Causal Inference with Covariate Balance Optimization

by

Yuying Xie

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, 2018

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:	Robert Platt Professor, McGill University
Supervisor(s):	Yeying Zhu Assistant Professor
	Cecilia Cotton Associate Professor
Internal Member:	Michael Wallace Assistant Professor
Internal Member:	Changbao Wu Professor
Internal-External Member:	Pierre Chaussé Associate Professor, Department of Economics

I hereby declare that I am the sole author of this 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.

Abstract

Causal inference is a popular problem in biostatistics, economics, and health science studies. The goal of this thesis is to develop new methods for the estimation of causal effects using propensity scores or inverse probability weights where weights are chosen in such a way to achieve balance in covariates across the treatment groups.

In Chapter 1, we introduce Neyman-Rubin Causal framework and causal inference with propensity scores. The importance of covariate balancing in causal inference is furthered discussed in this chapter. Besides, some general definitions and notations for causal inference are provided with many other popular propensity score approaches or weighting techniques in Chapter 2.

In Chapter 3, we describe a new model averaging approach to propensity score estimation in which parametric and nonparametric estimates are combined to achieve covariate balance. Simulation studies are conducted across different scenarios varying in the degree of interactions and nonlinearity in the treatment model. The results show that the proposed method produces less bias and smaller standard errors than existing approaches. They also show that a model averaging approach with the objective of minimizing the average Kolmogorov-Smirnov statistic leads to the best performance. The proposed approach is applied to a real data set in evaluating the causal effect of formula or mixed feeding versus exclusive breastfeeding in the first month of life on a child's BMI Z-score at age 4. The data analysis shows that formula or mixed feeding is more likely to lead to obesity at age 4, compared to exclusive breastfeeding.

In Chapter 4, we propose using kernel distance to measure balance across different treatment groups and propose a new propensity score estimator by setting the kernel distance to be zero. Compared to other balance measures, such as absolute standardized mean difference (ASMD) and Kolmogorov Smirnov (KS) statistic, kernel distance is one of the best bias indicators in estimating the causal effect. That is, the balance metric based on kernel distance is shown to have the strongest correlation with the absolute bias in estimating the causal effect, compared to several commonly used balance metrics. The kernel distance constraints are solved by generalized method of moments. Simulation studies are conducted across different scenarios varying in the degree of nonlinearity in both

the propensity score model and outcome model. The proposed approach produces smaller mean squared error in estimating causal treatment effects than many existing approaches including the well-known covariate balance propensity score (CBPS) approach when the propensity score model is misspecified. An application to data from the International Tobacco Control (ITC) policy evaluation project is provided.

Often interest lies in the estimation of quantiles other than the average causal effect. Other quantities such as quantiles or the quantile treatment effect may be of interest. In Chapter 5, we propose a multiply robust method for estimating marginal quantiles of potential outcomes by achieving mean balance in (1) the propensity score, and (2) the conditional distributions of potential outcomes. An empirical likelihood or entropy measure can be utilized instead of using inverse probability weighting. Simulation studies are conducted across different scenarios of correctness in both the propensity score models and outcome models. Our estimator is consistent if any of the models are correctly specified.

Acknowledgements

First and foremost I want to offer my sincerest gratitude to my supervisors, Dr. Yeying Zhu and Dr. Cecilia Cotton, who have supported me throughout this thesis with their patience, extensive knowledge and experience. They always support and encourage me to overcome any obstacles I have been facing through my PhD studies. Without their advice and insights, it would be impossible for me to complete this thesis. They became more of mentors and friends, than supervisors, which I truly appreciate. One simply could not wish for a better or friendlier supervisor, I am very grateful to have them as my supervisors.

I would also like to thank the rest of my thesis examining committee: Dr. Robert Platt, Dr. Pierre Chaussé and Dr. Changbao Wu, and Dr. Michael Wallace for taking their time to read my thesis, and for all the thoughtful and invaluable comments they give.

In addition, I am grateful to the department faculty and staff, for the knowledge I have learned and all the help and support I have received.

Many thanks go to my dear friends, who have supported me all the way. I am grateful for their friendship, encouragement, and the wonderful times we spent together.

Last but not the least, I would like to thank my parents and brother for their unconditional love. My special thanks go to Xinjie, who has always been encouraging me to follow my dreams.

This is dedicated to the one I love.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Neyman-Rubin Causal Framework

Causal inference is pervasive in many fields. In health research, researchers are interested in questions such as "What causes the disease?" or "Will aspirin reduce headaches?" In economics studies, researchers may ask "What are the factors driving increased gas prices?" In randomized trials, researchers measure the magnitude of causal effects by comparing the outcome when an action is applied with the outcome when no action is applied. The action is commonly known as a treatment, exposure or intervention. Confounders are variables that affect both the treatment assignment and the outcome. In randomized studies, treatments are randomly assigned to subjects. Random allocation ensures that there are no measured or unmeasured baseline characteristics as confounders. Figure 1.1 shows the causal diagram for a nonrandomized study with treatment assignment T, confounders $\mathbf{X}$, and outcome variable Y. For a randomized study, there is no edge from $\mathbf{X}$ to T. Estimating a causal effect in randomized studies is straightforward because any covariates that might influence the outcome can be assumed to have the same distribution across different treatment groups. Therefore, any difference in the outcome variable across the groups can be attributed to the different treatments.

There is increasing interest in observational studies to draw causal inference when the treatment assignment is not random. In non-randomized studies, there may be confounders

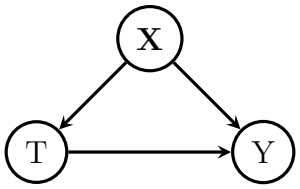

Figure 1.1: Causal Diagram with Confounders

that impact both the outcome and the treatment assignment. Baseline characteristics of treated subjects can differ greatly from the characteristics of untreated subjects. Ignoring these confounders may lead to biased estimation of the causal effects.

In this thesis, we work within the potential outcomes framework, also known as the Neyman-Rubin Causal Model, for estimating causal effects (Rubin, 1974; Splawa-Neyman et al., 1990; Holland et al., 1985; Imbens and Rubin, 2015). A potential outcome is the hypothetical value of the outcome variable for a subject under a specific treatment (Rubin, 2008). If we use i to index individuals and Y_i for the outcome of the ith subject, each individual has a pair of potential outcomes under the binary treatment setting. We use $Y_i(1)$ to denote the potential outcome under treatment, while $Y_i(0)$ denotes the potential outcome under the control.The control here refers to no treatment assignment throughout the thesis. The individual causal effect can be defined as $Y_i(1) - Y_i(0)$. Estimation of the individual causal effect is not practical since we can never observe both $\{Y_i(1), Y_i(0)\}$ simultaneously. If the treatment is applied to subject i, $Y_i(1)$ is observed and $Y_i(0)$ is sometimes called the counterfactual outcome. On the other hand, if the treatment is not applied to subject i, $Y_i(1)$ will be the counterfactual outcome and $Y_i(0)$ will be observed. The causal quantity we are interested in estimating is often the average causal effect, which is defined as the difference in the potential outcomes within the same subject averaged over a given population: $\mathrm{E}\{Y(1) - Y(0)\}$. The causal risk ratio ($\mathrm{E}\{Y(1)\}/\mathrm{E}\{Y(0)\} = P\{Y(1) = 1\}/P\{Y(0) = 1\}$) is also an interesting quantity when we want to study the incident proportion ratio between treatment and control groups.

Standard assumptions are essential under the Neyman-Rubin Causal Model to obtain consistent causal effect in observational studies (Rosenbaum and Rubin, 1983; Cole and Frangakis, 2009). Four key assumptions are given below:

- **Consistency**: A subject with an assigned treatment equal to t has observed outcome Y equal to its potential outcome $Y(t)$: $Y_i = T_i \times Y_i(1) + (1 - T_i) \times Y_i(0)$.
- **Strongly Ignorable Treatment Assignment**: The treatment T assigned is independent of the counterfactual outcomes given the observed characteristics:

$$\{Y_i(1), Y_i(0)\} \perp\!\!\!\perp T_i | \mathbf{X}_i.$$

- **Stable Unit Treatment Value Assumption (SUTVA)**: Each subject's potential outcomes should be unaffected by the actual treatment assignment of another subject.
- **Positivity**: $0 < P(T_i = 1 | \mathbf{X}_i = \mathbf{x}) < 1$.

The consistency assumption can be problematic when there are different versions of treatment (Hernan and Vander Weele, 2011). For example, we want to study the average effect between the heart transplant (treatment) and medical therapy (control) on patient's 5 year mortality. The doctor may conduct the heart transplant in different pre-operative procedures. The average effect of heart transplant in a study where the doctors used a traditional pre-operative procedure may differ from that in another study where doctors tried a novel pre-operative procedure. The treatment is not well defined in this case.

The strongly ignorable treatment assignment assumption is also known as the no unmeasured confounders assumption. It implies that conditional on baseline characteristics the treatment assignment is independent of the set of potential outcomes. If there is no unmeasured confounders, one can still obtain unbiased estimates of causal effects (Robins et al., 2000). The no unmeasured confounders assumption is not testable since $Y_i(0)$ is not observed if treatment is assigned or $Y_i(1)$ is not observed if no treatment is assigned. However, assessing the assumption is feasible (Imbens and Rubin, 2015).

The SUTVA excludes the possibilities of units interfering with each other and multiple versions of a treatment but there are circumstances in which it is not credible. For example, it may be violated in a family study about some infectious diseases such as flu. If one family member receives a vaccine, it may affect the others' change of getting flu.

The positivity assumption implies that each subject has a non-zero probability of receiving either treatment or control. If a particular subpopulation has zero probability of

being assigned to the treatment, the estimation of causal effects will be based on extrapolation and we may have to exclude such subpopulation from analysis (Imbens and Rubin, 2015).

1.2 Causal Inference with Propensity Scores

The propensity score is defined as the conditional probability of assignment to the treatment group given the observed covariates (Rosenbaum and Rubin, 1983): $e(\mathbf{x}) = P(T = 1|\mathbf{X} = \mathbf{x})$. Rosenbaum and Rubin (1983) also showed that, under the strongly ignorable treatment assignment assumption,

$$\{Y_i(1), Y_i(0)\} \perp\!\!\!\perp T_i | e_i(\mathbf{X})$$

which means, given $e(\mathbf{X})$, the distributions of potential outcomes are independent of the treatment assignment. It plays an important role in estimating the causal effect in observational studies (Rosenbaum and Rubin, 1983).

There are many approaches to estimate the average causal effect using propensity scores, such as inverse probability weighting (IPW; Hirano and Imbens (2001)), stratification (Rosenbaum and Rubin, 1983), and matching (Cochran and Rubin, 1973; Rubin, 1973a,b). Causal inference based on propensity scores is usually a two step procedure. First the propensity score is estimated (for example using logistic regression). Then the causal effect is estimated by weighting the outcome using the inverse of the estimated propensity score for each subject or matching control subjects to treated subjects based on the similarity of the estimated propensity scores. These methods have been shown to be effective in estimating the average causal effect (Hirano and Imbens, 2001; Stuart, 2010).

Logistic regression has frequently been used in observational studies to estimate the propensity scores for binary treatment. Traditional logistic regression is still found to be a good choice when the propensity score model is not heavily misspecified. Alternatively, Lee et al. (2010) suggest using classification and regression trees (CART) (Breiman et al., 1984), prune CART, and ensemble methods such as: bagged CART (Breiman, 1996), random forest (Breiman, 2001), and generalized boosted model (GBM) (McCaffrey et al.,

2004) for propensity score estimation. All the above mentioned approaches are built on decision trees. Among these, GBM and random forest showed improvement over traditional parametric methods in terms of absolute bias, coverage rate, and reduction in standard error (Lee et al., 2010). A random forest is a collection of decision trees based on bootstrapped samples of the original data. For each split of the tree, a random subset of the predictors is chosen and the best split is found among the chosen predictors (Friedman et al., 2001). GBM (Friedman et al., 2001) is an iterative tree method that fits a tree model using residuals from the previous tree. Therefore, the trees are grown sequentially but not independently such as random forest.

Despite the popularity of IPW, Kang and Schafer (2007) found that the estimators based on IPW are sensitive to the misspecification of the propensity score model. Many double robust estimators (Robins et al., 1994, 1995; Rotnitzky et al., 1998; Qin et al., 2008; Tan, 2010; Fan et al., 2016) have been proposed by modelling the treatment assignment mechanism and the relationship between covariates and outcome to improve the efficiency. These methods are robust to the misspecification of one of the treatment or outcome models.

1.3 Covariate Balancing

Covariate balance means that the distributions of measured covariates for observations in the treatment group and control group are similar to each other (Harder et al., 2010). Rubin (2007) advocated mimicking randomized experiments since randomization is considered the gold standard in estimating the average causal effect. It means analysis in observational studies can be modified to derive an estimator you would have derived when you run a randomized control trial. The baseline characteristics of observational studies are used to adjust the studies such that the subgroups of treatment and control units are similar to each other. Randomized design is usually infeasible due to many reasons including but not limited to time limit, ethical concerns, and massive expenses. However, this can be accomplished by propensity score approaches, matching methods (Stuart, 2010; Iacus et al., 2012) or weighting based covariate balancing methods (Qin and Zhang, 2007; Hainmueller, 2012; Imai and Ratkovic, 2014; Hazlett, 2015; Chan et al., 2016; Wong and Chan, 2017).

Hence, achieving balance in covariates across treatment groups is one of the most important targets for propensity score based approaches to achieve the similarity of subgroups. By achieving balance in covariates, the bias in estimating the average causal effect can be reduced (Harder et al., 2010). Mean balance is defined as the equivalence of the means of the potential outcomes between the treatment and control groups (Hazlett, 2015). Under strongly ignorable treatment assignment, Hazlett (2015) showed that if the mean balance on a linear outcome model is achieved by weighting, an unbiased estimator of the average causal effect on the treated can be obtained. In this section, we give a brief introduction to the literature specifically focusing on covariate balancing.

Recently, several propensity score modelling approaches have been proposed which target achieving balance in the covariates. For example, the GBM approach estimates the propensity score through generalized boosted regression (McCaffrey et al., 2004). The number of iterations in GBM is determined by minimizing average standardized absolute mean difference such that the covariate balance is optimized. The covariate balancing propensity score method (CBPS) was introduced to model the treatment assignment using a logistic model while simultaneously optimizing the covariate balance (Imai and Ratkovic, 2014). The CBPS method estimates the parameters of the propensity score model by solving estimating equations implied by the covariate balancing property while still incorporating the standard logistic regression estimation procedure (Imai and Ratkovic, 2014). This estimation is conducted using generalized method of moments (GMM) (Hansen, 1982) or empirical likelihood (Owen, 2001). The CBPS method optimizes the balance of covariates between treatment and control groups. The advantage of CBPS is that it reduces the effect of potential misspecification of a parametric propensity score model (Imai and Ratkovic, 2014). Fan et al. (2016) discussed the theoretical properties, optimal choice of covariate balancing function for CBPS methodology and proposed a double robust and efficient improved-CBPS (iCBPS) methodology.

In addition to the commonly used matching and inverse propensity score weighting approaches in observational studies, entropy balancing was proposed by Hainmueller (2012). The entropy measure used in entropy balancing method is the measure of divergence from one distribution to another distribution. It differs from other preprocessing approaches by directly focusing on the goal of achieving covariate balance, which can help to reduce model

dependence while retaining valuable information. The entropy balancing method relies on a maximum entropy reweighting scheme that directly incorporates covariate balance into the weight function. The reweighting scheme includes a large set of balance constraints so that the covariate distributions in the treatment and control groups will match exactly on all pre-specified moments. The entropy balancing method is shown to be double robust when either the logistic propensity score model or the outcome model is linear in some functions of covariates which is also used for balancing constraints. Hazlett (2015) also proposed a kernel balancing method to find the weight vector such that the mean balancing on the kernel matrix is satisfied.

There has been some recent work focusing on using kernel-based methods to estimate causal effects. Wong and Chan (2017) proposed a kernel-based method to achieve covariate functional balance for functions of covariates in a reproducing kernel Hilbert space (RKHS). This method shows that the infinite-dimensional optimization can be transformed into finite-dimensional optimization. The true outcome regression function is assumed to lie in the RKHS. The consistency of the causal effect estimator is achieved as long as the estimation error of the outcome regression function is $o_p(1)$. In the simulation studies of Wong and Chan (2017) to compare their proposed covariate functional balancing estimator with IPW estimator, it is shown that the empirical performance of both estimators are related to the degree of covariate balancing. The IPW estimator can be very unstable without any covariate balancing. Zhao (2016) proposed estimating the propensity score using a covariate balancing scoring rule under a logistic regression model and generalized the linear predictors into different model spaces such as RKHS. Kallus (2016) proposed a kernel optimal matching method under the framework of generalized optimal matching by minimizing a bias-dual-norm imbalance metric under the RKHS norm.

1.4 Outline of the Thesis

As we introduced before, there is a rich literature on statistical analysis of causal inference. The thesis addresses the causal inference problem focusing on covariate balancing. In Chapter 2, we introduce the basic setting for estimating average causal effect or other quantities and give a more detailed review of commonly used propensity score approaches

and weighting techniques such as IPW, CBPS, entropy balancing, and Targeted Maximum Likelihood Estimation.

In Chapter 3, we introduce a new model averaging approach for propensity score estimation which combines a parametric model with a nonparametric model. The proposed methodology is similar in spirit to super learner (van der Laan et al., 2007). Given a library of candidate models, super learner uses a weighted linear combination of all candidates to build a new estimator by minimizing a particular loss function, such as the mean squared error, using cross-validation (Pirracchio et al., 2015). Our approach differs in the sense that it is a linear combination of one parametric propensity score model with one nonparametric model and the weight on either model is determined by optimizing a selected balance metric, such as the average value of absolute standardized mean difference (ASMD) of all measured covariates or the mean Kolmogorov-Smirnov (KS) test statistic.

In Chapter 4, we introduce another approach to estimate the average causal effect using kernel distance which is also based on the idea of covariate balancing. The target is to find the optimal regression coefficients in logistic regression such that the kernel distance between the distributions of covariates under treatment and control groups are zero after inverse probability weighting. Similar to the CBPS approach, we employ estimating equations to find the optimal coefficient estimates based on logistic regression. However, we aim to achieve the balance in the whole distributions of covariates between treatment and control groups not only the first or second order moments.

Often interest lies in the estimation of quantities other than the population means. The quantiles of an outcome can be a more meaningful measure in asymmetric distributions for real life problems (Zhang et al., 2012; Díaz, 2015). Several methods have been proposed to estimate quantile treatment effects (QTE) , for example, the marginal quantiles of potential outcomes and their difference (Zhang et al., 2012). In Chapter 5, we aim to incorporate the idea of covariate balancing to the estimation of the quantile treatment effect (QTE) or quantile treatment effect in the treated (QTET). We aim to achieve balance in the conditional distributions of outcomes between treated and control groups and estimate the QTE or QTET simultaneously.

Chapter 2

Propensity Score Approaches and Beyond

In this chapter, we will introduce the general setting for causal inference which will be used throughout this thesis and some propensity score approaches or weighting methods for estimating causal quantities.

2.1 (Augmented) Inverse Probability Weighting

Let $\mathbf{X} = (1, X_1, \ldots, X_p)^\intercal$ be a vector of baseline covariates. Following the definition in Chapter 1, the data are denoted as $(T_i, Y_i, \mathbf{X}_i)$, $i = 1, \ldots, n$ and $n = n_0 + n_1$ where n_1 and n_0 are the numbers of observations in treatment and control groups separately. The quantities we are interested in estimating are the average causal effect (ACE):

$$\mu = \text{ACE} = E\left\{Y(1) - Y(0)\right\}$$

and the average causal effect in the treated (ACET):

$$\text{ACET} = E\left\{Y(1) - Y(0) | T = 1\right\}.$$

Throughout the thesis, we will keep with this basic setting. When estimating the ACE, a treated unit is assigned a weight of $\widehat{w}_i = 1/\widehat{e}(\mathbf{X}_i)$ where $\widehat{e}(\mathbf{X}_i)$ is the estimated

propensity score for that unit. The weight for a control unit is $\widehat{w}_i = 1/\{1-\widehat{e}(\mathbf{X}_i)\}$. When estimating the ACET, the weight for a treated unit is 1 and the weight for a control unit is $\widehat{w}_i = \widehat{e}(\mathbf{X}_i)/\{1-\widehat{e}(\mathbf{X}_i)\}$. The Horvitz-Thompson estimator (Horvitz and Thompson, 1952) based on the inverse probability weighting is formulated to estimate the ACE (Robins et al., 1994),

$$\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n}\left\{T_i Y_i \widehat{w}_i - (1-T_i)Y_i\widehat{w}_i\right\}.$$

The effect of weighting by inverse probability weighting (IPW) is to create a new pseudo-population with w_i copies of each subject so that treatment T is not confounded with covariates $\mathbf{X}$. However, an estimator based on augmented inverse probability weighting (AIPW) can achieve full efficiency compared to an estimator based on IPW if both treatment assignment mechanism and outcome regression model are correctly specified. Throughout this section, let $\mu_1 = \mathrm{E}\{Y(1)\}$ and $\mu_0 = \mathrm{E}\{Y(0)\}$, then $\mu = \mu_1 - u_0$. We also define the conditional expectation of potential outcome pairs:

$$\mu_{1|\mathbf{X}} = \mathrm{E}\{Y(1)|\mathbf{X}\},$$
$$\mu_{0|\mathbf{X}} = \mathrm{E}\{Y(0)|\mathbf{X}\}.$$

The AIPW estimator of ACE in Robins et al. (1994) is defined as:

$$\widehat{\mu}_R = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{T_i\{Y_i - \widehat{\mu}_{1|\mathbf{X}_i}\}}{\widehat{e}(\mathbf{X}_i)} - \frac{(1-T_i)\{Y_i - \widehat{\mu}_{0|\mathbf{X}_i}\}}{1-\widehat{e}(\mathbf{X}_i)} + \widehat{\mu}_{1|\mathbf{X}_i} - \widehat{\mu}_{0|\mathbf{X}_i}\right],$$

where $\widehat{\mu}_{1|\mathbf{X}_i}$ and $\widehat{\mu}_{0|\mathbf{X}_i}$ can be estimated by regressing the observed outcomes on covariates in treatment or control groups respectively.

Similar to Robins et al. (1994), Qin and Zhang (2007) proposed an empirical likelihood based estimator for the population mean in missing data problems by maximizing the biased sampling likelihood subject to covariate moment constraints. Under the assumptions given in Section 1.1, the estimation of the ACE can be treated as a two-sample missing data problem and the estimator for μ_1 has the following form:

$$\widehat{\mu}_{1,EL} = \frac{1}{n_1}\sum_{i=1}^{n}\frac{\widehat{\theta}\hat{e}^{-1}(\mathbf{X}_i)}{1+\lambda^{\intercal} r_1(\mathbf{X}_i, \widehat{\boldsymbol{\beta}}, \widehat{\theta}, \widehat{\mathbf{a}}_1)} T_i Y_i,$$

where $\widehat{\mathbf{a}}_1 = \sum_{i=1}^n \mathbf{a}_1(\mathbf{X}_i)$, $\mathbf{a}_1 = \mathrm{E}\{\mathbf{a}_1(\mathbf{X})\}$, $r_1(\mathbf{X}, \boldsymbol{\beta}, \theta, \mathbf{a}_1) = \begin{bmatrix} 1 - \theta e^{-1}(\mathbf{X}) \\ e^{-1}(\mathbf{X})\{\mathbf{a}_1(\mathbf{X}) - \mathbf{a}_1\} \end{bmatrix}$. Here $\mathbf{a}_1(\mathbf{X})$ is a user specified vector function of covariates. $\widehat{\lambda}$ is the Lagrange multiplier determined by

$$\sum_{i=1}^n \frac{T_i r_1(\mathbf{X}_i, \widehat{\boldsymbol{\beta}}, \widehat{\theta}, \widehat{a}_1)}{1 + \lambda^\intercal r_1(\mathbf{X}_i, \widehat{\boldsymbol{\beta}}, \widehat{\theta}, \widehat{\mathbf{a}}_1)} = 0.$$

$\widehat{\mu}_{0,EL}$ can be obtained in a similar way. The empirical likelihood based estimator also has the same double robust property as the AIPW estimator proposed by Robins et al. (1994). The double robust estimator will be consistent if either the propensity score model or the outcome model is correctly specified (Scharfstein et al., 1999). Moreover, if the true outcome regression model lies in the space spanned by some known functions which are also incorporated into the above $\mathbf{a}_1(\mathbf{X})$, the empirical likelihood based estimator can also achieve full efficiency. Both estimators improve the efficiency of estimators by fully incorporating available information of the outcome model.

2.2 Covariate Balancing Propensity Score

The IPW estimator is sensitive to misspecification of the propensity score model (Kang and Schafer, 2007; Zhu et al., 2014). Some methods have been proposed to address this problem including the covariate balancing propensity score (CBPS) (Tan, 2010; Hainmueller, 2012; Graham et al., 2012; Chan et al., 2016). The CBPS can also be extended to study general treatment regimes and the longitudinal analysis setting (Fong and Imai, 2014; Imai and Ratkovic, 2015). A common choice for the propensity score model is logistic regression:

$$\mathrm{logit}(\boldsymbol{X}) = \boldsymbol{X}^\intercal \boldsymbol{\beta}, \tag{2.1}$$

where $\boldsymbol{\beta} \in \Theta$ is an unknown parameter column vector. Generally, one will estimate $\boldsymbol{\beta}$ by maximizing the log-likelihood function. It can be solved from the following estimating equation based on the score vector:

$$\frac{1}{n} \sum_{i=1}^n \mathbf{S}_{\boldsymbol{\beta}}(T_i, \mathbf{X}_i) = \mathbf{0},$$

where $\mathbf{S}_{\boldsymbol{\beta}}(T_i, \mathbf{X}_i) = \{\frac{T_i}{e_{\boldsymbol{\beta}}(\mathbf{X}_i)} - \frac{1-T_i}{1-e_{\boldsymbol{\beta}}(\mathbf{X}_i)}\}e'_{\boldsymbol{\beta}}(\mathbf{X}_i)$ and $e'_{\boldsymbol{\beta}}(\mathbf{X}_i) = \partial e(\mathbf{X}_i)/\partial\boldsymbol{\beta}^\intercal$. In CBPS, Imai and Ratkovic (2014) proposed to replace $e'_{\boldsymbol{\beta}}(\mathbf{X}_i)$ by a user specified vector function $\mathbf{f}(\mathbf{X}_i)$. The CBPS method estimates $\boldsymbol{\beta}$ by solving the following m-dimensional estimation equations:

$$\bar{\mathbf{g}}_{\boldsymbol{\beta}}(T, \mathbf{X}) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{g}_{\boldsymbol{\beta}}(T_i, \mathbf{X}_i) = \mathbf{0}, \tag{2.2}$$

with $\mathbf{g}_{\boldsymbol{\beta}}(T_i, \mathbf{X}_i) = \{\frac{T_i}{e_{\boldsymbol{\beta}}(\mathbf{X}_i)} - \frac{1-T_i}{1-e_{\boldsymbol{\beta}}(\mathbf{X}_i)}\}\mathbf{f}(\mathbf{X}_i)$. Here $\mathbf{f}(\mathbf{X}_i)$ is called the covariate balancing function since Equation (2.2) can be also expressed as a covariate balancing equation:

$$\sum_{i=1}^{n} \frac{T_i}{e_{\boldsymbol{\beta}}(\mathbf{X}_i)}\mathbf{f}(\mathbf{X}_i) = \sum_{i=1}^{n} \frac{1-T_i}{1-e_{\boldsymbol{\beta}}(\mathbf{X}_i)}\mathbf{f}(\mathbf{X}_i).$$

The covariate balancing equation enables the robust and efficient estimation of ACE. If we choose the $\mathbf{f}(\mathbf{X}_i)$ to be $\{\mathbf{X}_i^\intercal, \ldots, (\mathbf{X}_i^m)^\intercal\}$, the estimating equations become the moment condition of covariates as we mentioned before. If the number of parameters equals the number of equations, it will be just-identified estimation. If the covariate balancing equation is combined with score function condition, an over-identified continuous updating generalized method of moments (GMM) estimator can be derived:

$$\widehat{\boldsymbol{\beta}}_{\text{GMM}} = \arg\min_{\boldsymbol{\beta}\in\Theta} \bar{\mathbf{g}}_{\boldsymbol{\beta}}(T, \mathbf{X})^\intercal \widehat{\mathbf{W}} \bar{\mathbf{g}}_{\boldsymbol{\beta}}(T, \mathbf{X}).$$

Here $\widehat{\mathbf{W}}$ is the $m \times m$ dimensional covariance matrix. The ACE can be estimated through IPW once $\boldsymbol{\beta}$ is estimated. The estimation can also be done through the empirical likelihood framework with the covariate balancing equation given in Equation (2.2) as constraints.

There was still one issue left unsolved in the CBPS method: what is the optimal choice of $\mathbf{f}(\mathbf{X})$? Fan et al. (2016) explored the optimal choice of the $\mathbf{f}(\mathbf{X})$ and the theoretical properties of the CBPS-based IPW estimator. If there exists an $\boldsymbol{\alpha}$ such that $\boldsymbol{\alpha}^\intercal\mathbf{f}(\mathbf{X})$ is equal to the weighted average of conditional mean functions of two potential outcomes, the CBPS-based IPW estimator is still consistent when the propensity score model is arbitrary misspecified. Once the covariate balancing function is constructed in this way, the CBPS-based IPW estimator will be semi-parametric efficient when the propensity score model is also correctly specified (Fan et al., 2016).

2.3 Entropy and Kernel Balancing

The IPW estimators including CBPS require specifying a propensity score model. Entropy balancing (Hainmueller, 2012) and kernel balancing (Hazlett, 2015) are proposed to achieve covariate balancing by a maximum entropy reweighting without estimating a propensity score model. Different from IPW approaches, entropy balancing directly focuses on covariate balancing. Hainmueller (2012) focuses on the estimation of ACET in observational studies with binary treatment. Since ACET $= E\{Y(1)|T=1\} - E\{Y(0)|T=1\}$, the first expectation can be estimated from the treatment group directly without weighting. The second expectation is a counterfactual mean and is unobserved. Weights for the control group need to be estimated. In the entropy balancing method, the weights for the control group are chosen to minimize a distance metric:

$$\min_{w_i} H(w) = \sum_{\{i|T_i=0\}} h(w_i)$$

subject to covariate balancing and normalizing constraints:

$$\sum_{\{i|T_i=0\}} w_i c_{ri}(\mathbf{X}_i) = \sum_{\{i|T_i=1\}} \frac{1}{n_1} c_{ri}(\mathbf{X}_i) \quad \text{with} \quad r = 1, \ldots, R, \tag{2.3}$$

$$\sum_{\{i|T_i=0\}} w_i = 1, \tag{2.4}$$

$$w_i \geq 0 \quad \text{for all} \quad T_i = 0, \tag{2.5}$$

where $h(\cdot)$ is a distance metric chosen from a class of empirical minimum discrepancy estimators proposed by Read and Cressie (2012). A backward Kullback-Leibler distance, also called entropy measure, which is the measure of divergence between two distributions (Kullback and Leibler, 1951), is employed (Kullback, 1959): $h(w_i) = w_i \log(w_i/q_i)$ with w_i being the control weight and q_i the base weight. A set of uniform base weights is usually utilized.

Equations (2.3) is the covariate balancing constraint with $c_{ri}(X_{ij}) = X_{ij}^r$, $r = 1, \ldots, R$. The balancing constraint aims to balance between the weighted average of the rth covariate moment in the control group and the same moment for treatment group up to the Rth moment. The covariate distributions between treatment and control group will match

exactly up to Rth order. Equation (2.4) and (2.5) are normalization constraints used by distance metrics.

The Lagrange multiplier method is used to estimate the weights subject to the constraints, then the ACET can be estimated by

$$\widehat{\text{ACET}} = \frac{1}{n}\sum_{i=1}^{n} T_i Y_i - \sum_{i=1}^{n} \{(1-T_i) Y_i \widehat{w}_i\}\,. \tag{2.6}$$

The entropy balancing searches for the optimal set of weights satisfying the covariate balance constraints while remaining as close as possible to the set of uniform base weights under the entropy measure (Hainmueller, 2012). The entropy balancing achieves a high degree of covariate balance directly by imposing up to Rth moment balancing constraints, while conventional IPW approaches still need to do balancing check after estimation (Hirano et al., 2003). Entropy balancing also has the double robust property as long as either propensity score model or outcome model is linear in $c_r(\mathbf{X})$ (Zhao and Percival, 2017).

Kernel balancing is based on a similar idea to entropy balancing, although kernel balancing relaxes the limitation on moments by imposing the kernel (Hazlett, 2015). In the entropy method, unbiasedness is assured only when the propensity score model or outcome model is linear in the pre-specified functions of observed covariates. However, kernel balancing seeks weights such that treatment and control groups have equal means on the set of bases implied by a kernel, which is proved to be a very large space of functions.

In kernel balancing, Hazlett (2015) focuses on the estimation of ACET. The conditional expectation of $Y(0)$ is assumed to be:

$$\mathrm{E}\left\{Y_i(0)|\mathbf{X}_i = \mathbf{x}\right\} = \phi(\mathbf{x})^\intercal \theta.$$

A mean balance condition on $\phi(\mathbf{X})$ similar to Equation (2.3) is derived:

$$\sum_{\{i|T_i=0\}} w_i \phi(\mathbf{X}_i) = \frac{1}{n_1} \sum_{\{i|T_i=1\}} \phi(\mathbf{X}_i).$$

Model assumptions are not required but a model space through a choice of kernel such as Gaussian kernel: $K(\mathbf{X}_i, \mathbf{X}_j) = e^{-\frac{\|\mathbf{X}_i - \mathbf{X}_j\|^2}{2\sigma^2}}$ is employed. A positive semi-definite kernel $\mathbf{K}$ matrix is defined with $K_{i,j} = K(\mathbf{X}_i, \mathbf{X}_j)$. The ith row of $\mathbf{K}$ can be written as

$\mathbf{K}_i = \{K(\mathbf{X}_i, \mathbf{X}_1), \ldots, K(\mathbf{X}_i, \mathbf{X}_n)\}$. By reordering the observations, the first n_1 rows of $\mathbf{K}$ corresponding to treated units:

$$\mathbf{K} = \begin{pmatrix} \mathbf{K_t} \\ \mathbf{K_c}, \end{pmatrix}$$

where $\mathbf{K_t}$ is $n_1 \times n$ and $\mathbf{K_c}$ is $n_0 \times n$.

Unlike pre-specified functions, the kernel function can be generalized to an inner product of infinite-dimensional eigenfunctions by Mercer's Theorem (Mercer, 1909): $K(\mathbf{X}_i, \mathbf{X}_j) = \langle \phi(\mathbf{X}_i), \phi(\mathbf{X}_i) \rangle$. Hence, the mean balance condition on $\phi(\mathbf{X})$ can be transformed to mean balance on $\mathbf{K}$:

$$\sum_{\{i|T_i=0\}} w_i \mathbf{K}_i = \frac{1}{n_1} \mathbf{K_t}^\intercal \mathbb{1}_{n_1}.$$

Here $\mathbb{1}_{n_1}$ is a uniform vector of length equal n_1. With the constraint, we can optimize the entropy measure to estimate the weights for the control group. An estimate of ACET can be obtained by Equation (2.6).

The kernel balancing finds the weights such that the weighted average rows of $\mathbf{K}_c$ is equal to the average rows of $\mathbf{K}_t$. Achieving the mean balance on the kernel matrix implies achieving the mean balance on a large set of smooth functions of covariates. This is very helpful when researchers have no knowledge about what function of covariates would drive the treatment assignment mechanism. Similar to entropy balancing, kernel balancing also avoids an iterative balance check which is a common practice in matching methods.

2.4 Targeted Maximum Likelihood Estimation

Targeted Maximum Likelihood Estimation (TMLE) is a well-established method to estimate many types of causal effects (van der Laan and Rose, 2011; van der Laan, 2014). TMLE estimates the parameter of interest in a way that reduces bias by considering the remaining parameters as nuisance parameters (van der Laan, 2014). It can be used to estimate the ACE by conditioning the remaining parameters in the likelihood function for the observed:

$$L = P(\mathbf{Y}|T, \mathbf{X})P(T|\mathbf{X})P(\mathbf{X}).$$

Define $Q(T, \mathbf{X}) \equiv \mathrm{E}(\mathbf{Y}|T, \mathbf{X})$, $g(\mathbf{X}) \equiv P(T|\mathbf{X})$. An initial Q_n^0 for $Q(T, \mathbf{X})$ can be estimated from observed data using many techniques such as: traditional regression models, machine learning approaches, and super learner (van der Laan et al., 2007). Here $g(\mathbf{X})$ is a nuisance parameter. Then Q_n^0 can be updated by

$$Q_n^1 = Q_n^0 + \epsilon h(\mathbf{X}),$$

where $h(\mathbf{X})$ is a function of nuisance parameter determined by the influence curve of the parameter of interest. We can estimate ϵ by regressing the observed outcome Y on $h(\mathbf{X})$ with intercept Q_n^0. Finally, the TMLE estimator for ACE, the parameter of interest, is given by

$$\widehat{\mu}_{\mathrm{TMLE}} = \frac{1}{n} \sum_{i=1}^{n} \{Q_n^1(1, \mathbf{X}_i) - Q_n^1(0, \mathbf{X}_i)\}.$$

The TMLE estimator is also double robust when either the nuisance parameter ($g(\mathbf{X})$) or outcome model ($Q(T, \mathbf{X})$) is correctly specified. It is also semi-parametric efficient when both are correctly specified. An R package, *tmle*, is available for the implementation of TMLE in statistical application (Gruber and van der Laan, 2011).

2.5 Summary

In this chapter, we introduce a variety of propensity score based approaches such as: IPW, AIPW, empirical likelihood based AIPW method, CBPS, iCBPS, TMLE, entropy and kernel balancing methods. In the following chapter, we will present our proposed methods based on covariate balancing.

Chapter 3

Model Averaging Approach

3.1 Proposed Approach

In the literature, model averaging or weighted estimation has been introduced in other research areas. For example, Olkin and Spiegelman (1987) proposed a semiparametric density estimator which combines the parametric maximum likelihood estimator and the nonparametric kernel estimator. In survival analysis, to estimate the hazard function, Kouassi and Singh (1997) proposed a weighted average of the parametric Weibull model and a kernel hazard estimator. Nottingham and Birch (2000) developed a model-robust quantal regression (MRQR) which combines a logistic regression model with a local linear regression. Motivated by a setting where a parametric model is insufficient to fit the entire data set, Mays et al. (2001) developed several semiparametric approaches to improve the fit, one of which combines the regression fit of ordinary least squares and the fit of a local linear regression. Finally, Zhu et al. (2016) considered combining multiple candidate models in estimating controlled direct effects and found that combining a parametric model with a nonparametric model leads to more accurate and efficient estimates.

The quantities we are interested in estimating average causal effect (ACE) and average causal effect in the treated (ACET). As we discussed in Chapter 2, there are many approaches to estimate the ACE or ACET using propensity scores, such as inverse probability weighting (IPW), stratification, and matching. Here, we focus on IPW, although

the proposed methodology can be extended to matching and stratification procedures.

We will consider two propensity score models: logistic regression (LR) and random forest (RF). The simulation by Lee et al. (2010) results show that LR and RF lead to the least biased causal estimates, among the collection of methods examined, when the propensity score models are not heavily misspecified. Therefore, averaging over LR and RF methods is a reasonable choice and both algorithms are easy to implement and computationally fast. The proposed model averaging propensity score estimate is a linear combination of the estimated propensity scores from these two models. Let $\widehat{e}_1(\mathbf{X})$ be the estimated propensity score from LR and $\widehat{e}_2(\mathbf{X})$ be the estimated propensity score from RF. Our proposed model averaging propensity score estimate is:

$$\widehat{e}_c(\mathbf{X}) = \lambda \widehat{e}_1(\mathbf{X}) + (1-\lambda)\widehat{e}_2(\mathbf{X}),$$

where λ is a mixing parameter between 0 and 1. The value of λ is chosen such that a certain balance statistic is minimized. The balance statistic is a measure on the covariates we use to assess similarity across the treatment groups. Four balance statistics from a list of balance statistics in Stuart et al. (2013) are proposed and for each criterion, the optimal λ is found via a grid searching methodology. For each balance statistic, grid searching is simply an exhaustive search where we calculate the balance statistic for each pre-specified λ value and the optimal λ is chosen such that the balance statistic is minimized. Here, the grid search runs from 0 to 1 at an increment of 0.01. We consider the following balance statistics based on absolute standardized mean difference (ASMD) and Kolmogorov-Smirnov (KS) statistic: mean ASMD, median ASMD, maximum ASMD, and mean KS statistic.

