AN OLS-BASED METHOD FOR CAUSAL INFERENCE IN OBSERVATIONAL STUDIES

Yuanfang Xu

Submitted to the faculty of the University Graduate School

in partial fulfillment of the requirements

for the degree

Doctor of Philosophy

in the Department of Biostatistics,

Indiana University

July 2019

Accepted by the Graduate Faculty, Indiana University, in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Ying Zhang, Ph.D., Chair

Doctoral Committee

Bin Huang, Ph.D.

Wanzhu Tu, Ph.D.

May 8, 2019

Giorgos Bakoyannis, Ph.D.

Yiqing Song, M.D., Sc.D.

ACKNOWLEDGMENTS

I would like to extend my most sincere gratitudes to the people who have made my Ph.D. study at Indiana University an amazing and memorable experience.

My heartiest thanks and appreciation go to Dr. Ying Zhang, who has been an excellent, dedicated mentor guiding me all the way through my thesis research. I am privileged and honored to be one of Dr. Zhang's Ph.D. students. What I have learned from Dr. Zhang's broad and profound knowledge in statistics, constant enthusiasm, and commitment to excellence in academic research will be my lifetime benefits. I can't thank Dr. Zhang enough for his belief in me and his encouragement.

I am greatly appreciative to my co-advisor Dr. Bin Huang, who introduced me to the area of causal inference research and opened the door to my research for this dissertation. Thanks a lot to Dr. Huang for her inspiration, her generosity in sharing her thoughts and project data, her care and support during my employment at the Cincinnati Children's Hospital Medical Center.

My heart is full of gratitude to other committee members Dr. Wanzhu Tu, Dr. Giorgos Bakoyannis and Dr. Yiqing Song for taking their precious time in reading this dissertation. I am so grateful for Dr. Tu's kindness in agreeing to serve on my committee on short notice. Thanks for Dr. Bakoyannis's help and insightful discussions with Dr. Zhang and me in the last year. Thanks to Dr. Song for being my minor advisor and allowing me to attend his lectures of meta-analysis course remotely in Cincinnati. It really saved me lots of time in traveling!

I would also like to thank our Department Chair Dr. Barry Katz and Dr. Ying Zhang as the Director of Education in our department, for granting me the

permission of pursuing Ph.D. while working full time at my current job during the past three years. Some special thanks go to Dr. Fang Li, Dr. Fei Tan, Dr. Wei Zheng from the Department of Mathematics for their help and mentoring during my study for master degree at school of science.

Finally, yet very importantly, I must say millions of thanks to my dear husband Yucheng Xiao for his love, support and sacrifice. We have gone through many tough times together. There is no doubt that I couldn't reach to this point without his selfless devotion to our family. Thanks to my lovely daughter Jingyu Xiao and son Keyi Xiao, who bring me so much love, joy, pleasure, strength, and pride. Thanks to my parents for their love, help in taking care of my kids and everything they have ever done for me. I am blessed to have such a wonderful family that I treasure. I will always be grateful for that blessing.

Yuanfang Xu

AN OLS-BASED METHOD FOR CAUSAL INFERENCE IN OBSERVATIONAL STUDIES

Observational data are frequently used for causal inference of treatment effects on prespecified outcomes. Several widely used causal inference methods have adopted the method of inverse propensity score weighting (IPW) to alleviate the influence of confounding. However, the IPW-type methods, including the doubly robust methods, are prone to large variation in the estimation of causal effects due to possible extreme weights. In this research, we developed an ordinary least-squares (OLS)-based causal inference method, which does not involve the inverse weighting of the individual propensity scores.

We first considered the scenario of homogeneous treatment effect. We proposed a two-stage estimation procedure, which leads to a model-free estimator of average treatment effect (ATE). At the first stage, two summary scores, the propensity and mean scores, are estimated nonparametrically using regression splines. The targeted ATE is obtained as a plug-in estimator that has a closed form expression. Our simulation studies showed that this model-free estimator of ATE is consistent, asymptotically normal and has superior operational characteristics in comparison to the widely used IPW-type methods. We then extended our method to the scenario of heterogeneous treatment effects, by adding in an additional stage of modeling the covariate-specific treatment effect function nonparametrically while maintaining the model-free feature, and the simplicity of OLS-based estimation. The estimated

covariate-specific function serves as an intermediate step in the estimation of ATE and thus can be utilized to study the treatment effect heterogeneity.

We discussed ways of using advanced machine learning techniques in the proposed method to accommodate high dimensional covariates. We applied the proposed method to a case study evaluating the effect of early combination of biologic & non-biologic disease-modifying antirheumatic drugs (DMARDs) compared to step-up treatment plan in children with newly onset of juvenile idiopathic arthritis disease (JIA). The proposed method gives strong evidence of significant effect of early combination at 0.05 level. On average early aggressive use of biologic DMARDs leads to around 1.2 to 1.7 more reduction in clinical juvenile disease activity score at 6-month than the step-up plan for treating JIA.

Ying Zhang, Ph.D., Chair

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

Introduction

1.1 Causal Inference

Causality is at the core of scientific inquiring. For instance, does cigarettes smoking increase the risk of lung cancer? Is a given drug effective in treating patient with hypertension? Is an employment program helpful in moving the job trainees into the job market? Investigation of causality arises in many domains of science, and some view it as the ultimate goal of many scientific research. Learning causality helps researchers to understand the underlying mechanism through which the causal effects take place from the observed data. The important understanding serves as a basis for treatment actions and policy making.

Ronald Fisher's theory of experimental design (Fisher, 1935) tells us that causal inference in statistics is fundamentally based on the randomized experiments. Randomized controlled trials (RCTs) are widely recognized as the gold standard approach and most powerful design for drawing valid causality conclusions. Randomization ensures that the distribution of baseline characteristics for treated subjects and untreated subjects is balanced at baseline therefore treatment assignment is independent of either observed or unobserved baseline characteristics. As a result, the effect of treatment on outcomes can be simply estimated by comparing outcomes between treated and untreated subjects and has causal interpretation. Unfortunately, in practice, RCTs might not be able to carry out as initially plotted. For example, as

researchers aim to seek the evidence for adoption of the treatment (intervention) into real-world clinical practice, participants are usually allowed to switch off the treatment or switch onto the treatment from control although the treatment is randomly assigned at baseline. This non-compliance issue arisen in clinical trials was well illustrated by (MOBILITY, 2015), a randomized trial designed for investigating the effectiveness of metformin (MET) in obese and overweight youth with Bipolar Spectrum Disorders (BSD) and treated with SGAs (second generation antipsychotics). It was found that out of subjects randomized to the MET group at the baseline around $\frac{1}{4}$ of them didn't adhere to the assignment while around $\frac{1}{3}$ of subjects from the control group switched onto taking MET at certain time point for various reasons related to SGA adherence and weight gain during the follow-up. The simple intention-to-treat (ITT) analysis approach may then fail to provide an unbiased estimate that reflects the causal effect of MET in weight control in real life. We essentially need to deal with post-randomization confounding to account for non-adherence based on post-baseline observational data.

Needless to say that randomized experiments usually require substantial effort and in many situations random treatment allocation is infeasible for ethical or practical reasons. Many studies are observational (non-randomization) by nature. In public health, medical research as well as social science, researchers must resort to existing resource of observational studies to address causality. Similar to the non-compliance issue in RCTs, observational studies are complicated with possible selection bias and confounding. Lack of treatment randomization leads to potential systematic differences in baseline characteristics between treated and untreated subjects. Ignoring such systematic differences will likely result in a biased estimation of causal effect.

One can imagine that in a biomedical study, physicians make use of patients historical medical records to answer the question of "is a given drug effective in treating patients with hypertension". Suppose it was observed that people who were treated with the drug had lower blood pressure on average than those who were treated with standard care (control) at the end of study. Is it then justifiable to claim that the drug is effective? Without controlling for confounding factors, the effect physicians observed might be biased and misleading. Perhaps patients in the drug treated group are younger and have less other comorbidity conditions that could worsen the blood pressure although the drug has no effect. Or maybe the statistical association of "drug" and "blood pressure" holding for the entire study population gets reversed in subpopulations stratified by certain confounders, just as the well-known classical example of Simpsons Paradox (Simpson, 1951).

Over the past few decades, the methodology development of causal inference has been an active research area in various disciplines, such as statistics, computer science, economics and epidemiology, in order to address the challenges of inferring causality using observational data. Most importantly, the potential outcomes framework (Neyman, 1990; Rubin, 1974) laid the theoretical foundation for studying causality from observational data and has been widely adopted in statistical research of causal inference. The introduction of potential outcomes, however, transfers the causal inference problem to missing data problem as only a fraction of potential outcomes are observed. The observed outcome is produced jointly by two correlated data generating models: the underlying science model for the two potential outcomes and the treatment selection mechanism. Estimation of treatment effects can be conducted by modeling outcome or modeling the treatment selection. In statistical literature,

methods of handling missing potential outcomes have been focusing on using propensity score (Austin, 2009, 2011; Williamson et al., 2012). Among a variety of propensity score methods, the inverse propensity score weighting (IPW) method has been particularly the topic of research interest since it is considered as being central to the semi-parametric theory in missing data problem (Tsiatis, 2006; van der Laan and Robins, 2003). As propensity score methods rely on a correctly specified working propensity score model, some recent research interest in causal inference shift to a class of doubly robust (DR) methods (Heejung and Robins, 2005) typically augmenting the simple IPW estimator and promise to provide dual protection again either outcome model misspecification or propensity score model misspecification. However, these widely used methods have drawbacks frequently discussed in literature. IPW is well-known for being unstable and highly variable when the true or estimated propensity scores are close to 0 or 1 while DR methods require at least one correct model to perform well. Since realistically we don't possess accurate knowledge of either outcome model or treatment assignment model, it calls the need to derive a consistent and robust estimator of causal treatment effect without concerns of parametric specification for either model. Motivated by this desire, this thesis is devoted to developing a conceptually simple approach of constructing an estimator of treatment effects that ensures consistency and robustness yet allow quite flexibility in employing non-parametric methods to relax the parametric model assumptions.

To fully appreciate our proposed method, we first provide an overview of ongoing research in causal inference literature in this chapter. Starting with some basic concepts of causal modeling under the potential outcomes framework and the widely used standard causal assumptions for identifying treatment effect, we discuss the two

schools of methods for estimating average treatment effect: outcome modeling and treatment selection modeling. For the methods of using propensity scores, we focus on the discussion of IPW and its role in the semiparametric theory of missing data problem as well as its drawback. The doubly robust methods and its limitation are also reviewed followed by brief description of some existing nonparametric causal inference methods in literature.

1.2 Causal Models

Standard statistical analysis, typically regression analysis, is commonly used to make inference of associations among variables. To distinguish "causation" from "association" and facilitate the inferences of causality, we need causal models as formal tools to frame the causal questions rigorously in mathematical language with transparent assumptions.

1.2.1 Neyman-Rubin Causal Model

The most popular causal model is the so-called potential outcome framework or Neyman-Rubin model. The idea of "potential outcomes" or "counterfactural" was originally introduced in Neyman's non-parametric model in which $Y_x(u)$ was used to denote the unit-based response variable as "the value that the outcome Y would observed in unit u, had treatment X been x (Neyman, 1990). Initially potential outcomes were brought up as a way to deal with treatment effects in randomized experiments. Later this concept gradually evolved into a general framework that is also applicable to observational studies and it became the fundamental framework for causal inference established in a series of papers by (Cochran, 1968, 1973; Holland,

1986; Robins et al., 2000; Rosenbaum, 2002; Rosenbaum and Rubin, 1983, 1984; Rubin, 1974, 1976a, 1990; Scharfstein et al., 1999). Most modern causal inference methodologies are built upon this framework with wide applications in the area of statistics, medicine, economics, political science, and social science, etc.

The most notable feature of the potential outcome framework is that the theoretical definition of causality applies clearly to a single unit being observed in a study. In the above example of drug for hypertension, a patient is a single unit. A causal effect is defined precisely at each unit level using a set of potential outcomes that could be observed if hypothetically the corresponding treatment status were realized in real world. In the simplest case with the treatment variable A being the binary indicator of treated or not, let $(Y_i^{(1)}, Y_i^{(0)})$ denotes the pair of two potential outcomes for unit i when $A_i = 1$ (treated) or $A_i = 0$ (not treated). The causal effect for unit i can then be simply calculated as the difference between the two potential outcomes denoted as $\tau_i = Y_i^{(1)} - Y_i^{(0)}$. It is obviously that we now face the fundamental difficulty of estimating the causal average treatment effect since $Y_i^{(1)}$ and $Y_i^{(0)}$ can't be observed simultaneously. In other words, it is impossible to measure causal effects at the individual level from the observed outcomes.

1.2.2 Causal Assumptions and Average Treatment Effect

The most widely studied average treatment effects is the population average treatment effect (ATE) defined as

$$\tau = E(Y^{(1)}) - E(Y^{(0)})$$

As potential outcomes on different units can be used to estimate $E(Y^{(1)})$ or $E(Y^{(0)})$, τ is estimable under some assumptions. We use X to denote a fixed vector of pre-treatment variables indicating the baseline characteristics observed for a single unit in addition to the outcome Y and treatment A. The two random variables$(Y^{(1)}, Y^{(0)})$ are typically related to X statistically. Therefore in more explicit form we may write τ in term of $\tau(X)$, the conditional treatment effect given X.

$$\tau = E(\tau(X)) = E_X(E(Y^{(1)} - Y^{(0)}|X))$$

Given n individuals from a target population in a study, the observed data consists of n independent and identical copies of random variable $D = (Y, X, A)$, which provide valid estimates for the following two conditional expectations:

$$E(Y^{(1)}|A = 1); E(Y^{(0)}|A = 0)$$

However in an observation study, due to selection bias, A is not independent of the joint distribution of $(Y^{(1)}, Y^{(0)})$, which results in

$$E(Y^{(1)}|A = 1) \neq E(Y^{(1)}); E(Y^{(0)}|A = 0) \neq E(Y^{(0)})$$

Hence the ATE can't be directly estimated by the valid estimates of $E(Y^{(1)}|A = 1)$ and $E(Y^{(0)}|A = 0)$.

In order to link $E(Y^{(a)}|A = a)$ with $E(Y^{(a)})$ and identify τ, a set of standard assumptions are built as an important part of the causal model.

The key assumption of "strong ignorability" (Rosenbaum and Rubin, 1983) include two conditions for each unit i . One is the condition of "no unmeasured confounder", i.e., $(Y^{(1)}, Y^{(0)}) \perp A|X$ where $\perp$ denotes independence. It states that X is sufficient to be adjusted for in order to remove all the possible confounding between the relationship of A and Y. The other condition is "positivity" or "overlap" requiring that the probability of being treated or not treated for any unit with all possible values of X is positive. This is to ensure that given X, $Y^{(0)}$ and $Y^{(1)}$ are well defined for every unit. A weaker version of the ignorability assumption is $E(Y^{(a)}|A = a, X) = E(Y^{(a)}|X); a = 1, 0$, which is called weak ignorability (Rosenbaum and Rubin, 1984) or mean independence assumption (Imbens, 2004). This assumption seems to be weaker but it is argued that in very rare settings that the weak ignorability holds while the storng ignorability does not hold (Vansteelandt and Joffe, 2014).

Two more assumptions in addition to ignorablity are Rubins's stable unit-treatment value (SUTVA) assumption (Rubin, 1980, 1986, 1990) and consistency assumption (Cole and Frangakis, 2009; Pearl, 2000, 2010). SUTVA states that the treatment levels are identical across all the units and the potential outcomes for any unit do not depend on the treatment or outcome of other units. Consistency assumption is used to link the potential outcome and observed outcome as $Y = Y^{(1)}A + Y^{(0)}(1 - A)$.

With these standard assumptions, τ is identifiable by the observed outcomes since

$$\begin{aligned}\tau &= E_X(E(Y^{(1)} - Y^{(0)}|X)) \\ &= E_X(E(Y^{(1)}|A=1,X) - E(Y^{(1)}|A=0,X))) \\ &= E_X(E(Y|A=1,X) - E(Y|A=0,X)))\end{aligned}$$

In this thesis, we mainly focus on the estimation of ATE, the population level average treatment effect. It is worth mentioning though that other quantities of average treatment effect might be of interest in some particular research context such as the average treatment effect on the treated (ATT) $E(Y^{(1)} - Y^{(0)}|A = 1)$, the average treatment effect on the untreated (ATU) $E(Y^{(1)} - Y^{(0)}|A = 0)$. In some applications, it may be unrealistic to estimate the effect of the treatment if it were applied to all the subjects, then the ATT or ATU may be of greater interest than the ATE.

1.3 Methods for Estimating Average Treatment Effect

With the imposed assumptions discussed above, clearly we can specify distribution of the observed data with density P_0 under some dominating measure given by

$$P(D) = P(Y|X,A)P(A|X)P(X)$$

An important observation about this density is that it basically implies two correlated generating processes of observed data: (i) $P(Y|X,A)$ is about how the potential outcomes generate given X; (ii) $P(A|X)$ defines the how treatment is assigned given

X. The problem of estimating average treatment effect therefore can be tackled from two angles: outcome modeling and treatment selection modeling.

1.3.1 Outcome Regression

Assume that we know the true outcome generating process, i.e., the true functional form of $\mu(X, A) = E(Y|X, A)$, estimating τ is straightforward since

$$\tau^{OR} = \int (\mu(X, 1) - \mu(X, 0)) \, \mathrm{d}P(x) \tag{1.1}$$

If the treatment effect is homogeneous in the study population, the individual treatment effect can be written as $\tau_i = \tau + \epsilon_i; i = 1, \cdots n$ where ϵ_i's are random noise independent of X_i, τ is simply $\mu(X, 1) - \mu(X, 0)$.

In practice we may propose an outcome regression model, typically a parametric model, $m(X, A; \alpha)$ with α being the parameters to be estimated by the observed data through some estimation mechanism. When $m(X, A; \alpha)$ correctly specifies $\mu(X, A)$, a consistent estimator $\hat{\tau}^{OR}$ can be obtained. This regression adjustment method is intuitive and has been used frequently for estimating effects of treatment historically.

There are however some practical concerns of using outcome regression method in the causal inference. Firstly, the consistency of $\hat{\tau}^{OR}$ relies on a correct model for $\mu(X, A)$, which can be challenging in some real applications when no prior knowledge is possessed about how outcome relates to treatment assignment and baseline covariates. It can become even more challenging when there exists potential extrapolation bias (Glynn and Quinn, 2010; King and Zeng, 2006). The issue of extrapolation bias

arises in the situation, for example, when some range of X is dominated by treated units such that prediction of the counterfactuals over this range of X based on the model fitted for the observed data would extrapolate for those untreated units. If the true outcome generating mechanisms for the treated and untreated units are very different over this problematic range, the extrapolation could result in substantial bias in the estimation. The second concern is regarding the interpretation. Unless treatment effect is homogeneous, the estimator $\hat{\mu}(X,1) - \hat{\mu}(X,0)$ directly from the outcome regression model of $m(X, A; \alpha)$ has only the interpretation of conditional treatment effect given X. To obtain τ, which is the marginal treatment effect integrating X out $\hat{\mu}(X,1) - \hat{\mu}(X,0)$ through directly modeling the outcome, one widely used algorithm is the "g-formula" (Robins, 1986; Taubman et al., 2009) that involves outcome regression modeling and resampling-based methodology.

1.3.2 Propensity Score

Estimation of τ, on the other hand, may utilize the information of $P(A|X)$, the treatment selection model. $\pi(X) = P(A = 1|X)$ was formally defined as propensity score by (Rosenbaum and Rubin, 1983) as the probability of treatment assignment conditional on observed baseline covariates. The strong ignorability implies

$$E(Y^{(a)}|\pi(x)) = E(Y|A = a, \pi(x)), a = 1, 0$$

Thus the propensity score has an attractive property of being a balancing score: conditional on the propensity score, the distribution of measured baseline covariates would be similar between treated and untreated subjects. Hence $A \perp (Y^{(1)}, Y^{(0)})|\pi(x)$

(Rosenbaum and Rubin, 1983, 1984) and

$$\begin{aligned}\tau &= E_{\pi(X)}\left[E\{Y^{(1)}|\pi(X)\} - E\{Y^{(0)}|\pi(X)\}\right] \\ &= E_{\pi(X)}\left[E\{Y|\pi(X), A=1\} - E\{Y|\pi(X), A=0\}\right]\end{aligned}$$

The balancing property has made propensity score popular in causal inference for several reasons. First, the introduction of propensity scores makes it possible to separate the study design from the analysis, similar to a randomized study (Austin, 2011). By modeling the treatment assignment, we may view an observation study as a study being designed for assigning the treatment among the subjects with certain probabilities determined by their baseline characteristics. Thus propensity score provides a way to mimic an observation study to an RCT at the design stage. Second, as Rubin showed, the true propensity score is the finest balancing score while X is the coarsest balancing score (Rosenbaum and Rubin, 1983). When the dimensionality of X is high, it is much more convenient to check the balance of baseline characteristics between treated and untreated subjects using the one-dimension propensity score other than X itself. Third, comparing to outcome regression, the diagnostics of propensity score model by examining the distribution of X between treated and untreated subjects is much easier than assessing whether the relation among Y, A, X is correctly specified. In some situations if we are more knowledgeable about the treatment assignment mechanism than the outcome generating model, it offers a good alternative way to estimate treatment effects. Propensity score methods increasingly become a part of the standard toolkit for controlling confounding in observational studies.

In observational study, the propensity score is typically unknown therefore must be estimated from the data. The most commonly used method is logistic regression modeling in which a parametric model $\pi(X;\beta)$ is proposed to model treatment selection. Recently, more advanced methods are developed to address the issue of possible model misspecification in parametric logistic regression, such as bagging, boosting, random forests etc. (Lee et al., 2010; McCaffrey et al., 2004; Setoguchi et al., 2008). These machine learning methods have the advantage of being capable of handling high dimensional baseline covariates and potentially finding a good model for predicting the treatment assignment. Another school of method is the covariates balancing propensity score (CBPS) method (Imai and Ratkovic, 2014; Wyss et al., 2014) which takes into account dual characteristics of the propensity score as the conditional probability of treatment assignment as well as a balancing score. CBPS models $p(A = 1|X)$ while optimizing the covariate balance through a set of moment conditions induced by the mean independence between the A and X after inverse propensity score weighting.

Using propensity scores in adjusting for observed confounding has been mostly discussed around four different approaches in literature: (i) matching (Abadie and Imbens, 2006; Rosenbaum, 1989; Rosenbaum and Rubin, 1985) based on propensity score is an intuitive way and is considered as a common approach for making less biased causal inference (Rubin, 1973, 1976b,c); (ii) subclassification or stratification by propensity scores (Hansen, 2004; Rosenbaum, 1991; Rosenbaum and Rubin, 1984) can be conceptualized as a meta-analysis of a set of quasi-RCTs (Austin, 2011); (iii) covariates adjustment method uses propensity scores as covariate in regression (Little and An, 2004); (iv) inverse propensity score weighting (Hirano et al., 2003;

Robins et al., 2000; Rosenbaum, 1987) aims to create a pseudo population in which the distribution of X is independent of A (Austin, 2011). Comprehensive reviews of these methods, their limitations and comparisons of their performance in empirical studies can be found in (Austin, 2011; Imbens, 2004; Lunceford and Davidian, 2004; Stuart, 2010).

1.3.3 Inverse Propensity Score Weighting

Among the aforementioned propensity score methods, the inverse propensity score weighting(IPW) has especially under extensive study in causal inference due to its role in semi-parametric theory in missing data setting. It is also our focus of discussion in this thesis.

The idea of inverse probability weighting was first proposed in the context of surveys in a paper (Horvitz and Thompson, 1952). The Horvitz-Thompson estimator of τ is

$$\tau^{HT} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y_i A_i}{\pi(X_i)} - \frac{Y_i(1-A_i)}{1-\pi(X_i)}\right) \tag{1.2}$$

A landmark paper (Robins et al., 1994) discovered its role in the semiparametric inference in missing data setting. This paper derived a class of all consistent and asymptotically normal semiparametric estimators for parameters in a semiparametric full data model when data are missing at random and showed the estimators in this class can be expressed as solutions to estimating equations that involve inverse probability weighting. (Lunceford and Davidian, 2004) used the theory of M-estimation to show that when $\pi(x)$ is correctly specified τ^{HT} is a consistent estimator of τ and has a limiting normal distribution with $\sqrt{n}$ converge rate to true τ. (Tsiatis, 2006) also

provides a very good discussion of inverse probability weighting and its connection to semi-parametric theory.

Regardless of its theoretical foundation and popularity, the poor performance of IPW method in numerical studies has also been recognized and frequently discussed in literatures. In (Tsiatis, 2006), he stated that "there is a technical condition that $\pi(X_i)$ be strictly greater than 0 for all values of X in the support of X in order that the IPW estimator be consistent and asymptotically normal" and "even if this technical condition holds true, if $\pi(X_i)$ is very small, then this gives undue influence to the i-th observation in the IPW estimator and could result in a very unstable estimator with poor performance with small to moderate sample sizes". (Gutman and Rubin, 2017) reviewed the commonly used statistical procedures for estimating ATE including matching, sub-classification, weighting and model-based adjustment. They showed under extensive simulations that the widely used propensity score method has "poor operating characteristics" and pointed out the potential drawback of inverse weighting: when some $\pi(X_i)$ are close to 0 or 1, a few observations dominate the estimated treatment effect resulting in large sampling variance and it can be even worse when the propensity score model is misspecified. Peter C. Austin has written a handful of research papers regarding propensity score methods. In his relatively recent paper (Austin and Stuart, 2015), he discussed the difficulty arising in the application of IPW method when treated subjects have a very low propensity score resulting in very large weights or some control subjects with a propensity score close to one can result in very large weights, which explicitly indicates uncertainty in the estimation of ATE.

1.4 Doubly Robust Estimation

The outcome regression or propensity score method requires specification of $\mu(X, A)$ and $\pi(X)$, respectively, typically with parametric modeling assumptions. Consistency of $\hat{\tau}^{OR}$ or $\hat{\tau}^{HT}$ depends on the correctly specified model of $\mu(X, A)$ or $\pi(X)$, which could be very challenging when distribution of the observed data is unknown and complicated. The doubly robust (DR) methods are a new class of methods designed to offer double protection against model misspecification by comprising the outcome model and propensity score model in a manner such that the estimator of τ is consistent if either $\mu(X, A)$ or $\pi(X)$ model is correctly specified (Scharfstein et al., 1999), thus named "doubly robust".

1.4.1 Augmented Inverse Probability Weighting

The most commonly used DR estimator, standard augmented IPW estimators (AIPW), takes the form as

$$\tau^{AIPW} = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{A_i Y_i}{\pi(X_i)} - \frac{(1-A_i)Y_i}{1-\pi(X_i)} - \{A_i - \pi(X_i)\}\left\{\frac{\mu_1(X)}{\pi(X)} + \frac{\mu_0(X)}{1-\pi(X)}\right\}\right] \quad (1.3)$$

where $\mu_1(X) = E(Y|X, A = 1)$ and $\mu_0(X) = E(Y|X, A = 0)$ are outcome regression among treated and untreated, respectively.

It is easy to show that when either $\pi(X) = \pi(X, \beta)$ or $\mu_a(X) = m(X, a, \alpha_a); a = 0, 1$ is correctly specified, $\tau^{AIPW} \to_p \tau$.

The standard AIPW is a special case of a class of augmented IPW estimators. As shown in (Robins et al., 1994), when model of $\pi(X)$ is correct, with no additional

assumptions on the distribution of the data, all consistent and asymptotically normal estimators are asymptotically equivalent to estimators of the general form of AIPW with some function $h(X)$:

$$\frac{1}{n}\sum_{i=1}^{n}\left[\frac{A_iY_i}{\pi(X_i)}-\frac{(1-A_i)Y_i}{1-\pi(X_i)}-\{A_i-\pi(X_i)\}\left\{\frac{h_1(X)}{\pi(X)}+\frac{h_0(X)}{1-\pi(X)}\right\}\right]$$

(Robins et al., 1994) also showed that the standard AIPW estimator with $h_a(X)$ being the true outcome regression model is the optimal DR estimator with asymptotically smallest variance. Hence the standard AIPW is said to be locally efficient: if both $\mu(X, A)$ and $\pi(X)$ are correctly specified, the asymptotic variance of the standard AIPW is the least among all the other AIPW estimator in general form. In other words, the local efficiency implies that if $\pi(X)$ is correctly specified, τ^{AIPW} gains efficiency over all AIPW estimators in the general form under the condition that the model $\mu(X, A)$ is also correctly specified. More discussion regarding asymptotic property of AIPW estimators can be found in many other publications, for example, (Cao et al., 2010; Heejung and Robins, 2005; Kang and Schafer, 2007; Lipsitz et al., 1999; Lunceford and Davidian, 2004; Tan, 2006, 2010; Tsiatis, 2006).

1.4.2 Limitation of DR Methods

The DR estimation through AIPW is "more constructively viewed as incorporating outcome model into the IPW than incorporating propensity model into the outcome regression" (Tan, 2007). When $\mu(X, A)$ is modeled correctly the introduction of inverse weighting may bring in larger variance (in large samples) than the simple outcome regression model, thus variance of τ^{AIPW} is no less than τ^{OR}. On

the other hand, when $\pi(X)$ is modeled correctly, it is anticipated that introduction of outcome model helps to improve the efficiency of basic IPW estimator made only through propensity score model with possible extreme propensity scores. As can be seen in τ^{AIPW}, inverse weighting is taken with respect to $A\left(Y-\mu_1(X)\right)$ or $(1-A)\left(Y-\mu_0(X)\right)$ instead of AY or $(1-A)Y$ in simple IPW. So when the outcome model are correctly specified or slightly misspecified, it is expected to see less instability issue when extreme propensity scores occur. However, In the situation that outcome may be poorly modeled due to uncertainty about distribution of the data, the AIPW estimator is found to even lose substantial efficiency compared to the IPW estimator (Ibrahim et al., 2005; Qin et al., 2009; Rubin and Van der Lann, 2008).

There has also been much efforts in developing improved AIPW estimator and other type of DR estimators. For example, the method of applying spline function to propensity scores in the outcome regression (Little and An, 2004), regression estimation with IPW coefficients (Kang and Schafer, 2007), "tilde" regression estimator (Tan, 010b), targeted maximum likelihood estimator (van der Laan and Gruber, 2010). Nevertheless, one common feature of all DR estimators is that DR methods require at least one of the two models, either outcome or propensity, is correctly specified to ensure double robustness and consistency. When one of the working model is misspecified, finite-sample bias can be amplified as demonstrated by (Carpenter et al., 2006; Kang and Schafer, 2007; Vansteelandt et al., 2012) in their numerical studies. Furthermore, two "bad" models are no better than one "bad" model. Severely biased estimator could result from the scenario when both model are wrong as illustrated in Kang and Schafer 's famous simulation study (Kang and Schafer, 2007).

1.5 Some Existing Nonparametric Methods in Causal Inference

To relax the parametric assumptions, some nonparametric approaches proposed in the estimation of either modeling outcome or propensity score or both have been discussed in causal inference literature.

(Hahn, 1998) established the fundamental result for semiparametric estimation of ATE: under certain regularity conditions,the asymptotic variance bound for τ is given by

$$E\left(\frac{\sigma_1^2(X)}{\pi(X)} + \frac{\sigma_0^2(X)}{1-\pi(X)} + (\tau(X)-\tau)^2\right) \tag{1.4}$$

where $\sigma_a^2(X) = var(Y^{(a)}|X); a = 0, 1.$

In this paper, Hahn also constructed an estimator of τ which can achieve this asymptotic variance bound by noparametrically estimating $E(AY|X = x); E((1 - A)Y|X = x); \pi(x)$ and compute ATE as

$$\hat{\tau}^{hahn} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\hat{E}(A_iY_i|X_i)}{\hat{\pi}(X_i)} - \frac{\hat{E}((1-A_i)Y_i|X_i)}{1-\hat{\pi}(X_i)}\right) \tag{1.5}$$

However, no numerical experiment was presented in the Hahn's paper for the evaluation of $\hat{\tau}^{hahn}$'s performance. Later, (Hirano et al., 2003) argued that the goal of achieving the theoretical asymptotic variance bound can be reached through inverse weighting of nonparametric estimator of $\pi(x)$ when the estimator of $\pi(x)$ is sufficiently flexible. They proposed to estimate $\pi(x)$ as a series logit estimator. Other similar work of applying nonparametric technique in propensity modeling includes (Liang et al., 2004) who considered nonparametric estimation of the component of a partial linear model with missing covariates using IPW , (Wang et al., 2010) who proposed a

class of AIPW kernel estimating equations for nonparametric outcome regression and non-parametric estimation of $\pi(X)$. There are also much research interest lying in using machine learning techniques such as classification, regression tree, random forest and Bayesian methods from the direct outcome modeling perspective. For example, CART for honest inference for treatment effect (Athey and Imbens, 2016), Bayesian tree methods for regression surface modeling (Hill, 2011). These methods often find applications in estimating conditional treatment effects or subgroup classification to deal with heterogeneity in causal inference.

