probability samples by taking advantage of one or multiple external reference datasets with possibly partially observed confounding variables.

The second topic of this thesis is discussed in Chapter 4, which explored robust estimation of the treatment harm rate (THR). The THR is a measure of treatment risk and assesses the treatment effect heterogeneity within a subpopulation. However, it is non-identifiable and can only be partially identified. We constructed in Chapter 4 a class of partitioning-based bounds for the THR which can be consistently estimated with RCT data. These bounds incorporate the information of auxiliary covariates and attain the information upper bound under the optimal partition. A cross-fitting strategy is adopted for learning the optimal partitioning rule and estimation of the bounds.

Chapter 5 dived into causal mediation analysis and investigated the partial identification issue in randomized controlled trials with noncompliance. We derived nonparametric bounds of causal mediation effects under minimal assumptions and extended existing results in causal mediation analysis with RCT data to the instrumental variable setting.

The three topics studied in the thesis address three distinct inferential problems under the common umbrella for causal inference. The statistical tool developed in Chapter 4, however, can be applied to improve the bounds derived in Chapter 5. We further discuss this potential line of new research in Section 6.2.2.

6.2 Future Work

6.2.1 Variable Selection for Causal Inference with Nonprobability Samples

In Section 3.4, we studied how to estimate the population average treatment effects when the data at hand are nonprobability samples by leveraging external probability data with information in the covariates. Particularly, we considered the scenario where the covariates in the original nonprobability data are only partially observed in the external reference datasets. As demonstrated in Proposition 3.4.1, when the sample inclusion indicator is ignorable with respect to the treatment indicator, the covariates in the reference dataset are sufficient for identification if they can fully explain the variation in $\tau(\boldsymbol{X})$. For example, if $Y(1) = X_1 + X_2 + X_3 + \varepsilon_1$ and $Y(0) = X_1 + X_2 + \varepsilon_0$ where ε_1 and ε_0 are random noises independent of the selection indicator variable R, then we can identify the population average treatment effect as long as X_3 is observed for the reference data. This has also been verified numerically in the simulation studies given in Section 3.4.5. The extended question of interest is whether we should also include X_1 in the estimation of the participation probability provided that X_1 is observed for the reference data? The answer of this question relies on whether the inclusion of X_1 can improve the efficiency of the resulting inverse weighted estimator.

With regard to variable selection for the estimation of propensity score, a consensus in causal inference is that the variables associated with the outcome should be involved because they reduce the estimator's variance (Brookhart et al., 2006; Rotnitzky et al., 2010; De Luna et al., 2011; Shortreed and Ertefaie, 2017). However, it is not known whether this principle is applicable to the estimation of the participation probability. This is potentially a future research topic.

6.2.2 Partitioning-Based Bounds for General Causal Parameters on Binary Outcomes in RCTs

In a statistical problem, we are often interested in estimating or making inference about some parameters of interest of a statistical model given the observed data. We can place different assumptions on the data, leading to different statistical models. A parameter is identifiable if different values of the parameter must yield different statistical models. Under mild assumptions, the identifiability of a parameter is equivalent to that we can consistently estimate the parameter based on infinite observed data generated from the model. The non-identifiability issue arises when the parameters of interest are not identifiable, which is common in econometrics.

In a RCT, for example, the average treatment effect (ATE) is non-identifiable in the presence of noncompliance under the exclusion-restriction assumption. There are three popular solutions to the non-identifiability issue: (1) imposing strong identification assumptions and focusing on similar parameters that are identifiable; see Angrist et al. (1996) for example, which assumes no defiers and shifts interest to the complier ATE; (2) conducting sensitivity analysis for the parameter (Robins et al., 2000); and (3) addressing the partial identification issue and deriving the bounds on the parameter (Balke and Pearl, 1997). The three solutions are not mutually exclusive; rather, they can complement each other and can be reported together in statistical analysis.

When the outcomes are binary, Balke (1996) showed that the sharp nonparametric bounds on the average treatment effect in RCTs with noncompliance are the solutions of linear programming problems, with the objective functions of which determined by the parameter of interest, and the constraints relying on the assumptions on the statistical model. Therefore, the Balke-Pearl algorithm can be easily generalized to address the

partial identification issue in other causal inference problems when the outcomes are binary. In Chapter 5, we have used the Balke-Pearl algorithm to derive sharp non-parametric bounds on the causal mediation effects with instrumental variables. In fact, the partial identification issue has been extensively studied in causal inference concerning principle stratification, instrumental variables, and mediation analysis. Many of the existing bounds of causal parameters on binary outcomes that have been derived are in a general form of taking the maximum or minimum value from a set of probability entries (Cheng and Small, 2006; Cai et al., 2008; Richardson et al., 2014; Swanson et al., 2018; Gabriel et al., 2023). As demonstrated in Section 5.7.1, those probability entries are essentially corresponding to the vertices of (symbolic) polyhedral under the Balke-Pearl algorithm. We show that these bounds of similar forms can be improved in the presence of covariates based on the trick of partitioning outlined in Chapter 4.