The ASMD for a covariate X_j is calculated as the absolute value of the difference in weighted means of the covariate between the treatment group and control group divided by the standard deviation of the covariate in the treated group (Stuart et al., 2013).

$$\text{ASMD}^{(j)} = |\bar{X}^w_{j,1}/sd(X_{j,1}) - \bar{X}^w_{j,0}/sd(X_{j,1})| \quad \text{for} \quad j = 1, \ldots, p.$$

Here, $\bar{X}^w_{j,1}$ is the weighted mean of the jth covariate in the treated group while $\bar{X}^w_{j,0}$ is the weighted mean in the control group,

$$\bar{X}^w_{j,0} = \frac{\sum_{i=1}^n (1-T_i) X_{ij} \widehat{w}_i}{\sum_{i=1}^n (1-T_i)\widehat{w}_i}, \quad \bar{X}^w_{j,1} = \frac{\sum_{i=1}^n T_i X_{ij} \widehat{w}_i}{\sum_{i=1}^n T_i \widehat{w}_i},$$

where the weights are based on the model averaged propensity score models and $sd(X_{j,1})$ is the sample standard deviation of X_j in the treatment group without weighting. For each continuous covariate there is a single corresponding ASMD value. The mean/median/max ASMD is the mean/median/max of all ASMD values over all covariates. For categorical covariates, there is a corresponding ASMD for each level of the covariate. To evaluate the corresponding ASMD for each level of a binary or categorical covariate, we create a binary variable indicating that the unit is equal to that level and then we calculate the ASMD of the indicator variable (Lee et al., 2010).

The KS statistic is defined as the maximum discrepancy of two empirical weighted cumulative distribution functions:

$$KS^{(j)} = \sup_x |F^w_{1,n_1}(X_{j,1}) - F^w_{0,n_0}(X_{j,0})| \quad \text{for} \quad j = 1, \ldots, p.$$

Here n_1 is the number of units in the treatment group and n_0 is the number of units in the control group. The two empirical weighted cumulative distribution functions are built based on the covariate values in the treatment group and control group separately. For categorical variables, the reported KS statistic is the difference in proportions for each level, between the treated and control groups (Ridgeway et al., 2015).

We have four model averaging or combined approaches based on the four balance statistics. After we find the optimal λ and corresponding new combined propensity scores, we estimate the weights using the IPW method and then the ACE or ACET based on the Equation (2.1) in Section 2.1. In Section 3.2, we show the consistency of $\widehat{\lambda}$ estimator, if $\widehat{\lambda}$ is chosen such that the mean KS statistic is minimized.

3.2 Theoretical Properties

Through the statement and proof of several lemmas, we will show that when the true propensity score follows the RF model and λ_0 is the true value of the mixing parameter, we have $\widehat{\lambda}_n \xrightarrow{a.s.} \lambda_0 = 0$ as $n \to \infty$ if $\hat{\lambda}_n$ is chosen such that the mean KS statistic is minimized. For simplicity, we ignore the randomness in $e_1(\cdot)$ and $e_2(\cdot)$. Given data

$(\mathbf{X}_i, T_i)$, $i = 1, \ldots, n$, the mean KS statistic can be written as:

$$G_n(\lambda) = \frac{1}{p} \sum_{j=1}^{p} g_{j,n}(\lambda),$$

where

$$g_{j,n}(\lambda) = \sup_{a \in A_j} \left| h_{j,n}(\lambda, a) \right|,$$

A_j is the support of the jth covariate and

$$h_{j,n}(\lambda, a) = \sum_{i=1}^{n} \frac{1}{n} \left[\frac{T_i I(a \geq X_{ij} T_i)}{\lambda e_1(\mathbf{X}_i) + (1-\lambda) e_2(\mathbf{X}_i)} - \frac{(1-T_i) I\{a \geq X_{ij}(1-T_i)\}}{1 - \lambda e_1(\mathbf{X}_i) - (1-\lambda) e_2(\mathbf{X}_i)} \right]$$

is the difference of two empirical weighted cumulative distribution function based on the inverse probability weights. Let $e_1(\mathbf{X}_i)$ and $e_2(\mathbf{X}_i)$ be the underlying models we assume for LR and RF respectively.

Hereafter, we assume the true propensity score model is the RF model, which means the true value of λ is zero, *i.e.*, $\lambda_0 = 0$. Hence the true propensity score is $e(\mathbf{X}) = e_2(\mathbf{X})$. We present the following theoretical properties:

Lemma 3.2.1. *Assume* $\mathbf{X} = (X_1, \ldots, X_p)$ *is a continuous random vector. We have:* $G_n(\lambda) \xrightarrow{a.s.} G(\lambda) = \frac{1}{p} \sum_{j=1}^{p} g_j(\lambda)$, *as* $n \to \infty$, *where*

$$g_j(\lambda) = \sup_{a \in A_j} |h_j(\lambda, a)|$$

and

$$h_j(\lambda, a) = \int_{A_1} \cdots \int_{(-\infty, a] \cap A_j} \cdots \int_{A_p} \frac{\lambda\{e_2(\mathbf{x}) - e_1(\mathbf{x})\}}{\{\lambda e_1(\mathbf{x}) + (1-\lambda) e_2(\mathbf{x})\}\{1 - \lambda e_1(\mathbf{x}) - (1-\lambda) e_2(\mathbf{x})\}} dF(x_1, \ldots, x_j, \ldots, x_p),$$

where $F(x_1, \ldots, x_j, \ldots, x_p)$ *is the joint cumulative distribution function of* $\mathbf{X} = (X_1, \ldots, X_p)$.

Proof. This property can be easily extended to discrete covariates. We only show the continuous case here. By the strong law of large numbers, it can be shown that $h_{j,n}(\lambda, a) \xrightarrow{a.s.} h_j(\lambda, a)$ as $n \to \infty$. Consequently by the continuous mapping theorem, $\sup_a |h_{j,n}(\lambda, a)| \xrightarrow{a.s.} \sup_a |h_j(\lambda, a)|$ as $n \to \infty$. Finally, $G_n(\lambda) \xrightarrow{a.s.} G(\lambda)$ as $n \to \infty$. □

Lemma 3.2.2. *$G(\lambda_0) = 0$ and $G(\lambda) > 0,\ \forall \lambda \neq \lambda_0$.*

Proof. This is true because there exists some $\mathbf{x}$ such that $e_1(\mathbf{x}) \neq e_2(\mathbf{x})$ which means that the integral, $h_j(\lambda, a)$, would not be zero unless $\lambda = \lambda_0 = 0$. The same conclusion applies to $G(\lambda)$. □

Theorem 3.2.3. *The minimizer of $G_n(\lambda)$, $\widehat{\lambda}_n$, will converge almost surely to λ_0 as $n \to \infty$.*

Proof. We prove this theorem by decomposing, Λ, the support of λ.
Assume $\Lambda = \{B_0, \ldots, B_k\}$, in which $\Lambda = \cup_{i=0}^{k} B_i$, $B_i \cap B_j = \emptyset$ for $i \neq j$, and $\lambda_0 \in B_0$. We also assume $G_n(B_i) = \inf_{\lambda \in B_i} G_n(\lambda)$ and $G(B_i) = \inf_{\lambda \in B_i} G(\lambda)$.

By a similar proof as in Lemma 3.2.1, we have $G_n(B_i) \xrightarrow{a.s.} G(B_i)$ as $n \to \infty$, for $i = 1, \ldots, k$ and $G_n(\lambda) \geq G_n(B_i)$. Assume $A = \{\omega : \lim_{n\to\infty} G_n(B_i) = G(B_i), \lim_{n\to\infty} G_n(\lambda_0) = G(\lambda_0), i = 1, \ldots, k\}$, then $P(A) = 1$. Therefore, $\forall \omega \in A$, $\exists N$, when $n > N$, we have $G_n(B_i) > G_n(\lambda_0)$ by Lemma 3.2.2 for $i = 1, \ldots, k$, so $\inf_{\lambda \in \cup_{i=1}^{k} B_i} G_n(\lambda) > G_n(\lambda_0)$, then we get $\widehat{\lambda}_n \in B_0$.

For any $\epsilon > 0$, if we choose $B_0 = (\lambda_0 - \epsilon, \lambda_0 + \epsilon)$ then $\widehat{\lambda}_n \in B_0$ leads to $|\widehat{\lambda}_n - \lambda_0| \leq \epsilon$, which means $\forall \omega \in A$, we also have $\lim_{n\to\infty} \widehat{\lambda}_n = \lambda_0$. Then, let $B = \{\omega : \lim_{n\to\infty} \widehat{\lambda}_n = \lambda_0\}$, we have $A \subset B$ and thus, $P(B) = 1$. Consequently, we have $\widehat{\lambda}_n \xrightarrow{a.s.} \lambda_0$ as $n \to \infty$. □

The consistency property of $\widehat{\lambda}$ ensures that the proposed IPW estimator is consistent when the LR model is misspecified but the RF is not.

3.3 Simulation Studies

3.3.1 Simulation Setup

In this section, we conduct a simulation study to compare the proposed approach to existing approaches for estimating propensity scores. The simulation setup comes from Lee et al.

(2010). The true effect of binary treatment T on continuous outcome Y is set to be -0.4. We generate ten covariates in total. Four of them (X_1, X_2, X_3, X_4) are confounders which are associated with both treatment T and the outcome variable Y. Three of them (X_8, X_9, X_{10}) are associated only with the outcome variable. The rest (X_5, X_6, X_7) are associated only with the treatment. Six of the covariates (X_1, X_3, X_5, X_6, X_8, X_9) are Bernoulli(0.5) and the others are $N(0,1)$ distributed. The covariates have a correlation structure as follows:

$$corr(X_1, X_5) = 0.2,\ corr(X_2, X_6) = 0.9,\ corr(X_3, X_8) = 0.2,\ corr(X_4, X_9) = 0.9.$$

All other correlations are set to 0. The binary treatment T is generated from a Bernoulli distribution with probability depending on the covariates:

$$\text{logit}\{P(T_i = 1|\mathbf{X}_i)\} = \boldsymbol{\beta}^\intercal f(\mathbf{X}_i). \tag{3.1}$$

Here, $\boldsymbol{\beta}$ is the corresponding coefficient vector and $f(\mathbf{X}_i)$ is a function of $\mathbf{X}_i$, the vector of covariates for unit i depending on the scenarios listed below. The continuous outcome Y_i is generated from a linear combination of T_i and $\mathbf{X_i}$,

$$Y_i = \boldsymbol{\alpha}^\intercal \mathbf{X}_i + \mu T_i + \epsilon_i,\ i = 1, \ldots, n.$$

Here, $\boldsymbol{\alpha}$ is the coefficient vector for the outcome model, μ is the treatment effect equal to -0.4 and ϵ_i has an independent $N(0,1)$ distribution.

We run simulations using seven scenarios that differ in the degree of linearity and additivity in the true treatment model (3.1). The true coefficient values for $(\boldsymbol{\beta}, \boldsymbol{\alpha}, \mu)$ and the definition of seven models follow Lee et al. (2010):

A: Additivity and linearity (main effects only);

B: Mild non-linearity (one quadratic term);

C: Moderate non-linearity (three quadratic terms);

D: Mild non-additivity (four two-way interaction terms);

E: Mild non-additivity and non-linearity (three two way interaction terms and one quadratic term);

F: Moderate non-additivity (ten two-way interaction terms);

G: Moderate non-additivity and non-linearity (ten two-way interaction terms and three quadratic terms).

The definition of seven propensity score models are given here:

$$\text{Model A:} \quad P(T=1|\mathbf{X}) = \text{expit}(\beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7),$$

$$\begin{aligned}\text{Model B:} \quad P(T=1|\mathbf{X}) = \text{expit}(&\beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 \\ &+ \beta_2 X_2^2),\end{aligned}$$

$$\begin{aligned}\text{Model C:} \quad P(T=1|\mathbf{X}) = \text{expit}(&\beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 \\ &+ \beta_2 X_2^2 + \beta_4 X_4^2 + \beta_7 X_7^2),\end{aligned}$$

$$\begin{aligned}\text{Model D:} \quad P(T=1|\mathbf{X}) = \text{expit}(&\beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 \\ &+ 0.5\beta_1 X_1 X_3 + 0.7\beta_2 X_2 X_4 + 0.5\beta_4 X_4 X_5 + 0.5\beta_5 X_5 X_6),\end{aligned}$$

$$\begin{aligned}\text{Model E:} \quad P(T=1|\mathbf{X}) = \text{expit}(&\beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 \\ &+ \beta_2 X_2^2 + 0.5\beta_1 X_1 X_3 + 0.7\beta_2 X_2 X_4 + 0.5\beta_4 X_4 X_5 + 0.5\beta_5 X_6),\end{aligned}$$

$$\begin{aligned}\text{Model F:} \quad P(T=1|\mathbf{X}) = \text{expit}(&\beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 \\ &+ 0.5\beta_1 X_1 X_3 + 0.7\beta_2 X_2 X_4 + 0.5\beta_3 X_3 X_5 + 0.7\beta_4 X_4 X_6 \\ &+ 0.5\beta_5 X_5 X_7 + 0.5\beta_1 X_1 X_6 + 0.7\beta_2 X_2 X_3 + 0.5\beta_3 X_3 X_4 \\ &+ 0.5\beta_4 X_4 X_5 + 0.5\beta_5 X_5 X_6),\end{aligned}$$

$$\begin{aligned}\text{Model G:}\quad P(T=1|\mathbf{X}) = \text{expit}(&\beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 \\ &+ \beta_2 X_2^2 + \beta_4 X_4^4 + \beta_7 X_7^2 + 0.5\beta_1 X_1 X_3 + 0.7\beta_2 X_2 X_4 + 0.5\beta_3 X_3 X_5 \\ &+ 0.7\beta_4 X_4 X_6 + 0.5\beta_5 X_5 X_7 + 0.5\beta_1 X_1 X_6 + 0.7\beta_2 X_2 X_3 \\ &+ 0.5\beta_3 X_3 X_4 + 0.5\beta_4 X_4 X_5 + 0.5\beta_5 X_5 X_6).\end{aligned}$$

where $\beta_1 = 0.8$, $\beta_2 = -0.25$, $\beta_3 = 0.6$, $\beta_4 = -0.4$, $\beta_5 = -0.8$, $\beta_6 = -0.5$, and $\beta_7 = 0.7$.

The outcome model is:

$$Y = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_3 + \alpha_4 X_4 + \alpha_5 X_8 + \alpha_6 X_9 + \alpha_7 X_{10} + \mu T + \epsilon,$$

where $\alpha_0 = -3.85$, $\alpha_1 = 0.3$, $\alpha_2 = -0.36$, $\alpha_3 = -0.73$, $\alpha_4 = -0.2$, $\alpha_5 = 0.71$, $\alpha_6 = -0.19$, and $\alpha_7 = 0.26$. μ is -0.4.

The LR model used to estimate propensity scores is the standard logistic regression with all main effects included. We also considered employing forward selection to include quadratic terms and two-way interaction terms in the logistic regression model but the results were not much different from the results with only the main effects model. This is because the selection biases are not very large and the average propensity scores are around 0.5 in all scenarios. The RF propensity score model is implemented using the *randomForest* package in R (Liaw and Wiener, 2002) with default settings. The default number of trees generated is 500. Here, covariate balancing propensity score (CBPS) approach and generalized boosted model (GBM) are conducted using the *CBPS* (Fong et al., 2014) and *gbm* (Ridgeway, 2015) packages respectively. The data are simulated with sizes $n = 100$, 1000, and 5000; 1000 data sets are generated for each scenario. For the proposed approach we find the optimal λ using each of the four proposed balance statistics. After the propensity scores are estimated, we employ IPW to estimate the ACE.

3.3.2 Performance Metrics

We evaluate and compare the performance of the four model averaging approaches with LR, RF, CBPS, and GBM using the following metrics:

1. Mean of absolute biases in percentage: the average absolute difference of the estimated ACE from the true treatment effect of $\mu = -0.4$ based on all data sets, divided by the true treatment effect,

$$\frac{\sum_{i=1}^{1000} \left| \widehat{\text{ACE}}_i - \mu \right|}{1000\,|\mu|}.$$

2. Empirical standard error: the standard deviation of 1000 treatment effect estimates.

3. Absolute bias of average ACE: the absolute value of the bias of the mean ACE, denoted as:

$$\left| \frac{1}{1000} \sum_{i=1}^{1000} \widehat{\text{ACE}}_i - \mu \right|.$$

4. Mean squared error (MSE): the average of the squared difference between the estimated ACE and its mean value.

5. Average λ: the average of all λ selected based on the model averaging method across the 1000 replications.

6. Coverage rate: the percentage of the 1000 confidence intervals that include the true treatment effect. The confidence interval for each data set is calculated as $\widehat{\text{ACE}}_i \pm 1.96 \times \widehat{\text{SE}}_i$, where $\widehat{\text{SE}}_i$ is the theoretical standard error calculated using the sandwich formula using the *survey* package (Lumley, 2004).

7. Weights: we also compare the inverse probability weights of ten simulated data sets under different methods.

3.3.3 Results

From scenarios A to G, the degree of misspecification in the logistic regression model increases so we expect that the model averaging methods will eventually outperform the other existing propensity score approaches on their own.

Table 3.1 gives the results of the simulation studies for the estimation of ACE at a sample size $n = 1000$. For the proposed model averaging method, C1 refers to use of

mean ASMD as the balance statistic, C2 to median ASMD as the balance statistic, C3 to maximum ASMD as the balance statistic, and C4 to mean KS statistic as the balance statistic.

The CBPS used here is based on the just-identified condition and the two-step GMM estimation approach (Imai and Ratkovic, 2014). We also tried fitting the CBPS model with a continuous updating GMM estimator and found the results to be similar. Incorporating second order and higher order terms in the balance condition may improve the performance of CBPS. The GBM does not show much improvement in estimating ACE but it does lead to better performance when estimating ACET in the simulation results in Lee et al. (2010).

From scenarios A to G, the model averaging methods result in smallest mean of absolute biases in percentage across all scenarios followed by LR, RF, and CBPS. Particularly in scenario G, the mean bias percentage of LR is about twice the size of C4. The GBM method has the largest mean of absolute bias compared to other approaches considered. This shows that the model averaging methods indeed averaged out the biases from LR and RF and increased efficiency in all scenarios.

In almost all scenarios, the empirical standard errors of all model averaging approaches are less than those from traditional approaches. Figure 3.1 also shows that the distributions of the ACE estimates from model averaging methods are more centralized than other methods. The LR and RF methods have more outliers in all scenarios.

The MSE is the sum of the squared bias and the empirical variance. Among all methods examined, the model averaging approaches have the smallest MSE. The GBM method produces the largest MSE in all scenarios.

The coverage rates of 95% confidence intervals are very high for all methods in all scenarios (results not shown). The reason may be because the confidence interval is based on the theoretical standard error calculated using the sandwich formula in the *survey* package in R (Lumley, 2004) which is much larger that the empirical standard error (results not shown). This suggests that the sandwich formula is too conservative. The usage of *survey* package should be taken cautiously since a design based method is used to construct the variance estimator while our method is model based. Theoretically, it is a wrong formula to estimate the variance of average causal effect. The *survey* package treats the propensity score as known when estimating the variance which is why it can be

Table 3.1: Performance of Measures by Propensity Score Models in 1000 Simulated Data Sets with $n = 1000$ (Estimation of ACE)

Measure	Method	Scenarios						
		A	B	C	D	E	F	G
Mean of absolute biases in percentage (×100)	C1	3.44	3.41	2.11	4.16	4.18	3.89	3.80
	C2	3.44	3.42	2.19	4.20	4.22	3.84	3.84
	C3	3.45	3.42	2.06	4.18	4.21	3.87	3.32
	C4	3.45	3.44	2.34	4.08	4.20	3.61	3.11
	LR	4.00	3.81	2.74	4.94	5.04	5.34	6.61
	RF	5.22	5.06	6.01	5.81	6.08	5.69	5.88
	CBPS	5.70	5.70	3.99	6.02	6.74	5.59	5.74
	GBM	10.76	10.69	10.99	12.52	12.62	12.04	11.86
Empirical standard error (×100)	C1	1.69	1.66	1.07	2.07	2.07	1.94	1.26
	C2	1.68	1.65	1.11	2.08	2.11	1.92	1.36
	C3	1.71	1.67	1.05	2.08	2.10	1.94	1.20
	C4	1.63	1.61	1.17	1.98	1.98	1.83	1.28
	LR	2.06	1.95	1.34	2.56	2.66	2.63	1.91
	RF	2.67	2.55	3.22	2.99	3.14	2.90	3.02
	CBPS	1.84	1.86	1.85	2.06	2.06	1.98	2.00
	GBM	2.01	2.04	2.20	2.17	2.27	2.13	2.22
Absolute bias of average ACE (×100)	C1	0.39	0.37	0.09	0.28	0.41	0.40	1.35
	C2	0.41	0.42	0.10	0.30	0.42	0.33	1.36
	C3	0.37	0.35	0.01	0.27	0.39	0.39	1.10
	C4	0.59	0.56	0.24	0.55	0.77	0.03	0.91
	LR	0.04	0.06	0.51	0.20	0.12	1.17	2.55
	RF	0.01	0.31	0.55	0.09	0.0007	0.05	0.11
	CBPS	2.12	2.08	0.77	2.12	2.48	1.96	2.05
	GBM	4.29	4.24	4.35	4.99	5.04	4.80	4.72
Mean squared error ($\times 10^4$)	C1	3.00	2.88	1.15	4.35	4.47	3.92	3.41
	C2	3.00	2.91	1.24	4.44	4.63	3.79	3.71
	C3	3.06	2.92	1.11	4.42	4.57	3.92	2.65
	C4	3.00	2.90	1.43	4.22	4.49	3.36	2.45
	LR	4.24	3.80	2.06	6.60	7.09	8.27	10.15
	RF	7.13	6.60	10.66	8.95	9.84	8.39	9.10
	CBPS	7.89	7.80	4.03	8.74	10.38	7.72	8.20
	GBM	22.45	22.17	23.74	29.63	30.52	27.60	27.26
Average λ (×100)	C1	94.23	95.03	95.49	94.10	94.79	92.68	93.21
	C2	92.45	93.31	95.30	92.00	93.61	90.66	92.83
	C3	93.69	94.67	94.26	93.66	94.30	91.73	90.74
	C4	85.58	88.84	85.91	88.39	88.48	85.23	88.57

In each cell, the numbers are multiplied by 100, except for mean squared error, the numbers are multiplied by 10000. C1: Model averaging method with mean ASMD ; C2: Model averaging method with median ASMD; C3: Model averaging method with max ASMD; C4: Model averaging method with mean KS statistic; LR: Logistic regression; RF: Random forest; CBPS: Covariate balancing propensity score; GBM: Generalized boosted model; CI: Confidence interval.

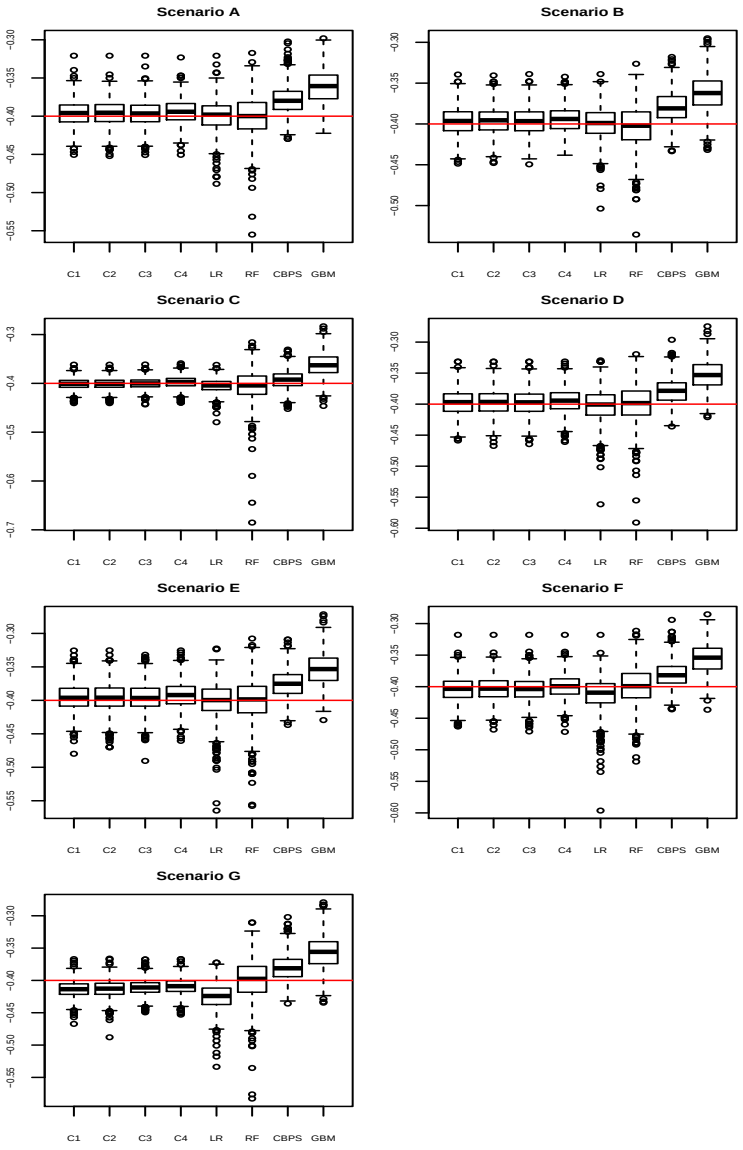

Figure 3.1: Boxplots of ACE Estimates across All Simulation Scenarios ($n = 1000$ units, 1000 replications)

C1: Model averaging method with mean ASMD as balance statistic; C2: Model averaging method with median ASMD as balance statistic; C3: Model averaging method with max ASMD as balance statistic; C4: Model averaging method with mean KS statistic as balance statistic; LR: Logistic regression method; RF: Random forest method; CBPS: Covariate balancing propensity score; GBM: Generalized Boosted Model.

too conservative in this case. The use of the bootstrap approach instead of the sandwich formula may yield more reliable confidence intervals. This is the approach we take in Section 3.4.

We also record the absolute bias of the average ACEs over 1000 replications. Among all approaches, the proposed model averaging methods reduced the bias in scenarios F & G, compared to LR, CBPS, and GBM. The RF method produces smallest biases but its empirical standard errors are the largest.

Figure 3.2 shows the distributions of λ for C1, C2, C3, and C4 methods from scenario A to G. Figure 3.3 shows the distributions of propensity score weights under different methods for randomly selected 10 data sets with $n = 1000$. Heavy or extreme weights can lead to biased and highly variable estimates of the ACE. Compared to LR and RF methods, the model averaging methods have fewer heavy weights especially in more complex scenarios. The GBM also has fewer extreme heavy weights.

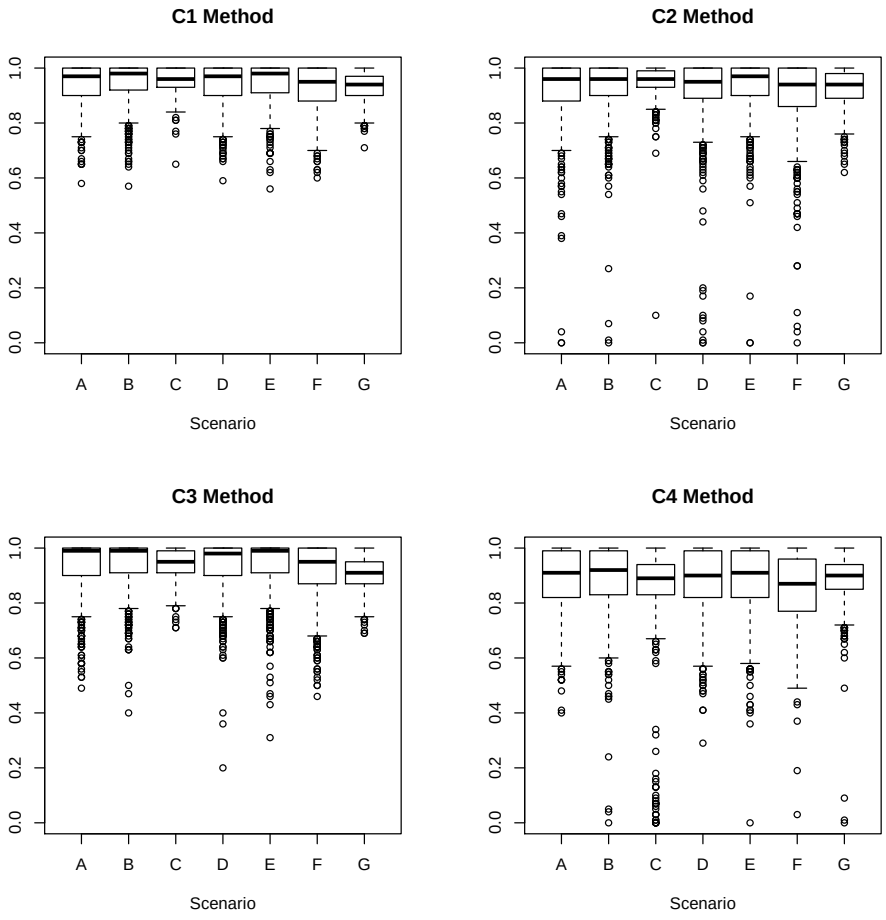

Figure 3.2: Boxplots of Estimated λ for the Four Model Averaging Methods across All Simulation Scenarios ($n = 1000$ units, 1000 replications)

C1: Model averaging method with mean ASMD as balance statistic; C2: Model averaging method with median ASMD as balance statistic; C3: Model averaging method with max ASMD as balance statistic; C4: Model averaging method with mean KS statistic as balance statistic.

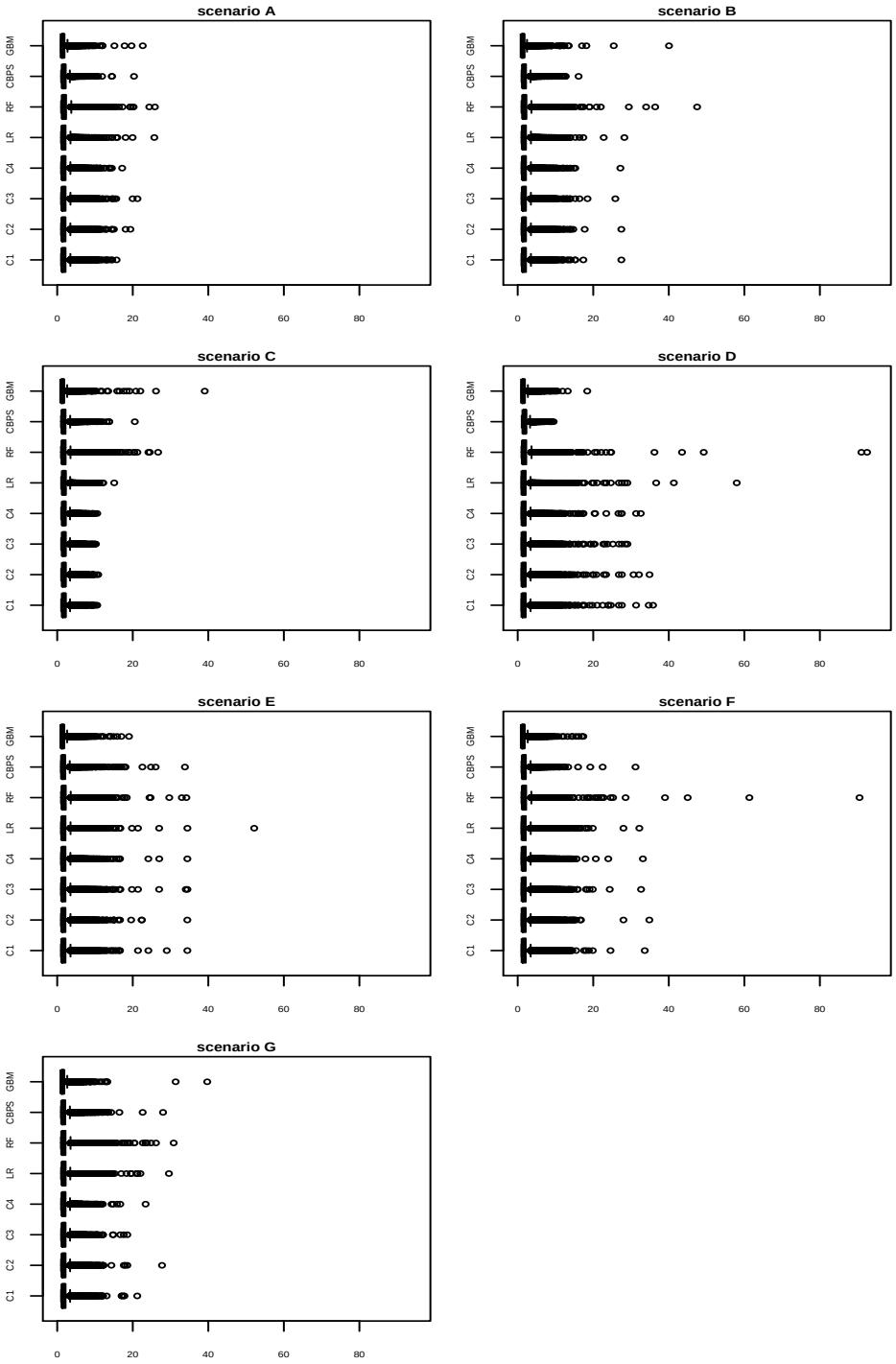

Figure 3.3: Distribution of Inverse Probability Weights for 10 Random Datasets of size $n = 1000$.

Comparison Across Proposed Model Averaging Methods

Simulation results for $n = 100$ (Table A.1) and $n = 5000$ (Table A.2) are given in the Appendix A.1.1. From Table A.1, we see that all model averaging methods have considerable lower mean of absolute biases, lower empirical standard error compared to LR, RF, GBM, and comparable empirical standard error with CBPS. When $n = 5000$, C4 methods have considerable smaller mean of absolute biases compared to LR, RF, GBM, and CBPS in scenarios D, F, and G. The C1, C2, C3, and GBM methods also have smaller empirical standard error than all other methods in all scenarios.

All model averaging methods perform similarly to each other. The C4 method based on the KS statistic is preferable to the other three combined methods. It has the smallest mean of absolute biases in most scenarios as shown in Tables 3.1, A.1, and A.2. In scenarios F and G, C4 also leads to the smallest mean biases. From Figure 3.1, we find all model averaging methods are more concentrated around the true treatment effect ($\mu = -0.4$), among which, C4 has the smallest variance and fewest outliers. In terms of MSE, the C4 method performs the best especially in scenarios F and G where the true propensity score model is highly nonlinear. As we know, the KS statistic measures the maximum discrepancy in the whole distribution between the treatment and the control group and thus is a better representative of covariate balance than ASMD which measures the difference in the first moment only. So the C4 method tends to average out more bias and when there is model misspecification in the LR model, more weight is placed on RF method.

3.4 Application

Childhood obesity has become an important health problem around the world and may lead to other severe obesity related diseases during adult years (Speiser et al., 2005). Recently, research has shown some evidence to support the benefit of early breastfeeding on reducing childhood obesity (Laurence et al., 2004; Ehrenthal et al., 2016). In this section, we apply our model averaging approach to data drawn from the Delaware Mother Baby Cohort (DMBC). Details of the study design have been reported elsewhere (Ehrenthal

et al., 2013). Data are available on 2232 mother-child pairs with 15 maternal and infant characteristics. There is no missing data. The treatment of interest is formula or mixed feeding versus exclusive breastfeeding in the first month after the child was born. There are two outcome variables in the data set: child's body mass index (BMI) Z-score at age 4 and a binary indicator that the child is obese at age 4. The remaining 12 covariates include mother's BMI group, participation in the Women, Infants, and Children program, whether the child is first birth or not, child's gestational age, child's birth weight, mother's age at delivery, mother's race, mother's education level, mother's citizenship, medicaid or uninsured, mother's marriage status, and an indicator of maternal smoking.

In preliminary analyses, we apply LR and RF methods to estimate propensity scores. A forward selection is applied within the LR method to select the most important main effects, two-way interactions, and quadratic terms. For the RF model, we set the number of trees to be 5000 and search for the optimal number of randomly selected covariates at each split of the tree. The propensity scores from model averaging approaches are the combination of propensity scores from LR with forward selection and RF method with "optimal" mixing parameters. We found that some observations had an estimated propensity score equal to zero or very close to zero in the treatment group. For those observations, infinite or large weights will be created by IPW. The large weights have great impact on results and make the estimated causal effect unstable. We used the weight trimming approach discussed by Lee et al. (2011) to improve the accuracy and performance on all methods and weighted combination of propensity scores: weights over 20 were set to 20. We calculated the ASMD values and estimated the ACE using the proposed model averaging approaches, as well as the LR and RF methods. We employed a bootstrap approach with a reproduction size of 500 to calculate the standard error and report the 90% pseudo-empirical likelihood ratio confidence interval based on asymptotic scaled χ^2 approximation with the bootstrap procedure. Details of the empirical likelihood approach for obtaining the confidence intervals can be found in Wu (2005) and Wu and Rao (2010). The theoretical justification for this method is based on the true inclusion probability. When the propensity scores are estimated by machine learning methods, there is no theoretical justification in the literature. Since we re-estimate the propensity scores within each bootstrap sample, the re-estimation usually captures the induced variation and the bootstrap procedure is a good approximation. The results are summarized in Tables

Table 3.2: Analysis of Breastfeeding Data with Continuous Outcome (Estimation of ACE, Samples=500)

Method	ACE	SE	90% PELR CI	Selected λ
C1	0.1336	0.0753	(0.0130, 0.2538)	0.99
C2	0.1336	0.0741	(0.0153, 0.2515)	0.99
C3	0.1357	0.0743	(0.0130, 0.2578)	0.86
C4	0.1345	0.0753	(0.0125, 0.2560)	0.78
LR	0.1331	0.0774	(0.0065, 0.2593)	1
RF	0.1173	0.0784	(-0.0271, 0.2607)	0

ACE: Average causal effect of formula or mixed feeding versus exclusive breastfeeding on a child's BMI Z-score at age 4;
SE: Standard error;
CI: Confidence interval;
PELR CI: Pseudo-Empirical Likelihood Ratio Confidence Interval of ACE;
SE and CI are based on 500 bootstrap samples.

3.2 and 3.3.

From Table 3.2, all the model averaging methods have positive estimates of ACE. Compared to bootstrap standard errors from LR and RF, the model averaging methods all result in slightly smaller standard errors. The ACE estimate based on C4 method, which lead to the best finite performance in the previous simulation study, is 0.1345. The 90% pseudo-empirical likelihood interval of all model averaging methods do not contain zero, and thus we can conclude that formula or mixed feeding increases BMI Z-score at a significance level of 0.1. Here, the selected mixing parameters λ are close to 1, which indicates that for this data set, the LR model outperforms the RF model. The results of the data analysis for the binary obesity outcome are given in Table 3.3. For the binary outcome, the causal effect refers to the causal risk ratio of being obese. The risk of a child being obese at age 4 is approximately 1.22 times higher if the child receives formula or mixed feeding versus exclusive breastfeeding. All 90% pseudo-empirical likelihood intervals by the model averaging approaches do not contain 1, which means formula or mixed feeding has a positive effect on obesity at a significance level of 0.1. The other results are similar

Table 3.3: Analysis of Breastfeeding Data with Binary Outcome (Estimation of RR, Samples=500)

Method	RR	SE	90% PELR CI	Selected λ
C1	1.2195	0.1953	(1.0118, 1.4717)	0.99
C2	1.2195	0.1946	(1.0135, 1.4692)	0.99
C3	1.2251	0.1945	(1.0175, 1.4769)	0.70
C4	1.2245	0.1953	(1.0150, 1.4789)	0.59
LR	1.2187	0.1962	(1.0060, 1.4781)	1
RF	1.2243	0.2464	(0.9951, 1.5085)	0

RR: Risk Ratio; ratio of the risk of being obese between formula or mixed feeding group and exclusive breastfeeding group;
SE: Bootstrap standard error of RR;
CI: Confidence interval of RR;
PELR CI: Pseudo-Empirical Likelihood Ratio Confidence Interval of RR;
SE and CI are based on 500 bootstrap samples.

to the results in Table 3.2.

Next we check the balance in each covariate after weighting. We used a cutoff value of 0.1 or 0.2 to evaluate the performance of covariate balancing by checking the ASMD (Stuart et al., 2013) and KS statistics. Figure 3.4 shows plots of weighted ASMD versus the unweighted ASMD for each model averaging approach. Unweighted ASMD is calculated by setting all weights to be 1 when calculating ASMD. As can be seen, several covariates are highly biased in the original data with ASMD > 0.2. Figure 3.5 shows plots of weighted KS statistics versus the unweighted KS statistics for each model averaging approach. The results are similar to Figure 3.4. After weighting, the selection bias has been removed and almost all the covariates have an ASMD less than 0.1, which means all proposed methods help to achieve covariate balance in ASMD and KS statistics.