1.6 Thesis Outline

The use of propensity score to estimate ATE has been widely adopted in existing literature of causal inference method. The IPW-type of methods requires the correct specification of the propensity score model to yield an asymptotically unbiased estimator of ATE. Although some nonparametric techniques have been considered to take into account the fact that the true association of covariates and outcome, covariates and treatment assignment are hard to be correctly specified through parametric modeling, the potential numerical instability issue originated from inverse weighting remains unaddressable. Therefore, we propose to develop a new method to address these issues. Similar to most causal inference methods, we develop our method on the foundation of the potential outcome framework and incorporate both outcome generating and treatment selection mechanism into the estimation process. We consider modeling the outcome and treatment selection nonparametrically to obtain the marginal mean score and propensity score, respectively. These two summary scores are combined in such a way that the causal treatment effects can be solved through

ordinary least squares method. Particularly, we make use of the estimated propensity scores to derive pseudo "covariate" in the least squared estimation which is totally different from inverse weighting approach to avoid the potential numerical issue of the IPW-type of methods. The remainder of this thesis is organized as follows.

In Chapter 2, we introduce the method of sieve nonparametric estimation as it is heavily used in the development of our proposed method. Starting with the concept of sieves, we discuss the general approach of sieve estimation followed by description of splines, a popular class of sieves and a commonly used technique in nonparametric regression.

Chapter 3 begins with the motivation behind our ordinary least squared based method and derives the core linear equation based on the standard causal assumptions for potential outcome framework. We then discuss in the rest of this chapter the two stage estimation procedure leading to a model-free estimator of τ in a relative simple scenario when the treatment effect is assumed to be homogeneous. We present two Monte Carlo studies to demonstrate the performance of our proposed estimator and its comparisons with other traditional IPW-type methods. To illustrate how the proposed OLS-based causal inference method is applied to real world data analysis, we introduce the JIA study and present the analysis results for evaluating the effect of early combination of biologic & non-biologic disease-modifying antirheumatic drugs (DMARD) compared to the non-biologic DMARD alone among children with newly onset of juvenile idiopathic arthritis disease under homogeneous treatment effect assumption.

We discuss in chapter 4 the way of extending our proposed method to more general setting when there exists heterogeneity in treatment effects. To maintain

the "model free" feature of our estimator, we describe the way of using the method of sieves to add in one more stage of nonparametric estimation of treatment effect function into the two-stage estimation procedure. The covariate specific treatment effects can be estimated through the sieve estimation and offers a way of examining the treatment effect heterogeneity. The model-free estimator of ATE is then computed as the empirical mean of the estimated covariate specific effects. A Monte Carlo simulation study is conducted to assess the performance of our proposed estimator of ATE and its comparison with other IPW-type methods. We apply this extended approach to the JIA study for the exploration of treatment effect heterogeneity.

In Chapter 5, we focus on the discussion of extending our proposed method furthermore to meet the challenge of estimation when covariates space is high dimensional. By incorporating the commonly used machine learning methods into the first stage for the estimation of the two summary scores, we demonstrate that our method is capable of making causal inference in applications with large set of covariates. We describe in great details how the machine learning methods are applied to obtain a feasible estimation of treatment effects in the JIA study. The thesis concludes with the conclusions and general discussion in Chapter 6 with outlines for the future directions of our research.

CHAPTER 2

Sieve Estimation and Splines

2.1 Nonparametric Regression

In Chapter 1, we used $m(X, A; \alpha)$ to denote a working potential outcome generating model in the outcome regression method and $\pi(X; \beta)$ to specify a working treatment selection model in the propensity score based methods. By writing in these forms, it is assumed the statistical model $m(X, A; \alpha)$ is the mean outcome Y for given A, X for observed data and is determined by a finite-dimensional real-valued parameter $\alpha \in R^p$; the model $\pi(X; \beta)$ is determined by a finite-dimensional of real-valued parameter $\beta \in R^q$ that characterizes the probability $P(A|X)$. For example, if X is a scalar random variable and no interaction of X and A is assumed to present, then the linear mean model is $E(Y|X, A) = \alpha_0 + \alpha_1 X + \alpha_2 A$ of which $\alpha = (\alpha_0, \alpha_1, \alpha_2) \in \mathbb{R}^3$ is a vector of 3 parameters that can be estimated through ordinary least squares (OLS). Similarly, the simplest logistic regression model for the propensity scores is $P(A = 1|X) = \frac{exp(\beta_0 + \beta_1 X)}{1 + exp(\beta_0 + \beta_1 X)}$ from which the 2 dimensional parameter $\beta = (\beta_0, \beta_1) \in \mathbb{R}^2$ can be estimated through maximum likelihood estimation (MLE). In both cases, restrictions (parametric assumptions) were explicitly imposed to the form of regression functions. Thus they are referred to as parametric models.

In reality, the imposed functional form restrictions may not reflect the true association among the variables in the model. Especially, in many biomedical and epidemiology studies, the outcome or the treatment assignment could be related to

the measured covariates in a complicated functional form so their true association is hard to be correctly specified through parametric modeling. Since valid inference for the causal treatment effects depends on either a correct model of outcome or a correct model of propensity score, we may resort to the nonparametric modeling techniques and try to avoid the unrealistic parametric assumptions.

Without loss of generality, we consider the following regression model

$$Y_i = f(Z_i) + u_i; i = 1, \cdots, n \tag{2.1}$$

where $Y_i \in \mathbb{R}$ is a scalar response variable, Z_i is a $p \times 1$ random vector, $f(.)$ is an unknown smooth function and u_i is the error term satisfying $E(u_i) = 0$ and $var(u_i) < \infty$. $\mathbb{F}$ denotes the class of functions that the unknown function $f(.)$ belongs to. When $\mathbb{F}$ is restrictive and the functional form of f is known up to some finite dimensional parameters, e.g. the simplest linear regression model described above, estimation of f reduces to the problem of parametric estimation. When the functional form of f is completely arbitrary, however, f cannot be summarized by a finite set of parameters. We try to use OLS to solve f by minimizing

$$\hat{f}(z) = \arg\min_{f \in \mathbb{F}} \frac{1}{n} \sum_{i=1}^{n} [Y_i - f(Z_i)]^2, \tag{2.2}$$

Different than minimizing over a finite number of parameters in the parametric scenario, the minimization problem in the nonparametric context should be carried out over a set of functions. With infinite-dimensional parameters to be estimated, the estimation of $\hat{f}(z)$ based on finite samples could be very difficult. Even if one may

handle the computation difficulty, the resulting estimator may have undesirable large sample properties, such as inconsistency or converging to the true regression function at very low rate (Chen, 2007). Moreover, since the minimum of $\sum_{i=1}^{n} [Y_i - f(Z_i)]^2$ is zero, to another extreme, in (2.2) if no restrictions is put on $\mathbb{F}$, the minimum 0 can be achieved by any function interpolating the observed data, which usually does not converge to f in any meaningful sense.

Many nonparametric techniques have been developed to address the question of estimating f. These techniques can be broadly classified as two types: Kernel and sieve estimation.

2.1.1 Kernel Method

The method of kernel smoothing provides a way to locally approximate f. Since this method is not our focus of study, we briefly introduce its idea and some popular estimators. This type of methods is said to be a local approximation because they estimate the regression function at a particular point by "locally" fitting a pth degree polynomial to the observed data via weighted least squares. The idea is quite intuitive: for any given point z, $f(z)$ is estimated by a local average of y_i associated with z_i's near the point z within a pre-defined bandwidth $|z_i - z| \leq b$. The bandwidth b controls the smoothingness of estimated function. The Nadaraya-Watson (NW) estimator (Nadaraya, 1964; Watson, 1964) is such a kernel estimator with degree $p = 0$, i.e., local constant kernel estimator.

$$\hat{f}_{NW}(z) = \frac{\sum_{i=1}^{n} y_i K(\frac{z_i - z}{b})}{\sum_{i=1}^{n} K(\frac{z_i - z}{b})} = \sum_{i=1}^{n} \frac{K(\frac{z_i - z}{b})}{\sum_{i=1}^{n} K(\frac{z_i - z}{b})} y_i \tag{2.3}$$

where $K(z)$ is a kernel function $K : \mathbb{R} \to \mathbb{R}$ satisfying $\int K(z)\,\mathrm{d}z = 1; \int zK(z)\,\mathrm{d}z = 0; 0 < \int z^2K(z)\,\mathrm{d}z < \infty$. The weight function is $w_i(z) = \frac{K(\frac{z_i - z}{b})}{\sum_{i=1}^{n} K(\frac{z_i - z}{b})}$ with $\sum_{i=1}^{n} w_i(z) = 1$.

Due to its mathematical simplicity, the NW estimator has found some applications in nonparametric regression. Other frequently used kernel estimators include local linear kernel estimator in the case of $p = 1$ and more generally, local polynomial kernel estimators (Cleveland, 1979; Fan and Gijbels, 1992; Muller and Stadtmuller, 1987; Stone, 1977). For example, in Robinsons double residual method for addressing the partially linear regression problem (Robinson, 1988), he considered estimating the mean function and propensity function using the kernel smoothing method. In employing nonparametric methods for missing data problem using IPW or AIPW type of estimator, many researchers have also proposed the kernel method (Liang et al., 2004; Wang et al., 2010).

2.1.2 Sieve Method

The method of sieves was originally proposed by (Grenander, 1981). In contrast to the Kernel method, it is a nonparametric technique of global approximation. Specifically, to resolve the problem of estimating functions over an infinite-dimensional functional space using finite samples, Grenander suggested to perform the optimization (minimization of the sum of square errors in OLS, maximization of the likelihood in MLE) within a subset of the parameter space of the unknown function, and then allow this subset to "grow" along with the sample size. This sequence of subsets from which the estimator is drawn is called a "sieve" and thus the resulting estimation procedure was named the "method of sieves".

The critical part in the sieve method is using the approximating subspaces, the sieve spaces, to reduce the dimension of functional space in which a functional estimator is sought. For the regression model of (2.1) , in practice, there is no way that we can search in the full space $\mathbb{F}$ for $f(z)$. So we turn to a simpler space of lower dimension, say, $\mathbb{F}_{q_n}$, defined as

$$\mathbb{F}_{q_n} = \left\{ f : f(z) = \sum_{j=1}^{q_n} \alpha_j \phi_j(z) \right\} \tag{2.4}$$

where ϕ_j is a sequence of approximating terms known as basis functions; q_n is the dimensionality of the sieve space indicating the number of basis functions lying in the sieve space given fixed sample size n. If q_n is allowed to grow to infinity when the sample size n goes to infinity and not grow too fast, the sieve space $\mathbb{F}_{q_n}$ is gradually expanding along with n in the sense that $\mathbb{F}_{q_1} \subset \mathbb{F}_{q_2} \subset \cdots \subset \mathbb{F}$. As a result of growing, the sieve space becomes denser and denser in $\mathbb{F}$ such that the best approximation to any function $f^* \in \mathbb{F}$ inside $\mathbb{F}_{q_n}$ must get arbitrarily close to f^* as $n \to \infty$ (Chen, 2007). In other words, functions inside the sieve space can be used to approximate various smooth functions at certain convergence rate depending on the complexity of the approximation. The global nature of sieve method is due to the fact that it estimates the function of interest over its functional space by restricting to certain type of sieves in a single step. Thus, comparing to the kernel method, the sieve method has the advantage of being computational easier with many known basis functions ready to be used.

2.2 Sieve Estimation

The sieve expansion of $\sum_{j=1}^{q_n} \alpha_j \phi_j(z)$ in (2.4) reduces the dimensionality of the optimization problem of (2.2) significantly since the number of basis functions q_n for $\phi_j(z); j = 1, \cdots, q_n$ required to approximate $f(z)$ grow much slower as sample size increases. The basis functions $\phi_j(z)$'s are typically non-linear functions of z but the sieve expansion is a linear combination of these bases. For such a linear span we can define the matrix of regressors with dimension $n \times q_n$ as

$$\Phi = \begin{bmatrix} \phi_1(z_1) & \phi_2(z_1) & \dots & \phi_{q_n}(z_1) \\ \phi_1(z_2) & \phi_2(z_2) & \dots & \phi_{q_n}(z_2) \\ \dots & \dots & \dots & \dots \\ \phi_1(z_n) & \phi_2(z_n) & \dots & \phi_{q_n}(z_n) \end{bmatrix}$$

Provided that the sieve is appropriate and the growth of sieve is sufficiently slow, $\sum_{j=1}^{q_n} \alpha_j \phi_j(z)$ is expected to approximate the true regression function $f(z)$ reasonable well such that $f(z) - \sum_{j=1}^{q_n} \alpha_j \phi_j(z) \to 0$ for all z as $n \to 0$. Then the nonparametric regression by least squares in (2.2) is asymptotically equivalent to the problem of minimizing

$$\arg\min_{\alpha} \frac{1}{n} \sum_{i=1}^{n} \left[Y_i - \sum_{j=1}^{q_n} \alpha_j \phi_j(z) \right]^2, \tag{2.5}$$

over the sieve $\mathbb{F}_{q_n}$. Within such sieve space, the estimation of α becomes a standard parametric regression problem and $\hat{\alpha}$ can be easily computed via OLS as

$$\hat{\alpha} = (\alpha_1, \alpha_2, \cdots, \alpha_{q_n})^T = (\Phi^T \Phi)^{-} \Phi^T Y \tag{2.6}$$

where $Y = (Y_1, Y_2, \cdots, Y_n)^T$ and $(D)^-$ denotes the generalized inverse of (D) satisfying $DD^-D = D$ and $D^-DD^- = D^-$ (Penrose, 1955). Hence the estimator of $f(z)$ is given by

$$\hat{f}(z) = \sum_{j=1}^{q_n} \hat{\alpha}_j \phi_j(z) \tag{2.7}$$

The idea of sieve least square estimation discussed above can be easily generalized to sieve maximum likelihood estimation, sieve generalized linear model etc. For example, if our goal is to estimate the propensity score function $\pi(x)$ using this sieve method, similarly we can approximate the unknown function $\pi(x)$ by a linear span of some suitable p_n dimensional basis functions $\gamma_j(x); j = 1, \cdots, p_n$ to form a sieve log-likelihood as

$$l(X;\beta) = \sum_{i=1}^{n} \left\{ a_i \sum_{j=1}^{p_n} \beta_j \gamma_j(x) + \sum_{i=1}^{n} log(1 + exp(\sum_{j=1}^{p_n} \beta_j \gamma_j(x))) \right\}, \tag{2.8}$$

where a_i indicates if the ith individual was assigned to treatment. Then by maximizing the sieve log-likelihood with respect to the unknown coefficients $\beta_j; j = 1, \cdots, p_n$ we obtain the sieve MLE of $\beta_j, \hat{\beta}_j; j = 1, \cdots, p_n$ that results in the sieve NPMLE of $\pi(X)$, $\hat{\pi}(x) = \sum_{j=1}^{p_n} \hat{\beta}_j \gamma_j(x)$.

In addition to its advantage of being generalizable to deal with many estimation problems, the sieve method is conceptually simple and can be easily implemented. Most often, the sieve estimation can be carried out in three steps: (i) choosing the dimension of sieve space; (ii) choosing appropriate basis functions; (iii) estimating the coefficients associated with the basis functions. In many cases, we just need to decide the number of those well-studied basis functions, and then conduct the standard analysis by treating the model as if it were a fully parametric model. As for the large

sample properties of sieve estimator, Geman and Hwang (1982) initially established the general consistency of sieve MLE. (Chen, 2007) discussed the regularity conditions for sieve MLE. There has been considerable work on asymptotic normality of sieve estimator (Andrews, 1991; Bierens, 2012; Chen, 2007; Chen and Pouzo, 2015; Newey, 1997; Stone, 1982, 1985, 1986). Discussion of asymptotic properties of sieve estimators in general is beyond the scope of this thesis.

2.3 Splines

2.3.1 Sieves for Sieve Estimation

Sieve method provides common strategy for nonparametric estimation. However, different sieves lead to different sieve estimators. A variety of sieve estimators are presented as demonstrating examples, in e.g. (Chen, 2007; Geman and Hwang, 1982; Grenander, 1981). As discussed in (Chen, 2007), in choosing appropriate sieves for various applications the most popular class of functions considered is the Hölder smoothness class or p-smooth class of functions whose rigorous mathematical definition was given in (Chen, 2007)(p5569-5570). To put it in simple language, Hölder class is a class of functions that have a well-behaved remainder term in Taylor expansion and thus can be well approximated by a variety linear sieves including power series, Fourier series, splines and wavelets, etc. In causal inference literatures, (Hahn, 1998) proposed power series estimation method for computing $\hat{E}(AY|X=x), \hat{E}((1-A)Y|X=x)$ and $\hat{\pi}(x)$. He suggested to construct a power series as $p^K = [p^K(X_1), \cdots, p^K(X_n)]^T$ with $p^K(x) = (p_{1K}(x), \cdots, p_{KK}(x))^T$. The element $p_{KK}(x)$ should satisfy the condition 3 of Theorem 6 presented in his paper,

which as the author pointed out, could be extremely difficult in practice. In the discussion of obtaining a non-parametrically estimated $\hat{\pi}(x)$ to feed into the IPW estimator of ATE, (Hirano et al., 2003) adopted the sieve approach and proposed a series logit estimator of $\hat{\pi}(x)$. He described the use of triangular power series in the approximation of log odds ratio to obtain such a sieve estimator. These proposed sieve estimators were mainly presented and elaborated for the purpose of theoretical justification while the implementation of them was not mentioned and sounds not so straightforward.

In this thesis, we specifically consider using spline functions in the sieve non-parametric estimation for causal inference problem. Spline has been widely recognized in the statistical and mathematical literature as a useful tool of nonparametric estimation (Stone, 1985, 1986). Its nature of being "smoothly joined" piecewise polynomial functions makes it perform very well in approximating quite arbitrary functions. Polynomial power series is capable of approximating arbitrary functions well with possibly very high order polynomial, but the design matrix formed by very high order polynomials is usually ill-conditioned and is likely to cause computation difficulty in the estimation. Spline, especially B-splines, makes the computation more stable, faster and thus leads to a more stable estimator. Large sample properties of spline-based sieve estimator in nonparametric and semi-parametric settings have also been well established in literatures e.g. (Huang, 2003; Zhang et al., 2010; Zhou et al., 1998). Moreover, splines are easy to implement with many software packages such as R developed for accommodating this convenience.

2.3.2 Spline Function

Our discussion is still within the context of nonparametric estimation $f(z)$ in model (2.1). Formally, let $\triangledown$ be a partition of Z into disjoint sets. In the case of Z being a scaler random variable, partition is made by certain number of data points within the range of Z and elements in $\triangledown$ are intervals. More generally, in tensor product splines or other multivariate splines, element of $\triangledown$ can be two or higher dimensional triangle or rectangle. For convenience of discussion, we consider only the univariate splines here.

A function $g : \mathbb{R} \to \mathbb{R}$ is said to be a r-th degree, equivalently, $(r+1)$-th order(r is an integer and $r \geq 0$) spline with a nondecreasing sequence of knots at $\kappa_1 < \cdots < \kappa_m$ if

- g is a polynomial of degree r on all the intervals $[-\infty, \kappa_1], \cdots, [\kappa_m, \infty]$.
- For $j = 0, 1, \cdots, r-1$, the jth derivate of g is continuous at each knot $\kappa_1, \cdots, \kappa_m$.

According to its definition, a spline function g is a set of piecewise polynomials joint smoothly at the knots and constitutes a linear space of dimension $m + r + 1$. When no knot is used, splines of degree r are simply polynomials of degree r. When $r = 0$, g is simply a step function with jump at the knots. $r = 3$ means g is a piecewise cubic curve that are continuous, and have continuous first as well as second derivatives at the knots, which makes it smooth in appearance as the slope and the rate of change in the slope are continuous at the knots. Visually, a cubic spline is a smooth curve, and it is the most commonly used spline when a smooth fit is desired.

With given degree r, g can be simply constructed by truncated power basis functions. The truncated polynomial of degree r associated with a knot $\kappa_j; j = 1, \cdots, m$ is usually denoted as the function $(z - \kappa_j)_+^r$ which takes values

$$(z - \kappa_j)_+^r = \begin{cases} 0 & \text{if } z < \kappa_j \\ (z - \kappa_j)^r & \text{if } z \geq \kappa_j \end{cases}$$

Thus a truncated power based spline function is

$$g(z) = \sum_{k=0}^{r} \alpha_k z^k + \sum_{j=1}^{m} b_j (z - \kappa_j)_+^r$$

where $\alpha_k, b_j \in \mathbb{R}$. The design matrix is then the $n \times (1 + r + m)$ matrix with entries:

$$\begin{bmatrix} 1 & z_1 & z_1^2 & \dots & z_1^r & (z_1 - \kappa_1)_+^r & (z_1 - \kappa_2)_+^r & \dots & (z_1 - \kappa_m)_+^r \\ 1 & z_2 & z_2^2 & \dots & z_2^r & (z_2 - \kappa_1)_+^r & (z_2 - \kappa_2)_+^r & \dots & (z_2 - \kappa_m)_+^r \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 1 & z_n & z_n^2 & \dots & z_n^r & (z_n - \kappa_1)_+^r & (z_n - \kappa_2)_+^r & \dots & (z_n - \kappa_m)_+^r \end{bmatrix}$$

Truncated power function basis has the advantage of being conceptually simple and the parameters in a model to these corresponding basis functions have clear interpretation. However, since the values in the above design matrix may get very large when r is large and the columns of the design matrix may be highly correlated, computation of spline coefficients may be numerically unstable. Also, when z_i's span a wide range and there are a large number of knots used to fit the regression spline,

truncated power bases may cause numerical issue of inefficiency and instability. In practice, the B-spline bases are most commonly used in applications.

2.3.3 B-Splines

Comprehensive reviews regarding the definition and properties of B-splines can be found in (De Boor, 2001; Dierckx, 1993; Eilers and Marx, 1996; Ruppert et al., 2003). We give a brief summary of this type of spline as follows.

Given degree r and a non-decreasing sequence of knots $\kappa_0 \leq \kappa_1 \leq \cdots \kappa_m$, a B-spline curve of $S(z)$ is defined as

$$S(z) = \sum_{i=0}^{n} \alpha_i B_i^r(z); z \in [\kappa_r, \kappa_{n+1}) \tag{2.9}$$

where $n = m - r - 1$ and B_i^r are recursively defined basis functions in the form of

$$B_i^0(z) = \begin{cases} 1 & \text{if } \kappa_i \leq z < \kappa_{i+1} \\ 0 & \text{if } \textit{otherwise} \end{cases}$$

$$B_i^j(z) = \frac{z - \kappa_i}{\kappa_{i+j} - \kappa_i} B_{i,j-1}(z) + \frac{\kappa_{i+j+1} - z}{\kappa_{i+j+1} - \kappa_{i+1}} B_{i+1,j-1}(z)$$

The knots interval within $[\kappa_r, \kappa_{n+1})$ are referred to as interior knots and knots $\kappa_0, \cdots, \kappa_{r-1}$ and $\kappa_{n+2}, \cdots, \kappa_{n+r+1}$ at the two ends are called boundary knots typically achieved by repeating the first used knot (κ_r) and last used knot (κ_{n+1}) r times, respectively. The great advantage of B-spline basis functions is that each of them has only small support region that makes computation easier and faster than the truncated power base.

One can easily see from the definition above that B_i^j is non-zero over an interval that spans at most $r+1$ knots, which means that the design matrix is made of columns each of whose rows contain at most $r+1$ adjacent nonzero entries. This explain the fact that B-splines are usually numerical stable and effective and therefore preferable in applications.

Choosing knots number and locations is critical in regression spline estimation. A popular choice for the knots location in regression spline is to place knots at equally spaced sample quantiles of Z with the two boundary knots set to be the minimum and maximum of Z in the sample, respectively. Optimal knots number can be determined by leave-subjects-out K-fold cross-validation (Huang, 2003) or Akaike information criterion (AIC) (Xue and Liang, 2009). Alternatively, a rule of thumb based on the convergence rate of regression spline estimators (Zhang et al., 2010; Zhou et al., 1998) can be adopted to determine the right number of knots. In applications, we may explore the optimal number of knots with the guideline $\geq Cn^{1/v}$ where n is the sample size , v takes certain value between 2 and 5 and C is a constant for tuning.

CHAPTER 3

An OLS-based Method for Casual Inference - Homoscedastic Treatment Effect Case

3.1 Setup

The major goal of many observational studies is to estimate the average effect of a binary treatment on a scalar outcome. Unless otherwise noted, in the remainder of this thesis we consider the following basic setting for a causal inference question. Some notation in Chapter 1 are reiterated here. We have a random sample of size n from an observational study, from which estimating the average causal effect at population level is of interest. For unit $i(i = 1, \cdots, n)$ in this random sample, let A_i be an indicator of whether the active treatment (treatment for the evaluation of average causal effect) is received, with $A_i = 1$ if unit i is in the treatment arm, and $A_i = 0$ if unit i is in the control arm. The outcome of interest Y is a continuous variable. For each unit i, a vector of pre-treatment covariates denoted as X is measured. The observed n independent and identically distributed copies of data are made of triples $D_i = (A_i, Y_i, X_i)$.

Using potential outcome notation, $Y_i^{(1)}$ and $Y_i^{(0)}$ are the outcome for unit i under active treatment and control, respectively. $(Y_i^{(1)}, Y_i^{(0)})$ does not depend on A_j or $(Y_j^{(1)}, Y_j^{(0)})$ if $i \neq j$. The individual treatment effect for unit i is $\tau_i = Y_i^{(1)} - Y_i^{(0)}$. The covariate-specific ATE for a subpopulation with covariate X being x is $\tau(x) = E(Y^{(1)} - Y^{(0)} | X = x)$. The population average treatment effect is $\tau = E(Y^{(1)} - Y^{(0)}) =$

$E_X\{\tau(x)\}$, where $E_X(.)$ is the expectation taking over the population with respect to X. Following most of the causal inference literatures, we use the standard assumptions for the identification of $\tau(x)$ and τ.

- Consistency

$$Y = AY^{(1)} + (1 - A)Y^{(0)} \tag{3.1}$$

- Positivity

$$0 < P(A = 1|X = x) < 1; \forall x \in X \tag{3.2}$$

- Ignorability (weak version)

$$E(Y^{(a)}|A, X) = E(Y^{(a)}|X); a = 0, 1 \tag{3.3}$$

3.2 Motivation

The observed outcome Y is associated with both A and X. The expectation of observed Y conditional on both A and X is measurable. Since X acts as confounder between A and Y, we can't directly use $E(Y|A, X)$ to compute $E(Y^{(a)}|X)$ or $E(Y^{(a)})$. But notice that

$$E(Y|A = a, X) = aE(Y|A = 1, X) + (1 - a)E(Y|A = 0, X) \tag{3.4}$$

With (3.3) we can further write (3.4) as

$$\begin{aligned} E(Y|A=a,X) &= aE(Y|A=1,X)+(1-a)E(Y|A=0,X) \\ &= E(Y^{(0)}|X)+aE(Y^{(1)}-Y^{(0)}|X) \\ &= E(Y^{(0)}|X)+a\tau(X) \end{aligned} \tag{3.5}$$

On the other hand, based on (3.1) and (3.3), we have

$$\begin{aligned} E(Y|X) &= E(Y|A=1,X)P(A=1|X)+E(Y|A=0,X)P(A=0|X) \\ &= E(Y^{(1)}|X)E(A|X)+E(Y^{(0)}|X)(1-E(A|X)) \\ &= E(Y^{(0)}|X)+E(A|X)E(Y^{(1)}-Y^{(0)}|X) \\ &= E(Y^{(0)}|X)+\pi(X)\tau(X) \end{aligned} \tag{3.6}$$

Where $\pi(X)$ is the propensity score function which is bounded away from 0 and 1 guaranteed by (3.2).

Subtracting (3.6) from (3.5) results in

$$E(Y|A=a,X)=m(X)+(a-\pi(X))\tau(X) \tag{3.7}$$

where $m(X)=E(Y|X)$, the mean score function.

(3.7) serves as our core structural equation for the estimation of τ. When the treatment effect is homoscedastic in the study population, we have (i) $\tau_i=\tau+\epsilon_i$; (ii) ϵ_i is independent of X_i; (iii) $E(\epsilon_i)=0$ so the study population is homogeneous in terms of treatment effect. The assumption of homogeneous treatment effect is indeed

frequently adopted in causal inference literature to simplify the problem of estimating ATE. A typical example is presented in (Robins et al., 1992) with the associated method of G-estimation. Robins's G-estimator was discussed in the context of an observation study with the goal of evaluating the effect of being current cigarette smoker on the level of forced expiration volume in one second (FEV1) in a cohort of 2713 adult white male. It was assumed in the development of his proposed estimator that there is no interaction between the exposure, current smoking, and other confounders such as past smoking history, past respiratory symptoms, age, coexistent heart disease etc. No existence of interaction between treatment and confounder, in other words, means that the confounder X only plays role of being an effect mediator but not an effect modifier. Under this scenario, $\tau(x) = \tau$ thus (3.7) can be further simplified to

$$E(Y|A = a, X) = m(X) + (a - \pi(X))\tau \tag{3.8}$$

(3.8) is built on two summary scores: $m(X)$ and $\pi(X)$. It clearly displays a linear structure and motivates a simple OLS method to estimate τ by solving

$$\arg\min_{\tau} \sum_{i=1}^{n} \{(Y_i - m(X_i)) - (A_i - \pi(X_i))\tau\}^2. \tag{3.9}$$

If both the mean and propensity scores are known, it is straightforward that τ can be obtained as

$$\hat{\tau} = \frac{\sum_{i=1}^{n} \{(Y_i - m(X_i))(A_i - \pi(X_i))\}}{\sum_{i=1}^{n} (A_i - \pi(X_i))^2} \tag{3.10}$$

To examine the property of $\hat{\tau}$ assuming $m(x_i)$ and $\pi(x_i)$ are known, let $\omega_i = Y_i - m(X_i) - (A_i - \pi(X_i))\tau = Y_i - E(Y_i|A_i, X_i)$, then it immediately follows that $E(\omega_i|X_i, A_i) = 0$ and $var(\omega_i|X_i, A_i) = var(Y|X, A)$ since data are i.i.d. $\hat{\tau}$ can be rewritten as $\hat{\tau} = \tau + \frac{\sum_{i=1}^{n} \omega_i(A_i - \pi(X_i))}{\sum_{i=1}^{n} (A_i - \pi(X_i))^2}$. Further,

$$\sqrt{n}(\hat{\tau} - \tau) = \sqrt{n} \frac{\frac{1}{n} \sum_{i=1}^{n} \omega_i(A_i - \pi(X_i))}{\frac{1}{n} \sum_{i=1}^{n} (A_i - \pi(X_i))^2}$$

Since by law of large numbers,

$$\frac{1}{n} \sum_{i=1}^{n} (A_i - \pi(X_i))^2 \rightarrow_p E\left((A_i - \pi(X_i))^2\right) = \pi(X)(1 - \pi(X))$$

Also,

$$\begin{aligned} E\left(\omega_i(A_i - \pi(X_i))\right) &= E\left(E\left(\omega_i(A_i - \pi(X_i))\right)|A_i, X_i\right) \\ &= E\left((A_i - \pi(X_i))E(\omega_i|A_i, X_i)\right) \\ &= 0 \end{aligned}$$

and

$$\begin{aligned} E\left(\omega_i^2(A_i - \pi(X_i))^2\right) &= E\left((A_i - \pi(X_i))^2 var(\omega_i|A_i, X_i)\right) \\ &= var(\omega_i|X_i, A_i)E((A_i - \pi(X_i))^2) \\ &= var(Y|X, A)\pi(X)(1 - \pi(X)) \end{aligned}$$

By central limit theory (CLT) ,

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}\omega_i(A_i-\pi(X_i))\right)\rightarrow_d N\left(0, var(Y|X,A)\pi(X)(1-\pi(X))\right)$$

Then according to Slutsky's theorem we have

$$\sqrt{n}(\hat{\tau}-\tau)\rightarrow_d N\left(0,\frac{var(Y|X,A)}{\pi(X)(1-\pi(X))}\right)$$

However, $\hat{\tau}$ is not a feasible estimator as the mean and propensity scores are unknown in practice. Therefore the estimation of τ needs to be accomplished in two steps. Once the mean and propensity scores are consistently estimated from the observed data, we may derive a plug-in estimator of τ with $\pi(X)$ and $m(X)$ in (3.10) replaced by their corresponding estimates.

3.3 Two-Stage Estimation Method

Estimation of τ in OLS-based manner as (3.10) requires knowledge of $m(x)$ and $\pi(x)$, we propose the following two-stage estimation procedure and argue that it leads to a model-free estimator of τ.

3.3.1 Stage 1: Nonparametric Estimation of Mean and Propensity Scores

To ensure consistent estimation of the two summary scores, in the first stage, we adopt the regression-spline based nonparametric sieve estimation methods as discussed in Chapter 2 to estimate $m(x)$ and $\pi(x)$, respectively.

- For estimating $m(x)$, we use a regression-spline based nonparametric sieve least-squares estimation method. First, we seek to estimate the mean score in a sieve

space spanned by B-splines, that is

$$m(x) = \sum_{j=1}^{q_n} \alpha_j B_j(x)$$

where $B_j(x)$ is the pre-specified spline basis functions for $j = 1, \cdots, q_n$ and q_n is the number of spline basis functions that increases as sample size increases. The spline coefficients $\alpha = (\alpha_1, \cdots, \alpha_{q_n})$ are then estimated by $\hat{\alpha} = (\hat{\alpha}_1, \cdots, \hat{\alpha}_{q_n})$ via the OLS method

$$\hat{\alpha} = \arg\min_{\alpha} \sum_{i=1}^{n} \left\{ Y_i - \sum_{j=1}^{q_n} \alpha_j B_j(X_i) \right\}^2$$

with design matrix formed by the spline basis functions as

$$B = \begin{bmatrix} B_1(X_1) & B_2(X_1) & \dots & B_{q_n}(X_1) \\ B_1(X_2) & B_2(X_2) & \dots & B_{q_n}(X_2) \\ \dots & \dots & \dots & \dots \\ B_1(X_n) & B_2(X_n) & \dots & B_{q_n}(X_n), \end{bmatrix}$$

the OLS-estimation of α is given by

$$\hat{\alpha} = (\hat{\alpha}_1, \cdots, \hat{\alpha}_{q_n})^T = (B^T B)^{-} B^T Y$$

where $Y = (Y_1, \cdots, Y_n)^T$ and D^- denotes the generalized inverse matrix of D. The sieve estimator of $\hat{m}(x)$ is $\hat{m}(x) = \sum_{j=1}^{q_n} \hat{\alpha}_j B_j(x)$.