Suppose θ is the parameter of interest. Let $LB(\theta)$ and $UB(\theta)$ denote respectively the lower and upper bound of θ. We assume that $LB(\theta)$ and $UB(\theta)$ take the forms $LB(\theta) = \max \mathcal{P}_l$ and $UB(\theta) = \min \mathcal{P}_u$, where $\mathcal{P}_l$ and $\mathcal{P}_u$ are collections of symbolic probability entries that can be identified from the observed RCT data; see Section 5.3 for illustrative examples of such bounds. Let $\boldsymbol{X} \in \mathbb{X}$ denote baseline covariates. Consider a partition of $\mathbb{X}$ as $\Delta = \{\mathbb{X}_1, \cdots, \mathbb{X}_J\}$ and define $\mathcal{P}_{l,\mathbb{X}_j}$ and $\mathcal{P}_{u,\mathbb{X}_j}$ as collections of probabilities that are replaced by their conditional counterparts given $\boldsymbol{X} \in \mathbb{X}_j$ — any probability expression $\Pr(\cdot)$ in $\mathcal{P}_l$ will be replaced by $\Pr(\cdot \mid \boldsymbol{X} \in \mathbb{X}_j)$ in $\mathcal{P}_{l,\mathbb{X}_j}$. The partitioning-based bounds (PBBs) for θ are given by

$$L(\theta, \Delta) = \sum_{j=1}^{J} \Pr(\boldsymbol{X} \in \mathbb{X}_j) \max \mathcal{P}_{l,\mathbb{X}_j}, \quad U(\theta, \Delta) = \sum_{j=1}^{J} \Pr(\boldsymbol{X} \in \mathbb{X}_j) \min \mathcal{P}_{u,\mathbb{X}_j}.$$

The PBBs possess two desirable. First, by Jensen's inequality, it follows that $LB(\theta) \leq$

$L(\theta, \Delta) \leq U(\theta, \Delta) \leq UB(\theta)$, implying the partitioning-based are improved after incorporating the information of covariates. Second, provided a partition, the partitioning-based bounds can be consistently estimated as long as J is not too large.

The choice of the partition Δ is critical to the PBBs. A simple and commonly used approach is to define strata based on categorical covariates, such as gender or race. However, traditional stratification is limited to discrete covariates, and the number of strata can become prohibitively large when the dimensionality of the covariates for stratification increases. There are indeed optimal partitions for the PBBs, or equivalently, optimal rules for partitioning.

Suppose $\mathcal{P}_{l,\boldsymbol{X}} = \{p_{l1}(\boldsymbol{X}), \cdots, p_{lL}(\boldsymbol{X})\}$ and $\mathcal{P}_{u,\boldsymbol{X}} = \{p_{u1}(\boldsymbol{X}), \cdots, p_{uU}(\boldsymbol{X})\}$. The optimal partition for the partitioning-based lower bound of θ is given by $\Delta_{opt,low} = \{\mathbb{X}_{l1}, \cdots, \mathbb{X}_{lL}\}$ where $\mathbb{X}_{lk} = \{\boldsymbol{x} \in \mathbb{X} : p_{lk}(\boldsymbol{x}) \geq p_{li}(\boldsymbol{x}) \text{ for } i = 1, \cdots, L\}$; that is, in every cell of the optimal partition, one of the probabilities entries in the collection dominate the others. Analogously, the optimal partition for $U(\theta, \Delta)$ is given by $\Delta_{opt,upp} = \{\mathbb{X}_{u1}, \cdots, \mathbb{X}_{uL}\}$ where $\mathbb{X}_{uk} = \{\boldsymbol{x} \in \mathbb{X} : p_{uk}(\boldsymbol{x}) \geq p_{ui}(\boldsymbol{x}) \text{ for } i = 1, \cdots, U\}$. The $[L(\theta, \Delta_{opt,low}), U(\theta, \Delta_{opt,upp})]$ are the tightest PBBs, which are equal to $[\mathrm{E}(\max \mathcal{P}_{l,\boldsymbol{X}}), \mathrm{E}(\min \mathcal{P}_{u,\boldsymbol{X}})]$.

The optimal rules for partitioning can be learned from external datasets. Alternatively, we can employ the cross-fitting strategy presented in Section 4.4.3 to construct data-driven partitions based on estimated optimal rules. The details have been outlined in Chapter 4.