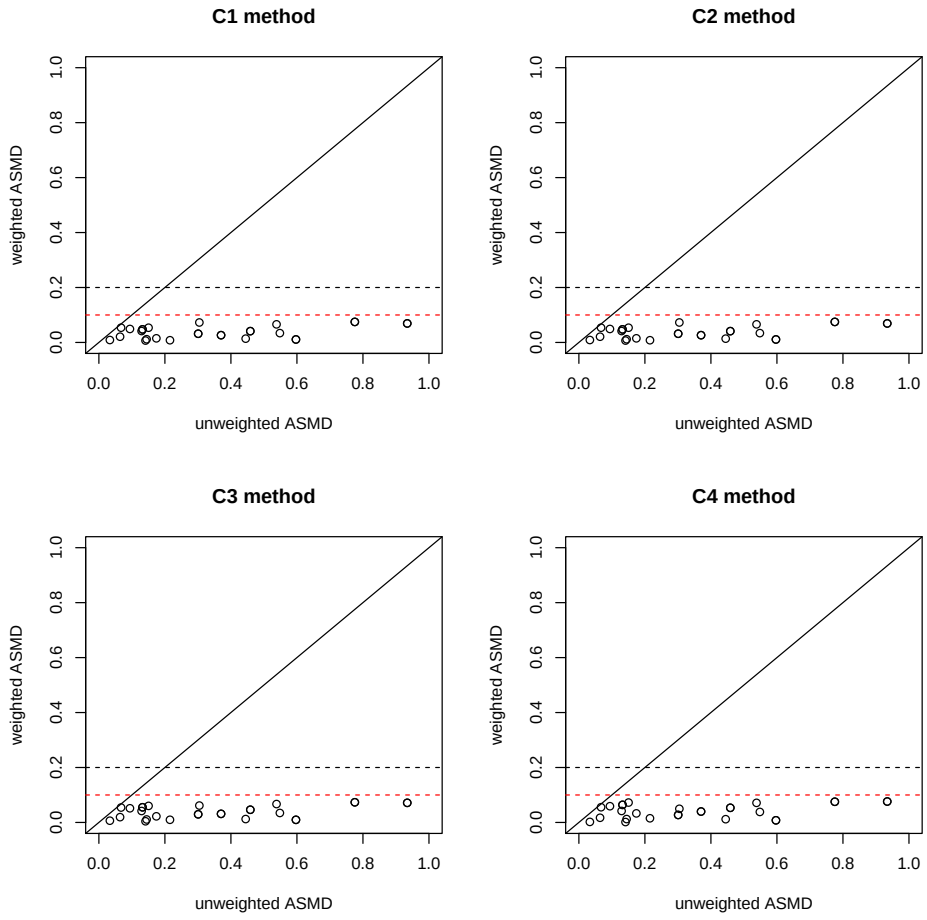

Figure 3.4: Plots of Weighted ASMD versus Unweighted ASMD (Red Line for Cut-off Value 0.1, Black Line for Cut-off Value 0.2)

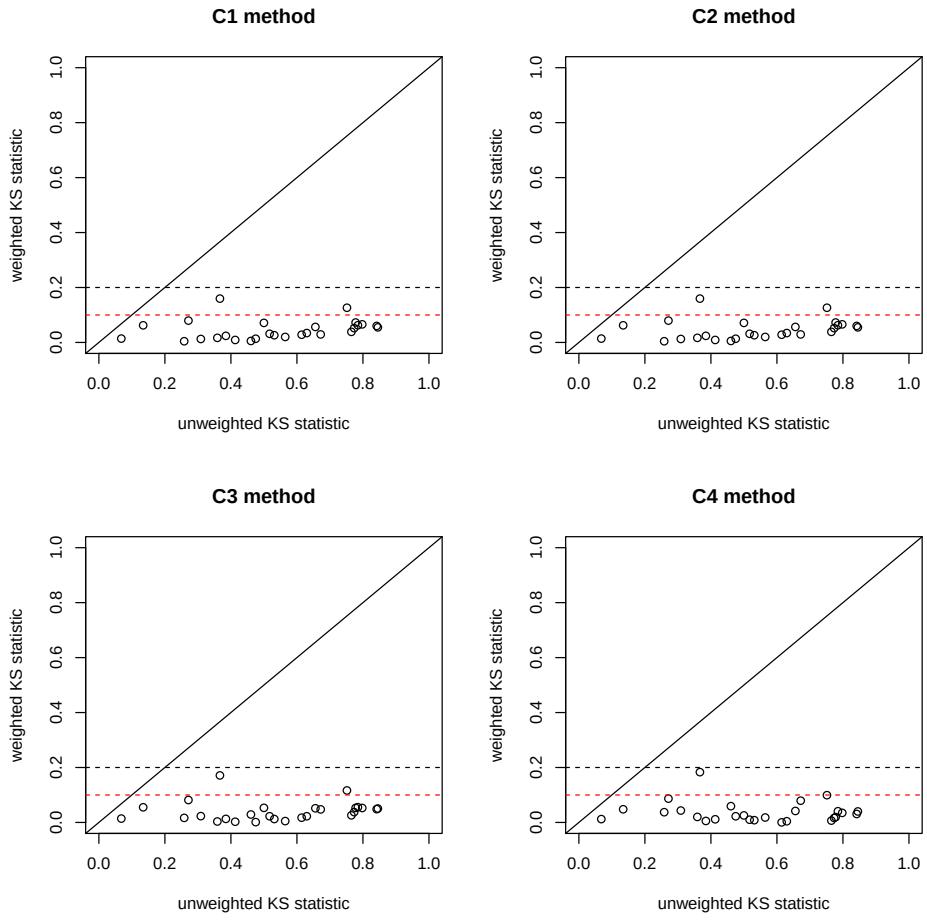

Figure 3.5: Plots of Weighted KS Statistics versus Unweighted KS Statistics (Red Line for Cut-off Value 0.1, Black Line for Cut-off Value 0.2)

3.5 Conclusion

In this chapter, we proposed a new propensity score estimator which combines LR and RF models. This approach optimizes the covariate balance through a grid search for a mixing parameter. One issue we did not explore in this chapter is what kind of covariates should be balanced. Zhu et al. (2015) showed that in addition to the real confounders, by balancing covariates that are predictive of the outcome, precision of the causal estimator can be improved. We use the same simulation setup as Lee et al. (2010) to evaluate the performance of the newly proposed model averaging approach with sample sizes $n = 100$, $n = 1000$, and $n = 5000$. With the varying degree of nonlinearity of treatment models, we find that all the model averaging methods, especially the model averaging method with the objective to minimize mean KS statistic, have smaller bias and standard error than existing propensity score modelling approaches, especially in the case when there is model misspecification in the logistic regression model. We also applied our proposed approach to a real data set of 2232 mother-child pairs. We assess the average causal effect of formula or mixed feeding versus exclusive breastfeeding in the first month on a child's BMI Z-score at age 4. Our results show that formula or mixed feeding instead of exclusive breastfeeding in the first month of life leads to higher BMI Z-scores at age 4.

There are several advantages to our proposed model averaging approach of propensity score estimation. First, compared to the traditional LR and RF methods, model averaging methods are more likely to produce estimates with smaller standard errors and biases, especially in small sample sizes (Table A.1). Second, the proposed approach also helps to reduce the imbalance in the covariates thus reducing selection bias due to measured covariates. Third, the model averaging approach can reduce the bias especially under misspecified propensity score models in which the traditional parametric approach has larger bias.

In this chapter, we used grid searching to find the optimal λ. The generalized cross-validation approach (GCV; Brookhart and van der Laan (2006)) can be implemented to improve selection of optimal λ. We can target the MSE of ACE to select the optimal λ under the GCV approach. The parametric and nonparametric models we chose to construct the model averaging approach are LR and RF. Since many machine learning techniques

have been introduced for estimating propensity scores, new model averaging approaches can be introduced by combining other parametric and nonparametric propensity score models. The parametric component can be improved by incorporating model selection techniques, like forward or backward selection, although our simulation study does not show great improvement when we employ this technique.

Our focus is on the IPW procedure. As Imbens and Rubin (2015) pointed out, IPW is not recommended when the treatment groups differ considerably in terms of covariates. It would be interesting to extend the proposed methodology to other causal inference estimation procedures, like matching based on propensity scores and double robust estimation. For example, in matching, we could choose the mixing parameter λ by optimizing the balance in the covariates for the post matched sample. In addition, instead of treating all covariates equally, one can prioritize certain covariates as is done in Genetic Matching (Diamond and Sekhon, 2013). An evolutionary search algorithm can be applied to assign the weight to each covariate and we can evaluate the balance statistics by incorporating the covariate weights. For a double robust estimator, we can plug the proposed propensity score estimator into a targeted maximum likelihood estimator (van der Laan, 2014). Then, the resulting causal effect estimator is asymptotically linear if the propensity score estimator, as well as the outcome regression estimator converge to the true value with a rate faster than $n^{-1/4}$. One advantage is that the asymptotic variance of the proposed causal estimator can be constructed using the variance of the influence function, overcoming the inference challenge faced by the IPW estimator.

Finally, to achieve the covariate balance, we focus on minimizing the ASMD or KS statistic. There exist other balance statistics that may be more appropriate and can be directly implemented in our approach. For example, a balance statistic based on prognostic score (Stuart et al., 2013) could be used. This balance statistic is shown to be more correlated with bias in treatment effect estimates, compared to ASMD and KS statistic.

Chapter 4

Kernel Distance Propensity Score Approach

In Chapter 3, we proposed a model averaging approach that reduces the imbalance of covariates between treatment and control groups based on absolute standardized mean difference (ASMD) and Kolmogorov-Smirnov (KS) statistics. Alternatively, there are many other balance measures that can be used like kernel distance (Zhu et al., 2018). Here, we propose a new propensity score approach that aims to minimize the kernel distance between the covariate distributions of treatment and control groups without assuming a functional space for the outcome regression function. We mainly use the kernel function in the definition of kernel distance to measure the distance of two probability measures, while the object function in Wong and Chan (2017) is the empirical validity measure. The method derived by Kallus (2016) differs from our method by minimizing a different imbalance metric with respect to weights directly.

A kernel function can be decomposed into an inner product of two infinite-dimensional basis functions so balancing kernel distance is not just balancing first moment or second moment. In the kernel-based method derived by Wong and Chan (2017), the true outcome regression function is assumed to lie within the reproducing kernel Hilbert space (RKHS) and is expressed in terms of kernel functions while our method optimizes the kernel distance in terms of the kernel function without assuming a functional space for the outcome

regression function.

The kernel distance comes from the integral probability metrics (IPM), which is a popular family of distance measures on probabilities (Zolotarev, 1983; Sriperumbudur et al., 2012):

$$\gamma(\mathbb{P}, \mathbb{Q}) = \sup_{f \in \mathcal{F}} \left| \int f d\mathbb{P} - \int f d\mathbb{Q} \right|, \tag{4.1}$$

where $\mathcal{F}$ is a class of functions on function space and $\mathbb{P}$ and $\mathbb{Q}$ are two probability measures. The IPMs have been used in many situations with appropriate choices of $\mathcal{F}$ including the kernel test (Gretton et al., 2012). Equation (4.1) is called kernel distance or maximum mean discrepancy when $\mathcal{F} = \{f : \|f\|_{\mathcal{K}} \leq 1\}$, where $\mathcal{K}$ represents a RKHS (Aronszajn, 1950; Gretton et al., 2012). The kernel distance has been shown to be a good measure of covariates balance (Zhu et al., 2018). A RKHS is a Hilbert space of functions with a reproducing kernel (Shawe-Taylor and Cristianini, 2004) in which its evaluation operators are bounded linear operators. The balance metric based on kernel distance has been shown to have the strongest correlation with the absolute bias in estimating the causal effect, compared to several commonly used balance metrics (Zhu et al., 2018).

4.1 Proposed Approach

Let the data $(T_i, \mathbf{X}_i, Y_i)$ be defined as in Chapter 3 and $\mathbb{P}_{n_1}$ and $\mathbb{Q}_{n_0}$ be two probability measures of covariates in the treatment and control groups. Define $\| \ \|_{\mathcal{K}}$ to be the norm for RKHS. The empirical estimator (Gretton et al., 2012; Sriperumbudur et al., 2012) of the kernel distance defined in Equation (4.1) is:

$$\widehat{\gamma}_k(\mathbb{P}_{n_1}, \mathbb{Q}_{n_0}) = \| \sum_{i=1}^{n} T_i^* K(\cdot, \mathbf{X}_i) \|_{\mathcal{K}} = \sqrt{\sum_{i=1}^{n} \sum_{j=1}^{n} T_i^* T_j^* K(\mathbf{X}_i, \mathbf{X}_j)}, \tag{4.2}$$

where $n = n_1 + n_0$ and we are reminded that n_1 and n_0 are the numbers of observations for treatment and control groups, respectively. The weight is $T_i^* = T_i/n_1 - (1 - T_i)/n_0$.

Gretton et al. (2012) showed that the convergence rate of the kernel distance estimator is free of the dimension of the covariates. It outperforms other multivariate two-sample

tests in high dimensional data settings with a lower Type II error when distinguishing two different distributions from each other. In addition, Gretton et al. (2012) claimed that kernel distance has a reasonable computation cost compared to other two-sample tests with a cost of $O(n^2)$ in Equation (4.2), free of the dimension of $\mathbf{X}$. If we rely on the inverse probability weighting (IPW) method to estimate the causal effect, we propose the weighted kernel distance with an appropriate modification of T_i^*: $T_i^* = \widehat{w}_i / \sum_{j=1}^n T_j \widehat{w}_j$ if $T_i = 1$ and $T_i^* = -\widehat{w}_i / \sum_{j=1}^n (1 - T_j)\widehat{w}_j$ if $T_i = 0$, where w_i is the inverse probability weight for subject i. It can be easily found that the expectation of $\sum_{i=1}^n T_i \widehat{w}_i$ or $\sum_{i=1}^n (1 - T_i)\widehat{w}_i$ is n if the correct propensity score model is fitted. For simplicity, we replace $\sum_{j=1}^n (1 - T_j)\widehat{w}_j$ by n so

$$T_i^* = \frac{T_i \widehat{w}_i}{n} - \frac{(1 - T_i)\widehat{w}_i}{n}.$$

The kernel function, $K(\cdot, \cdot) : \mathbb{R}^p \times \mathbb{R}^p \mapsto \mathbb{R}$, will produce a measure of distance between two vectors. In order for the method to be applied under the Generalized Method of Moments (GMM) framework, a positive definite kernel function is required. The Gaussian kernel is one of the most popular positive definite kernels:

$$K(\mathbf{X}_i, \mathbf{X}_j) = e^{-\frac{\|\mathbf{X}_i - \mathbf{X}_j\|^2}{2\sigma^2}},$$

where $\|\cdot\|$ is the l^2 norm. The σ^2 can be treated as a scale parameter. If $\mathbf{X}$ is standardized, we usually choose σ^2 to be $\dim(\mathbf{X})$. If X is not standardized, we usually choose σ^2 to be $\mathrm{E}\{||\mathbf{X}_i - \mathbf{X}_j||^2\}$ or $\mathrm{median}\{||\mathbf{X}_i - \mathbf{X}_j||^2\}$. However, Equation (4.2) is a biased estimator of kernel distance. Gretton et al. (2012) therefore suggest an unbiased estimator of squared kernel distance:

$$\widehat{\gamma}^2(\mathbb{P}_{n_1}, \mathbb{Q}_{n_0}) = \sum_{i=1}^n \sum_{j=1, j \neq i}^n T_i^* T_j^* K(\mathbf{X}_i, \mathbf{X}_j). \tag{4.3}$$

Sriperumbudur et al. (2012) state that $\widehat{\gamma}_k(\mathbb{P}_{n_1}, \mathbb{Q}_{n_0}) = 0$ if and only if $\mathbb{P}_{n_1} = \mathbb{Q}_{n_0}$. If we aim to minimize the discrepancy between the covariate distributions of the treatment and control groups after weighting, we just need to set Equation (4.3) to zero.

Similar to the covariate balancing propensity score (CBPS) method, we assume the propensity score follows a logistic regression model:

$$e(\boldsymbol{X}) = \frac{\exp(\boldsymbol{X}^\intercal \boldsymbol{\beta})}{1 + \exp(\boldsymbol{X}^\intercal \boldsymbol{\beta})}.$$

We aim to find the optimal $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_p)^\intercal$ such that $\sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} T_i^* T_j^* K(X_{i,m}, X_{j,m}) = 0$ for $m = 0, 1, \ldots, p$, where $K(X_{i,m}, X_{j,m})$ measures the discrepancy for two observations in the mth dimension. Hence, covariate balance is achieved under the optimal $\widehat{\boldsymbol{\beta}}$. We have $p+1$ unknown parameters and $p+1$ equations to solve. We use the estimating equations to find the optimal $\boldsymbol{\beta}$ under GMM.

Our target functions are:

$$G^{(m)} = \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} 2T_i^* T_j^* K(X_{i,m},\, X_{j,m}),\; m = 0, \ldots, p, \tag{4.4}$$

or in vector form when scaled by a constant $n^2/n(n-1)$:

$$\mathbf{G}_{\boldsymbol{\beta}} = \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} n^2 T_i^* T_j^* \begin{Bmatrix} K(X_{i,0},\, X_{j,0}) \\ K(X_{i,1},\, X_{j,1}) \\ \vdots \\ K(X_{i,p},\, X_{j,p}) \end{Bmatrix}. \tag{4.5}$$

There is a total $n^2 - n$ terms in the above target function. The purpose of rescaling in Equation (4.5) is to make the objective function an average. To use the GMM, we construct the new observation $\mathbf{Z}_k = (X_{i,0}, \ldots, X_{i,p}, X_{j,0}, \ldots, X_{j,p})^\intercal$ for $i > j$, and $\mathbf{T}_k = (T_i, T_j)^\intercal$ where $k = (i-2)(i-1)/2 + j$ is a 1 to 1 mapping from (i, j) to k. The $\mathbf{Z}_k$ is the kth unique combination of two observations, $k = 1, \ldots, (n^2 - n)/2$, $i, j = 1, \ldots, n$.

We define

$$\begin{aligned} q_{\boldsymbol{\beta}}(T_i,\, \mathbf{X}_i) &= nT_i^*, \\ h_{\boldsymbol{\beta}}(\mathbf{T}_k,\, \mathbf{Z}_k) &= q_{\boldsymbol{\beta}}(T_i,\, \mathbf{X}_i) q_{\boldsymbol{\beta}}(T_j,\, \mathbf{X}_j) \\ &= q_{\boldsymbol{\beta}}(\pi_3(\mathbf{T}_k),\, \pi_1(\mathbf{Z}_k)) q_{\boldsymbol{\beta}}(\pi_4(\mathbf{T}_k),\, \pi_2(\mathbf{Z}_k)), \end{aligned}$$

where $\pi_1(\mathbf{Z}_k) = \mathbf{X}_i$ and $\pi_2(\mathbf{Z}_k) = \mathbf{X}_j$ are the projection functions of the first and second groups of observations. The $\pi_3(\mathbf{T}_k)$ and $\pi_4(\mathbf{T}_k)$ are the projection functions of the first

and second component of $\mathbf{T}_k$. Here is an example:

$$\pi_1 : \mathbf{Z}_k = \begin{pmatrix} X_{i,0} \\ \vdots \\ X_{i,p} \\ X_{j,0} \\ \vdots \\ X_{j,p} \end{pmatrix} \mapsto \begin{pmatrix} X_{i,0} \\ \vdots \\ X_{i,p} \end{pmatrix}.$$

We also define

$$\begin{Bmatrix} K(X_{i,0},\, X_{j,0}) \\ K(X_{i,1},\, X_{j,1}) \\ \vdots \\ K(X_{i,p},\, X_{j,p}) \end{Bmatrix} = \begin{Bmatrix} K(f_0(\mathbf{Z}_k)) \\ K(f_1(\mathbf{Z}_k)) \\ \vdots \\ K(f_p(\mathbf{Z}_k)) \end{Bmatrix} = \mathbf{F}(\mathbf{Z}_k),$$

where,

$$f_m(\mathbf{Z}_k) : \begin{pmatrix} X_{i,0} \\ \vdots \\ X_{i,p} \\ X_{j,0} \\ \vdots \\ X_{j,p} \end{pmatrix} \mapsto \begin{pmatrix} X_{i,m} \\ X_{j,m} \end{pmatrix},\ m = 0, \ldots, p.$$

Let $(n^2 - n)/2 = N$, So Equation (4.5) becomes

$$\mathbf{G}_{n,\boldsymbol{\beta}} = \frac{1}{N} \sum_{k=1}^{N} h_{\boldsymbol{\beta}}(\mathbf{T}_k,\, \mathbf{Z}_k)\mathbf{F}(\mathbf{Z}_k). \tag{4.6}$$

We define the vector function and its mth element as:

$$\begin{aligned} \mathbf{g}_{\boldsymbol{\beta}}(\mathbf{T}_k,\, \mathbf{Z}_k) &= h_{\boldsymbol{\beta}}(\mathbf{T}_k,\, \mathbf{Z}_k)\mathbf{F}(\mathbf{Z}_k) \\ g_{k,m} &= h_{\boldsymbol{\beta}}(\mathbf{T}_k,\, \mathbf{Z}_k)K(f_m(\mathbf{Z}_k)). \end{aligned} \tag{4.7}$$

Consequently, Equation (4.6), which is also the target function, becomes

$$\mathbf{G}_{n,\boldsymbol{\beta}} = \bar{\mathbf{g}}_{\boldsymbol{\beta}}(\mathbf{T},\, \mathbf{Z}) = \frac{1}{N} \sum_{k=1}^{N} \mathbf{g}_{\boldsymbol{\beta}}(\mathbf{T}_k,\, \mathbf{Z}_k). \tag{4.8}$$

We set the target function to zero and solve for $\boldsymbol{\beta}$. The estimating equation is considered the empirical estimator of the expectation of estimating equation $E\{\mathbf{g}_{\boldsymbol{\beta}}(\mathbf{T}_k, \mathbf{Z}_k)\}$. Hence we can use the GMM framework to solve $\widehat{\boldsymbol{\beta}}$. The efficient GMM estimator is

$$\widehat{\boldsymbol{\beta}}_{\text{GMM}} = \arg\min_{\boldsymbol{\beta}\in\boldsymbol{\theta}} \bar{\mathbf{g}}_{\boldsymbol{\beta}}(\mathbf{T}, \mathbf{Z})^{\intercal}\Sigma_{\boldsymbol{\beta}}(\mathbf{T}, \mathbf{Z})^{-1}\bar{\mathbf{g}}_{\boldsymbol{\beta}}(\mathbf{T}, \mathbf{Z}), \tag{4.9}$$

where $\Sigma_{\boldsymbol{\beta}}(\mathbf{T}, \mathbf{Z})^{-1} \equiv \widehat{\mathbf{A}}$ is the inverse of variance-covariance matrix estimator. The variance-covariance matrix estimator of $\bar{\mathbf{g}}_{\boldsymbol{\beta}}(\mathbf{T}, \mathbf{Z})$ is given by

$$\Sigma_{\boldsymbol{\beta}}(\mathbf{T}, \mathbf{Z}) = \frac{1}{N}\sum_{k=1}^{N} \mathrm{E}\left\{\mathbf{g}_{\boldsymbol{\beta}}(\mathbf{T}_k, \mathbf{Z}_k)\mathbf{g}_{\boldsymbol{\beta}}(\mathbf{T}_k, \mathbf{Z}_k)^{\intercal}|\mathbf{Z}_k\right\}.$$

The efficient GMM estimator $\widehat{\boldsymbol{\beta}}_{\text{GMM}}$ is derived based on just-identified condition. The variance-covariance estimator can be ignored if we only study the just-identified case. More covariate balancing equations such as score functions can be incorporated to obtain a set of over-identified conditions. It is suggested that over-identified conditions will improve the asymptotic efficiency in the estimation of the GMM estimator (Imai and Ratkovic, 2014).

4.1.1 Remarks

In addition to using only one kernel function, combining different positive definite kernels is an efficient way to improve the precision of the kernel estimator (Gönen and Alpaydın, 2011). Many approaches have been proposed to combine several methods together to improve the performance of a particular estimator such as super learner (van der Laan et al., 2007). Zhan and Ghosh (2015) combined several kernels to improve prediction using kernel ridge regression on an outcome variable with auxiliary information. In mediation analysis, Zhu et al. (2016) proposed combining multiple candidate models like machine learning algorithms in estimating the controlled direct effects. In Xie et al. (2017), a parametric propensity score model and a nonparametric propensity score model are combined to estimate the ACE. The combined method is shown to be more accurate than using a single model. Following similar ideas, we suggest the $K(X_i, X_j)$ in Equation (4.4) can be constructed as $\sum_{m=1}^{M} \frac{1}{M} K_m(X_i, X_j)$, where we can combine several positive definite kernels,

$K_m(\cdot,\cdot)$, such as Gaussian kernel, Laplacian kernel ($\exp\{-||x_i - x_j||/\sigma\}$), exponential kernel ($\exp\{-||x_i - x_j||/(2\sigma^2)\}$) and so on. It will be an interesting topic to explore how to weight each kernel instead of averaging them. The theoretical properties remain the same for the combined kernel distance approach.

4.2 Theoretical Properties

The GMM framework requires the expectation of the estimating function to be zero. We discuss the theoretical properties including $\mathrm{E}\{\mathbf{g}_{\boldsymbol{\beta}}(\mathbf{T}_k, \mathbf{Z}_k)\} = \mathbf{0}$ in this section and Section 4.6. First, we introduce the assumptions following from Fan et al. (2016).

Assumptions 1.

(1) Weak common support condition: there exists a constant $0 < a_0 < 1/2$ *such that with probability approaching one,* $a_0 < e(\mathbf{X}_i) < 1 - a_0$.

(2) There exists a positive definite matrix $\mathbf{A}^*$ *and* $\widehat{\mathbf{A}} \xrightarrow{p} \mathbf{A}^*$.

(3) $\boldsymbol{\beta}$ *takes value from a compact set* Θ.

(4) $g_{k,m}$ *is continuous in* $\boldsymbol{\beta}$.

(5) $\boldsymbol{\beta}^*$ *is the unique minimizer of* $\mathrm{E}\{\bar{\mathbf{g}}_{\boldsymbol{\beta}}(\mathbf{T}, \mathbf{Z})\}^\intercal \mathbf{A}^* \mathrm{E}\{\bar{\mathbf{g}}_{\boldsymbol{\beta}}(\mathbf{T}, \mathbf{Z})\}$.

Under the assumptions given above, we will introduce the consistency and asymptotic normality of both $\widehat{\boldsymbol{\beta}}_{\mathrm{GMM}}$ and $\widehat{\mu}_{\widehat{\boldsymbol{\beta}}}$ in this section. For simplicity, we will drop the subscript of $\widehat{\boldsymbol{\beta}}_{\mathrm{GMM}}$ and use $\widehat{\boldsymbol{\beta}}$ instead throughout this chapter.

Theorem 4.2.1. *Under Assumptions 1 and Lemma 4.6.2 in Section 4.6, we have* $\widehat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}^*$ *as* $n \to \infty$. *Moreover, if the propensity score model is correctly specified, i.e.* $P(T = 1|\mathbf{X} = \mathbf{x}) = e_{\boldsymbol{\beta}_0}(\mathbf{x})$, *then* $\boldsymbol{\beta}_0 = \boldsymbol{\beta}^*$ *and* $\widehat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}_0$ *as* $n \to \infty$.

Theorem 4.2.1 is important and leads to the consistency of the proposed kernel estimator. The construction of Theorem 4.2.1 is also quite different from the general derivation

since the constructed random variables are not independent. It enables us to construct the consistency of the proposed kernel estimator, $\widehat{\mu}_{\widehat{\boldsymbol{\beta}}}$. The detailed proof of this theorem is given in Section 4.6.

Theorem 4.2.2 (Consistency of the ACE Estimator)**.** *Under Assumptions 1, if the propensity score model is correctly specified and $\widehat{\boldsymbol{\beta}}$ is obtained through Equation (4.9), then $\widehat{\mu}_{\widehat{\boldsymbol{\beta}}} \xrightarrow{p} \mu$ as $n \to \infty$.*

This theorem implies that our kernel based estimator is consistent when $P(T = 1|\mathbf{X} = \mathbf{x}) = e_{\boldsymbol{\beta}_0}(\mathbf{x})$. In addition to the consistency of the ACE estimator, we have the asymptotic normality of ACE estimator.

Theorem 4.2.3. *Under Assumptions 1 and the mean value theorem in calculus, we can derive that $\mathbf{G}_{n,\widehat{\boldsymbol{\beta}}}/n^2 = \mathbf{G}_{n,\boldsymbol{\beta}_0}/n^2 + \widetilde{\mathbf{B}}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$. With Lemma 4.6.3 in Section 4.6, we further conclude that*

$$n(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{d} N(\mathbf{0}, \mathbf{B}^{-1}\mathbf{V}(\mathbf{B}^{-1})^{\intercal}), \quad \text{as } n \to \infty,$$

where $\mathbf{B}$ is the limiting matrix of $\widetilde{\mathbf{B}}$, which is the derivative of $\mathbf{G}_{n,\boldsymbol{\beta}}$ evaluating at a value, $\widetilde{\boldsymbol{\beta}}$, between $\boldsymbol{\beta}_0$ and $\widehat{\boldsymbol{\beta}}$. $\mathbf{V}$ is the variance covariance matrix derived in Lemma 4.6.3.

Theorem 4.2.4. *Under Assumptions 1 and the mean value theorem, there is a vector $\widetilde{\mathbf{h}}(\widetilde{\boldsymbol{\beta}})^{\intercal}$ such that $\sqrt{n}\widehat{\mu}_{\widehat{\boldsymbol{\beta}}} = \sqrt{n}\widehat{\mu}_{\boldsymbol{\beta}_0} + \widetilde{\mathbf{h}}(\widetilde{\boldsymbol{\beta}})^{\intercal}\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$ we can further conclude that*

$$\sqrt{n}(\widehat{\mu}_{\widehat{\boldsymbol{\beta}}} - \mu) \xrightarrow{d} N(0, \Omega_\mu), \quad \text{as } n \to \infty,$$

where Ω_μ is the asymptotic variance.

The randomness of $\widehat{\mu}_{\widehat{\boldsymbol{\beta}}}$ mainly consists of randomness from $\widehat{\mu}_{\boldsymbol{\beta}_0}$ and $\widehat{\boldsymbol{\beta}}$. In Subsections 4.6.7 and 4.6.8, we prove that the randomness from $\widehat{\boldsymbol{\beta}}$ is ignorable as $n \to \infty$. The variance of $\widehat{\mu}_{\widehat{\boldsymbol{\beta}}}$ is close to the variance of $\widehat{\mu}_{\boldsymbol{\beta}_0}$, which can be estimated by the sandwich variance estimator as is commonly used in estimating equations (Lumley et al., 2004; Schafer and Kang, 2008). Hirano et al. (2003) found that IPW with the true propensity score is not efficient while IPW with their proposed estimated propensity score is semi-parametric efficient. Our proposed kernel distance propensity score estimator is not semi-parametric efficient due to the correlation of $\mathbf{g}_{\boldsymbol{\beta}}(\mathbf{T}_k, \mathbf{Z}_k)$ with $\mathbf{g}_{\boldsymbol{\beta}}(\mathbf{T}_j, \mathbf{Z}_j)$.

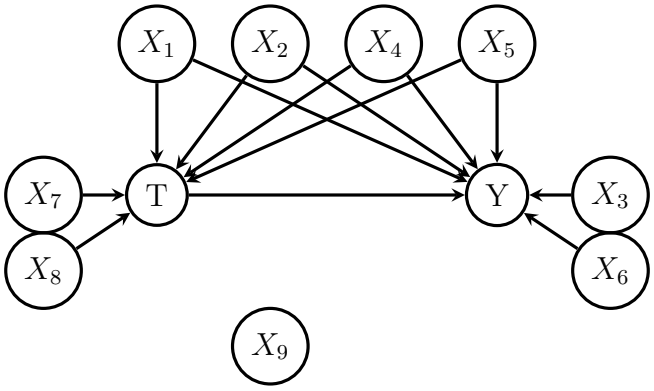

Figure 4.1: Causal Diagram among Variables in the Simulation Setup

4.3 Simulation Studies

4.3.1 Simulation Setup

In this section, we conduct a set of simulation studies to evaluate the performance of the proposed kernel distance propensity score approach compared with other existing approaches. The simulation is based on Stuart et al. (2013) with some modifications. There are nine continuous covariates: four are confounders, two are only related to the treatment indicator T, and two are only related to the outcome variable Y. The last one is neither related to treatment nor outcome (Figure 4.1). The six covariates related to the treatment indicator follow a mixture normal distribution: $1/2 \times N(-1, 1) + 1/2 \times N(1, 1)$. The others follow a $N(0, 1)$ distribution. With the mixture distribution, the imbalance between the distributions of covariates in the treatment and control groups can be considerable. The discrepancy between the empirical cumulative distribution functions of covariates of treated and control groups is shown in Figure 4.2. Although we only include continuous covariates here, our approach is not limited to this. Joshi et al. (2011) stated that the kernel distance can be used to compare discrete distributions.

We explore the performance of our estimator while varying in the degree of misspecification of both the propensity score model and the outcome model. We generate T from

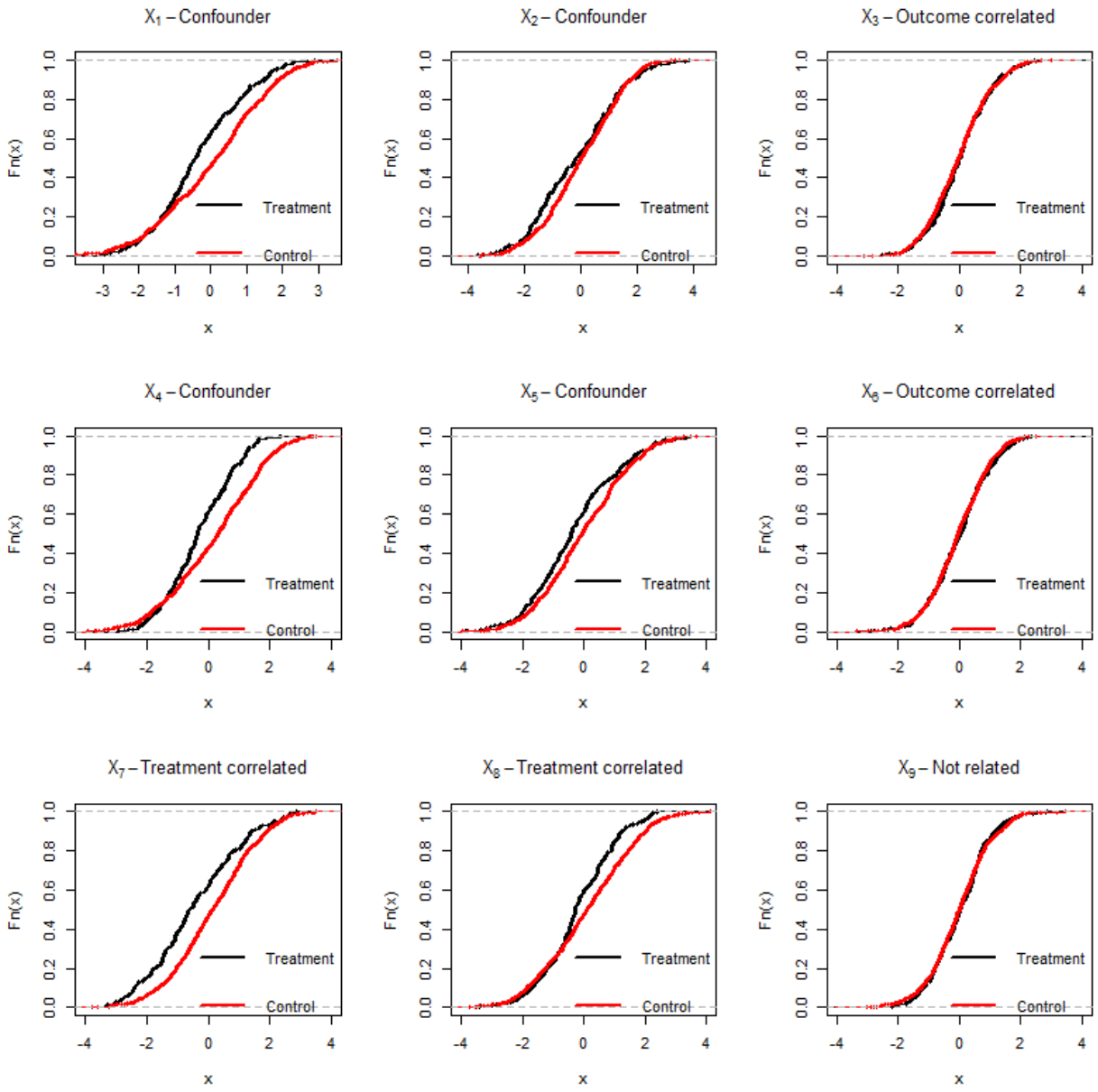

Figure 4.2: Empirical Cumulative Distribution Functions for Treatment and Control Groups across Covariates under Simulation Scenario 1A

the following propensity score model:

$$\text{logit}\{P(T=1|\mathbf{X})\} = f(\mathbf{X})^\intercal \boldsymbol{\beta}.$$

We also generate Y from the following outcome model:

$$Y = \phi(\mathbf{X})^\intercal \boldsymbol{\alpha} + \mu T + \epsilon,$$

where $\epsilon \sim N(0,1)$ independently.

The true treatment effect is $\mu = 3$. $f(\mathbf{X})$ and $\phi(\mathbf{X})$ depend on the propensity score and outcome models used. We have three propensity score models used in generating the data.

$$\text{Model 1:} \quad \text{logit}\{P(T=1|\mathbf{X})\} = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_4 + \beta_4 X_5 + \beta_5 X_7 + \beta_6 X_8,$$

$$\begin{aligned}\text{Model 2:} \quad \text{logit}\{P(T=1|\mathbf{X})\} &= \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_4 + \beta_4 X_5 + \beta_5 X_7 + \beta_6 X_8 + \beta_7 X_2 X_4 \\ &\quad + \beta_8 X_2 X_7 + \beta_9 X_7 X_8 + \beta_{10} X_4 X_5 + \beta_{11} X_1^2 + \beta_{12} X_7^2,\end{aligned}$$

$$\begin{aligned}\text{Model 3:} \quad \text{logit}\{P(T=1|\mathbf{X})\} &= \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_4 + \beta_4 X_5 + \beta_5 X_7 + \beta_6 X_8 + \beta_7 X_2 X_4 \\ &\quad + \beta_8 X_2 X_7 + \beta_9 X_7 X_8 + \beta_{10} X_4 X_5 + \beta_{11} X_1^2 + \beta_{12} X_4^2 + \beta_2 X_8^2 \\ &\quad + \frac{\beta_7}{2} X_1 X_2 + \frac{\beta_8}{2} X_1 X_4 + \frac{\beta_9}{2} X_1 X_5 + \frac{\beta_1}{2} X_1 X_7 + \frac{\beta_7}{2} X_1 X_8 \\ &\quad + \frac{\beta_8}{2} X_2 X_4 + \frac{\beta_9}{2} X_2 X_5 + \frac{\beta_1}{2} X_2 X_8 + \frac{\beta_7}{2} X_4 X_5,\end{aligned}$$

where the coefficients are shown in Table 4.1.

Table 4.1: Coefficients for Propensity Score Models

β_1	β_2	β_3	β_4	β_5	β_6
log(2)	log(1.4)	log(2)	log(1.4)	log(2)	log(1.4)
β_7	β_8	β_9	β_{10}	β_{11}	β_{12}
log(1.2)	log(1.4)	log(1.6)	log(1.2)	log(1.4)	log(1.6)

There are two outcome models for generation:

$$\text{Model A:} \quad Y = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_3 + \alpha_4 X_4 + \alpha_5 X_5 + \alpha_6 X_6 + \mu T + \epsilon,$$

$$\begin{aligned}\text{Model B:} \quad Y = {} & \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_3 + \alpha_4 X_4 + \alpha_5 X_5 + \alpha_6 X_6 \\ & + \alpha_7 X_2 X_4 + \alpha_8 X_3 X_5 + \alpha_9 X_3 X_6 + \alpha_{10} X_4 X_5 + \mu T + \epsilon,\end{aligned}$$

where the coefficients are shown in Table 4.2.

