Fall 2020

Causal Effect Random Forest Of Interaction Trees For Learning Individualized Treatment Regimes In Observational Studies: With Applications To Education Study Data

Luo Li
Claremont Graduate University

CAUSAL EFFECT RANDOM FOREST OF INTERACTION TREES FOR LEARNING INDIVIDUALIZED TREATMENT REGIMES IN OBSERVATIONAL STUDIES: WITH APPLICATIONS TO EDUCATION STUDY DATA

By

Luo Li

Claremont Graduate University and San Diego State University

2020

APPROVAL OF THE REVIEW COMMITTEE

This dissertation has been duly read, reviewed, and critiqued by the Committee listed below,

which herby approves the manuscript of

Luo Li

as fulfilling the scope and quality requirements for meriting the degree of

Doctor of Philosophy.

Juanjuan Fan, Chair
Department of Mathematics and Statistics, San Diego State University
Professor

Richard Levine
Department of Mathematics and Statistics, San Diego State University
Professor

Barbara Bailey
Department of Mathematics and Statistics, San Diego State University
Associate Professor

John Angus
Institute of Mathematical Sciences, Claremont Graduate University
Professor

Qidi Peng
Institute of Mathematical Sciences, Claremont Graduate University
Research Assistant Professor

ABSTRACT OF THE DISSERTATION

CAUSAL EFFECT RANDOM FOREST OF INTERACTION TREES
FOR LEARNING INDIVIDUALIZED TREATMENT REGIMES
IN OBSERVATIONAL STUDIES:
WITH APPLICATIONS TO EDUCATION STUDY DATA
by
LUO LI
Doctor of Philosophy in Computational Science-Statistics
Claremont Graduate University and San Diego State University, 2020

Learning individualized treatment regimes (ITR) using observational data holds great interest in various fields, as treatment recommendations based on individual characteristics may improve individual treatment benefits with a reduced cost. It has long been observed that different individuals may respond to a certain treatment with significant heterogeneity. ITR can be defined as a mapping between individual characteristics to a treatment assignment. The optimal ITR is the treatment assignment that maximizes expected individual treatment effects. Rooted from personalized medicine, many studies and applications of ITR are in medical fields and clinical practice. Heterogeneous responses are also well documented in educational interventions. However, unlike the efficacy study in medical studies, educational interventions are often not randomized. Study results often suffer greatly from self-selection bias. Besides the intervention itself, the efficacy and effectiveness of interventions usually interact with a wide range of confounders.

In this study, we propose a novel algorithm to extend random forest of interaction trees to Casual Effect Random Forest of Interaction Trees (CERFIT) for learning individualized treatment effects and regimes. We first consider the study under a binary treatment setting. Each interaction tree recursively partitions the data into two subgroups with greatest heterogeneity of treatment effect. By integrating propensity score into the tree growing process, subgroups from the proposed CERFIT not only have maximized treatment effect differences, but also similar baseline covariates. Thus it allows for the estimation of the individualized treatment effects using observational data. In addition, we also propose to use residuals from linear models instead of the original responses in the algorithm. By doing so, the numerical stability of the algorithm is greatly improved, which leads to an improved prediction accuracy. We then consider the learning problem under non-binary treatment settings. For multiple treatments, through recursively partitioning data into two subgroups with greatest treatment effects heterogeneity with respect to two randomly selected treatment groups, the algorithm transforms the multiple learning ITR into a binary task. Similarly, continuous treatment can be handled through recursively partitioning the data into subgroups with greatest homogeneity in terms of the association between the response and the treatment within a child node. For all treatment settings, the CERFIT provides variable importance ranking in terms of treatment effects. Extensive simulation studies for assessing estimation accuracy and variable importance ranking are presented. CERFIT demonstrates competitive performance among all competing methods in simulation studies. The methods are also illustrated through an assessment of a voluntary education intervention for binary treatment setting and learning optimal ITR among multiple interventions for non-binary treatments using data from a large public university.

DEDICATION

I dedicate this dissertation to my family - my husband, my daughter and son, and my parents.

I love you all.

Your word is a lamp to my feet and a light to my path.

– Psalm 119:105

ACKNOWLEDGMENTS

First and foremost, I would like to express my deepest appreciation to my advisor, Dr. Juanjuan Fan. I could not imagine having a better advisor for my Ph.D. study other than her. I sincerely enjoyed every meeting with her. Her wide knowledge in the subject and excellent suggestions always results in an improved version of my paper. Without her encouragement and continuous guidance, I could not have finished this dissertation.

I would also like to extend my appreciation to all my dissertation committee members. I want to thank Dr. Richard Levine for his support and guidance in my dissertation research. Without his support, I would not get involved in education data mining, which I later found was extremely intriguing. I would also like to thank Dr. John Angus for his encouragement and knowledge. I truly enjoyed our discussions, which spark my interest in computational statistics. I also feel grateful that Dr. John Angus help me to access the high-performance computer in the Math Department at CGU to finish some of my simulation studies. My special thanks go to Dr. Barbara Bailey and Dr. Qidi Peng for providing me with their encouragement, valuable advice and for teaching me.

Furthermore, I would like to thank Dr. Jose Castillo and the Computational Science Research Center for a rich academic environment and funding my Ph.D. program. Also, I want to thank Dr. James Otto for setting up all the R environment I needed in SDSU high-performance computing clusters. Without his help, I could not accomplish all intense simulation studies for this dissertation.

Last but not least, I want to save my thanks to my family and my parents for their extraordinary support, patience and emotional comfort. I am forever in debt for their unconditional love to me.

TABLE OF CONTENTS

PAGE

LIST OF TABLES

PAGE

LIST OF FIGURES

PAGE

CHAPTER 1
INTRODUCTION

1.1 INDIVIDUALIZED TREATMENT REGIMES AND EDUCATION INTERVENTIONS

Learning individualized treatment regimes (ITR) using observational data holds great interest in various fields, as treatment recommendations based on individual characteristics may improve individual treatment benefits with a reduced cost. It has long been observed that different individuals may respond to a certain treatment with significant heterogeneity. [50, 2, 26, 42]. ITR can be defined as a mapping between individual characteristics to a treatment assignment [34]. The optimal ITR is the treatment assignment that maximizes expected treatment effects at an individual level. Rooted from personalized medicine, many studies and applications of ITR are in the medical research and clinical practice, such as depressive disorder, substance use disorder and sputum positive tuberculosis [44, 59, 71]. Heterogeneous responses are also well documented in educational interventions. Taking educational supplemental instruction (SI) research as an example, studies show that prior academic achievement, motivations, genders and minority status all affect SI effectiveness [60, 61, 57, 18]. Academic advising personalized to individual student's specific characteristics can maximize intervention effects and help students to achieve academic success.

However, unlike the efficacy study in the field of medication, educational interventions are often not randomized. Study results often suffered greatly from self-selection bias [5]. Special statistical adjustment, such as propensity score methods, need to be considered in order to achieve an unbiased estimate for observational data. Besides the intervention itself, the effectiveness of interventions usually interacts with a wide range of confounders, such as students' demographic, social economic background, academic status and even other intervention programs. Multiple interventions with similar educational objectives are also common. Oftentimes, interventions such as SI are provided along with other interventions, such as tutoring or recital supplement course. Therefore despite extensive studies on the effect of SI and its wide appeal, whether or not SI is effective continues to be controversial. After their systematic review of SI studies between 2001 to 2010, Dawson,Van Der Meer, Skalicky and Cowley concluded that SI seemed to work on some levels for some groups of students, as many studies are not methodologically sound or lack enough information. Existing methods are often parametric, and the nonlinear effects are often ignored in the model or rely on ad hoc approaches. The complex confounding, interaction and nonlinear relationship remain unveiled as studies rarely address multiple explanatory variables in one study, and are subject to model misspecification using the traditional parametric methods [5]. In addition, even though majority studies agree on the benefit of educational interventions, such as SI, claims are all based on the average treatment effect (ATE) on group level [18] and ignore the fact that not all the program attendees benefit from the intervention. Academic advising based on ATE can adversely impact an individual student‘s academic success considering the opportunity cost. In other words, if the student is spending time on an ineffective intervention, the student may lose the opportunity to benefit from another effective intervention. To improve the efficacy and effectiveness of the treatment, recommendation should be based on individualized treatment effects (ITE) rather than ATE.

Tree based machine learning methods have gained a great popularity due to the model flexibility with few statistical assumptions, their ability to handle a variety of data structures, and the interpretability and the exceptional predictive power. It is also one of leading methods for causal treatment effect estimation. Many criticisms discussed above, such as nonlinear issues and interaction effects, can be automatically handled in tree based methods. Therefore, in this study, we propose a novel algorithm extended from the tree based method, random forest of interaction trees [69] to access the intervention’s effectiveness on individual students using observational education study data. The method is data driven without explicit model specification. It has the versatility to handle both binary treatment and non-binary treatment settings. Furthermore, it also has advantages in dealing with high dimensional data, as opposed to the traditional parametric methods. By integrating general propensity scores into the tree growing process, this proposed method could be applied to both randomized and observational studies.

In the following sections, we first review essential concepts in tree based methods, which are the building blocks of the proposed algorithm.

1.2 CLASSIFICATION AND REGRESSION TREES

CART(classification and regression trees) is a tree based method proposed [8]. Tree based method recursively partitions the data using the best binary split until some criterion is met. Specifically, a tree is grown by splitting the root node into two child nodes that maximize between-node heterogeneity, or equivalently, minimize within-node impurity. In each split, CART algorithm searches all possible variables and all possible values. The same procedure is repeated for each child node until reaching a point where further splitting no longer decreases the impurity or a predetermined stopping rule is reached. A node that cannot be split any further is called a terminal node, an important attribute of a tree with respect to prediction. Specifically, each terminal node is a distinct partition of the sample based on the input variables. In other words, each terminal node is characterized by a unique combination of the attributes of an observation or patient characteristics. Given that each terminal node contains information on the outcome, predictions can easily be obtained given a set of patient characteristics.

To illustrate the process, lets consider a hypothetical example. To simplify the example, we only consider using Age (X_1), Gender (X_2), Pretest (X_3), and average quiz scores Quiz(X_4) to predict students' final exam scores (Y). As Figure 1.1 shows the root node (the whole data set) is split into two child nodes based on the splitting rule whether a student's quiz score is less than or equal to 7 $(X_4 \leq 7)$ out of possible points 10. If the answer is yes, the observation goes to the left child node, otherwise, goes to the right node. Then conditioned on this, the left child node is split into another two child nodes with protest score less than or equal to 30 $(X_3 \leq 65)$. This partition process continued by splitting the lower level right child node with whether an observation under that node is male $(X_2 = 0)$ or female $(X_2 = 1)$. The splitting continues recursively until a predetermined stopping rule is reached.

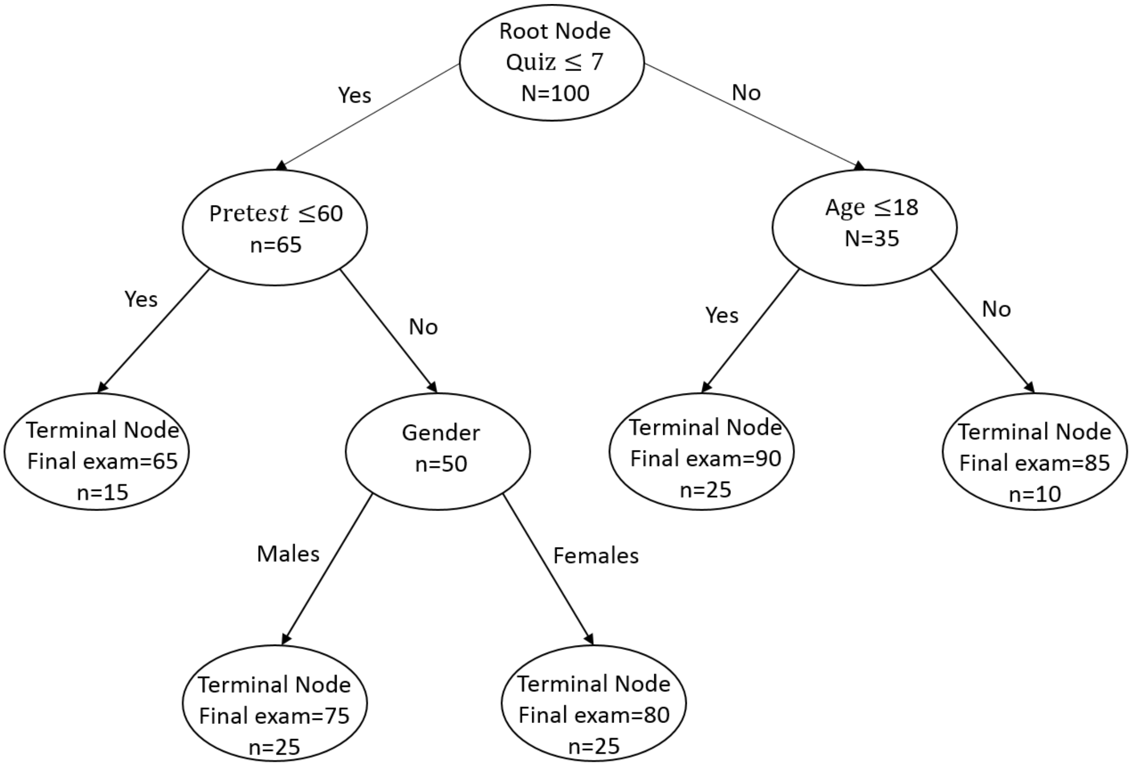

Figure 1.1. Hypothetical tree using age, gender, pretest, quiz to predict final exam scores.