References

Ackerman, B., Lesko, C. R., Siddique, J., Susukida, R., and Stuart, E. A. (2021), "Generalizing randomized trial findings to a target population using complex survey population data," *Statistics in Medicine*, 40, 1101–1120.

Altman, D. G., Schulz, K. F., Moher, D., Egger, M., Davidoff, F., Elbourne, D., Gøtzsche, P. C., Lang, T., and Group, C. (2001), "The revised CONSORT statement for reporting randomized trials: explanation and elaboration," *Annals of Internal Medicine*, 134, 663–694.

Angrist, J. D. and Imbens, G. W. (1995), "Two-stage least squares estimation of average causal effects in models with variable treatment intensity," *Journal of the American Statistical Association*, 90, 431–442.

Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996), "Identification of causal effects using instrumental variables," *Journal of the American Statistical Association*, 91, 444–455.

Aronow, P. M., Green, D. P., and Lee, D. K. (2014), "Sharp bounds on the variance in randomized experiments," *The Annals of Statistics*, 42, 850–871.

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

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

Austin, P. C. and Stuart, E. A. (2015), "Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies," *Statistics in Medicine*, 34, 3661–3679.

Ayer, M., Brunk, H. D., Ewing, G. M., Reid, W. T., and Silverman, E. (1955), "An empirical distribution function for sampling with incomplete information," *The Annals of Mathematical Statistics*, 641–647.

Balke, A. and Pearl, J. (1994), "Counterfactual probabilities: Computational methods, bounds and applications," in *Uncertainty in Artificial Intelligence*, Elsevier, pp. 46–54.

— (1997), "Bounds on treatment effects from studies with imperfect compliance," *Journal of the American Statistical Association*, 92, 1171–1176.

Balke, A. A. (1996), "Probabilistic counterfactuals: semantics, computation, and applications," Ph.D. thesis, University of California at Los Angeles, USA, uMI Order No. GAX96-10837.

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

Bareinboim, E. and Pearl, J. (2016), "Causal inference and the data-fusion problem," *Proceedings of the National Academy of Sciences*, 113, 7345–7352.

Beaumont, J.-F. (2005), "Calibrated imputation in surveys under a quasi-model-assisted approach," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67, 445–458.

Bhattacharya, D. (2005), "Asymptotic inference from multi-stage samples," *Journal of Econometrics*, 126, 145–171.

Boyd, S., Boyd, S. P., and Vandenberghe, L. (2004), *Convex Optimization*, Cambridge: Cambridge University Press.

Breiman, L. (2001), "Random forests," *Machine Learning*, 45, 5–32.

Brick, J. M. (2015), "Compositional model inference," in *Proceedings of the Section on Survey Research Methods*, Joint Statistical Meetings, American Statistical Association Alexandria, VA, pp. 299–307.

Brookhart, M. A., Schneeweiss, S., Rothman, K. J., Glynn, R. J., Avorn, J., and Stürmer, T. (2006), "Variable selection for propensity score models," *American Journal of Epidemiology*, 163, 1149–1156.

Buchanan, A. L., Hudgens, M. G., Cole, S. R., Mollan, K. R., Sax, P. E., Daar, E. S., Adimora, A. A., Eron, J. J., and Mugavero, M. J. (2018), "Generalizing evidence from randomized trials using inverse probability of sampling weights," *Journal of the Royal Statistical Society Series A: Statistics in Society*, 181, 1193–1209.

Busing, F. M. (2022), "Monotone regression: A simple and fast O (n) PAVA implementation," *Journal of Statistical Software*, 102, 1–25.

Cai, Z., Kuroki, M., Pearl, J., and Tian, J. (2008), "Bounds on direct effects in the presence of confounded intermediate variables," *Biometrics*, 64, 695–701.

Cao, W., Tsiatis, A. A., and Davidian, M. (2009), "Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data," *Biometrika*, 96, 723–734.

Carlson, M. D. and Morrison, R. S. (2009), "Study design, precision, and validity in observational studies," *Journal of Palliative Medicine*, 12, 77–82.

Chambers, R. and Clark, R. (2012), *An Introduction to Model-Based Survey Sampling with Applications*, New York, NY: Oxford University Press.

Chan, K. C. G., Yam, S. C. P., and Zhang, Z. (2016), "Globally efficient non-parametric inference of average treatment effects by empirical balancing calibration weighting," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 78, 673–700.

Chattopadhyay, A., Hase, C. H., and Zubizarreta, J. R. (2020), "Balancing vs modeling approaches to weighting in practice," *Statistics in Medicine*, 39, 3227–3254.