- For estimating $\pi(x)$, as the outcome considered is binary data A, we adopted a regression-spline based nonparametric maximum likelihood estimation (NPMLE) method. First, we model the propensity score using the regression splines

$$\pi(x) = \frac{\exp\left(\sum_{j=1}^{q_n} \beta_j B_j(x)\right)}{1 + \exp\left(\sum_{j=1}^{q_n} \beta_j B_j(x)\right)}$$

where $B_j(x)$'s have the same definition as in the estimation of $m(X)$. We then estimated the regression coefficients $\beta = (\beta_1, \cdots, \beta_{q_n})$ by $\hat{\beta} = (\hat{\beta}_1, \cdots, \hat{\beta}_{q_n})$ via the MLE,

$$\hat{\beta} = \arg\max_{\beta} \sum_{i=1}^{n} \left\{ A_i \sum_{j=1}^{q_n} \beta_j B_j(X_i) - \log\left[1 + \exp\left(\sum_{j=1}^{q_n} \beta_j B_j(X_i)\right)\right]\right\}$$

We adopted the Newton-Raphson algorithm to compute $\hat{\beta} = (\hat{\beta}_1, \cdots, \hat{\beta}_{q_n})$ from a sequence of equations

$$\sum_{i=1}^{n} B_j(X_i) \left\{ A_i - \frac{\exp\left(\sum_{j=1}^{q_n} \beta_j B_j(X_i)\right)}{1 + \exp\left(\sum_{j=1}^{q_n} \beta_j B_j(X_i)\right)} \right\} = 0; j = 1, \cdots, q_n$$

$\hat{\pi}(x)$ is thus obtained as $\hat{\pi}(x) = \frac{\exp\left(\sum_{j=1}^{q_n} \hat{\beta}_j B_j(x)\right)}{\left\{1+\exp\left(\sum_{j=1}^{q_n} \hat{\beta}_j B_j(x)\right)\right\}}$.

3.3.2 Stage 2: Plug-in Estimator of $\hat{\tau}$

The idea for the estimation of τ in the second stage is quite simple. The estimated $\hat{m}(X), \hat{\pi}(X)$ from stage 1 were substituted for the corresponding $m(X)$ and $\pi(X)$ in (3.10), respectively, to form a plug-in estimator of $\hat{\tau}$. Same as in (3.10),

this OLS estimator can be written in an explicit form as

$$\hat{\tau}^{mf} = \frac{\sum_{i=1}^{n} \{(Y_i - \hat{m}(X_i))(A_i - \hat{\pi}(X_i))\}}{\sum_{i=1}^{n} (A_i - \hat{\pi}(X_i))^2}. \tag{3.11}$$

There are several key points about our proposed model-free method for estimating the average treatment effect τ:

- In our core linear equation (3.8), the potential outcomes $Y^{(1)}$ and $Y^{(0)}$ are embedded in τ and treated as an integrated whole. As τ is exactly our target estimand of interest, it suggests that there is no need to estimate $E(Y^{(1)})$ and $E(Y^{(0)})$ separately in order to derive an estimator for τ, which distinguishes our method from all other IPW type methods. In the inverse weighting methods, after the propensity scores are obtained, we first need to estimate $E(Y^{(1)})$ and $E(Y^{(0)})$ based on $E\left(\frac{AY}{\pi(X)}\right)$ and $E\left(\frac{(1-A)Y}{1-\pi(X)}\right)$, respectively, then compute τ as the difference between $\hat{E}(Y^{(1)})$ and $\hat{E}(Y^{(0)})$.
- $\hat{\tau}^{mf}$ is not a type of inverse probability weighting estimator, hence it does not suffer the numerical instability issue as frequently seen in IPW estimator (1.2) and AIPW estimator (1.3) due to the estimated propensity score close to 0 or 1 at some covariate values X. It is obvious that in $\hat{\tau}^{mf}$ the estimated propensity scores are summed across all the individuals before taking inverse. This is totally different from the way of taking the inverse of the each estimated individual propensity score in (1.2) or (1.3) . Therefore, it is expected that $\hat{\tau}^{mf}$ won't be inflated by individual propensity scores near 0 or 1. This is the great advantage of $\hat{\tau}^{mf}$ over all other conventional IPW and AIPW estimators.

- It is worth of mentioning that Robinson's semi-parametric estimation approach in partially linear model results in the same estimator of τ. A semi-parametric partially linear model is given by

$$Y_i = g(X_i) + \tau A_i + u_i, i = 1, \cdots, n \tag{3.12}$$

where u_i is the random error term, $g(.)$ is an unknown smooth function and τ can be interpreted as average treatment effect of A on Y. Taking the conditional expectation of both sides of (3.12) given X_i, we have

$$E(Y_i|X_i) = g(X_i) + \tau E(A_i|X_i) \tag{3.13}$$

Subtracting (3.13) from (3.12) yields

$$Y_i - m(X_i) = (A_i - \pi(X_i))\tau + u_i \tag{3.14}$$

Based on (3.14) we can also derive an plug-in estimator of $\hat{\tau}$ with closed form as (3.11). Although estimator from this approach coincides with the our proposed estimator, the two are different in the sense that our estimator is derived without model assumption of (3.12). That is: (3.8) doesn't require (3.12) and is derived entirely based on the standard casual assumptions with additional homogeneity assumption. The development of $\hat{\tau}^{mf}$ involves no other parametric model assumptions either. In the first stage we use sieve estimation techniques to estimate $m(X)$ and $\pi(X)$ both non-parametrically. Thus $\hat{\tau}^{mf}$ is model-free and enjoys robustness against model misspecification. This avoids the issue aris-

ing from the DR type methods which requires at least one model is correctly specified and could be severely biased when both models are wrong.

3.4 Simulation Studies

We conducted two simulation studies to evaluate the performance of our proposed model-free estimator of ATE and to compare it with estimators from other conventional methods including IPW and AIPW at various scenarios.

3.4.1 Simulation Study (1)

3.4.1.1 Design of Study (1)

The first Monte Carlo simulation was designed to mimic a hypothetical cohort study from which we are interested in studying the average casual effect of a dichotomous treatment A on the outcome Y. There were two observed baseline covariates $X = (X_1, X_2)$ served as confounders in the causal relationship of A to Y. We assume that there is no other observed or unobserved confounders. X_1 is a continuous covariate generated from $uniform(-3, 3)$ and X_2 is a binary group indicator generated from $Bernoulli(0 \cdot 5)$ and independent of X_1. For each subject in a random sample, the probability of being assigned to the treatment group $A = 1$ was modelled by the following logistic model

$$\begin{aligned} \pi(X) &= P(A = 1|X_1, X_2) \\ &= \frac{\exp(-1.2 + 0.2X_1^2 + 0.3X_2 + 0.15X_2X_1)}{1 + \exp(-1.2 + 0.2X_1^2 + 0.3X_2 + 0.15X_2X_1)} \end{aligned} \tag{3.15}$$

and the treatment assignment $A = 1$ was generated from $Bernoulli(\pi(X))$. (3.15) gives the true treatment selection mechanism. With this mechanism, the proportion of subjects in the treated group is approximately 40% (on average). Particularly, (3.15) was designed so that the true propensity scores fall closely within the range of $(0.2, 0.8)$ to avoid possible extreme weights in IPW type methods. Given the treatment assignment indicator A and covariates (X_1, X_2), the outcome Y was generated according to the following model

$$Y = -2 - 5X_1X_2 + 0 \cdot 6X_1^2 + 2e^{X_1} + 6A + \epsilon \tag{3.16}$$

where the random error ϵ is independent of $X = (X_1, X_2)$ and is normally distributed with $N(0, 1)$. (3.16) gives the true potential outcome generating model. As we are specifically interested in the homogeneous treatment effect scenario, the true ATE is set to be a constant 6.

(3.15) and (3.16) together result in true mean scores given by

$$m(X) = -2 - X_1X_2 + 0 \cdot 6X_1^2 + 2e^{X_1} + \frac{6\exp(-1 \cdot 2 + 0 \cdot 2X_1^2 + 0 \cdot 3X_2 + 0 \cdot 15X_2X_1)}{1 + \exp(-1 \cdot 2 + 0 \cdot 2X_1^2 + 0 \cdot 3X_2 + 0 \cdot 15X_2X_1)} \tag{3.17}$$

3.4.1.2 Cubic B-Spline Approximation of Mean and Propensity scores

We estimated the mean and propensity scores at the first stage using the cubic B-spline based sieve method. Since X_2 is a binary variable, cubic B spline was only applied to X_1. For a study sample with n observations of X_1 contained in a closed interval $[a, b]$, we divided this interval into $q_n - 3$ subintervals made by a sequence of

spline knots given by

$$a = \kappa_1 = \kappa_2 = \xi_3 = \kappa_4 < \kappa_5 < \cdots < \kappa_{q_n} < \kappa_{q_n+1} = \kappa_{q_n+2} = \kappa_{q_n+3} = \kappa_{q_n+4} = b,$$

According to the discussion of knots selection for regression splines in Chapter 2, we set the number of knots q_n to be $\frac{[n^{1/3}]}{2}$, the largest integer less than $\frac{n^{1/3}}{2}$ and the knots were placed at the $q_n - 3$ quantiles of X_1. The mean scores and propensity scores were modelled by

$$m(X) = \sum_{j=1}^{q_n} \alpha_j^{(1)} B_j(X_1) + \sum_{j=1}^{q_n} \alpha_j^{(2)} B_j(X_1) X_2$$

and

$$\log\left\{\frac{\pi(X)}{1-\pi(X)}\right\} = \sum_{j=1}^{q_n} \beta_j^{(1)} B_j(X_1) + \sum_{j=1}^{q_n} \beta_j^{(2)} B_j(X_1) X_2,$$

respectively, where $B_j(X)$ is the normalized B-spline basis functions at the knots κ_j for $j = 1, \cdots, q_n$.

To ensure that the first stage yields consistent nonparametric estimation of mean scores and propensity scores using the cubic B splines, we examined the agreement of true propensity score curve, true mean score curve with the estimated propensity score function, estimated mean function, respectively, and showed them in Figure 3.1. A sequence of 200 data points in the interval [-3,3] were created as D_1. A binary indicator D_2 was created in two scenarios (i) all 1; (ii) all 0 with size 200. For each pair of $(D_{1(i)}, D_{2(i)} = 1)$ and $(D_{1(i)}, D_{2(i)} = 0)$, we evaluated their true propensity score $\pi(D_{1(i)}, D_{2(i)})$ and true mean score $m(D_{1(i)}, D_{2(i)})$ based on (3.15) and (3.17), respectively. We then generated 1000 datasets each with size 200 according to the

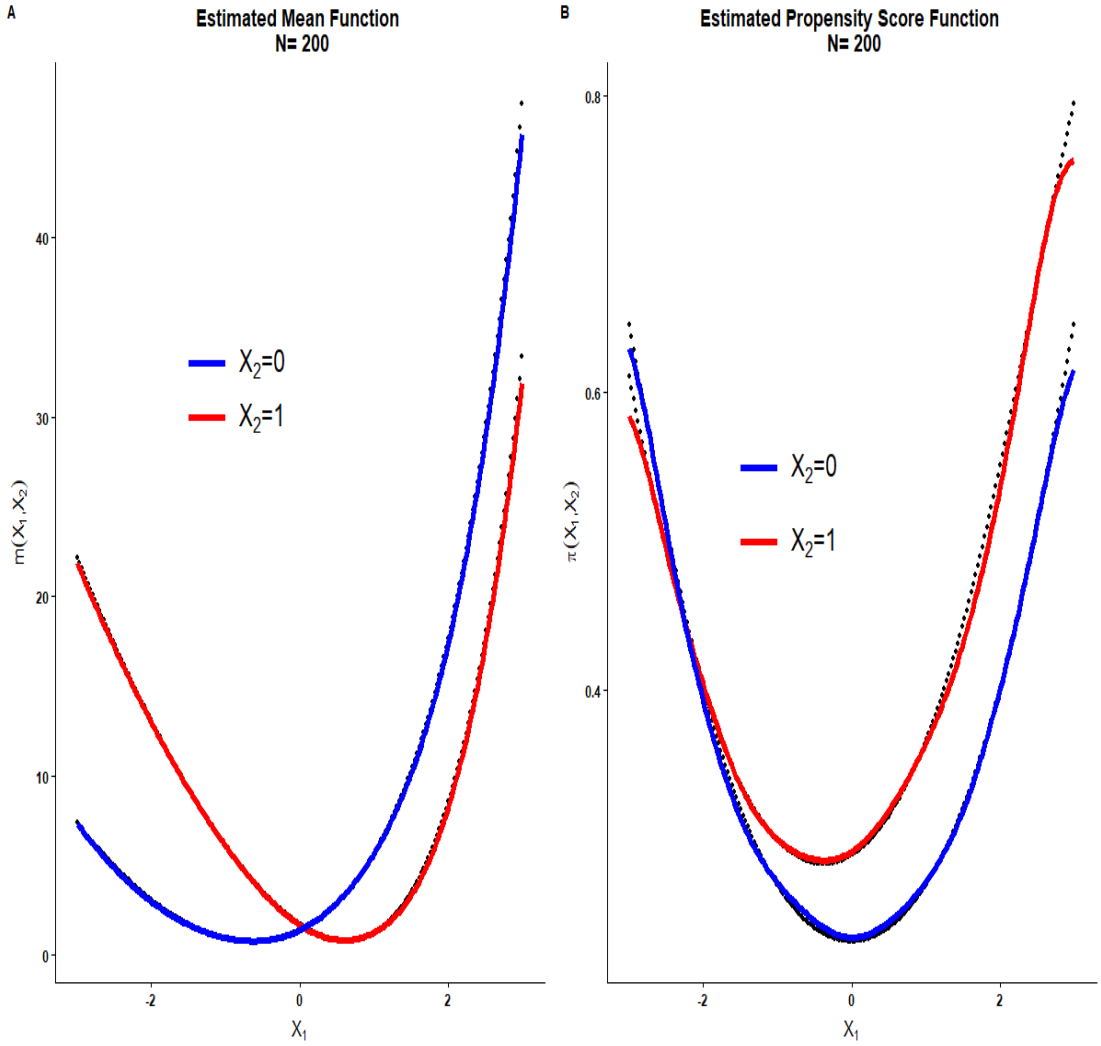
simulation design described above. With sample size being 200, the inner knots number for cubic B spline of X_1 was chosen to be 2, therefore $\alpha^{(1)}$ is a vector of 6 parameters. So are $\alpha^{(2)}$, $\beta^{(1)}$, $\beta^{(2)}$. For each of the 1000 random samples, we obtained $(\hat{\alpha}^{(1)}, \hat{\alpha}^{(2)})$ using the OLS-based method for Y and use them to predict $m(D)$. Similarly, $(\hat{\beta}^{(1)}, \hat{\beta}^{(2)})$ were obtained using the spline-based MLE for A and were used to predict $\pi(D)$. The average over the 1000 predicted mean and propensity scores were then plotted, respectively, in the left and right panels of Figure 3.1, where the black dots presents the true curves and the red line is for the average of estimated curves for $D_2 = 1$; the blue line is for the average of estimated curves for $D_2 = 0$. Figure 3.1 clearly indicates that the cubic B-splines sieve method yields consistent estimation of the true scores.

3.4.1.3 Competing Methods

Benchmark. We included a hypothetical scenario that we knew exactly the outcome model (3.16) and implemented the MLE method to achieve an efficient estimation of ATE of 6. MLE from the true outcome model is expected to be the most efficient estimator of τ. Therefore, result from this hypothetical analysis is used as our benchmark to evaluate all the competing methods.

IPW. In addition to the classic Horvitz-Thompson IPW estimator of ATE, which we denoted as $\tau_{IPW}^{(1)}$, we also considered its two variants $\tau_{IPW}^{(2)}$ and $\tau_{IPW}^{(3)}$ discussed in (Lunceford and Davidian, 2004). These two versions of IPW differ from the $\tau_{IPW}^{(1)}$ in the way of creating inverse weights from the estimated propensity scores. As (Lunceford and Davidian, 2004) showed, $\tau_{IPW}^{(2)}$ is helpful in stabilizing inverse probability weights and $\tau_{IPW}^{(3)}$ improves in precision over $\tau_{IPW}^{(1)}$ and $\tau_{IPW}^{(2)}$. More details of

FIGURE 3.1: Simulation study (1): consistent estimation of the mean and propensity scores using cubic B-splines in stage 1



the relevant theoretical justification can be found in their paper.

$$\tau^{(2)}_{IPW} = \left(\sum_{i=1}^{n} \frac{A_i}{\hat{\pi}(X_i)}\right)^{-1} \sum_{i=1}^{n} \frac{Y_i A_i}{\hat{\pi}(X_i)} - \left(\sum_{i=1}^{n} \frac{1-A_i}{1-\hat{\pi}(X_i)}\right)^{-1} \sum_{i=1}^{n} \frac{Y_i(1-A_i)}{1-\hat{\pi}(X_i)} \tag{3.18}$$

$$\begin{aligned}\tau^{(3)}_{IPW} = &\left(\sum_{i=1}^{n} \frac{A_i}{\hat{\pi}(X_i)}(1 - \frac{C_1}{\hat{\pi}(X_i)})\right)^{-1} \sum_{i=1}^{n} \frac{Y_i A_i}{\hat{\pi}(X_i)} \left(1 - \frac{C_1}{\hat{\pi}(X_i)}\right) - \\ &\left(\sum_{i=1}^{n} \frac{1-A_i}{1-\hat{\pi}(X_i)}(1 - \frac{C_0}{1-\hat{\pi}(X_i)})\right)^{-1} \sum_{i=1}^{n} \frac{Y_i(1-A_i)}{1-\hat{\pi}(X_i)} \left(1 - \frac{C_0}{1-\hat{\pi}(X_i)}\right)\end{aligned} \tag{3.19}$$

where

$$C_1 = \frac{\sum_{i=1}^{n} \left\{\frac{A_i - \hat{\pi}(X_i)}{\hat{\pi}(X_i)}\right\}}{\sum_{i=1}^{n} \left\{\frac{A_i - \hat{\pi}(X_i)}{\hat{\pi}(X_i)}\right\}^2}; C_0 = -\frac{\sum_{i=1}^{n} \left\{\frac{A_i - \hat{\pi}(X_i)}{1-\hat{\pi}(X_i)}\right\}}{\sum_{i=1}^{n} \left\{\frac{A_i - \hat{\pi}(X_i)}{1-\hat{\pi}(X_i)}\right\}^2}$$

For each of the three versions of IPW estimator listed above, we considered two scenarios with propensity scores from (i) true parametric propensity model (IPW-pT) given in (3.15); (ii) wrongly specified ordinary logistic regression model (IPW-pW) with covariates X_1, X_2 and the interaction term X_1X_2.

AIPW. For AIPW, we examined all four possible scenarios: (i) both the mean and propensity score models were specified correctly (AIPW-mT&pT) as given in (3.17) and (3.15), respectively; (ii) the mean score model was specified correctly as given in (3.17) but the propensity score model was specified wrongly as for the IPW-pW estimator (AIPW-mT&pW); (iii) the propensity score model was specified correctly as given in (3.15) but the mean score was wrongly specified as the ordinary linear regression model (AIPW-mW&pT) with covariates X_1, X_2 and X_1X_2;

(iv) both the mean and propensity score models were wrongly specified as aforementioned (AIPW-mW&pW).

Hahn's estimator. We implemented Hahn's estimator given by (1.5) and evaluated its performance in two scenarios (i) the true propensity score are known: (Hahn-Tp); (ii) propensity scores are estimated from the correctly specified propensity given in (3.15): (Hahn-pT). To obtain the nonparametric estimators for the conditional mean of AY given X and $(1-A)Y$ given X , in both scenarios, we adopted the cubic B-spline sieve estimation method and computed $\hat{E}(AY|X)$ and $\hat{E}((1-A)Y|X)$ as

$$\hat{E}(AY|X) = (B^T B)^{-} B^T Y^*$$

$$\hat{E}((1-A)Y|X) = (B^T B)^{-} B^T Y^{**}$$

where B is made of the same basis functions as we use in the estimation of $m(X)$ and $\pi(X)$; $Y^* = (A_1 Y_1, \cdots, A_n Y_n)^T$; $Y^{**} = ((1-A_1)Y_1, \cdots, (1-A_n)Y_n)^T$.

3.4.1.4 Results from Study (1)

We presented the simulation results for sample size 200, 400, and 800 in Table 3.1 with summary of the estimation bias (Bias), Monte-Carlo standard deviation (M-C SD) based on 1000 repetitions, average standard error (ASE) based on 100 bootstrap samples, and 95% coverage probability (95% CP) of the Wald 95% confidence interval. The results from our proposed method were labeled as "MF" in the table. To visualize the asymptotic property for all the estimators, we plotted the distribution of all estimators based on 1000 repetitions and presented them in Figure 3.2 and Figure 3.3. Some comments regarding the results are made as follows.

TABLE 3.1: Simulation study (1): comparison of bias, Monte Carlo standard deviation, asymptotic standard error and 95% coverage probability among all the methods

	Bias			**M-C SD**			**ASE**			95% **CP**		
	N=200	N=400	N=800	N=200	N=400	N=800	N=200	N=400	N=800	N=200	N=400	N=800
IPW(1)-pT	0·018	0·007	0·015	0·975	0·680	0·469	0·991	0·674	0·468	0·944	0·952	0·951
IPW(1)-pW	3·531	3·490	3·511	1·098	0·731	0·532	1·187	0·770	0·528	0·112	0·003	0·000
IPW(2)-pT	0·025	0·008	0·010	0·931	0·650	0·453	0·920	0·640	0·448	0·939	0·950	0·944
IPW(2)-pW	3·412	3·411	3·453	1·051	0·712	0·522	1·072	0·738	0·514	0·092	0·003	0·000
IPW(3)-pT	0·017	0·012	0·011	0·923	0·646	0·450	0·900	0·633	0·445	0·934	0·949	0·944
IPW(3)-pW	3·406	3·409	3·452	1·045	0·711	0·522	1·050	0·734	0·514	0·087	0·003	0·000
AIPW-p&mT	0·002	0·003	0·002	0·154	0·111	0·075	0·154	0·108	0·076	0·941	0·947	0·953
AIPW-pW&mT	0·002	0·003	0·001	0·153	0·110	0·075	0·154	0·107	0·075	0·940	0·952	0·950
AIPW-pT&mW	0·081	0·028	0·030	0·505	0·330	0·231	0·567	0·350	0·235	0·959	0·952	0·945
AIPW-pW&mW	3·470	3·458	3·488	1·049	0·719	0·523	1·090	0·744	0·519	0·095	0·003	0·000
HAHN(pT)	0·064	0·037	0·015	0·591	0·405	0·275	0·636	0·410	0·279	0·960	0·960	0·946
HAHN(Tp)	0·153	0·086	0·007	1·042	0·781	0·523	1·047	0·750	0·532	0·944	0·936	0·945
MLE	0·003	0·003	0·001	0·152	0·109	0·075	0·151	0·106	0·075	0·943	0·948	0·948
MF	0·003	0·002	0·002	0·156	0·111	0·076	0·157	0·109	0·076	0·950	0·945	0·951

M-C SD: Monte Carlo standard deviation from 1000 iterations
ASE: standard deviation from 100 bootstrapping samples
95% CP: 95% coverage probability

(i) Our model-free estimator is comparable to MLE and AIPW with the mean model be correctly specified, and performs much better than IPW even when the propensity model is correctly specified. It is numerically stable with virtually ignorable estimation bias. The Monte-Carlo standard deviation is very small, only slightly bigger than that based on the MLE method when the outcome model is completely known to allow the maximum likelihood estimation, indicating a minor loss of estimation efficiency. Moreover, the average standard error estimate is very close to the M-C SD even with sample size 200 and the coverage probability of 95% CI is also close to the nominal value of 0·95. The histogram of $\hat{\tau}^{mf}$ from 1000 random samples shows that our proposed method leads to an asymptoticly normally distributed estimator in this numerical experiment. It suggests that our method allows the standard statistical inference procedure to be applied for making causal inference on average treatment effect in finite sample.

(ii)When the propensity model is wrong the IPW estimators are severely biased and not reliable. When the propensity model is correctly specified, all the three versions of IPW yield consistent estimator of ATE in this simulation study. The coverage probability of the 95% CI are fine. Actually, biases from $\hat{\tau}^{(1)}_{IPW}, \hat{\tau}^{(2)}_{IPW}, \hat{\tau}^{(3)}_{IPW}$ are very similar, but their corresponding variances are in the order of $var(\hat{\tau}^{(1)}_{IPW}) > var(\hat{\tau}^{(2)}_{IPW}) > var(\hat{\tau}^{(3)}_{IPW})$, which is exactly what (Lunceford and Davidian, 2004) argued in their paper. We only presented the histogram of $\hat{\tau}^{(1)}_{IPW}$ in Figure 3.2 since the other two are very similar to it. Although in this experiment, the inverse probability weighting method results in asymptotic normally distributed estimator of ATE, its variation is apparently much bigger compared to our proposed method. For example,

for sample size 200, the ASE of our proposed estimator is 0.156 while for all the three IPW estimators the ASE is around 1.

(iii) As anticipated, the AIPW method performs much better than the IPW method, especially when the mean model is correctly specified. Its estimation bias is virtually ignorable even when sample size is only 200, comparable to MLE; the average standard error is close to the Monte-Carlo standard deviation, particularly when sample size increases to 800; the coverage probability of the 95% CI is around the nominal value of 0·95. The double robustness property of the AIPW is also demonstrated in the settings of AIPW-mT&pW and AIPW-mW&pT for this simulation study, as the estimation bias is very close to zero. It appears that when the mean score model is correctly specified, the estimation results for the two AIPW scenarios are very similar regardless whether the propensity model is correctly specified or not. In this experiment, we indeed find that AIPW with wrong mean score model but correct propensity model improves the precision over the simple IPW with correct propensity model. However, when both scores are wrongly specified, the estimation is totally off the mark resulting in a very large estimation bias, similarly as the simple IPW estimator with wrongly specified propensity model.

(iv) Hahn's proposed estimator can be viewed as an IPW type estimator therefore we see its similar performance compared to IPW in this simulation. First, it is interesting but not surprising to note that the ASE of τ^{Hahn} based on the true propensity score is larger than the estimator with propensity score estimated from the correct propensity model. (Tsiatis, 2006) already showed that even if the propensity score is known the IPW estimator based on an estimated propensity score is at least as efficient as the IPW estimator that uses the known propensity score. This is to say

that the variance of the influence function for IPW with estimated propensity scores is not larger than the variance of the influence function for known propensity score (Kennedy, 2016). Thus even when the propensity scores are known, it is preferable to estimate the propensity score from the data according to the correct model. Second, with $E(AY|X)$ and $E((1-A)Y|X)$ estimated using nonparametric technique and consistently estimated propensity scores from correctly specified propensity model, τ^{Hahn} is unbiased and asymptotically normally distributed. But as it is based on inverse probability weighting, it still yields much larger variance compared to our proposed estimator, as clearly shown in Table 3.1 and Figure 3.2. In his paper, Hahn proposed to estimate propensity scores also using nonparametric method. Although we didn't present the result in the Table 3.2 and Figure 3.2, we did compute the τ^{Hahn} using the same $\hat{\pi}(X)$'s as used in the calculation for τ^{mf} and found that its variation is even larger than that based on the estimated propensity score from the true parametric propensity model.

3.4.2 Simulation Study (2)

In simulation study (1), we constructed the true propensity score function to create a relatively ideal scenario for IPW type estimators with the true propensity scores are within a nice range so that the estimated propensity scores were expected to stay away from 0 and 1. In many situations, it is possible that the true propensity scores could be close to 0 or 1. The positivity assumption only states that propensity score should be bounded away from 0 or 1, which does not guarantee that estimated propensity scores staying away from 0 and 1. Therefore, we conducted this second simulation study in order to investigate the performance of our proposed estimator

FIGURE 3.2: Histogram of estimated treatment effects from all methods in study (1) with sample size 400: true effect is 6

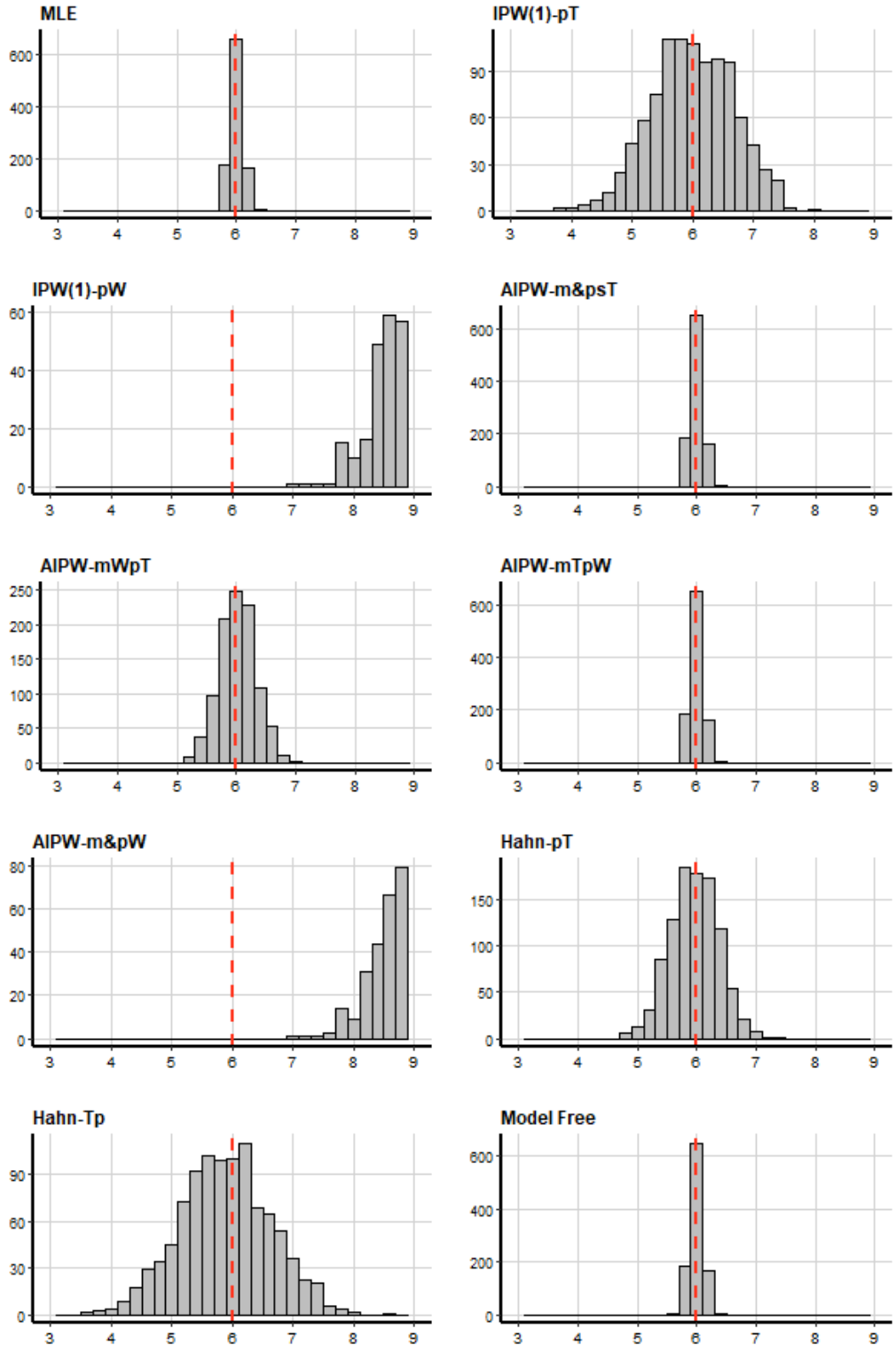
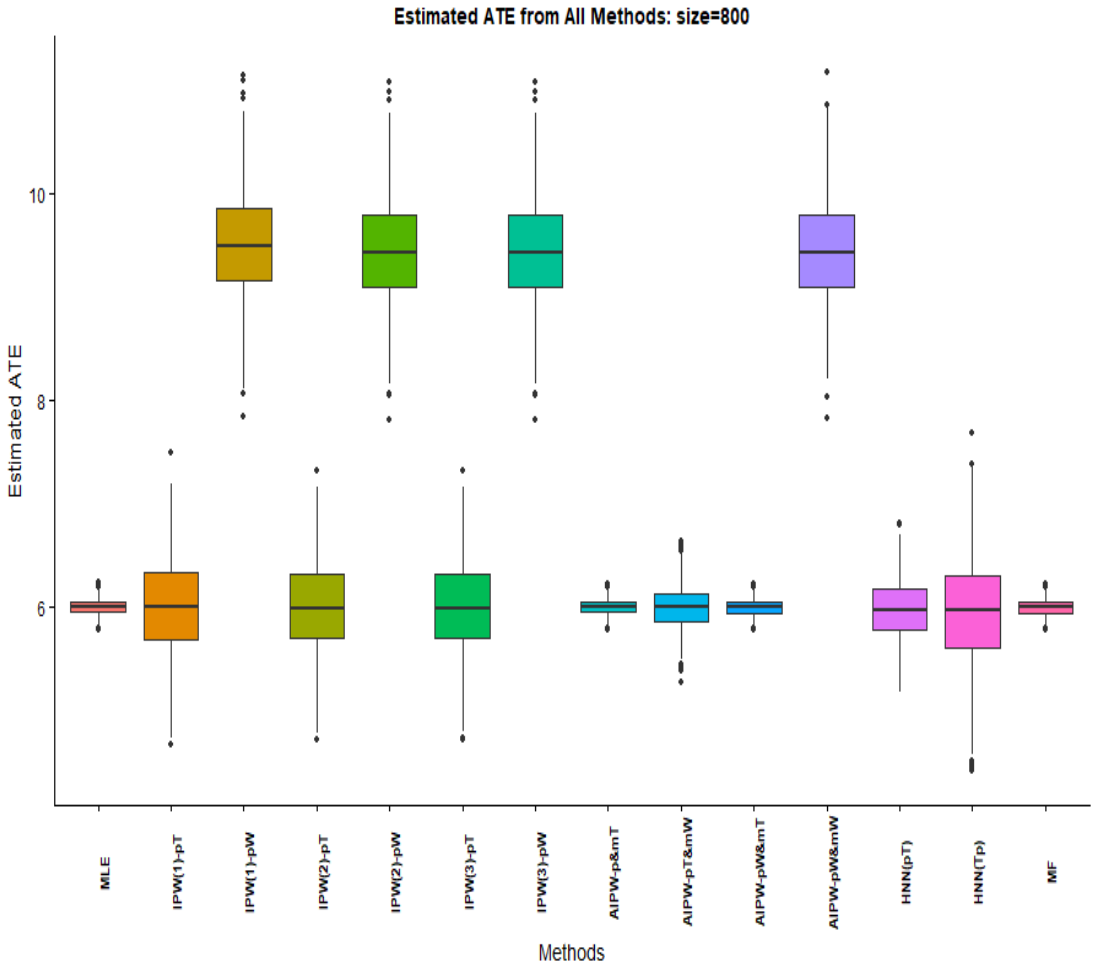


FIGURE 3.3: Boxplots of estimated treatment effects from all methods in study (1) with sample size 800: true effect is 6



and all other competing methods when the true propensity scores may possibly be very close to 0 or 1.