Table 4.2: Coefficients for Outcome Models

α_0	α_1	α_2	α_3	α_4	α_5	α_6	α_7	α_8	α_9	α_{10}	μ
-2.4	1.68	1.68	1.68	3.47	3.47	3.47	0.91	1.68	2.35	0.91	3

So there is a combination of 6 scenarios: 1A, 2A, 3A, 1B, 2B, and 3B, where the number indicates the propensity score model being used and the letter indicates the outcome model used for generation. Thus, scenarios 1A, 2A, and 3A, share the same outcome model with only main effects but they have different propensity score models. Propensity score model 1 only includes main effects; model 2 has a few more squared or interaction terms compared to model 1; model 3 has more squared and interaction terms compared to model 2. Comparing 1A with 1B, they have the same propensity score model but different outcome models in the sense that model B has some extra nonlinear terms. When estimating the propensity scores, we assume no knowledge of the true models and include all covariates in the logistic regression model as main effects. Thus, in scenarios 1A and 1B, the propensity score model is correctly specified for our approach and CBPS. On the other hand, there is model misspecification in all other scenarios. The performance metrics used are empirical standard error, bias, and mean squared error.

4.3.2 Simulation Results

We evaluate the performance of our proposed kernel distance propensity score (KDPS) and combined kernel distance propensity score (CKDPS) approaches in comparison with random forest (RF), CBPS, two different super learner (SL1 & SL2) methods using sample

sizes $n = 500$, $n = 1000$, and $n = 1500$. Each scenario is based on 1000 Monte Carlo simulations. The CKDPS includes the average of three positive definite kernels: Gaussian, Laplacian, and Exponential kernels. SL1 is a combination of logistic regression, random forest, elastic net regression, and generalized additive model, while SL2 excludes random forest. The ACE estimate results are shown in Table 4.3 for $n = 1500$. Results with $n = 1000$, $n = 500$ are shown in Table 4.4 and Table 4.5.

For the proposed KDPS and CKDPS, our choice of σ^2 is set to the median of all pairwise distances of each covariate. We evaluate the performance of KDPS using different $\widehat{\sigma}^2$ in Tables 4.6, 4.7, and 4.8. In KDPS, the default $\widehat{\sigma}^2$ is set to the median of all pairwise distances of each covariate. We defined KDPS_ds ($s = 5, 10, 20$) to be the estimator with the $\widehat{\sigma}^2$ used in KDPS being replaced by $\widehat{\sigma}^2/s$. We also defined KDPS_s ($s = 5, 10, 20$) to be the estimator with the $\widehat{\sigma}^2$ being replaced by $s\widehat{\sigma}^2$. Generally, the bias is the smallest when it is around the median. The standard error has a decreasing trend as $\widehat{\sigma}^2$ increases. The MSE shows a u-shaped trend. Using the median of all pairwise distances of each covariate is reasonable based on the bias-variance trade-off.

In terms of evaluating bias, from Table 4.3, we see that KDPS and CKDPS outperform all other methods in scenarios 2A, 3A, 2B, and 3B. All approaches become much more biased while the bias of our proposed method generally increases less than the other approaches from scenario 2A to 3A or 2B to 3B. When the propensity score model is correctly specified (1A and 1B), CBPS is less biased compared to other methods.

In terms of empirical standard error, all methods show a decreasing trend in empirical standard error from 1A to 3A or 1B to 3B except the RF method. In our data generation, when the propensity score model changes from 1 to 3, the standard error of the true propensity score decreases, which explains why the empirical standard error of the ACE estimator decreases. When the propensity score model is correctly specified, KDPS and CKDPS have larger standard errors compared to CBPS, which is supported by Theorem 4.2.4. The KDPS and CKDPS both have larger standard errors compared to other approaches as the degree of misspecification increases..

In terms of mean squared error, the proposed KDPS and CKDPS approaches both outperform CBPS in scenarios 2A, 3A, 2B, and 3B. It also outperforms RF and SL1 in all the scenarios.

Table 4.3: Performance for Estimation of ACE by Propensity Score Approaches ($n = 1500$)

		1A	2A	3A	1B	2B	3B
	KDPS	-0.78	-0.72	-0.90	-0.76	-1.74	-1.76
Average	CKDPS	-0.83	-0.76	-0.92	-0.80	-1.76	-1.78
Bias	RF	-2.49	-1.92	-1.56	-2.52	-2.35	-2.29
	CBPS	-0.63	-1.54	-1.83	-0.62	-2.48	-2.63
	SL1	-1.06	-4.42	-2.87	-0.99	-5.02	-3.61
	SL2	-0.34	-0.94	-1.37	-0.32	-1.97	-2.38
	KDPS	0.70	0.72	0.60	0.81	0.74	0.63
Empirical	CKDPS	0.71	0.78	0.67	0.81	0.79	0.67
Standard	RF	0.43	0.45	0.44	0.48	0.59	0.48
Error	CBPS	0.45	0.31	0.22	0.58	0.47	0.33
	SL1	1.13	0.41	0.39	1.21	0.46	0.45
	SL2	0.91	0.34	0.22	0.92	0.47	0.38
	KDPS	1.09	1.04	1.17	1.23	3.59	3.50
Mean	CKDPS	1.20	1.18	1.30	1.30	3.73	3.63
Squared	RF	6.38	3.90	2.63	6.60	5.88	5.47
Error	CBPS	0.60	2.46	3.41	0.73	6.38	7.05
	SL1	2.41	19.72	8.38	2.44	25.42	13.21
	SL2	0.94	1.00	1.94	0.96	4.09	5.79

KDPS: Kernel distance propensity score approach; CKDPS: Combined kernel distance propensity score approach; RF: Random forest approach; CBPS: Covariate balancing propensity score approach; SL1: Super learner method; SL2: Super learner method without RF.

Table 4.4: Performance for Estimation of ACE by Propensity Score Approaches ($n = 1000$)

		1A	2A	3A	1B	2B	3B
	KDPS	-1.01	-0.84	-0.96	-0.99	-1.81	-1.82
Average	CKDPS	-1.05	-0.87	-0.97	-1.05	-1.83	-1.83
Bias	RF	-2.72	-2.09	-1.65	-2.75	-2.54	-2.34
	CBPS	-0.81	-1.57	-1.82	-0.80	-2.47	-2.60
	SL1	-1.26	-4.95	-2.92	-1.20	-5.48	-3.64
	SL2	-0.44	-0.99	-1.38	-0.38	-2.00	-2.36
	KDPS	0.77	0.80	0.71	0.91	0.83	0.71
Empirical	CKDPS	0.87	0.88	0.77	0.93	0.89	0.78
Standard	RF	0.50	0.51	0.47	0.55	0.59	0.58
Error	CBPS	0.49	0.38	0.28	0.64	0.54	0.42
	SL1	1.40	0.46	0.50	1.45	0.52	0.54
	SL2	1.00	0.42	0.29	1.23	0.59	0.45
	KDPS	1.61	1.34	1.42	1.79	3.95	3.83
Mean	CKDPS	1.86	1.54	1.53	1.97	4.16	3.95
Squared	RF	7.66	4.64	2.94	7.84	6.83	5.80
Error	CBPS	0.90	2.60	3.40	1.05	6.41	6.96
	SL1	3.55	24.74	8.78	3.55	30.27	13.52
	SL2	1.19	1.16	1.97	1.65	4.33	5.78

KDPS: Kernel distance propensity score approach; CKDPS: Combined kernel distance propensity score approach; RF: Random forest approach; CBPS: Covariate balancing propensity score approach; SL1: Super learner method; SL2: Super learner method without RF.

Table 4.5: Performance for Estimation of ACE by Propensity Score Approaches ($n = 500$)

		1A	2A	3A	1B	2B	3B
	KDPS	-1.55	-1.19	-1.23	-1.53	-2.12	-1.99
Average	CKDPS	-1.57	-1.20	-1.24	-1.51	-2.11	-1.94
Bias	RF	-3.10	-2.31	-1.73	-3.10	-2.88	-2.43
	CBPS	-1.13	-1.58	-1.80	-1.16	-2.50	-2.57
	SL1	-1.55	-3.88	-2.67	-1.61	-4.59	-3.42
	SL2	-0.71	-1.11	-1.41	-0.74	-2.11	-2.36
	KDPS	0.93	0.95	0.90	1.05	1.04	1.00
Empirical	CKDPS	1.05	1.08	1.02	1.16	1.14	1.09
Standard	RF	0.66	0.69	0.63	0.71	0.82	0.78
Error	CBPS	0.60	0.47	0.40	0.80	0.71	0.59
	SL1	1.54	0.83	0.64	1.64	0.88	0.76
	SL2	1.31	0.59	0.43	1.39	0.79	0.68
	KDPS	3.27	2.34	2.32	3.45	5.57	4.97
Mean	CKDPS	3.57	2.60	2.57	3.63	5.74	4.98
Squared	RF	10.04	5.82	3.39	10.10	8.94	6.53
Error	CBPS	1.64	2.73	3.41	1.98	6.77	6.98
	SL1	4.77	15.75	7.53	5.26	21.82	12.30
	SL2	2.22	1.59	2.18	2.47	5.07	6.03

KDPS: Kernel distance propensity score approach; CKDPS: Combined kernel distance propensity score approach; RF: Random forest approach; CBPS: Covariate balancing propensity score approach; SL1: Super learner method; SL2: Super learner method without RF.

Table 4.6: Bias for KDPS with Different σ^2

	1A	2A	3A	1B	2B	3B
KDPS_d20	-1.37	-1.01	-1.01	-1.35	-1.91	-1.87
KDPS_d10	-1.28	-0.91	-1.00	-1.24	-1.85	-1.86
KDPS_d5	-1.21	-0.86	-0.99	-1.14	-1.83	-1.84
KDPS	-1.01	-0.84	-0.96	-0.99	-1.81	-1.82
KDPS_5	-1.03	-0.92	-0.97	-0.98	-1.84	-1.81
KDPS_10	-1.01	-0.94	-1.07	-0.97	-1.87	-1.91
KDPS_20	-0.99	-0.99	-1.21	-0.96	-1.92	-2.03

Table 4.7: Standard Error for KDPS with Different σ^2

	1A	2A	3A	1B	2B	3B
KDPS_d20	1.03	1.12	0.98	1.12	1.10	0.98
KDPS_d10	0.96	1.05	0.91	1.03	1.04	0.92
KDPS_d5	0.87	0.98	0.84	0.96	0.99	0.86
KDPS	0.77	0.80	0.71	0.91	0.83	0.71
KDPS_5	0.67	0.63	0.57	0.82	0.67	0.62
KDPS_10	0.67	0.66	0.58	0.83	0.70	0.63
KDPS_20	0.69	0.71	0.57	0.82	0.72	0.62

Table 4.8: Mean Squared Error for KDPS with Different σ^2

	1A	2A	3A	1B	2B	3B
KDPS_d20	2.94	2.28	1.99	3.08	4.86	4.45
KDPS_d10	2.56	1.93	1.83	2.58	4.49	4.29
KDPS_d5	2.22	1.70	1.68	2.22	4.32	4.11
KDPS	1.61	1.34	1.42	1.79	3.95	3.83
KDPS_5	1.50	1.24	1.27	1.64	3.83	3.68
KDPS_10	1.48	1.31	1.48	1.62	3.99	4.05
KDPS_20	1.44	1.49	1.79	1.58	4.21	4.53

Table 4.9: Simulation Results for Estimation of ACE, Absolute Bias between ASE and ESE

		1A	2A	3A	1B	2B	3B
KDPS	n=200	0.24	0.09	0.09	0.16	0.08	0.02
	n=500	0.19	0.03	0.05	0.17	0.02	0.07
	n=1000	0.14	0.08	0.11	0.09	0.06	0.06
	n=1500	0.03	0.15	0.17	0.05	0.12	0.13
CKDPS	n=200	0.09	0.09	0.25	0.01	0.11	0.21
	n=500	0.07	0.09	0.16	0.07	0.07	0.15
	n=1000	0.09	0.12	0.12	0.06	0.11	0.10
	n=1500	0.07	0.18	0.19	0.05	0.16	0.13

ASE: Average of standard errors reported from *survey* package, ESE: Empirical standard error.

In Table 4.9, we evaluate the bias between the average standard error (ASE) using the average of the sandwich variance estimates from the *survey* package and the empirical standard error (ESE) of the ACE estimates. Recall in Theorem 4.2.4, we stated that the asymptotic variance of the KDPS can be approximated by the sandwich variance. Under scenario 1A and 1B where the propensity score model is correct, the biases are closer to zero as n increases, which is consistent with our theoretical results in Theorem 4.2.4.

4.4 Application

The International Tobacco Control (ITC) project is an annual longitudinal survey signed and ratified by the Framework Convention on Tobacco Control started in 2002. It is designed to evaluate the effectiveness of national-level tobacco control policies in selected countries. It started in four countries: Canada, USA, Australia and UK. More detail about the ITC project can be found in Chen et al. (2018) or Thompson et al. (2006). Starting from 2011 (wave-8), a web survey was added, so participants could choose to answer the same questionnaire through either a web survey or a telephone survey. Our goal is to examine whether there is a significant mode effect (web versus telephone) on the number of cigarettes smoked per day reported by participants. We apply our proposed methods to the ITC wave-8 Canada data set. There is a total of 901 observations with 12 variables. The outcome variable (Y) is the number of cigarettes smoked per day for each participant. For our purpose, the treatment variable (T) is the mode of data collection with 1 indicating the participant completes the web survey or 0 if the participant answers the telephone survey. The primary goal is to examine the average difference in the number of cigarettes reported between web survey participants versus telephone survey participants. Most of the questions on the questionnaire are multiple choice but they are summarized by binary variables. The available ten covariates include: gender of participant (X_1: 1 for male, 0 otherwise), ethnicity (X_3: 1 if "White, English only", 0 otherwise), whether they visited their doctor since the last survey (X_4: 1 if yes, 0 otherwise), self description of health status (X_5: 1 if "Very good", 0 otherwise), measure of depression (X_6: 1 if "Little interest or pleasure" or "Feeling down or hopeless", 0 otherwise), frequency of alcohol drinks consumed in the last 12 months (X_7: 1 if "At least one day a week", 0 otherwise), income categories (X_8: 1 if "Low", 0 otherwise), education categories (X_9: 1 if "Low", 0 otherwise), and marital status (X_{10}: 1 if married, 0 otherwise). Age (X_2) is the only continuous covariate. It is feasible to assume that confounders related to choosing the web survey over telephone survey and the number of cigarettes smoked reported per day are among the basic demographic variables and all other measured covariates.

Covariate adjustment is necessary due to confounding. To find potential confounders, we perform univariate linear regression and logistic regression of the outcome and treatment on each covariate and report the p-value for the covariates (Table 4.10). Many covariates

Table 4.10: Summary for the Significance of Each Covariate in the ITC wave=8 Canada Dataset from Univariate Regression Models

	P-alue $(Y \sim X)$	Mean difference	P-value $(T \sim X)$	Odds ratio
Gender (X_1)	0.1000	1.0399	0.7792	1.0386
Age (X_2)	0.0004	0.0928	0.0006	-0.0197
Ethnicity (X_3)	0.0223	2.6950	0.3752	0.8007
Doctor visit (X_4)	0.4744	0.5150	0.0530	1.3523
Health status (X_5)	1.920×10^{-6}	-3.1690	0.3077	1.1568
Depression (X_6)	0.0466	1.2923	0.0218	0.73251
Alcohol consumption (X_7)	0.0627	-1.2615	0.0081	1.4671
Income categories (X_8)	0.0263	1.6047	< 0.0001	0.4830
Education categories (X_9)	0.0011	2.0768	< 0.0001	0.4763
Marital status (X_{10})	0.0180	-1.5022	0.0503	1.3065

Mean difference: Sample mean difference for ith covariate between $Y[X_i = 1]$ and $Y[X_i = 0]$; Odds ratio: The odds ratio for response $Y = 1$ comparing $X_i = 1$ vs $X_i = 0$.

(X_2, X_6, X_7, X_8, X_9, and X_{10}) show strong association with both the treatment variable and the outcome variable. Here, X_2 is the only continuous covariate and the mean difference for X_2 is actually the estimated increase in Y when X_2 is increased by one unit. The odds ratio for X_2 is the estimated ratio of the odds of $Y = 1$ when X_2 is increased by one unit as well. In this illustration, we include all covariates into the propensity score models for further causal effect estimation.

We show the mean difference and odds ratio to explore the expected nature of confounders. For example, individuals with $X_{10} = 1$ (marital status: 1 if married, 0 otherwise) have a higher probability being in treatment group but smaller outcome value compared to individuals with $X_{10} = 0$. Individuals with $X_9 = 1$ (education categories: 1 if "Low", 0 otherwise) have a lower probability being in treatment group but higher outcome value compared to individuals with $X_9 = 0$. The existence of confounders will lead to the biased estimation of average causal effect without balancing.

Estimates of the ACE based on our proposed method as well as other traditional approaches are given in Table 4.11. Both bootstrap and sandwich variance estimates from the *survey* package are used to estimate the standard error. Our choice of σ^2 is set to the median of all pairwise distances of each covariate. All methods except RF result in a negative value of the ACE estimate. The reason for a positive effect derived by RF may be due to the instability of RF method. The point estimates from KDPS and CKDPS are -0.2656 and -0.4696, respectively. The bootstrap standard error and theoretical standard error are very close to each other except for RF. In CKDPS, we combine Gaussian, Laplacian, and Exponential kernels. We also apply the KDPS to the data set with only Laplacian kernel or Exponential kernel (not shown). In both cases, the ACE estimates are closer to CKDPS compared to KDPS with Gaussian which we recommend here.

The negative ACE estimate indicates that the number of cigarettes smoked per day reported by web survey participants is less than telephone survey participants, although all methods show non-significant results at $\alpha = 0.05$ level since the 95% confidence intervals include zero. Therefore we conclude that there is no significant effect of the mode of survey collection on the number of cigarettes smoked per day reported by the participants.

Figure 4.3 shows the variability of the inverse probability weights across different propensity score approaches. An outlier (over 30) is removed from the weights of RF method. We can see that the RF method results in more extreme weights than other methods, which can also explain the larger standard error shown in Table 4.11. To evaluate the performance of our proposed approaches on balancing covariates, we report absolute standardized mean difference (ASMD) as a metric to evaluate covariate balance. Generally speaking, an ASMD value smaller than 0.1 indicates the covariate is balanced between the treatment and control groups (Stuart et al., 2013).

From Figure 4.4, we can see the ASMD for most covariates is reduced under 0.1 after applying our approach. This indicates our proposed approach achieved balance in all measured covariates in this application.

Figure 4.5 here compares the estimated kernel distance achieved by our method in comparison with all other methods after balancing. We can see that our proposed method has smaller kernel distance compared to all other method for most covariates. This is reasonable because our method aims to optimize the kernel distance. In terms of standardized

Table 4.11: Average Causal Effect between Observations Who Participated through Web Survey versus Telephone Survey

	ACE	BSE	TSE	Bootstrap 95%CI
KDPS	-0.4696	0.6612	0.6318	(-1.4770, 1.0192)
CKDPS	-0.2656	0.6695	0.6504	(-1.3707, 1.1728)
RF	0.8584	0.9073	0.8289	(-1.2839, 2.4555)
CBPS	-0.2011	0.6353	0.6287	(-1.4705, 0.9422)
SL1	-0.2921	0.6438	0.6303	(-1.4499, 1.0746)
SL2	-0.3055	0.6164	0.6300	(-1.5358, 0.8492)

BSE: Standard error based on 500 bootstraps samples; TSE: Theoretical standard error evaluated by *survey* package; CI: Confidence intervals based on quantiles of bootstrap estimates.

differences, our approach does not consistently produce lower values. The RF method does not reduce the imbalance in many covariates in term of standardized difference.

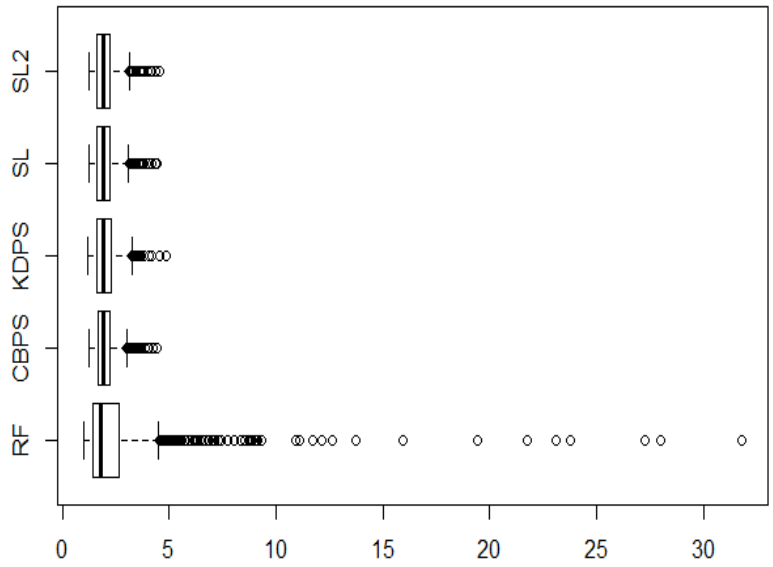

Figure 4.3: Boxplots of Weights by Propensity Score Approaches in ITC Data Analysis

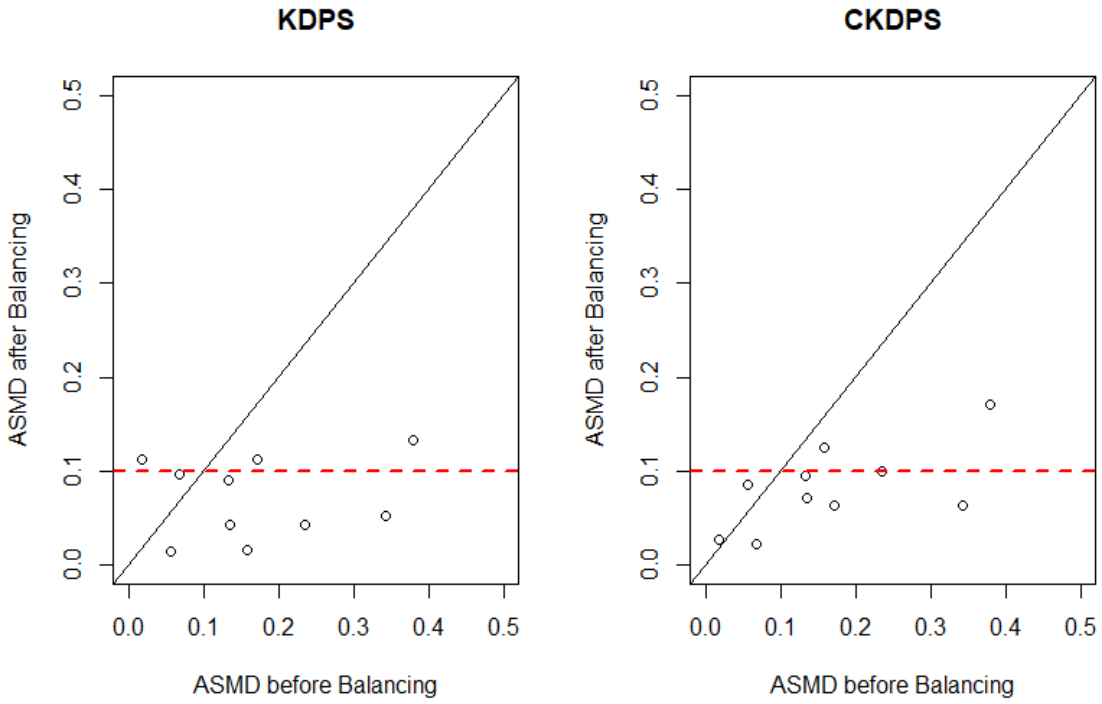

Figure 4.4: ASMD Values for Each Covariate under KDPS and CKDPS before and after Balancing in ITC Data Analysis

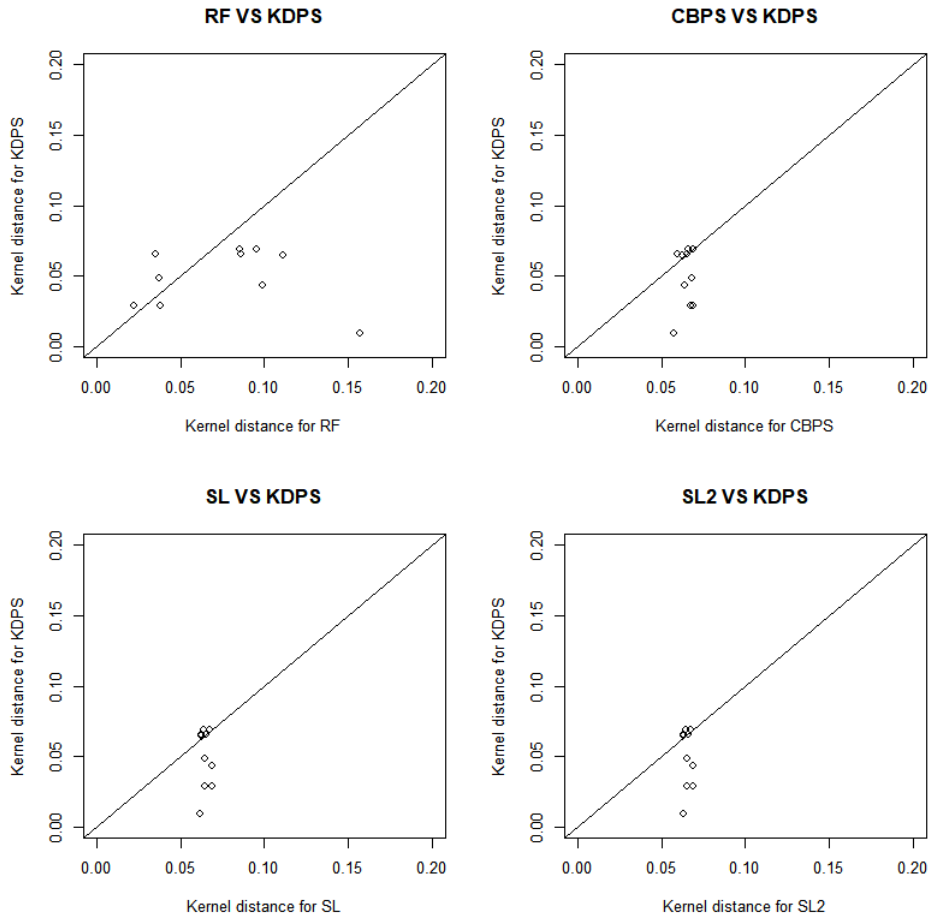

Figure 4.5: Kernel Distance for KDPS vs Other Methods

4.5 Conclusion

The KDPS approach uses the kernel distance as a measure of covariate balance and achieves balance in the covariate distributions between treatment and control groups. Our method improves CBPS in the sense that CBPS achieves the balance only in finite moment conditions but not the overall distributions. A kernel function can be decomposed into an inner product of two infinite-dimensional basis functions, so balancing kernel distance is not just balancing finite moment of covariates. We optimize the weighted kernel distance through the GMM framework. The simulation results show that when the propensity score is misspecified, our kernel approach yields causal estimators with smaller bias compared with other approaches.

There are several advantages to our proposed KDPS approach for propensity score estimation. First, our approach yields estimates with a smaller bias and mean squared error than all other approaches especially when the propensity score model is misspecified. Second, our kernel distance approach helps to reduce the imbalance in the covariates so that propensity score based approaches are valid to estimate the causal effects.

Similar to the KDPS approach, we can also construct another approach using empirical likelihood or pseudo empirical likelihood methods. Wu (2005) employs the pseudo empirical likelihood method for the analysis of complex survey data. In the case of responses missing at random, Qin and Zhang (2007) also seek to construct the constrained empirical likelihood estimation of mean response. Instead of estimating $\boldsymbol{\beta}$, we can estimate w_i through the empirical likelihood framework directly under the constraint: $\sum_{i=1}^{n}\sum_{j=1,i\neq j}^{n} T_i^* T_j^* K(X_{im}, X_{jm})=0$ for $m = 0, 1, \ldots, p$. Let $f(t, \mathbf{x}_i)$ be the joint density function of treatment T_i and covariate vector $\mathbf{X}_i$. Here w_i is the inverse probability weight through propensity score defined before. So $w_i = 1/P(T_i = 1|\mathbf{x}_i)$ for treated units or $w_i = 1/P(T_i = 0|\mathbf{x}_i)$ for control units. Thus, the empirical likelihood for the whole sample data can be simplified:

$$L = \prod_{i=1}^{n} f(t, \mathbf{x}_i) = \prod_{i=1}^{n} P(T_i = t|\mathbf{x}_i) f(\mathbf{x}_i) = \prod_{i=1}^{n} \frac{1}{w_i} f(\mathbf{x}_i). \tag{4.10}$$

The corresponding log likelihood function is:

$$\log L = \sum_{i=1}^{n} \log f(\mathbf{x}_i) - \log w_i, \tag{4.11}$$

where the constraints are $\sum_{i=1}^{n} w_i = 2n$ and the target is Equation (4.4), where $f(\cdot)$ is the density function of $\mathbf{X}$. By maximizing the log likelihood function under the constraints, we can derive the optimal $\widehat{w}_i$ and estimate the causal effect. After obtaining the weights, we can also employ targeted maximum likelihood estimation (TMLE) (van der Laan, 2014) to estimate the causal effect, which is shown to be double robust and super-efficient under certain conditions (van der Laan and Gruber, 2010). A few other double robust estimators may also be employed here. For example, using augmented IPW instead of IPW can be more efficient (Qin and Zhang, 2007). The kernel distance is defined by summing across all pairs but it is also worth exploring to sum across only the discordant pairs when one is treated and the other is not.

4.6 Theorems and Proofs

4.6.1 Proof of Theorem 4.6.1

The following theorem is required for the application of GMM framework and relies on Assumption 1. This theorem helps us to construct the consistency of $\widehat{\boldsymbol{\beta}}$.

Theorem 4.6.1 (Estimating Equation Condition)**.** *Suppose the correct propensity score is given as in Equation (2.1) and $\boldsymbol{\beta}_0$ is the true vector of logistic regression coefficients, then under Assumptions 1,*

$$\mathrm{E}\{\mathbf{g}_{\boldsymbol{\beta}_0}(\mathbf{T}_k, \mathbf{Z}_k)\} = \mathbf{0}.$$

The definition of $\mathbf{g}_{\boldsymbol{\beta}_0}(\mathbf{T}_k, \mathbf{Z}_k)$ can be found in Equation (4.7).

Proof. We only need to prove $\mathrm{E}(g_{k,m}) = 0$, where $\mathbf{g}_{\boldsymbol{\beta}_0}(\mathbf{T}_k, \mathbf{Z}_k) = (g_{k,0}, \ldots, g_{k,p})^\intercal$ for $m =$

$0, \ldots, p$.

$$\begin{aligned}
\mathrm{E}[g_{k,m}] &= \mathrm{E}\{h_{\boldsymbol{\beta}}(\mathbf{T}_k, \mathbf{Z}_k)K(f_m(\mathbf{Z}_k))\} \\
&= \mathrm{E}\{q_{\boldsymbol{\beta}}(T_i, \mathbf{X}_i)q_{\boldsymbol{\beta}}(T_j, \mathbf{X}_j)K(X_{i,m}, X_{j,m})\} \\
&= \mathrm{E}\{n^2 T_i^* T_j^* K(X_{i,m}, X_{j,m})\} \\
&= n^2 \mathrm{E}\left[\mathrm{E}\{T_i^* T_j^* K(X_{i,m}, X_{j,m})|\mathbf{X}_i, \mathbf{X}_j\}\right] \\
&= n^2 \mathrm{E}\{K(X_{i,m}, X_{j,m})\mathrm{E}(T_i^* T_j^*|\mathbf{X}_i, \mathbf{X}_j)\} \\
&= n^2 \mathrm{E}\{K(X_{i,m}, X_{j,m})\mathrm{E}(T_i^*|\mathbf{X}_i)\mathrm{E}(T_j^*|\mathbf{X}_j)\}, \quad \text{for } i < j,
\end{aligned}$$

where $T_i^* = \frac{T_i w_i}{n} - \frac{(1-T_i)w_i}{n}$, and $w_i = 1/\left[T_i e_{\boldsymbol{\beta}_0}(\mathbf{X}_i) + (1-T_i)\{1 - e_{\boldsymbol{\beta}_0}(\mathbf{X}_i)\}\right]$. We will drop the subscript $\boldsymbol{\beta}_0$ for simplicity. When estimating ACE, we can derive the conditional distribution of $T_i w_i|\mathbf{X}_i$ and $(1-T_i)w_i|\mathbf{X}_i$,

$$T_i w_i|\mathbf{X}_i = \begin{cases} \frac{1}{e(\mathbf{X}_i)} & \text{with probability} \quad e(\mathbf{X}_i) \\ 0 \quad \text{with probability} & 1 - e(\mathbf{X}_i) \end{cases},$$

$$(1-T_i)w_i|\mathbf{X}_i = \begin{cases} 0 \quad \text{with probability} & e(\mathbf{X}_i) \\ \frac{1}{1-e(\mathbf{X}_i)} & \text{with probability} \quad 1 - e(\mathbf{X}_i) \end{cases}.$$

Hence, $\mathrm{E}(T_i w_i|\mathbf{X}_i) = \mathrm{E}\{(1-T_i)w_i|\mathbf{X}_i\} = 1$ and $\mathrm{E}(T^*|\mathbf{X}_i) = 0$. This proves $\mathrm{E}(g_{k,m}) = 0$, so we conclude that $\mathrm{E}\{\mathbf{g}_{\boldsymbol{\beta}_0}(\mathbf{T}_k, \mathbf{Z}_k)\} = \mathbf{0}$. □

4.6.2 Proof of Lemma 4.6.2

To ensure the GMM framework can be applied here, we also prove that $\mathbf{G}_{n,\boldsymbol{\beta}}$ or $\bar{\mathbf{g}}_{\boldsymbol{\beta}}(\mathbf{T}, \mathbf{Z})$ converges in probability to $\mathrm{E}\{\mathbf{g}_{\boldsymbol{\beta}}(\mathbf{T}_k, \mathbf{Z}_k)\}$. This is quite different from the general law of large numbers where all variables are independent. The $\mathbf{g}_{\boldsymbol{\beta}}(\mathbf{T}_k, \mathbf{Z}_k)$ may depends on $\mathbf{g}_{\boldsymbol{\beta}}(\mathbf{T}_r, \mathbf{Z}_r)$ when $\mathbf{Z}_k$ and $\mathbf{Z}_r$ share the same component such as $\mathbf{Z}_k = (\mathbf{X}_i^\intercal, \mathbf{X}_j^\intercal)^\intercal$ and $\mathbf{Z}_r = (\mathbf{X}_j^\intercal, \mathbf{X}_l^\intercal)^\intercal$.

Lemma 4.6.2. *Let $\bar{g}_m(\boldsymbol{\beta}) = \sum_{k=1}^N g_{k,m}/N$ and $E_m(\boldsymbol{\beta}) = \mathrm{E}(g_{k,m})$, then $\bar{g}_m(\boldsymbol{\beta}) \xrightarrow{p} E_m(\boldsymbol{\beta})$ for $m = 0, 1, \ldots, p$. We can further conclude that $\bar{\mathbf{g}}_{\boldsymbol{\beta}} = (\bar{g}_0(\boldsymbol{\beta}), \ldots, \bar{g}_p(\boldsymbol{\beta}))^\intercal$ is a consistent estimator of $\mathrm{E}\{\mathbf{g}_{\boldsymbol{\beta}}(\mathbf{T}_k, \mathbf{Z}_k)\}$ uniformly for $\boldsymbol{\beta} \in \Theta$:*

$$\bar{\mathbf{g}}_{\boldsymbol{\beta}} \xrightarrow{p} \mathrm{E}\{\mathbf{g}_{\boldsymbol{\beta}}(\mathbf{T}_k, \mathbf{Z}_k)\}, \quad \text{as } n \to \infty.$$

We use the Chebyshev's inequality to prove the consistency of estimating functions in Lemma 4.6.2. It is also true for any $\boldsymbol{\beta}$ even when the propensity score model is not correctly specified.

Proof. We only need to prove $\bar{g}_m(\boldsymbol{\beta}) \xrightarrow{p} E_m(\boldsymbol{\beta})$ as $n \to \infty$. We will denote $E_m(\boldsymbol{\beta})$ by a_m. It can be easily derived that $E(g_{k,m}^2)$ is a finite constant not depending on n. We use $b_m = \mathrm{E}(g_{k,m}^2)$, where b_m is this constant. We prove by Chebyshev's inequality,

$$\begin{aligned}
&P\left(\left|\frac{1}{N}\sum_{k=1}^{N} g_{k,m} - a_m\right| > \epsilon\right) \\
\leq &\frac{\mathrm{Var}\left(\sum_{k=1}^{N} g_{k,m}/N\right)}{\epsilon^2} \\
= &\frac{\sum_{k=1}^{N}\mathrm{Var}(g_{k,m}) + \sum_{k\neq r}\mathrm{Cov}(g_{k,m}, g_{r,m})}{N^2\epsilon^2} \\
= &\frac{Nb_m + Na_m^2 + \sum_{k\neq r}\mathrm{Cov}(g_{k,m}, g_{r,m})}{N^2\epsilon^2}.
\end{aligned}$$

In the third term of the above formula, the covariance is either zero or a constant. here $g_{k,m}$ and $g_{r,m}$ are both functions of two observations, let $g_{k,m} = h(\mathbf{X}_i, \mathbf{X}_j)$ and $g_{r,m} = h(\mathbf{X}_l, \mathbf{X}_s)$, where $i < j$, $l < s$. Then $g_{k,m}$ and $g_{r,m}$ can only share one component. The covariance is nonzero when they share one component. Each $g_{k,m}$ will share one component with at most $(2n-4)$ other $g_{r,m}$ terms. So there are at most $N \times (2n-4)$ nonzero covariance terms. Let $\mathrm{Cov}(g_{k,m}, g_{r,m}) = c_m$. then we have

$$P\left(\left|\frac{1}{N}\sum_{k=1}^{N} g_{k,m} - a_m\right| > \epsilon\right) \leq \frac{Nb_m + Na_m^2 + N(2n-4)c_m}{N^2\epsilon^2} \tag{4.12}$$

Hence, $\bar{g}_m(\boldsymbol{\beta}) \xrightarrow{p} a_m$ as $n \to \infty$ and $N \to \infty$. □

4.6.3 Proof of Theorem 4.2.1

Theorem. *Under Assumptions 1 and Lemma 4.6.2, we have* $\widehat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}^*$ *as* $n \to \infty$. *Moreover, if the propensity score model is correctly specified, i.e.* $P(T = 1|\mathbf{X} = \mathbf{x}) = e_{\boldsymbol{\beta}_0}(\mathbf{x})$, *then* $\boldsymbol{\beta}_0 = \boldsymbol{\beta}^*$ *and* $\widehat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}_0$ *as* $n \to \infty$.

Proof. We prove the consistency of $\widehat{\boldsymbol{\beta}}$ in $\boldsymbol{\beta}^*$ by Theorem 2.1 in Newey and McFadden (1994). The conditions (i), (ii), and (iii) are satisfied by (3)-(5) of our Assumptions 1. Based on Assumption (2) and Lemma 4.6.2, $\bar{\mathbf{g}}_{\boldsymbol{\beta}}(\mathbf{T}, \mathbf{Z})^\intercal \Sigma_{\boldsymbol{\beta}}(\mathbf{T}, \mathbf{Z})^{-1} \bar{\mathbf{g}}_{\boldsymbol{\beta}}(\mathbf{T}, \mathbf{Z})$ converges in probability to $\mathrm{E}\{\bar{\mathbf{g}}_{\boldsymbol{\beta}}(\mathbf{T}, \mathbf{Z})\}^\intercal \mathbf{A}^* \mathrm{E}\{\bar{\mathbf{g}}_{\boldsymbol{\beta}}(\mathbf{T}, \mathbf{Z})\}$. So the condition (iv) is also satisfied. We conclude that $\widehat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}^*$.