Chen, J. and Sitter, R. (1999), "A pseudo empirical likelihood approach to the effective use of auxiliary information in complex surveys," *Statistica Sinica*, 385–406.

Chen, T. and Guestrin, C. (2016), "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining*, pp. 785–794.

Chen, Y., Li, P., and Wu, C. (2020), "Doubly robust inference with nonprobability survey samples," *Journal of the American Statistical Association*, 115, 2011–2021.

Cheng, J. and Small, D. S. (2006), "Bounds on causal effects in three-arm trials with noncompliance," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 68, 815–836.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018), "Double/debiased machine learning for treatment and structural parameters," *The Econometrics Journal*, 21, C1–C68.

Chow, C. (1970), "On optimum recognition error and reject tradeoff," *IEEE Transactions on Information Theory*, 16, 41–46.

Cole, S. R. and Stuart, E. A. (2010), "Generalizing evidence from randomized clinical trials to target populations: the ACTG 320 trial," *American Journal of Epidemiology*, 172, 107–115.

Cortes, C. and Vapnik, V. (1995), "Support vector networks," *Machine Learning*, 20, 273–297.

Couper, M. P. (2000), "Web surveys: A review of issues and approaches," *The Public Opinion Quarterly*, 64, 464–494.

Crump, R. K., Hotz, V. J., Imbens, G. W., and Mitnik, O. A. (2009), "Dealing with limited overlap in estimation of average treatment effects," *Biometrika*, 96, 187–199.

Cuzick, J., Edwards, R., and Segnan, N. (1997), "Adjusting for non-compliance and contamination in randomized clinical trials," *Statistics in Medicine*, 16, 1017–1029.

Dahabreh, I. J., Robertson, S. E., Steingrimsson, J. A., Stuart, E. A., and Hernan, M. A. (2020), "Extending inferences from a randomized trial to a new target population," *Statistics in Medicine*, 39, 1999–2014.

Dahabreh, I. J., Robertson, S. E., Tchetgen, E. J., Stuart, E. A., and Hernán, M. A. (2019), "Generalizing causal inferences from individuals in randomized trials to all trial-eligible individuals," *Biometrics*, 75, 685–694.

De Luna, X., Waernbaum, I., and Richardson, T. S. (2011), "Covariate selection for the nonparametric estimation of an average treatment effect," *Biometrika*, 98, 861–875.

Deville, J.-C. and Särndal, C.-E. (1992), "Calibration estimators in survey sampling," *Journal of the American Statistical Association*, 87, 376–382.

Diemert, E., Betlei, A., Renaudin, C., and Amini, M.-R. (2018), "A large scale benchmark for uplift modeling," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*.

Diemert, E., Betlei, A., Renaudin, C., Amini, M.-R., Gregoir, T., and Rahier, T. (2021), "A large scale benchmark for individual treatment effect prediction and uplift modeling," *arXiv preprint arXiv:2111.10106*.

Ding, P. (2024), *A First Course in Causal Inference*, New York, NY: Chapman and Hall/CRC.

Ding, P. and Vanderweele, T. J. (2016), "Sharp sensitivity bounds for mediation under unmeasured mediator-outcome confounding," *Biometrika*, 103, 483–490.

Dong, N., Stuart, E. A., Lenis, D., and Quynh Nguyen, T. (2020), "Using propensity score analysis of survey data to estimate population average treatment effects: A case study comparing different methods," *Evaluation Review*, 44, 84–108.

DuGoff, E. H., Schuler, M., and Stuart, E. A. (2014), "Generalizing observational study results: applying propensity score methods to complex surveys," *Health Services Research*, 49, 284–303.

DuMouchel, W. H. and Duncan, G. J. (1983), "Using sample survey weights in multiple re-

gression analyses of stratified samples," *Journal of the American Statistical Association*, 78, 535–543.

Durrett, R. (2010), *Probability: Theory and Examples*, Cambridge: Cambridge University Press, 4th ed.

Elliott, M. R. and Valliant, R. (2017), "Inference for nonprobability samples," *Statistical Science*, 32, 249–264.

Evans, A. and Kalra, L. (2001), "Are the results of randomized controlled trials on anticoagulation in patients with atrial fibrillation generalizable to clinical practice?" *Archives of Internal Medicine*, 161, 1443–1447.

Fan, J., Imai, K., Lee, I., Liu, H., Ning, Y., and Yang, X. (2021), "Optimal covariate balancing conditions in propensity score estimation," *Journal of Business & Economic Statistics*, 1–14.

Foygel Barber, R. (2024), "Hoeffding and Bernstein inequalities for weighted sums of exchangeable random variables," *Electronic Communications in Probability*, 29, 1–13.