3.4.2.1 Design of Study (2)

In this Monte Carlo simulation, we considered only one covariate X generated from $N(0,1)$. The following logistic model defines the true propensity score as a function of X

$$\pi(X) = P(A = 1|X) = \frac{\exp(-1 \cdot 5 - 1 \cdot 2X + X^2)}{1 + \exp(-1 \cdot 5 - 1 \cdot 2X + X^2)} \tag{3.20}$$

As $\pi(X)$ is completely determined by X, by adjusting the value of X we can force the corresponding propensity scores to be within a certain range (l, u). To do so, we generated X from $N(0,1)$, evaluated $\pi(X)$ using (3.20) and kept X's only if they satisfy $l <= \pi(X) <= u$ for proceeding to the next step of generating A. Thus, by varying the values of l and u, the true propensity scores can be controlled as desired. With a selected X, the treatment assignment indicator A was then generated from $Bernoulli(\pi(X))$. Based on X and A, the outcome Y was generated from the following model

$$Y = -2 + 1 \cdot 6X - 0 \cdot 4X^2 + 1 \cdot 2Xe^{-X} + 6A + \epsilon \tag{3.21}$$

where the random error ϵ is independent of (A, X) and is normally distributed with $N(0, 2)$. So the true ATE is still 6.

Figure 3.4 displays a randomly sample (size=400) with the range of propensity scores specified as $(0.01, 0.99)$. From the plot of true propensity score functional curve,

we can see that at the two ends when X is close to -2 or 3, the true propensity scores reach 0.99, very close to 1. We considered this as a typical example in which possible extreme propensity scores occur.

3.4.2.2 Results from study (2)

We examined all the competing methods described in simulation study (1). As our main purpose was to investigate their performance in presence of extreme propensity scores, for IPW and AIPW methods, we presented the Bias, M-C SD, ASE, 95% CP statistics only for the scenario when propensity model or mean score model was correctly specified in Table 3.2. Table 3.2 and Figure 3.5 reveal that when the true propensity scores are restricted in the range of (0.1, 0.9), all the competing methods have similar performance as seen in the simulation study (1), i.e, unbiased and asymptotically normal. However, when some of the true propensity scores are possibly as low as 0.01 or as high as 0.99, all IPW methods turned out to perform poorly due to some extreme weights, as shown in Figure 3.6 and Figure 3.7. Although the bias is still negligible except IPW(3)-pT, the average of ASE is not close to the M-C SD therefore the 95% CP is off 0.95. As for AIPW, when the mean model for outcome was correctly specified, the estimation of ATE is fine. But when AIPW relies on the correct sepecification of the propensity model, it has similar behavior as IPW. The same argument applies to Hahn's estimators. On the contrast, our proposed estimator shows its robustness against the extreme values of the propensity score and hence has a great advantage in applications. Thus we conclude that the proposed model-free estimator outperforms all IPW type estimators as it removes the impact of individual propensity score weights in the computation of ATE.

FIGURE 3.4: A typical sample generated according to the design of study (2) with extreme true propensity scores

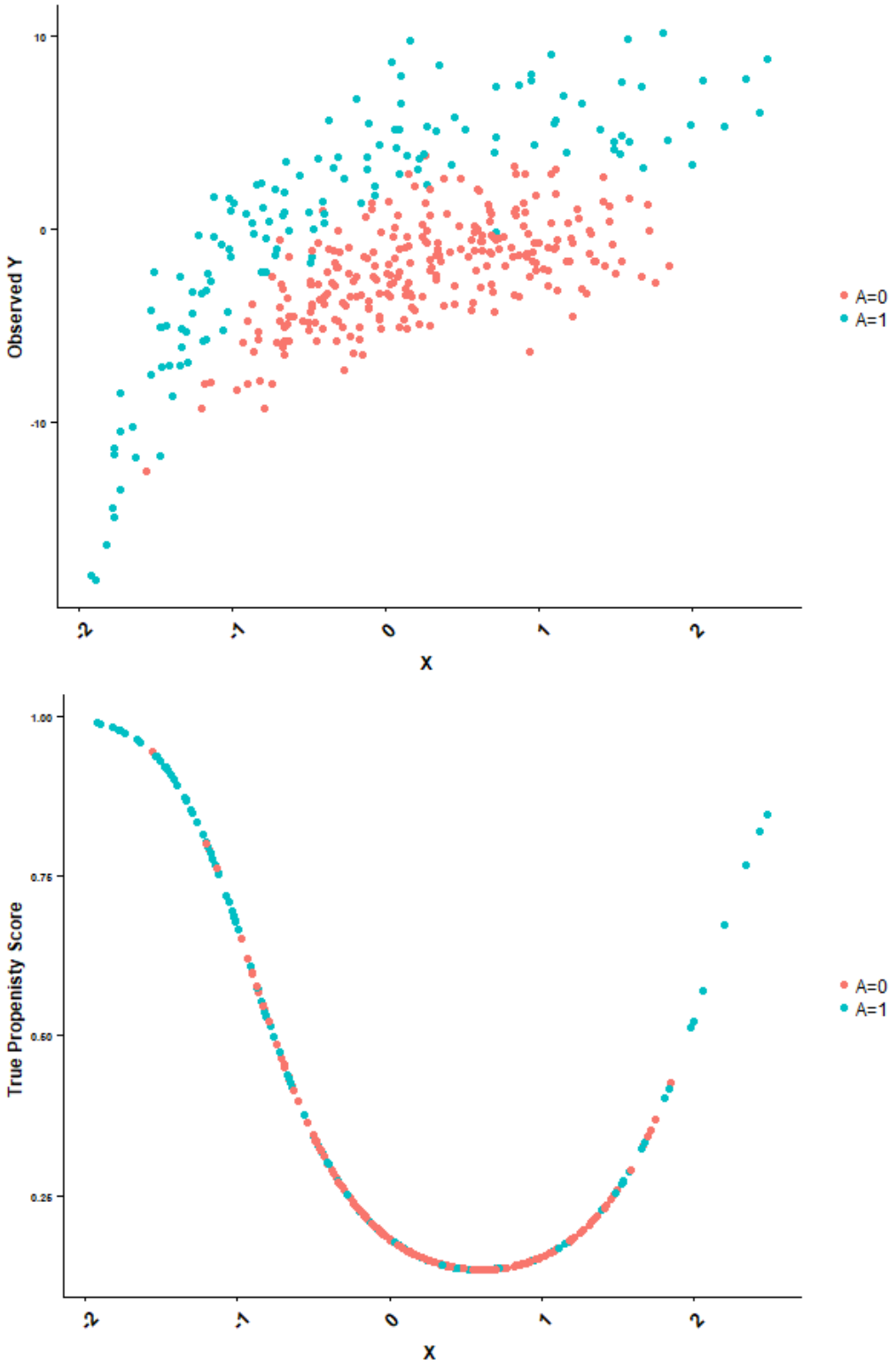


TABLE 3.2: Simulation study (2) under two scenarios: comparison of bias, Monte Carlo standard deviation, asymptotic standard error and 95% coverage probability among all the methods

	Bias	**M-C SD**	**ASE**	95% **CP**
True propenisty score range (0·1 - 0·9)				
IPW(1)-pT	0·010	0·296	0·299	0·960
IPW(2)-pT	0·005	0·287	0·285	0·950
IPW(3)-pT	0·002	0·273	0·268	0·942
AIPW-p&mT	0·007	0·256	0·254	0·953
AIPW-pW&mT	0·007	0·253	0·252	0·955
AIPW-pT&mW	0·001	0·276	0·273	0·945
Hahn-pT	0·039	0·287	0·293	0·957
Hahn-Tp	0·072	0·501	0·494	0·928
MLE	0·007	0·245	0·243	0·953
MF	0·005	0·246	0·245	0·950
True propenisty score range (0·01- 0·99)				
MLE	-0·012	0·250	0·248	0·941
IPW(1)-pT	0·009	1·380	0·797	0·788
IPW(2)-pT	0·063	0·908	0·578	0·807
IPW(3)-pT	0·198	0·417	0·362	0·838
AIPW-p&mT	0·011	0·287	0·293	0·946
AIPW-pW&mT	0·020	0·272	0·288	0·954
AIPW-pT&mW	0·026	0·942	0·568	0·857
Hahn-pT	0·133	1·250	1·152	0·864
Hahn-Tp	0·108	1·215	0·938	0·797
MF	0·017	0·251	0·250	0·943

M-C SD: Monte Carlo standard deviation from 1000 iterations

ASE: standard deviation from 100 bootstrapping samples

95% CP: 95% coverage probability

FIGURE 3.5: Distribution of estimated treatment effect from all methods (scenario 1: true propensity scores range $0.1 - 0.9$)

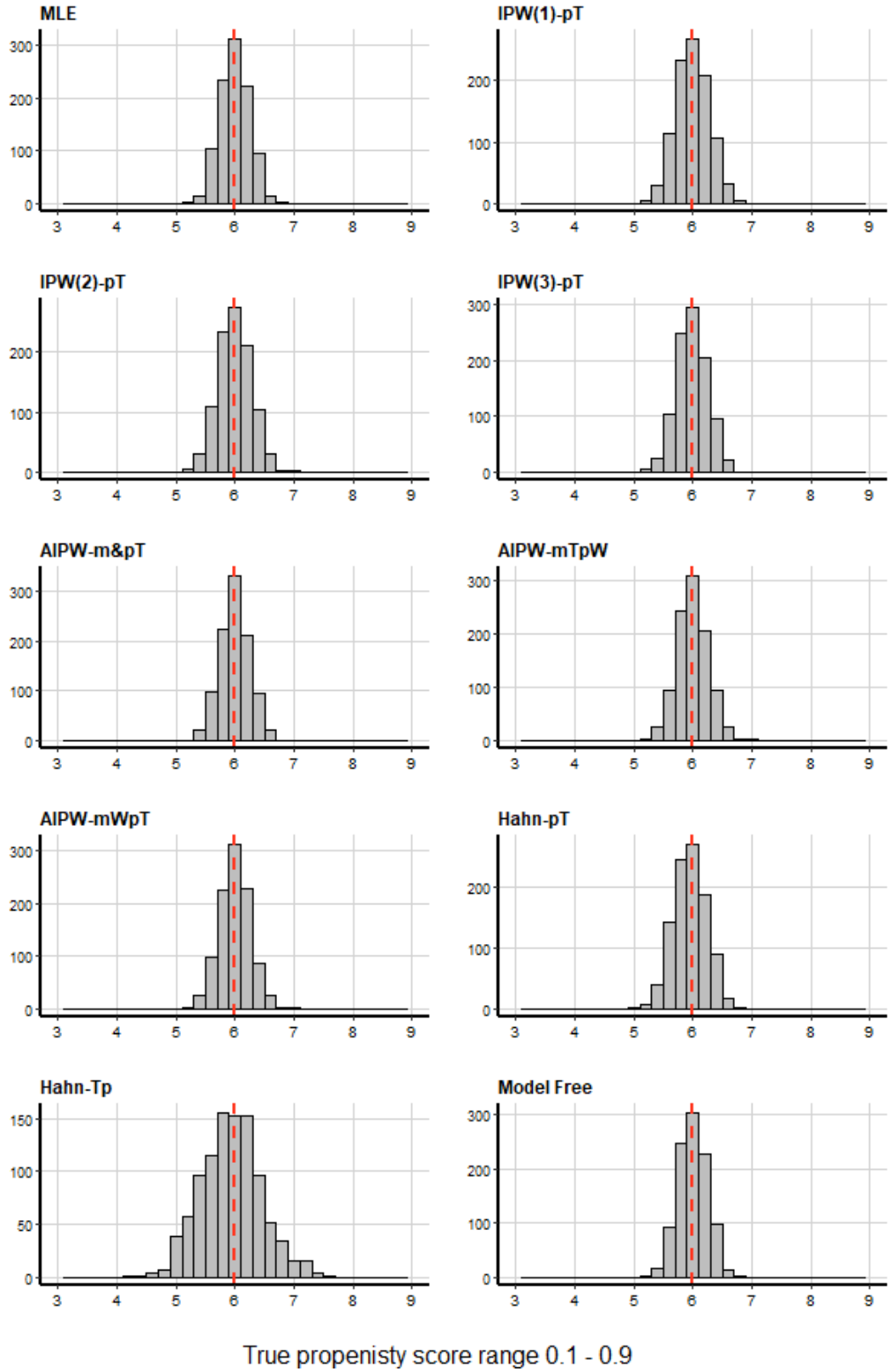


FIGURE 3.6: Distribution of estimated treatment effect from all methods (scenario 2: true propensity scores range $0.01 - 0.99$)

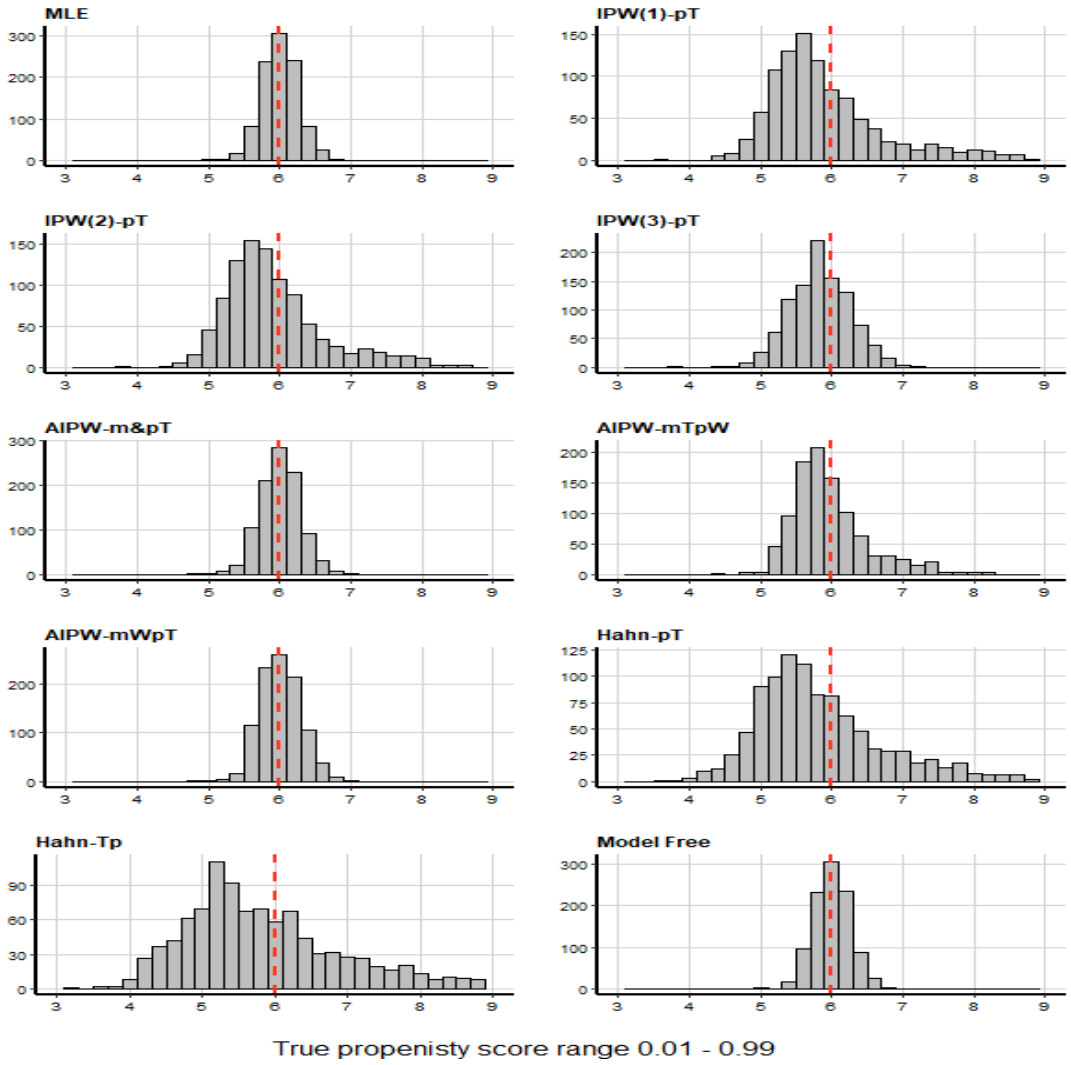
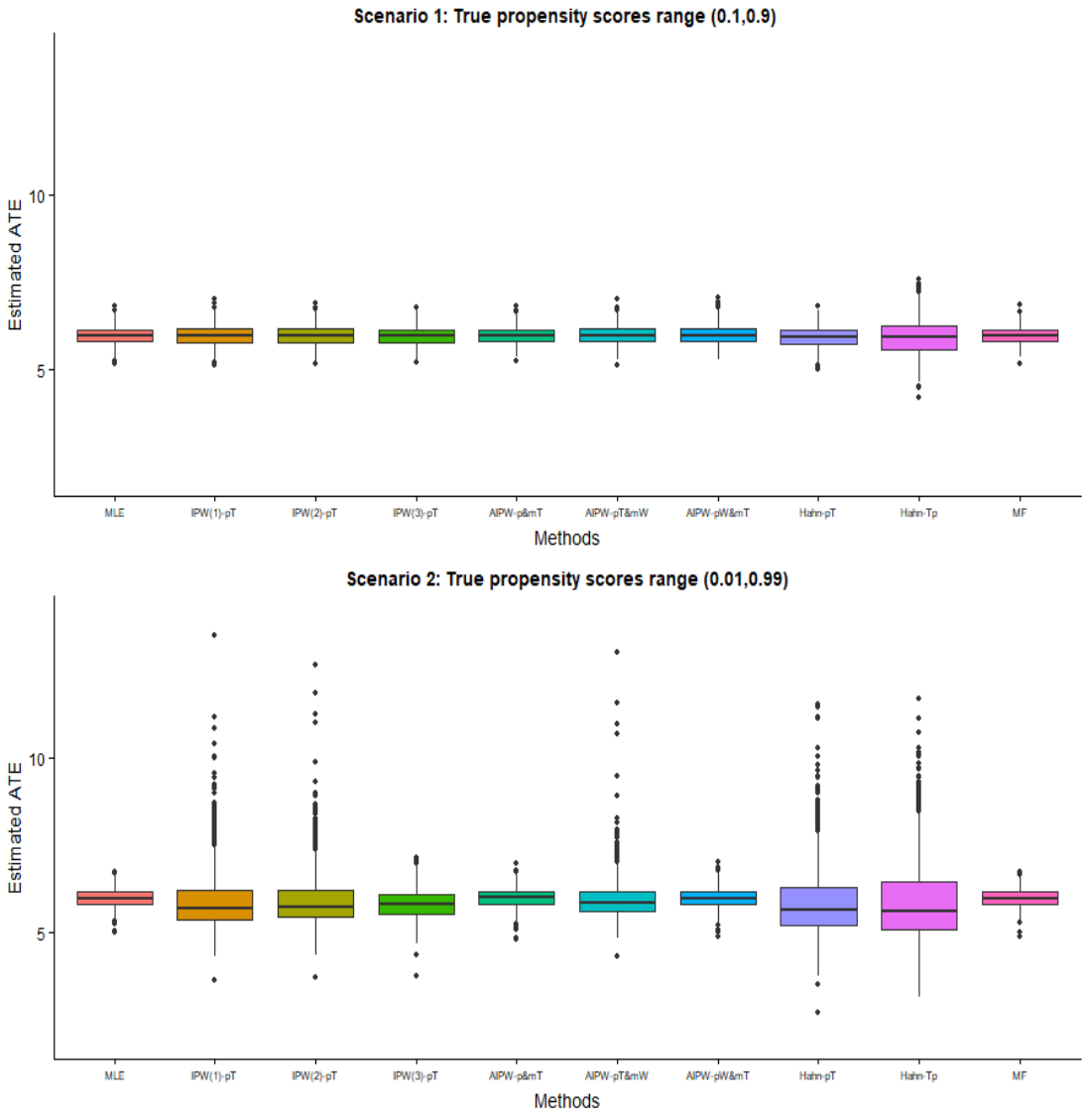


FIGURE 3.7: Estimated treatment effect from all methods for simulation study 2 with sample size 400



3.5 A Case Study

3.5.1 JIA Study Background

Juvenile idiopathic arthritis (JIA) describes a type of chronic inflammatory disease characterized by arthritis that begins before age 16 and persists for a minimum of 6 weeks. As one of the most common type of rheumatologic disease in children, JIA may result in disability and has incident rate of approximately 10 per 100,000 for girls and 5.7 per 100,000 for boys (Oberle et al., 2014). Currently, the cause of childhood arthritis is unknown with limited understanding of the disease etiology and pathogenesis (JIA, 2017). A variety of therapies have been used in treating JIA. Nonsteroidal anti-inflammatory drugs (NSAIDs), including ibuprofen and naproxen are the conventional early treatment for children with JIA to ease pain and inflammation. However, since NSAIDs do not prevent joint damage, they are not considered as disease-modifying agents and usually used in combination with other disease-modifying antirheumatic drugs (DMARDs). DMARDs are anti-inflammatory medicines capable of preventing joint damage, such as cartilage and bone destruction. There are two types of DMARDs: non-biologic DMARD and biologic DMARD. The most commonly prescribed non-biologic DMARD is methotrexate among others such as sulfasalazine, leflunomide and hydroxychloroquine. Biologic DMARDs including abatacept, adalimumab, canakinumab etc. are targeted to alter a specific step in the pathogenesis of the inflammatory response associated with the disease (Guo et al., 2018). Currently the most prevalent clinical practice for the treatment plan of JIA is to start patient on a non-biologic DMARD as the first line of treatment, then step-up by switching to or adding biologic DMARDs if the patients fail to make sufficient

progress. The effectiveness of early combination of DMARDs vs. the monotherapy of non-biologic DMARDs has been reported in studies in adult RA population (Gabay et al., 2015). In pediatric population the evidence has been limited. There were, however, studies with focus on pediatric population suggesting that there is a window of opportunity where early effective treatment could address underlying disease pathophysiology, prevent structural damage in joins, and thus promise for earlier and sustainable control of disease (Wallace et al., 2012).

This case study was designed with the goal to evaluate the effectiveness of the early aggressive (early combination of nonbioloigc DMARD and biologic DMARD) vs. step-up consensus treatment plan (CTP, starting on a non-biologic DMARD followed by switching to or adding biologic DMARDs) among pediatric patients with newly onset of JIA disease. The primary data resource is the electronic medical records (EMR) extracted from the Cincinnati children's hospital medical center(CCHMC)'s Epic system. All the rheumatology clinical encounters for patients diagnosed with pcJIA (Polyarticular-course JIA) were extracted from Epic between January 1st 2009 and December 31, 2017. Patients diagnosed with pcJIA for at least two distinct visits by the pediatric rheumatologists were identified as pcJIA patients. The study sample was made of 509 eligible pcJIA patients who were 1-19 years old,and newly diagnosed ($<$ 6 months) with pcJIA following the CARRA operational definition based on the ILAR (international league of association for Rheumatology, http://www.ilar.org/) code. The study individuals received prescription of either early combination DMARDs or non-biologic DMARD monotherapy as the first line treatment within 9 months of diagnosis and didn't have the comorbid conditions of inflammatory bowel disease (IBD), celiac disease, and trisomy 2. This study was approved by the IRB at Cincin-

nati children's hospital medical center, and was registered at CT.gov (NCT02524340) and HSRProj (20153590).

3.5.2 Baseline Characteristics of the Study Population

Out of a total of 509 eligible patients, there were 407 in either the early aggressive treatment arm or in the Step-up arm. Patients treatment identification was obtained by retrospective chart review of the medical records. The baseline visit for each patient was defined as the time that the patient initialized his/her first DMARD. The primary outcome of interest is the clinical juvenile disease activity score (cJADAS) within the 6 month window (4 to 8 months) after baseline visit. cJADAS is a summary score made of 3 components : physician's global rating of overall activity (MD global), parent/child ratings of well-being (Wellbeing), and counts of active joints (AJC). Each of these 3 components takes values from 0 to 10 with 10 indicating the most severe disease status. So cJADAS has the possible highest value of 30. As there were 80 subjects missing their 6-month cJADAS the final sample for analysis was further reduced to the 327 patients with observed 6-month outcome, among whom 225 from the early aggressive treatment arm and 102 from the step-up arm. Since some of the patients with available 6-month cJADAS outcome miss some of the baseline characteristics, we adopted the popular R package $\{MICE\}$ (van Buuren and Groothuis-Oudshoorn, 2011) to impute the baseline information from the predicted mean matching approach (PMM method) and obtained 5 datasets with complete baseline information for the following analysis.

Table 3.3 provides the summary statistics of two baseline characteristics, cJADAs and pain, of the patients grouped by the treatment arms. P values presented are based

TABLE 3.3: Baseline cJADAs and pain by treatment arms

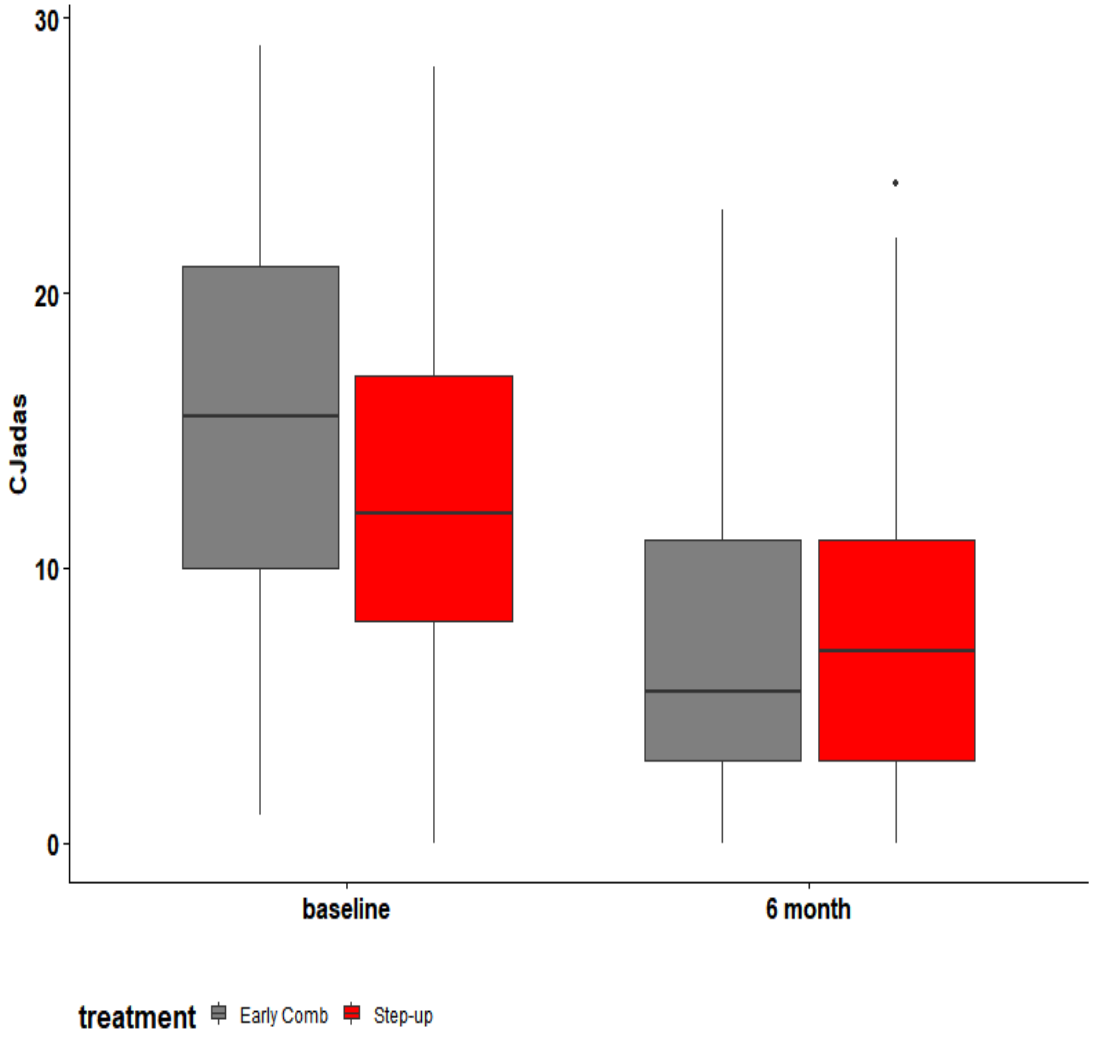
Baseline Covariates	Step-up	Early Combination	P value
cJADAs	12.8 ± 6.07	15.4 ± 7.34	0.001
Pain	4.23 ± 2.67	5.07 ± 2.75	0.01

on Wilcoxon test. Pain is the most common symptom of JIA disease and has been suggested to be linked with disease activity. Pain is also a score ranging from 0 to 10 with 10 indicating the highest severity in pain. Table 3.3 indicates that selection of patients into the early-combination group is associated with the patient's prognosis. patients with severe baseline disease, i.e. large score of baseline cJADAs and pain, were more likely to be assigned to the early combination group. Figure 3.8 gives an indication of beneficial effect of biological DMARD on lowing down the cJADAS score in this study population: the change of cJADAS from baseline to 6-month visit is -8.22 and -5.18 for the early combination group and step-up group, respectively.

3.5.3 Estimation of the ATE of Early Use of Biologic DMARD

In this exploratory analysis we considered the two baseline covariates of cJADAS and pain, as shown in Table 3.3 to be two potential confounders . To evaluate the average effect of the early aggressive use of biologic DMARD, we first dichotomized the pain score using cutoff 3 which makes about 42% patients falling in the low-pain group (pain$\leq$3). Therefore, baseline cJADAS and binary indicator of low pain are the continuous and binary covariates of X_1 and X_2, respectively, as described in the simulation study (1). We used the same estimating strategy described in that simulation study to estimate both marginal mean score and propensity score: applying cubic B-spline only to baseline cJADAS but including its interaction with low pain in

FIGURE 3.8: cJADAs at baseline and 6 month by treatment group



the modeling. With sample size being 327 the interior knots number for constructing such cubic B spline was chosen to be 3 and knots were placed at the 25%, 50%, 75% percentiles of cJADAS.