Depending on the nature of the outcome, CART can be applied to both classification and regression problems. Unlike the conventional parametric model, which assumes one correct specified global model for the whole data set. In CART, the data complexity is reduced through data partition. The simplest model then could be applied into the smaller homogeneous subgroup. In this example, the original 100 data in the root node is partitioned into 5 terminal nodes with much smaller data size (10 to 25). In addition, the variable selection, transformation and interaction problems in parametric model can be handled automatically in CART. For instance, the interaction between Pretest and Gender is automatically captured in the hierarchical tree structure. The variables splitting on the top levels of the tree are the variables with higher prediction power.

1.3 Bagging and Random Forest

Bagging is an earlier ensemble method. After the large initial tree achieved, CART usually requires a prune back process because of the overfitting problem. In addition, large individual tree has lower bias but high variance. These problems can be effectively addressed through bagging (bootstrap aggregating) [9]. Instead of growing a tree with the whole data set, bagging builds a tree on each bootstrap sample. The final aggregate classifier can be obtained by averaging (regression) or majority voting (classification). Single classifier based on one tree is unstable with high variance. Bagging reduces the variance using aggregated classifier [6].

In addition to bootstrapping, Random Forest embraces the idea of random subspace [31]. At each split, a random subset of predictors is considered as possible candidates, which further reduces the variance by de-correlating the trees [6]. Thus, Random Forest improve the prediction accuracy by further reducing variance of estimators.

Depending on the nature of the outcome, classification, regression, or survival trees can be grown in Random Forest. The Random Forest algorithm is summarized as following:

1. Draw bootstrap sample from the original data. On average 63% of the original sample will be included in the bootstrap sample and 37% will be left out, which is called out-of-bag data (OOB).

2. A classification, regression or survival tree is grown on the bootstrapped data.

   (a) At each node, randomly select a subset *mtry* of the total p predictors to consider splitting the data on. The default *mtry* is $\sqrt{p}$ for classification and survival trees, while $p/3$ for regression trees.

   (b) Among the *mtry* predictors selected, the optimal split-point is typically identified to minimize the within node error (regression), Gini index (classification), or maximize the log-rank statistic (survival).

    (c) Repeat steps a and b until between-node heterogeneity/within-node impurity ceases to improve or a stopping rule is reached (e.g., minimum node sample size needed to partition the data).

3. Repeat steps 1 and 2 for *ntree* bootstrap samples/trees as desired.

Once a Random Forest is constructed using the above steps, predictions are based on averaging the predicted values from each tree in the forest. Random Forest has a built-in tool for evaluating prediction accuracy that avoids overly optimistic estimates of accuracy because the data used to build the Random Forest is separate from the data used to evaluate its accuracy. Specifically, each tree in a Random Forest is constructed from a subsample of the data (due to bootstrapping) known as in-bag data, the left over OOB data not used to construct each tree are used to evaluate prediction accuracy. Commonly used measures of prediction error are the Brier score (regression), misclassification error (classification), and 1-Harells [27] index of concordance (survival).

1.4 Outline of the Dissertation

The organization of the dissertation is as follows. In Chapter 2, we first give an introduction of random forest of interaction trees (RFIT) and propensity score methods. Then we discuss the causal effect RFIT (CERFIT) algorithm in detail. We also present simulation studies to assess CERFIT's performance with respect to prediction accuracy and variable importance ranking. At last, we illustrate CERFIT through the analysis of an educational dataset from a large public university. The chapter concludes with a brief discussion on the strength of proposed method and the direction of future work.

In Chapter 3, we discuss the rationale and provide evidence on the benefit of using residuals from linear regression model to replace the responses in the algorithm. Then we conduct the simulation studies to investigate the numerical stability benefits of using linear residuals. In Chapter 4, we introduce CERFIT algorithm for non-binary treatment. We start the chapter with introductions on general propensity score methods for the multiple treatments and continuous treatment settings, then present how to integrate the general propensity score into the CERFIT algorithm. We also conduct simulation studies to assess the CERFIT's performance by prediction accuracy and variable importance ranking for multiple and continuous treatment settings, respectively. The application of CERFIT under non-binary treatments is demonstrated through learning optimal ITR under multiple parallel education interventions.

In Chapter 5, we summarize the method and discussthe R implementations through introducing main functions used in the proposed R program. The final chapter provides a discussion on the proposed methods and suggestions for future work.

CHAPTER 2
CERFIT FOR BINARY TREATMENTS

2.1 INTRODUCTION

The estimation of causal effects is challenging since we can not observe both the responses under the intervention and the responses without the intervention for any individual unit. Typically, causal inference relies on Rubin's potential outcomes framework [64, 65]. It is assumed that, for each study unit, there exists potential outcomes under the opposite treatment assignment regime. The casual treatment effects are identifiable under the "zero bias" or "no confounding" condition. Theoretically, confounding can be controlled either through the research design or in data analysis processes. The estimation of causal effects can be done at different levels, such as population, subpopulation, and unit [32]. The estimand of interest varies for different causal inference levels, such as the average treatment effect (ATE) at the population or subpopulation levels, and the individual treatment effect (ITE) at the unit level.

Under Robin's framework, recent endeavors to estimate the individualized treatment effect can be generally categorized into two different approaches. The first approach is the separate counterfactual model. Let $Y_i(1)$ and $Y_i(0)$ denote the treated and untreated outcomes for an individual unit with a set of baseline covariates X_i. Then the individualized treatment effect (ITE) is defined as the conditional difference between the two outcomes: $\mathcal{T}_i = E[Y_i(1)|X_i] - E[Y_i(0)|X_i]$, where $\mathcal{T}_i$ is the ITE for an individual unit i. Under this model, the estimation of ITE is deemed as a general "regression" problem. For any individual unit i, we observe only one Y_i under one of the treatment options; but the dataset contains units under either treatment group. Using available data, $Y_i(1)$ and $Y_i(0)$ can be modeled and estimated separately. Then the ITE can be estimated by the differences between the two outcomes. This framework usually involves two separate models, $Y(1) = f_1(X) + \epsilon_1$ and

$Y(0) = f_0(X) + \epsilon_0$. Representative papers under this framework include separate regression [76, 23], counterfactual synthetic random forest [49], and Bayesian additive regression trees or BART [29]. Under this approach confounding issues are arguably bypassed through the precise estimation of the responses or outcomes under two treatment regimes [29]. If the response is correctly modeled and precisely estimated for the two groups, the ITE can be estimated without bias. However, the existence of strong selection bias in some observational studies may compromise the prediction accuracy. For instance, with a dataset collected from a program enrolled mostly with male participants, we may have an unreliable estimation of female participant's responses.

Another approach is the so-called direct estimation model. As opposed to the separate counterfactual model, the direct approach estimates treatment effects in one model ($\mathcal{T} = \delta(X) + \epsilon$) using all of the data. The estimation of ITE is treated as an approximation problem. The primary idea of this method is that ITE can be approximated by the average treatment effect (ATE) of a subgroup g ($\hat{\mathcal{T}}_i \approx \widehat{ATE_g}$) when the subgroup g is small enough that it contains only subjects having homogeneous treatment effects. Representative papers under this approach include causal random forest [77] and random forest of interaction trees [43, 69, 68]. Although both methods are tree-based and involve recursively partitioning the data into two child nodes with greatest heterogeneity of treatment effects, their similarities end there. Causal random forest (CRF) splits the data by maximizing the variance of the treatment effect, while random forest of interaction trees (RFIT) chooses the split that maximizes the interaction effect with the treatment. The advantage of direct modeling is that it utilizes the data more efficiently, learning one model instead of two models under the separate counterfactual approach. Hence, it may achieve higher prediction accuracy. In addition, direct estimation model approaches allow variable importance estimates with respect to differential treatment effects. Methods under the counterfactual framework (such as BART) can provide variable importance rankings only with respect to treated or untreated outcomes, which is not directly relevant for establishing subgroups with the most differential treatment effects (where the treatment or intervention is the most or least useful). However, as the direct approach involves estimating the average treatment effect (in small groups with homogenous treatment effect) , the strong ignorability condition requires it to address issues of confounding when using observational study data. This usually involves appropriately controlling for or specifying the treatment assignment mechanism. However, neither CRF nor RFIT explicitly addresses this issue.

In this chapter, we propose to extend random forest of interaction trees (RFIT) to a causal effect RFIT (CERFIT) using propensity scores. With no random assignment of treatment in observational studies, selection bias can result in systematic differences in baseline covariates between the two groups. Propensity score adjustment is one of the most frequently used methods to address selection bias in observational studies [4, 63]. By integrating propensity score into the tree growing process, subgroups from the proposed causal effect random forest of interaction trees (CERFIT) not only maximize treatment effect differences, but also achieve similar baseline covariates within each terminal node. Thus it allows for estimation of the individualized treatment effect using observational data as well as variable importance rankings with respect to differential treatment effects.

This chapter is structured as follows. In Section 2.2, we first give an introduction of random forest of interaction trees (RFIT) and the propensity score, then present the causal effect RFIT (CERFIT) algorithm in detail. In Section 2.3 we present simulation studies to assess CERFIT's performance with respect to prediction accuracy and variable importance ranking. In Section 2.4, we illustrate CERFIT through the analysis of an educational dataset from a large public university. Section 2.5 concludes this chapter with a brief discussion.

2.2 Random Forest of Interaction Trees

Random forest [9] is a nonparametric machine learning method. The basis of random forest is classification and regression trees, or CART [8]. Tree based methods recursively partition the data using binary splits until some stopping criteria are met. Specifically, a tree is grown by splitting the root node into two child nodes that maximize between-node heterogeneity, or equivalently, minimize within-node impurity. A node that cannot be split any further is called a terminal node. Each terminal node is characterized by a unique combination of the attributes. Random forest (RF) is an ensemble method. Instead of growing a tree with the whole data set as with CART, RF builds multiple trees. Each of the trees is grown with a bootstrap sample through bagging. In addition, RF embraces the idea of "random subspace" [31] in the splitting process. In each split, the CART algorithm searches all possible variables and all possible cut-points; while in RF, only a random subset of predictors are considered as possible splitting candidates. By doing so, RF increases the variance through de-correlating the trees [6], and hence improves the prediction accuracy. Predictions in RF are based on averaging the predicted values from each tree in the forest. Depending on the nature of the outcome, different splitting rules can be defined for different types of RF. Typically, the optimal split-point is identified to minimize the within node error for regression trees, minimize the Gini index for classification trees [9], or maximize the log-rank test statistic for survival trees [37].

Random forest of interaction trees, or RFIT [69], essentially follows the routine of RF. Instead of creating regression or classification trees, RFIT uses bootstrap samples to create interaction trees. The optimal split-point is chosen by splitting the data into two child nodes that maximizes the differences in treatment effects. Consider a candidate binary split s that divides a node c into two child nodes, left child node c_L and right child node c_R, within which $\{Y_L(1), Y_R(1)\}$ are treated responses and $\{Y_L(0), Y_R(0)\}$ are untreated responses.

	c_L	c_R
$T = 1$	$Y_L(1), n_1, s_1^2$	$Y_R(1), n_3, s_3^2$
$T = 0$	$Y_L(0), n_2, s_2^2$	$\mathrm{Y}_R(0), n_4, s_4^2$

Here T denotes the treatment assignment, which is 1 for treated and 0 otherwise, and $\{n_1, \ldots, n_4\}$ and $\{s_1^2, \ldots, s_4^2\}$ denote sample sizes and sample variances in the four cells determined by the split and the treatment indicator. The average treatment effect for the two child nodes can be defined as $ATE_L = \frac{1}{n_1}\sum_{i=1}^{n_1} Y_L(1) - \frac{1}{n_2}\sum_{i=1}^{n_2} Y_L(0)$ and $ATE_R = \frac{1}{n_3}\sum_{i=1}^{n_3} Y_R(1) - \frac{1}{n_4}\sum_{i=1}^{n_4} Y_R(0)$. The t-test statistic for split s is defined as

$$t(s) = \frac{ATE_L - ATE_R}{\hat{\sigma}\sqrt{1/n_1 + 1/n_2 + 1/n_3 + 1/n_4}}, \tag{2.1}$$

where $\hat{\sigma}^2$ is the pooled estimator of the constant variance ($\hat{\sigma}^2 = \sum_{i=1}^{4} \frac{(n_i-1)s_i^2}{n-4}$, $n = \sum_{i=1}^{4} n_i$). The best split s^* is chosen by maximizing $t^2(s^*)$. By simple deduction, it is not hard to show that (2.1) is equivalent to the Wald test statistic for $H_0 : \beta_3 = 0$ in the interaction model

$$Y_i = \beta_0 + \beta_1 I(T_i = 1) + \beta_2 I(X_{ij} \leq c) + \beta_3 I(T_i = 1) I(X_{ij} \leq c) + \varepsilon_i, \tag{2.2}$$

where $I(\cdot)$ is the indicator function. T_i is the treatment assignment for the i^{th} subject, $I(X_{ij} \leq c)$ is the indicator for a binary cut based on covariate X_j, and $\varepsilon_i \sim N(0, \sigma)$.