Franc, V., Prusa, D., and Voracek, V. (2023), "Optimal strategies for reject option classifiers," *Journal of Machine Learning Research*, 24, 1–49.

Fréchet, M. (1935), "Généralisation du théoreme des probabilités totales," *Fundamenta Mathematicae*, 25, 379–387.

Gabriel, E. E., Sjölander, A., and Sachs, M. C. (2023), "Nonparametric bounds for causal effects in imperfect randomized experiments," *Journal of the American Statistical Association*, 118, 684–692.

Gadbury, G. L., Iyer, H. K., and Albert, J. M. (2004), "Individual treatment effects in randomized trials with binary outcomes," *Journal of Statistical Planning and Inference*, 121, 163–174.

Greenhouse, J. B., Kaizar, E. E., Kelleher, K., Seltman, H., and Gardner, W. (2008), "Generalizing from clinical trial data: a case study. The risk of suicidality among pediatric antidepressant users," *Statistics in Medicine*, 27, 1801–1813.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. (2017), "On calibration of modern neural networks," in *Proceedings of the 34th International Conference on Machine Learning*, eds. Precup, D. and Teh, Y. W., PMLR, vol. 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330.

Hainmueller, J. (2012), "Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies," *Political Analysis*, 20, 25–46.

Hammer, S. M., Katzenstein, D. A., Hughes, M. D., Gundacker, H., Schooley, R. T., Haubrich, R. H., Henry, W. K., Lederman, M. M., Phair, J. P., Niu, M., et al. (1996), "A trial comparing nucleoside monotherapy with combination therapy in HIV-infected adults with CD4 cell counts from 200 to 500 per cubic millimeter," *New England Journal of Medicine*, 335, 1081–1090.

Han, P. and Wang, L. (2013), "Estimation with missing data: beyond double robustness," *Biometrika*, 100, 417–430.

Hansen, L. P. (1982), "Large sample properties of generalized method of moments estimators," *Econometrica: Journal of the Econometric Society*, 1029–1054.

Hartley, H. and Rao, J. (1962), "Sampling with unequal probabilities and without replacement," *The Annals of Mathematical Statistics*, 350–374.

Hastie, T., Tibshirani, R., and Friedman, J. (2009), *The Elements of Statistical Learning*, New York, NY: Springer, 2nd ed.

Herbei, R. and Wegkamp, M. H. (2006), "Classification with reject option," *The Canadian Journal of Statistics*, 709–721.

Hirano, K. and Imbens, G. W. (2001), "Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization," *Health Services and Outcomes Research Methodology*, 2, 259–278.

Hirano, K., Imbens, G. W., and Ridder, G. (2003), "Efficient estimation of average treatment effects using the estimated propensity score," *Econometrica*, 71, 1161–1189.

Hirano, K., Imbens, G. W., Rubin, D. B., and Zhou, X.-H. (2000), "Assessing the effect of an influenza vaccine in an encouragement design," *Biostatistics*, 1, 69–88.

Hirt, C. W., Amsden, A. A., and Cook, J. L. (1974), "An arbitrary Lagrangian-Eulerian computing method for all flow speeds," *Journal of Computational Physics*, 14, 227–253.

Hong, H., Leung, M. P., and Li, J. (2020), "Inference on finite-population treatment effects under limited overlap," *The Econometrics Journal*, 23, 32–47.

Horvitz, D. G. and Thompson, D. J. (1952), "A generalization of sampling without replacement from a finite universe," *Journal of the American Statistical Association*, 47, 663–685.

Huang, E. J., Fang, E. X., Hanley, D. F., and Rosenblum, M. (2017), "Inequality in treatment benefits: Can we determine if a new treatment benefits the many or the few?" *Biostatistics*, 18, 308–324.

— (2019), "Constructing a confidence interval for the fraction who benefit from treatment, using randomized trial data," *Biometrics*, 75, 1228–1239.

Imai, K., Keele, L., and Yamamoto, T. (2010), "Identification, inference and sensitivity analysis for causal mediation effects," *Statistical Science*, 25, 51 – 71.

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

Imbens, G. (2014), "Instrumental variables: An econometrician's perspective," Tech. rep., National Bureau of Economic Research.

Imbens, G. W. and Rubin, D. B. (1997), "Bayesian inference for causal effects in randomized experiments with noncompliance," *The Annals of Statistics*, 305–327.

— (2015), *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*, Cambridge: Cambridge University Press.

Isaki, C. T. and Fuller, W. A. (1982), "Survey design under the regression superpopulation model," *Journal of the American Statistical Association*, 77, 89–96.

Josey, K. P., Juarez-Colunga, E., Yang, F., and Ghosh, D. (2021), "A framework for covariate balance using Bregman distances," *Scandinavian Journal of Statistics*, 48, 790–816.