The next step is to show that $\boldsymbol{\beta}_0 = \boldsymbol{\beta}^*$ when $P(T = 1|\mathbf{X} = \mathbf{x}) = e_{\boldsymbol{\beta}_0}(\mathbf{x})$. By Theorem 4.6.1, we have $\mathrm{E}\{\mathbf{g}_{\boldsymbol{\beta}_0}(\mathbf{T}_k, \mathbf{Z}_k)\} = \mathbf{0}$ and $\mathrm{E}\{\bar{\mathbf{g}}_{\boldsymbol{\beta}_0}(\mathbf{T}, \mathbf{Z})\}^\intercal \mathbf{A}^* \mathrm{E}\{\bar{\mathbf{g}}_{\boldsymbol{\beta}_0}(\mathbf{T}, \mathbf{Z})\} = 0$. Since $\mathbf{A}^*$ is a positive definite matrix, $\boldsymbol{\beta}^*$ and $\boldsymbol{\beta}_0$ are both minimizers of $\mathrm{E}\{\bar{\mathbf{g}}_{\boldsymbol{\beta}}(\mathbf{T}, \mathbf{Z})\}^\intercal \mathbf{A}^* \mathrm{E}\{\bar{\mathbf{g}}_{\boldsymbol{\beta}}(\mathbf{T}, \mathbf{Z})\}$. Based on Assumption (5), $\boldsymbol{\beta}^* = \boldsymbol{\beta}_0$. Therefore, we have $\widehat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}_0$ as $n \to \infty$. □

4.6.4 Proof of Theorem 4.2.2

Theorem. *Under Assumptions 1, if the propensity score model is correctly specified and $\widehat{\boldsymbol{\beta}}$ is obtained through Equation (4.9), then $\widehat{\mu}_{\widehat{\boldsymbol{\beta}}} \xrightarrow{p} \mu$ as $n \to \infty$.*

Proof. This proof is similar to what is done in Theorem 3.1 of Fan et al. (2016). When the propensity score model is correctly specified, let $e_{\widehat{\boldsymbol{\beta}}}(\mathbf{X}_i) = \widehat{e}_i$, then

$$\begin{aligned}\widehat{\mu}_{\widehat{\boldsymbol{\beta}}} &= \frac{1}{n}\left\{\sum_{i=1}^{n} T_i Y_i \widehat{w}_i - \sum_{i=1}^{n}(1-T_i) Y_i \widehat{w}_i\right\} \\ &= \frac{1}{n}\left\{\sum_{i=1}^{n} \frac{T_i Y_i}{\widehat{e}_i} - \sum_{i=1}^{n} \frac{(1-T_i) Y_i}{1-\widehat{e}_i}\right\}.\end{aligned}$$

Let $r_{\boldsymbol{\beta}}(T_i, X_i, Y_i) = T_i Y_i / e_i - (1-T_i) Y_i/(1-e_i)$ and $e_i = e(\mathbf{X}_i)$. By Assumptions (1) and (6), we get that $|T_i Y_i / e_i - (1-T_i) Y_i/(1-e_i)| < 2|Y_i|/a_0$. Hence by Lemma 2.4 in Newey and McFadden (1994), we have $\sup_{\boldsymbol{\beta}\in\Theta} |1/n \sum_{i=1}^{n} r_{\boldsymbol{\beta}}(T_i, X_i, Y_i) - \mathrm{E}\{r_{\boldsymbol{\beta}}(T_i, X_i, Y_i)\}| = o_p(1)$. We also have $\widehat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}_0$ from Theorem 4.2.1. Hence by the dominated convergence theorem, we have

$$\widehat{\mu}_{\widehat{\boldsymbol{\beta}}} = \mathrm{E}\left\{\frac{T_i Y_i}{e_i^0} - \frac{(1-T_i) Y_i}{1-e_i^0}\right\} + o_p(1),$$

where $e_i^0 = e_{\boldsymbol{\beta}_0}(\mathbf{X}_i)$. Since $Y_i = T_i \times Y_i(1) + (1 - T_i) \times Y_i(0)$, then

$$
\begin{aligned}
&\mathrm{E}\left\{\frac{T_i Y_i}{e_i^0} - \frac{(1-T_i)Y_i}{1-e_i^0}\right\} \\
=&\mathrm{E}\left\{\frac{T_i Y_i(1)}{e_i^0} - \frac{(1-T_i)Y_i(0)}{1-e_i^0}\right\} \\
=&\mathrm{E}\left[\frac{\mathrm{E}\{T_i Y_i(1)|\mathbf{X}_i\}}{e_i^0}\right] - \mathrm{E}\left[\frac{\mathrm{E}\{(1-T)_i Y_i(0)|\mathbf{X}_i\}}{1-e_i^0}\right] \\
=&\mathrm{E}\left[\frac{\mathrm{E}(T_i|\mathbf{X}_i)\mathrm{E}\{Y_i(1)|\mathbf{X}_i\}}{e_i^0}\right] - \mathrm{E}\left[\frac{\{1-\mathrm{E}(T_i|\mathbf{X}_i)\}\mathrm{E}\{Y_i(0)|\mathbf{X}_i\}}{1-e_i^0}\right] \\
=&\mathrm{E}\left[\mathrm{E}\{Y_i(1)|\mathbf{X}_i\}\right] - \mathrm{E}\left[\mathrm{E}\{Y_i(0)|\mathbf{X}_i\}\right] \\
=&\mathrm{E}\{Y(1) - Y(0)\} = \mu.
\end{aligned}
$$

Therefore, we can conclude that $\widehat{\mu}_{\widehat{\boldsymbol{\beta}}} \xrightarrow{p} \mu$ as $n \to \infty$.

□

4.6.5 Proof of Lemma 4.6.3

Lemma 4.6.3. *Let* $\mathbf{G}_{n,\boldsymbol{\beta}_0} = (G^{(0)}_{n,\boldsymbol{\beta}_0}, \ldots, G^{(p)}_{n,\boldsymbol{\beta}_0})^\intercal$ *be the estimating function evaluated at* $\boldsymbol{\beta}_0$. *Given the data* $(T_i, \mathbf{X}_i, Y_i)$, $i = 1, \ldots, n$, *it can be verified that the estimating function* $G^{(m)}_{n,\boldsymbol{\beta}_0}$ *is a martingale where* $G^{(m)}_{n,\boldsymbol{\beta}_0} = \sum_{l=2}^{n} U_l^{(m)}$, $m = 0, \ldots, p$, *and* $U_l^{(m)} = \sum_{1 \le j < l} q_{l,j}^{(m)}$, $q_{l,j}^{(m)} = 2n^2 T_l^* T_j^* K(X_{l,m}, X_{j,m})$.

With the proof of Lemma 4.6.3, we can use the martingale central limit theorem (Heyde and Brown, 1970) to establish the asymptotic normality of the estimating function in Lemma 4.6.4 in the following and further the asymptotic normality of the $\widehat{\boldsymbol{\beta}}$ and $\widehat{\mu}_{\widehat{\boldsymbol{\beta}}}$ in Theorems 4.2.3 and 4.2.4 of the thesis, respectively.

In our setting $U_l = \sum_{1 \le j < l} q_{l,j}$, so there is no U_1 and the martingale starts from U_2. For simplicity in the proof here, we drop the superscript, m and subscript, $\boldsymbol{\beta_0}$, and use G_n, U_l and $q_{l,j}$ instead of $G^{(m)}_{n,\boldsymbol{\beta}_0}$, $U_l^{(m)}$, and $q_{l,j}^{(m)}$. Also $K(X_{i,m}, X_{j,m})$ is simplified to $K_{i,j}$. We only need to prove the martingale difference of G_n is zero.

Proof. Consider,

$$\begin{aligned}
&\mathrm{E}(G_{n+1} - G_n | G_2, \ldots, G_n) \\
=&\mathrm{E}(U_{n+1} | U_2, \ldots, U_n) \\
=&\mathrm{E}\{ \sum_{1 \le j \le n+1} q_{n+1,j} | (\mathbf{X}_1, \ldots, \mathbf{X}_n), (T_1, \ldots, T_n) \} \\
=& \sum_{1 \le j \le n+1} \mathrm{E}\{ n^2 T^*_{n+1} T^*_j K_{n+1,j} | (\mathbf{X}_j, \mathbf{X}_n), (T_j, T_n) \},
\end{aligned}$$

and

$$\begin{aligned}
&\mathrm{E}\{ n^2 T^*_{n+1} T^*_j K_{n+1,j} | (\mathbf{X}_j, \mathbf{X}_n), (T_j, T_n) \} \\
=&\mathrm{E}[\mathrm{E}\{ n^2 T^*_{n+1} T^*_j K_{n+1,j} | (\mathbf{X}_j, \mathbf{X}_n, \mathbf{X}_{n+1}), (T_j, T_n) \}],
\end{aligned}$$

$K_{n+1,j}$ is only a function of $\mathbf{X}_{n+1}$ and $\mathbf{X}_j$, while T^*_j is a function of $\mathbf{X}_j$ and T_j, and T^*_{n+1} is also a function of $\mathbf{X}_{n+1}$ and T_{n+1}. So we have:

$$\begin{aligned}
&\mathrm{E}(G_{n+1} - G_n | G_2, \ldots, G_n) \\
=&\mathrm{E}[\mathrm{E}\{ n^2 T^*_{n+1} T^*_j K_{n+1,j} | (\mathbf{X}_j, \mathbf{X}_n, \mathbf{X}_{n+1}), (T_j, T_n) \}] \\
=&\mathrm{E}\{ n^2 K_{n+1,j} T^*_j \mathrm{E}(T^*_{n+1} | \mathbf{X}_{n+1}) \} \\
=&0.
\end{aligned}$$

Since we have $\mathrm{E}(T^*_{n+1} | \mathbf{X}_{n+1}) = 0$ from proof of Theorem 4.2.1. Based on Assumption (1), we can also have $\mathrm{E}|G_n| < \infty$, so G_n is martingale. □

4.6.6 Proof of Lemmas 4.6.4 - 4.6.7

From Lemmas 4.6.4 - 4.6.7, we show the asymptotic normality of estimating function G_n based on martingale central limit theorem.

Lemma 4.6.4. *Let* $\sigma^2_l = \mathrm{E}(U^2_l | \mathcal{F}_{l-1})$ *and* $s^2_n = \sum_{l=2}^n \mathrm{E}(\sigma^2_l)$ *where* $l \ge 2$ *and* $\mathcal{F}_{l-1}$ *is the* σ*-field generated by* $U_2, \ldots, U_{l-1}$*. Following from the martingale central limit theorem (Heyde*

and Brown, 1970) and Lemma 4.6.3, there exists a finite constant K_1 depending only on δ, such that

$$\sup_x |P(G_n \leq s_n x) - \Phi(x)|$$
$$\leq K_1 \left\{ s_n^{-2-2\delta} (\sum_{l=2}^{n} \mathrm{E}|U_l|^{2+2\delta} + \mathrm{E}|(\sum_{l=2}^{n} \sigma_l^2 - s_n^2)|^{1+\delta}) \right\}^{1/(3+2\delta)},$$

where $\Phi(x)$ is the cumulative distribution function of standard normal distribution and G_n is the mth element of $\mathbf{G}_{n,\boldsymbol{\beta}_0}$ for simplicity. If the following conditions are satisfied:

$$\lim_{n\to\infty} s_n^{-2-2\delta} \sum_{l=2}^{n} \mathrm{E}|U_l|^{2+2\delta} = 0 \tag{4.13}$$

$$\lim_{n\to\infty} \mathrm{E}|s_n^{-2}(\sum_{l=2}^{n} \sigma_l^2 - 1)|^{1+\delta} = 0, \tag{4.14}$$

then $\lim_{n\to\infty} P(G_n \leq s_n x) = \Phi(x)$ or $\frac{n(G_n/n^2)}{s_n/n} \xrightarrow{d} N(0,1)$.
We can further conclude that $n(\mathbf{G}_{n,\boldsymbol{\beta}_0}/n^2) \xrightarrow{d} N(\mathbf{0}, \mathbf{V})$ in vector form, where $\mathbf{V}$ is the limiting variance covariance matrix.

The asymptotic normality of G_n depends on the two conditions, Equations (4.13) and (4.14). Here, we prove that the conditions are satisfied under our setting when $\delta = 1$. In the following proof, we constructed three Lemmas to finish the proof of Lemma 4.6.4.

Lemma 4.6.5. *For $i, j, s, t, l = 1, \ldots, n$,*

(1) If $s \neq t$, then $\mathrm{E}(q_{l,j}^2 q_{l,s} q_{l,t}) = 0$ where $j, s, t < l$.

(2) If at least one of i, j, s, t is not equal to the rest, then $\mathrm{E}(q_{l,i} q_{l,j} q_{l,s} q_{l,t}) = 0$ where $i, j, s, t < l$.

(3) If $(l, i) \neq (s, j)$, then $\mathrm{E}(q_{l,i} q_{s,j}) = 0$ where $i < l$ and $j < s$.

Proof. For the first part, let's assume $t \neq s$,

$$\begin{aligned}
&\mathrm{E}(q_{l,j}^2 q_{l,s} q_{l,t}) \\
=&\mathrm{E}[q_{l,j}^2 q_{l,s} \mathrm{E}\{q_{l,t} | (\mathbf{X}_l, \mathbf{X}_j, \mathbf{X}_s, \mathbf{X}_t), (T_l, T_j, T_s)\}] \\
=&\mathrm{E}[q_{l,j}^2 q_{l,s} \mathrm{E}\{2n^2 T_l^* T_t^* K_{l,t} | (\mathbf{X}_l, \mathbf{X}_j, \mathbf{X}_s, \mathbf{X}_t), (T_l, T_j, T_s)\}] \\
=&\mathrm{E}\{2n q_{l,j}^2 q_{l,s} T_l^* K_{l,t} \mathrm{E}(T_t^* | \mathbf{X}_t)\} \\
=&0, \quad \text{Since } \mathrm{E}(T_t^* | \mathbf{X}_t) = 0.
\end{aligned}$$

Similarly, we can also prove $\mathrm{E}(q_{l,i} q_{l,j} q_{l,s} q_{l,t}) = 0$ and $\mathrm{E}(q_{l,i} q_{s,j}) = 0$. □

Lemma 4.6.6. $\lim_{n\to\infty} s_n^{-2-2\delta} \sum_{l=1}^n \mathrm{E}|U_l|^{2+2\delta} = 0$ *when* $\delta = 1$.

Proof.

$$\begin{aligned}
s_n^{-4} \sum_{l=1}^n \mathrm{E}|U_l|^4 =& s_n^{-4} \sum_{l=1}^n \mathrm{E}\left\{ \left(\sum_{1\leq j<l} q_{l,j} \right)^4 \right\} \\
=& s_n^{-4} \sum_{l=1}^n \mathrm{E}\left\{ \left(\sum_{1\leq j<l} q_{l,j}^2 + 2 \sum_{i<j<l} q_{l,i} q_{l,j} \right)^2 \right\}
\end{aligned}$$

In the following to expand the squared term inside the expectation, the expectation for the cross product within $\sum_{1\leq j<l} q_{l,j}^2$ (for example: $\mathrm{E}(q_{l,j}^2 q_{l,i}^2)$) is non zero. The expectation for the cross product within $\sum_{i<j<l} q_{l,i} q_{l,j}$ is zero by Lemma 4.6.5 (2) (for example: $\mathrm{E}(q_{l,i} q_{l,j} q_{l,s} q_{l,t})$). The cross products between the two summations (for example: $\mathrm{E}(q_{l,j}^2 q_{l,s} q_{l,t})$) are zero by Lemma 4.6.5 (1), so we have

$$\begin{aligned}
s_n^{-4} \sum_{l=1}^n \mathrm{E}|U_l|^4 =& s_n^{-4} \sum_{l=1}^n \mathrm{E}\left(\sum_{1\leq j<l} q_{l,j}^4 + 6 \sum_{i<j<l} q_{l,i}^2 q_{l,j}^2 \right) \\
=& s_n^{-4} (G_I + G_{II}),
\end{aligned}$$

where $G_I = \sum_{1\leq j<l\leq n} \mathrm{E}(q_{l,j}^4) = O(n^2)$ and $G_{II} = 6 \sum_{1\leq i<j<l<n} \mathrm{E}(q_{l,i}^2 q_{l,j}^2) = O(n^3)$.

The $\mathrm{E}(q_{l,j}^4)$ and $\mathrm{E}(q_{l,i}^2 q_{l,j}^2)$ are finite. We can also find that,

$$
\begin{aligned}
s_n^2 &= \sum_{l=2}^{n} \mathrm{E}(\sigma_l^2) \\
&= \sum_{l=2}^{n} \mathrm{E}(U_l^2) - 0 \\
&= \sum_{l=2}^{n} \mathrm{E}(U_l^2) - \sum_{l=2}^{n} \mathrm{E}^2(U_l) \\
&= \sum_{l=2}^{n} \mathrm{Var}(U_l) - \sum_{l \neq s}^{n} \mathrm{Cov}(U_l, U_s) \\
&= \mathrm{Var}\left(\sum_{l=2}^{n} U_l \right) = \mathrm{Var}(G_n), \text{ where } \mathrm{Cov}(U_l, U_s) = 0 \text{ by Lemma 4.6.5 (3)}, \\
&= \mathrm{Var}\left(\sum_{1 \leq j < l \leq n} q_{l,j} \right) \\
&= \mathrm{E}\left\{ \left(\sum_{1 \leq j < l \leq n} q_{l,j} \right)^2 \right\} - \mathrm{E}^2 \left(\sum_{1 \leq j < l \leq n} q_{l,j} \right) \\
&= \mathrm{E}\left\{ \left(\sum_{1 \leq j < l \leq n} q_{l,j} \right)^2 \right\} - 0 \\
&= \mathrm{E}\left(\sum_{1 \leq j < l \leq n} q_{l,j}^2 \right) + \mathrm{E}\left(\sum_{(l,j) \neq (s,t)} q_{l,j} q_{s,t} \right) \\
&= \sum_{1 \leq j < l \leq n} \mathrm{E}\left(q_{l,j}^2 \right) \\
&= O(n^2)
\end{aligned}
$$

$\mathrm{E}(q_{l,j}^2)$ is finite. Based on the above derivation, we get that $\lim_{n \to \infty} s_n^{-4} \sum_{l=1}^{n} \mathrm{E}|U_l|^4 = 0$. □

Lemma 4.6.7. $\lim_{n \to \infty} \mathrm{E}|s_n^{-2}(\sum_{l=1}^{n} \sigma_l^2 - 1)|^{1+\delta} = 0$ *when* $\delta = 1$.

Proof.

$$
\begin{aligned}
&\mathrm{E}\left\{\left(s_n^{-2}\sum_{l=2}^{n}\sigma_l^2-1\right)^2\right\}\\
=&\mathrm{E}\left\{\left(s_n^{-2}\sum_{l=2}^{n}\sigma_l^2-1\right)^2\right\}-\mathrm{E}^2\left(s_n^{-2}\sum_{l=2}^{n}\sigma_l^2-1\right), \text{since } \mathrm{E}^2\left(s_n^{-2}\textstyle\sum_{l=2}^{n}\sigma_l^2-1\right) \text{ is zero,}\\
=&\mathrm{Var}\left(s_n^{-2}\sum_{l=2}^{n}\sigma_l^2\right)\\
=&s_n^{-4}\mathrm{Var}\left(\sum_{l=2}^{n}\sigma_l^2\right).
\end{aligned}
$$

$$
\begin{aligned}
&\mathrm{Var}\left(\sum_{l=2}^{n}\sigma_l^2\right)\\
=&\mathrm{Var}\left\{\sum_{l=2}^{n}\mathrm{E}(U_l^2|\mathcal{F}_{l-1})\right\}\\
=&\mathrm{Var}\left[\sum_{l=2}^{n}\mathrm{E}\left\{U_l^2|(\mathbf{X}_1,\ldots,\mathbf{X}_{l-1}),(T_1,\ldots,T_{l-1})\right\}\right]\\
\leq&\mathrm{Var}\left(\sum_{l=2}^{n}U_l^2\right)\\
=&\sum_{l=2}^{n}\mathrm{Var}(U_l^2)+2\sum_{l<s}\mathrm{Cov}(U_l,U_s)\\
=&\sum_{l=2}^{n}\mathrm{Var}(U_l^2), \text{ By Lemma 4.6.5 (3), the second term is zero,}\\
=&\sum_{l=2}^{n}\mathrm{Var}\left(\sum_{1\leq j<l}q_{l,j}^2+\sum_{1\leq i<j<l}q_{l,j}q_{l,i}\right)
\end{aligned}
$$

Similarly, the covariance between the above two summation terms within variance opera-

tion are zero due to Lemma 4.6.5, so we have

$$\begin{aligned}
\mathrm{Var}\left(\sum_{l=2}^{n}\sigma_l^2\right) =& \sum_{l=2}^{n}\left\{\mathrm{Var}\left(\sum_{1\le j<l} q_{l,j}^2\right)+\mathrm{Var}\left(\sum_{1\le i<j<l} q_{l,j}q_{l,i}\right)\right\}\\
=& \sum_{l=2}^{n}\left\{\mathrm{Var}\left(\sum_{1\le j<l} q_{l,j}^2\right)+\sum_{1\le i<j<l}\mathrm{Var}\left(q_{l,j}q_{l,i}\right)\right\}\\
=& \sum_{l=2}^{n}\mathrm{Var}\left(\sum_{1\le j<l} q_{l,j}^2\right)+\sum_{1\le i<j<l\le n}\mathrm{Var}(q_{l,j}q_{l,i})\\
=& O(n^3)+O(n^3)
\end{aligned}$$

$s_n^{-4}=O(n^{-4})$, so that $\lim_{n\to\infty}\mathrm{E}|s_n^{-2}(\sum_{l=2}^{n}\sigma_l^2-1)|^2=0$. □

Based on Lemma 4.6.5, Lemma 4.6.6, and Lemma 4.6.7, we conclude that $\lim_{n\to\infty} P(G_n\le s_n x)=\Phi(x)$.

4.6.7 Proof of Theorem 4.2.3

Theorem. *Under Assumptions 1 and the mean value theorem in calculus, we can derive that* $\mathbf{G}_{n,\widehat{\boldsymbol{\beta}}}/n^2=\mathbf{G}_{n,\boldsymbol{\beta}_0}/n^2+\widetilde{\mathbf{B}}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)$. *With Lemma 4.6.4, we further conclude that*

$$n(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)\xrightarrow{d} N(\mathbf{0},\mathbf{B}^{-1}\mathbf{V}(\mathbf{B}^{-1})^{\intercal}),\quad \text{as } n\to\infty,$$

where $\mathbf{B}$ *is the limiting matrix of* $\widetilde{\mathbf{B}}$*, which is the derivative of* $\mathbf{G}_{n,\boldsymbol{\beta}}$ *evaluating at a value,* $\widetilde{\boldsymbol{\beta}}$*, between* $\boldsymbol{\beta}_0$ *and* $\widehat{\boldsymbol{\beta}}$*.* $\mathbf{V}$ *is the variance covariance matrix derived in Lemma 4.6.4.*

Proof. By dominated convergence theorem and consistency of $\widehat{\boldsymbol{\beta}}$, we can get that $\widetilde{\mathbf{B}}=\mathbf{B}+o_p(1)$, where $\widetilde{\mathbf{B}}=\frac{\partial(G_{\boldsymbol{\beta}}/n^2)}{\partial\boldsymbol{\beta}}|_{\boldsymbol{\beta}=\widetilde{\boldsymbol{\beta}}}$ and $\widetilde{\boldsymbol{\beta}}$ is the intermediate value between $\boldsymbol{\beta}_0$ and $\widehat{\boldsymbol{\beta}}$. We can estimate $\mathbf{B}$ by $\frac{\partial(G_{\boldsymbol{\beta}}/n^2)}{\partial\boldsymbol{\beta}}|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}}$.

By the assumptions and asymptotic normality of $\mathbf{G}_{\boldsymbol{\beta}_0}/s_n$, it is implied that $G_{\widehat{\boldsymbol{\beta}}} = 0$ with probability approaching one. Hence

$$0 = n\frac{\mathbf{G}_{\boldsymbol{\beta}_0}}{n^2} + \widetilde{\mathbf{B}}n(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$$
$$n(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) = -\widetilde{\mathbf{B}}^{-1}n\frac{\mathbf{G}_{\boldsymbol{\beta}_0}}{n^2} \xrightarrow{d} N(\mathbf{0}, \mathbf{B}^{-1}V\left(\mathbf{B}^{-1}\right)^{\intercal}).$$

□

4.6.8 Proof of Theorem 4.2.4

Theorem. *Under Assumptions 1 and the mean value theorem, there is a vector* $\widetilde{\mathbf{h}}(\widetilde{\boldsymbol{\beta}})^{\intercal}$ *such that* $\sqrt{n}\widehat{\mu}_{\widehat{\boldsymbol{\beta}}} = \sqrt{n}\widehat{\mu}_{\boldsymbol{\beta}_0} + \widetilde{\mathbf{h}}(\widetilde{\boldsymbol{\beta}})^{\intercal}\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$ *we can further conclude that*

$$\sqrt{n}(\widehat{\mu}_{\widehat{\boldsymbol{\beta}}} - \mu) \xrightarrow{d} N(0, \Omega_\mu), \quad \text{as } n \to \infty,$$

where Ω_μ *is the asymptotic variance.*

Proof. $\widetilde{\mathbf{h}}(\widetilde{\boldsymbol{\beta}}) = \mathbf{h}_0 + o_p(1)$ by dominated convergence theorem.

$$\begin{aligned}
\sqrt{n}(\widehat{\mu}_{\widehat{\boldsymbol{\beta}}} - \mu) &= \sqrt{n}(\widehat{\mu}_{\boldsymbol{\beta}_0} - \mu) + \widetilde{\mathbf{h}}(\widetilde{\boldsymbol{\beta}})^{\intercal}\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \\
&\quad \text{Since} \quad n(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{d} N(\mathbf{0}, \mathbf{B}^{-1}V\left(\mathbf{B}^{-1}\right)^{\intercal}), \\
&\quad \text{we have} \quad \sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{d} 0, \text{ and } \sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) = o_p(1). \text{ So} \\
\sqrt{n}(\widehat{\mu}_{\widehat{\boldsymbol{\beta}}} - \mu) &= \sqrt{n}(\widehat{\mu}_{\boldsymbol{\beta}_0} - \mu) + \mathbf{h}_0^{\intercal}o_p(1) \\
&= \sqrt{n}(\widehat{\mu}_{\boldsymbol{\beta}_0} - \mu) + o_p(1)
\end{aligned}$$

Hence, $\sqrt{n}(\widehat{\mu}_{\widehat{\boldsymbol{\beta}}} - \mu) \xrightarrow{d} N(0, \Omega_\mu)$, where $\Omega_\mu = \mathrm{E}\left[\left\{\frac{T_iY_i}{\pi_{i,0}} - \frac{(1-T_i)Y_i}{1-\pi_{i,0}}\right\}^2\right] - \mu^2$, which can be estimated by sandwich variance.

□

Chapter 5

Multiple Robust Estimation of Causal Quantile Treatment Effects

In addition to estimate the average causal effect (ACE) or average causal effect in the treated (ACET), other quantities such as quantiles may be of interest. This is a popular problem in both health science and economics literature where a policy-maker may be interested in the effect of treatment on the lower tail of the distributions of potential outcomes (Firpo, 2007). For an example in medical study, the Consortium on Safe Labour (CSL) is a large observational study designed to describe contemporary labour progression in the United States (Zhang et al., 2010). The research question is to examine the causal effect of epidural analgesia on the duration of the second stage of labour. Because of the skewed distributions of outcomes, the median or other quantiles may be more appropriate measures than the mean and variance. Obstetricians are particularly interested in higher percentiles (e.g., 95th) of the labour duration potential outcome (Zhang et al., 2012). Zhang et al. (2012) proposed a set of quantile treatment effect (QTE) estimators, such as outcome regression estimator, inverse probability weighting (IPW) estimator, stratified estimator, double robust (DR) estimator, and hybrid estimator using the propensity score and empirical cumulative distribution function. Xu et al. (2017) also proposed a QTE estimator using a Bayesian additive regression tree to estimate the propensity score and a Dirichlet process mixture of normals model to estimate the density function.

Han and Wang (2013) proposed a multiple robust estimator for missing data problems. Multiple nonresponse models and multiple imputation models are also fitted for nonresponse in surveys (Chen and Haziza, 2017). Based on the idea of multiple robustness, we also propose a multiple robust method for estimating marginal quantiles of potential outcomes and the quantile treatment effect by achieving mean balance in (1) the propensity score, and (2) the conditional distributions of potential outcomes. An estimating equations approach can be employed if we employ a parametric propensity score model. We can also use empirical likelihood or entropy balancing approaches if we want to estimate the weights directly for each observation without modelling the propensity score model.

In this chapter, we aim to achieve balance in the conditional distributions of outcomes between treated and control groups and estimate the QTE or quantile treatment effect in the treated (QTET) simultaneously. An empirical likelihood or entropy measure approach can be utilized instead of using inverse probability weighting.

5.1 Framework for Quantile Treatment Effect

In this chapter, we use p to denote the probability level and m to denote the dimension of the covariate vector. Let $F_t(\cdot)$ be the marginal cumulative distribution function (CDF) of $Y(t)$, the potential outcome under treatment $T = t$. Let $F_{t|s}(\cdot)$ be the conditional distribution of $Y(t)$ given $T = s$. In a binary treatment setting, the quantile treatment effect for the $100 \times p$th $(0 < p < 1)$ percentile (p^{th} quantile) is defined as the difference in population quantiles between potential outcomes:

$$\delta_p = F_1^{-1}(p) - F_0^{-1}(p) = \xi_{p,1} - \xi_{p,0}$$

where $F_t^{-1}(p) = \inf\{q : F_t(q) \geq p\}$ for $t = 0, 1$.

The quantile treatment effect in the treated for the $100 \times p$th percentile is defined as:

$$\delta_{p|1} = F_{1|1}^{-1}(p) - F_{0|1}^{-1}(p).$$

$F_{1|1}(q)$ can be estimated using the empirical CDF without adjusting but not $F_{0|1}(q)$.

QTE or QTET evaluated at different p levels are not necessarily equivalent. Let us take QTE as our quantity of interest. Within the potential outcome framework, we can only estimate $F_{1|1}(q)$ or $F_{0|0}(q)$ based on the observed outcomes. We define the difference in quantiles (DIQ) as:

$$\Delta_p = F_{1|1}^{-1}(p) - F_{0|0}^{-1}(p).$$

We can now see that DIQ would equal QTE if

$$F_{1|1}(p) = F_1(p) \tag{5.1}$$

$$F_{0|0}(p) = F_0(p). \tag{5.2}$$

5.2 Proposed Approach

The DIQ can be modified to equal the QTE or QTET with appropriate weights adjustment. Let $G_t(q|\mathbf{x})$ be the distribution function of $Y(t)$ conditioned on covariate vector $\mathbf{x}$ with dimension m. Let $H(\mathbf{x})$ be the joint CDF of the covariates and $H(\mathbf{x}|T = t)$ be the conditional CDF of $\mathbf{X}|T = t$, then:

$$\begin{aligned} F_t(q) &= P(Y(t) \leq q) \\ &= \mathrm{E}\left[\mathrm{E}\left[I\{Y(t) \leq q|\mathbf{x}\}\right]\right] \\ &= \mathrm{E}\left[P\left\{Y(t) \leq q|\mathbf{x}\right\}\right] \\ &= \int G_t(q|\mathbf{x}) dH(\mathbf{x}) \\ &= \mathrm{E}_{\mathbf{x}}\left\{G_t(q|\mathbf{x})\right\}, \end{aligned}$$

and

$$\begin{aligned} F_{t|t}(q) &= P\left\{Y(t) \leq q|T = t\right\} \\ &= \int G_t(q|\mathbf{x}) dH(\mathbf{x}|T = t) \\ &= \mathrm{E}_{\mathbf{x}|T=t}\left\{G_t(q|\mathbf{x})\right\}, \quad t = 0, 1. \end{aligned}$$

The Equations (5.1) and (5.2) implies that

$$\mathrm{E}_{\mathbf{x}}\left\{G_t(q|\mathbf{x})\right\} = \mathrm{E}_{\mathbf{x}|T=t}\left\{G_t(q|\mathbf{x})\right\}, \quad \text{for} \quad t = 0, 1.$$

We make $F_t(q)$ and $F_{t|t}(q)$ equivalent so that we can estimate QTE by estimating DIQ. We can achieve the mean balance of $G_t(q|\mathbf{x})$ by weighting when estimating the QTE. Hence, the sample condition for the above balancing condition is given by

$$\frac{1}{n}\sum_{i=1}^{n} G_t(q|\mathbf{x}_i) = \sum_{i\in S_t} G_t(q|\mathbf{x}_i)w_i, \quad t = 0, 1, \tag{5.3}$$

where $S_t = \{i : T_i = t, i = 1, \ldots, n\}$ is the set of observations for treatment or control groups. The conditional distribution function $G_t(q|\mathbf{x})$ can be modelled by

$$G_t(q|\mathbf{x}, \boldsymbol{\beta}_t) = \Phi(q - \mathbf{g}(\mathbf{x})^\intercal \boldsymbol{\beta}_t)$$

for $t = 0, 1$ where $\Phi(\cdot)$ is the CDF of standard normal distribution. Here $\mathbf{g}(\mathbf{x})$ is a vector function of covariates. In addition to standard normal CDF, other CDF functions are available and standard normal CDF through Box-Cox transformation can also be applied (Zhang et al., 2010). We can estimate $G_t(q|\mathbf{x})$ as an empirical CDF with parameter $\boldsymbol{\beta}$.

In addition to having constraints on $G_t(q|\mathbf{x})$, we may also want to incorporate information from a propensity score model. Let $\pi^l(\mathbf{x})$ be an arbitrary propensity score model and $\pi(\mathbf{x})$ be the true propensity score model. We have the following fact:

$$\mathrm{E}\left[\frac{T}{\pi(\mathbf{x})}\left[\pi^l(\mathbf{x}) - \mathrm{E}\{\pi^l(\mathbf{x})\}\right]\right] = 0.$$

It can be easily verified that the above equation can be simplified to

$$\mathrm{E}\left[\frac{1}{\pi(\mathbf{x})}\left[\pi^l(\mathbf{x}) - \mathrm{E}\{\pi^l(\mathbf{x})\}\right] | T = 1\right] = 0. \tag{5.4}$$

Replacing $\frac{1}{\pi(\mathbf{x})}$ by a general weight w_i which sum to 1 in set S_t, we can have the empirical version of Equation (5.4):

$$\sum_{i\in S_t} w_i \pi^l(\mathbf{x}_i) = \frac{1}{n}\sum_{i=1}^{n} \pi^l(\mathbf{x}_i). \tag{5.5}$$

Equation (5.5) is the balancing condition on arbitrary propensity score model, which is also derived in Qin and Zhang (2007) and Han and Wang (2013) and similar to the covariate balancing equation in covariate balancing propensity score (Imai and Ratkovic,

2014). Based on the idea of multiple robust estimator, combining Equations (5.3) for multiple candidate conditional CDFs of $Y(t)$ using different forms of $\mathbf{g}(\mathbf{x})$ and balancing constraints (5.5) on multiple candidate propensity score models, we use the empirical likelihood approach to estimate w_i by maximizing $\Pi_{i\in S_t}^n w_i$ subject to balancing constraints:

$$\sum_{i\in S_t}^{n} w_i = 1, \tag{5.6}$$

$$\frac{1}{n}\sum_{i=1}^{n} G_t^k(\widehat{q}_{p,t}^k|\mathbf{x}_i, \widehat{\boldsymbol{\beta}}^k) = \sum_{i\in S_t} G_t^k(\widehat{q}_{p,t}^k|\mathbf{x}_i, \widehat{\boldsymbol{\beta}}^k)w_i, \quad k = 1,\ldots,K, \tag{5.7}$$

$$\frac{1}{n}\sum_{i=1}^{n} \pi^l(\widehat{\boldsymbol{\alpha}}^l, \mathbf{x}_i) = \sum_{i\in S_t} \pi^l(\widehat{\boldsymbol{\alpha}}^l, \mathbf{x}_i)w_i, \quad l = 1,\ldots,L, \tag{5.8}$$

when $t = 1$ or $t = 0$, we can estimate the weights for the treatment units or control units separately. Here $\pi^l(\widehat{\boldsymbol{\alpha}}^l, \mathbf{x}_i)$ denotes the lth candidate propensity score model with estimated parameter $\widehat{\boldsymbol{\alpha}}^l$. $G_t^k(\widehat{q}_{p,t}^k|\mathbf{x}_i, \widehat{\boldsymbol{\beta}}^k)$ denotes the kth candidate conditional CDF with the corresponding estimated $\widehat{\boldsymbol{\beta}}^k$ and estimated p^{th} quantile $\widehat{q}_{p,t}^k$ of kth conditional CDF solved from $\frac{1}{n}\sum_{i\in S_t}\Phi\left(q-(1,\mathbf{x}_i^\intercal)\widehat{\boldsymbol{\beta}}^k\right) = p$. Let n_t be the number of observations in treatment (t=1) or control (t=0) groups.

By the Lagrange multiplier method, we can find that

$$\widehat{w}_i = \frac{1}{n_t}\frac{1}{1+\widehat{\boldsymbol{\lambda}}^\intercal\widehat{\mathbf{r}}_i(\widehat{\boldsymbol{\alpha}},\widehat{\boldsymbol{\beta}},\widehat{\mathbf{q}}_{p,t})} \Bigg/ \sum_{i\in S_t}\frac{1}{n_t}\frac{1}{1+\widehat{\boldsymbol{\lambda}}^\intercal\widehat{\mathbf{r}}_i(\widehat{\boldsymbol{\alpha}},\widehat{\boldsymbol{\beta}},\widehat{\mathbf{q}}_{p,t})}.$$

A simple derivation can show that the denominator above is 1 and simplify the above equation:

$$\widehat{w}_i = \frac{1}{n_t}\frac{1}{1+\widehat{\boldsymbol{\lambda}}^\intercal\widehat{\mathbf{r}}_i(\widehat{\boldsymbol{\alpha}},\widehat{\boldsymbol{\beta}},\widehat{\mathbf{q}}_{p,t})}.$$

By defining $\widehat{\psi}^k = \frac{1}{n}\sum_{i=1}^n G_t^k(\widehat{q}_{p,t}^k|\mathbf{x}_i, \widehat{\boldsymbol{\beta}}^k)$, $\widehat{\theta}^l = \frac{1}{n}\sum_{i=1}^n \pi^l(\widehat{\boldsymbol{\alpha}}^l, \mathbf{x}_i)$, we get

$$\widehat{\mathbf{r}}_i(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}, \widehat{\mathbf{q}}_{p,t}) = \begin{Bmatrix} \pi^1(\widehat{\boldsymbol{\alpha}}^1, \mathbf{x}_i) - \widehat{\theta}^1 \\ \vdots \\ \pi^L(\widehat{\boldsymbol{\alpha}}^L, \mathbf{x}_i) - \widehat{\theta}^L \\ G_t^1(\widehat{q}_{p,t}^1|\mathbf{x}_i, \widehat{\boldsymbol{\beta}}^1) - \widehat{\psi}^1 \\ \vdots \\ G_t^K(\widehat{q}_{p,t}^K|\mathbf{x}_i, \widehat{\boldsymbol{\beta}}^K) - \widehat{\psi}^K \end{Bmatrix},$$

where $\widehat{\boldsymbol{\alpha}} = (\widehat{\boldsymbol{\alpha}}^1, \ldots, \widehat{\boldsymbol{\alpha}}^L)^\intercal$, $\widehat{\boldsymbol{\beta}} = (\widehat{\boldsymbol{\beta}}^1, \ldots, \widehat{\boldsymbol{\beta}}^K)^\intercal$, and $\widehat{\mathbf{q}}_{p,t} = (\widehat{q}_{p,t}^1, \ldots, \widehat{q}_{p,t}^K)^\intercal$. Here $\widehat{\boldsymbol{\lambda}}^\intercal = (\widehat{\lambda}_1, \ldots, \widehat{\lambda}_{K+L})^\intercal$ satisfies

$$\sum_{i \in S_t} \frac{1}{n_t} \frac{\widehat{\mathbf{r}}_i(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}, \widehat{\mathbf{q}}_{p,t})}{1 + \widehat{\boldsymbol{\lambda}}^\intercal \widehat{\mathbf{r}}_i(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}, \widehat{\mathbf{q}}_{p,t})} = \mathbf{0}. \tag{5.9}$$

Equation (5.9) may have multiple roots for $\boldsymbol{\lambda}^\intercal$. For implementation, an easy way to solve Equation (5.9) is to find the unique $\boldsymbol{\lambda}^\intercal$ that minimizes

$$H = -\sum_{i \in S_t} \log\left\{1 + \widehat{\boldsymbol{\lambda}}^\intercal \widehat{\mathbf{r}}_i(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}, \widehat{\mathbf{q}}_{p,t})\right\}$$

under the constraint $1 + \widehat{\boldsymbol{\lambda}}^\intercal \widehat{\mathbf{r}}_i(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}, \widehat{\mathbf{q}}_{p,t}) > 0$, where H is a convex function. The derivative of H is proportional to the left hand side of Equation (5.9). Finally the estimated $\widehat{\delta}_p$ is derived

$$\widehat{\delta}_p = \widehat{\xi}_{p,1} - \widehat{\xi}_{p,0}$$

and covariate balance is achieved simultaneously, where $\widehat{\xi}_{p,t} = \widehat{F}_{n,t}^{-1}(p) = \inf\{q : \widehat{F}_{n,t}(q) \geq p\}$ and $\widehat{F}_{n,t}(q) = \sum_{i \in S_t} \widehat{w}_i I(y_i \leq q)$, $t = 0, 1$.