After the best split s^* is chosen, the tree is grown recursively until it reaches the predetermined maximum tree depth or minimum terminal node size. The estimation of the individualized treatment effect, $\mathcal{T}_i$, is based on the average treatment effect $ATE_t = E[Y_t(1)] - E[Y_t(0))]$ within the terminal node t, to which the i^{th} subject belongs. And the final prediction for each subject is the average prediction across all trees [69].

2.3 Propensity Score Methods

The estimation of causal treatment effect under the potential outcome framework usually assumes the strong ignorability condition: $\{Y(1), Y(0)\} \perp\!\!\!\perp T|X$ and $0 < Pr(T = 1|X) < 1$, where T is the treatment indicator and X is a set of baseline covariates. This condition assumes that the treatment assignment is independent of the potential outcomes given a set of covariates, and the probability of treatment selection is between 0 and 1, exclusive. It is a sufficient condition for causal effect estimation [55], $E[Y(0)|X] = E[Y|X, T = 0]$ and $E[Y(1)|X] = E[Y|X, T = 1]$. Note that, in observational studies, treatment is not randomly assigned among study subjects, but associated with the study subject's baseline covariates. Propensity score adjustment is one of the most frequently used methods to address this confounding issue. The propensity score, e, is defined as the probability of treatment conditional on a set of covariates, $e = Pr(T = 1|X)$. In their seminal work, Rosenbaum and Rubin (1983) show that an unbiased estimate of the average treatment effect can be obtained by conditioning on the propensity score alone, instead of the set of covariates: $\{Y(1), Y(0)\} \perp\!\!\!\perp T|e(X)$. This approach has been widely applied in causal inference and several methods based on propensity score have been proposed, such as matching, stratification, inverse probability of treatment weighting (IPTW), and covariate adjustment [3, 4]. With propensity score matching, treated and untreated units are matched based on similar values of the propensity score. Then the matched samples can be analyzed as if data were obtained from a randomized trial. The stratification method approximates a quasi-randomized experiment. Study subjects are stratified into several strata based on the quantile values of the propensity score. The baseline covariates of the units within the same stratum will be similar if the propensity score is correctly modeled. Under IPTW adjustment, a weight based on the propensity score is used to create a synthetic sample within which the covariate distribution for one treatment group is similar to that for the other treatment group. Thus the distribution of the confounders is independent of the treatment assignment, allowing for an unbiased estimate of the average treatment effect. Under the covariate adjustment approach, the propensity score is used as a new covariate in the modeling. This method reduces confounding and allows the estimation of the outcome associated with treatment while adjusting for the propensity score, which contains information on a set of confounders [3].

2.4 CERFIT for Binary Treatments Algorithm

The random forest of interaction trees, or RFIT, algorithm was developed for estimating subgroup average treatment effects (ATEs) using data from randomized trials [69].

To extend the RFIT for use with observational study data, we propose to use the propensity score to address the confounding issues through three methods.

The fundamental differences between a randomized trial and an observational study is the treatment assignment mechanism. In observational studies, the baseline characteristics for individual units in the treated group often differ systematically from their counterparts in the control group due to self selection of the treatment assignment. To address this issue, we propose to use IPTW to adjust the data first and grow the trees using subsamples with more balanced baseline covariates between two treatment groups. Specifically, the IPTW weight is defined as $w = \frac{T}{e} + \frac{1-T}{1-e}$. A unit with lower probability to be included in the treatment group ($T = 1$), will receive a higher weight to be included in the bootstrap sample, and vice versa. In the same manner, a unit with higher probability to be include in the control group ($T = 0$) will receive a lower weight, and vice versa. In random forest (RF) or random forest of interaction trees (RFIT), each tree is grown based on a bootstrap sample. In our proposed causal effect random forest of interaction trees (CERFIT), each tree is built based on a weighted bootstrap sample selected using weights w. However, previous work by Xu et al.[78] found that weighting each subject by IPTW may inflate the sample size and type I error rate under the null model of no treatment effect. A stabilized IPTW is usually suggested to address this issue [62]:

$$w_s = \frac{T \cdot Pr(T = 1)}{e} + \frac{(1 - T) \cdot Pr(T = 0)}{1 - e}, \tag{2.3}$$

where $Pr(T = 0)$ and $Pr(T = 1)$ are marginal probabilities for each of the treatment groups. Stabilized IPTW helps mitigate a common issue in the practice of IPTW, extremely large weights, by reducing the weights in general. An observation with a very low propensity score in the treated group or a very high propensity score in the control group will receive a very large weight. In the context of random forest, this will increase the similarity of subsamples used for each tree, and thus compromise the ensemble accuracy. Therefore, several studies suggested further truncating the w_s using the quantiles of the weight distribution [17, 47]. We recommend using the 10^{th} and 90^{th} percentiles as thresholds for the CERFIT algorithm.

A second extension to accommodate observational data is to use the propensity score to adjust the RFIT splitting rule, further controlling issues of confounding during the tree growing process. To this end, we use the propensity score as a blocking covariate in the interaction model

$$Y_i = \beta_0 + \beta_1 I(T_i = 1) + \beta_2 I(X_{ij} \leq c) + \beta_3 I(T_i = 1) I(X_{ij} \leq c) + \beta_4 e_i + \varepsilon_i, \tag{2.4}$$

where e_i is the propensity score for the i^{th} subject and ε_i is $iid\ N(0, \sigma)$.

Thirdly, we address issues of confounding by utilizing the weighted average treatment effect (ATE). When the model for estimating a propensity score is correctly specified, Lunceford and Davidian [51] demonstrates that the ATE can be consistently estimated by two alternative weighted ATE's [51]: 1) $\frac{1}{n_1}\sum_{i=1}^{n_1} w_i(1)Y_i(1) - \frac{1}{n_0}\sum_{i=1}^{n_0} w_i(0)Y_i(0)$ or 2) $\frac{\sum_{i=1}^{n_1} w_i(1)Y_i(1)}{\sum w_i(1)} - \frac{\sum_{i=1}^{n_0} w_i(0)Y_i(0)}{\sum w_i(0)}$, where n_1 and n_0 are the sample sizes for the treated and untreated groups. In CERFIT, we utilize the second estimator since it has smaller variance compared to the first one [51]. Using the truncated weights, the average treatment effect in the terminal node of CERFIT is calculated as

$$ATE_w = \frac{\sum w_i'(1)Y_i(1)}{\sum w_i'(1)} - \frac{\sum w_i'(0)Y_i(0)}{\sum w_i'(0)}, \tag{2.5}$$

where w_i' is the truncated stabilized weight for the i^{th} subject.

The propensity score plays a critical role in the proposed CERFIT algorithm, thus it is essential to utilize a robust method in the propensity score estimation. We recommend random forest (RF) for the propensity score analysis as a large body of work has demonstrated random forest's superior prediction accuracy in various data situations [13, 12]. Moreover, Lee, Lessler, and Stuart [46] show that the propensity scores estimated using RF are able to balance covariates better than those estimated using logistic regression. In the *randomForest* R package [7], the default $mtry$ value for the regression problem is set as $mtry = p/3$, where p is the number of predictors. We recommend using $max\{3, p/6\}$ as the default $mtry$ value in CERFIT. The reason for this recommendation is that using weighted bootstrap samples may increase the similarity of the trees. A smaller $mtry$ can help further de-correlate the trees. In particular, $p/6$ is half of the default $mtry$ value in the *randomForest* [7], R package and the value of 3 in our recommendation is designed to stay away from an $mtry$ value that is too small, when p is not very large, in order to preserve the quality of splits and ultimately the prediction accuracy of the random forest. The default terminal node size is set at 10, with minimum size for each of the treatment groups set at 5. The detailed CERFIT procedure is summarized in the CERFIT algorithm.

Table 2.1. The CERFIT Algorithm for Binary Treatments

```
The CERFIT algorithm for binary treatments
1. Estimate propensity scores using random forest with treatment indicator as outcome.
2. Draw bootstrap samples from the data using $w'$ (truncated $w_s$) as sampling weights.
3. Grow an interaction tree based on each weighted bootstrap sample.
    −3.1 At each node, randomly select a subset $mtry$ of the total $p$ covariates from which to determine a split rule. The default value of $mtry$ is set at $max\{3, p/6\}$.
    −3.2 Among the $mtry$ covariates selected, the optimal split is identified by maximizing the squared Wald test statistic for testing $H_0 : \beta_3 = 0$, in equation (2.4).
    −3.3 Repeat steps $3.1$ and $3.2$ until reaching a pre-specified stopping rule (e.g., maximum tree depth, minimum terminal node size).
4. Repeat steps 1 to 3 for $ntree$ trees as desired, with a default value set at $500$.
```

2.5 Variable Importance Ranking

Compared to the separate counterfactual framework for estimating the individualized treatment effect, one important advantage of the proposed CERFIT algorithm is the availability of variable importance rankings. Permutation based variable importance score (VIMP) is one of the most frequently used methods of obtaining variable importance rankings. In random forest [9], variable importance is ranked using VIMP. The relative VIMP for a particular variable is calculated by comparing the difference in the prediction error of the out-of-bag (OOB) data to the prediction error when the variable is noised up by randomly permuting its values. The larger the VIMP value, the higher the predictive power of the variable. Following the variable importance ranking scheme from Random Forest, we permute the product of a variable and the treatment in CERFIT, and calculate the change in the squared Wald test statistic before and after permutation. The specified steps in finding VIMP (ζ^j) a covariate X_j $(j = 1, 2 \cdots, p)$ are described in the variable importance algorithms below.

Table 2.2. The Variable Importance: Permutation Based Algorithm

Variable importance algorithm: permutation based variable importance score (VIMP)

1. Let Γ_b denote a tree b $(b = 1, \cdots, ntree)$ and O denote the out-of-bag sample that has not been used in creating tree Γ_b.

2. Send O down tree Γ_b and calculate $ST^j(b) = \sum_{i=1}^{m} t^2(s_i)$, the summation of squared Wald test statistics for all m splits ($i = 1, 2, \cdots, m$) within tree Γ_b.

3. Permute the product of X_j and treatment indicator T. Repeat step 2 using permuted $(X_jT)_*$ in place of the original X_jT in O, and calculate $ST^j_*(b)$ with the permuted data.

4. Compute $VI^j_b = \frac{ST^j(b)-ST^j_*(b)}{ST(b)}$.

5. Repeat steps 1 to 4 for every tree in the forest and calculate the mean of the VI^j_b over the forest, $\zeta^j = (\sum_{b=1}^{ntree} VI^j_b)/b$.

2.6 SIMULATION STUDIES FOR PREDICTION ACCURACY

2.6.1 Simulation models

We simulate data with 20 covariates, among which X_j $(j = 1, ..., 11)$ are generated from the standard normal distribution $N(0, 1)$, and X_j $(j = 12, ..., 20)$ are generated from the Bernoulli distribution with $p = 0.5$ as in Lu et al.[49]. Simulation of the treatment selection is modified from the framework used by Setoguchi, Schneeweiss, Brookhart, Glynn, and Cook [66], which has been used by several other studies [4]. We select two models from their originally proposed seven models. We slightly modify the intercept to generate the marginal probability of treatment assignment $Pr(T = 1) \approx 0.2$, which is close to the proportion of treated subjects in our application. The original intercept in their models was all set to 0; we modified it to -1.8 and kept all the other coefficients the same as in their paper. In both models, treatment selection is associated with 7 covariates: $X_1, X_2, X_3, X_4, X_{11}, X_{12}$ and X_{13}. The first model is an additive and linear model, and the second model is a non-additive and non-linear model with four two-way interaction terms and one quadratic term.