We also implemented outcome regression, IPW and AIPW approaches with the same two covariates to estimate the treatment effect. To make a fair comparison, for other methods, the outcome and treatment assignment were also modeled nonparametrically as:

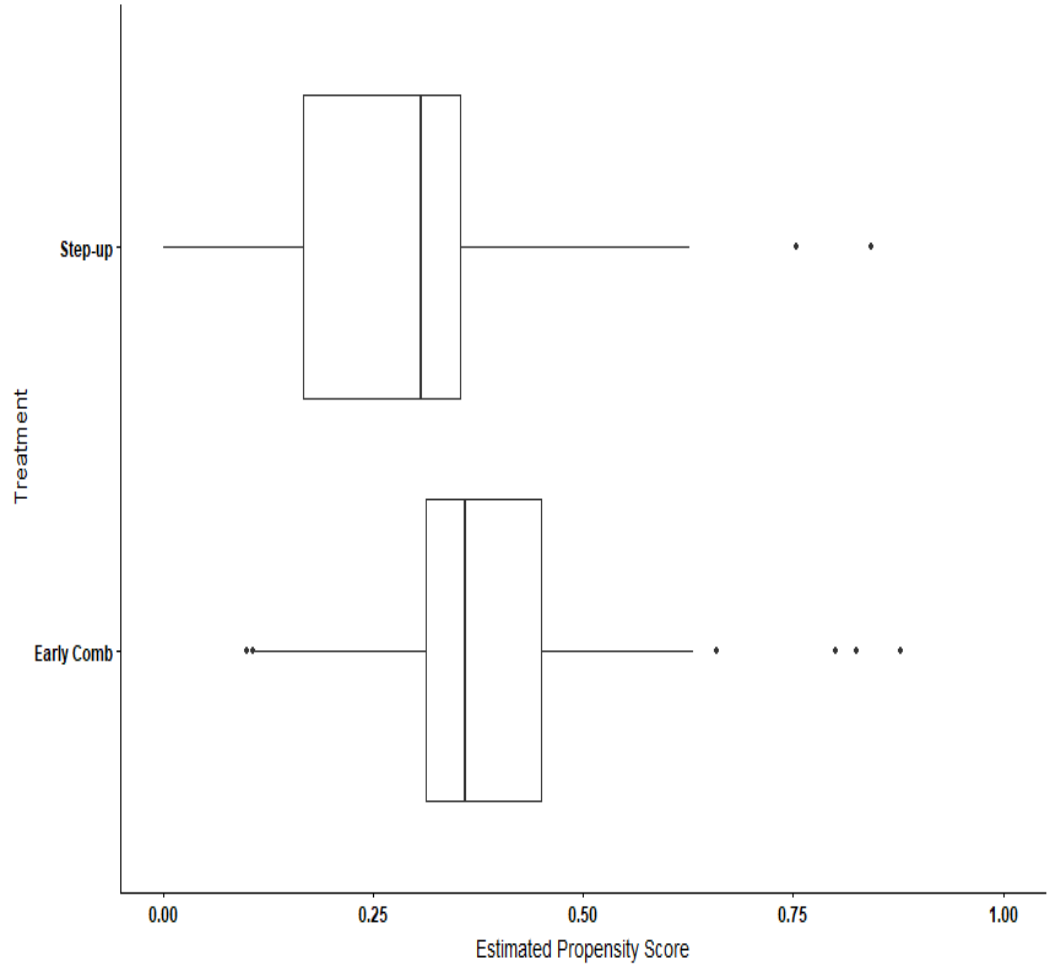
(i) $Y = f(X_1, X_2) + \alpha A + \epsilon$ where $f(X_1, X_2)$ was approximated by adopting exactly the same sieve estimating strategy as used in mean score and propensity score estimation in our proposed method. This is essentially a partial linear regression approach to estimate $\hat{\alpha}$ as the targeted treatment effect.

(ii) All the three versions of IPW using the nonparametricly estimated propensity scores in our proposed method.

(iii) AIPW with outcome modeled in the same way in (i) and propensity scores the same as in our proposed method and IPW.

For the purpose of assessing the validity of the propensity score estimation based on the cubic B-spline regression, we looked into the degree of overlapping in the distribution of estimated propensity scores for the two groups and covariate balancing statistics. As shown in Figure 3.9, the overlapping is quite satisfactory and greater balance is achieved for cJADAs and low pain indicator: in unweighted sample, standardized mean difference for cJADAs and low pain are 0.41 and 0.12, respectively. In weighted sample, standardized mean difference for cJADAs and low pain are reduced to 0.03 and 0.005, respectively. These balance diagnosis statistics indicate that the adequacy of the specification of the propensity-score model is desirable.

FIGURE 3.9: Distribution of the estimated propensity scores using cubic B-splines of baseline cJADAs and its interaction with low pain indicator in JIA study



We presented results of the estimated average treatment effect, the estimated standard error as well as the 95% confidence intervals in Table 3.4. The average treatment effects showed in Table 3.4 were averaged across the 5 imputed dataset and we used 500 bootstrap samples to obtain an estimate of the within-imputation standard error for each imputed dataset and adopted Rubin's rule of multiple imputation (Barnard and Rubin, 1999) to combine estimators and standard errors from the 5 datasets. Under the assumption of homogeneous treatment effect, our proposed model-free method gives the estimated average treatment effect -1.27 with 95% CI $(-2.65, 0.11)$, in contrast to partial linear regression estimator $-1.22(-2.6, 0.15)$, IPW^1 $-1.07(-3.21, 1.08)$, IPW^2 $-1.06(-2.41, 0.46)$, IPW^3 $-1.06(-2.49, 0.36)$, AIPW $-1.04(-2.57, 0.48)$, Hahn's method $-0.76(-2.5, 0.97)$. The results indicate that early aggressive use of biologic DMARDs can be effective. The averaged treatment effect estimated using the proposed method suggests that the early aggressive use of biologic DMARDs leads to about 1.27 point reduction in cJADAs in treating children with newly diagnosed pcJIA. Partial linear outcome regression yields similar results with comparable standard error and confidence interval. All the IPW type methods give effects around -1 but with larger standard errors and therefore wider confidence intervals, especially from IPW^1.

TABLE 3.4: Estimated average causal effects of early aggressive use of biologic DMARD based on baseline cJADAs and low pain indicator in JIA study assuming homogeneous treatment effect

Method	Estimate	Ste.	95%lower limit	95%upper limit
Regression	-1.222	0.701	-2.595	0.152
IPW[1]	-1.066	1.093	-3.208	1.076
IPW[2]	-1.063	0.777	-2.587	0.460
IPW[3]	-1.064	0.723	-2.488	0.361
AIPW	-1.042	0.779	-2.568	0.484
HAHN	-0.765	0.886	-2.501	0.971
Model Free	-1.271	0.702	-2.648	0.106

CHAPTER 4

Extension of the OLS-based Method to Estimate ATE in Heterogeneous Treatment Effect Scenario

4.1 Heterogeneity of Treatment Effects

In many real-world applications, individuals in the study population usually differ in their background characteristic as well as how they respond to a given treatment. For example, in a typical biomedical observational study using medical records from healthcare databases, patients possess diverse characteristics such as age, gender, race, disease etiology and severity, presence of comorbidities, and some genetic risk factors etc.. As a given treatment might affect the outcome of interest for patients with different characteristics in different ways, the causal treatment effects are potentially modified by these varying patient characteristics instead of being homogeneous across the population. Heterogeneity of treatment effects, in other words, indicates that there exists interactions between treatment and certain patient characteristics. So the individual treatment effects $\tau_i = \tau + \epsilon_i$ with the error term ϵ_i dependent on covariates. τ_i's are not identifiable due to the fundamental problem of missingness in potential outcomes but as it was argued in Chapter 3, under standard causal assumptions the conditional average treatment effect can be used to characterize the subpopulation average treatment effects given the covariates. In heterogeneous treatment effect scenario, these conditional average treatment effects can then be aggregated to yield population average treatment effect.

In Section 3.2 of Chapter 3, the core linear equation for estimating ATE in our proposed method was given as

$$E(Y|A=a,X) = m(X) + (a - \pi(X))\tau(X) \tag{4.1}$$

where $\tau(X)$ denotes the conditional average treatment effect or covariates specific treatment effect function. When the homogeneous treatment effect assumption is not warranted, i.e. $\tau(X) \neq \tau$, in order to obtain an estimator of τ based on (4.1), it is necessary that the estimation of $\tau(X)$ should be accomplished first. One may see that an intuitive way to achieve this goal is simply to impose a parametric model of $\tau(X)$ so that the estimation of τ becomes analogous to the homogeneous treatment effect case. As an illustrating example, suppose X represents age and we wish to conduct subgroup analysis assuming that the treatment has different effect on $age > 60$ group and $age \leq 60$ group. Then simply, we can define $\tau(X) = a_0 + a_1 I(X > 60)$ where $1(.)$ is the indicator function. With such formulation of $\tau(X)$, the two-stage estimation method discussed in Chapter 3 can be directly applied to obtain estimators of $\hat{a}_0$ and $\hat{a}_1$ for a_0 and a_1, respectively. The population ATE τ can then be taken as the average effect over the two age groups with $\hat{a}_0$ and $\hat{a}_1$.

Parametric modeling of $\tau(X)$ greatly simplifies the task of estimating $\tau(X)$ and τ. It however introduces the additional assumption of heterogeneity mechanism that deviates from the spirit of model-free. Just like the fact that in reality we don't possess accurate knowledge of outcome generating process and the treatment assignment mechanism, we usually are not knowledgeable about the underlying heterogeneity pattern of the treatment effect and should try to avoid making parametric

assumptions. In this Chapter, we describe the way of extending our proposed model free method to heterogeneous treatment effect scenario. With the desire to maintain the model-free feature of our proposed estimator of τ in the heterogeneous treatment effect scenario, we propose to add in one more stage of sieve estimation of $\tau(X)$ into the two-stage estimation procedure. The nice linear structure of (4.1) makes such extension natural and easily implementable. With the help of splines as useful non-parametric estimation tool, we show that the proposed spline-based nonparametric extension maintains the model-free feature of the average treatment effect estimator. Moreover, we demonstrate that such extension is not only capable of estimating ATE consistently but also has the great advantage of studying treatment effect heterogeneity through the estimated covariate-specific treatment effect function, which may be more meaningful than studying the population ATE in many applications.

4.2 Model-Free Method for Heterogeneous Treatment Effect

4.2.1 Motivation

To implement our model-free estimation strategy for τ in heterogeneous treatment effect scenario, we first assume both the mean and propensity scores are known in (4.1). It is obvious that the estimation of $\tau(X)$ with known $m(X)$ and $\pi(X)$ becomes exactly the same non-parametric regression problem as (2.1) discussed in Chapter 2. Therefore, the regression-spline based nonparametric sieve estimation method can be directly applied to estimate $\tau(X)$.

Following the same argument of sieve estimation discussed before, we seek to estimate $\tau(x)$ in a sieve space spanned by B-splines

$$\tau(x) = \sum_{j=1}^{q_n} \delta_j B_j(x)$$

where $B_j(x), j = 1, \cdots, q_n$, the pre-specified spline basis functions, are given in Section 2.3 of Chapter 2. q_n specifies the dimensionality of the sieve space and increases as sample size increases. The spline coefficients $\delta = (\delta_1, \cdots, \delta_{q_n})$ are then estimated by $\hat{\delta} = (\hat{\delta}_1, \cdots, \hat{\delta}_{q_n})$ based on OLS

$$\hat{\delta} = \arg\min_{\delta} \sum_{i=1}^{n} \left\{ Y_i - m(X_i) - \{(A_i - \pi(X_i))\} \sum_{j=1}^{q_n} \delta_j B_j(X_i) \right\}^2 \tag{4.2}$$

Denote

$$D = \begin{bmatrix} A_1 - \pi(X_1) & & \\ & \ddots & \\ & & A_n - \pi(X_n) \end{bmatrix}$$

$$B = \begin{bmatrix} B_1(X_1) & B_2(X_1) & \dots & B_{q_n}(X_1) \\ B_1(X_2) & B_2(X_2) & \dots & B_{q_n}(X_2) \\ \dots & \dots & \dots & \dots \\ B_1(X_n) & B_2(X_n) & \dots & B_{q_n}(X_n) \end{bmatrix} = \begin{bmatrix} B(X_1) \\ B(X_2) \\ \vdots \\ B(X_n) \end{bmatrix}$$

where $B(x) = (B_1(x), \cdots, B_{q_n}(x))$. Thus the OLS-estimation of δ is given by

$$\hat{\delta} = (\hat{\delta}_1, \cdots, \hat{\delta}_{q_n})^T = ((DB)^T DB)^- (DB)^T (Y - M) = (B^T D^2 B)^- (DB)^T (Y - M)$$

where $Y = (Y_1, \cdots, Y_n)^T, M = (m(X_1), \cdots, m(X_n)^T$ and

$$D^2 = \begin{bmatrix} (A_1 - \pi(X_1))^2 & & \\ & \ddots & \\ & & (A_n - \pi(X_n))^2 \end{bmatrix}$$

The sieve estimator of $\hat{\tau}(x)$ is then

$$\hat{\tau}(x) = \sum_{j=1}^{q_n} \hat{\delta}_j B_j(x) \tag{4.3}$$

To obtain the final estimator of τ, since $\tau = E_X(\tau(X))$, we propose to estimate τ using the empirical mean of $\hat{\tau}(X)$. That is

$$\hat{\tau} = \frac{1}{n} \sum_{i=1}^{n} \hat{\tau}(X_i) = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{q_n} \hat{\delta}_j B_j(X_i) \tag{4.4}$$

4.2.2 Three-Stage Estimation Procedure

Since $m(X)$ and $\pi(X)$ are unknown in reality, we need to replace them with their corresponding estimators in order to compute $\hat{\tau}(x)$ in (4.3) and $\hat{\tau}$ in (4.4). These can be accomplished in the following three-stage estimation procedure.

- Stage 1: Nonparametric estimation of mean and propensity scores.

  This stage remains the same as what we described in Section 3.3.1 of Chapter 3.

  By using regression-spline based nonparametric sieve least-squares estimation

method, $\hat{m}(x)$ may be obtained as

$$\hat{m}(x) = \hat{\alpha}^T B = \sum_{j=1}^{q_n} \hat{\alpha}_j B_j(x)$$

where $\hat{\alpha}_j, j = 1, \cdots, q_n$ are the estimated spline coefficients

$$\hat{\alpha} = (\hat{\alpha}_1, \cdots, \hat{\alpha}_{q_n})^T = (B^T B)^{-} B^T Y$$

Similarly, the regression-spline based nonparametric maximum likelihood estimation (NPMLE) method yields $\hat{\pi}(x)$ with

$$\hat{\pi}(x) = \frac{\exp\left(\sum_{j=1}^{q_n} \hat{\beta}_j B_j(x)\right)}{1 + \exp\left(\sum_{j=1}^{q_n} \hat{\beta}_j B_j(x)\right)}$$

where $\hat{\beta}_j, j = 1, \cdots, q_n$ are solved from

$$\sum_{i=1}^{n} B_j(X_i) \left\{ A_i - \frac{\exp\left(\sum_{j=1}^{q_n} \beta_j B_j(X_i)\right)}{1 + \exp\left(\sum_{j=1}^{q_n} \beta_j B_j(X_i)\right)} \right\} = 0; j = 1, \cdots, q_n$$

by Newton-Raphson algorithm.

- Stage 2: Sieve estimation of $\hat{\tau}(X)$ with plug-in estimated mean and propensity score.

  The estimated $\hat{\pi}(X)$ from Stage 1 is then substituted for the corresponding $\pi(X)$ in D for the sieve OLS estimation of $\hat{\delta}$; and the estimated $\hat{m}(X)$ from

Stage 1 replaces the corresponding $m(X)$ in the vector M. Let

$$\hat{D} = \begin{bmatrix} A_1 - \hat{\pi}(X_1) & & \\ & \ddots & \\ & & A_n - \hat{\pi}(X_n) \end{bmatrix} \text{ and } \hat{M} = (\hat{m}(X_1), \cdots, \hat{m}(X_n)^T,$$

then

$$\hat{\delta}^* = (B^T \hat{D}^2 B)^{-} (\hat{D}B)^T (Y - \hat{M})$$

that yields a model free estimator of $\tau(X)$ given by

$$\hat{\tau}^{mf}(x) = \sum_{j=1}^{q_n} \hat{\delta}_j^* B_j(x)$$

- Stage 3: Empirical mean of $\hat{\tau}^{mf}(x)$ as the final estimator of τ.

  For $X_i = x_i; i = 1, \cdots, n$ we compute its empirical mean and obtain the model-free estimator of τ by

$$\hat{\tau}^{mf} = \frac{1}{n} \sum_{i=1}^{n} \hat{\tau}^{mf}(X_i)$$

4.2.3 Some Comments

- In the proposed three-stage procedure for the estimation of τ, the estimation of $\tau(X)$ acts as an intermediate step. This intermediate step actually provides rich and meaningful information in the analysis of heterogeneous treatment effect other than estimating the overall average treatment effect. If the main purpose of a causal study is to compare treatment effect among different groups of people such as precision medicine, examining treatment heterogeneity will be a

central task. The function $\hat{\tau}^{mf}(X)$ can be used to examine how the treatment effect varies according to covariates X. So by making use of this estimated covariates-specific treatment effect function, we are able to evaluate the specific treatment effect for a given individual when the relevant covariates of the individual X is provided. It is also easy to use $\hat{\tau}^{mf}(X)$ to characterize the group of patients who are expected to benefit most from the treatment so that in practice researchers may learn from the data whom are the subjects that should be treated. The IPW-type methods, however, cannot be directly applied to examine heterogeneity through inverse weighting. Note that

$$E\left(\frac{AY}{\pi(X)}\right) = E\left(E\left(\frac{AY^{(1)}}{\pi(X)}|X\right)\right) = E\left(\frac{Y^{(1)}E(A|X)}{\pi(X)}\right) = E(Y^{(1)}),$$

and $E\left(\frac{(1-A)Y}{1-\pi(X)}\right) = E(Y^{(0)})$. It is clear that IPW doesn't give estimators of $E(Y^{(a)}|X); a = 1, 0$ and results in an estimator with interpretation of population ATE. Although this makes it simple to estimate ATE without bothering treatment heterogeneity, it has the disadvantage of being inflexible to look into the picture of heterogeneity.

- Another interesting question regarding the estimation of ATE in the presence of heterogeneity is to see if our proposed method yields a consistent estimation of τ when the treatment heterogeneity is ignored. Homogeneous treatment effect can be considered as a special case of heterogeneous treatment effect. Since

when $\tau(X) = \tau$, $B = {I_n}^T$, $\hat{D}B = (A_1 - \hat{\pi}(X_1), \cdots, A_1 - \hat{\pi}(X_n))^T$ and

$$\begin{aligned} \hat{\tau} &= \frac{1}{n}\sum_{i=1}^{n} \hat{\tau}^{mf}(X_i) \\ &= \frac{\sum_{i=1}^{n} \{(Y_i - \hat{m}(X_i))(A_i - \hat{\pi}(X_i))\}}{\sum_{i=1}^{n}(A_i - \hat{\pi}(X_i))^2} \end{aligned} \tag{4.5}$$

On the other hand, in the case of $\tau(X) \neq \tau$ if we ignore this fact of heterogeneity and estimate τ by (4.5), then since (4.1) can be rewritten as

$$E(Y|A, X) = m(X) + (A - \pi(X))\tau + (A - \pi(X))\,(\tau - \tau(X)) \tag{4.6}$$

and $(A - \pi(X))\,(\tau - \tau(X)) \neq 0$. So the estimating equation derived from the ordinary least-squares method is not unbiased and hence generally it won't result in a consistent estimator of τ.

4.3 Simulation Studies

4.3.1 Design

In this numerical study, we adopted similar simulation design as simulation study (1) in Chapter 3. Two observed baseline covariates $X_1 \sim uniform(-3, 3)$ and $X_2 \sim Bernoulli(0 \cdot 4)$ act as confounders in the causal relationship of treatment A to outcome Y without other observed or unobserved confounders. The true propensity function is defined as

$$\begin{aligned} \pi(X) &= P(A = 1|X_1, X_2) \\ &= \frac{\exp(-2 + 0.4X_1^2 + 0.3log(X_2X_1^2 + 1)}{1 + \exp(-2 + 0.4X_1^2 + 0.3log(X_2X_1^2 + 1)} \end{aligned} \tag{4.7}$$

$A = 1$ associated with X was generated from $Bernoulli(\pi(X))$. Under this treatment assignment mechanism, the true propensity scores are expected to fall in the range of $(0 \cdot 1, 0 \cdot 9)$. To create a heterogeneous treatment effect scenario, given the treatment assignment indicator A and covariates (X_1, X_2), the outcome Y was generated based on the following model

$$Y = -2 + 1.2X_1^2 + X_1^3 X_2 + \tau(X_1, X_2)A + \epsilon \tag{4.8}$$

where the random error $\epsilon \sim N(0, 1.5)$ is independent of $X = (X_1, X_2)$. We especially designed the true conditional average treatment effect function $\tau(X_1, X_2)$ to reflect the general situation in real world applications when the treatment heterogeneity pattern is complicated and hard to predict: $\tau(X_1, X_2) = 1 - 4\exp(\frac{X_1}{2}) + 2 \cdot 5X_2(X_1 + 1) + 2X_1^2$. The true population ATE τ is given by $E(\tau(X_1, X_2)) = \int \tau(X_1, X_2)\ \mathrm{d}f(X_1, X_2)$. So true τ is

$$\tau = 1 - 4/3 \times \frac{exp(3/2) - exp(-3/2)}{6} + 7 \approx 2 \cdot 32$$

4.3.1.1 Cubic B-spline Approximation of $\tau(X_1, X_2)$

For the implementation of the first stage estimation of mean and propensity scores, we still adopted the cubic B-spline based sieve method. Cubic B-spline was only applied to X_1 as X_2 is a binary group indicator. The sequence of q_n inner knots $(a = \kappa_1 = \kappa_2 = \xi_3 = \kappa_4 < \kappa_5 < \cdots < \kappa_{q_n} < \kappa_{q_n+1} = \kappa_{q_n+2} = \kappa_{q_n+3} = \kappa_{q_n+4} = b; a <= X_1 <= b)$ for generating B-spline basis functions were placed at the $q_n - 3$ quantiles of X_1 with $q_n = \frac{\lceil n^{1/3} \rceil}{2}$. The mean and propensity scores were modeled the

same fashion as in homogeneous case

$$m(X) = \sum_{j=1}^{q_n} \alpha_j^{(1)} B_j(X_1) + \sum_{j=1}^{q_n} \alpha_j^{(2)} B_j(X_1) X_2$$

and

$$\log\left\{\frac{\pi(X)}{1-\pi(X)}\right\} = \sum_{j=1}^{q_n} \beta_j^{(1)} B_j(X_1) + \sum_{j=1}^{q_n} \beta_j^{(2)} B_j(X_1) X_2,$$

where $B_j(X)$ is the normalized B-spline basis functions at the knots κ_j for $j = 1, \cdots, q_n$.

With $\hat{m}(X), \hat{\pi}(X)$ obtained from the first stage, we examined the performance of sieve estimation of $\hat{\tau}(X_1, X_2)$ in the second step by modeling $\tau(X)$ as

$$\tau(X) = \sum_{j=1}^{q_n} \delta_j^{(1)} B_j(X_1) + \sum_{j=1}^{q_n} \delta_j^{(2)} B_j(X_1) X_2 \tag{4.9}$$

We present the comparison of true curves and the corresponding estimated $\hat{\tau}(X_1, X_2)$ for sample size 300 in Figure 4.1. To plot the true curves and the estimated functional curves shown in Figure 4.1, we first created a sequence of 300 data points $Z_1(i), i = 1, \cdots, 300$ in the interval $[-3, 3]$. A binary indicator Z_2 was created in two scenarios (i) all 1; (ii) all 0 with size 300. The true individual causal effects for each pair of $(Z_{1(i)}, Z_{2(i)} = 1)$ and $(Z_{1(i)}, Z_{2(i)} = 0)$ were evaluated based on $\tau(Z_1, Z_2)$, respectively. To obtain the estimated δ's, we then generated 1000 datasets each with size 300 and conducted the first stage estimation of mean scores and propensity scores with the inner knots number for cubic B-spline of X_1 being 3 within each simulated sample. Thus

$\hat{D}$ and $\hat{M}$ for each of the 1000 random samples are

$$\hat{D} = \begin{bmatrix} A_1 - \hat{\pi}(X_{1(1)}, X_{2(1)}) & & \\ & \ddots & \\ & & A_{300} - \hat{\pi}(X_{1(300)}, X_{2(300)}) \end{bmatrix};$$

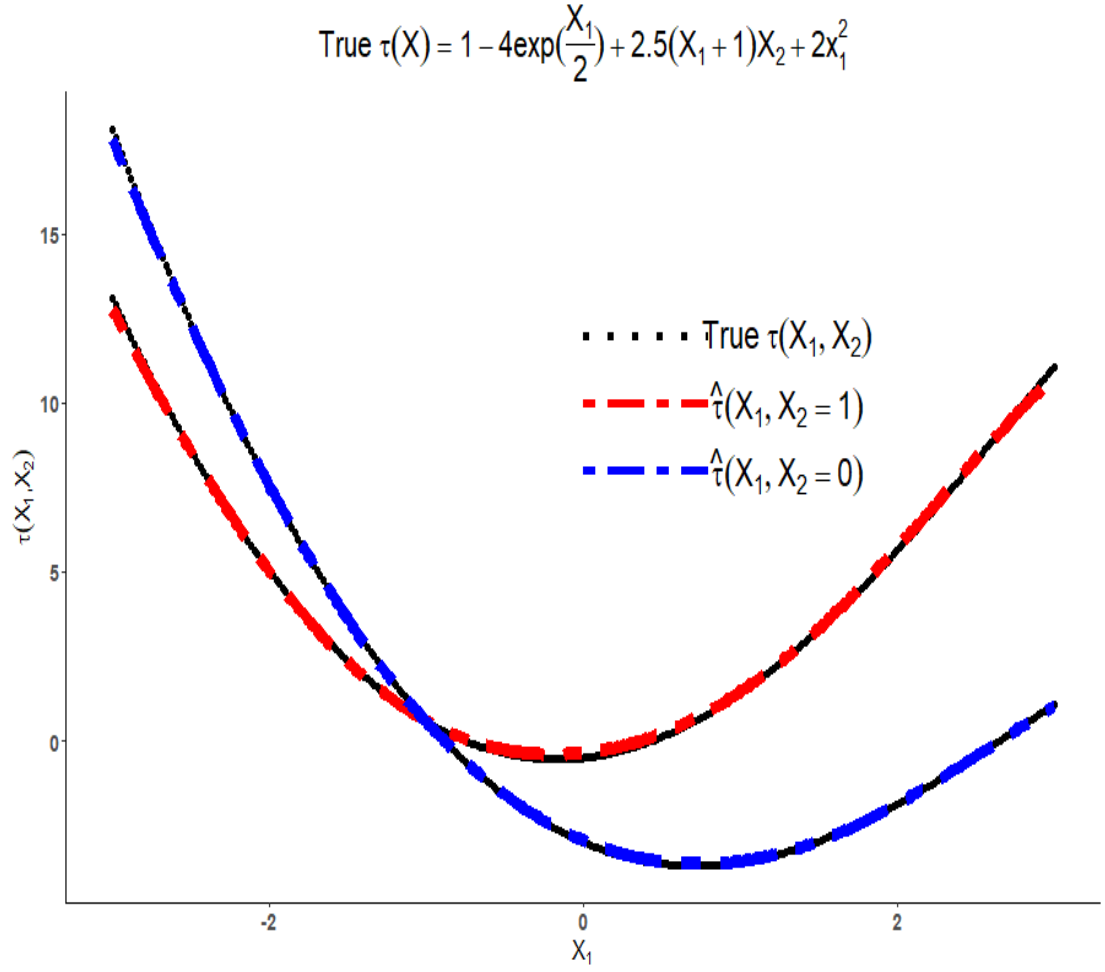
$$\hat{M} = (\hat{m}(X_{1(1)}, X_{2(1)}), \cdots, \hat{m}(X_{1(300)}, X_{2(300)})^T$$

For each sample, we obtained $(\hat{\delta}^{(1)}, \hat{\delta}^{(2)})$ for (4.9) based on $(\overline{B}^T \hat{D}^2 \overline{B})^- (\hat{D}\overline{B})^T (Y - \hat{M})$ where $\overline{B}$ is

$$\overline{B} = \begin{bmatrix} B_1(X_{1(1)}) & \dots & B_7(X_{1(1)}) & B_1(X_{1(1)})X_{2(1)} & \dots & B_7(X_{1(1)})X_{2(1)} \\ B_1(X_{1(2)}) & \dots & B_7(X_{1(2)}) & B_1(X_{1(2)})X_{2(2)} & \dots & B_7(X_{1(2)})X_{2(2)} \\ \dots & & & & & \\ B_1(X_{1(300)}) & \dots & B_7(X_{1(300)}) & B_1(X_{1(300)})X_{2(300)} & \dots & B_7(X_{1(300)})X_{2(300)} \end{bmatrix}$$

The predicted $\hat{\tau}(Z_1, Z_2)$ was then obtained as $\hat{\tau}(Z_1, Z_2) = \sum_{j=1}^{q_n} \hat{\delta}_j^{(1)} B_j(Z_1) + \sum_{j=1}^{q_n} \hat{\delta}_j^{(2)} B_j(Z_1) Z_2$. The average over the 1000 predicted $\hat{\tau}(Z_1, Z_2)$ were plotted vs. Z_1 for $Z_2 = 1$ and $Z_2 = 0$, respectively. In Figure 4.1, the black dots present the true curves of $\tau(Z_1, Z_2)$ and the red dashed line is for the estimated curve of $\hat{\tau}(Z_1, 1)$; the blue dashed line is for the curve of $\hat{\tau}(Z_1, 0)$. Figure 4.1 indicates that cubic B-spline estimator of $\tau(Z_1, Z_2)$ in Stage 2 results in quite consistent estimation of the true treatment effect function. As can be clearly seen in this figure, the estimated functional curves of $\hat{\tau}(Z_1, 1)$ and $\hat{\tau}(Z_1, 0)$ are nearly the same as their corresponding

FIGURE 4.1: Estimation of $\tau(X_1, X_2)$ with sample size 300



true curves. This means that (i) $\hat{\tau}(Z_1, Z_2)$ is a consistent estimator of the heterogeneity; (ii) $\frac{1}{300}\sum_{i=1}^{300}\hat{\tau}(Z_{1(i)}, Z_{2(i)})$, the empirical mean of $\hat{\tau}(Z_1, Z_2)$, is expected to be a consistent estimator of τ.

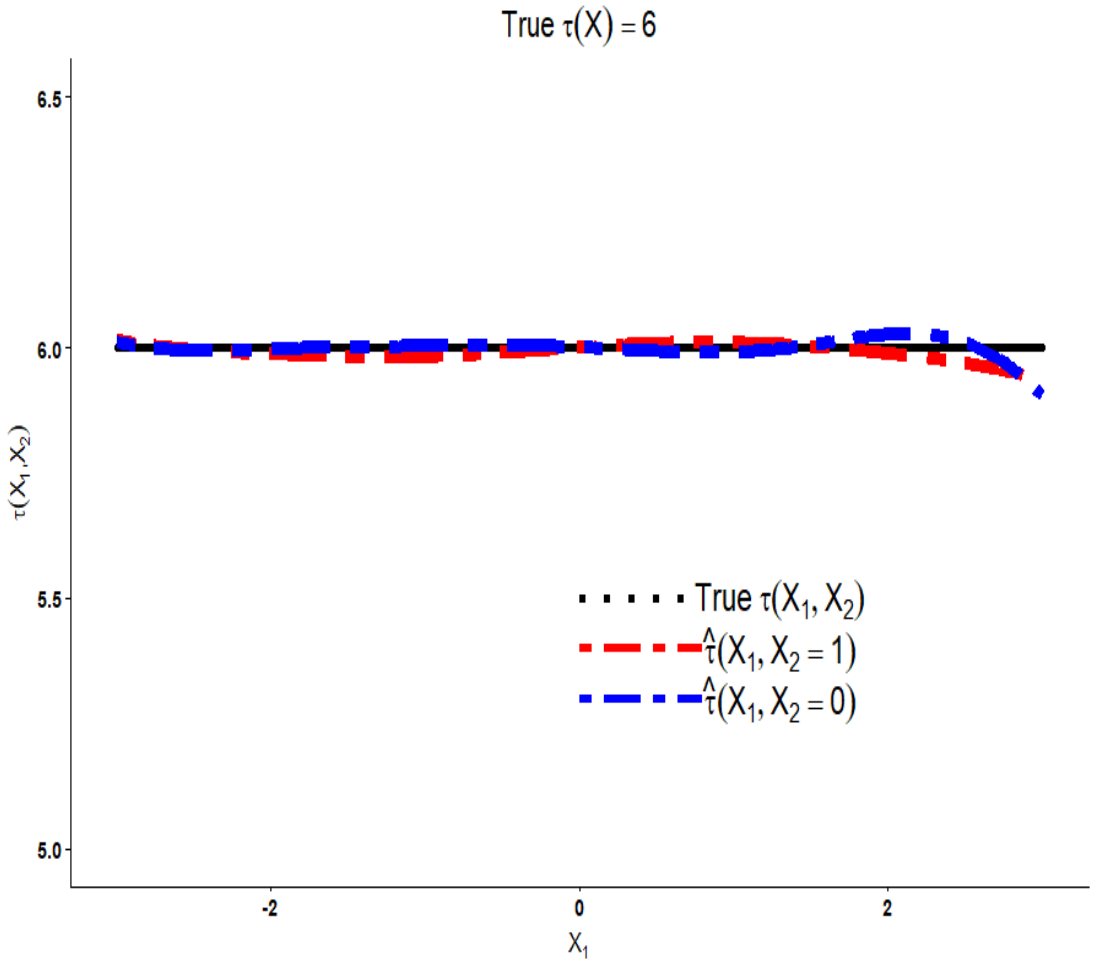
An interesting point regarding the sieve estimation of $\hat{\tau}(X)$ is whether it also works in homogeneous treatment effect scenario. Theoretically, homogeneity can be viewed as a special case of heterogeneity. Since in real applications the nature of true underlying treatment effect pattern is usually unknown, we may hesitate to make strong assumption of homogeneity. The question is when the true $\tau(X)$ is a constant but we compute $\hat{\tau}(X)$ following the three-stage estimation procedure for

dealing with treatment heterogeneity, will $\hat{\tau}(X)$ yield a consistent estimator of the true constant effect? We used the simulation study (1) introduced in Chapter 3 to gain some insights into this question. In simulation study (1), the true $\tau(X_1, X_2)$ was set to be a constant 6. To estimate $\hat{\tau}(X_1, X_2)$ (although now it is a constant) based on the three-stage procedure in this simulation setting, we adopted the same sets of $(Z_1(i), Z_{2(i)} = 1), (Z_1(i), Z_{2(i)} = 0), i = 1, \cdots, 300$ and modeling strategies for creating Figure 4.1 and presented the two estimated curves in Figure 4.2. The true $\tau(Z_1, Z_2 = 1)$ and true $\tau(Z_1, Z_2 = 0)$ are 6 as depicted by the black horizontal line in Figure 4.2. The red and blue dashed lines were plotted based on the estimated $\hat{\tau}(Z_1, 1)$ and $\hat{\tau}(Z_1, 0)$, respectively. As shown in Figure 4.2, the black, red and blue lines overlap with each other very well except for some slight discrepancies at the two ends of X_1. The average of $\hat{\tau}(Z_1, 1)$ and $\hat{\tau}(Z_1, 0)$ are reasonably close to the true 6, 6.023 and 6.006, respectively. This experiment gives empirical evidence that in general we may employ the three-stage estimation procedure for the purpose of estimating average treatment effect in both homogeneous and heterogeneous treatment effect scenarios.