5.2.1 Entropy Measure Approach

In addition to empirical likelihood, other distance measures like entropy measure are also available. We propose to use the entropy measure as the objective function under the same

sets of constraints (5.6), (5.7), and (5.8). The entropy measure to minimize is defined as:

$$L = \sum_{i \in S_t} w_i \log(w_i/p_i) \tag{5.10}$$

under initial weights p_i. The loss function, $w \log(w/p)$, belongs to a general class of empirical minimum discrepancy estimators defined by Cressie–Read (CR) divergence (Read and Cressie, 2012). The entropy measure approach involves a reweighting scheme that searches for the optimal set of weights. The entropy measure is minimized when $w_i = p_i$. By minimizing the entropy measure, the scheme searches for a set of weights that is adjusted far enough to satisfy the balancing constraints and as close as possible to the initial weights in order to retain efficiency (Hainmueller, 2012). Normally, we set uniform initial weights: $p_i = 1/n_t$. With the Lagrange multiplier method, we can also derive the optimal weight:

$$\widehat{w}_i = p_i e^{-\widehat{\boldsymbol{\lambda}}^\intercal \widehat{\mathbf{r}}_i} \Big/ \sum_{i \in S_t} p_i e^{-\widehat{\boldsymbol{\lambda}}^\intercal \widehat{\mathbf{r}}_i},$$

where $\widehat{\mathbf{r}}_i = \widehat{\mathbf{r}}_i(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}, \widehat{\mathbf{q}}_{p,t})$. The $\widehat{\boldsymbol{\lambda}}^\intercal$ satisfies

$$\sum_{i \in S_t} p_i e^{-\widehat{\boldsymbol{\lambda}}^\intercal \widehat{\mathbf{r}}_i} \widehat{\mathbf{r}}_i = \mathbf{0}.$$

The Lagrange multiplier can also be solved in a similar way. With the optimal weights, we can estimate the QTE through weighted empirical cumulative distribution functions of observed outcomes of treatment and control groups.

5.3 Theoretical Properties

The derivations here are based on weights derived under empirical likelihood method, although the theoretical results can also be extended to the entropy measure approach.

5.3.1 Consistency of the Quantile Estimator

The following theorems establish the consistency of the estimated quantile of $F_t(\xi)$: $\widehat{\xi}_{p,t} = \inf\{q : \widehat{F}_{n,t}(q) \geq p\}$. Hence, the corresponding QTE estimator, $\widehat{\delta}_p$, is also consistent under the conditions. The theorems can also be applied to the QTET estimator.

Theorem 5.3.1. *If one of the propensity score models, $\pi^l(\mathbf{x})$ for $l = 1, \ldots, L$, is correctly specified, $\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}$, and $\widehat{\mathbf{q}}_{p,t}$ are estimated from their corresponding specified models, then $\widehat{\xi}_{p,t} \xrightarrow{p} \xi_{p,t}$ as $n \to \infty$ for $t = 0, 1$.*

The detailed derivation can be found in Section 5.7. The following theorem shows the consistency when one of the conditional CDFs is correctly specified.

Theorem 5.3.2. *If one of the conditional CDFs, $G_t^k(q|\mathbf{x})$ for $k = 1, \ldots, K$, is correctly specified, $\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}$, and $\widehat{\mathbf{q}}_{p,t}$ are estimated from their corresponding specified models, then $\widehat{\xi}_{p,t} \xrightarrow{p} \xi_{p,t}$ as $n \to \infty$ for $t = 0, 1$.*

The detailed derivation is provided in Section 5.7. The consistency of the quantile estimators can be easily extended to the consistency of QTE estimator. Combining Theorems 5.3.1 and 5.3.2, we can conclude the proposed QTE estimator is multiple robust as long as one of the models is correctly specified regardless of the correctness of all other models.

5.3.2 Asymptotic Normality of the Quantile Estimator

The asymptotic normality depends on which of the candidate propensity score model or conditional CDF is correctly specified. Under the suitable regularity conditions, we can have the asymptotic normal distributions for our estimator.

Theorem 5.3.3. *When one of the propensity score model is correctly specified, then $\widehat{\xi}_{p,t} \xrightarrow{d} N\left(0, \frac{\sigma_F^2}{f_t^2(\xi_{p,t})}\right)$ as $n \to \infty$ for $t = 0, 1$.*

Here σ_F^2 is the asymptotic variance of empirical CDF which will be introduced in Section 5.7 and $f_t(\cdot)$ is the density function of $F_t(\cdot)$. The asymptotic normality of $\widehat{\xi}_{p,t}$ depends on the asymptotic normality of empirical CDF. The asymptotic normality of empirical CDF is well established, although our estimator is based on weighted empirical CDF. We establish the asymptotic normality of weighted empirical CDF in the proof of Theorem 5.3.3 in Section 5.7.

Theorem 5.3.4. *When one of the conditional CDFs is correctly specified, then* $\widehat{\xi}_{p,t} \xrightarrow{d} N\left(0\frac{\sigma_F^2}{f_t^2(\xi_{p,t})}\right)$ *as* $n \to \infty$ *for* $t = 0, 1$.

Here σ_F^2 is the asymptotic variance of empirical CDF which will be introduced in Section 5.7 but the construction is a little bit different from σ_F^2 in Theorem 5.3.3. A sketch of proof is provided in the Subsection 5.7.4.

5.4 Simulation Studies

5.4.1 Simulation Setup

In this section, we conduct a set of simulation studies to evaluate the performance of the proposed QTE estimator under the empirical likelihood framework compared with the double robust estimator proposed by Zhang et al. (2012). The simulation setting is the combination of Stuart et al. (2013) and Kang and Schafer (2007) with some modifications. There are nine continuous covariates: four are confounders, two are related only to the treatment indicator T, and two are only related to the outcome variable Y. The last one is neither related to treatment nor outcome. There are no unmeasured confounders. The causal diagram is the same as what is shown in Figure 4.1. The six covariates related to the treatment indicator follow a mixture normal distribution: $1/2 \times N(-1, 1) + 1/2 \times N(1, 1)$ except X_1. Here X_1 and the rest of the covariates follow a $N(0, 1)$ distribution.

Three propensity score models (model 1, 2, and 3) and three outcome models (model A, B, and C) are used with different settings of $h(\mathbf{X})$ and $g(\mathbf{X})$ such that:

$$\text{logit}\left\{P(T = 1|\mathbf{X})\right\} = h(\mathbf{X})^\intercal \boldsymbol{\alpha}$$

and

$$Y = g(\mathbf{X})^\intercal \boldsymbol{\beta} + \mu X_1 * T + \epsilon.$$

where ϵ follows $N(0,1)$. The simulation propensity score models are given below:

Model 1: $\text{logit}\{P(T=1|\mathbf{X})\} = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_4 + \beta_4 X_5 + \beta_5 X_7 + \beta_6 X_8$,

Model 2: $\text{logit}\{P(T=1|\mathbf{X})\} = \beta_1 W_1 + \beta_2 X_2 + \beta_3 W_4 + \beta_4 W_5 + \beta_5 X_7 + \beta_6 W_8$,
with transformation: $W_1 = \exp(X_1/3)$, $W_4 = X_1 - X_4 + 3$,
$W_5 = X_1 X_5/10 + 0.5$, $W_8 = -\exp(X_8/2)$,

Model 3: $\text{logit}\{P(T=1|\mathbf{X})\} = \beta_1 W_1 + \beta_2 W_2 + \beta_3 W_4 + \beta_4 W_5 + \beta_5 X_7 + \beta_6 W_8$,
with transformation: $W_1 = \exp(X_1)$, $W_2 = X_2^2$, $W_4 = -2X_4/(1+\exp(X_1)) + 0.5$,
$W_5 = -X_4 X_5/2 - 2$, $W_8 = |X_8| - 1$.

The outcome models are given below:

Model A: $Y = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_3 + \alpha_4 X_4 + \alpha_5 X_5 + \alpha_6 X_6 + \mu X_1 + \epsilon$.

Model B: $Y = \alpha_0 + \alpha_1 X_1 + \alpha_2 W_2 + \alpha_3 X_3 + \alpha_4 W_4 + \alpha_5 W_5 + \alpha_6 X_6 + \alpha_1 X_1^2 + \mu X_1 + \epsilon$,
with transformation: $W_2 = \exp(X_2/2)$, $W_4 = \exp(X_4/3)$, $W_5 = X_5 X_6$.

Model C: $Y = \alpha_0 + \alpha_1 X_1 + \alpha_2 W_2 + \alpha_3 W_3 + \alpha_4 W_4 + \alpha_5 X_5 + \alpha_6 W_6 + \alpha_1 X_1^3/3 + \mu X_1 + \epsilon$,
with transformation: $W_2 = X_2^2$, $W_3 = \exp(X_3/3)$, $W_4 = |X_4|$, $W_6 = X_6^2$.

Here μ is set to 5 while other parameters, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, follow the same values as in Tables 4.1 and 4.2.

We explore the performance of our estimator using different combinations of the above models. There are five scenarios for data generation: 1A, 2A, 1B, 2B, and 3C. The first four scenarios are enough to verify the consistency of the quantile estimator and the last scenario is used to study the performance of proposed estimator when no model is correctly specified. For example, if 1A is the true data generating scenario, then 1B is enough to study the consistency when only propensity score model is correctly specified and 1C is not necessary. There is no explicit solution for the true quantile treatment effect, δ_p. To

evaluate the performance of our proposed QTE estimator, we use a Monte Carlo simulation of $10,000$ replications with a sample size of $1,000,000$ to get a numerical solution to δ_p as a benchmark. We evaluate our proposed QTE estimator and DR estimator using this benchmark. In this study, we focus on the quantile treatment effect for 50th percentile. Simulation studies on other percentiles (25th, 75th, and 95th) can also be found in Appendix A.2. There are five proposed QTE estimators and four double robust (DR) estimators. Each DR estimator includes one propensity score model and one outcome model. For example, DR_1A indicates propensity score model 1 and outcome model A are specified in this estimator. Estimator DR_1A has at least one model correctly specified in data scenarios 1A, 2A, and 1B. Our QTE estimators can include more than two models. We use a notation like QTE_12A to indicate that propensity score models 1, 2 and outcome model A are specified as constraints to construct the corresponding estimator.

5.4.2 Simulation Results

The simulation results with 5 scenarios are shown in Tables 5.1 - 5.5 with 1000 replications and a sample size of $n = 1000$. More simulation results (other quantiles, $n = 5000$, and $n = 200$) are given in Appendix A.2. We evaluated our proposed QTE estimators by including some of the correctly specified propensity score models or outcome models. By matching the scenarios, each QTE estimator has at least one model correctly specified. While for DR estimator, for example, DR_2B does not have any models correctly specified in scenario 1A. The results are evaluated using bias, empirical standard error (ESE), root mean squared error (RMSE), bootstrapped standard error (BSE), and coverage rate (CR) based on bootstrapped standard error. We include scenario 3C when studying the quantile treatment effect for 50th percentile.

First, we compared the performance in scenarios 1A, 2A, 1B, and 2B. In terms of bias, our proposed QTE estimators achieve consistency in all scenarios. DR_1A in scenario 2B, DR_2A in scenario 1B, DR_1B in scenario 2A, and DR_2B in scenario 1A are not consistent. For DR_1B in scenario 2A, the bias looks small but actually both the estimated quantiles for treatment and control happen to have nearly equal biases in the same direction. In Appendix A.2.1, we show the simulation results for other percentiles (25th, 75th, and 95th).

Some DR estimators have large biases even when one of the models is correctly specified for 25th, 75th, and 95th percentiles with sample size equal to 1000. A further study (results not shown) shows that the biases from those DR estimators approach zero when we increase the sample size to 5000. On the other hand, most of our proposed estimators always have biases close to zero no matter whether the sample size is 1000 or 5000 for low or high percentiles. The reason for the large biases is due to the sparsity of observations at 25th, 75th or 95th percentiles of the potential outcome distributions. The DR estimators show large biases when sample size is small, which is also verified by the small sample size studies in the Appendix A.2. The empirical performance on other quantiles showed that the convergence rate of QTE estimators is higher than the rate of DR estimators. QTE estimator is more robust against the sparsity of observations at the tails of the observed outcome distributions.

In terms of RMSE, QTE estimators show small RMSE or the same efficiency as most DR estimators. The QTE estimators have smaller ESE compared to DR estimators when the true outcome model is included. In terms of coverage rate for 95% bootstrapped confidence interval, all QTE estimators have coverage rates close to 95%, while the inconsistent DR estimators show much lower coverage rates.

We include the scenario 3C in Tables 5.5, A.19, and A.24 when estimating quantile treatment effect for 50th percentile in which none of the estimators are correctly specified. When no model is correctly specified, none of the DR estimators or QTE estimators are consistent. The QTE estimators generally result in smaller standard errors.

5.5 Application

In this section, we considered an investigation by Koenker and Hallock (2001) and Abrevaya (2001) on the impact of various demographic characteristics and maternal behaviour on the birthweight of infants born in United States. It is based on the June 1997 Detailed Natality Data published by National Center for Health Statistics. The smoking status of mother is considered the treatment indicator here. The birthweight of infants is the outcome variable and is recorded in grams. The quantile treatment effect is defined as the quantile difference of infants' birthweight distributions between treatment (mother smokes)

Table 5.1: Scenario 1A: Simulation Results of QTE and DR Estimators with 50th Percentile ($n = 1000$)

	BIAS	ESE	RMSE	BSE	CR
QTE_12A	0.0010	0.3592	0.3592	0.3641	0.944
QTE_12B	-0.0174	0.4790	0.4793	0.4860	0.941
QTE_1AB	0.0017	0.3589	0.3589	0.3635	0.945
QTE_2AB	0.0047	0.3505	0.3505	0.3570	0.952
QTE_12AB	0.0013	0.3590	0.3590	0.3640	0.945
DR_1A	0.0030	0.3763	0.3763	0.4029	0.964
DR_2A	-0.0059	0.3797	0.3797	0.4365	0.965
DR_1B	-0.0411	0.5180	0.5197	0.5596	0.954
DR_2B	1.0682	0.5505	1.2017	0.5674	0.514

RMSE: Root mean squared error, ESE: Empirical standard error, BSE: Bootstrapped standard error, CR: Coverage rate for 95% bootstrapped confidence interval.

Table 5.2: Scenario 2A: Simulation Results of QTE and DR Estimators with 50th Percentile ($n = 1000$)

	BIAS	ESE	RMSE	BSE	CR
QTE_12A	0.0147	0.4181	0.4183	0.4252	0.943
QTE_12B	0.0051	0.7329	0.7330	0.7059	0.937
QTE_1AB	0.0127	0.4169	0.4171	0.4239	0.948
QTE_2AB	0.0133	0.4053	0.4056	0.4207	0.948
QTE_12AB	0.0126	0.4082	0.4084	0.4245	0.947
DR_1A	0.0053	0.3467	0.3468	0.3764	0.958
DR_2A	0.0258	0.4354	0.4362	0.4780	0.968
DR_1B	0.0263	0.5025	0.5032	0.5314	0.965
DR_2B	-0.0027	0.7705	0.7706	0.8022	0.966

RMSE: Root mean squared error, ESE: Empirical standard error, BSE: Bootstrapped standard error, CR: Coverage rate for 95% bootstrapped confidence interval.

Table 5.3: Scenario 1B: Simulation Results of QTE and DR Estimators with 50th Percentile ($n = 1000$)

	BIAS	ESE	RMSE	BSE	CR
QTE_12A	0.0385	0.4757	0.4772	0.4791	0.946
QTE_12B	0.0043	0.3210	0.3210	0.3402	0.945
QTE_1AB	0.0055	0.3234	0.3235	0.3396	0.944
QTE_2AB	-0.0001	0.3149	0.3149	0.3301	0.947
QTE_12AB	0.0045	0.3240	0.3240	0.3399	0.945
DR_1A	0.0283	0.5490	0.5498	0.5922	0.956
DR_2A	0.6621	1.7098	1.8335	0.8085	0.841
DR_1B	0.0067	0.3886	0.3886	0.3895	0.970
DR_2B	0.0034	0.3207	0.3208	0.3671	0.967

RMSE: Root mean squared error, ESE: Empirical standard error, BSE: Bootstrapped standard error, CR: Coverage rate for 95% bootstrapped confidence interval.

Table 5.4: Scenario 2B: Simulation Results of QTE and DR Estimators with 50th Percentile ($n = 1000$)

	BIAS	ESE	RMSE	BSE	CR
QTE_12A	0.0179	0.5719	0.5721	0.5993	0.954
QTE_12B	-0.0073	0.3674	0.3674	0.3813	0.948
QTE_1AB	-0.0010	0.3733	0.3733	0.3800	0.945
QTE_2AB	-0.0032	0.3621	0.3621	0.3761	0.943
QTE_12AB	-0.0044	0.3662	0.3662	0.3799	0.949
DR_1A	0.6086	0.4696	0.7687	0.5145	0.803
DR_2A	0.0224	0.6667	0.6670	0.7838	0.973
DR_1B	0.0026	0.3101	0.3101	0.3518	0.972
DR_2B	0.0223	0.4385	0.4391	0.4873	0.976

RMSE: Root mean squared error, ESE: Empirical standard error, BSE: Bootstrapped standard error, CR: Coverage rate for 95% bootstrapped confidence interval.

Table 5.5: Scenario 3C: Simulation Results of QTE and DR Estimators with 50th Percentile ($n = 1000$)

	BIAS	ESE	RMSE	BSE	CR
QTE_12A	-0.8458	0.6800	1.0852	0.7051	0.774
QTE_12B	-0.9043	0.7452	1.1717	0.7588	0.779
QTE_1AB	-0.7007	0.7018	0.9917	0.7168	0.826
QTE_2AB	-0.9568	0.6495	1.1564	0.6770	0.702
QTE_12AB	-0.8383	0.6788	1.0786	0.7034	0.780
DR_1A	-0.6531	0.7186	0.9710	0.8126	0.888
DR_2A	-0.8990	0.8478	1.2357	1.0524	0.886
DR_1B	-0.4525	0.8088	0.9268	0.8989	0.839
DR_2B	-0.8843	0.9714	1.3137	1.1203	0.902

RMSE: Root mean squared error, ESE: Empirical standard error, BSE: Bootstrapped standard error, CR: Coverage rate for 95% bootstrapped confidence interval.

and control (mother does not smoke) groups. Previous studies show that low birthweight is associated with a wide range of subsequent health issues, economic issues due to high cost, and long duration of health care (Koenker and Hallock, 2001). Consequently, there has been a great deal of interest in studying the low tail quantile treatment effect on infant birthweight. The empirical distributions of birthweight between treatment and control groups are shown in Figure 5.1.

There is a total of $50,000$ observations in the data set with eight covariates. The eight covariates include: Married (mother's marriage status), Black (mother's race recorded as either black or white), Boy (gender of the infant, 1 for boy), MomAge (mother's age), CigsPerDay (number of cigarettes smoked per day), MomWtGain (mother's pregnancy weight gain), MomEdLevel (mother's education level: less than high school, high school, some college, and college graduate), and Visit (prenatal visit with four categories: no visit, first visit in the first trimester of the pregnancy, first visit in the second trimester of the pregnancy, and first visit in the last trimester). CigsPerDay is removed from analysis due to the collinearity with treatment indicator. We assumed all confounders related to

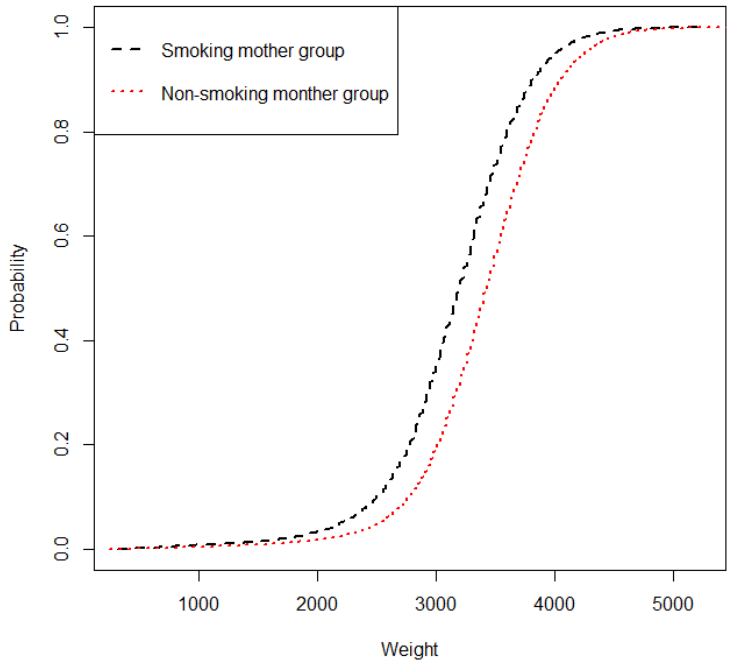

Figure 5.1: Infant Birthweight Distributions for Smoking and Non-smoking Mother Groups

the treatment indicator and outcome variable are among the basic demographic variables and all other measured covariates. For example, MomEdLevel, as a socioeconomic factor, usually will affect both mothers' smoking status and their babies' birthweight. A direct comparison of infants birthweight between treatment and control groups is not reliable due to the existence of confounders.

We evaluated our proposed QTE estimator in comparison with the DR estimator. We employ only one QTE estimator which includes all the propensity score models and outcome models. For the propensity score models, three link functions are employed including logit link, cloglog link, and probit link. For each link function, we also have three different model specifications. For example, propensity score model is denoted as PS1, PS2, and PS3 models under logit link. PS1 model only includes main effects; PS2 model has a few more squared and interaction terms compared to PS1 model; PS3 model has more squared and interaction terms compared to PS2 model. Similarly, we apply the same model specifications under each of the other two link functions (PS4-PS6 under probit link and PS7-PS9 under cloglog link). Hence, we have a total of nine propensity score models.

For outcome models, we apply the same three model specifications described in the propensity score models for the deterministic component of our outcome models. We only use the identity link function for outcome models. The outcome models are denoted as: OCM1, OCM2, and OCM3. A histogram of the outcome variable shows that the birthweight is normally distributed so that any other link function is not useful. We conducted preliminary studies on comparing the propensity scores generated by different propensity score models (results not shown). The three propensity score models using the same link function generally generated quite different propensity score for each other. The propensity scores generated from models with the same model deterministic component but different link functions are very close to each other. On the other hand, the DR estimator can only include one propensity score model and outcome model. So we specified three different DR estimators based on different link functions used for propensity score model: DR_logit, DR_probit, and DR_cloglog. Each DR estimator is a combination of PS3, PS6, and PS9 combined with OCM3.

The results are shown in Table 5.6 focusing on 50th percentile with bootstrapped standard errors based on 500 bootstrapped samples. The quantile treatment effect estimates

Table 5.6: Quantile Treatment Effect Estimate for 50th Percentile with Birthweight Data

Method	Estimate	BSE	CI
QTE	-206.0	15.8	(-237.0, -175.0)
DR_logit	-194.0	22.3	(-237.7, -150.3)
DR_probit	-218.8	22.8	(-263.5, -174.1)
DR_cloglog	-215.0	21.7	(-257.5, -172.5)

BSE: Bootstrapped standard error, CI: 95% bootstrapped confidence interval.

from QTE and all DR estimators are within $(-220, -190)$. The negative value of QTE estimate indicates that the 50th percentile of the infant birthweight if the mother smokes is about 206 grams smaller than the 50th percentile of infant birthweight if the mother does not smoke. All the methods have no zero included in their 95% bootstrapped confidence intervals which indicates a significant effect between mother's smoking behaviour and the baby's birthweight. Due to the inclusion of more propensity score models and outcome models, our proposed QTE method is more efficient than DR estimator here. We also study the quantile treatment effects ranging from 5th to 95th percentiles shown in Figure 5.2 with 95% confidence bounds for our proposed QTE method. The confidence bounds for our proposed method are much wider at lower or higher tail quantile treatment effects. The lower tail quantile treatment effect is about 20-30 grams smaller compared to 50th quantile treatment effect but the difference is not significant based on the figure.

5.6 Conclusion

The current statistical literature focuses on the population mean of potential outcomes. We propose a quantile treatment effect estimator which achieves balance in the covariates using a set of conditions on: (1) the propensity score, and (2) the conditional distributions of potential outcomes. There are several nice features of our proposed methods. Our proposed method shows more protection against model misspecification compared to DR estimator. When the true outcome model is correctly specified, our proposed QTE estimator achieves

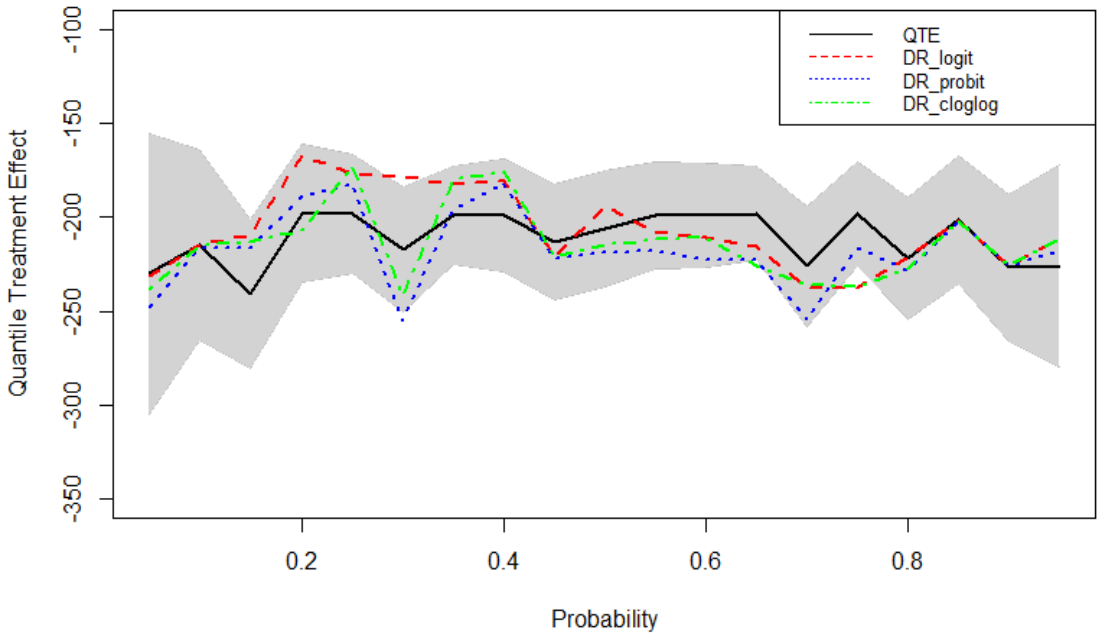

Figure 5.2: Quantile Treatment Effect Estimates for Different Probability Levels

a higher efficiency compared to DR estimator. Our proposed estimators also have faster convergence rates compared to DR estimators since many DR estimators show biased results when estimating low or high percentiles at small sample sizes even with true model included. Besides, our proposed estimators tend to have good simulation performance with higher efficiency even when no model is correctly specified as pointed out by Han and Wang (2013) and verified by our simulation studies.

Both empirical likelihood and entropy measure can be used with the same theoretical properties. Some other measures between two probability measures may also be utilized including kernel distance, Kullback-Leibler distance, chi-squared distance, and generalized pseudo empirical likelihood (Tan and Wu, 2015) .

5.7 Theorems and Proofs

5.7.1 Proof of Theorem 5.3.1

Theorem. *If one of the propensity score models, $\pi^l(\mathbf{x})$ for $l = 1, \ldots, L$, is correctly specified, $\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}$, and $\widehat{\mathbf{q}}_{p,t}$ are estimated from their corresponding specified models, then $\widehat{\xi}_{p,t} \xrightarrow{p} \xi_{p,t}$ as $n \to \infty$ for $t = 0, 1$.*

We prove this theorem for $t = 1$, the derivation is same for $t = 0$ and we assume a true propensity score model $\pi^1(\boldsymbol{\alpha}_0^1) = \pi(\mathbf{x})$ and $\boldsymbol{\alpha}_0^1$ is the true parameter in the correctly specified propensity score model. For simplification, let $\pi_i^1(\widehat{\boldsymbol{\alpha}}^1) = \pi^1(\widehat{\boldsymbol{\alpha}}^1, \mathbf{x}_i)$.

Proof. We reparameterize the Lagrange multiplier $\widehat{\boldsymbol{\lambda}}^\intercal$ by $\widehat{\boldsymbol{\lambda}}^\intercal = (\frac{\widehat{\tau}_1+1}{\widehat{\theta}^1}, \frac{\widehat{\tau}_2}{\widehat{\theta}^1}, \ldots, \frac{\widehat{\tau}_{K+L}}{\widehat{\theta}^1}) = \frac{1}{\widehat{\theta}^1}\{(1, 0, \ldots, 0) + \widehat{\boldsymbol{\tau}}^\intercal\}$, where $\widehat{\theta} = \sum_{i=1}^n \pi_i^1(\widehat{\boldsymbol{\alpha}}^1)/n$ and $\theta_0^1 = \mathrm{E}\{\pi^1(\boldsymbol{\alpha}_0^1)\} = P(T = 1)$.

From Equation (5.9), we have

$$
\begin{aligned}
\mathbf{0} &= \sum_{i \in S_1} \frac{1}{n_1} \frac{\widehat{\mathbf{r}}_i(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}, \widehat{\mathbf{q}}_{p,1})}{1 + \widehat{\boldsymbol{\lambda}}^\intercal \widehat{\mathbf{r}}_i(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}, \widehat{\mathbf{q}}_{p,1})} \\
&= \sum_{i \in S_1} \frac{1}{n_1} \frac{\widehat{\mathbf{r}}_i(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}, \widehat{\mathbf{q}}_{p,1})}{1 + \dfrac{\pi_i^1(\widehat{\boldsymbol{\alpha}}^1) - \widehat{\theta}^1}{\widehat{\theta}^1} + \dfrac{\widehat{\boldsymbol{\tau}}^\intercal}{\widehat{\theta}^1} \widehat{\mathbf{r}}_i(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}, \widehat{\mathbf{q}}_{p,1})} \\
&= \sum_{i \in S_1} \frac{1}{n_1} \frac{\widehat{\mathbf{r}}_i(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}, \widehat{\mathbf{q}}_{p,1})}{1 + \dfrac{\pi_i^1(\widehat{\boldsymbol{\alpha}}^1) - \widehat{\theta}^1}{\widehat{\theta}^1} + \dfrac{\widehat{\boldsymbol{\tau}}^\intercal}{\widehat{\theta}^1} \widehat{\mathbf{r}}_i(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}, \widehat{\mathbf{q}}_{p,1})} \frac{\widehat{\theta}^1}{\widehat{\theta}^1} \\
&= \widehat{\theta}^1 \frac{n}{n_1} \frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{T_i}{\pi_i^1(\widehat{\boldsymbol{\alpha}}^1)} \frac{\widehat{\mathbf{r}}_i(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}, \widehat{\mathbf{q}}_{p,1})}{1 + \widehat{\boldsymbol{\tau}}^\intercal \widehat{\mathbf{r}}_i(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}, \widehat{\mathbf{q}}_{p,1}) / \pi_i^1(\widehat{\boldsymbol{\alpha}}^1)} \right\} \\
&= D_n(\widehat{\boldsymbol{\tau}}^\intercal) \xrightarrow{p} D(\boldsymbol{\tau}^\intercal), \\
&\text{since } \widehat{\theta}^1 \xrightarrow{p} P(T=1) \text{ and } n/n_1 \xrightarrow{p} 1/P(T=1).
\end{aligned}
$$

Using the results of White (1982), $\widehat{\boldsymbol{\alpha}}^l \xrightarrow{p} \boldsymbol{\alpha}_*^l$, $\widehat{\boldsymbol{\beta}}^k \xrightarrow{p} \boldsymbol{\beta}_*^k$, $\boldsymbol{\alpha}_*^1 = \boldsymbol{\alpha}_0^1$, and $\widehat{q}_{p,1}^k \xrightarrow{p} q_{1,p,*}^k$. And

$$
D(\boldsymbol{\tau}^\intercal) = \mathrm{E}\left\{ \frac{T}{\pi^1(\boldsymbol{\alpha}_0^1)} \frac{\mathbf{r}_i(\boldsymbol{\alpha}_*, \boldsymbol{\beta}_*, \mathbf{q}_{p,1,*})}{1 + \boldsymbol{\tau}^\intercal \mathbf{r}_i(\boldsymbol{\alpha}_*, \boldsymbol{\beta}_*, \mathbf{q}_{p,1,*}) / \pi_i^1(\boldsymbol{\alpha}_0^1)} \right\}
$$

is the limiting value of $D_n(\widehat{\boldsymbol{\tau}}^\intercal)$ with

$$
\mathbf{r}_i(\boldsymbol{\alpha}_*, \boldsymbol{\beta}_*, \mathbf{q}_{p,1,*}) = \begin{Bmatrix} \pi^1(\alpha_0^1, \mathbf{x}_i) - \theta_0^1 \\ \vdots \\ \pi^L(\alpha_*^L, \mathbf{x}_i) - \theta^L \\ G_t^1(q_{p,1,*}^1 | \mathbf{x}_i, \boldsymbol{\beta}_*^1) - \psi^1 \\ \vdots \\ G_t^K q_{p,1,*}^K | \mathbf{x}_i, \boldsymbol{\beta}_*^K) - \psi^K, \end{Bmatrix}.
$$

Here θ^l and ψ^k are limiting values of $\widehat{\theta}^l$ and $\widehat{\psi}^k$ and $\boldsymbol{\tau}_*^\intercal = \mathbf{0}$ is a solution to $D(\boldsymbol{\tau}^\intercal) = \mathbf{0}$. Based on empirical likelihood theory and suitable regularity conditions, it can be easy verified that $\widehat{\boldsymbol{\tau}}^\intercal \xrightarrow{p} \mathbf{0}$ by Qin and Lawless (1994) and Han and Wang (2013) and even $\widehat{\boldsymbol{\tau}}^\intercal = O_p(n^{-1/2})$ similar to Theorem 1 of Qin and Lawless (1994).

Here, we show how to simply the estimated weight w_i by showing that $\sum_{i\in S_t}\frac{1}{n_t}\frac{1}{1+\widehat{\boldsymbol{\lambda}}^\intercal\widehat{\mathbf{r}}_i(\widehat{\boldsymbol{\alpha}},\widehat{\boldsymbol{\beta}},\widehat{\mathbf{q}}_{p,t})}=1$:

$$\begin{aligned}\sum_{i\in S_t}\frac{1}{n_t}\frac{1}{1+\widehat{\boldsymbol{\lambda}}^\intercal\widehat{\mathbf{r}}_i(\widehat{\boldsymbol{\alpha}},\widehat{\boldsymbol{\beta}},\widehat{\mathbf{q}}_{p,t})}&=\sum_{i\in S_t}\frac{1}{n_t}\frac{1+\widehat{\boldsymbol{\lambda}}^\intercal\widehat{\mathbf{r}}_i(\widehat{\boldsymbol{\alpha}},\widehat{\boldsymbol{\beta}},\widehat{\mathbf{q}}_{p,t})-\widehat{\boldsymbol{\lambda}}^\intercal\widehat{\mathbf{r}}_i(\widehat{\boldsymbol{\alpha}},\widehat{\boldsymbol{\beta}},\widehat{\mathbf{q}}_{p,t})}{1+\widehat{\boldsymbol{\lambda}}^\intercal\widehat{\mathbf{r}}_i(\widehat{\boldsymbol{\alpha}},\widehat{\boldsymbol{\beta}},\widehat{\mathbf{q}}_{p,t})}\\&=\frac{1}{n_t}\sum_{i\in S_t}\left\{1-\frac{\widehat{\boldsymbol{\lambda}}^\intercal\widehat{\mathbf{r}}_i(\widehat{\boldsymbol{\alpha}},\widehat{\boldsymbol{\beta}},\widehat{\mathbf{q}}_{p,t})}{1+\widehat{\boldsymbol{\lambda}}^\intercal\widehat{\mathbf{r}}_i(\widehat{\boldsymbol{\alpha}},\widehat{\boldsymbol{\beta}},\widehat{\mathbf{q}}_{p,t})}\right\}\\&=1.\end{aligned}$$

Next we are going to prove the consistency of weighted empirical CDF: $\widehat{F}_{n,1}(q)\xrightarrow{p}F_1(q)$. The estimated weight can be expressed as

$$\widehat{w}_i=\frac{1}{n_1}\frac{\widehat{\theta}^1/\pi_i^1(\widehat{\boldsymbol{\alpha}}^1)}{1+\widehat{\boldsymbol{\tau}}^\intercal\widehat{\mathbf{r}}_i(\widehat{\boldsymbol{\alpha}},\widehat{\boldsymbol{\beta}},\widehat{\mathbf{q}}_{p,1})/\pi_i^1(\widehat{\boldsymbol{\alpha}}^1)}.$$

$$\begin{aligned}\widehat{F}_{n,1}(q)&=\sum_{i\in S_1}\widehat{w}_iI(Y_i\le q)\\&=\sum_{i\in S_1}\frac{1}{n_1}\frac{\widehat{\theta}^1/\pi_i^1(\widehat{\boldsymbol{\alpha}}^1)}{1+\widehat{\boldsymbol{\tau}}^\intercal\widehat{\mathbf{r}}_i(\widehat{\boldsymbol{\alpha}},\widehat{\boldsymbol{\beta}},\widehat{\mathbf{q}}_{p,1})/\pi_i^1(\widehat{\boldsymbol{\alpha}}^1)}I(Y_i\le q)\\&=\frac{\widehat{\theta}^1}{n_1/n}\frac{1}{n}\sum_{i=1}^{n}\frac{T_i}{\pi_i^1(\widehat{\boldsymbol{\alpha}}^1)}\frac{1}{1+\widehat{\boldsymbol{\tau}}^\intercal\widehat{\mathbf{r}}_i(\widehat{\boldsymbol{\alpha}},\widehat{\boldsymbol{\beta}},\widehat{\mathbf{q}}_{p,1})/\pi_i^1(\widehat{\boldsymbol{\alpha}}^1)}I(Y_i\le q).\end{aligned}$$

All parameters converge in probability to some limiting values. So we have $T_i/\pi_i^1(\widehat{\boldsymbol{\alpha}}^1)=T_i/\pi_i^1(\boldsymbol{\alpha}_0^1)+o_p(1)$, $1\Big/\left\{1+\widehat{\boldsymbol{\tau}}^\intercal\widehat{\mathbf{r}}_i(\widehat{\boldsymbol{\alpha}},\widehat{\boldsymbol{\beta}},\widehat{\mathbf{q}}_{p,1})/\pi_i^1(\widehat{\boldsymbol{\alpha}}^1)\right\}=1+o_p(1)$ and $\frac{\widehat{\theta}^1}{n_1/n}=1+o_p(1)$. So we can simplify $\widehat{F}_{n,1}(q)$:

$$\begin{aligned}\widehat{F}_{n,1}(q)&=\frac{1}{n}\sum_{i=1}^{n}\left\{\frac{T_i}{\pi_i^1(\boldsymbol{\alpha}_0^1)}+o_p(1)\right\}\{1+o_p(1)\}I(Y_i\le q)\\&=\frac{1}{n}\sum_{i=1}^{n}\frac{T_i}{\pi_i^1(\boldsymbol{\alpha}_0^1)}I(Y_i\le q)+o_p(1)\\&\text{Using } T_iI(Y_i\le q)=TI\{Y(1)\le q\},\\&\xrightarrow{p}\mathrm{E}\left[\frac{T}{\pi^1(\boldsymbol{\alpha}_0^1)}I\{Y(1)\le q\}\right]\\&=P\{Y(1)\le q\}=F_1(q).\end{aligned}$$

For the last step, we are going to prove $\widehat{\xi}_{p,1} \xrightarrow{p} \xi_{p,1}$ as $n \to \infty$ through the definition. First we have $\widehat{F}_{n,1}(q) - F_1(q) = o_p(1)$ for any q, which means $\forall\, \epsilon, \delta > 0$, $\exists\, n$, when $n \geq N$, we have $P\left\{|\widehat{F}_{n,1}(q) - F_1(q)| > \epsilon\right\} < \delta$ for any q.