Kallus, N. (2022), "What's the harm? sharp bounds on the fraction negatively affected by treatment," *Advances in Neural Information Processing Systems*, 35, 15996–16009.

Kang, J. D. and Schafer, J. L. (2007), "Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data," *Statistical Science*, 22, 523–539.

Kim, J. K. and Haziza, D. (2014), "Doubly robust inference with missing data in survey sampling," *Statistica Sinica*, 24, 375–394.

Kim, J. K. and Kim, J. J. (2007), "Nonresponse weighting adjustment using estimated response probability," *Canadian Journal of Statistics*, 35, 501–514.

Kim, J. K. and Park, M. (2010), "Calibration estimation in survey sampling," *International Statistical Review*, 78, 21–39.

Lê Cook, B., McGuire, T. G., Lock, K., and Zaslavsky, A. M. (2010), "Comparing methods of racial and ethnic disparities measurement across different settings of mental health care," *Health Services Research*, 45, 825–847.

Lee, B. K., Lessler, J., and Stuart, E. A. (2010), "Improving propensity score weighting using machine learning," *Statistics in Medicine*, 29, 337–346.

Lee, D., Yang, S., Dong, L., Wang, X., Zeng, D., and Cai, J. (2023), "Improving trial generalizability using observational studies," *Biometrics*, 79, 1213–1225.

Lee, S. and Valliant, R. (2009), "Estimation for volunteer panel web surveys using propensity score adjustment and calibration adjustment," *Sociological Methods & Research*, 37, 319–343.

Lei, J. (2014), "Classification with confidence," *Biometrika*, 101, 755–769.

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

Li, X. and Ding, P. (2016), "Exact confidence intervals for the average causal effect on a binary outcome," *Statistics in Medicine*, 35, 957–960.

— (2017), "General forms of finite population central limit theorems with applications to causal inference," *Journal of the American Statistical Association*, 112, 1759–1769.

Li, Y. and Li, L. (2021), "Propensity score analysis methods with balancing constraints: A Monte Carlo study," *Statistical Methods in Medical Research*, 30, 1119–1142.

Little, R. J. (2004), "To model or not to model? Competing modes of inference for finite population sampling," *Journal of the American Statistical Association*, 99, 546–556.

Little, R. J. and Rubin, D. B. (2019), *Statistical Analysis with Missing Data*, vol. 793, Hoboken, NJ: John Wiley & Sons.

Lumley, T. and Scott, A. (2017), "Fitting regression models to survey data," *Statistical Science*, 265–278.

Manski, C. F. (1990), "Nonparametric bounds on treatment effects," *The American Economic Review*, 80, 319–323.

Matheiss, T. and Rubin, D. S. (1980), "A survey and comparison of methods for finding all vertices of convex polyhedral sets," *Mathematics of Operations Research*, 5, 167–185.

Mattheiss, T. H. (1973), "An algorithm for determining irrelevant constraints and all vertices in systems of linear inequalities," *Operations Research*, 21, 247–260.

McCaffrey, D. F., Griffin, B. A., Almirall, D., Slaughter, M. E., Ramchand, R., and Burgette, L. F. (2013), "A tutorial on propensity score estimation for multiple treatments using generalized boosted models," *Statistics in Medicine*, 32, 3388–3414.

McCaffrey, D. F., Ridgeway, G., and Morral, A. R. (2004), "Propensity score estimation with boosted regression for evaluating causal effects in observational studies." *Psychological Methods*, 9, 403.

Minderer, M., Djolonga, J., Romijnders, R., Hubis, F., Zhai, X., Houlsby, N., Tran, D., and Lucic, M. (2021), "Revisiting the calibration of modern neural networks," *Advances in Neural Information Processing Systems*, 34, 15682–15694.

Nelder, J. A. and Wedderburn, R. W. (1972), "Generalized linear models," *Journal of the Royal Statistical Society Series A: Statistics in Society*, 135, 370–384.

Neyman, J. (1923a), "On the application of probability theory to agricultural experiments. Essay on principles," *Ann. Agricultural Sciences*, 1–51.

— (1923b), "Sur les applications de la théorie des probabilités aux experiences agricoles: Essai des principes," *Roczniki Nauk Rolniczych*, 10, 1–51.

Niculescu-Mizil, A. and Caruana, R. (2005), "Predicting good probabilities with supervised learning," in *Proceedings of the 22nd International Conference on Machine Learning*, New York, NY, USA: Association for Computing Machinery, ICML '05, p. 625–632.