4.3.1.2 Estimators of τ from Various Competing Methods

The same set of competing methods described in Chapter 3 were used to compare the performance of our proposed estimator with other conventional methods. The hypothetical scenario that the outcome model (4.8)is known including the true function form of $\tau(X_1, X_2)$ and we simply performed the maximum likelihood estimation method was used as the benchmark for comparison as it gives the efficient estimator of τ. For IPW methods, we considered two scenarios (i) true parametric

FIGURE 4.2: Spline-based sieve estimator of $\tau(X_1, X_2)$ when the true $\tau(X_1, X_2)$ is constant



propensity model (IPW-pT) given in (4.7); (ii) wrongly specified ordinary logistic regression model (IPW-pW) with covariates X_1, X_2 and the interaction term X_1X_2. For AIPW, we again examined all four possible scenarios (i) AIPW-mT&pT: both mean and propensity scores where modeled correctly by (4.8) and (4.7), respectively; (ii) AIPW-mT&pW: the mean score model was specified correctly as given in (4.8) but the propensity score model was specified wrongly as for the IPW-pW estimator; (iii) AIPW-mW&pT: the propensity score model was specified correctly as given in (4.7) but the mean score was wrongly specified as the ordinary linear regression model with covariates X_1, X_2, X_1X_2, AX_1, AX_2 and AX_1X_2; (iv) AIPW-mW&pW: both the mean and propensity score models were wrongly specified as aforementioned. Simulation results for sample size 200, 400, and 800 are presented in Table 4.1, in which we summarized Bias, M-C SD based on 1000 repetitions, ASE based on 100 bootstrap samples, and 95% CP of the Wald 95% confidence interval. Our proposed method was implemented in two ways: (i) MF(HME): estimating τ ignoring heterogeneity, i.e. treating it as homogeneous treatment effect scenario; (ii) MF(HTE): modeling $\tau(X_1, X_2)$ by (4.9).

Firstly, for our proposed method, Table 4.1 suggests that simply ignoring heterogeneity by assuming $\tau(X_1, X_2)$ to be constant (MF(HME)) results in much larger bias and variance than the estimators from modeling $\tau(X_1, X_2)$ nonparametrically (i.e, MF(HTE)). This is not surprising since the nonparametric estimator of $\tau(X_1, X_2)$ is consistent that it has a small bias for a wide range of underlying regression functions, but a constant estimator for $\tau(X_1, X_2)$ will only have a low bias if the assumption of homogeneity holds i.e., $\tau(X_1, X_2) = \tau$, and can otherwise suffer from large bias. Apparently, increasing sample size did not help for the bias. So when

there exists heterogeneity MF(HME) is not a consistent estimator as shown in Figure 4.3 and its the coverage probability of 95% CI is totally off from 0·95. In contrast, Table 4.1 indicates that MF(HTE) yields consistent estimator of τ with both bias and variability only slightly inferior to MLE and AIPW with correctly specified mean model. When sample size is large, the ASE and the M-C SD are quite close with the coverage probability of 95% CI fairly close to 0·95. Asymptotic normality property of MF(HTE) was also clearly shown in Figure 4.3.

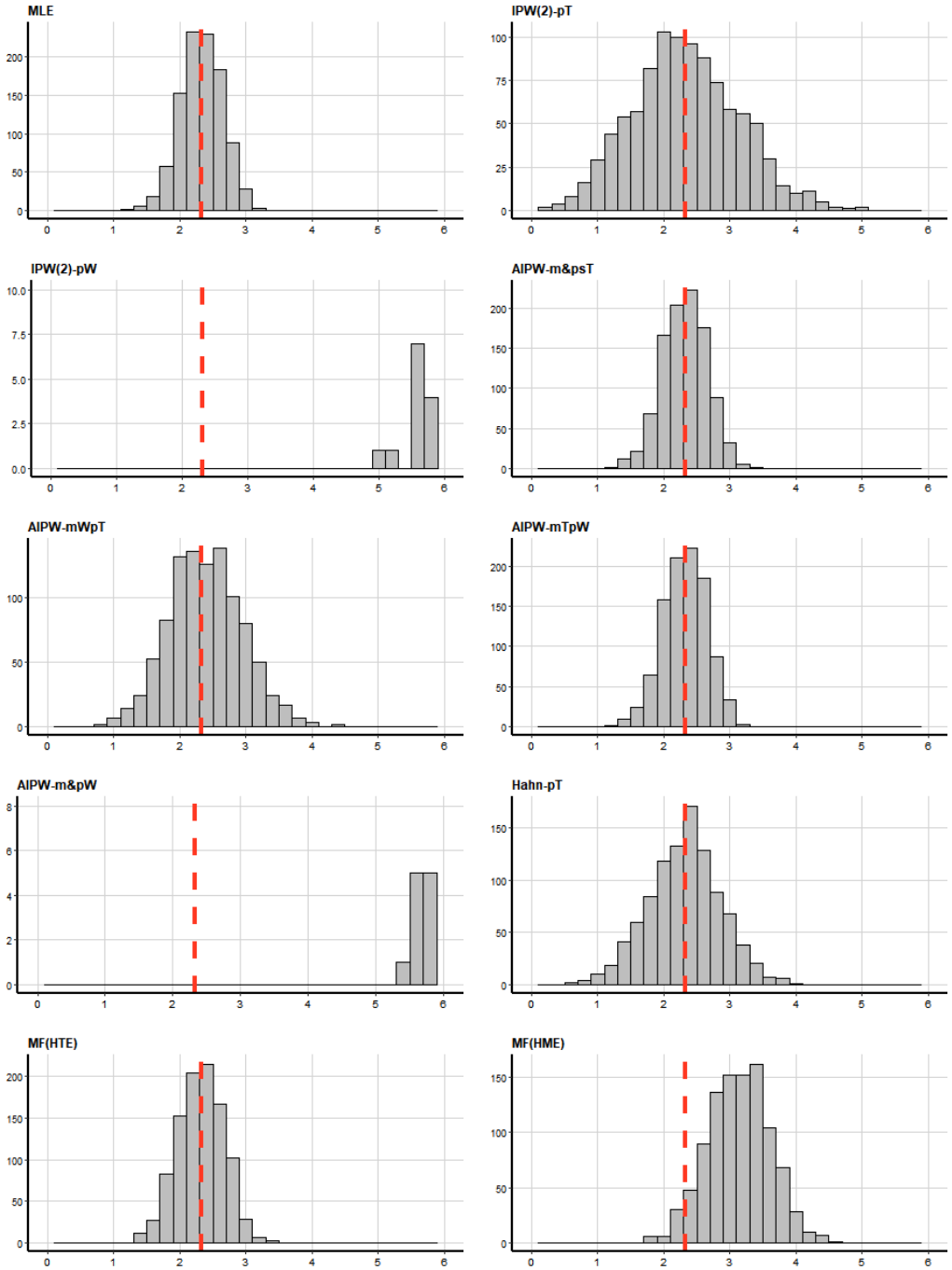
Secondly, the IPW type methods including simple IPW, AIPW, Hahn's method, display the same pattern of bias, variability, coverage probability as in Table 3.1 for the homogeneous treatment effect case: with correctly specified propensity model, IPW resulted in unbiased estimators but are severely biased and not reliable when propensity model was misspecified. Although in our simulation design the true propensity scores are not expected to be extreme, we can still see much higher variability (comparing to MLE and MF(HTE)) even when the propensity score was correctly specified, as indicated in Figure 4.3. The AIPW method with correctly specified mean model performed much better than the IPW method and was helpful in reducing the high variability due to inverse weighting. AIPW estimators with both misspecified models for mean and propensity scores are severely biased just as in IPW-pW. In this simulation setting, Hahn's methods seemed to perform well with propensity scores estimated from correctly specified treatment assignment model and outcome modeled nonparametrically. However, comparing to our proposed estimator, it still gave quite larger variability due to inverse weighting.

TABLE 4.1: Summary table of results from simulation study with heterogeneous treatment effect: comparison of bias, Monte Carlo standard deviation, asymptotic standard error and 95% coverage probability among all the methods

	Bias			**M-C SD**			**ASE**			95% **CP**		
	N=200	N=400	N=800	N=200	N=400	N=800	N=200	N=400	N=800	N=200	N=400	N=800
IPW(1)-pT	0·013	0·007	0·011	1·143	0·807	0·555	1·141	0·796	0·559	0·945	0·953	0·941
IPW(1)-pW	5·565	5·562	5·504	1·364	0·896	0·601	2·519	0·969	0·621	0·095	0·000	0·000
IPW(2)-pT	0·027	0·009	0·014	1·167	0·818	0·565	1·162	0·810	0·568	0·949	0·952	0·946
IPW(2)-pW	5·425	5·503	5·476	1·273	0·873	0·599	1·491	0·916	0·610	0·029	0·000	0·000
IPW(3)-pT	0·033	0·015	0·017	1·148	0·806	0·556	1·130	0·795	0·559	0·945	0·954	0·943
IPW(3)-pW	5·409	5·497	5·476	1·181	0·852	0·594	1·244	0·862	0·599	0·002	0·000	0·000
AIPW-p&mT	0·004	0·014	0·002	0·440	0·328	0·238	0·495	0·325	0·227	0·951	0·953	0·935
AIPW-pT&mW	0·004	0·014	0·003	0·435	0·326	0·237	0·494	0·322	0·225	0·957	0·952	0·928
AIPW-pW&mT	0·138	0·066	0·034	0·867	0·571	0·397	0·944	0·591	0·400	0·964	0·966	0·949
AIPW-pW&mW	5·518	5·522	5·495	1·128	0·786	0·562	1·388	0·795	0·548	0·029	0·000	0·000
HNN(pT)	0·017	0·001	0·002	0·777	0·559	0·384	0·803	0·546	0·379	0·946	0·951	0·934
HNN(Tp)	0·024	0·013	0·006	0·874	0·642	0·460	0·856	0·623	0·442	0·937	0·941	0·931
MLE	0·006	0·016	0·002	0·411	0·305	0·221	0·435	0·303	0·213	0·955	0·941	0·940
MF(HTE)	0·011	0·013	0·003	0·498	0·338	0·244	0·990	0·389	0·235	0·993	0·970	0·941
MF(HME)	0·747	0·827	0·861	0·694	0·497	0·329	0·704	0·478	0·332	0·824	0·593	0·252

M-C SD: Monte Carlo standard deviation from 1000 iterations
ASE: standard deviation from 100 bootstrapping samples
95% CP: 95% coverage probability

FIGURE 4.3: Distribution of Estimators from Various Methods for the simulation study with heterogeneous treatment effect: True $\tau \approxeq 2.32$



4.4 Application of Model-Free Method in Estimating Heterogeneous Treatment Effect in JIA Study

In Chapter 3, we described the exploratory analysis of estimating the treatment effect of early aggressive use of biologic DMARDs vs. the Step-up treatment plan among the eligible 327 pcJIA pediatric patients with complete 6-month cJADAs outcome. Specifically, we assumed that the effect of early combination of biologic DMARDs and non-biologic DMARDs in treating JIA is homogeneous across the study sample and estimated the causal effect based on two clinical relevant baseline characteristics: cJADAs, binary indicator of low pain (pain ≤ 3). Under the homogeneous treatment effect assumption, our propose model-free method gives an effect of -1.271 with 95% confidence interval $(-2.648, 0.106)$. Since in reality we are not certain that this assumption of homogeneity holds, it is desirable to apply our extended model-free method discussed in this chapter to infer the heterogeneity in the treatment effects. We explored the effectiveness of early aggressive use of biologic DMARDs by addressing two questions: (i) whether patients with different baseline disease severity are likely to respond differently to biologic DMARDs treated at early stage ; (ii) by taking into the fact of heterogeneity whether or how much the estimated ATE will differ?

We still considered the two covariates cJADAs and low pain indicator in examining the potential heterogeneous pattern of treatment effect. Therefore, we modeled the mean and propensity scores in the same way as in the simulation study discussed in section 4.3. The cubic B- spline of cJADAs was constructed exactly the same as described in Chapter 3: 3 interior knots were placed at 25%, 50%, 75% percentiles of

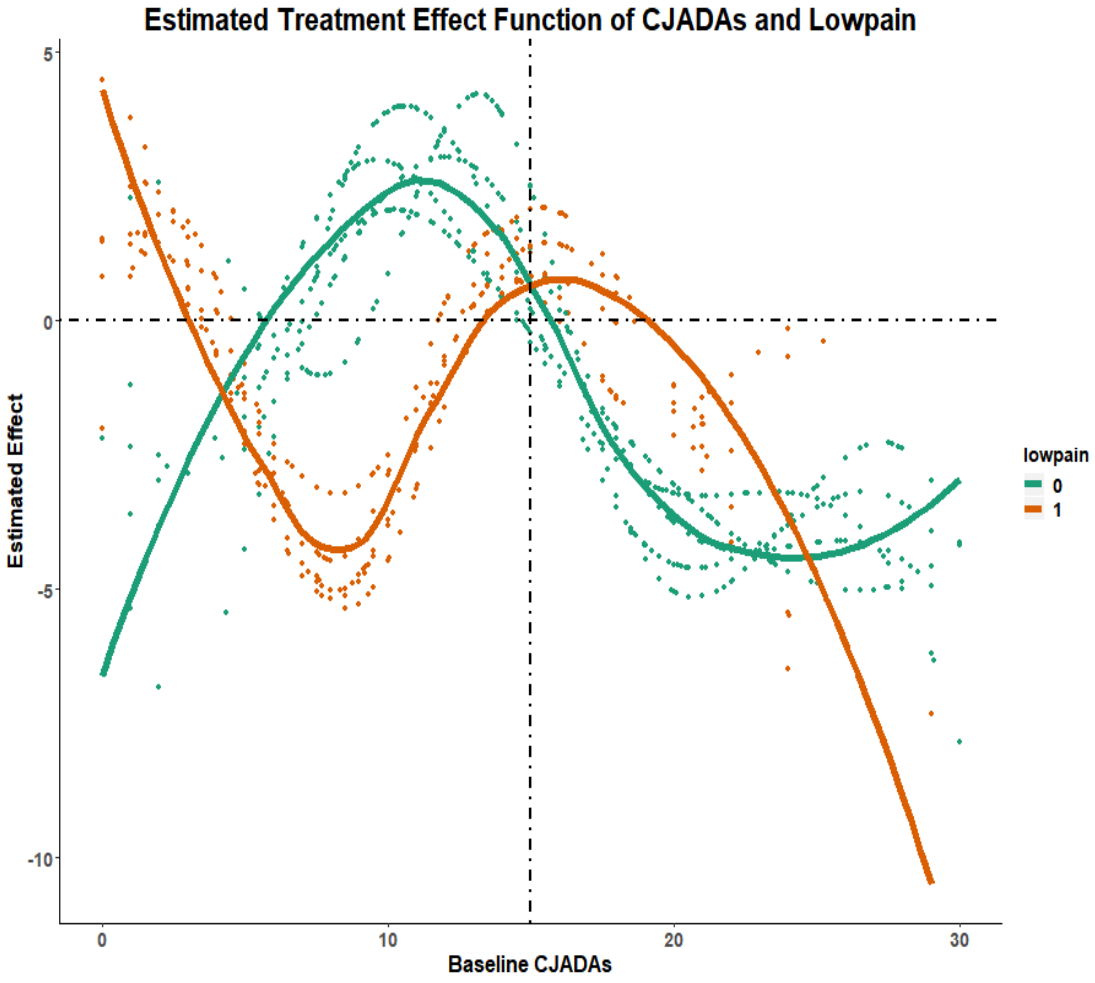
the baseline cJADAS. For each of the 5 imputed datasets, we conducted the spline estimation of $\tau(X)$ at the second stage using three-stage estimation procedure and presented all the estimated individual treatment effects based on the 5 datasets in Figure 4.3 with red dots denoting patients from the low pain group and green dots for patients from the moderate to severe pain group. The green line and red line are the smoothing curves over the green dots and red dots, respectively, using LOESS technique (with smoothing parameter span defined as 0.75). The vertical black dashed line in Figure 4.3 represents median value of cJADAs (score 15) and the horizontal black dashed line is for borderline effect (0).

Figure 4.3 implies that the treatment effect is likely to be heterogeneous in terms of baseline cJADAs and severity of pain among the study patients. The average treatment effect for the low pain group was estimated to be about -1.47 while for its reference group about -0.97. The green dots and red dots clearly display different patterns in term of baseline cJADAs. When cJADAs is above median (15 or more), both the red line and green line are below the horizontal reference line and going down as cJADAs increases, which indicates that both groups benefit from the early use of biologic DMARDs if their baseline cJADAs score is 15+ and the effect is seen to be even more pronounced with higher cJADAs scores for patients in low pain. However, when cJADAs is below 15, the trend of treatment effects are quite different and tells different story for the two groups. Early use of biologic DMARDs started to show effect to reduce the cJADAs for the low pain group when their baseline cJADAs score is 4 and above and for those patients experiencing low pain and cJADAs score around 7 to 10, the benefit from biologic DMARDs is likely to reach the maximum of about -5. For the patients in the moderate to high group, it seems to be opposite as most

of the green dots locate above the horizonatal reference line of border effect when the baseline cJADAs is between 7 and 15, which is interesting. We can also see from this plot that the moderate to high pain group also tended to have extreme larger baseline cJADAs scores (> 25) while patients in the low pain group are more likely to have baseline cJADAs scores < 5 so data is sparse at the two ends for either groups although the distribution of low pain show good balance when cJADAs is between 5 and 25. As our analysis was limited to these two covariates and the sample size is only around 300, not very large, care might need to be taken to interpret these findings of treatment heterogeneity in terms of baseline cJADAs. Nevertheless, there is an indication based on this analysis that patients with different baseline disease severity might respond to biologic DMARDs differently. At lower level of the baseline cJADAs, patients experiencing no or mild pain will likely to benefit more from early use of biologic DMARDs in treating their JIA disease while at high level of cJADAs this benefit seems to be more sounding for patients who experienced moderate to high pain.

To make statistical inference about the average treatment effect while taking into the consideration of heterogeneity, we implemented the three stage estimation procedure for each of the 5 imputed datasets and used 500 bootstrap samples to obtain standard errors. The result shows that with modeling of treatment heterogeneity in baseline cJADAS low pain, the population average treatment effect was a bit smaller than the estimator assuming homogeneous effect: -1.154. The estimated standard error is round 0.732, slightly larger than that in homogeneous effect scenario. The corresponding 95% confidence interval is $(-2.589, 0.281)$.

FIGURE 4.4: Estimated functional (in terms of baseline cJADAs and low pain) curves of heterogeneous treatment effects



CHAPTER 5

Application of OLS-based Method in Observational Studies with A Large Set of Covariates

5.1 Curse of Dimensionality and Machine Learning Methods in Causal Inference

It was emphasized in the previous analyses of JIA study that only two baseline characteristics were incorporated in the estimation of the treatment effect of early combination of biological DMARDs. As introduced in Chapter 3, in addition to these two selected covariates, there are many more baseline characteristics that could be potential confounders and modifiers of the treatment effect. A more accurate estimator of treatment effect relies on a scrutinization of the complete set of confounders and treatment effect modifiers. With many baseline characteristics taken into consideration, it is not feasible to apply traditional nonparametric methods as discussed in the two-stage or three-stage estimation procedure in order to estimate the treatment effect. Due to curse of dimensionality, this limitation does not just apply to our proposed model-free method but also to all methodologies involving nonparametric estimation.

This curse of dimensionality in nonparametric estimation can be illustrated using the data from JIA study. With baseline cJADAs and binary indicator of low pain, in the cubic B-spline regression with 3 interior knots for estimating mean scores (also propensity scores and treatment effect function), there are $7+7 = 14$ spline basis

functions constructed so 14 coefficients needs to be estimated based on 327 observed samples. Suppose now we are knowledgeable that another measure of baseline disease severity, loss motion of range (lom for short), is also an important counfounder and better to be taken into account in the propensity score estimation. If cubic B-splines are to be applied with 3 interior knots for lom, there are 7 more basis functions from this additional covariate. And if we would like to consider the pairwise interactions (tensor splines) among cJADAs, lom, low pain indicator, there will be $14 \times 7 = 98$ coefficients in total to be estimated based on the 327 observed samples. As can be imagined, with more covariates added into the nonparametric estimation process, the dimension of model parameter space increases in a polynomial order and causes computing trouble with a moderate data set in traditional MLE or OLS approaches.

For modeling with large number of covariates, one often considers generalized additive models (Hastie and Tibshirani, 1986) as it compromises the curse of dimensionality in nonparametric modeling techniques. Suppose we want to make inference about an unknown function f that predicts the average treatment effect τ using a p dimensional vector of inputs $X = (x_1, \cdots, x_p)$, i.e

$$\tau(X) = f(x_1, \cdots, x_p)$$

For an additive model, $f(x_1, \cdots, x_p)$ is modeled by

$$f(x_1, \cdots, x_p) = g_1(x_1) + \cdots + g_p(x_p)$$

With the additive model, the dimension of model parameter space would not grow in a polynomial order of the number of covariates. Hence it may make the nonparametric estimation implementable. In the above example, if model $\tau(X)$ is modeled additively with the 3 covariates: cJADAs, lom and low pain. With the same strategy of applying cubic B-splines to cJADAs and lom, the total coefficients to be estimated will be reduced to $7 + 7 + 1 = 15$, which is much smaller than 98 and make the computation much easier.

Alternatively, some machine learning methodologies can be adopted for case studies with large number of covariates to obtain practical solutions. Machine learning is a modern technology that has been widely used in the statistical analysis including prediction, classification, learning association, regression etc. for high dimensional data. This thesis is not intended to give a thorough discussion regarding the topics of machine learning or developing a new approach. Rather, we focus on the application of a couple of existing popular machine learning methods and discuss how their strengths in dealing with a large set of covariates can be borrowed into our proposed model-free method in estimating the two summary scores. Compared to the additive modeling approach, machine learning methods are more powerful as they typically also take care of interactions among a large set of covariates. For making prediction in both regression and classification problems, tree-based machine learning methods have become quite popular due to its flexibility in fitting interactions and nonlinearities. Many tree-based algorithms, for example, boosting (Freund and Schapire, 1997; Friedman, 2001) , random forests (Breiman, 2001) and bagging (Breiman, 1996), generalized boosted models or gradient boosting machine (GBM) (Natekin and Knoll, 2013; Ridgeway, 2005), Bayesian additive regression tree (BART) (Chipman et al.,

2010) have been developed with computing softwares available in R. We choose BART and GBM for study in this thesis and briefly introduce the two methods as follows.

- **BART** is a Bayesian method for estimating a nonparametric function using sums of regression trees. Consider the problem of nonparametric estimation of some unknown function f in the regression problem of $Y = f(Z) + u$ given in (2.1) in Chapter 2 where $var(u) = \sigma^2 < \infty$ and $Z = (z_1, \cdots, z_p)$ is a p dimension of predictor space. BART models $f(Z)$ as a sum of distinct m trees

$$f(z) = \sum_{i=1}^{m} g_i(z; T, H)$$

where each function $g_i(z; T, H)$ represents a binary tree whose structure denoted by T and terminal nodes (leaves) given by $H = \{\mu_1, \cdots, \mu_l\}$. Each tree $g_i(z; T, H)$ specifies how an observation goes through according to the splitting rules (typically in the form of $z_p \leq c; p = 1, \cdots, d$ with c being a threshold value within the range of values of z_p) defined by its internal nodes until reaching its terminal node. In other words, each tree $g_i(Z; T, H)$ contains information of how an observation should be partitioned and stops at a certain terminal value. The sum of all such trees, i.e $\sum_{i=1}^{m} g_i(z; T, H)$ is also called an ensemble-of-trees. Different from other ensemble-tree-based machine learning methods, BART adopts three Bayesian nonparametric priors for T, H and σ^2, respectively. The joint prior probability can be expressed as

$$\begin{aligned} P(g_i(z; T, H), \cdots, g_m(z; T, H), \sigma^2) &= \left[\prod_{i=1}^{m} P(H^i|T^i)P(T^i)\right] P(\sigma^2) \\ &= \left[\prod_{i=1}^{m}\prod_{h=1}^{l} P(H_h^i|T^i)P(T^i)\right] P(\sigma^2) \end{aligned}$$

$P(T^i)$ controls the complexity of the ith tree with two important parameters: the relative location of a nonterminal node to the root, also called the depth; $(P_1, \cdots, P_p); \sum_{j=1}^{p} P_j = 1$, a vector of specified probabilities of being chosen for serving as the splitting variable among all the available predictors. $P(H^i|T^i)$ defines the prior for leaf parameters and in regression problem it typically adopts a normal distribution resulting in the "best guess" of $\hat{y} = \mu_l$ in the partition of predictor space. The prior for error variance $P(\sigma^2)$ is often chosen to shrink $P(H^i|T^i)$ towards the center of the distribution of y and plays a role in model regularization. BART relies on Markov Chain Monte Carlo (MCMC) technique for sampling from posterior distributions. At each iteration step, given the three sets of priors the algorithm specifies the likelihood of y in the leaves as $y \sim N(\mu_l, \hat{\sigma}^2)$ where μ_l is the current best fit of partition and $\hat{\sigma}^2$ is the current best guess of the variance. With the total number of trees to be fitted being m, to obtain posterior distribution of $P(g_i(z; T, H), \cdots, g_m(z; T, H), \sigma^2|y)$ and make prediction, BART uses the Gibbs sampler (Geman and Geman, 1984) to conduct Bayesian back-fitting (Hastie and Tibshirani, 2000), in which the ith tree fitted iteratively while keeping the other $m - 1$ tress fixed to construct partial residuals that explained only by the ith tree. Then by drawing a large number of MCMC samples over the ensemble-of-trees model space evaluated at z_j, a posterior mean estimate of the target function f for a given value of z_j can be obtained by taking the average of these draws. Detailed discussion regarding BART algorithm, how the MCMC is implemented as well as how BART deals with classification problem can be found in (Chipman et al., 2010; George and Jensen, 2014; Kapelner and Bleich, 2016). Some popular R packages

for BART include $\{BART\}$ (McCulloch et al., 2019), $\{bartMachine\}$ (Kapelner and Bleich, 2018).

- **GBM** is a popular tree-based machine learning method using gradient-descent based boosting algorithms (Freund and Schapire, 1997; Friedman, 2001). Fundamentally, estimation of $f(Z)$ in the above nonparametric regression problem based on boosting is achieved by minimizing some type of specified loss function, such as the classical L_2 squared loss function when y is continuous. Using the notation in (Natekin and Knoll, 2013), $\hat{f}(z) = \arg\min_{f(z)} E_z\left[E_y(\Psi(y, f(z))|z\right]$ where $\Psi(y, f)$ is the loss function. Like in BART, the unknown function f is first parametrized in the additive form of $\sum_{i=0}^{m} f_i(z)$ with $f_0(z)$ being the initial guess and m the number of total iterations called boosts. To initiate the iterative process, a base-learner function denoted as $h(z, \theta)$ is first defined based on which the successive functional increments are constructed through greedy search. The optimization rule for the function estimate at the dth iteration is defined by $\hat{f}_{d-1} + \rho_d h(z, \theta_d) \to \hat{f}_d$ and (ρ_d, θ_d) are estimated through optimization

$$\arg\min_{\rho,\theta} \sum_{i=1}^{N} \Psi(y_i, \hat{f}_{d-1}(z_i)) + \rho h(z_i, \theta)$$

where $\Psi(y_i, \hat{f}_{d-1}(z_i))$ are the residuals from the previous step of fitting. In the implementation of the gradient-descent boosting algorithm, particularly we first evaluate the negative gradient $\{g_d(z_i)\}$ over the observed data points of $(z_i, y_i); i = 1, \cdots, n$ as

$$g_d(z) = E_z\left[\frac{\partial \Psi(y, f(z))}{\partial f(z)}|z\right]_{f(z)=\hat{f}_{(d-1)(z)}}$$

and then choose a new function $h(z, \theta_d)$ such that

$$(\rho_d, \theta_d) = \arg\min_{\rho,\theta} \sum_{i=1}^{n} -g_d(z_i) + \rho h(z_i, \theta)$$

Simply speaking, GBM repetitively trains the residuals from the previous step of fitting and improve the overall fit by adding new increments to the ensemble of trees sequentially. For the classification problems using GBM, the rational behind the iterative process is similar as in regression problem except that the loss function should be defined appropriately for categorical outcome. Challenges of applying GBM in real application arise as it is tricky to choose the tuning hyperparameters for optimal fit and prediction. The most commonly used tuning hyperparameters in GBM implementations include the total number of trees to fit, the depth of trees controlling the complexity of the boosted ensemble, the learning rate or shrinkage controlling the speed the algorithm proceeds down the gradient descent and subsampling controlling whether or not a fraction of the available training observations is used for fitting. There are also a few choices of software in R for implementing GBM, such as $\{gbm\}$ (Hijmans et al., 2019), $\{dismo\}$ (Hijmans et al., 2017).

BART or GBM can be used in multiple ways for the purpose of estimating treatment effect depending on the methods we shall adopt to solve the problem. One intuitive way is to use them in the outcome regression and based on the idea of G-computation to make prediction for $E(Y^1)$ and $E(Y^0)$, respectively, and then compute $\hat{\tau} = \hat{E}(Y^1) - \hat{E}(Y^0)$. That is, we consider applying BART or GBM to estimate an unknown function g that predicts outcome Y using the treatment indicator and a p di-

mensional vector of inputs $X = (x_1, \cdots, x_p)$ (i.e, $E(Y|X, A) = g(A, x_1, \cdots, x_p)$). Alternatively, both methods can be applied to estimate an unknown function f that estimates the propensity scores using a q dimensional vector of inputs $X^* = (x_1, \cdots, x_q)$ (i.e, $P(A = 1|X^*) = E(A|X^*) = f(x_1, \cdots, x_q)$). X could be the same as X^* but not necessarily. The non-parametrically estimator $\hat{P}(A = 1|X^*)$ can then be fed into either IPW or AIPW methods to compute $\hat{\tau}$. It is straightforward that both of BART and GBM can also be incorporated easily into the first stage of our proposed model free method to estimate mean and propensity scores, the only difference is that to estimate mean scores, we seek to estimate an unknown function g^* such that $E(Y|X) = g^*(x_1, \cdots, x_p)$. So in general applying our proposed method to studies involving large set of covariates X will necessarily include the following steps:

(i) using the existing machine learning methods to compute $\hat{m}(X)$ and $\hat{\pi}(X)$.

(ii) determining the set of covariates X^o used for the estimation of $\tau(X^o)$ based on prior knowledge of potential treatment modifiers. In this step, to avoid computation difficulties, we will apply additive models or at most consider only the interactions of some binary covariates and continuous covariates.

(iii) obtaining the final estimator of average treatment effect $\hat{\tau}$ by taking average of $\hat{\tau}(X^o)$.

(iv) computing the standard error through bootstrapping to make statistical inference.

5.2 Application of Various Machine Learning Methods in JIA Study

The purpose of this final analysis discussed in this Chapter is to obtain a more accurate estimator of the treatment effect of using early combination of biologic

DMARDs compared to the step-up treatment plan by considering all the baseline characteristics listed in Table 5.1 in addition to baseline cJADAs and pain. P values in Table 5.1 are from Wilcoxon test (for continuous variables) or Fisher's exact test (for categorical). These baseline characteristics include baseline age (age), indicator of private insurance (private), gender (Female), ANA positive indicator (ANA positive), loss range of motion (lom), race (white), well being, md global assessment (md assessment), active joint count (AJC), esr, indicator of B27 being positive (B27), indicator of rheumatoid factor positive (RF positive), time to diagnosis relative to baseline visit (timediag), morning stiff levels (1: no stiffness, 2: 15 min or less, 3: $>$ 15 minutes) (ms1, ms2), Jra disease subtypes (1: RF (-), 2: RF (+), 3: oligoarticular, 4: other) (subtype1, subtype2, subtype3). We used the widely used R packages $\{bartMachine\}$, $\{gbm\}$ to implement the BART, GBM methods in the estimation of treatment effect of interest for JIA study. Equipped with these tools, we can easily incorporate all the baseline characteristics listed above into the estimation process. In addition, for the propensity score estimation we also used the R package $\{CBPS\}$ (Fong et al., 2019) to obtain the so-called covariate balancing propensity scores, which are not based on machine learning methods but have the desired property of providing optimal balancing for the covariate being considered.