Treatment selection model I Additivity and linearity.

$$\begin{aligned} logit(Pr(T = 1|X)) = &- 1.8 + 0.8X_1 - 0.25X_2 + 0.6X_3 - 0.4X_4 \\ &- 0.8X_{11} - 0.5X_{12} + 0.7X_{13} \end{aligned} \tag{2.6}$$

Treatment selection model II Quadratic and interactions

$$\begin{aligned} logit(Pr(T = 1|X)) = &- 1.8 + 0.8X_1 - 0.25X_2 + 0.6X_3 - 0.4X_4 - 0.8X_{11} \\ &- 0.5X_{12} + 0.7X_{13} - 0.25X_2^2 + 0.8X_1X_3 - 0.175X_2X_4 \\ &- 0.2X_4X_{12} - 0.4X_{12}X_{13}. \end{aligned} \tag{2.7}$$

The continuous outcomes are simulated based on the models proposed by Lu et al.[49].

Model I

$$f_1(X,T) = 2.445 - I_{\{T=0\}} \times m(X) - I_{\{T=1,g(X)>0\}} + \varepsilon \tag{2.8}$$

Model II

$$f_2(X,T) = 2.445 - I_{\{T=0\}} \times \sin(m(X)) - I_{\{T=1,g(X)>0\}} + \varepsilon \tag{2.9}$$

Model III

$$f_3(X,T) = 2.445 - I_{\{T=0\}} \times \sin(m(X)) - I_{\{T=1,h(X)>0\}} + \varepsilon, \tag{2.10}$$

where $m(X) = 0.4X_1 + 0.154X_2 - 0.152X_{11} - 0.126X_{12}$,
$g(X) = 0.254X_2^2 - 0.152X_{11} - 0.4X_{11}^2 - 0.126X_{12}$,
$h(X) = 0.254X_3^2 - 0.152X_4 - 0.126X_5 - 0.4X_5^2$, and $\varepsilon \sim N(0,1)$. Note that in Lu's model, the random error term was set up as $\varepsilon \sim N(0,0.1)$.

The complexity of the models increases from model I to model III. In model I, $Y(0)$ has a linear association with covariates X_1, X_2, X_{11} and X_{12}. In models II and III, $Y(0)$ and $Y(1)$ are both nonlinear models. In addition, in model III, $Y(0)$ is associated with X_1, X_2, X_{11} and X_{12}, but $Y(1)$ is simulated with non-overlapping covariates X_4 and X_5. Thus there are four confounders: X_1, X_2, X_{11} and X_{12} in models I and II, but six confounders in model III with two additional confounders, X_4 and X_5.

2.6.2 Simulation settings and parameters

Training data with three sample sizes $n = 500$, $n = 1000$ and $n = 2000$ are used to estimate the individualized treatment effect. The performance is assessed by the mean squared error (MSE) using a test sample $n' = 1000$ based on 200 simulation runs. For each scenario, a total of 500 trees are grown and the value of $mtry$ is set at 3.

We compare the performance of CERFIT with three other methods: synthetic random forest (synRF) [49], Bayesian additive regression trees (BART) [29], and causal random forest

(CRF) [77]. Both synRF and BART are separate counterfactual models, while both CFR and CERFIT are direct models.

synRF is designed to improve the prediction accuracy of random forest (RF) through synthetic features. Specifically, synRF grows random forests using different values for $mtry$ and terminal $nodesize$. It then calculates the predicted values based on each RF, which are the so-called synthetic features. The prediction of the individualized treatment effect (ITE) is based on the two separate synthetic forests fitted with synthetic features and original features [40]. The simulation is implemented using the *randomForestSRC* R-package [38]. Two separate forests are constructed for the two treatment groups, each with $ntree = 1000$. The parameters for each of the trees are all the combinations of nodesize values $\{1, \cdots, 10, 20, 30, 50, 100\}$ and $mtry$ values $\{1, 10, 20\}$ as recommended by Lu et al. (2018).

BART is a sum-of-tree model based on Bayesian regularized trees. Each consecutive tree refits the residuals that are not explained by the other trees. Fitting and inference procedures are done through the iterative Bayesian backfitting MCMC algorithm [15]. The implementation of BART is performed through the R-package *BayesTree* [16] with default settings $ntree = 200$ and 1000 MCMC iterations.

CRF is another type of random forest that modifies the splitting rule to maximize the between nodes treatment effect heterogeneity. It differs from random forest of interaction trees by choosing the split that maximizes the variance of $\hat{\mathcal{T}}$. In addition, different from regular RF, CRF builds double-sample trees to obtain honest splitting rules. Specifically, a randomly selected subset of data is first divided into two equal halves. A base learner is grown with one half of the data, and the estimation is based on the other half of the data. The splitting rule is honest because the treatment effect is estimated by Y_i without being used for evaluating the best split [77]. For CRF, we use the R-package *grf* [73] with default settings $ntree = 2000$ and $mtry = p/3$.

2.6.3 Simulation results

Figure 2.1 presents the box plots of MSE obtained from 200 simulation runs for the four methods. The boxes are color coded to reflect the three different sample sizes used to learn the model. In general, the prediction accuracy is significantly improved as sample size increases from 500 to 2000 for all four methods. It can be seen that the proposed method, CERFIT, outperforms the other three methods consistently under all the scenarios considered. CFR is the runner-up under models I and II with treatment selection I, and under almost all scenarios with moderate sample sizes ($n = 500$ and $n = 1000$). synRF's performance is quickly improved as sample size increases from $n = 500$ to $n = 2000$. This improvement is especially prominent under model III. BART has the worst performance when sample size is smaller than $n = 1000$, but catches up under sample size $n = 2000$. The MSE for BART is similar to CRF when sample size reaches $n = 2000$. In addition, it is worth mentioning that, when sample size is small, models under the direct approach are superior to the models under separate counterfactual approach in general. One plausible explanation is that the separate counterfactual approach models are learning with insufficient data, since the data has to be further divided into two subsets based on treatment status. The performance of the separate counterfactual approach models improves as sample size increases. When $n = 2000$, syRF has the second best performance behind CERFIT under most scenarios considered.
Comparing the performance under two different treatment selection models, the advantage of CERFIT is more significant under treatment selection model II, under which the treatment selection is simulated with a quadratic regression model with interaction terms. This can be explained by the fact that CERFIT is the only method that directly addresses the selection bias issue in observational studies among the four methods.

2.7 Simulation Studies for Variable Importance Ranking

To evaluate the variable importance algorithm outlined in Section 2.4, we simulate 8 covariates from the standard normal distribution $X_i \sim N(0, 1), i = 1, 2, \cdots, 8$. Treatment selection is based on a linear additive model using four covariates X_1, X_2, X_5 and X_6,

$$\text{logit}(Pr(T = 1|X)) = -1.5 + 0.8X_1 - 0.25X_2 + 0.6X_5 - 0.4X_6. \tag{2.11}$$

The outcomes are simulated with two linear models with a common response model for the control group ($T = 0$), but two different models for ITE, $\delta(X)$. Let

$$f(X, T) = 0.5 + 0.5X_3 + 0.5X_4 + 0.5X_5 + 0.5X_6 + I_{T=1}\delta(X) + \varepsilon. \tag{2.12}$$

The individual treatment effects $\delta(X)$ are determined by the two models

$$\delta_A(X) = N(0, 1) \tag{2.13}$$

$$\delta_B(X) = 0.5 + X_1 + 1.5X_2 + 2X_3 + 2.5X_4. \tag{2.14}$$

Model A, $\delta_A(X)$, is a null model, in which the treatment effects are random numbers generated from the standard normal distribution. Therefore, none of the covariates in model A have any effect on the treatment. With model B, $\delta_B(X)$, there are two confounding variables (X_1 and X_2) that affect both treatment selection and the outcome. Two covariates (X_3 and X_4) affect the outcome alone. The ITE in model B is associated with only four covariates (X_j, $j = 1, \cdots 4$). The coefficients are assigned in a way that the true variable importance increases from X_1, X_2, X_3 to X_4, with X_4 having the highest predictive power. The rest of the covariates, X_5 to X_8, are equally unimportant.

Simulation results based on the permutation based variable importance score with 200 simulation runs are presented in Figure 2.2. For both models A and B, CERFIT variable importance algorithms correctly identified the variables that impact the treatment effects. The relative variable importance measures are consistent with the underlying truth. Figure 2.2 presents the permutation based variable importance scores. The box plot on the left shows that the distributions of the variable importance scores are similar for all the variables as they should be in model A. On right panel in Figure 2.2, we can see that covariate X_4 is being identified as the most important, followed by X_3, X_2, and X_1. Although the covariates X_5 and X_6 are associated with treatment selection and the outcome, they have no impact on the ITE and therefore are identified as being equally unimportant as X_7 and X_8.

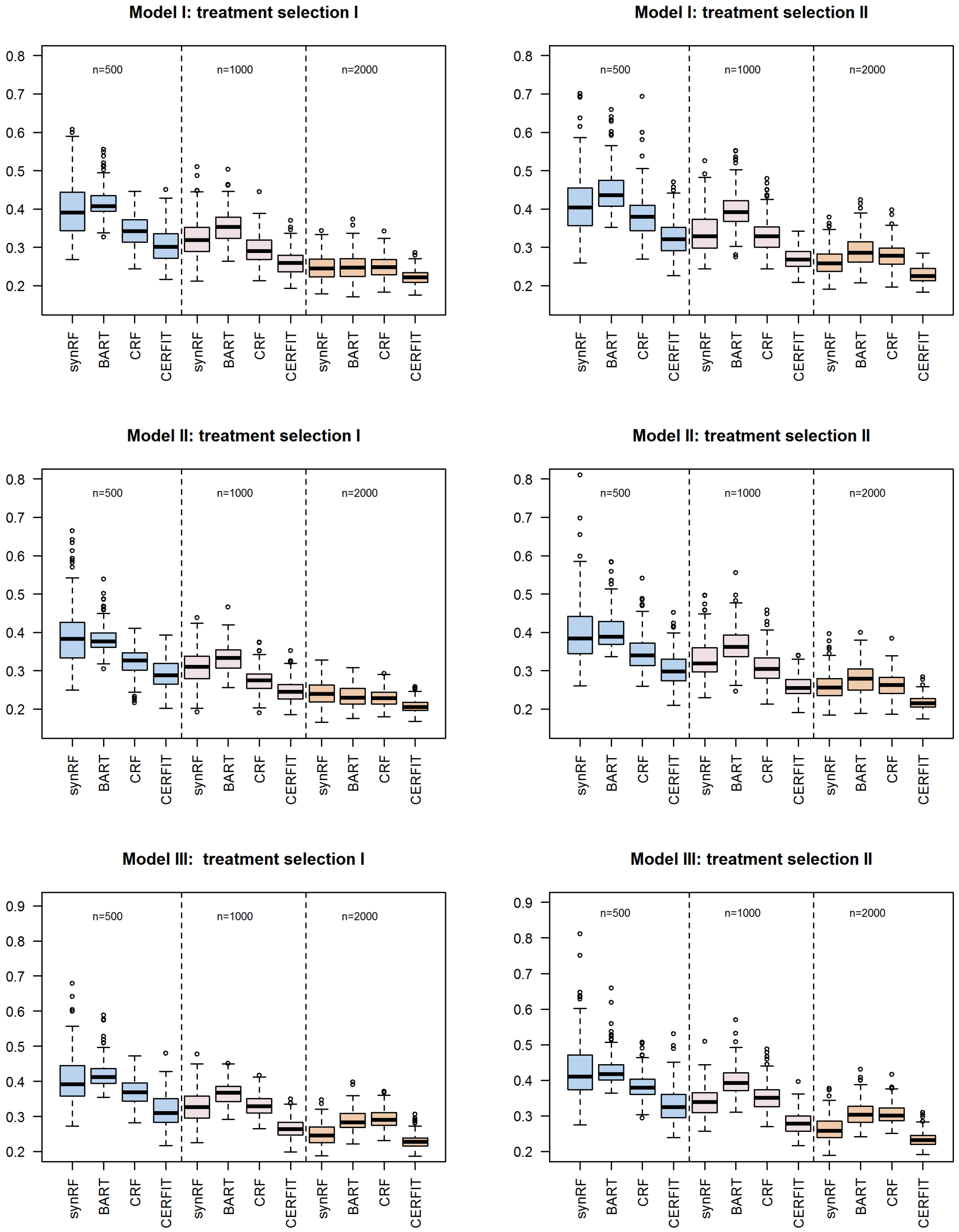

Figure 2.1. Simulation results: comparison of four methods in terms of prediction accuracy. Training data with three sample sizes (500,1000, and 2000) are used to estimate the individualized treatment effect (ITE). The MSE for the ITE is based on a test sample of size 1000, and 200 simulation runs.