O'Muircheartaigh, C. and Hedges, L. V. (2014), "Generalizing from unrepresentative experiments: a stratified propensity score approach," *Journal of the Royal Statistical Society Series C: Applied Statistics*, 63, 195–210.

Pearl, J. (2001), "Direct and indirect effects," in *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., UAI'01, p. 411–420.

— (2009), *Causality*, Cambridge: Cambridge University Press, 2nd ed.

Pearl, J., Glymour, M., and Jewell, N. P. (2016), *Causal Inference in Statistics: A Primer*, Chichester: John Wiley & Sons.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011), "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, 12, 2825–2830.

Platt, J. et al. (1999), "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," *Advances in Large Margin Classifiers*, 10, 61–74.

Qin, J. and Zhang, B. (2007), "Empirical-likelihood-based inference in missing response problems and its application in observational studies," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 69, 101–122.

Rafei, A., Elliott, M. R., and Flannagan, C. A. (2022), "Robust and efficient bayesian inference for non-probability samples," *arXiv preprint arXiv:2203.14355*.

Richardson, A., Hudgens, M. G., Gilbert, P. B., and Fine, J. P. (2014), "Nonparametric bounds and sensitivity analysis of treatment effects," *Statistical Science*, 29, 596 – 618.

Ridgeway, G., Kovalchik, S. A., Griffin, B. A., and Kabeto, M. U. (2015), "Propensity score analysis with survey weighted data," *Journal of Causal Inference*, 3, 237–249.

Rivers, D. (2007), "Sampling for web surveys," in *Proceedings of the Section on Survey Research Methods*, Joint Statistical Meetings, American Statistical Association Alexandria, VA, pp. 1–26.

Robins, J. M. and Greenland, S. (1996), "Identification of causal effects using instrumental variables: comment," *Journal of the American Statistical Association*, 91, 456–458.

Robins, J. M., Rotnitzky, A., and Scharfstein, D. O. (2000), "Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models," in *Statistical Models in Epidemiology, the Environment, and Clinical Trials*, New York, NY: Springer, pp. 1–94.

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

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

Rothwell, P. M. (2005), "External validity of randomised controlled trials: 'to whom do the results of this trial apply?'," *The Lancet*, 365, 82–93.

Rotnitzky, A., Li, L., and Li, X. (2010), "A note on overadjustment in inverse probability weighted estimation," *Biometrika*, 97, 997–1001.

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

— (1980), "Randomization analysis of experimental data: The Fisher randomization test comment," *Journal of the American Statistical Association*, 75, 591–593.

— (2005), "Causal inference using potential outcomes," *Journal of the American Statistical Association*, 100, 322–331.

Rubin-Bleuer, S. and Kratina, I. S. (2005), "On the two-phase framework for joint model and design-based inference," *The Annals of Statistics*, 33, 2789 – 2810.

Rudolph, K. E., Díaz, I., Rosenblum, M., and Stuart, E. A. (2014), "Estimating population treatment effects from a survey subsample," *American Journal of Epidemiology*, 180, 737–748.

Rudolph, K. E., Sofrygin, O., and van der Laan, M. J. (2021), "Complier stochastic direct effects: identification and robust estimation," *Journal of the American Statistical Association*, 116, 1254–1264.

Sadinle, M., Lei, J., and Wasserman, L. (2019), "Least ambiguous set-valued classifiers with bounded error levels," *Journal of the American Statistical Association*, 114, 223–234.

Shen, C., Jeong, J., Li, X., Chen, P.-S., and Buxton, A. (2013), "Treatment benefit and treatment harm rate to characterize heterogeneity in treatment effect," *Biometrics*, 69, 724–731.

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

Sjölander, A. (2009), "Bounds on natural direct effects in the presence of confounded intermediate variables," *Statistics in Medicine*, 28, 558–571.

Sparapani, R., Spanbauer, C., and McCulloch, R. (2021), "Nonparametric machine learning and efficient computation with Bayesian additive regression trees: The BART R package," *Journal of Statistical Software*, 97, 1–66.

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

Splawa-Neyman, J., Dabrowska, D. M., and Speed, T. P. (1990b), "On the application of probability theory to agricultural experiments. Essay on principles. Section 9," *Statistical Science*, 5, 465 – 472.

Stefanski, L. A. and Boos, D. D. (2002), "The calculus of M-estimation," *The American Statistician*, 56, 29–38.

Stuart, E. A., Ackerman, B., and Westreich, D. (2018), "Generalizability of randomized trial results to target populations: design and analysis possibilities," *Research on Social Work Practice*, 28, 532–537.

Stuart, E. A., Bradshaw, C. P., and Leaf, P. J. (2015), "Assessing the generalizability of randomized trial results to target populations," *Prevention Science*, 16, 475–485.