5.2.1 Notation for Estimators from Various Methods

In order to compare the estimators among various methods, we consider the following combinations:

- Outcome regression from BART and GBM, denoted as BART-G and GBM-G, respectively.

TABLE 5.1: Baseline characteristics of the JIA study population by the two treatment arms

Baseline Covariates	Step-up	Early Combination	P value
Age	10.2 ± 5.07	10.2 ± 4.61	0.88
Female	161 (71.6%)	74 (72.5%)	0.89
White	195(86.7%)	91(89.2%)	0.59
Private insurance	153(68%)	65(63.7%)	0.45
Jra subtype	Polyarticular RF (-) 79(35.1%)	Polyarticular RF (-) 46(45.1%)	0.3
	Polyarticular RF (+) 17(7.6%)	Polyarticular RF (+) 9(8.8%)	
	oligoarticular 68(30.2%)	oligoarticular 24(23.5%)	
	other 61(27.1%)	other 23(22.5%)	
Time to diagnosis (in month)	1.30 ± 1.93	1.41 ± 2.16	0.58
Active joint count	4.92 ± 3.58	6.13 ± 3.72	0.01
Wellbeing	3.50 ± 2.42	4.36 ± 2.63	0.007
MD Assessment	4.37 ± 2.46	4.96 ± 2.84	0.12
Morning stiffness	no stiffness 65(28.9%)	no stiffness 17(16.7%)	0.04
	15 min or less 35(15.6%)	15 min or less 15(14.7%)	
	$>$ 15 minutes 125(55.6%)	$>$ 15 minutes 70(68.6%)	
Loss range of motion	5.60 ± 7.19	10.1 ± 12.2	0.0001
Esr	20.6 ± 19.0	27.2 ± 26.1	0.06
Rheumatoid Factor postive	14(6.2%)	13 (12.7%)	0.05
Antinuclear Antibodies positve	17(7.6%)	15(14.7%)	0.07
HLA-B27(Yes)	9(4%)	9(8.8%)	0.11

- IPW-psCB, IPW-psBART, IPW-psGBM with propensity scores from CBPS, BART, GBM, respectively.
- All the combinations of methods in outcome and propensity score estimation for AIPW: for example AIPW-BART-psCB denotes the AIPW with the outcome modeled with BART and propensity scores with CBPS. The others options are: AIPW-BART-psBART, AIPW-BART-psGBM, AIPW-GBM-psCB, AIPW-GBM-psBART, AIPW-GBM-psGBM.
- As for our proposed method, we considered heterogeneous treatment effects along with all the combinations of different methods in mean scores and propensity scores estimation. So we came up with 6 estimators: MF(BART-psCB), MF(BART-psBART), MF(BART-psGBM), MF(GBM-psCB), MF(GBM-psBART) and MF(GBM-psGBM). For example, MF(BART-psCB) denotes the method with mean scores based on BART, propensity score from CBPS.

  To define the set of covariates for $\tau(X)$, in addition to baseline cJADAs and low pain indicator considered in the previous analyses of heterogeneous treatment effects, we added in two more variables: the actual follow-up time (denoted as diffVisits) for the assessment of 6-month outcome and the morning stiffness. The 6-month outcome assessment time for each study patient was defined as the closest visit time to 6 month after baseline visit within the window of 4 to 8 months and if a patient didn't have follow-up visit within this window but had 3 month visit, we then imputed the 6-month outcome by carrying over the 3-month outcome forward. As a result, the 6-month outcome assessment time varies across the study sample. Out of the 327 patients with defined 6-month cJADAs score, it was found that about 75% of them actually were

assessed before 6 month with about 25 % were even around 3 months. Since it is highly likely that the treatment effect of early use of biological DMARDs is time dependent, we consider the diffVisits as a potentially important variable for examining treatment effect heterogeneity. Morning stiffness is also an important measure of disease severity that may associate with the effect of biological DMARDs. With three levels of morning stiffness, one dummy variable was created as $I(MF = 3)$. Thus, we have two continuous variables cJADAs and diffVisits, two binary indicators including low pain, morning stiffness level being 3. We adopted the following modeling techniques in the spline estimation of $\tau(X)$

$$E(\tau|x_1, x_2, x_3, x_4) = f_1(x_1) + f_2(x_2) + \alpha x_3 + f_3(x_1) \times x_4$$

where $f_1(x_1)$ and $f_2(x_2)$ are cubic B-spline expansion of cJADAs and diffVisits, respectively. x_3 and x_4 are dummy variables of $I(MF = 3)$ and $I(pain \leq 3)$, respectively. This means that x_1, x_2, x_3 are additive but as we found possible interaction between cJADAs and low pain indicator in previous analysis, we still kept this interaction term in the model of $\tau(X)$.

- Reg-parm: estimator from parametric multiple linear outcome regression while considering interactions between treatment and cJADAs, treatment and diffVisits, treatment and low pain, treatment and morning stiffness.

5.2.2 Relative Importance of Variables and Interactions in the Estimation of Mean and Propensity Scores

The introduction of machine learning methods aims at estimating mean and propensity scores at the first stage. We started with a complete set of baseline covariates (a total of 19) and let BART/GBM do the rest job of estimation. Although our main purpose is to obtain $\hat{m}(X)$ and $\hat{\pi}(X)$ without bothering of digging into the details about how the models being fitted, investigation of the important variables in predicting outcome as well as treatment assignment is helpful for gaining some insights into the observed data and the underlying associations among these variables. Moreover, as two different machine learning algorithms were used, the question of whether these two approaches are comparable in the outcome regression or propensity score modeling is worth of exploration. To answer this question, we inspected the relative importance of variables for the mean score estimation and propensity score estimation based on BART and GBM, respectively. In ensemble-tree-based methods, the relative importance of variables is considered as an important guideline for variable selection as it is calculated according to the "inclusion proportion" in the trees during the iteration process of model fitting. In GBM, the measure of relative importance for each variable is based on the number of times that variable is selected for splitting, weighted by the squared improvement to the model as a result of each split, and averaged over all trees (Natekin and Knoll, 2013). Similarly, in BART, for any given predictor this quantity is calculated as the proportion of times that this predictor is involved in defining a splitting rule out of all rules appearing in the sum-of-trees model among the posterior draws(Kapelner and Bleich, 2016).

We randomly chose one out of the 5 imputed datasets for the illustration of variable selection in BART and GBM. It was found that the performance of prediction is not much sensitive to the chosen tuning hyper parameters in this case. We report the final results based on the following specified values for tuning parameters (i) BART: default values in bartMachine package for the hyperparameters of defining the prior probabilities. The number of trees to be grown: 200; number of burn-in iterations: 500; post-burn-in samples for prediction: 1000 (ii) GBM: tree complexity: 3; learning rate or shrinkage 0.005; bagging fraction: 0.5, number of trees to be grown depends on the tree complexity and shrinkage. Relative importance of all covariates, two-way interactions in estimation of the two summary scores from both methods are represented in Figure 5.1 and Figure 5.2.

- According to Figure 5.1, the ranks (from highest to least relative importance) of the top 10 selected variables based on BART are baseline cJADAs, lom, private, subtype3, white, subtype1, ms1, RF postive, age,subtype2. Based on GBM, this order is baseline cJADAs, age, lom, md assessment, pain, timediag, subtype3, well being, white, esr. So the two methods agree on the most important variable, i.e the baseline cJADAs and there are 5 common variables among the top 10 selected variables by each method. The Pearson correlation of the predicted mean scores from the two methods is about 0.93. As for the relative importance of two-way interactions among the 19 variables, in BART, results indicate that the top 5 important two-way interactions are: subtype1 by white, subtype3 by baseline cJADAs, lom by baseline cJADAs, ANA positive by baseline cJADAs, white by pain. In GBM, however, the variables from the

top 5 two-way interactions involving baseline cJADAs include baseline age, lom, white, timediag and subtype3.

- Figure 5.2 shows that in the propensity model fitting, the top 10 important variables selected by BART are lom, baseline cJADAs, B27, ANA positive, RF positive, timediag, ms2, well being, pain and age, while in GBM these top 10 important variables are baseline cJADAs, lom, timediag, esr, age, pain, well being, ANA positive, md assessment and pain. Although there are 8 common variables, the most important variable given by the two methods don't agree with each other. The Pearson correlation of the predicted propensity scores from the two methods is about 0.88. For the two-way interactions, the top 5 important ones from BART are B27 by lom, lom by RF positive, timediag by ANA postive, subtype1 by private and B27 by timediag, from GBM are esr by lom, timediag by baseline cJADAs, timediag by lom and baseline cJADAs by esr.

We also present the scatter plot of estimated mean/propensity scores from BART against that from GBM in Figure 3. Figure 3 suggests that although there exists discrepancies in the variable selection procedures of the two methods, the resulted mean scores and propensity scores are highly correlated and comparable.

5.2.3 Diagnosis of Propensity Scores from Different Methods

In real applications involving propensity score estimation, to assess whether propensity score modeling resulting in valid propensity scores, it is a common practice to examine the degree of overlapping of the estimation propensity scores between the two treatment arms, and the baseline covariates balancing before and after being

FIGURE 5.1: Relative variable importance of the baseline covariate and their two way interactions in mean score estimation

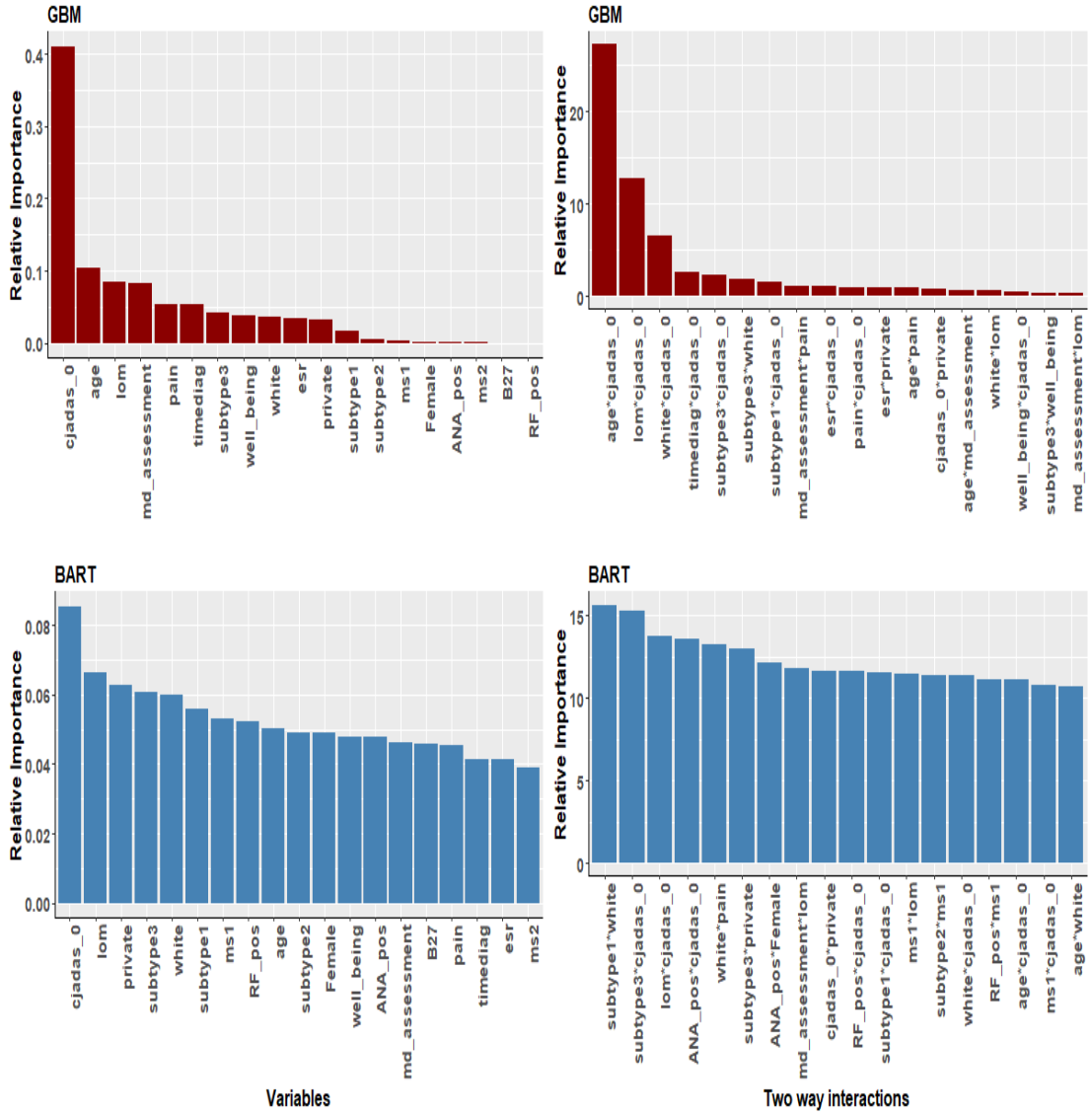


FIGURE 5.2: Relative variable importance of the baseline covariate and their two way interactions in propensity score estimation

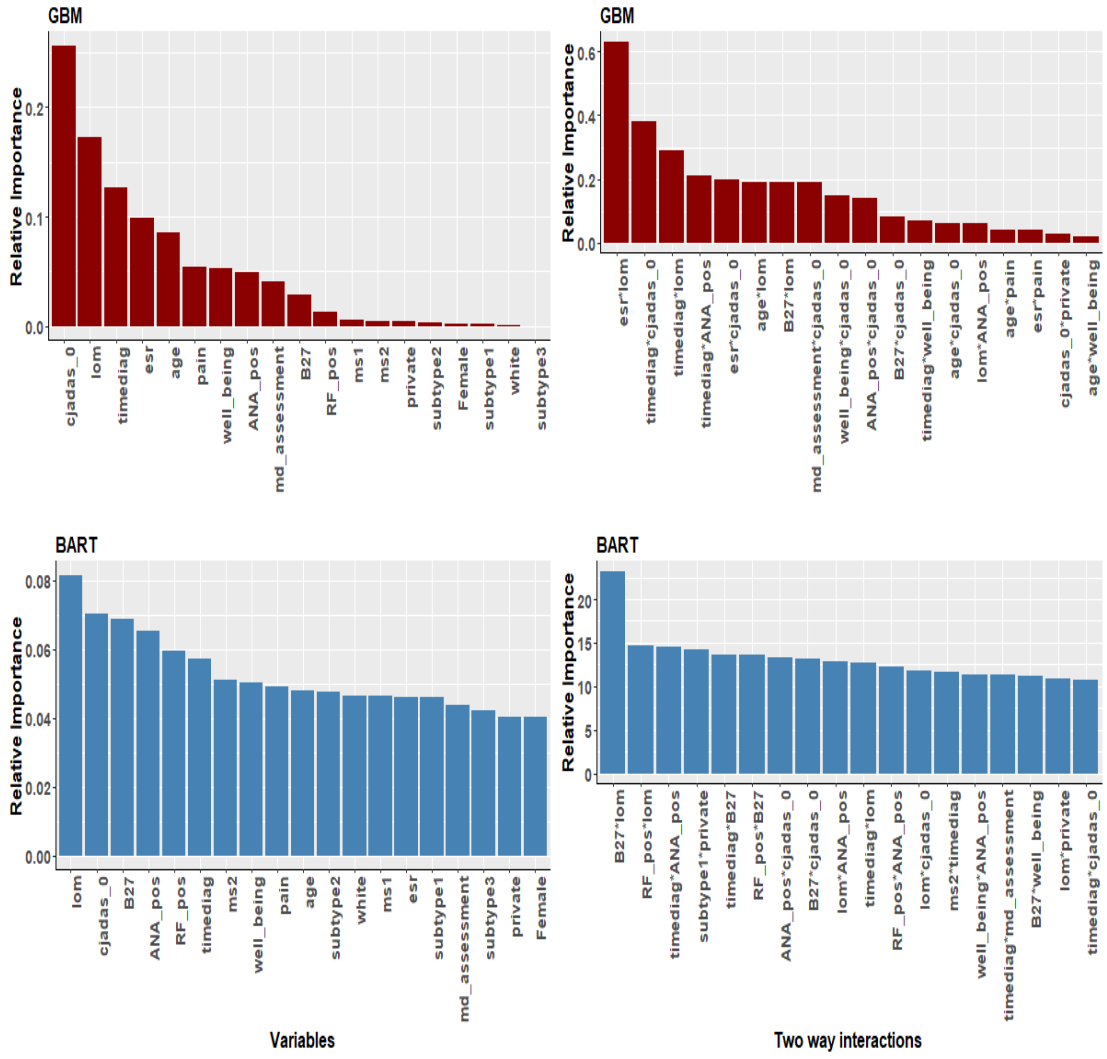
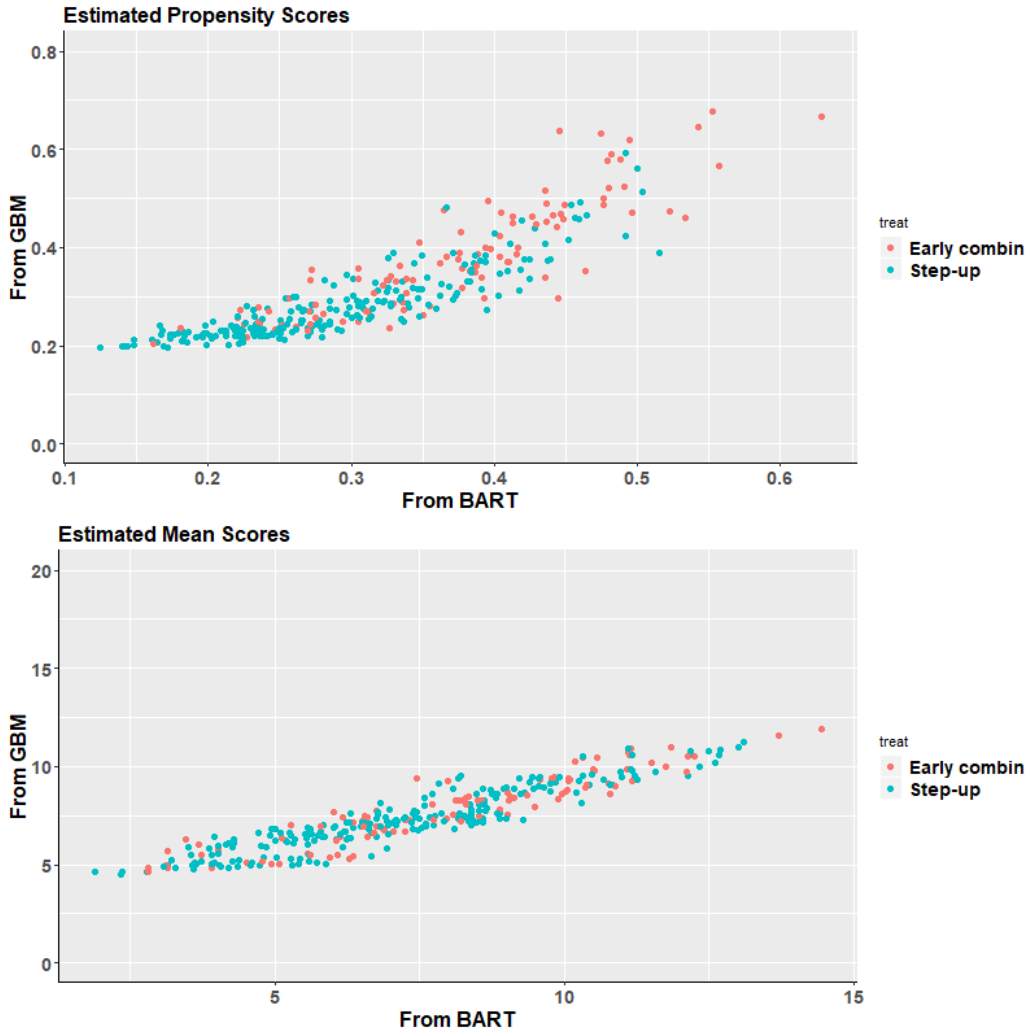


FIGURE 5.3: Scatter plots of estimated mean and propensity score from BART verse from GBM



weighted by the estimated propensity scores. We showed in Figure 5.3 that BART and GBM yield comparable propensity scores. To make comparison of all the three different methods, i.e. CBPS, BART and GBM adopted in propensity scores modeling, we used the boxplots to present the distributions of estimated propensity scores in Figure 5.4. As indicated in Figure 5.4, the overlapping of the two treatment groups for all three methods are satisfactory although the GBM method gave relatively higher propensity scores for the early combination group (therefore more separated) than the other two methods. The covariates balancing diagnosis was illustrated in Figure 5.5. As the three methods are comparable, we only presented the statistics based on BART method in this plot. The threshold of standardized mean difference for assessing the balance of covariates between the two treatment arms is usually 0.2. Figure 5.5 suggests that more balance for the majority of the baseline characteristics was achieved in the weighted sample after incorporating the estimated propensity scores from BART. The standardized mean difference for all the 20 variables are less than 0.2 in the weighted sample. This indicates that the estimated propensity scores are valid and appropriate for use in the inference for causal treatment effects.

5.2.4 Heterogeneous Treatment Effects

With the variables diffVisits and morning stiffness added into the treatment effect function modeling, we reexamined the heterogeneous pattern of the treatment effects and presented the heterogeneity from different perspectives in Figure 5.6. The dots in Figure 5.6 are the estimated individual treatment effects and the colored curves are the corresponding smoothing LOESS curves. The left two panels from up to down are for the estimated curves of treatment effect as a function of follow-up

FIGURE 5.4: Distribution of estimated propensity scores from CBPS, BART and GBM, respectively, by treatment groups

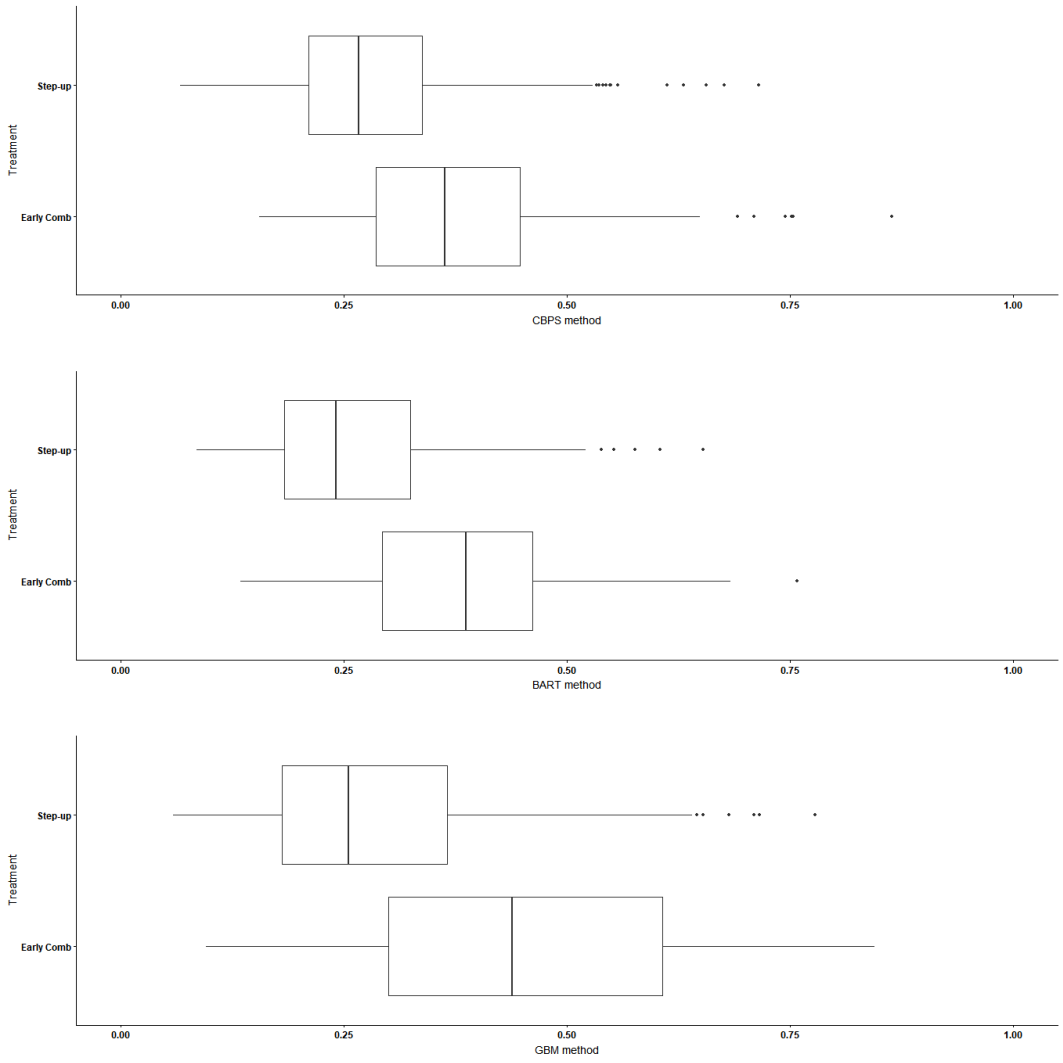
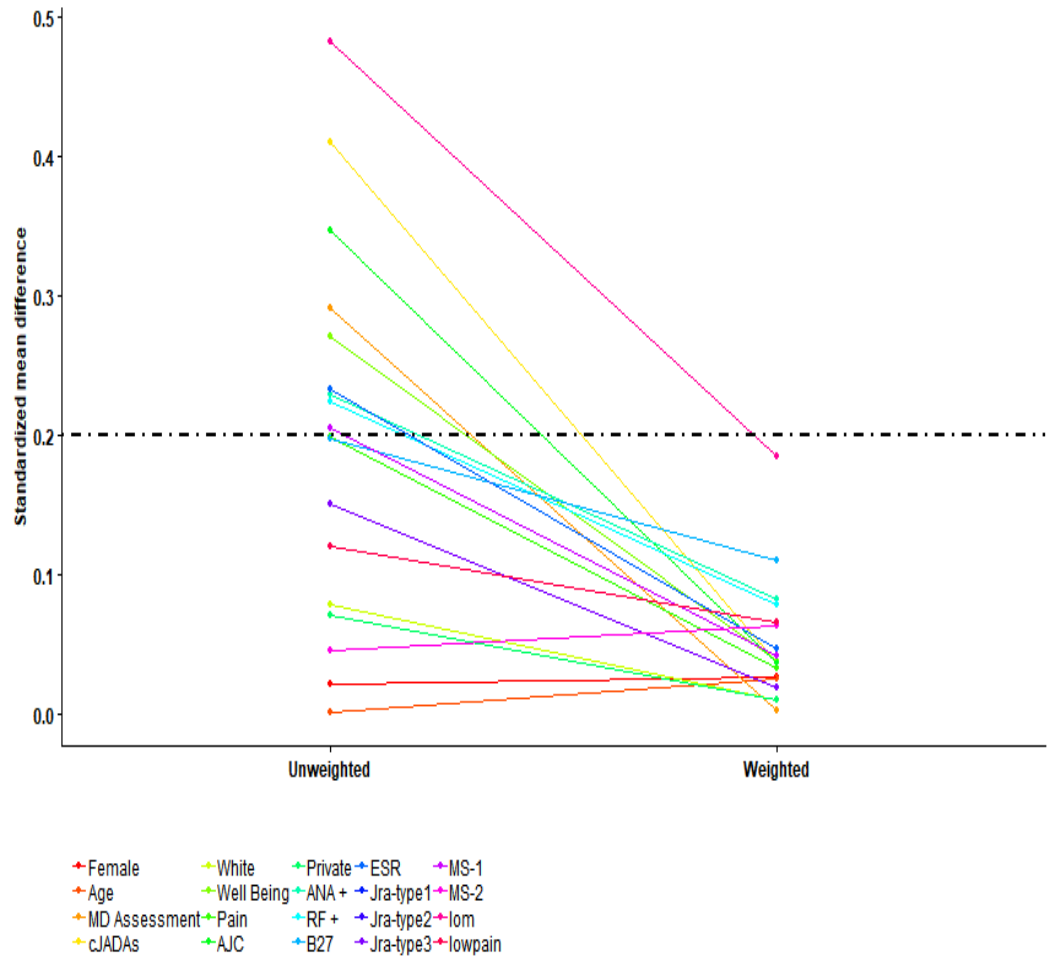


FIGURE 5.5: Diagnosis of estimated propensity scores for assessing the covariates balancing



time stratified by morning stiffness and low pain, respectively. In the right two panels from up to down, we plotted the estimated curves of treatment effect as a function of baseline cJADAs by morning stiffness and low pain, respectively. We also used the vertical line at exactly 6 month to detect possible different trend before and after 6 month. The reference score for cJADAs is still 15, the median. First, the left two plots imply that treatment effect is kind of stable before 5 month of follow up and starts getting stronger after around 6.5 month. There is much variation in the effects between 5 and 6 month, but the overall effect is positive. The trends for the two morning stiffness groups and the two pain groups are similar over time. Second, the right two plots show similar trend of treatment effect with respect to cJADAs as seen in Figure 4.3: more beneficial effect of early use of biologic DMARDs was seen among those patients with cJADAs score 15+ than those had less cJADAs score at baseline. Patients with morning stiffness being 3 ($>$ 15 minutes) seem to enjoy less benefit from early use of biologic DMARDs than the other two reference groups. Also, the interaction effect of cJADAs and low pain indicator still presents in the right bottom plot and biologic DMARDs could be effective in lowing down the cJADAs during follow up for those patients experiencing moderate to sever pain with cJADAs score 15+.

5.2.5 Estimated Average Treatment Effects

Table 5.2 presents a summary of the statistics of all the estimators from various methods described in section 5.2.1. For each of these estimators, we used 200 bootstrapping samples to compute its standard error, based on which the 95% confidence

FIGURE 5.6: Examination of heterogeneous treatment effects with respect to baseline cJADAs, 6-month outcome assessment time, low pain and severe morning stiffness

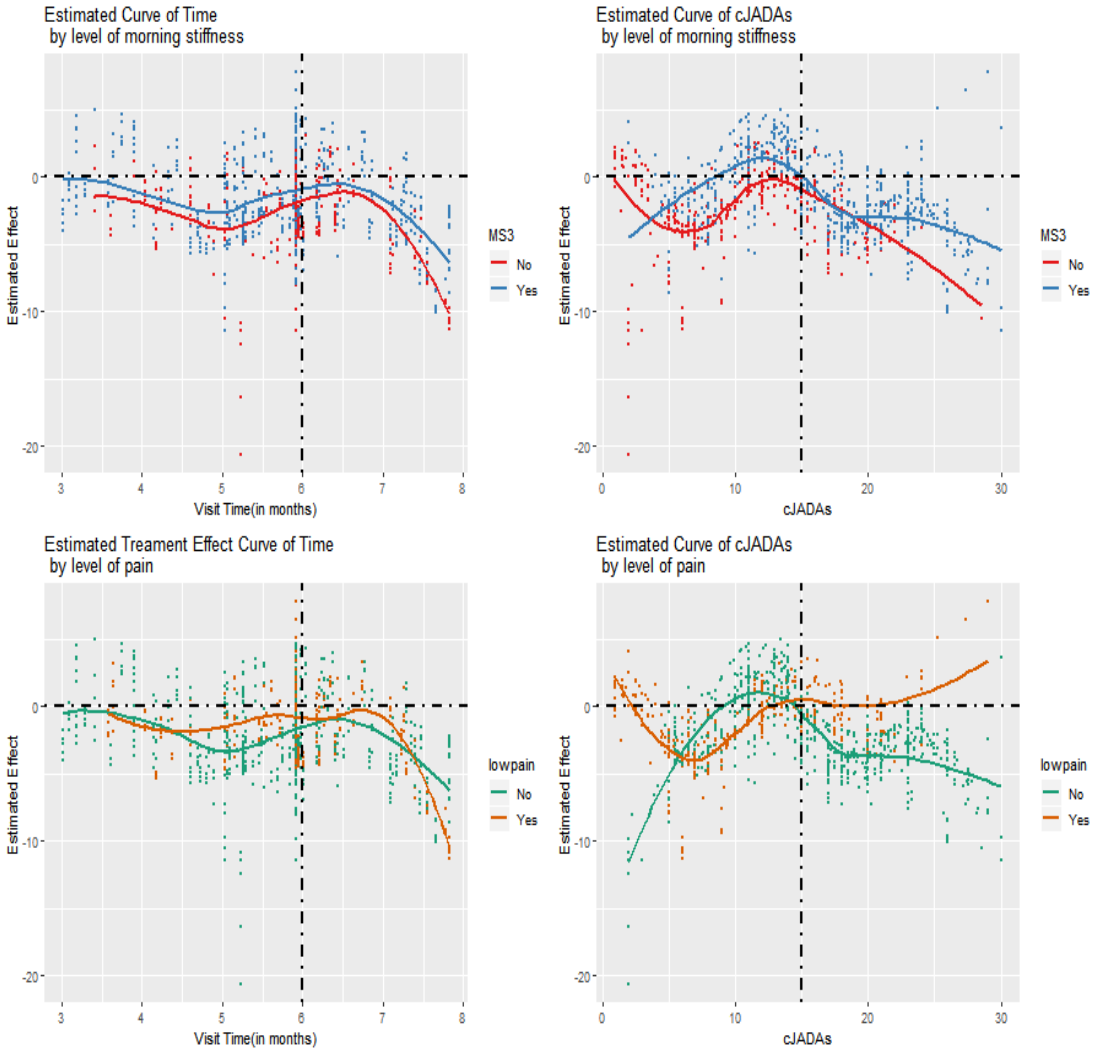
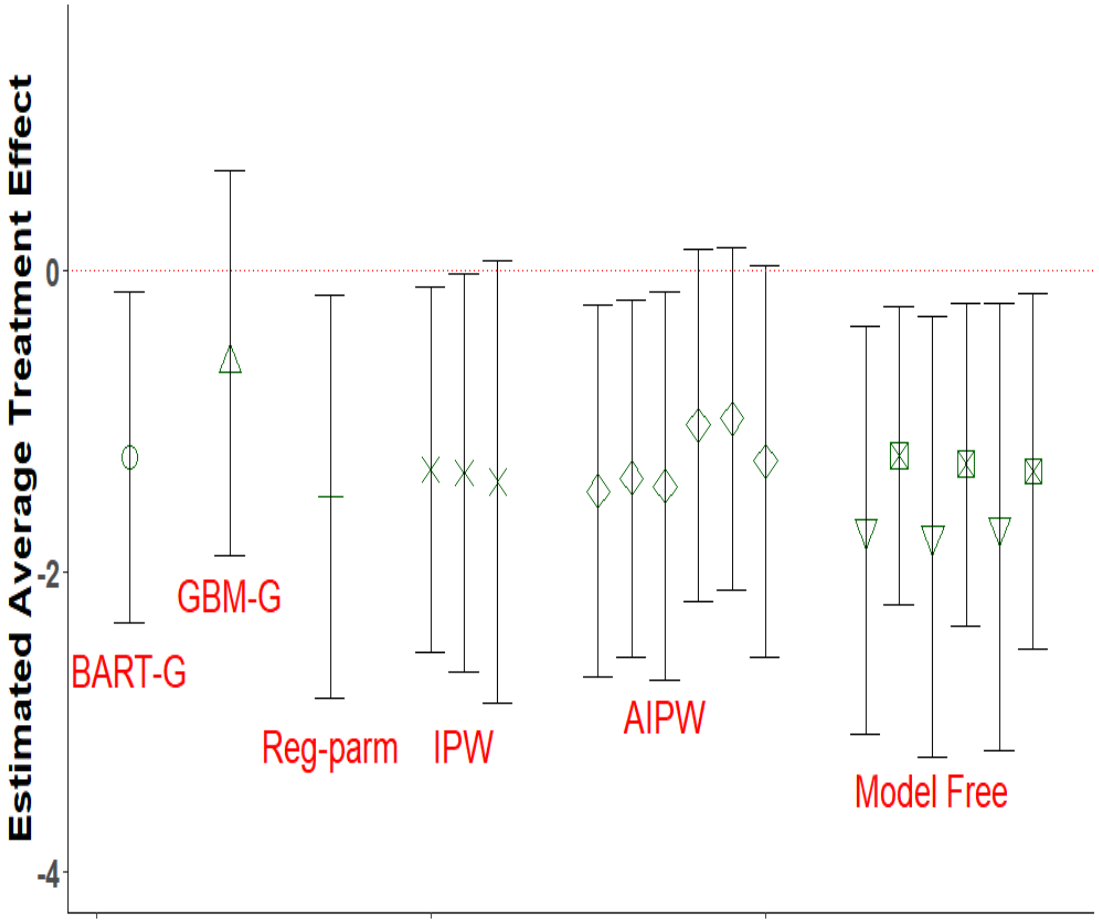


FIGURE 5.7: Estimators of treatment effect and 95% confidence intervals



intervals was constructed. We also plotted all the estimators with their corresponding 95% confidence intervals in Figure 5.7.