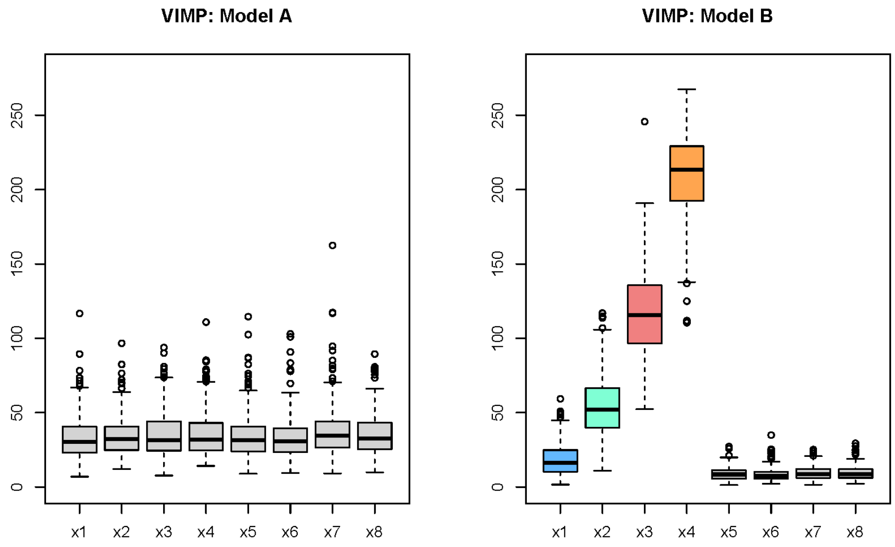

Figure 2.2. Simulation results: permutation based variable importance score (VIMP) with the proposed CERFIT algorithm.

2.8 Application Study: Assessment of a Voluntary Supplemental Instruction Course

The California State University Chancellor's Office identified Introductory Statistics as a bottleneck course, in particular the relatively high failure and repeat rate delaying student graduation across the system. As a first phase to addressing the issue, the CSU offered so-called Course Redesign with Technology grants to consider alternative instructional modalities towards improving student success. As part of one such grant, San Diego State University introduced a one-unit supplemental instruction course. Funding was available to offer the course to only 20% of Introductory Statistics students. Students voluntarily enrolled in the course and met twice per week in a small group, active learning environment to review the topic of the week, discuss conceptual issues, and work on extra but related statistics problem sets and data analyses. In this Section, we use CERFIT to assess this supplemental instruction section with respect to student performance in the course. One particular actionable outcome is to determine financial resources devoted to the section: either increase offerings if successful, or introduce/consider an alternative intervention if not successful.

There were $n = 976$ students enrolled in the course, among which 182 students (18.65%) enrolled in the one-unit supplemental instruction section ("treatment group"). A total of 37 covariates are considered in the model besides the treatment status. There are 8 continuous variables, 18 binary variables, 7 ordinal, and 4 nominal variables. The data covers students' demographics, university information, course specific information, as well as admission information; see Table 2.4 for variable descriptions and missing data information. The outcome variable is the final exam score ranging from 0 to 300. The missing values are imputed using the R-package *mice* [75]. The propensity score for each of the students is estimated using random forest.

To evaluate the impact of the supplemental instruction course on individual students and identify important variables that impact the treatment effect, we use a 100-fold cross validation procedure. The original data is randomly split into 100 almost equal-sized groups, with each group containing 9 or 10 students. Then we use data from 99 groups as training data to grow the CERFIT and leave out one group of data as testing data to make predictions. A total of $ntree = 500$ trees are constructed for each forest and 100 forests of interaction trees are built for the 100-fold cross validation. Using our default $mtry$ formula, $mtry$ is set at 6. Variable importance is measured by the average of VIMP for each variable across 100 forests.

A histogram of the predicted individualized treatment effect (ITE) is presented in Figure 2.3. The average predicted ITE is around 10.7. This result suggests that the supplemental instruction course has a small but positive impact on students' performance overall. The predicted ITE values around zero indicate no treatment effect on those students' performance. There are 741 (75.92%) students with positive ITEs and the predicted maximum score gain on the final exam is around 66. There are also 235 (24.08%) students with negative ITEs and the minimum predicted value is -29, which indicates that the program has no impact or even adverse impact on some of the students.

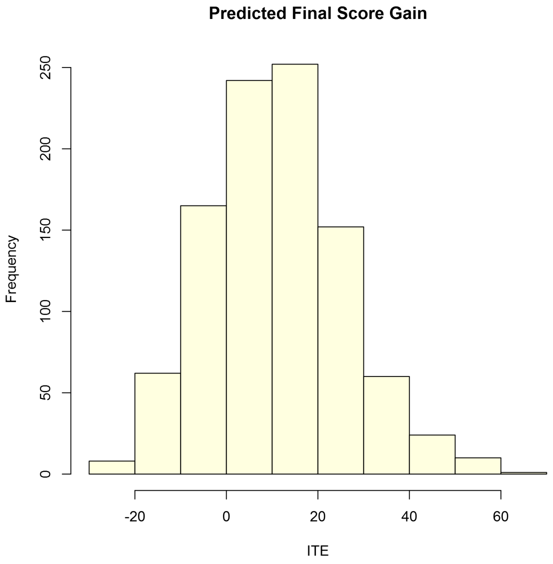

Figure 2.3. Predicted individualized treatment effect (ITE) using CERFIT. The outcome is the final exam score in Introductory Statistics, ranging from 0 to 300. The treatment is the one-unit, self-selected supplemental instruction course.

The variable importance rankings based on the permutation based variable importance score (VIMP) are presented in Figure 2.4. Variables with higher importance are on the top of the figure. Note that larger values on the permutation based variable importance score indicate more important variables. From Figure 2.4, we can see that SAT math and verbal scores, high school GPA, and age are identified as the most important predictors. Other important covariates include college description proxy to major, Homework 1 time and score, indicator of a first-generation student with parents having some college experience, low income based on EFC, location of highest math class taken and highest math class complete, and number of units attempted for the term. Variables with least predictive power are whether the student is disabled, in an honor or scholarship program, or at full-time status. Although the literature is not consistent on how student's academic ability impacts on the treatment effect from academic supplemental instruction program[18], student's academic readiness reflected by the student's SAT performance, high school GPA, and first homework performance are generally considered as top factors that impact the treatment effect. In addition, student's socioeconomic background measured by the first-generation status and low income are also key predictors [56]. Another interesting finding is that the number of units attempted plays a critical role on whether a student can benefit from enrolling in the supplemental instruction course. Lower number of units attempted associates with a higher ITE.

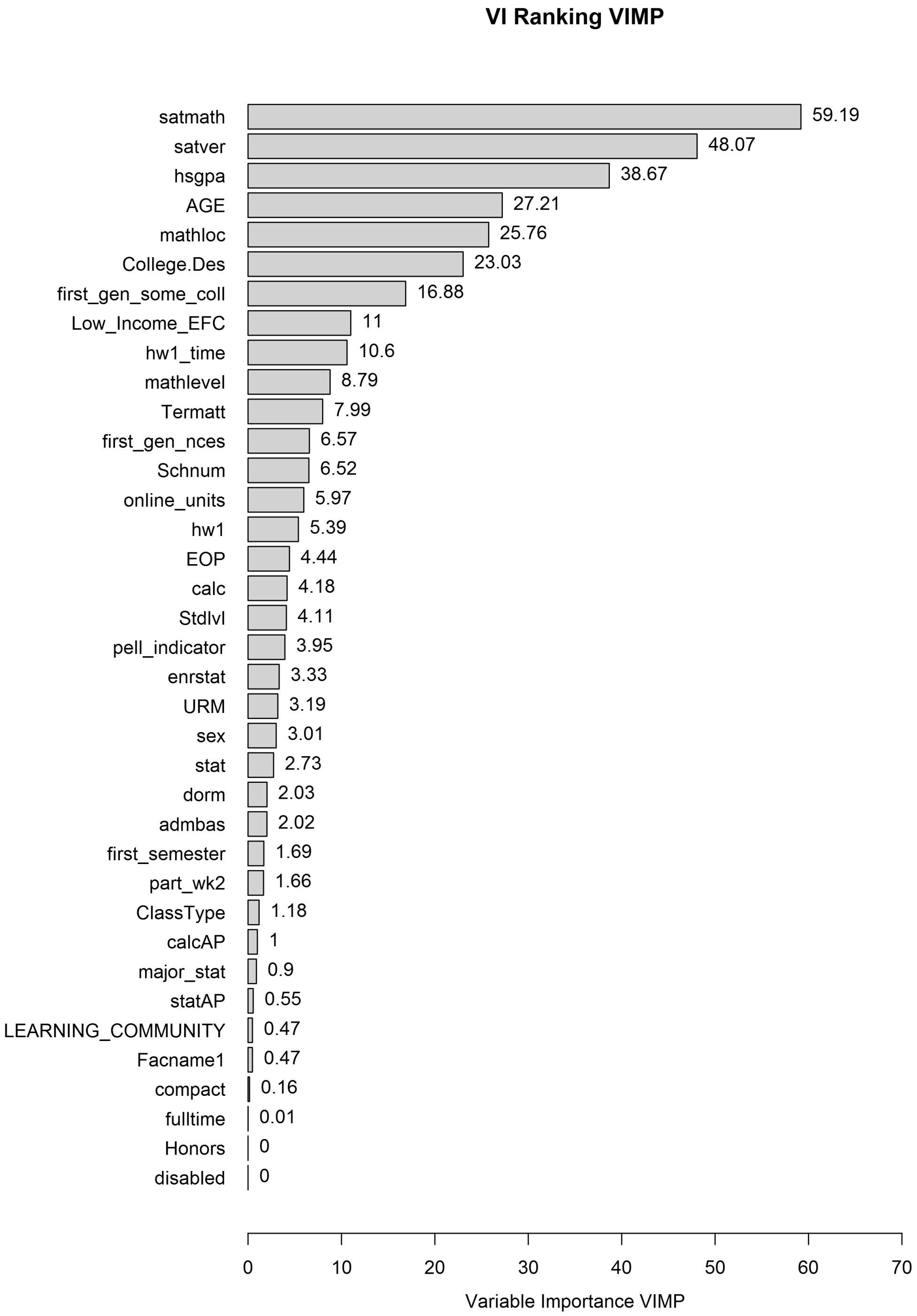

Figure 2.4. Important predictors based on variable importance score.

Table 2.3 presents a profile of students with the largest and smallest predicted ITEs. The four cases that are most negatively impacted as well as the four cases that are most positively impacted by enrolling in the supplemental instruction course are profiled individually. In addition, students with the bottom 10% as well as top 10% ITE's are profiled as groups.

From Table 2.3 we can see that students' with the most negative ITEs have comparatively higher SAT math scores. Very few of them are first-generation college students. Although a negative ITE may seem counter-intuitive, it may be that stronger students who enroll the supplemental instruction course gain little benefit from the class material. In fact, these students may be substituting the supplemental instruction course for study time and thus perform worse in the course. Along these lines as well, these students who enroll in the supplemental instruction course may be less motivated students and effectively using the course to avoid the harder work of studying. Alternatively, although these students are stronger on average, students with a negative ITE may find the Introductory Statistics course material harder, thus motivated to enroll in the supplemental instruction course.

From Table 2.3 as well, we see that students with the most positive predicted ITEs have lower SAT math scores. A large proportion of them are first-generation college students. These latter students are also older, spent longer time to finish the first homework, and enrolled in relatively fewer units for the term. Even though this group of students have similar high school GPA, they are weaker in math preparation as judged by the SAT math score. One may conjecture that by enrolling in the supplemental instruction course, they improve their success in the course through greater time-on-task and more specifically further review of course material and extra practice in statistical problem solving. In addition, their comparable GPA in high school might suggest that these students with lower socioeconomic status on average are generally good and slightly more mature students who are motivated to improve performance in the introductory statistics course through extra efforts in the supplemental instruction course. Students with good math readiness could not benefit from the extra practice problems, especially when they already have a higher course load for the term. They could even be adversely impacted by enrolling in the program since, again, they may benefit more from studying by themselves.

Table 2.3. Profile of Students with the Largest and Smallest Predicted ITE.

Rank	*ITE*	*SATmath*	*hsGPA*	*Age*	*HW1time (hour)*	1^{st}*Generation College*	*#Units*
Bottom four	-29	680	3.62	18.54	1	0	12
	-28	540	3.50	18.75	1	0	15
	-25	580	3.46	18.65	1	0	15
	-25	570	3.50	18.70	1	0	15
Bottom 10%	-13.15	616.3	3.57	18.89	1.5	0.04	14.57
Top four	54	330	3.67	18.24	1.5	1	12
	56	480	4.00	18.51	2.5	1	13
	58	480	3.75	18.08	2	1	12
	66	490	3.93	18.27	2	1	12
Top 10%	38.53	454.1	3.47	20.41	1.65	0.76	13.54

2.9 Conclusions

Estimation of causal treatment effects, especially individualized treatment effects (ITEs) using observational data, holds great interest in various research fields. However, causal inference under the counterfactual framework usually requires the assumption of strong ignorability. Without properly addressing issues of confounding when using observational study data, the estimation of ITE can be biased and unreliable. Propensity score based methods are theoretically appealing and are one of the most widely used methods in causal inference involving confounders. Current machine learning methods with high prediction accuracy may help mitigate the confounding issue significantly, however, integrating the propensity score into machine learning methods enables us to take advantage of propensity score methods to improve predictive performance under causal inference.