So $\forall \epsilon^*$, $\delta > 0$, we want to prove $P(|\widehat{\xi}_{p,1} - \xi_{p,1}| > \epsilon^*) < \delta$, we prove in one direction first.

$$
\begin{aligned}
P(\widehat{\xi}_{p,1} - \xi_{p,1} > \epsilon^*) &= P(\widehat{\xi}_{p,1} > \xi_{p,1} + \epsilon^*) \\
\widehat{\xi}_{p,1} > \xi_{p,1} + \epsilon^* \text{ iff } &\widehat{F}_{n,1}(\epsilon^* + \xi_{p,1}) < p \\
&= P\left\{\widehat{F}_{n,1}(\epsilon^* + \xi_{p,1}) < p\right\} \\
&= P\left\{\widehat{F}_{n,1}(\epsilon^* + \xi_{p,1}) - F_1(\epsilon^* + \xi_{p,1}) < p - F_1(\epsilon^* + \xi_{p,1})\right\} \\
&\text{Notice that } p - F_1(\epsilon^* + \xi_{p,1}) < 0 \\
&\leq P\left\{\left|\widehat{F}_{n,1}(\epsilon^* + \xi_{p,1}) - F_1(\epsilon^* + \xi_{p,1})\right| > |p - F_1(\epsilon^* + \xi_{p,1})|\right\} \\
&< \delta \quad \text{If } \epsilon \leq |p - F_1(\epsilon^* + \xi_{p,1})|, \\
&\text{which means } \forall\, \epsilon^* \text{ we just need to select } \epsilon \text{ such that } \epsilon \leq F_1(\epsilon^* + \xi_{p,1}) - p.
\end{aligned}
$$

Similarly we can get if $\epsilon \leq p - F_1(\xi_{p,1} - \epsilon^*)$ then $P(\widehat{\xi}_{p,1} - \xi_{p,1} < -\epsilon^*) < \delta$. If we select $\epsilon = min\left\{F_1(\epsilon^* + \xi_{p,1}) - p, p - F_1(\xi_{p,1} - \epsilon^*)\right\}$, then we finish the proof for the consistency of QTE estimator.

□

5.7.2 Proof of Theorem 5.3.2

Theorem. *If one of the conditional CDFs, $G_t^k(q|\mathbf{x})$ for $k = 1, \ldots, K$, is correctly specified, $\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}$, and $\widehat{\mathbf{q}}_{p,t}$ are estimated from their corresponding specified models, then $\widehat{\xi}_{p,t} \xrightarrow{p} \xi_{p,t}$ as $n \to \infty$ for $t = 0, 1$.*

Proof. Let's assume $G_1^1(q|\mathbf{x}, \boldsymbol{\beta}_0^1)$ is the true specified conditional CDF for treatment group with true parameter $\boldsymbol{\beta}_0^1$. So $\widehat{\boldsymbol{\beta}}^1 \xrightarrow{p} \boldsymbol{\beta}_0^1$ as $n \to \infty$.

$$\begin{aligned}\widehat{F}_{n,1}(q) &= \sum_{i\in S_1} \widehat{w}_i I(Y_i \le q) \\ &= \sum_{i\in S_1} \widehat{w}_i \left\{ I(Y_i \le q) - G_1^1(q|\mathbf{x}_i, \widehat{\boldsymbol{\beta}}^1) \right\} + \sum_{i\in S_1} \widehat{w}_i \widehat{G}_1^1(q|\mathbf{x}, \widehat{\boldsymbol{\beta}}^1)\end{aligned}$$

We replace the second term according to constraint (5.7),

$$= \frac{n}{n_1}\frac{1}{n}\sum_{i=1}^{n} \frac{T_i \left\{ I(Y_i \le q) - G_1^1(q|\mathbf{x}_i, \widehat{\boldsymbol{\beta}}^1) \right\}}{1 + \widehat{\boldsymbol{\lambda}}^\intercal \widehat{\mathbf{r}}_i(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}, \widehat{\mathbf{q}}_{p,1})} + \frac{1}{n}\sum_{i=1}^{n} G_1^1(q|\mathbf{x}_i, \widehat{\boldsymbol{\beta}}^1).$$

Replace $TI(Y \le q)$ by $TI(Y(1) \le q)$, the above equation becomes

$$\begin{aligned}\widehat{F}_{n,1}(q) &\xrightarrow{p} P(T=1)\mathrm{E}\left[\frac{T\left\{ I(Y(1) \le q) - G_1^1(q|\mathbf{x}, \boldsymbol{\beta}_0^1) \right\}}{1 + \boldsymbol{\lambda}_*^\intercal \mathbf{r}(\boldsymbol{\alpha}_*, \boldsymbol{\beta}_*, \mathbf{q}_{p,1,*})} \right] + \mathrm{E}\left\{ G_1^1(q|\mathbf{x}, \boldsymbol{\beta}_0^1) \right\} \\ &= P(T=1)\mathrm{E}\left[\mathrm{E}\left[\frac{T\left\{ I(Y(1) \le q) - G_1^1(q|\mathbf{x}, \boldsymbol{\beta}_0^1) \right\}}{1 + \boldsymbol{\lambda}_*^\intercal \mathbf{r}(\boldsymbol{\alpha}_*, \boldsymbol{\beta}_*, \mathbf{q}_{p,1,*})} \,\middle|\, \mathbf{x} \right] \right] + F_1(q) \\ &= P(T=1)\mathrm{E}\left[\frac{\mathrm{E}\left(T|\mathbf{x}\right)}{1 + \boldsymbol{\lambda}_*^\intercal \mathbf{r}(\boldsymbol{\alpha}_*, \boldsymbol{\beta}_*, \mathbf{q}_{p,1,*})} \mathrm{E}\left[I\{Y(1) \le q\} - G_1^1(q|\mathbf{x}) \,\middle|\, \mathbf{x} \right] \right] + F_1(q) \\ &= P(T=1)\mathrm{E}\left[\frac{\mathrm{E}\left(T|\mathbf{x}\right)}{1 + \boldsymbol{\lambda}_*^\intercal \mathbf{r}(\boldsymbol{\alpha}_*, \boldsymbol{\beta}_*, \mathbf{q}_{p,1,*})} \left\{ G_1^1(q|\mathbf{x}) - G_1^1(q|\mathbf{x}) \right\} \right] + F_1(q) \\ &= F_1(q)\end{aligned}$$

After we derive the consistency of $\widehat{F}_{n,1}(q)$ to $F_1(q)$, we can similarly construct that $\widehat{\xi}_{p,1} \xrightarrow{p} \xi_{p,1}$ as $n \to \infty$ when a conditional CDF is correctly specified. □

5.7.3 Proof of Theorem 5.3.3

Theorem. *When one of the propensity score model is correctly specified, then* $\widehat{\xi}_{p,t} \xrightarrow{d} N\left(0, \frac{\sigma_F^2}{f_t^2(\xi_{p,t})}\right)$ *as* $n \to \infty$ *for* $t = 0, 1$.

Proof. We derive the asymptotic normality for $t = 1$. The proofs are not different for $t = 0$. Let's assume $f_1(y)$ is the probability density function for $F_1(y)$. For any c,

$$
\begin{aligned}
&P\left\{\sqrt{n}(\widehat{\xi}_{p,1} - \xi_{p,1}) \le c\right\} \\
&= P\left\{\widehat{\xi}_{p,1} \le \xi_{p,1} + c/\sqrt{n}\right\} \\
&\widehat{\xi}_{p,1} \le \xi_{p,1} + c/\sqrt{n} \quad \text{iff} \quad p \le \widehat{F}_{n,1}(\xi_{p,1} + c/\sqrt{n}) \\
&= P\left\{\widehat{F}_{n,1}(\xi_{p,1} + c/\sqrt{n}) \ge p\right\} \\
&= P\left\{F_1(\xi_{p,1} + c/\sqrt{n}) - \widehat{F}_{n,1}(\xi_{p,1} + c/\sqrt{n}) \le F_1(\xi_{p,1} + c/\sqrt{n}) - p\right\} \\
&= P\left[\sqrt{n}\left\{F_1(\xi_{p,1} + c/\sqrt{n}) - \widehat{F}_{n,1}(\xi_{p,1} + c/\sqrt{n})\right\} \le \sqrt{n}\left\{F_1(\xi_{p,1} + c/\sqrt{n}) - p\right\}\right]
\end{aligned}
$$

We notice that

$$
\sqrt{n}\left\{F_1(\xi_{p,1} + c/\sqrt{n}) - p\right\} = \sqrt{n}\left\{F_1(\xi_{p,1} + c/\sqrt{n}) - F_1(\xi_{p,1})\right\} = \sqrt{n} f_1(\eta_{p,n}) * \frac{c}{\sqrt{n}},
$$

where $\eta_{p,n}$ is a value between $\xi_{p,1} + c/\sqrt{n}$ and $\xi_{p,1}$. By mean value theorem, we have $\eta_{p,n} \to \xi_{p,1}$ as $n \to \infty$. So we can write $\sqrt{n}\left\{F_1(\xi_{p,1} + c/\sqrt{n}) - p\right\} = c f_1(\xi_{p,1}) + o_p(1)$. We can further derive that

$$
\begin{aligned}
&P\left\{\sqrt{n}(\widehat{\xi}_{p,1} - \xi_{p,1}) \le c\right\} \\
&= P\left[\sqrt{n}\left\{F_1(\xi_{p,1} + c/\sqrt{n}) - \widehat{F}_{n,1}(\xi_{p,1} + c/\sqrt{n})\right\} + o_p(1) \le c f_1(\xi_{p,1})\right]
\end{aligned}
$$

Hence we only need to prove that $\sqrt{n}\left\{F_1(\xi_{p,1} + c/\sqrt{n}) - \widehat{F}_{n,1}(\xi_{p,1} + c/\sqrt{n})\right\} + o_p(1) \xrightarrow{d} N(0, \sigma_F^2)$. σ_F^2 is the asymptotic variance and will be defined in the following derivation.

Let $\widehat{\theta}_0^1 = \sum_{i=1}^n \pi_i^1(\boldsymbol{\alpha}_0^1)$. We have the following Taylor expansion of $\widehat{F}_{n,1}(\xi_{p,1} + c/\sqrt{n})$ at

$(0, \boldsymbol{\alpha}_*, \boldsymbol{\beta}_*, \mathbf{q}_{p,1,*})$ where $\boldsymbol{\alpha}_* = (\boldsymbol{\alpha}_0^1, \boldsymbol{\alpha}_*^2, \ldots, \boldsymbol{\alpha}_*^L)^\intercal$ and $\boldsymbol{\tau} = \mathbf{0}$:

$$
\begin{aligned}
&\widehat{F}_{n,1}(\xi_{p,1} + c/\sqrt{n}) \\
&= \frac{\widehat{\theta^1}}{n_1} \sum_{i=1}^{n} \left\{ \frac{T_i}{\pi_i^1(\widehat{\boldsymbol{\alpha}}^1)} \frac{I(Y_i \le \xi_{p,1} + c/\sqrt{n})}{1 + \widehat{\boldsymbol{\tau}}^\intercal \widehat{\mathbf{r}}_i(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}, \widehat{\mathbf{q}}_{p,1}) / \pi_i^1(\widehat{\boldsymbol{\alpha}}^1)} \right\} \\
&= \frac{n\widehat{\theta}_0^1}{n_1} \frac{1}{n} \sum_{i=1}^{n} \frac{T_i I(Y_i \le \xi_{p,1} + c/\sqrt{n})}{\pi_i^1(\boldsymbol{\alpha}_0^1)} \\
&\quad - \frac{n\widehat{\theta}_0^1}{n_1} \frac{1}{n} \sum_{i=1}^{n} \frac{T_i I(Y_i \le \xi_{p,1} + c/\sqrt{n})}{\pi_i^1(\boldsymbol{\alpha}_0^1)} \frac{\widehat{\mathbf{r}}_i^\intercal(\boldsymbol{\alpha}_*, \boldsymbol{\beta}_*, \mathbf{q}_{p,1,*})}{\pi_i^1(\boldsymbol{\alpha}_0^1)} (\widehat{\boldsymbol{\tau}} - \mathbf{0}) \\
&\quad + \frac{\frac{1}{n}\sum_{h=1}^{n} \frac{\partial \pi_h^1(\boldsymbol{\alpha}_0^1)}{\partial \boldsymbol{\alpha}^{1,\intercal}}}{n1/n} \frac{1}{n} \sum_{i=1}^{n} \frac{T_i I(Y_i \le \xi_{p,1} + c/\sqrt{n})}{\pi_i^1(\boldsymbol{\alpha}_0^1)} (\widehat{\boldsymbol{\alpha}}^1 - \boldsymbol{\alpha}_0^1) \\
&\quad - \frac{\frac{1}{n}\sum_{h=1}^{n} \pi_h^1(\boldsymbol{\alpha}_0^1)}{n1/n} \frac{1}{n} \sum_{i=1}^{n} \frac{T_i I(Y_i \le \xi_{p,1} + c/\sqrt{n})}{\{\pi_i^1(\boldsymbol{\alpha}_0^1)\}^2} \frac{\partial \pi_i^1(\boldsymbol{\alpha}_0^1)}{\partial \boldsymbol{\alpha}^{1,\intercal}} (\widehat{\boldsymbol{\alpha}}^1 - \boldsymbol{\alpha}_0^1) \\
&\quad + O_p(n^{-1})
\end{aligned}
$$

By using the fact that $\frac{\widehat{\theta}_0^1}{n_1/n} = 1 + o_p(1)$, $\frac{1}{n}\sum_{h=1}^{n} \frac{\partial \pi_h^1(\boldsymbol{\alpha}_0^1)}{\partial \boldsymbol{\alpha}^{1,\intercal}} = \mathrm{E}\left\{\partial \pi^1(\boldsymbol{\alpha}_0^1)/\partial \boldsymbol{\alpha}^{1,\intercal}\right\} + o_p(1)$, $\frac{1}{n}\sum_{h=1}^{n} \pi_h^1(\boldsymbol{\alpha}_0^1) = \theta_0 + o_p(1)$, and $\widehat{\mathbf{r}}_i^\intercal(\boldsymbol{\alpha}_*, \boldsymbol{\beta}_*, \mathbf{q}_{p,1,*}) = \mathbf{r}_i^\intercal(\boldsymbol{\alpha}_*, \boldsymbol{\beta}_*, \mathbf{q}_{p,1,*}) + o_p(1)$, we can simplify $\widehat{F}_{n,1}(\xi_{p,1} + c/\sqrt{n})$:

$$
\begin{aligned}
&\widehat{F}_{n,1}(\xi_{p,1}+c/\sqrt{n}) \\
&= \frac{1}{n}\sum_{i=1}^{n}\frac{T_i I(Y_i \le \xi_{p,1}+c/\sqrt{n})}{\pi_i^1(\boldsymbol{\alpha}_0^1)} - \frac{1}{n}\sum_{i=1}^{n}\frac{T_i I(Y_i \le \xi_{p,1}+c/\sqrt{n})\mathbf{r}_i^{\intercal}(\boldsymbol{\alpha}_*,\boldsymbol{\beta}_*,\mathbf{q}_{p,1,*})}{\left\{\pi_i^1(\boldsymbol{\alpha}_0^1)\right\}^2}(\widehat{\boldsymbol{\tau}}-\mathbf{0}) \\
&\quad + \frac{\mathrm{E}\left\{\partial\pi^1(\boldsymbol{\alpha}_0^1)/\partial\boldsymbol{\alpha}^{1,\intercal}\right\}}{\theta_0}\frac{1}{n}\sum_{i=1}^{n}\frac{T_i I(Y_i \le \xi_{p,1}+c/\sqrt{n})}{\pi_i^1(\boldsymbol{\alpha}_0^1)}(\widehat{\boldsymbol{\alpha}}^1-\boldsymbol{\alpha}_0^1) \\
&\quad - \frac{1}{n}\sum_{i=1}^{n}\frac{T_i I(Y_i \le \xi_{p,1}+c/\sqrt{n})}{\left\{\pi_i^1(\boldsymbol{\alpha}_0^1)\right\}^2}\frac{\partial\pi_i^1(\boldsymbol{\alpha}_0^1)}{\partial\boldsymbol{\alpha}^{1,\intercal}}(\widehat{\boldsymbol{\alpha}}^1-\boldsymbol{\alpha}_0^1)+o_p(n^{-1/2}) \\
&= \frac{1}{n}\sum_{i=1}^{n}\frac{T_i I(Y_i \le \xi_{p,1}+c/\sqrt{n})}{\pi_i^1(\boldsymbol{\alpha}_0^1)} - \mathrm{E}\left[\frac{TI(Y \le \xi_{p,1}+c/\sqrt{n})\mathbf{r}^{\intercal}(\boldsymbol{\alpha}_*,\boldsymbol{\beta}_*,\mathbf{q}_{p,1,*})}{\left\{\pi^1(\boldsymbol{\alpha}_0^1)\right\}^2}\right](\widehat{\boldsymbol{\tau}}-\mathbf{0}) \\
&\quad + \frac{\mathrm{E}\left\{\partial\pi^1(\boldsymbol{\alpha}_0^1)/\partial\boldsymbol{\alpha}^{1,\intercal}\right\}}{\theta_0}\mathrm{E}\left\{\frac{TI(Y \le \xi_{p,1}+c/\sqrt{n})}{\pi^1(\boldsymbol{\alpha}_0^1)}\right\}(\widehat{\boldsymbol{\alpha}}^1-\boldsymbol{\alpha}_0^1) \\
&\quad - \mathrm{E}\left[\frac{TI(Y \le \xi_{p,1}+c/\sqrt{n})}{\left\{\pi^1(\boldsymbol{\alpha}_0^1)\right\}^2}\frac{\partial\pi^1(\boldsymbol{\alpha}_0^1)}{\partial\boldsymbol{\alpha}^{1,\intercal}}\right](\widehat{\boldsymbol{\alpha}}^1-\boldsymbol{\alpha}_0^1)+o_p(n^{-1/2})
\end{aligned}
$$

To simply the above equation, let $\mathbf{A} = \mathrm{E}\left[\frac{TI(Y \le \xi_{p,1}+c/\sqrt{n})\mathbf{r}^{\intercal}(\boldsymbol{\alpha}_*,\boldsymbol{\beta}_*,\mathbf{q}_{p,1,*})}{\left\{\pi^1(\boldsymbol{\alpha}_0^1)\right\}^2}\right]$,
$\mathbf{B} = \frac{\mathrm{E}\left\{\partial\pi^1(\boldsymbol{\alpha}_0^1)/\partial\boldsymbol{\alpha}^{1,\intercal}\right\}}{\theta_0}\mathrm{E}\left\{\frac{TI(Y \le \xi_{p,1}+c/\sqrt{n})}{\pi^1(\boldsymbol{\alpha}_0^1)}\right\} - \mathrm{E}\left[\frac{TI(Y \le \xi_{p,1}+c/\sqrt{n})}{\left\{\pi^1(\boldsymbol{\alpha}_0^1)\right\}^2}\frac{\partial\pi^1(\boldsymbol{\alpha}_0^1)}{\partial\boldsymbol{\alpha}^{1,\intercal}}\right]$.

Then we get:

$$
\begin{aligned}
\sqrt{n}\widehat{F}_{n,1}(\xi_{p,1}+c/\sqrt{n}) = \sqrt{n}\frac{1}{n}\sum_{i=1}^{n}&\frac{T_i I(Y_i \le \xi_{p,1}+c/\sqrt{n})}{\pi_i^1(\boldsymbol{\alpha}_0^1)} \\
&- \mathbf{A}\sqrt{n}(\widehat{\boldsymbol{\tau}}^{\intercal}-\mathbf{0}) + \mathbf{B}\sqrt{n}(\widehat{\boldsymbol{\alpha}}^1-\boldsymbol{\alpha}_0^1)+o_p(1).
\end{aligned}
$$

In the following steps, we are able to construct the asymptotic normality of

$\sqrt{n}\left\{F_1(\xi_{p,1}+c/\sqrt{n}) - \widehat{F}_{n,1}(\xi_{p,1}+c/\sqrt{n})\right\}$ with the asymptotic normality of $\sqrt{n}\widehat{F}_{n,1}(\xi_{p,1}+c/\sqrt{n})$.

$$
\begin{aligned}
&\sqrt{n}\left\{F_1(\xi_{p,1}+c/\sqrt{n})-\widehat{F}_{n,1}(\xi_{p,1}+c/\sqrt{n})\right\}\\
&=\sqrt{n}\frac{1}{n}\sum_{i=1}^{n}\left\{F_1(\xi_{p,1}+c/\sqrt{n})-\frac{T_i}{\pi_i^1(\boldsymbol{\alpha}_0^1)}I(Y_i\le\xi_{p,1}+c/\sqrt{n})\right\}\\
&\quad+\mathbf{A}\sqrt{n}(\widehat{\boldsymbol{\tau}}^{\intercal}-\mathbf{0})-\mathbf{B}\sqrt{n}(\widehat{\boldsymbol{\alpha}}^1-\boldsymbol{\alpha}_0^1)+o_p(1)\\
&=\sqrt{n}\frac{1}{n}\sum_{i=1}^{n}Z_{n,i}+\mathbf{A}\sqrt{n}(\widehat{\boldsymbol{\tau}}^{\intercal}-\mathbf{0})-\mathbf{B}\sqrt{n}(\widehat{\boldsymbol{\alpha}}^1-\boldsymbol{\alpha}_0^1)+o_p(1)
\end{aligned}
$$

where $Z_{n,i}=F_1(\xi_{p,1}+c/\sqrt{n})-\frac{T_i}{\pi_i^1(\boldsymbol{\alpha}_0^1)}I(Y_i\le\xi_{p,1}+c/\sqrt{n})$.

So $\{Z_{n,i}, i=1,\ldots,n; n=1,\ldots,\infty\}$ is triangular array. The Lyapunov central limit theorem can be employed here.

We will prove that $\sum_{i=1}^{n}\{Z_{n,i}-\mathrm{E}(Z_{n,i})\}\Big/\{\sum_{i=1}^{n}\mathrm{Var}(Z_{n,i})\}^{1/2}\xrightarrow{p}N(0,1)$ by checking the Lyapunov condition. First we can derive that $\mathrm{E}(Z_{n,i})=0$. It can be verified that $T_iI(Y_i\le\xi_{p,1}+c/\sqrt{n})=T_iI\{Y_i(1)\le\xi_{p,1}+c/\sqrt{n}\}$ under strongly ignorable treatment assignment given in Chapter 1. Also we can derive the variance,

$$
\begin{aligned}
&\mathrm{Var}(Z_{n,i}) \\
&= \mathrm{E}\left[\frac{T_i^2}{\{\pi_i^1(\boldsymbol{\alpha}_0^1)\}^2} I\{Y_i(1) \le \xi_{p,1} + c/\sqrt{n}\}\right] - \mathrm{E}^2\left[\frac{T_i}{\pi_i^1(\boldsymbol{\alpha}_0^1)} I\left\{Y_i(1) \le \xi_{p,1} + c/\sqrt{n}\right\}\right] \\
&= \mathrm{E}\left[\mathrm{E}\left[\frac{T_i}{\{\pi_i^1(\boldsymbol{\alpha}_0^1)\}^2} I\left\{Y_i(1) \le \xi_{p,1} + c/\sqrt{n}\right\} \,\middle|\, \boldsymbol{x}_i\right]\right] \\
&\quad - \left[\mathrm{E}\left[\mathrm{E}\left[\frac{T_i}{\pi_i^1(\boldsymbol{\alpha}_0^1)} I\left\{Y_i(1) \le \xi_{p,1} + c/\sqrt{n}\right\} \,\middle|\, \boldsymbol{x}_i\right]\right]\right]^2 \\
&= \mathrm{E}\left[\frac{\mathrm{E}\left(T_i|\boldsymbol{x}_i\right)}{\{\pi_i^1(\boldsymbol{\alpha}_0^1)\}^2} \mathrm{E}\left[I\left\{Y_i(1) \le \xi_{p,1} + c/\sqrt{n}\right\} \,\middle|\, \boldsymbol{x}_i\right]\right] \\
&\quad - \left[\mathrm{E}\left[\frac{\mathrm{E}\left(T_i|\boldsymbol{x}_i\right)}{\pi_i^1(\boldsymbol{\alpha}_0^1)} \mathrm{E}\left[I\left\{Y_i(1) \le \xi_{p,1} + c/\sqrt{n}\right\} \,\middle|\, \boldsymbol{x}_i\right]\right]\right]^2 \\
&= \mathrm{E}\left\{\frac{1}{\pi_i^1(\boldsymbol{\alpha}_0^1)} G_1(\xi_{p,1} + c/\sqrt{n}|\mathbf{x}_i)\right\} - \left[\mathrm{E}\left[\mathrm{E}\left[I\left\{Y_i(1) \le \xi_{p,1} + c/\sqrt{n}\right\} \,\middle|\, \boldsymbol{x}_i\right]\right]\right]^2 \\
&= \mathrm{E}\left\{\frac{1}{\pi^1(\boldsymbol{\alpha}_0^1)} G_1(\xi_{p,1} + c/\sqrt{n}|\mathbf{x})\right\} - \left\{F_1(\xi_{p,1} + c/\sqrt{n})\right\}^2,
\end{aligned}
$$

which is a finite constant and denoted by σ_z^2. Next we are going to check

$$
\lim_{n\to\infty} \sum_{i=1}^n \mathrm{E}\left\{|Z_{n,i} - \mathrm{E}(Z_{n,i})|^{2+\epsilon}\right\} \Big/ \left\{\sum_{i=1}^n \mathrm{Var}(Z_{n,i})\right\}^{1+\epsilon/2} = 0 \text{ for some } \epsilon > 0.
$$

Based on the weak common support assumption: $a < \pi(x) < 1 - a$ for some $a > 0$, we know

$$
\begin{aligned}
Z_{n,i} &= F_1(\xi_{p,1} + c/\sqrt{n}) - \frac{T_i}{\pi_i^1(\boldsymbol{\alpha}_0^1)} I(Y_i \le \xi_{p,1} + c/\sqrt{n}) \\
&\le F_1(\xi_{p,1} + c/\sqrt{n}) + \frac{T_i}{\pi_i^1(\boldsymbol{\alpha}_0^1)} I(Y_i \le \xi_{p,1} + c/\sqrt{n}) \\
&\le 1 + \frac{1}{a}
\end{aligned}
$$

Similarly, we get $|Z_{n,i}| \leq 1 + 1/a$. So $|Z_{n,i} - 0|^{2+\epsilon} \leq |1 + 1/a|^{2+\epsilon}$, so we can get

$$\sum_{i=1}^{n} \mathrm{E}\left\{|Z_{n,i} - \mathrm{E}(Z_{n,i})|^{2+\epsilon}\right\} \Big/ \left\{\sum_{i=1}^{n} \mathrm{Var}(Z_{n,i})\right\}^{1+\epsilon/2}$$
$$\leq n|1 + \frac{1}{a}|^{2+\epsilon} \Big/ \left(n\sigma_z^2\right)^{1+\epsilon/2} \to 0$$

as $n \to \infty$. So we can conclude that $\frac{\sqrt{n}\frac{1}{n}\sum_{i=1}^{n} Z_{n,i}}{\sqrt{\sigma_z^2}} \xrightarrow{d} N(0,1)$. Since $\mathbf{A}\sqrt{n}(\widehat{\boldsymbol{\tau}}^\intercal - \mathbf{0})$ and $\mathbf{B}\sqrt{n}(\widehat{\boldsymbol{\alpha}}^1 - \boldsymbol{\alpha}_0^1)$ all approximate normal distribution, we can conclude that $\sqrt{n}\left\{F_1(\xi_{p,1} + c/\sqrt{n}) - \widehat{F}_{n,1}(\xi_{p,1} + c/\sqrt{n})\right\} \xrightarrow{d} N(0, \sigma_F^2)$, where σ_F^2 is the asymptotic variance which will include both variances and covariances among $\sqrt{n}\frac{1}{n}\sum_{i=1}^{n} Z_{n,i}$, $\sqrt{n}(\widehat{\boldsymbol{\tau}}^\intercal - \mathbf{0})$, and $\sqrt{n}(\widehat{\boldsymbol{\alpha}}^1 - \boldsymbol{\alpha}_0^1)$ .Finally by Slutsky's theorem

$$\begin{aligned}
&P\left\{\sqrt{n}(\widehat{\xi}_{p,1} - \xi_{p,1}) \leq c\right\} \\
&= P\left[\sqrt{n}\left\{F(\xi_{p,1} + c/\sqrt{n}) - \widehat{F}_{n,1}(\xi_{p,1} + c/\sqrt{n})\right\} + o_p(1) \leq c f_1(\xi_{p,1})\right] \\
&\to P\left\{N(0, \sigma_F^2/f_1^2(\xi_{p,1})) \leq c\right\}
\end{aligned}$$

$$\sqrt{n}(\widehat{\xi}_{p,1} - \xi_{p,1}) \xrightarrow{d} N\left(0, \frac{\sigma_F^2}{f_1^2(\xi_{p,1})}\right).$$

□

5.7.4 Proof of Theorem 5.3.4

Theorem. *When one of the conditional CDFs is correctly specified, then* $\widehat{\xi}_{p,t} \xrightarrow{d} N\left(0 \frac{\sigma_F^2}{f_t^2(\xi_{p,t})}\right)$ *as* $n \to \infty$ *for* $t = 0, 1$.

Proof. For $t = 1$, assume $G_1^1(q|\mathbf{x})$ is the correctly specified conditional CDF and $\boldsymbol{\beta}_0^1$ is the true parameter vector, we reparameterize the Lagrange multiplier $\widehat{\boldsymbol{\lambda}}^\intercal$ by $\widehat{\boldsymbol{\lambda}}^\intercal =$

$(\frac{\widehat{\tau}_1+1}{\widehat{\psi}^1}, \frac{\widehat{\tau}_2}{\widehat{\psi}^1}, \ldots, \frac{\widehat{\tau}_{K+L}}{\widehat{\psi}^1}) = \frac{1}{\widehat{\psi}^1}[(1,0,\ldots,0) + \widehat{\boldsymbol{\tau}}^\intercal]$ and rearrange:

$$\widehat{\mathbf{r}}_i(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}, \widehat{\mathbf{q}}_{p,t}) = \begin{Bmatrix} G_1^1(\widehat{q}_{p,1}^1 | \mathbf{x}_i, \widehat{\boldsymbol{\beta}}^1) - \widehat{\psi}^1 \\ \vdots \\ G_1^K(\widehat{q}_{p,1}^K | \mathbf{x}_i, \widehat{\boldsymbol{\beta}}^K) - \widehat{\psi}^K \\ \pi^1(\widehat{\boldsymbol{\alpha}}^1, \mathbf{x}_i) - \widehat{\theta}^1 \\ \vdots \\ \pi^L(\widehat{\boldsymbol{\alpha}}^L, \mathbf{x}_i) - \widehat{\theta}^L \end{Bmatrix}.$$

Let $G_{1,i}^1(\widehat{\boldsymbol{\beta}}^1) = G_1^1(\widehat{q}_{p,1}^1 | \mathbf{x}_i, \widehat{\boldsymbol{\beta}}^1)$. Similar to the proof in Theorem 5.3.3, we can derive the asymptotic normality of
$\sqrt{n}\left\{F_1(\xi_{p,1} + c/\sqrt{n}) - \widehat{F}_{n,1}(\xi_{p,1} + c/\sqrt{n})\right\} + o_p(1)$ by Taylor expansion of

$$\begin{aligned} \widehat{F}_{n,1}(\xi_{p,1} + c/\sqrt{n}) &= \frac{1}{n_1} \sum_{i=1}^{n} \frac{T_i \left\{I(Y_i \le \xi_{p,1} + c/\sqrt{n})\right\}}{1 + \widehat{\boldsymbol{\lambda}}^\intercal \widehat{\mathbf{r}}_i(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}, \widehat{\mathbf{q}}_{p,1})} \\ &= \frac{\widehat{\psi}^1}{n_1} \sum_{i=1}^{n} \left\{ \frac{T_i}{G_{1,i}^1(\widehat{\boldsymbol{\beta}}^1)} \frac{I(Y_i \le \xi_{p,1} + c/\sqrt{n})}{1 + \widehat{\boldsymbol{\tau}}^\intercal \widehat{\mathbf{r}}_i(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}, \widehat{\mathbf{q}}_{p,1}) / G_{1,i}^1(\widehat{\boldsymbol{\beta}}^1)} \right\} \end{aligned}$$

at $(0, \boldsymbol{\alpha}_*, \boldsymbol{\beta}_*, \mathbf{q}_{p,1,*})$ where $\boldsymbol{\beta}_* = (\boldsymbol{\beta}_0^1, \boldsymbol{\beta}_*^2, \ldots, \boldsymbol{\beta}_*^K)^\intercal$. Hence, we can also derive that

$$\widehat{\xi}_{p,1} \xrightarrow{d} N\left(0 \frac{\sigma_F^2}{f_1^2(\xi_{p,1})}\right) \quad \text{as} \quad n \to \infty.$$

σ_F^2 is the asymptotic variance of $\sqrt{n}\left\{F_1(\xi_{p,1} + c/\sqrt{n}) - \widehat{F}_{n,1}(\xi_{p,1} + c/\sqrt{n})\right\}$.

□

Chapter 6

Discussion and Future work

6.1 Discussion

In this thesis, we proposed causal inference approaches to estimate average causal effects or quantile treatment effects specifically focusing on covariate balancing problems. First, we combined a parametric and a nonparametric model to estimate the propensity score. Second, we proposed to optimize kernel distance between the covariate distributions of treatment and control groups. Finally, we studied the quantile treatment effect estimator with covariate balancing constraints under the empirical likelihood or entropy measure framework.

The model averaging approach combines logistic regression and random forest models to estimate the propensity score with a mixing parameter λ. The optimal λ is selected using a grid search such that a certain covariate balance measure is optimized such as average value of absolute standardized mean difference (ASMD) and mean Kolmogorov-Smirnov (KS) test statistic. With the varying degree of nonlinearity of treatment models, model averaging methods, especially the model averaging method with the objective to minimize mean KS statistic, have consistently better performance than conventional propensity score approaches, especially when there is model misspecification in the logistic regression model. Combining two or more different propensity score models has the advantage to reduce bias, standard error, and imbalance in covariates without doing balance checking. Second,

combining several propensity score models can provide more protection against model misspecification.

In addition to commonly used balance measures like ASMD and KS statistic, the kernel distance is known to measure the discrepancy between two distributions. The kernel distance is zero as long as two distributions are equivalent. Previous simulation studies have shown that it has the strongest correlation with absolute bias in estimating the causal effect compared to other balance measures. The kernel distance propensity score approach used the kernel distance as a measure of covariate balance and achieved the balance in the covariate distributions between treated and control groups. Optimal parameters can be derived to optimize the modified kernel distance under the generalized method of moments or empirical likelihood framework. The Kernel function can be expressed as an inner product of two infinite-dimensional basis functions, so minimizing kernel distance is not just balancing finite moment of covariates. It can be treated as balancing all infinite dimensions. The covariate balancing propensity score (CBPS) is proposed to achieve the balance only in first or second moment conditions but not the overall distributions. The simulation results show that when the imbalance between treated and control groups are considerable, our kernel approaches are more likely to reduce the bias and variance.

In the third project, we studied the quantile treatment effect (QTE) estimation with the same idea of covariate balancing. QTE estimation is another hot topic in statistical practice but not many methods have been developed to balance the quantiles of conditional cumulative distribution functions of outcomes in observational studies. We propose a quantile treatment effect estimator which achieves balance in the covariates using two sets of constraints on: (1) the propensity score, and (2) the conditional distributions of potential outcomes. The inverse probability weighting (IPW) approach is sensitive to model misspecification. Unlike IPW approaches, we can either optimize the empirical likelihood or entropy measure with the balancing constraints to obtain the optimal weights which will reduce the effect of misspecification. Based on the idea of multiple robustness, there are several nice features from our proposed methods. Our proposed method shows more protection against model misspecification compared to double robust estimator. Our proposed estimator also have faster convergence rate compared to double robust estimator when estimating low or high percentiles. These quantities are of more interest than median

or average causal effect in many real applications.

Throughout this thesis, we explored the covariate balancing problems in causal inference. However, our proposed methods in this thesis are constructed within the potential outcome framework with certain conditions. The conditions should always be checked carefully. The assumptions about the absence of unmeasured confounding and others should be carefully qualified to emphasize the assumptions. In any application to real world data set, discussion about the possibility of unobserved confounding and other assumptions is essential and arguments should be provided about why it can safely be ignored.

6.2 Future work

There are some interesting extensions and questions left unsolved from this thesis that we can employ in the exploration of future work. The order of these suggestions is listed in accordance with the order of the projects:

In the simulation studies of model averaging approach, the standard error from the sandwich formula is too conservative and leads to higher than the nominal coverage rates of true treatment effect. Therefore, in the data analysis of Chapter 3, we focus on the bootstrapping approach to obtain standard errors. Some future work may include the construction of a valid variance estimator.

The covariates that affect the treatment assignment are the most important to be balanced. Covariate balancing should focus more on treatment related covariates. Assessing the correlation to treatment assignment mechanism is important. As stated in the conclusion of Chapter 3, we can assign weights to each covariate to focus on achieving balance on important covariates.

More than two models can be combined, e.g., $\widehat{e}_c(\mathbf{X}) = \sum_{i=1}^{k} \lambda_i \widehat{e}_i(\mathbf{X})$ subject to $\sum_{i=1}^{k} \lambda_i = 1$, where $\widehat{e}_i(\mathbf{X})$ is the estimated propensity score from the ith model. It is believed that more propensity score models can provide more protection to misspecification and higher efficiency.

In Chapter 4, we use combined kernel functions to replace a single Gaussian kernel. In the combined estimator estimator, we treat each kernel function equally. How to weight

each kernel and how to choose the tuning parameters like the σ^2 in Gaussian kernel are also worth exploration. Hazlett (2015) discussed the bias-variance trade-off of the σ^2 on mean balance of outcome variable in the control group. A larger σ^2 may leads to smaller variance but larger bias, while a smaller σ^2 can lead to smaller bias but larger variance. Although our simulation studies showed that median of σ^2 can be a good choice, a further study can explore a data driven and automatic selection for σ^2.

The first two projects only consider the modification based on IPW approach to reduce the effect of misspecification. Actually there are many double robust methods available. Targeted maximum likelihood estimation (TMLE) has been developed to estimate average causal effect. We can also incorporate our approaches with the TMLE method to estimate the causal effect we are interested in.

Application of covariate balancing idea in dynamic treatment regimes (DTR) problem can be a future direction. Multi-stage treatment is necessary in cancer treatment and some chronic illnesses. A dynamic treatment regime consists of a sequence of decision rules on the treatment assignment. Finding the optimal combination of treatment is important and relies on the correct estimation of treatment effect at each stage. IPW approaches and some double robust estimation of DTR have been applied to study optimal DTR problem (Moodie et al., 2012; Wallace and Moodie, 2015). A further study may include incorporating the idea of covariate balancing optimization and multiple robustness into each stage of DTR problem to adjust for confounding effect and reduce model misspecification effect since traditional IPW methods are sensitive to model misspecification.

References

Abrevaya, J. (2001). The effects of demographics and maternal behavior on the distribution of birth outcomes. *Empirical Economics* **26,** 247–257.

Aronszajn, N. (1950). Theory of reproducing kernels. *Transactions of the American Mathematical Society* **68,** 337–404.

Breiman, L. (1996). Bagging predictors. *Machine Learning* **24,** 123–140.

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

Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. (1984). *Classification and Regression Trees*. Chapman & Hall, London.

Brookhart, M. A. and van der Laan, M. J. (2006). A semiparametric model selection criterion with applications to the marginal structural model. *Computational Statistics & Data Analysis* **50,** 475–498.

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.

Chen, M., Thompson, M. E., and Wu, C. (2018). Empirical likelihood methods for complex surveys with data missing-by-design. *Statistica Sinica* **28,** 2027–2048.

Chen, S. and Haziza, D. (2017). Multiply robust imputation procedures for the treatment of item nonresponse in surveys. *Biometrika* **104,** 439–453.

Cochran, W. G. and Rubin, D. B. (1973). Controlling bias in observational studies: A review. *Sankhyā: The Indian Journal of Statistics, Series A* **35,** 417–446.

Cole, S. R. and Frangakis, C. E. (2009). The consistency statement in causal inference: a definition or an assumption? *Epidemiology* **20,** 3–5.

Diamond, A. and Sekhon, J. S. (2013). Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies. *Review of Economics and Statistics* **95,** 932–945.

Díaz, I. (2015). Efficient estimation of quantiles in missing data models. *arXiv preprint arXiv:1512.08110* .

Ehrenthal, D. B., Maiden, K., Rao, A., West, D. W., Gidding, S. S., Bartoshesky, L., Carterette, B., Ross, J., and Strobino, D. (2013). Independent relation of maternal prenatal factors to early childhood obesity in the offspring. *Obstetrics & Gynecology* **121,** 115–121.