According to Table 5.2 and Figure 5.7, the estimated treatment effects given by all the estimators are below 0 with the biggest effect given by MF(BART-psBART) -1.769 and the smallest effect from GBM-G -0.611. All the estimators based on our proposed methods , i.e. MF(BART-psCB), MF(GBM-psCB), MF(BART-psBART), MF(GBM-psBART), MF(BART-psGBM) and MF(GBM-psGBM) give 95% confidence intervals below 0. It suggests that there is strong evidence of significant treatment effect of early aggressive use of biological DMARDs among this study popu-

TABLE 5.2: Comparison of estimated average treatment effect of early combination among all the methods

	Estimator	Ste.	95%lower	95%upper
BART-G	-1.240	0.564	-2.346	-0.134
GBM-G	-0.611	0.653	-1.892	0.669
Reg-parm	-1.504	0.687	-2.850	-0.158
IPW-psBART	-1.322	0.624	-2.545	-0.100
IPW-psCB	-1.344	0.679	-2.674	-0.014
IPW-psGBM	-1.404	0.752	-2.879	0.070
AIPW-BART-psCB	-1.465	0.632	-2.702	-0.227
AIPW-BART-psBART	-1.380	0.608	-2.572	-0.188
AIPW-BART-psGBM	-1.434	0.659	-2.726	-0.142
AIPW-GBM-psCB	-1.025	0.600	-2.201	0.151
AIPW-GBM-psBART	-0.983	0.584	-2.129	0.162
AIPW-GBM-psGBM	-1.268	0.667	-2.575	0.039
MF(BART-psCB)	-1.725	0.692	-3.082	-0.369
MF(GBM-psCB)	-1.229	0.505	-2.218	-0.240
MF(BART-psBART)	-1.769	0.748	-3.234	-0.304
MF(GBM-psBART)	-1.285	0.549	-2.361	-0.209
MF(BART-psGBM)	-1.705	0.759	-3.194	-0.217
MF(GBM-psGBM)	-1.333	0.605	-2.519	-0.148

lation at level 0.05. As presented in Chapter 3 and 4, with only baseline cJADAs and low pain indicator involved in the treatment effect estimation, the treatment effect is found to be about -1.1 with non-statistical significance at level 0.05. Therefore more pronounced treatment effect was revealed after taking into account a more complete set of baseline covariates. Although we found similarity in the estimated mean/propensity scores between the two machine learning methods, it is quite interesting to notice that the three estimators with mean score estimated from BART (MF(BART-psCB), MF(BART-psBART) and MF(BART-psGBM)) give comparable effects around -1.7 while the three estimators with mean score based on GBM (MF(GBM-psCB), MF(GBM-psBART) and MF(GBM-psGBM)) yield similar results close to -1.3. And, the standard error associated with MF(BART-psCB), MF(BART-psBART) and MF(BART-psGBM) are larger than the three estimators with mean score based on GBM therefore resulting in wider confidence intervals. The most narrow 95% confidence interval is given by MF(GBM-psCB) while the biggest point estimator of treatment effect is given by MF(BART-psBART). The results also imply that in this specific case study the estimated treatment effect is more driven by the modeling strategy of mean scores than by modeling of propensity scores.

Among the three estimators not involving propensity scores, BART-G and Reg-parm result in significant average treatment effects at level of 0.05. The variables used in specifying treatment effect heterogeneity in Reg-parm are the same as in estimators from our proposed method but the resulting treatment effect from Reg-parm is about -1.5, which is different from those given by our proposed method. As shown in Figure 5.6 and discussed in the section 5.2.4 of heterogeneous treatment effects, the estimated curves of treatment effects as function of diffvisit, baseline cJADAs

may not be simply linear. So effect given by Reg-parm might not be reliable and not necessarily preferable than the effects given by our proposed methods. Interestingly, GBM-G gives much smaller effect, -0.61, than BART-G.

Among the three estimators based on IPW, IPW-psGBM doesn't yield a significant treatment effect at 0.05 level with widest confidence interval while IPW-psBART and IPW-psCB give very similar treatment effects, around -1.3. For AIPW, the three estimators associated with the outcome being modeled by BART show significance at 0.05 level with the estimated treatment effect around -1.4. With the outcome modeled by GBM in AIPW the estimated treatment effects are a bit smaller with no evidence of statistical significance at level 0.05. Since the estimated propensity scores fall in the range about $(0.15, 0.8)$, no severe issue of extreme inverse probability weights were observed for the IPW and AIPW. It is however noticeable that when GBM is used in outcome modeling our proposed method results in smaller bootstrapping standard errors than its counterpart of AIPW. For example, standard error associated with AIPW-GBM-psBART is 0.584, larger than that of MF(GBM-psBART) 0.549. Although when BART is used in outcome modeling our proposed method gives larger standard errors than those AIPW, but both methods yield significant treatment effects.

CHAPTER 6

Conclusion And Discussion

6.1 Conclusion

In this thesis we proposed some nonparametric spline-based sieve estimation methods for causal treatment effects in the spirit of ordinary least-squares method. We elaborated on how this proposed OLS-based method can be applied to estimate homogeneous treatment effect, and to study heterogeneous effects pattern as well as to estimate average treatment effect when there exists heterogeneity in treatment effects. When X is a low-dimensional covariate vector, our method does not need to specify parametric functional forms for both the mean and propensity score models and hence can be regarded as a robust and model-free estimation method. The most notable feature of our method is that unlike the IPW-type methods, the proposed approach does not need to inversely weight the individual propensity scores and hence prevents the inflation of the observed outcomes associated with the estimated propensity scores near 0 or 1. As demonstrated in the numerical studies, this methodological advantage of incorporating the outcome and treatment assignment mechanism results in numerical stability in estimating the casual treatment effect in comparison with other IPW-type methods. We showed through multiple simulations that in both homogeneous and heterogeneous treatment effect scenarios, the proposed model-free estimator is consistent with limiting normal distribution. We demonstrated that the ordinary asymptotic normality theory based inference is valid in our proposed

approach for estimating average treatment effect with moderate sample size. It is nearly as efficient as the MLE method when the complete stochastic mechanism for outcomes is known, which is, of course, not practically feasible. In contrast to IPW-type methods, our approach is capable of studying treatment effect heterogeneity as well as inferring average treatment effect, which offers a great advantage when applied to answer casual questions regarding treatment effect heterogeneity. Moreover, when X is high-dimensional, we showed that the advanced machine learning techniques may be integrated into our method and make the implementation feasible and thus more generalizable in real world applications. All these nice features lead the proposed method to be a practically desired approach in making causal inference for treatment effects in observational studies.

6.2 Discussion

Although we only considered the scenario of comparison between treatment and control for the sake of simplicity in presentation and for the reason that it is the most common situation in biomedical studies, the proposed methodology can be readily extended to a scenario with multiple treatment levels. Suppose we have K treatment levels (a_k for $k = 1, \cdots, K$) in a study with A indicating the treatment that a subject received. Let $Y(a_k)$ denote the potential outcome associated with treatment level a_k for $k = 1, \cdots, K$. Let a_K chosen to be the reference treatment level for comparison and $\tau_k = E(Y(a_k) - Y(a_K))$ denote the causal treatment effect for treatment level a_k in comparison with the reference level a_K for $k = 1, \cdots, K-1$. Let $\pi_k(X) = pr(A = a_k|X)$ denote the multivariate propensity scores for $k = 1, \cdots, K-1$. In the scenario of homogeneous causal treatment effects, the equation of the expected

observed outcome given treatment indicator A and covariate X in simple treatment-control case can be similarly derived as

$$E(Y|A,X) = m(X) + \sum_{k=1}^{K-1} \left(1[A = a_k] - \pi_k(X)\right) \tau_k. \tag{6.1}$$

the two-stage model-free estimator of $\tau = (\tau_1, \cdots, \tau_{K-1})$ can be explicitly obtained by solving for the least-squares problem

$$\arg\min_{\tau} \sum_{i=1}^{n} \left\{ (Y_i - m(X_i)) - \sum_{j=1}^{K-1} (1[A_i = a_j] - \pi_j(X_i))\tau_j \right\}^2,$$

after obtaining the nonparametric spline-based sieve estimates for $m(X)$ and $\pi_k(X)$ for $k = 1, \cdots, K-1$ via the least-squares model and multinomial logistic regression model conducted in the first stage.

When treatment effects are heterogeneous, the three-stage estimation procedure is still implementable. We first use the sieve method to estimate the covariate-specific function $\tau_k(X) = \sum_{j=1}^{q_{n(k)}} \delta_{j(k)} B_j(x)$ for $k = 1, \cdots, K-1$ where $B_j(x)$'s are still the spline basis functions. The design matrix for computing the spline coefficients $(\delta_{1(1)}, \cdots, \delta_{q_n(1)}, \cdots, \delta_{1(K-1)}, \cdots, \delta_{q_n(K-1)})$ is thus

$$\Upsilon_{n \times (q_{n(1)} + \cdots + q_{n(K-1)})} = \left[\begin{array}{c|c|c|c|c|c} B_1 & B_2 & B_3 & \ldots & B_{K-2} & B_{K-1} \end{array} \right]$$

where for $k = 1, \cdots, K-1$

$$B_k = \begin{bmatrix} (A_{1k} - \hat{\pi}_{1k}(X_1))B_{1(k)}(X_1) & \dots & (A_{1k} - \hat{\pi}_{1k}(X_1))B_{q_{n(k)}}(X_1) \\ (A_{2k} - \hat{\pi}_{2k}(X_2))B_{1(k)}(X_2) & \dots & (A_{2k} - \hat{\pi}_{2k}(X_2))B_{q_{n(k)}}(X_2) \\ \dots & \dots & \dots \\ (A_{nk} - \hat{\pi}_{nk}(X_n))B_{1(k)}(X_n) & \dots & (A_{nk} - \hat{\pi}_{nk}(X_n))B_{q_{n(k)}}(X_n) \end{bmatrix}$$

So $(\hat{\delta}_{1(1)}, \cdots, \hat{\delta}_{q_n(1)}, \cdots, \hat{\delta}_{1(K-1)}, \cdots, \hat{\delta}_{q_n(K-1)}) = (\Upsilon^T\Upsilon)^-\Upsilon^T(Y - \hat{M})$ where $\hat{M}$ is the vector of estimated mean scores. And the average treatment effect of the $k-$th treatment level τ_k can thus be estimated by $\hat{\tau}_k = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{q_{n(k)}}\hat{\delta}_{j(k)}B_j(x_i)$.

In this thesis, we only considered the continuous outcome. When the outcome is binary or count, the average treatment effect $\tau = E(Y(1) - Y(0))$ or the covariate-specific treatment effect function $\tau(X) = E(Y(1) - Y(0)|X)$ can be estimated in the same way as in the continuous outcome case. However, they might not bear conventional interpretation of causal effects, such as odds ratio for the binary outcome. Our core linear model derived from the standard causal assumptions will need to be extended to a "generalized" linear equation with the appropriate link function for other types of outcome in order to make inference for the conventional causal effects. This will be the future direction for this research.

BIBLIOGRAPHY

Abadie, A. and G. W. Imbens (2006). Large sample properties of matching estimators for average treatment effects. *Journal of the American Statistical Association 74*(1), 235–267.

Andrews, D. W. K. (1991). Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica 59*(3), 817–858.

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

Austin, P. C. (2009). The relative ability of different propensity score methods to balance measured covariates between treated and untreated subjects in observational studies. *Medical Decision Making 29*(6), 661–677.

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

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

Barnard, J. and D. B. Rubin (1999). Small-sample degrees of freedom with multiple imputation. *Biometrika 86*, 948–955.

Bierens, H. J. (2012). Consistency and asymptotic normality of sieve estimators under weak and verifiable conditions. https://capcp.la.psu.edu/papers/2011/bierensconsistency.pdf.

Breiman, L. (1996). Bagging predictors. https://statistics.berkeley.edu/sites/default/files/tech-reports/421.pdf.

Breiman, L. (2001). Random forests. https://www.stat.berkeley.edu/~breiman/randomforest2001.pdf.

Cao, W., A. Tsiatis, and M. Davidian (2010). Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data. *Biometrika 96*(3), 723–734.

Carpenter, J. R., M. G. Kenward, and S. Vansteelandt (2006). A comparison of multiple imputation and doubly robust estimation for analyses with missing data. *Journal of the Royal Statistical Society, Series A 169*, 571584.

Chen, X. (2007). Large sample sieve estimation of semi-nonparametric models. *Handbook of Econometrics 6B, Part B*, 5549–5632.

Chen, X. and D. Pouzo (2015). Sieve wald and qlr inferences on semi/nonparametric conditional moment models. *Econometrica 83*, 1013–1079.

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

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association 74*(368), 829–836.

Cochran, W. G. (1968). The effectiveness of adjustment by subclassication in removing bias in observational studies. *Biometrics 24*, 295–313.

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

Cole, S. R. and C. E. Frangakis (2009). Commentary: The consistency statement in causal inference: A definition or an assumption? *Epidemiology 20*(1), 3–5.

De Boor, C. (2001). *A practical guide to splines*. New York: Springer.

Dierckx, P. (1993). *Curve and Surface Fitting with Splines*. New York: Oxford University Press.

Eilers, P. H. C. and B. D. Marx (1996). Flexible smoothing with b-splines and penalties. *Statistical Science 11*(2), 89–121.

Fan, J. and I. Gijbels (1992). Variable bandwidth and local linear regression smoothers. *Annals of Statistics 20*(4), 2008–2036.

Fisher, R. A. (1935). *The Design of Experiments (8th ed., 1966)*. New York: Hafner Press.

Fong, C., M. Ratkovic, K. Imai, C. Hazlett, X. Yang, and S. Peng (2019). Package cbps. `https://cran.r-project.org/web/packages/CBPS/CBPS.pdf`.

Freund, Y. and R. E. Schapire (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences 55*, 119–139.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annual of Statistics 29*(5), 1189–1232.

Gabay, C., M. Riek, A. Scherer, and A. Finckh (2015). Effectiveness of biologic dmards in monotherapy versus in combination with synthetic dmards in rheumatoid arthritis: data from the swiss clinical quality management registry. *Rheumatology 54*, 16641672.

Geman, S. and D. Geman (1984). Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Mathematical Intelligence 6*, 721–741.

Geman, S. and C.-R. Hwang (1982). Nonparametric maximum likelihood estimation by the method of sieves. *Annals of Statistics 10*(2), 401–414.

George, E. I. and S. T. Jensen (2014). Variable selection for bart: An application to gene regulation. *The Annals of Applied Statistics 8*(3), 17501781.

Glynn, A. N. and K. M. Quinn (2010). An introduction to the augmented inverse propensity weighted estimator. *Political Analysis 18*, 36–56.

Grenander, U. (1981). *Abstract Inference*. New York: Wiley.

Guo, Q., Y. Wang, D. Xu, J. Nossent, N. J. Pavlos, and J. Xu (2018). Rheumatoid arthritis: Pathological mechanisms and modern pharmacologic therapies. `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5920070/pdf/41413_2018_Article_16.pdf`.

Gutman, R. and D. B. Rubin (2017). Estimation of causal effects of binary treatments in unconfounded studies with one continuous covariate. *Statistical Methods in Medical Research 26*(3), 1199–1215.

Hahn, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica 66*, 315331.

Hansen, B. B. (2004). Full matching in an observational study of coaching for the sat. *Journal of the American Statistical Association 99*(467), 609–618.

Hastie, T. and R. Tibshirani (1986). Generalized additive models. *Statistical Science 1*(3), 297–310.

Hastie, T. and R. Tibshirani (2000). Bayesian backfitting (with comments and a rejoinder by the authors). *Statistical Science 15*(3), 196–223.

Heejung, B. and J. Robins (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics 61*, 962–972.

Hijmans, R. J., S. Phillips, J. Leathwick, and J. Elith (2017). Package dismo. `https://cran.r-project.org/web/packages/dismo/dismo.pdf`.

Hijmans, R. J., S. Phillips, J. Leathwick, and J. Elith (2019). Package gbm. `https://cran.r-project.org/web/packages/gbm/gbm.pdf`.

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

Hirano, K., G. W. Imbens, and G. Ridder (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica 71*(4), 1161–1189.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association 81*, 945–960.

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

Huang, J. Z. (2003). Local asymptotics for polynomial spline regression. *Annals of Statistics 31*(5), 1600–1635.

Ibrahim, J. G., M. Chen, S. R. Lipsitz, and A. Herring (2005). Missing data methods for generalized linear models: A comparative review. *Journal of the American Statistical Association 100*, 332346.

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

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

JIA (2017). Juvenile idiopathic arthritis. `https://www.cincinnatichildrens.org/health/j/jra`.

Kang, J. D. Y. and J. L. Schafer (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science 22*(4), 523–539.

Kapelner, A. and J. Bleich (2016). bartmachine: Machine learning with bayesian additive regression trees. `https://www.jstatsoft.org/article/view/v070i04`.

Kapelner, A. and J. Bleich (2018). Package bartmachine. `https://cran.r-project.org/web/packages/bartMachine/bartMachine.pdf`.

Kennedy, E. H. (2016). Semiparametric theory and empirical processes in causal inference. `https://arxiv.org/pdf/1510.04740.pdf`. [Book Chapter].

King, G. and L. Zeng (2006). The dangers of extreme counterfactuals. *Political Analysis 14*, 131–159.

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

Liang, H., S. Wang, J. Robins, and R. J. Carroll (2004). Estimation in partially linear models with missing covariates. *Journal of the American Statistical Association 99*, 357–367.

Lipsitz, S. R., J. G. Ibrahim, and L. P. Zhao (1999). A weighted estimating equation for missing covariate data with properties similar to maximum likelihood. *Journal of the American Statistical Association 94*(448), 1147–1160.

Little, R. and H. An (2004). Robust likelihood-based analysis of multivariate data with missing values. *Statistica Sinica 14*, 949–968.

Lunceford, J. K. and M. Davidian (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study. *Statistics in Medicine 23*, 2937–2960.

McCaffrey, D. F., G. Ridgeway, and A. R. Morral (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological Methods 9*, 403–425.

McCulloch, R., R. Sparapani, R. Gramacy, C. Spanbauer, and M. Pratola (2019). Package bart. `https://cran.r-project.org/web/packages/BART/BART.pdf`.

MOBILITY (2015). Moblity study. `https://med.uc.edu/mobility`. [Project Website].

Muller, H.-G. and U. Stadtmuller (1987). Variable bandwidth kernel estimators of regression curves. *Annals of Statistics 15*(1), 182–201.

Nadaraya, E. A. (1964). On estimating regression. *Theory of Probability and its Applications 9*, 141–142.

Natekin, A. and A. Knoll (2013). Gradient boosting machines, a tutorial. *Frontiers in Neurorobotics 7*, Article 21.

Newey, W. K. (1997). Convergence rates and asymptotic normality for series estimators. *Journal of Econometrics 79*, 147–168.

Neyman, J. (1990). On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science 5*(4), 465–480.

Oberle, E. J., J. G. Harris, and J. W. Verbsky (2014). Polyarticular juvenile idiopathic arthritis epidemiology and management approaches. *Clinical Epidemiology 6*, 379393.

Pearl, J. (2000). *Causality: Models, reasoning, and inference*. New York: Cambridge University Press.

Pearl, J. (2010). On the consistency rule in causal inference: Axiom, definition, assumption, or theorem? *Epidemiology 21*(6), 872–875.

Penrose, R. (1955). A generalized inverse for matrices. *Mathematical Proceedings of the Cambridge Philosophical Society 51*(3), 406–413.

Qin, J., B. Z. Zhang, and D. H. Y. Leung (2009). Empirical likelihood in missing data problems. *Journal of the American Statistical Association 104*, 1492–1503.

Ridgeway, G. (2005). Generalized boosted models: A guide to the gbm package.

Robins, J. (1986). A new approach to causal inference in mortality studies with a sustained exposure period application to control of the healthy worker survivor effect. *Mathematical Modelling 7*, 13931512.

Robins, J., M. Hernan, and B. Brumback (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology 11*(5), 550–560.

Robins, J., S. D. Mark, and W. K. Newey (1992). Estimating exposure effects by modelling the expectation of exposure conditional on confounders. *Biometrics 48*, 479–495.

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

Robinson, P. M. (1988). Root-n-consistent semiparametric regression. *Econometrica 56*, 931–954.

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

Rosenbaum, P. R. (1989). Optimal matching for observational studies. *Journal of the American Statistical Association 84*(408), 1024–1032.

Rosenbaum, P. R. (1991). A characterization of optimal designs for observational studies. *Journal of the Royal Statistical Society. Series B (Methodological) 53*(3), 597–610.

Rosenbaum, P. R. (2002). *Observational Studies* (2nd ed. ed.). New York: Springer.

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

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

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

Rubin, D. B. (1973). The use of matching 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*(5), 688–701.

Rubin, D. B. (1976a). Inference and missing data. *Biometrika 63*, 581–592.

Rubin, D. B. (1976b). Matching methods that are equal percent bias reducing: Some examples. *Biometrics 32*, 109–120.

Rubin, D. B. (1976c). Multivariate matching methods that are equal percent bias reducing: Maximums on bias reduction for fixed sample sizes. *Biometrics 32*, 121–132.

Rubin, D. B. (1980). Randomization analysis of experimental data: The fisher randomization test comment. *Journal of the American Statistical Association 75*(371), 591–593.

Rubin, D. B. (1986). Comment: What if's have causal answers. *Journal of the American Statistical Association 81*, 961–962.

Rubin, D. B. (1990). Comment: Neyman (1923) and causal inference in experiments and observational studies. *Statistical Science 5*, 472–480.

Rubin, D. B. and M. Van der Lann (2008). Empirical efficiency maximization: Improved locally efficient covariate adjustment in randomized experiments and survival analysis. *The International Journal of Biostatistics 4*(1), Article 5.

Ruppert, D., M. Wand, and R. Carroll (2003). *Semiparametric Regression*. New York: Cambridge University Press.

Scharfstein, D. O., R. Andrea, and J. Robins (1999). Adjusting for non-ignorable drop-out using semiparametric non-response models. *Journal of the American Statistical Association 94*, 1096–1120.

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

Setoguchi, S., S. Schneeweiss, M. A. Brookhart, R. J. Glynn, and E. F. Cook (2008). Evaluating uses of data mining techniques in propensity score estimation: A simulation study. *Pharmacoepidemiolgy and Drug Safety 17*, 546–555.

Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society, Series B. 13*, 238–241.

Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and akaike's criterion. *Journal of the Royal Statistical Society. Series B (Methodological) 39*, 44–47.

Stone, M. (1982). Optimal global rates of convergence for nonparametric regression. *Annals of Statistics 10*, 1040–1053.

Stone, M. (1985). Additive regression and other nonparametric models. *Annals of Statistics 13*, 689–705.

Stone, M. (1986). The dimensionality reduction principle for generalized additive models. *Annals of Statistics 14*, 590–606.

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

Tan, Z. (2006). A distributional approach for causal inference using propensity scores. *Journal of the American Statistical Association 101*, 1619–1637.

Tan, Z. (2007). Comment: Understanding or, ps and dr. *Statistical Science 22*(4), 560568.

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

Tan, Z. (2010b). Nonparametric likelihood and doubly robust estimating equations for marginal and nested structural models. *The Canadian Journal of Statistics 38*(4), 609–632.

Taubman, S., J. Robins, M. Mittleman, and M. Hernan (2009). Intervening on risk factors for coronary heart disease: An application of the parametric g-formula. *International Journal of Epidemiology 38*, 1599–1611.

Tsiatis, A. A. (2006). *Semiparametric Theory and Missing Data*. New York: Springer.

van Buuren, S. and K. Groothuis-Oudshoorn (2011). mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software 45*(3), 1–67.

van der Laan, M. and S. Gruber (2010). Collaborative double robust targeted maximum likelihood estimation. *The International Journal of Biostatistics 6*(1), Article 17.

van der Laan, M. and J. Robins (2003). *Unified Methods for Censored Longitudinal Data and Causality*. Berlin: Springer.

Vansteelandt, S., M. Bekaert, and G. Claeskens (2012). On model selection and model misspecification in causal inference. *Statistical Methods in Medical Research 21*(1), 7–30.

Vansteelandt, S. and M. M. Joffe (2014). Structural nested models and g-estimation: the partially realized promise. *Statistical Science 29*(4), 707–731.

Wallace, C. A., E. H. Giannini, S. J. Spalding, P. J. Hashkes, K. M. ONeil, A. S. Zeft, I. S. Szer, S. Ringold, H. I. Brunner, L. E. Schanberg, R. P. Sundel, D. Milojevic, M. G. Punaro, P. Chira, B. S. Gottlieb, G. C. Higgins, N. T. Ilowite, Y. Kimura, S. Hamilton, A. Johnson, B. Huang, and D. J. Lovell (2012). Trial of early aggressive therapy in polyarticular juvenile idiopathic arthritis. *Arthritis Rheumatology 64*(6), 2012–2021.

Wang, L., A. Rotnitzky, and X. Lin (2010). Nonparametric regression with missing outcomes using weighted kernel estimating equations. *Journal of the American Statistical Association 105*(491), 11351146.

Watson, G. (1964). Smooth regression analysis. *Sankhya, The Indian Journal of Statistics 26*, 359–372.

Williamson, E., R. Morley, A. Lucas, and J. Carpenter (2012). Propensity scores: from naiveeenthusiasm to intuitive understanding. *Statistical Methods in Medical Research 21*(3), 273–293.

Wyss, R., A. R. Ellis, M. A. Brookhart, G. CJ, M. Jonsson-Funk, L. RJ, and T. Sturmer (2014). The role of prediction modeling in propensity score estimation: An evaluation of logistic regression, bcart, and the covariate-balancing propensity score. *American Journal of Epidemiology 180*(6), 645–655.

Xue, L. and H. Liang (2009). Consistent variable selection in additive models. *Statistica Sinica 19*, 1281–1296.

Zhang, Y., L. Hua, and J. Huang (2010). A spline-based semiparametric maximum likelihood estimation method for the cox model with interval-censored data. *Scandinavian Journal of Statistics 37*, 338 – 354.

Zhou, S., X. Shen, and D. A. Wolfe (1998). Local asymptotics for regression splines and confidence regions. *Annals of Statistics 26*(5), 1760–1782.

CURRICULUM VITAE

Yuanfang Xu

EDUCATION

- Ph.D. in Biostatistics, Minor in Epidemiology, Indiana University, Indianapolis, IN, 2019
- M.S. in Applied Statistics, Purdue University, Indianapolis, IN, 2014
- M.S. in Computer Science & Engineering, Guangxi University, Nanning, China, 2004

WORKING EXPERIENCE

- Biostatistician, Cincinnati Children's Hospital Medical Center, Cincinnati, OH Jun 2016 - May 2019
- Graduate Research Assistant, Department of Biostatistics, Indiana University, Indianapolis, IN Jul 2015 - May 2016
- Graduate Teaching Assistant, Department of Mathematics, Indiana University, Indianapolis, IN Aug 2014 - Jun 2015
- Summer Intern, The Polis Center, Indiana University, Indianapolis, IN May 2014 - Aug 2014
- Lecturer, Changsha University of Science & Technology, Changsha, China Jul 2004 - Feb 2009

SELECT PUBLICATIONS

- Yuanfang Xu, Giorgos Bakoyannis, Bin Huang, Ying Zhang. An OLS-Based Model-Free Estimator of Causal Treatment Effect. In submission to *Biometrika*, 2019
- Chao Niu, Yuanfang Xu, Yunjie Huang, Hossain MM, Guilbert Theresa. Evaluation of Risk Scores to Predict Pediatric Severe Asthma Exacerbations. In submission to *Journal of Allergy and Clinical Immunology in practice*, 2019.
- Zhihui He, Keren Armoni Domany, Leonardo Nava-Guerra, Leonardo Nava-Guerra, Yuanfang Xu, Md Monir Hossain, Raouf Amin. Phenotype of Ventilatory Control in Children with Moderate to Severe Persistent Asthma and Obstructive Sleep Apnea. (2019). *Sleep, under review.*
- Keren Armoni Domany, Hadas Nahman-Averbuch, Christopher D King, Thomas Dye, Yuanfang Xu, Md Monir Hossain, Andrew D. Hershey, Narong Simakajornboon. (2019). Clinical presentation, diagnosis and polysomnographic findings in children with migraine referred to sleep clinics. *Sleep Medicine, Accepted.*
- Armoni Domany K, Zhihui He, Nava-Guerra L, Khoo MC, Xu Y, Hossain MM, DiFrancesco MW, McConnell K, Amin RS (2019). The Effect of Adenotonsillectomy on Ventilatory Control in Children with Obstructive Sleep Apnea. *Sleep, In press.* https://doi.org/10.1093/sleep/zsz045.
- Armoni Domany K, Hantragool S, Smith DF, Xu Y, Hossain M, Simakajornboon N.(2018). Sleep Disorders and Their Management in Children with Ehlers-Danlos Syndrome Referred to Sleep Clinics. *Journal of Clinical Sleep Medicine.* **14(4)**: 624-629.

- Armoni Domany K, Hossain MM, Nava-Guerra L, Khoo MC, McConnell K, Carroll JL, Xu Y, Di Francesco M, Amin RS.(2018). Cardioventilatory Control in Preterm Born Children and the Risk of Obstructive Sleep Apnea. *American Journal of Respiratory and Critical Care Medicine*. **197(12)**: 1596-1603.
- Dow DF, Mehta K, Xu Y, England E. (2018). The Relationship between Body Mass Index and Fatty Infiltration in the Shoulder Musculature. *Journal of Computer Assisted Tomography*. **42(2)**:323-329.
- Yuanfang Xu, Yang Zhou, Hua Zhen. (2006). The Research on the Realization of Back Propagation Network Based on MATLAB. *Microcomputer Applications*. **8**: 41-44.
- Yang Zhou, Yuanfang Xu. (2006). An Introduction to Computer Forensics Technologies. *Modern Computer*. **3**: 51-54.
- Yuanfang Xu, Ling Mo. (2004). A Study on Messages Matching Problem in Content-based Publish/Subscribe Model. *Computer and Modernization*. **11**: 67-69.

POSTER

- An Ordinary Least Squares Spline-Based Method for Causal Inference in Observational Studies. 2019 Midwest Biopharmaceutical Statistics Workshop. May, 2019

AWARDS

- Outstanding Graduating Master in Statistics Program, Purdue University, School of Science, Indianapolis. April, 2014