Simulation results show that CERFIT outperforms other competing methods consistently under all the scenarios considered in the study. Unlike the models under the separate counterfactual approach that require two separate models on two targeted responses, CERFIT's learning target is set as the treatment effects themselves. Fitting one model using all of the data results in a higher prediction accuracy, as we see in CERFIT. Previous research has found that synRF and BART have superior performance to CRF [49] even when the sample size is moderate ($n = 500$). One reason might be that in this study, the error terms for the responses are simulated with $\varepsilon \sim N(0, 1)$ rather than $\varepsilon \sim N(0, 0.1)$ as in the previous study. The bigger size of error terms increases the learning difficulty on the responses, but not necessarily the treatment effects. Therefore, models under the direct approach, such as CRF and CERFIT, have competitive edges over their competitors. In addition, we observe that CERFIT's superiority is especially significant when the treatment selection has a nonlinear relationship with the covariates; in this case the benefits of adjusting confounding through propensity scores are prominent.

The CERFIT algorithm also produces accurate variable importance ranking, which cannot be achieved by the methods using separate counterfactual modeling. Machine learning methods such as RF are often called "black-box" methods. Variable importance ranking is one of the most important tools to help in interpreting the findings from the "black-box". As we showed in the application section, the accurate prediction of ITE is fundamental and plays a critical role in advising students in their enrollment decisions. However, to understand why and how a student can benefit most from an intervention program is another key question sought by program designers and participants. For instance, based on our analysis, students from lower socioeconomic background with higher high school GPA but lower SAT math score should be encouraged to enroll in the statistics supplemental instruction section.

CERFIT can be applied to various areas where estimation of personalized treatment effects using observational education data is of interest. Since CERFIT depends on IPTW, its performance can be greatly impacted by the accuracy of the propensity score estimation. The selection of proper prediction model for the propensity score is critical. We recommend RF or other machine learning methods such as boosting [53] and neural networks because of their superior prediction accuracy, rather than classic logistic regression. Currently, CERFIT can be applied only in the setting of binary treatment groups. In addition, just as random forest variable importance measures are in favor of variables with many possible splits [67], CERFIT's variable importance measures may suffer from the same bias. Since CERFIT uses half of $mtry$ than the Breiman random forest default, this issue is slightly mitigated. For studies dealing with variables of different types, approaches suggested by [67] and [11] can be combined with the current CERFIT's variable importance algorithms to achieve unbiased variable selection.

Table 2.4. Detailed Variable Description for the Application Data .

Covariates	Description
Demographics	
Sex	Sex (Male or Female)
Age	Age in years
URM	Underrepresented Minority (yes or no)
Disabled	Disabled (yes or no)
LowIncome	Low Income based on EFC (yes or no)
PellGrant	Pell Grant Recipient (yes or no)
FirstGen_NCES	First Generation Student based on the NCES (yes or no)
FirstGen_SomeCollege	First Generation with some college experience (yes or no)
Course information	
SectionNum	Section Number (4 levels)
Instructor	Instructors Name (2 faculties)
ClassFormat	Class Format (Traditional or Hybrid)
PartWK2	Week 2 Participation
HW1_Score	Homework 1 Score
HW1_Time	Time to complete Homework 1(minutes)
Admission information	
SATmathˆ	SAT math score
SATverbˆ	SAT verbal score
Hsgpa	High School GPA
statAP*	AP Statistics taken (yes or no)
calcAP*	AP Calculus taken (yes or no)
MathLoc$^{\#}$	Location of highest math class taken (3 levels)
MathLevel$^{\#}$	Highest math class completed
Calc	Level of Calculus taken
Stat	Number of statistics classes taken (0, 1 or 2)
Online	Number of online unites attempted
Univerisity information	
StdLvl	Level of student at university(4 levels)
Enroll	Enrollment Status (3 levels)
CollegeDes	College description proxy to major(8 levels)
majorStat	Admitted to major (yes or no)
AdmBas	Admission basis (4 levels)
EOP	Part of Educational Opportunity Program (yes or no)
FirstSemester	First Semester (yes or no)
Dorm	On-Campus Housing in Dorms (yes or no)
Honors	Honors Program (yes or no)
LearningComm	Learning community - specialized dorms (yes or no)
Compact	Compact for Success - scholarship program (yes or no)
Fulltime	Full-time status (yes or no)
TermAtt	Units Attempted Number of units attempted this term

Note missing data: SAT$\hat{(}n = 149)$, Hsgpa($n = 89$), AP*($n = 121$), Math$^{\#}$($n = 67$), Calc($n = 43$) and Stat ($n = 43$)

CHAPTER 3
RESIDUAL CERFIT

3.1 Improve Numerical stability Using Residuals

One of the criticisms in interaction tree algorithm is the splitting rule, which is determined by selecting an individual covariate to estimate the average treatment difference without control the other covariates [1]. The concern of this criterion is that the treatment difference might be affected by covariates other than treatment. To address this issue, following the idea presented by [24], we propose to replace the original responses with residuals estimated from linear regression models to improve the numerical stability of the splitting rule. For an estimation of treatment effects, we can assume the following general model for responses Y:

$$Y = \alpha_0 + g(X) + Td(X) + \varepsilon \tag{3.1}$$

where α_0 is the overall mean, $g(X)$ is a function of baseline covariates affect responses, and $Td(X)$ is the interaction between the treatment and covariates. In the subgroup identification, we are interested in the $Td(X)$ only, and the baseline covariates effect $g(X)$ can be removed.

Let us recap the splitting rule in the interaction tree. The best split s^* is chosen by maximizing $t_s^2(s^*)$

$$argmax\{t^2(s)\} = \frac{ATE_L - ATE_R}{\hat{\sigma}\sqrt{1/n_1 + 1/n_2 + 1/n_3 + 1/n_4}} \tag{3.2}$$

where $\hat{\sigma}^2$ is the pooled estimator of the constant variance, and ATE_L and ATE_R are the average treatment effects for the two child nodes. The ATE under the general model as shown in Equation 3.1 can be expressed by

$$ATE = E(Y|X, T = 1) - E(Y|X, T = 0) = Td(X) \tag{3.3}$$

Regardless the form of $g(X)$, the only component that has an impact on the treatment effect is the treatment covariate interaction term $Td(X)$. Furthermore, similar as the work in [24], we can show the numerical stability benefits of removing $g(X)$ as follow:

$$\mathbb{E}_N(Y) = \frac{1}{N}\sum Td(X_i) + \mathcal{N}(\frac{1}{N}\sum h(x_i), \frac{1}{n^2}\sum g^2(X_i)) + o_p(1) \tag{3.4}$$

where $h(X) = \alpha_0 + g(X)$. Both $g(X)$ and $d(X)$ are centered to 0. $\mathbb{E}$ is to calculate sample average. When $h(X) \gg d(X)$, the second term in the above equation dominates the results. If we have a good estimator of $h(X)$ as $\widehat{h}(X)$, it is easy to show that we can stabilize the solution by eliminating the impact of $h(X)$.

$$\mathbb{E}_N(Y - \widehat{h}(X)) = \frac{1}{n}\sum Td(X_i) + o_p(1) \tag{3.5}$$

Fu et al. [24] suggest that without an optimal $h(X)$, a simple linear regression can improve the numerical stability of the algorithm. In our study we propose to follow their idea to fit a linear model with Y and X, and then, instead of using Y to evaluate the splitting rule, we use linear residuals ($\widetilde{Y}$) to determine the optimal split:

$$\widetilde{Y}_i = \beta_0 + \beta_1 I(T_i = 1) + \beta_2 I(X_{ij} \leq c) + \beta_3 I(T_i = 1)I(X_{ij} \leq c) + \beta_4 e_i + \varepsilon_i, \tag{3.6}$$

3.2 Simulation Studies: Residual numerical stability

3.2.1 Simulation models

We simulate 8 covariates from the standard normal distribution $X_i \sim N(0,1), i = 1, 2, \cdots, 8$. Treatment selection is based on a linear additive model using four covariates X_1, X_2, X_5 and X_6,

$$\text{logit}(Pr(T = 1|X)) = -1.5 + 0.8X_1 - 0.25X_2 + 0.6X_5 - 0.4X_6; \tag{3.7}$$

The response Y is simulated with three models. For the first two models, we simulate a common linear model for individualized treatment effect $Td_1(X)$, and two separate models for main effects $h_I(X)$ and $h_{II}(X)$. One is a simple linear model, and another is quadratic regression model.

$$h_I(X) = 2 + 0.5X_3 + 0.5X_4 + 0.5X_5 + 0.5X_6; \tag{3.8}$$

$$h_{II}(X) = 2 + 0.5X_3 + 0.5X_4 + 0.5X_5^2 + 0.5X_6; \tag{3.9}$$

The individual treatment effects $\delta(X)$ are determined by a linear additive model

$$Td_1(X) = I(T = 1)(0.5 + X_1 + 1.5X_2 + 2X_3 + 2.5X_4); \tag{3.10}$$

To further test the performance of replacing original responses with linear residuals, we then simulate the third linear model, a degree 3 polynomial model for the main effect($h_{III}(X)$). The interaction term $Td_2(X)$ is a tree type model with an interaction term:

$$h_{III}(X) = -5 - 2X_1 - 2X_2^2 + 2X_3^3; \tag{3.11}$$

$$Td_2(X) = I(T = 1)(-2 + 2I(X_1 \leq 0.5) + 2I(X_2 \leq 0.5)I(X_3 \leq 0.5)); \tag{3.12}$$

where $I(\cdot)$ is the indicator function.

$$Y = h(X) + Td(X) + \varepsilon \tag{3.13}$$

3.2.2 Simulation results

The numerical stability benefits using linear regression residuals are measured by the mean squared error (MSE) of individualized treatment effect. A smaller MSE indicates a higher prediction accuracy. The simulation results presented in Figure 3.1 are obtained from 100 simulation runs with training data $n = 1000$ and testing data $tn = 1000$. For all three simulated main effect models, using residuals generated from a simple linear regression model to replace original responses in tree growing process greatly increased the prediction accuracy. The benefits are significant even under the linear main effect models with second order or third order polynomial terms.

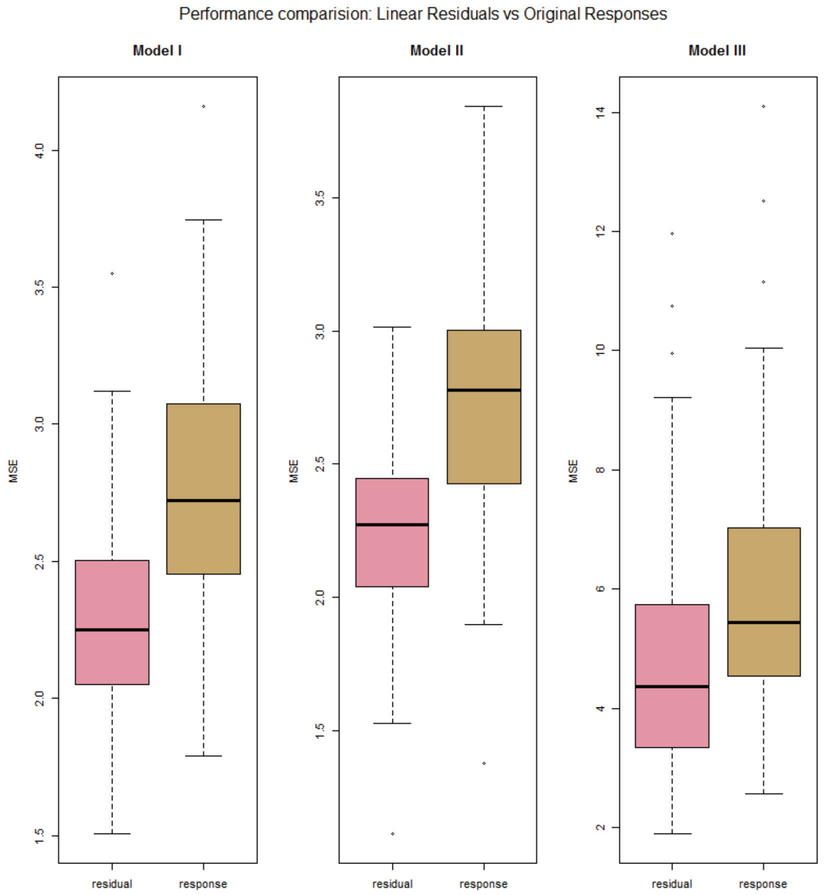

Figure 3.1. Numerical stability benefits by replacing original responses with residuals from linear regression model. A smaller MSE indicates a higher prediction accuracy.