Swanson, S. A., Hernán, M. A., Miller, M., Robins, J. M., and Richardson, T. S. (2018), "Partial identification of the average treatment effect using instrumental variables: review of methods for binary instruments, treatments, and outcomes," *Journal of the American Statistical Association*, 113, 933–947.

Tan, Z. (2010), "Bounded, efficient and doubly robust estimation with inverse weighting," *Biometrika*, 97, 661–682.

Vaicenavicius, J., Widmann, D., Andersson, C., Lindsten, F., Roll, J., and Schön, T. (2019), "Evaluating model calibration in classification," in *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, eds. Chaudhuri, K. and Sugiyama, M., PMLR, vol. 89 of *Proceedings of Machine Learning Research*, pp. 3459–3467.

Valliant, R. and Dever, J. A. (2011), "Estimating propensity adjustments for volunteer web surveys," *Sociological Methods & Research*, 40, 105–137.

VanderWeele, T. J. (2011), "Controlled direct and mediated effects: definition, identification and bounds," *Scandinavian Journal of Statistics*, 38, 551–563.

VanderWeele, T. J. and Chiba, Y. (2014), "Sensitivity analysis for direct and indirect effects in the presence of exposure-induced mediator-outcome confounders," *Epidemiology, Biostatistics, and Public Health*, 11, e9027.

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

Wager, S., Du, W., Taylor, J., and Tibshirani, R. J. (2016), "High-dimensional regression adjustments in randomized experiments," *Proceedings of the National Academy of Sciences*, 113, 12673–12678.

Wong, R. K. and Chan, K. C. G. (2018), "Kernel-based covariate functional balancing for observational studies," *Biometrika*, 105, 199–213.

Wu, C. (2003), "Optimal calibration estimators in survey sampling," *Biometrika*, 90, 937–951.

— (2022), "Statistical inference with non-probability survey samples," *Survey Methodology*, 48, 283–311.

Wu, C. and Sitter, R. R. (2001), "A model-calibration approach to using complete auxiliary information from survey data," *Journal of the American Statistical Association*, 96, 185–193.

Wu, C. and Thompson, M. E. (2020), *Sampling Theory and Practice*, Cham: Springer.

Wu, P., Ding, P., Geng, Z., and Liu, Y. (2024), "Quantifying individual risk for binary outcome: Bounds and inference," *arXiv preprint arXiv:2402.10537*.

Wu, Y. and Fuller, W. A. (2005), "Preliminary testing procedures for regression with survey samples," in *Proceedings of the Section on Survey Research Methods*, American Statistical Association, pp. 3683–3688.

Xie, Y. (2013), "Population heterogeneity and causal inference," *Proceedings of the National Academy of Sciences*, 110, 6262–6268.

Ye, T., Shao, J., Yi, Y., and Zhao, Q. (2023), "Toward better practice of covariate adjustment in analyzing randomized clinical trials," *Journal of the American Statistical Association*, 118, 2370–2382.

Yin, Y., Liu, L., and Geng, Z. (2018), "Assessing the treatment effect heterogeneity with a latent variable," *Statistica Sinica*, 115–135.

Yuan, K.-H. and Jennrich, R. I. (1998), "Asymptotics of estimating equations under natural conditions," *Journal of Multivariate Analysis*, 65, 245–260.

Yuan, M. and Wegkamp, M. (2010), "Classification Methods with Reject Option Based on Convex Risk Minimization," *Journal of Machine Learning Research*, 11.

Zadrozny, B. and Elkan, C. (2001), "Obtaining calibrated probability estimates from decision trees and naive Bayesian classifiers," in *Proceedings of the Eighteenth International Conference on Machine Learning*, San Francisco, CA: Morgan Kaufmann Publishers Inc., ICML '01, p. 609–616.

Zhang, Z., Wang, C., Nie, L., and Soon, G. (2013), "Assessing the heterogeneity of treat-

ment effects via potential outcomes of individual patients," *Journal of the Royal Statistical Society Series C: Applied Statistics*, 62, 687–704.

Zhao, Q. (2019), "Covariate balancing propensity score by tailored loss functions," *The Annals of Statistics*, 47, 965–993.

Zhao, Q. and Percival, D. (2017), "Entropy balancing is doubly robust," *Journal of Causal Inference*, 5, 20160010.

Zhou, Y., Matsouaka, R. A., and Thomas, L. (2020), "Propensity score weighting under limited overlap and model misspecification," *Statistical Methods in Medical Research*, 29, 3721–3756.

Zubizarreta, J. R. (2015), "Stable weights that balance covariates for estimation with incomplete outcome data," *Journal of the American Statistical Association*, 110, 910–922.