Ehrenthal, D. B., Wu, P., and Trabulsi, J. (2016). Differences in the protective effect of exclusive breastfeeding on child overweight and obesity by mother's race. *Maternal and Child Health Journal* **20,** 1971–1979.

Fan, J., Imai, K., Liu, H., Ning, Y., and Yang, X. (2016). Improving covariate balancing oropensity score: a doubly robust and efficient approach. Technical report, Princeton University.

Firpo, S. (2007). Efficient semiparametric estimation of quantile treatment effects. *Econometrica* **75,** 259–276.

Fong, C. and Imai, K. (2014). Covariate balancing propensity score for general treatment regimes. *Princeton Manuscript* .

Fong, C., Ratkovic, M., and Imai, K. (2014). *CBPS: R package for covariate balancing propensity score*. R package version 0.9.

Friedman, J., Hastie, T., and Tibshirani, R. (2001). *The Elements of Statistical Learning*. Springer, Berlin.

Gönen, M. and Alpaydın, E. (2011). Multiple kernel learning algorithms. *Journal of Machine Learning Research* **12,** 2211–2268.

Graham, B. S., de Xavier Pinto, C. C., and Egel, D. (2012). Inverse probability tilting for moment condition models with missing data. *The Review of Economic Studies* **79,** 1053–1079.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. (2012). A kernel two-sample test. *Journal of Machine Learning Research* **13,** 723–773.

Gruber, S. and van der Laan, M. J. (2011). tmle: An r package for targeted maximum likelihood estimation.

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

Han, P. (2014). Multiply robust estimation in regression analysis with missing data. *Journal of the American Statistical Association* **109,** 1159–1173.

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* **50,** 1029–1054.

Harder, V. S., Stuart, E. A., and Anthony, J. C. (2010). Propensity score techniques and the assessment of measured covariate balance to test causal associations in psychological research. *Psychological Methods* **15,** 234–249.

Hazlett, C. (2015). Kernel balancing: A flexible non-parametric weighting procedure for estimating causal effects. *Unpublished manuscript* .

Hernan, M. A. and Vander Weele, T. J. (2011). Compound treatments and transportability of causal inference. *Epidemiology* **22,** 368–377.

Heyde, C. and Brown, B. (1970). On the departure from normality of a certain class of martingales. *The Annals of Mathematical Statistics* **41,** 2161–2165.

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.

Holland, P. W., Glymour, C., and Granger, C. (1985). *Statistics and Causal Inference*. Wiley Online Library.

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.

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

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

Imai, K. and Ratkovic, M. (2015). Robust estimation of inverse probability weights for marginal structural models. *Journal of the American Statistical Association* **110,** 1013–1023.

Imbens, G. W. and Rubin, D. B. (2015). *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press, New York.

Joshi, S., Kommaraji, R. V., Phillips, J. M., and Venkatasubramanian, S. (2011). Comparing distributions and shapes using the kernel distance. In *Proceedings of the Twenty-seventh Annual Symposium on Computational Geometry*, pages 47–56. ACM.

Kallus, N. (2016). Generalized optimal matching methods for causal inference. *arXiv preprint arXiv:1612.08321* .

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.

Koenker, R. and Hallock, K. (2001). Quantile regression: An introduction. *Journal of Economic Perspectives* **15,** 43–56.

Kouassi, D. A. and Singh, J. (1997). A semiparametric approach to hazard estimation with randomly censored observations. *Journal of the American Statistical Association* **92,** 1351–1355.

Kullback, S. (1959). *Statistics and Information theory.* J Wiley Sons, New York.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics* **22,** 79–86.

Laurence, M., Strawn, G., and Zuguo, M. (2004). Does breastfeeding protect against pediatric overweight? analysis of longitudinal data from the center for disease control an prevention pediatric nutrition surveillance system. *Pediatrics* **113,** e81–e86.

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

Lee, B. K., Lessler, J., and Stuart, E. A. (2011). Weight trimming and propensity score weighting. *PLoS ONE* **6,** e18174.

Liaw, A. and Wiener, M. (2002). Classification and regression by randomforest. *R News* **2,** 18–22.

Lumley, T. (2004). Analysis of complex survey samples. *Journal of Statistical Software* **9,** 1–19.

Lumley, T. et al. (2004). Analysis of complex survey samples. *Journal of Statistical Software* **9,** 1–19.

Mays, J. E., Birch, J. B., and Alden Starnes, B. (2001). Model robust regression: combining parametric, nonparametric, and semiparametric methods. *Journal of Nonparametric Statistics* **13,** 245–277.

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,** 403–425.

Mercer, J. (1909). Functions of positive and negative type, and their connection the theory of integral equations. *Philosophical Transactions of The Royal Society of London, Series A* **209,** 415–446.

Moodie, E. E., Chakraborty, B., and Kramer, M. S. (2012). Q-learning for estimating optimal dynamic treatment rules from observational data. *Canadian Journal of Statistics* **40,** 629–645.

Newey, W. K. and McFadden, D. (1994). Large sample estimation and hypothesis testing. *Handbook of Econometrics* **4,** 2111–2245.

Nottingham, Q. J. and Birch, J. B. (2000). A semiparametric approach to analysing dose–response data. *Statistics in Medicine* **19,** 389–404.

Olkin, I. and Spiegelman, C. H. (1987). A semiparametric approach to density estimation. *Journal of the American Statistical Association* **82,** 858–865.

Owen, A. B. (2001). *Empirical Likelihood*. CRC press, London/Boca Raton.

Pirracchio, R., Petersen, M. L., and van der Laan, M. (2015). Improving propensity score estimators' robustness to model misspecification using super learner. *American Journal of Epidemiology* **181,** 108–119.

Qin, J. and Lawless, J. (1994). Empirical likelihood and general estimating equations. *The Annals of Statistics* **22,** 300–325.

Qin, J., Shao, J., and Zhang, B. (2008). Efficient and doubly robust imputation for covariate-dependent missing responses. *Journal of the American Statistical Association* **103,** 797–810.

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.

Ramdas, A., Reddi, S. J., Póczos, B., Singh, A., and Wasserman, L. A. (2015). On the decreasing power of kernel and distance based nnonparametric hypothesis tests in high dimensions. In *The Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI-15)*, pages 3571–3577.

Read, T. R. and Cressie, N. A. (2012). *Goodness-of-fit Statistics for Discrete Multivariate Data*. Springer Science & Business Media.

Rehkopf, D. H., Glymour, M. M., and Osypuk, T. L. (2016). The consistency assumption for causal inference in social epidemiology: when a rose is not a rose. *Current Epidemiology Reports* **3,** 63–71.

Ridgeway, G. (2015). *gbm: Generalized Boosted Regression Models*. R package version 2.1.1.

Ridgeway, G., McCaffrey, D., Morral, A., Ann, B., and Burgette, L. (2015). *twang: Toolkit for Weighting and Analysis of Nonequivalent Groups*. R package version 1.4-9.3.

Robins, J. M., Hernan, M. A., and Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology* **11,** 550–560.

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

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association* **90,** 106–121.

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.

Rotnitzky, A., Robins, J. M., and Scharfstein, D. O. (1998). Semiparametric regression for repeated outcomes with nonignorable nonresponse. *Journal of the American Statistical Association* **93,** 1321–1339.

Rubin, D. B. (1973a). Matching to remove bias in observational studies. *Biometrics* **29,** 159–183.

Rubin, D. B. (1973b). The use of matched sampling and regression adjustment to remove bias in observational studies. *Biometrics* **29,** 185–203.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* **66,** 688–701.

Rubin, D. B. (2007). The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials. *Statistics in Medicine* **26,** 20–36.

Rubin, D. B. (2008). Causal inference using potential outcomes: design, modeling, decisions. *American Statistician* **62,** 277–278.

Schafer, J. L. and Kang, J. (2008). Average causal effects from nonrandomized studies: a practical guide and simulated example. *Psychological Methods* **13,** 279.

Scharfstein, D. O., Rotnitzky, A., and Robins, J. M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association* **94,** 1096–1120.

Shawe-Taylor, J. and Cristianini, N. (2004). *Kernel Methods for Pattern Analysis*. Cambridge: Cambridge University Press.

Speiser, P. W., Rudolf, M. C., Anhalt, H., Camacho-Hubner, C., Chiarelli, F., Eliakim, A., Freemark, M., Gruters, A., Hershkovitz, E., Iughetti, L., et al. (2005). Childhood obesity. *The Journal of Clinical Endocrinology & Metabolism* **90,** 1871–1887.

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

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

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science* **25,** 1–21.

Stuart, E. A., Lee, B. K., and Leacy, F. P. (2013). Prognostic score–based balance measures can be a useful diagnostic for propensity score methods in comparative effectiveness research. *Journal of Clinical Epidemiology* **66,** S84–S90.

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

Tan, Z. and Wu, C. (2015). Generalized pseudo empirical likelihood inferences for complex surveys. *Canadian Journal of Statistics* **43,** 1–17.

Thompson, M. E., Fong, G. T., Hammond, D., Boudreau, C., Driezen, P., Hyland, A., et al. (2006). Methods of the international tobacco control (itc) four country survey. *Tobacco Control* **15,** iii12–iii18.

van der Laan, M. J. (2014). Targeted estimation of nuisance parameters to obtain valid statistical inference. *The International Journal of Biostatistics* **10,** 29–57.

van der Laan, M. J. and Gruber, S. (2010). Collaborative double robust targeted maximum likelihood estimation. *The International Journal of Biostatistics* **6,** 1–69.

van der Laan, M. J., Polley, E. C., and Hubbard, A. E. (2007). Super learner. *Statistical Applications in Genetics and Molecular Biology* **6,** 1544–6115.

van der Laan, M. J. and Rose, S. (2011). *Targeted learning: Causal Inference for Observational and Experimental Data.* Springer Science & Business Media.

Wallace, M. P. and Moodie, E. E. (2015). Doubly-robust dynamic treatment regimen estimation via weighted least squares. *Biometrics* **71,** 636–644.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica* **50,** 1–25.

Wong, R. K. and Chan, K. C. G. (2017). Kernel-based covariate functional balancing for observational studies. *Biometrika* **105,** 199–213.

Wu, C. (2005). Algorithms and r codes for the pseudo empirical likelihood method in survey sampling. *Survey Methodology* **31,** 239–243.

Wu, C. and Rao, J. (2010). Bootstrap procedures for the pseudo empirical likelihood method in sample surveys. *Statistics & Probability Letters* **80,** 1472–1478.

Xie, Y., Zhu, Y., Cotton, C. A., and Wu, P. (2017). A model averaging approach for estimating propensity scores by optimizing balance. *Statistical Methods in Medical Research.* doi: 10.1177/0962280217715487.

Xu, D., Daniels, M. J., and Winterstein, A. G. (2017). Causal inference on quantiles with application to electronic ealth records. *Submitted to Biometrics* .

Zhan, X. and Ghosh, D. (2015). Incorporating auxiliary information for improved prediction using combination of kernel machines. *Statistical Methodology* **22,** 47–57.

Zhang, J., Troendle, J., Reddy, U. M., Laughon, S. K., Branch, D. W., Burkman, R., Landy, H. J., Hibbard, J. U., Haberman, S., Ramirez, M. M., et al. (2010). Contemporary cesarean delivery practice in the united states. *American Journal of Obstetrics and Gynecology* **203,** 326e1–326e10.

Zhang, Z., Chen, Z., Troendle, J. F., and Zhang, J. (2012). Causal inference on quantiles with an obstetric application. *Biometrics* **68,** 697–706.

Zhao, Q. (2016). Covariate balancing propensity score by tailored loss functions. *arXiv preprint arXiv:1601.05890* .

Zhao, Q. and Percival, D. (2017). Entropy balancing is doubly robust. *Journal of Causal Inference* **5,** 2193–3685.

Zhu, Y., Ghosh, D., Coffman, D. L., and Savage, J. S. (2016). Estimating controlled direct effects of restrictive feeding practices in the 'early dieting in girls' study. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **65,** 115–130.

Zhu, Y., Ghosh, D., Mitra, N., and Mukherjee, B. (2014). A data-adaptive strategy for inverse weighted estimation of causal effects. *Health Services and Outcomes Research Methodology* **14,** 69–91.

Zhu, Y., Savage, J. S., and Ghosh, D. (2018). A kernel-based metric for bblance assessment. *Journal of Causal Inference.* doi: 10.1177/0962280217715487.

Zhu, Y., Schonbach, M., Coffman, D. L., and Williams, J. S. (2015). Variable selection for propensity score estimation via balancing covariates. *Epidemiology* **26,** e14–e15.

Zolotarev, V. M. (1983). Probability metrics. *Teoriya Veroyatnostei i ee Primeneniya* **28,** 264–287.

APPENDICES

Appendix A

More Simulation Results

A.1 Appendix for Chapter 3

A.1.1 More Simulation Results for Chapter 3

Table A.1: Performance of Measures by Propensity Score Models in 1000 Simulated Data Sets with $n = 100$ (Estimation of ACE)

Measure	Method	Scenarios						
(×100)		A	B	C	D	E	F	G
Mean of absolute biases in percentage	C1	13.02	13.01	10.75	15.20	15.43	15.16	11.62
	C2	13.44	13.30	10.98	15.54	15.59	15.22	11.82
	C3	13.32	13.11	10.93	15.46	15.46	14.78	11.62
	C4	13.65	13.13	11.21	15.09	15.38	14.72	11.31
	LR	16.18	16.71	14.17	20.83	20.17	19.71	17.73
	RF	20.10	18.95	19.14	20.20	20.14	19.71	18.73
	CBPS	17.54	16.90	16.08	18.39	18.63	19.15	16.85
	GBM	26.42	24.31	23.63	27.62	25.81	26.22	24.05
Empirical standard error	C1	6.51	6.32	5.59	7.64	7.58	7.58	6.03
	C2	6.79	6.48	5.71	7.83	7.73	7.70	6.21
	C3	6.66	6.49	5.72	7.72	7.66	7.26	6.10
	C4	6.62	6.16	5.64	7.38	7.36	7.11	5.81
	LR	9.44	9.54	8.63	12.42	11.59	11.26	11.45
	RF	9.86	9.19	9.57	10.1	10.51	10.0	9.47
	CBPS	6.90	6.80	6.77	7.06	7.51	7.28	7.15
	GBM	9.55	8.85	9.58	9.63	9.79	9.43	9.80
Absolute bias of average ACE	C1	2.02	2.28	1.09	1.93	2.40	2.19	0.01
	C2	2.12	2.36	1.13	1.93	2.42	2.04	0.01
	C3	2.06	2.18	1.21	1.97	2.31	2.38	0.04
	C4	2.51	2.72	1.55	2.49	3.01	2.67	0.53
	LR	0.82	1.08	0.08	0.16	1.13	0.53	2.54
	RF	2.16	2.31	1.54	1.94	1.97	2.09	1.04
	CBPS	5.24	4.93	4.25	5.69	5.32	5.92	4.31
	GBM	8.76	8.12	7.08	9.38	8.42	9.13	6.70
Mean squared error	C1	0.46	0.45	0.32	0.62	0.63	0.62	0.36
	C2	0.51	0.48	0.34	0.65	0.66	0.63	0.39
	C3	0.49	0.47	0.34	0.63	0.64	0.58	0.37
	C4	0.50	0.45	0.34	0.61	0.63	0.58	0.34
	LR	0.90	0.92	0.74	1.54	1.36	1.27	1.38
	RF	1.02	0.90	0.94	1.06	1.14	1.04	0.91
	CBPS	0.75	0.71	0.64	0.82	0.85	0.88	0.70
	GBM	1.68	1.44	1.42	1.81	1.67	1.72	1.41
Average λ	C1	85.95	86.76	88.56	87.05	87.20	87.79	87.74
	C2	80.43	80.22	84.46	81.85	82.47	81.54	83.76
	C3	82.83	82.81	85.21	82.08	82.15	81.43	82.92
	C4	70.54	71.17	77.09	73.15	72.11	74.07	77.30

In each cell, all the numbers are multiplied by 100. C1: Model averaging method with mean ASMD ; C2: Model averaging method with median ASMD; C3: Model averaging method with max ASMD; C4: Model averaging method with mean KS statistic; LR: Logistic regression; RF: Random forest; CBPS: Covariate balancing propensity score; GBM: Generalized boosted model; CI: Confidence interval.

Table A.2: Performance of Measures by Propensity Score Models in 1000 Simulated Data Sets with $n = 5000$ (Estimation of ACE)

Measure	Method	Scenarios						
		A	B	C	D	E	F	G
Mean of absolute biases in percentage (×100)	C1	1.55	1.54	0.92	2.02	1.85	1.96	2.81
	C2	1.56	1.57	1.00	2.02	1.90	1.95	3.46
	C3	1.51	1.51	0.91	2.03	1.86	2.02	2.30
	C4	1.51	1.58	3.41	1.91	2.08	1.62	2.24
	LR	1.69	1.66	1.51	2.48	2.23	3.20	6.32
	RF	5.64	5.40	5.94	6.52	5.20	6.29	4.83
	CBPS	2.13	2.38	1.95	2.36	3.16	2.01	2.81
	GBM	5.20	5.85	6.40	6.52	7.58	6.48	7.84
Empirical standard error (×100)	C1	0.77	0.75	0.45	1.00	0.93	0.78	0.53
	C2	0.78	0.77	0.48	1.00	0.95	0.83	0.94
	C3	0.75	0.73	0.46	1.00	0.93	0.80	0.49
	C4	0.75	0.77	1.31	0.96	1.05	0.78	1.12
	LR	0.87	0.83	0.59	1.21	1.14	1.08	0.83
	RF	1.30	1.43	1.90	1.80	1.78	1.80	1.91
	CBPS	0.86	0.88	0.75	1.05	1.02	0.87	0.90
	GBM	0.74	0.81	0.98	0.88	0.90	0.86	1.02
Absolute bias of average ACE (×100)	C1	0.07	0.19	0.13	0.13	0.15	0.56	1.12
	C2	0.06	0.19	0.18	0.14	0.14	0.52	1.31
	C3	0.08	0.20	0.04	0.14	0.16	0.60	0.91
	C4	0.10	0.20	1.07	0.003	0.08	0.21	0.08
	LR	0.02	0.08	0.49	0.39	0.14	1.15	2.53
	RF	2.22	2.08	2.19	2.53	1.88	2.39	1.54
	CBPS	0.62	0.81	0.58	0.62	1.15	0.56	1.04
	GBM	2.08	2.34	2.55	2.60	3.03	2.59	3.14
Mean squared error ($\times 10^4$)	C1	0.60	0.60	0.22	1.02	0.89	0.92	1.54
	C2	0.61	0.63	0.26	1.02	0.92	0.96	2.60
	C3	0.57	0.57	0.21	1.02	0.89	1.00	1.07
	C4	0.57	0.63	2.86	0.92	1.11	0.65	1.26
	LR	0.76	0.70	0.59	1.62	1.32	2.49	7.09
	RF	6.62	6.37	8.41	9.64	6.70	8.95	6.02
	CBPS	1.12	1.43	0.90	1.49	2.36	1.07	1.89
	GBM	4.87	6.13	7.46	7.53	9.99	7.45	10.90
Average λ (×100)	C1	95.77	96.52	95.56	94.31	95.90	90.42	90.82
	C2	93.01	94.69	95.77	91.03	92.84	82.40	86.47
	C3	94.30	95.62	94.19	93.25	95.32	90.40	88.94
	C4	90.03	83.95	10.47	83.03	63.06	78.21	24.99

In each cell, the numbers are multiplied by 100, except for mean squared error, the numbers are multiplied by 10000. C1: Model averaging method with mean ASMD ; C2: Model averaging method with median ASMD; C3: Model averaging method with max ASMD; C4: Model averaging method with mean KS statistic; LR: Logistic regression; RF: Random forest; CBPS: Covariate balancing propensity score; GBM: Generalized boosted model; CI: Confidence interval.

A.2 Appendix for Chapter 5

A.2.1 More Simulation Results

Simulation with $n = 1000$ for other Quantiles

Table A.3: Scenario 1A: Simulation Results of QTE and DR Estimators with 25th Percentile ($n = 1000$)

	BIAS	ESE	RMSE	BSE	CR
QTE_12A	0.0086	0.3982	0.3983	0.4120	0.962
QTE_12B	-0.0022	0.5516	0.5516	0.5651	0.946
QTE_1AB	0.0086	0.4007	0.4008	0.4111	0.961
QTE_2AB	0.0097	0.3965	0.3967	0.4065	0.963
QTE_12AB	0.0077	0.3986	0.3986	0.4114	0.965
DR_1A	-0.0390	0.6718	0.6729	0.7008	0.964
DR_2A	-0.1678	0.7968	0.8143	0.7910	0.949
DR_1B	-0.0616	0.9085	0.9106	0.9344	0.967
DR_2B	-0.9280	1.0667	1.4138	1.1223	0.874

RMSE: Root mean squared error, ESE: Empirical standard error, BSE: Bootstrapped standard error, CR: Coverage rate for 95% bootstrapped confidence interval.

Table A.4: Scenario 2A: Simulation Results of QTE and DR Estimators with 25th Percentile ($n = 1000$)

	BIAS	ESE	RMSE	BSE	CR
QTE_12A	-0.0110	0.4408	0.4410	0.4566	0.951
QTE_12B	0.0224	0.7748	0.7752	0.7437	0.928
QTE_1AB	-0.0029	0.4397	0.4398	0.4496	0.952
QTE_2AB	-0.0098	0.4403	0.4404	0.4539	0.949
QTE_12AB	-0.0103	0.4418	0.4419	0.4550	0.952
DR_1A	0.2165	0.7965	0.8254	0.7014	0.938
DR_2A	-0.0840	0.9080	0.9118	0.8779	0.953
DR_1B	0.9787	0.7390	1.2264	0.7415	0.743
DR_2B	-0.0647	1.0932	1.0951	1.1687	0.968

RMSE: Root mean squared error, ESE: Empirical standard error, BSE: Bootstrapped standard error, CR: Coverage rate for 95% bootstrapped confidence interval.

Table A.5: Scenario 1B: Simulation Results of QTE and DR Estimators with 25th Percentile ($n = 1000$)

	BIAS	ESE	RMSE	BSE	CR
QTE_12A	0.0437	0.4977	0.4997	0.4942	0.930
QTE_12B	0.0076	0.3245	0.3246	0.3346	0.955
QTE_1AB	0.0075	0.3251	0.3252	0.3342	0.960
QTE_2AB	0.0067	0.3171	0.3172	0.3271	0.954
QTE_12AB	0.0075	0.3254	0.3255	0.3347	0.958
DR_1A	0.1672	1.3324	1.3429	1.2196	0.968
DR_2A	0.9229	2.0104	2.2122	1.1880	0.844
DR_1B	0.0403	0.6235	0.6248	0.7702	0.979
DR_2B	-0.0159	0.6534	0.6536	0.7960	0.979

RMSE: Root mean squared error, ESE: Empirical standard error, BSE: Bootstrapped standard error, CR: Coverage rate for 95% bootstrapped confidence interval.

Table A.6: Scenario 2B: Simulation Results of QTE and DR Estimators with 25th Percentile ($n = 1000$)

	BIAS	ESE	RMSE	BSE	CR
QTE_12A	0.0235	0.5583	0.5588	0.5541	0.934
QTE_12B	0.0056	0.3376	0.3376	0.3583	0.958
QTE_1AB	0.0041	0.3401	0.3401	0.3519	0.958
QTE_2AB	0.0024	0.3337	0.3337	0.3558	0.960
QTE_12AB	0.0051	0.3364	0.3364	0.3577	0.959
DR_1A	0.7380	1.0137	1.2539	1.1228	0.910
DR_2A	-0.0001	1.1040	1.1040	1.1087	0.973
DR_1B	0.0667	0.6837	0.6870	0.7646	0.984
DR_2B	-0.0221	0.7374	0.7377	0.7714	0.984

RMSE: Root mean squared error, ESE: Empirical standard error, BSE: Bootstrapped standard error, CR: Coverage rate for 95% bootstrapped confidence interval.

Table A.7: Scenario 1A: Simulation Results of QTE and DR Estimators with 75th Percentile ($n = 1000$)

	BIAS	ESE	RMSE	BSE	CR
QTE_12A	-0.0100	0.4069	0.4070	0.4282	0.951
QTE_12B	-0.0415	0.5712	0.5727	0.5971	0.944
QTE_1AB	-0.0149	0.4056	0.4059	0.4268	0.949
QTE_2AB	-0.0073	0.3999	0.3999	0.4213	0.955
QTE_12AB	-0.0150	0.4035	0.4037	0.4273	0.948
DR_1A	-0.0546	0.6924	0.6946	0.7252	0.967
DR_2A	-0.1364	0.8440	0.8550	0.8170	0.947
DR_1B	-0.0453	0.8195	0.8208	0.9489	0.962
DR_2B	-1.6571	0.9458	1.9080	0.9633	0.570

RMSE: Root mean squared error, ESE: Empirical standard error, BSE: Bootstrapped standard error, CR: Coverage rate for 95% bootstrapped confidence interval.

Table A.8: Scenario 2A: Simulation Results of QTE and DR Estimators with 75th Percentile ($n = 1000$)

	BIAS	ESE	RMSE	BSE	CR
QTE_12A	0.0057	0.4750	0.4750	0.4713	0.936
QTE_12B	-0.0211	0.7842	0.7845	0.7931	0.947
QTE_1AB	-0.0009	0.4741	0.4741	0.4673	0.939
QTE_2AB	0.0063	0.4628	0.4629	0.4678	0.939
QTE_12AB	0.0029	0.4695	0.4695	0.4696	0.934
DR_1A	-0.1449	0.6203	0.6370	0.6686	0.971
DR_2A	0.0463	0.8580	0.8592	0.8259	0.970
DR_1B	-0.7435	0.7271	1.0399	0.7578	0.832
DR_2B	0.0664	1.0383	1.0405	1.176	0.982

RMSE: Root mean squared error, ESE: Empirical standard error, BSE: Bootstrapped standard error, CR: Coverage rate for 95% bootstrapped confidence interval.

Table A.9: Scenario 1B: Simulation Results of QTE and DR Estimators with 75th Percentile ($n = 1000$)

	BIAS	ESE	RMSE	BSE	CR
QTE_12A	0.0300	0.6375	0.6382	0.6599	0.955
QTE_12B	0.0093	0.4149	0.4150	0.4507	0.956
QTE_1AB	0.0077	0.4158	0.4158	0.4499	0.958
QTE_2AB	0.0057	0.4023	0.4023	0.4372	0.961
QTE_12AB	0.0079	0.4150	0.4150	0.4500	0.956
DR_1A	0.4607	1.9962	2.0486	2.0338	0.964
DR_2A	1.5120	4.0672	4.3392	2.9629	0.950
DR_1B	0.1625	1.3719	1.3815	1.3957	0.965
DR_2B	0.1303	1.3526	1.3589	1.4482	0.969

RMSE: Root mean squared error, ESE: Empirical standard error, BSE: Bootstrapped standard error, CR: Coverage rate for 95% bootstrapped confidence interval.

Table A.10: Scenario 2B: Simulation Results of QTE and DR Estimators with 75th Percentile ($n = 1000$)

	BIAS	ESE	RMSE	BSE	CR
QTE_12A	-0.0036	0.8519	0.8519	0.8646	0.948
QTE_12B	0.0025	0.4753	0.4753	0.5026	0.947
QTE_1AB	-0.0065	0.4952	0.4953	0.5031	0.938
QTE_2AB	0.0042	0.4739	0.4739	0.4982	0.943
QTE_12AB	-0.0002	0.4774	0.4774	0.5022	0.941
DR_1A	-0.5870	1.4110	1.5282	1.4702	0.956
DR_2A	0.6528	2.1846	2.2801	2.2491	0.958
DR_1B	-0.2286	1.2058	1.2273	1.2119	0.965
DR_2B	0.3690	1.6036	1.6455	1.5336	0.952

RMSE: Root mean squared error, ESE: Empirical standard error, BSE: Bootstrapped standard error, CR: Coverage rate for 95% bootstrapped confidence interval.

Table A.11: Scenario 1A: Simulation Results of QTE and DR Estimators with 95th Percentile ($n = 1000$)

	BIAS	ESE	RMSE	BSE	CR
QTE_12A	0.0469	0.8023	0.8037	0.8679	0.964
QTE_12B	-0.0510	1.3685	1.3695	1.3484	0.918
QTE_1AB	0.0402	0.7991	0.8001	0.9068	0.975
QTE_2AB	0.1214	0.7962	0.8054	0.9162	0.970
QTE_12AB	0.0403	0.8005	0.8015	0.9167	0.977
DR_1A	0.0942	1.0157	1.0200	1.1220	0.972
DR_2A	-0.1186	1.0907	1.0971	1.1444	0.969
DR_1B	0.0936	1.6344	1.6371	1.6937	0.963
DR_2B	-1.9303	1.7712	2.6198	1.7689	0.805

RMSE: Root mean squared error, ESE: Empirical standard error, BSE: Bootstrapped standard error, CR: Coverage rate for 95% bootstrapped confidence interval.

Table A.12: Scenario 2A: Simulation Results of QTE and DR Estimators with 95th Percentile ($n = 1000$)

	BIAS	ESE	RMSE	BSE	CR
QTE_12A	-0.0268	0.7775	0.7779	0.8067	0.950
QTE_12B	-0.0549	1.3104	1.3115	1.3166	0.939
QTE_1AB	-0.0328	0.7719	0.7726	0.8087	0.957
QTE_2AB	-0.0313	0.7961	0.7968	0.8175	0.957
QTE_12AB	-0.0288	0.7942	0.7948	0.8178	0.959
DR_1A	-0.3326	0.8208	0.8856	0.9387	0.963
DR_2A	0.1132	1.0412	1.0473	1.1349	0.970
DR_1B	-1.3326	1.1545	1.7631	1.2849	0.815
DR_2B	0.1000	1.5586	1.5618	1.6996	0.975

RMSE: Root mean squared error, ESE: Empirical standard error, BSE: Bootstrapped standard error, CR: Coverage rate for 95% bootstrapped confidence interval.

Table A.13: Scenario 1B: Simulation Results of QTE and DR Estimators with 95th Percentile ($n = 1000$)

	BIAS	ESE	RMSE	BSE	CR
QTE_12A	0.0004	1.7401	1.7401	1.7114	0.943
QTE_12B	0.0207	0.9364	0.9366	0.9908	0.960
QTE_1AB	0.0263	0.9440	0.9444	1.0076	0.964
QTE_2AB	0.0178	0.9252	0.9254	1.0078	0.967
QTE_12AB	0.0209	0.9418	0.9420	1.0136	0.962
DR_1A	0.2486	2.9823	2.9927	2.6733	0.947
DR_2A	0.6615	4.1385	4.1910	3.4288	0.948
DR_1B	0.1421	1.2084	1.2167	1.4297	0.974
DR_2B	0.2063	1.4297	1.4445	1.5918	0.978

RMSE: Root mean squared error, ESE: Empirical standard error, BSE: Bootstrapped standard error, CR: Coverage rate for 95% bootstrapped confidence interval.

Table A.14: Scenario 2B: Simulation Results of QTE and DR Estimators with 95th Percentile ($n = 1000$)

	BIAS	ESE	RMSE	BSE	CR
QTE_12A	-0.1375	2.0235	2.0281	1.9224	0.934
QTE_12B	-0.0271	0.9844	0.9847	1.0708	0.973
QTE_1AB	-0.0435	0.9834	0.9844	1.1427	0.977
QTE_2AB	-0.0374	0.9902	0.9909	1.1413	0.976
QTE_12AB	-0.0311	0.9947	0.9952	1.1588	0.973
DR_1A	-4.2657	1.9529	4.6914	1.9865	0.311
DR_2A	-0.0324	2.9554	2.9556	2.910	0.952
DR_1B	-0.3879	1.1983	1.2595	1.2661	0.957
DR_2B	0.0670	1.4033	1.4049	1.4993	0.979

RMSE: Root mean squared error, ESE: Empirical standard error, BSE: Bootstrapped standard error, CR: Coverage rate for 95% bootstrapped confidence interval.

Simulation with Large Sample Size: $n = 5000$

Table A.15: Scenario 1A: Simulation Results of QTE and DR Estimators with 50th Percentile ($n = 5000$)

	BIAS	ESE	RMSE
QTE_12A	-0.0031	0.1535	0.1536
QTE_12B	-0.0036	0.2045	0.2045
QTE_1AB	-0.0029	0.1535	0.1535
QTE_2AB	-0.0035	0.1482	0.1482
QTE_12AB	-0.0031	0.1536	0.1536
DR_1A	-0.0050	0.1587	0.1588
DR_2A	-0.0074	0.1692	0.1693
DR_1B	-0.0042	0.2260	0.2261
DR_2B	-1.0471	0.2416	1.0746

RMSE: Root mean squared error, ESE: Empirical standard error.

Table A.16: Scenario 2A: Simulation Results of QTE and DR Estimators with 50th Percentile ($n = 5000$)

	BIAS	ESE	RMSE
QTE_12A	0.0103	0.1869	0.1872
QTE_12B	0.0046	0.3250	0.3251
QTE_1AB	0.0134	0.2345	0.2349
QTE_2AB	0.0095	0.1850	0.1852
QTE_12AB	0.0098	0.1874	0.1877
DR_1A	0.0044	0.1551	0.1551
DR_2A	0.0062	0.1912	0.1913
DR_1B	0.0335	0.2295	0.2319
DR_2B	0.0068	0.3398	0.3399

RMSE: Root mean squared error, ESE: Empirical standard error.

Table A.17: Scenario 1B: Simulation Results of QTE and DR Estimators with 50th Percentile ($n = 5000$)

	BIAS	ESE	RMSE
QTE_12A	-0.0017	0.2092	0.2092
QTE_12B	-0.0056	0.1416	0.1417
QTE_1AB	-0.0057	0.1417	0.1418
QTE_2AB	-0.0042	0.1417	0.1417
QTE_12AB	-0.0058	0.1416	0.1417
DR_1A	-0.0030	0.2299	0.2299
DR_2A	0.6945	2.0006	2.1178
DR_1B	-0.0058	0.1495	0.1496
DR_2B	-0.0046	0.1375	0.1375

RMSE: Root mean squared error, ESE: Empirical standard error.

Table A.18: Scenario 2B: Simulation Results of QTE and DR Estimators with 50th Percentile ($n = 5000$)

	BIAS	ESE	RMSE
QTE_12A	0.0032	0.2520	0.2520
QTE_12B	-0.0059	0.1652	0.1653
QTE_1AB	-0.0042	0.1933	0.1933
QTE_2AB	-0.0062	0.1635	0.1636
QTE_12AB	-0.0058	0.1641	0.1642
DR_1A	0.6287	0.2096	0.6627
DR_2A	0.0021	0.2592	0.2592
DR_1B	-0.0016	0.1451	0.1451
DR_2B	-0.0052	0.1729	0.1730

RMSE: Root mean squared error, ESE: Empirical standard error.

Table A.19: Scenario 3C: Simulation Results of QTE and DR Estimators with 50th Percentile ($n = 5000$)

	BIAS	ESE	RMSE
QTE_12A	-2.7931	0.3135	2.8106
QTE_12B	-2.8283	0.3404	2.8487
QTE_1AB	-2.5969	0.3452	2.6197
QTE_2AB	-2.8644	0.3002	2.8800
QTE_12AB	-2.7763	0.3122	2.7938
DR_1A	-2.5469	0.3194	2.5668
DR_2A	-2.8299	0.3955	2.8574
DR_1B	-2.3356	0.3847	2.3671
DR_2B	-2.8915	0.4349	2.9241

RMSE: Root mean squared error, ESE: Empirical standard error.

Simulation with Small Sample Size: $n = 200$

Table A.20: Scenario 1A: Simulation Results of QTE and DR Estimators with 50th Percentile ($n = 200$)

	BIAS	ESE	RMSE	BSE	CR
QTE_12A	-0.0021	0.7495	0.7495	0.8502	0.073
QTE_12B	-0.0259	1.0903	1.0906	1.1848	0.970
QTE_1AB	0.0033	0.7459	0.7459	0.8505	0.972
QTE_2AB	0.0082	0.7407	0.7407	0.8413	0.968
QTE_12AB	-0.0051	0.7498	0.7498	0.8545	0.970
DR_1A	-0.0365	0.8220	0.8228	1.0230	0.989
DR_2A	-0.0407	0.8814	0.8824	1.0550	0.978
DR_1B	0.0224	1.3896	1.3898	1.6112	0.985
DR_2B	-0.9911	1.2184	1.5706	1.4856	0.934

RMSE: Root mean squared error, ESE: Empirical standard error, BSE: Bootstrapped standard error, CR: Coverage rate for 95% bootstrapped confidence interval.

Table A.21: Scenario 2A: Simulation Results of QTE and DR Estimators with 50th Percentile ($n = 200$)

	BIAS	ESE	RMSE	BSE	CR
QTE_12A	-0.0081	0.8626	0.8627	0.9590	0.969
QTE_12B	0.0937	1.5214	1.5243	1.5349	0.960
QTE_1AB	-0.0110	0.8538	0.8538	1.0111	0.975
QTE_2AB	-0.0090	0.8583	0.8583	1.0012	0.974
QTE_12AB	-0.0128	0.8617	0.8618	1.0440	0.974
DR_1A	-0.0357	0.8016	0.8024	0.9430	0.983
DR_2A	-0.0431	1.0378	1.0387	1.1773	0.983
DR_1B	0.0461	1.1464	1.1473	1.4441	0.988
DR_2B	0.0468	1.8903	1.8909	1.9645	0.973

RMSE: Root mean squared error, ESE: Empirical standard error, BSE: Bootstrapped standard error, CR: Coverage rate for 95% bootstrapped confidence interval.

Table A.22: Scenario 1B: Simulation Results of QTE and DR Estimators with 50th Percentile ($n = 200$)

	BIAS	ESE	RMSE	BSE	CR
QTE_12A	0.1807	1.0704	1.0855	1.1330	0.961
QTE_12B	0.0150	0.7245	0.7246	0.7830	0.964
QTE_1AB	0.0177	0.7250	0.7252	0.7828	0.962
QTE_2AB	0.0081	0.6976	0.6976	0.7683	0.965
QTE_12AB	0.0219	0.7143	0.7146	0.7852	0.964
DR_1A	0.2873	1.4366	1.4650	1.6664	0.981
DR_2A	0.7314	1.4243	1.6011	1.6589	0.967
DR_1B	0.0656	0.8221	0.8248	1.0196	0.979
DR_2B	0.0507	0.7673	0.7690	0.9572	0.980

RMSE: Root mean squared error, ESE: Empirical standard error, BSE: Bootstrapped standard error, CR: Coverage rate for 95% bootstrapped confidence interval.

Table A.23: Scenario 2B: Simulation Results of QTE and DR Estimators with 50th Percentile ($n = 200$)

	BIAS	ESE	RMSE	BSE	CR
QTE_12A	0.1282	1.4204	1.4262	1.4608	0.960
QTE_12B	-0.0017	0.8004	0.8004	0.8789	0.967
QTE_1AB	-0.0085	0.7890	0.7891	0.9059	0.969
QTE_2AB	-0.0153	0.7876	0.7877	0.8965	0.969
QTE_12AB	-0.0142	0.7963	0.7964	0.9414	0.972
DR_1A	0.4553	1.1600	1.2462	1.4727	0.975
DR_2A	0.3122	1.7746	1.8018	1.8862	0.974
DR_1B	-0.0515	0.7166	0.7185	0.8884	0.979
DR_2B	0.0763	1.0265	1.0293	1.1279	0.975

RMSE: Root mean squared error, ESE: Empirical standard error, BSE: Bootstrapped standard error, CR: Coverage rate for 95% bootstrapped confidence interval.

Table A.24: Scenario 3C: Simulation Results of QTE and DR Estimators with 50th Percentile ($n = 200$)

	BIAS	ESE	RMSE	BSE	CR
QTE_12A	-0.8620	1.6408	1.8535	1.7346	0.918
QTE_12B	-0.9034	1.7983	2.0125	1.9202	0.927
QTE_1AB	-0.7826	1.6362	1.8137	1.7230	0.916
QTE_2AB	-1.0007	1.5436	1.8396	1.6495	0.895
QTE_12AB	-0.8271	1.6432	1.8396	1.7527	0.916
DR_1A	-0.9912	2.3116	2.5152	2.1615	0.970
DR_2A	-0.8293	2.2777	2.4240	2.4778	0.964
DR_1B	-0.8727	2.1542	2.3243	2.3227	0.946
DR_2B	-0.7914	2.4588	2.5830	2.6190	0.966

RMSE: Root mean squared error, ESE: Empirical standard error, BSE: Bootstrapped standard error, CR: Coverage rate for 95% bootstrapped confidence interval.