CHAPTER 4
CERFIT FOR NON-BINARY TREATMENTS

4.1 INTRODUCTION

Interventions with non-binary treatments are common in various fields. Non-binary treatments may include more than two different treatments, multiple levels of one treatment, or even treatment with continuous values. In medical research, comparing efficacy of several drugs or determine effect drug doses are common. In education interventions such as supplemental instruction, treatment level is usually measured by the number of sessions that students attended. Under non-binary treatment setting, simple dichotomized treatment variable with treated and untreated will introduce subjective bias and information loss [21]. Learning individualized treatment effects is a more challenging task for non-binary treatments. While with binary treatments, it only requires learning one treatment effect between the treated and control group to identify the optimal treatment assignment; with non-binary treatments, it requires learning multiple treatment effects in order to achieve the same goal.

Giving increasing interests in individualized treatment regimes (ITR), diverse statistical methods have been developed in recent years. Because of their flexibility in modeling with few statistical assumptions, and their ability to handle a variety of data structures [8, 70], multiple machine learning tree based methods were proposed in ITR research. These methods can be broadly categorized into two approaches. The first one takes a regression-based subgroup identification approach. Through examining interactions between treatment and covariates, subgroups with similar or adverse treatment effects are identified. The optimal ITR can be determined by the treatment with maximized treatment benefits. Qualitative interaction tree (QUINT) [19], causal random forest [77], and random forest of interaction trees [43, 69, 68] are representative methods under this approach. Interaction based subgroups are often highly interpretable. However, methods under this approach are primarily designed for binary treatments. The other approach in ITR learning is growing a tree by maximizing a specified value function. While the aforementioned methods directly search for subgroups through interactions, methods under this approach construct a tree via maximizing a value function associated with treatment effects. Non-binary treatments and observational data can be handled under the value function framework. For instance, Zhao, Zeng, Rush and Kosorok [83] introduced the framework of outcome weighted learning (OWL) to directly find the optimal treatment rule for binary treatment. Tao, Wang and Almirall [71] proposed an adaptive contrast weighted learning (ACWL) algorithm by maximizing or minimizing an objective function. Recently, Chen, Tian, Cai and Yu [14] proposed a general framework by weighting and A-learning for subgroup identification to recover the optimal ITR through minimizing convex loss functions. Minimum impurity decision assignments (MIDAs) method [44] also falls in this category. Estimation of ITR is achieved through minimizing purity measures in a recursive algorithm in MIDAs. Both ACWL and Chen et al.'s [14] works can be applied for multiple treatments. MIDAs is the only method that can also be applied for continuous treatment setting. However, value function based methods could be model-dependent and less interpretable. The direct estimation of individualized treatment effects may not be available. Furthermore, methods under this framework also lack the ability of ranking variable importance with respect to treatment effects.

In this chapter, we propose algorithms to transform non-binary treatments optimization problems into a binary like problem using Random Forest of Interaction Trees [69]. The non-binary treatments considered in the algorithm include both multiple treatments (nominal or ordered) and continuous treatment. For multiple treatments, through recursive partitioning data into two subgroups with greatest treatment effects heterogeneity with respect to two randomly selected treatment groups, the algorithm transforms the multiple learning ITR into a binary task. Similarly, continuous treatment can be handled through recursively partitioning the data into subgroups with greatest homogeneity in terms of the association between the response and the treatment within a child node. The method is flexible, and the results are easy to interpret. Individualized treatment effects can be directly estimated along with variable importance ranking. In addition, by integrating general propensity scores into the tree growing process, the proposed method could be applied to both randomized and observational studies.

The remainder of this chapter is structured as follows. In Section 4.2, we introduce the Causal Effect RFIT (CERFIT) algorithm for non-binary treatments. In Section 4.3, we first review the general propensity score estimation methods for the multiple treatments and continuous treatment settings, and then present how to integrate the general propensity score into the CERFIT algorithm. Section 4.4 and 4.5 contain simulation studies to assess the CERFIT's performance by prediction accuracy and variable importance ranking for multiple and continuous treatment settings, respectively. In Section 4.6, we illustrate CERFIT's application under non-binary treatments through the analysis of multiple education interventions. Section 4.7 concludes the chapter with a brief discussion.

4.2 CERFIT for Non-Binary Treatments Causal Inference

The RFIT algorithm was developed for estimating subgroup average treatment effects (ATEs) using data from two armed randomized trials [69]. In our previous study, we proposed an algorithm to extend the RFIT for estimating individualized treatment effect for observational data. Due to the recursive feature of the tree algorithm, RFIT can be easily transformed and adapt to a multiple treatment situation. For each individual i in $(1, \cdots, n)$, we observe (Y_i, X_i, T_i). Y_i is the response, X_i is a set of baseline covariates, $T_i = t$ is the treatment assignment ($t \in \mathcal{T}$), where $\mathcal{T}$ is a collection of m treatment options. Under the potential outcomes framework, each individual has $(m - 1)$ potential outcomes $(Y_i^1, \cdots, Y_i^{m-1})$. The individualized treatment effects (ITE) between treatment u and v is given by the treatment difference $(Y_i^u - Y_i^v)$. The multiple treatment effects can be viewed as

$\frac{m(m-1)}{2}$ pairwise contrasts. The interaction tree is grown by recursively splitting the data across values of random selected covariates. The value that maximizes the difference in treatment effects between the two binary treatments is selected as the best split. Each terminal node retains individuals with maximized treatment difference from the other terminal nodes. Under binary treatment, each split is based on treatment differences between the treated and control group. Under a multiple treatments setting, pairwise contrasts among treatments can be implemented into sequential splitting by randomly selecting a pair of treatments at each split without modifying the splitting rule. As a tree grows bigger, each pairwise contrast has the same chance of being selected to evaluate the splitting rule. Within a tree, different pairs of treatments encounter different splitting depths, but these depths are randomized among all trees in a forest. Similar to the binary treatment setting, the terminal nodes contain individuals with maximized treatment effect heterogeneity compared with the individuals within other terminal nodes but evaluated through multiple treatments. The ensemble of the tree yields a smoothed estimation of treatment outcome for each treatment. The splitting rule is essentially the same as under binary treatment setting, but with a randomly selected pair of treatments $T = t_u$ and $T = t_v$. For each partitioning, the evaluation of the splitting rule is through a subset of observations received treatment assignment $T_k = t_u$ or $T_k = t_v$, where k in $(1, \cdots, n_{uv})$.

$$Y_k = \beta_0 + \beta_1 I(T_k = t_u) + \beta_2 I(X_{kj} \leq c) + \beta_3 I(T_k = t_u) I(X_{kj} \leq c) + \varepsilon_k, \tag{4.1}$$

where $I(\cdot)$ is the indicator function. T_k is the treatment assignment for the k^{th} subject, $I(X_{kj} \leq c)$ is the indicator for a binary cut based on covariate X_j, and ε_k is iid $N(0, \sigma)$.

An alternative algorithm for a multiple treatments setting is to only focus on the $m - 1$ contrasts, instead of viewing the problem as $\frac{m(m-1)}{2}$ pairwise contrasts. Since the estimated ITE should be transitive, we only need the information from $m - 1$ pairwise ITEs to identify the optimal ITR. The advantage of this alternative algorithm is apparent. With fewer required pairs of contrasts, the tree is grown more efficiently in learning the ITEs between each of the treatments to the reference treatment. However, the accuracy might be impacted since the algorithm only focuses on the comparison of treatment effects to one predefined reference treatment. In terms of the variable importance ranking, under the alternative algorithm, variable importance is ranked based on the impact of each of covariates on the treatment differences with respect to the reference treatment only. This alternative algorithm is denoted as CERFIT2 and the aforementioned algorithm is called CERFIT1 here after.

Under a continuous treatment setting, by slightly modifying the treatment variable T_i in equation 2.2, we have

$$Y_i = \beta_0 + \beta_1 T_i + \beta_2 Z_i + \beta_3 T_i Z_i + \varepsilon_i, \tag{4.2}$$

where $T_i = t$ ($t \in \mathcal{T}$) is the treatment received by the i^{th} subject, Z_i is the indicator for a binary cut based on the covariate X_j, $Z_i = I(X_{ij} \leq c)$, and ε_i iid $N(0, \sigma)$. A significant Wald test for $H_0 : \beta_3 = 0$ from equation 4.2 indicates the association between the treatment variable (T) and the response (Y) is significantly different between subgroup $X_j \leq c$ and $X_j > c$. With sequential splitting and growing, a particular terminal node retains observations who have similar associations between T and Y.

4.3 General Propensity Scores (GPS)

To expand the application of RFIT to observational data, we propose to integrate the propensity score into the tree growing process to address confounding issues in observational studies. Under a binary treatment setting, the propensity score e is defined as the probability of the treatment conditional on a set of covariates, $e = Pr(T = 1 \mid X)$. In their seminal work, Rosenbaum and Rubin (1983) show that an unbiased estimate of the average treatment effect can be obtained by conditioning on the propensity score alone, instead of a set of covariates: $\{Y(1), Y(0)\} \perp\!\!\!\perp T \mid e(X)$. This approach has been widely applied in causal inference. Several methods based on propensity score have been proposed, such as matching, stratification, inverse probability of treatment weighting (IPTW), and covariate adjustment [3, 4]. Applications of propensity score methods are usually limited to a binary treatment setting. Recent works on general propensity scores (GPS) [25, 30, 36, 21] extend applications of propensity score methods into a general treatment setting for causal inference.

4.3.1 GPS for multiple treatments (GPSm)

Let $r(t, X) = Pr(T = t|X)$ as the conditional probability of receiving a particular treatment, then GPSm can be defined as $R(X) = (r(t_1, X), ..., r(t_m, X)$ [25]. In practice, GPSm are usually estimated using multinomial logistic, multinomial probit models for nominal treatments or proportional odds models for ordered treatments [25, 36, 48]. Causal inference for observational data with multiple treatments can be implemented using pairwise matching [52, 45], vector matching [48] or IPTW methods [20, 54]. For ITR study under a multiple treatments setting, IPTW is the most frequently used method. For instance, the ACWL method proposed by Tao et al.[71] uses multinomial logistic regression to estimate propensity scores and form a double robust augmented inverse probability weighted estimator. similarly, Angle-based direct learning [58] and the personalized benefit scoring system from the general framework of subgroup identification [14] both integrate the multinomial logistic regression based propensity scores in the value function for learning ITR in observational studies. Similar to a binary treatment setting, the estimation of propensity score for multiple treatments are also subject to model misspecification when using parametric multinomial logistic regression. The problem becomes more prominent as the number of treatments increases. To mitigate estimation errors and extreme weights in IPTW methods McCaffrey et al. [53] proposed using a general boosted model(GBM) to estimate GPSm with a stopping rule that maximizes the resulting covariate balance. The method was later extended to multiple treatments [53] and continuous treatment settings [84] by modifying the stopping rules to yield a match between the target group and the entire sample. The covariate balancing propensity score (CBPS) method proposed by Imai and Ratkovic [35] can also be extended to a setting of multiple treatments. The GPSm in CBPS method [22] is estimated such that the general covariate balancing conditions are satisfied: $E(\frac{I\{T_i=t_j\}X_i}{r(t_j,X_i)}) = E(X_i)$, where $I(\cdot)$ is the indicator function, $T_i = t_j$ is the treatment assignment for the i^{th} observation.

4.3.2 GPS for continuous treatment (GPSc)

Let $r(t, X), t \in \mathcal{T}$ be the conditional density of the treatment given observed covariates, $r(t, X) = f_{T|X}(t \mid X)$, then GPSc can be written as $R = r(T, X)$, where T is a random variable denotes the treatment received and t is a specific level of T. The GPSc has a balancing property similar to the standard propensity score. The probability that a subject received a treatment assignment $T = t$ is independent to the value of X within strata with the same value of $r(t, x)$: $X \perp\!\!\!\perp I(T = t) \mid r(t, X)$, where $I(\cdot)$ is the indicator function. Together with the assumption of weak unconfoundedness:$Y(t) \perp\!\!\!\perp T \mid X$ for all $t \in \mathcal{T}$, where $Y(t)$ is a random variable that maps a particular potential treatment t to a potential outcome. Hirano and Imbens [30] show that GPSc can be used to eliminate any biases associated with differences in covariates. As a popular practice, to estimate the GPSc, one could assume that the treatment T or its transformation $m(T)$ is normally distributed given covariate X: $m(T) \mid X \sim N(\gamma' X, \sigma^2)$, where the parameters γ and σ^2 can be estimated by maximum likelihood [30]. Thus GPSc can be estimated by the normal density function:

$$\widehat{R}_i = \frac{1}{\sqrt{2\pi\widehat{\sigma}^2}} exp(-\frac{1}{2\widehat{\sigma}^2}(T_i - \widehat{\gamma}' X)^2) \tag{4.3}$$

The underlying assumption for this method is that the conditional distribution of the treatment or its transformation given the observed covariates needs to be approximately normal. With high dimensional covariates, this two-stepped parametric density estimation may suffer from the course of dimensionality and model misspecification. To address this issue, Zhu, Coffman and Ghosh [84] extended the GBM based propensity score estimation approach to a novel boosting algorithm for GPSc estimation. They take a more general approach to assume $T_i = m(X_i) + \varepsilon, \varepsilon \sim N(0, \sigma^2)$, where $m(X)$, the mean function of T given X is estimated using a nonparametric boosting algorithm with a stopping criterion such that the treatment assignment and covariates are independent in the weighted sample. The degree of independence between the treatment and each covariate can be measured by commonly used correlation matrices, such as Pearson, Spearman, Kendall and distance. The CBPS method can also be extended to a continuous treatment setting. Fong and Imai [21] propose a parametric and a noneparametric versions of CBPS for a continuous treatment. The CBPS does not involve direct estimation of GPSc but uses an empirical likelihood approach to choose weights that achieve a sample data with zero correlations between the treatment assignment and the covariates.

4.3.3 Incorporating GPS in the algorithm for observational study data

In random forest (RF) or random forest of interaction trees (RFIT), each tree is grown based on a bootstrap sample. In our proposed causal effect random forest of interaction trees (CERFIT), each tree is built based on a weighted bootstrap sample selected using weights $w_i = \frac{1}{r(t,X_i)}$. By doing so, we can address the confounding issues in observational data by growing a tree using subsamples with minimized association between treatments and baseline covariates.

Secondly, we propose to use the GPS to adjust the RFIT splitting rule to further control the confounding issues during the tree growing process. To this end, we use the GPS as a control covariate in the interaction model.

IPTW using propensity scores leads to an unbiased estimation of treatment effects. Several studies show conventional variance estimator under the IPTW method can be biased, because using weights induces within subject correlation. Consequently, it is suggested that a robust standard error should be considered with IPTW based regression models [28, 41]. Therefore, we propose to use a sandwich type robust standard error (RSE) to adjust the Walt test statistics, which helps to account for the lack of independence in replications of subjects induced by IPTW weighting:

$$t_s(s) = \frac{\hat{\beta}_3}{RSE} \tag{4.4}$$

where RSE is the robust standard error estimated using

$$RSE = [(X'X)^{-1}(X'\text{diag}(\varepsilon_i^2)X)(X'X)^{-1}]^{\frac{1}{2}} \tag{4.5}$$

Additionally, based on our simulation studies in Chapter 3, we use residuals estimated from the linear regression model to replace the original responses in the proposed algorithm to improve the prediction accuracy. The best split s^* is chosen by maximizing the squared Wald test statistics for $H_0 : \beta_3 = 0$ in Equation 4.6 and 4.7.

Multiple treatments

$$\widetilde{Y_k} = \beta_0 + \beta_1 I(T_k = t_u) + \beta_2 I(X_{kj} \leq c) + \beta_3 I(T_k = t_u) I(X_{kj} \leq c) + \beta_4 r(T_k = t_u, X_k) + \varepsilon_k, \tag{4.6}$$

where $\widetilde{Y_i}$ is the linear residuals; $I(\cdot)$ is the indicator function. T_k is the treatment assignment for the k^{th} subject, k in $(1, \cdots, n_{uv})$;$I(X_{kj} \leq c)$ is the indicator for a binary cut based on covariate X_j, and ε_k is $iid\ N(0, \sigma)$.

Continuous treatments

$$\widetilde{Y_i} = \beta_0 + \beta_1 T_i + \beta_2 Z_i + \beta_3 T_i Z_i + \beta_4 r(T_i, X_i) + \varepsilon_i, \tag{4.7}$$

Where $\widetilde{Y_i}$ is the linear residuals; $T_i = t$ ($t \in \mathcal{T}$) is the treatment received by the i^{th} subject, Z_i is the indicator for a binary cut based on the covariate X_j, $Z_i = I(X_{ij} \leq c)$, and ε_i is $iid\ N(0, \sigma)$.

4.3.4 CERFIT for Non-Binary Treatments Algorithm

The detailed tree algorithms for settings of non-binary treatments are summarized in the following tables. In particular, the default $mtry$ value for the regression problem is set as $mtry = p/3$ in the *randomForest* R package [7], where p is the number of predictors. We recommend to use $max\{3, p/2\}$ and $max\{3, p/6\}$ as the default $mtry$ value in CERFIT for multiple treatments and continuous treatment, respectively. For multiple treatments, we propose a larger value of $mtry$ since randomly selecting a pair of treatments to evaluate at each split, greatly increases the estimation variance. A bigger value of $mtry$ helps mitigate this impact. With the same underlying reason, we recommend using a smaller $mtry$ for continuous treatments, since using weighted bootstrap samples may increase the similarity of the trees, which decreases the estimation variance among trees. A smaller $mtry$ can help de-correlate the trees. In particular, $p/6$ is half of the default $mtry$ value in the *randomForest* [7], R package and the value of 3 in our recommendation is designed to stay away from an $mtry$ value that is too small, when p is not very large, in order to preserve the quality of splits and ultimately the prediction accuracy of the random forest. The default terminal node size is set at 30.

To estimate the treatment effects for multiple treatments, within the terminal node, the interaction effect for each treatment $T = t_j$ is estimated using a weighted average $\widehat{\widetilde{Y}}(t_j) = \frac{\Sigma w_i(t_j)\widetilde{Y}_i(t_j)}{\Sigma w_i(t_j)}$, where i is the i^{th} observation in the terminal node with treatment assignment t_j. For each subject, the final prediction for treatment t_j is the average prediction across all $ntree$ trees.

To estimate the optimal treatment regime for continuous treatment, we model the conditional expectation of the interaction effect $\widetilde{Y}_i$ given T_i using an additive regression model. In their original work, Hirano and Imai [30] recommend using a flexible Gaussian

Table 4.1. The CERFIT Algorithm for Multiple Treatments

The CERFIT algorithm for multiple treatments

1. Estimate propensity score $R(T, X)$ and inverse probability weight w.

2. Calculate residual $\widetilde{Y}$ using linear regression model and use $\{\widetilde{Y}, T, R(T, X), X\}$ as input data.

3. Draw bootstrap samples from the data using w as sampling weights.

4. Grow an interaction tree based on each weighted bootstrap sample.

−4.1a At each node, randomly select a pair of treatment $T = t_u$ and $T = t_v$; define $t_u = 1$ and $t_v = 0$, then subset the data that only contains these two treatments.

Or using step $4.1b$ as an alternative algorithm for $4.1a$

−4.1b Predetermine a reference treatment $T = t_{ref}$ and set $t_{ref} = 0$. At each node, randomly select another treatment $T = t_u$ and define $t_u = 1$, then subset the data that only contains these two treatments.

−4.2 Randomly select $mtry$ of total p covariates from which to determine a split rule. The default value of $mtry$ is set at $p/2$.

−4.3 Among the $mtry$ covariates selected, the optimal split is identified by maximizing the adjusted squared Wald test statistic for testing $H_0 : \beta_3 = 0$, in Equation (4.6) using data generated in step 4.1a or 4.1b.

−4.4 Repeat steps 4.1 to 4.3 until reaching a pre-specified stopping rule (e.g., maximum tree depth, minimum terminal node size).

5. Repeat steps 3 and 4 for $ntree$ trees as desired, with a default value set at 500.

quadratic regression model:

$$E[Y_i|T_i, r(T_i, X_i)] = \eta_0 + \eta_1 T_i + \eta_2 T_i^2 + \eta_3 r(T_i, X_i) + \eta_4 r(T_i, X_i)^2 + \eta_5 T_i r(T_i, X_i) \quad (4.8)$$

Where $r(T_i, X_i)$ is the GPSc.

Table 4.2. The CERFIT Algorithm for Continuous Treatments

The CERFIT algorithm for continuous treatments

1. Estimate GPSc and inverse probability weight w_i.

2. Calculate residual $\widetilde{Y}$ using a linear regression model and use $\widetilde{Y}, T, R(T, X), X$ as input data.

3. Draw bootstrap samples from the data using w_i as sampling weights.

4. Grow an interaction tree based on each weighted bootstrap sample.

$-$4.1 At each node, randomly select a subset $mtry$ of the total p covariates from which to determine a split rule. The default value of $mtry$ is set at $max\{3, p/6\}$.

$-$4.2 Among the $mtry$ covariates selected, the optimal split is identified by maximizing the squared Wald test statistic for testing $H_0 : \beta_3 = 0$, in Equation (4.7).

$-$4.3 Repeat steps 4.1 and 4.2 until reaching a pre-specified stopping rule (e.g., maximum tree depth, minimum terminal node size).

5. Repeat steps 3 and 4 for $ntree$ trees as desired, with a default value set at 500.

Within the terminal node of the CERFIT tree, the data structure is relatively simple with a smaller sample size. Therefore, we use the following parsimonious yet flexible model:

$$E[\widetilde{Y}|T, r(T_i, X_i)] = \eta_0 + \eta_1 T_i + \eta_2 T_i^2 + \eta_3 T_i^3 \tag{4.9}$$

To reduce overfitting, a LASSO penalty [74] is used, where the penalization is determined by minimizing 10-fold cross-validated prediction error. The final treatment prediction for each subject is the average prediction across all trees.

For a discrete continuous treatment. The optimal treatment could be determined by the treatment associated with a maximized treatment outcome. For studies with interests in continuous treatment on the interval $[a, b]$, within each terminal node, a standard optimization routine [10] available in the base R function $optimize$ can be applied to render an optimal treatment level within each terminal node. Then the final optimized ITR can be determined by an ensemble results across a forest.

Variable importance ranking in CERFIT is implemented using minimal depth (MD) procedure [39]. MD assesses the predictive power of a variable by the depth of the first split of a variable relative to the root node of a tree. A smaller MD indicates a covariate has higher predictive power. The specified steps in finding MD (MD_j) for a covariate X_j $(j = 0, 1, 2 \cdots, p)$ is described in the variable importance algorithm below.

Table 4.3. The Variable Importance: Minimal Depth Algorithm

Variable importance algorithm: Minimal Depth

1. Let Γ_b denotes tree b $(b = 0, 1, 2 \cdots, ntree)$ and D_i^j $(i = 0, 1, 2 \cdots, r)$ denotes the distance from the root node to the nodes split on a covariate X_j for all r splits on covariate X_j within tree Γ_b

2. Sort D_i^j for each covariate X_j, and find minimal depth D_b^j for X_j within tree Γ_b using $Min(D_i^j)$.

3. For covariate X_j is not used in the tree growth, define $D_j = MaxD_b + 1$, where $MaxD_b$ denotes the maximum tree depth for tree b.

4. Repeat steps 1 to 3 for every tree in the forest and calculate the mean of the D_b^j over the forest $MD_j = \frac{\sum_{b=1}^{ntree} D_b^j}{ntree}$ for each of the X_j.

4.4 Simulation Studies for Multiple Treatments

We simulate data with 10 covariates X_j $(j = 1, ..., 10)$ from the uniform distribution from -1 to 1. Simulation of the treatment selection is modified from the framework used by Huling and Yu [34]. Three treatments were generated in their study. In our simulation, four treatments are generated.

Treatment selection model.

$$logit(Pr(T=1|X)) = 0.1 + 0.5X2 - 0.25X3 \tag{4.10}$$

$$logit(Pr(T=2|X)) = 0.1 - 0.5X1 + 0.25X4 \tag{4.11}$$

$$logit(Pr(T=3|X)) = 0.1 + 0.5X3 - 0.25X1 \tag{4.12}$$

$$Pr(T=4|X)) = 1 - Pr(T=1|X)) - Pr(T=2|X)) - Pr(T=3|X)) \tag{4.13}$$

The continuous outcomes are simulated based on the models proposed by Qi, Liu, Fu and Liu [58] and Zhang, Laber, Davidian and Tsiatis[82]. Let $Y = h_m(X) + c_m(X,T) + \varepsilon$, where $h_m(X)$ is the main effect that has no contribution to define the true ITR. The second term $c_m(X,T)$ defines the optimal ITR for each of observations. The random error term is set up as $\varepsilon \sim N(0,1)$.

$$h_m(X) = 2 + X1 + X3 + X5 + X7 \tag{4.14}$$

$$\begin{aligned} c_{m1}(X,T) =& (1 + X_1 + X_2 + X_3 + X_4)I(T=1) + (1 + X_1 - X_2 - X_3 + X_4)I(T=2) + \\ & (1 + X_1 - X_2 + X_3 - X_4)I(T=3) + (1 - X_1 - X_2 + X_3 + X_4)I(T=4) \end{aligned} \tag{4.15}$$

$$\begin{aligned} c_{m2}(X,T) =& (3I(X_2 < 0) - I(X_1 \geq= -0.3))I(T=1) + (4I(X_1 > 0) - 2)I(T=2) + \\ & (I(X_1 \leq 0) - 2)(2I(X_2 \leq -0.3) - 1)I(T=3) + (3I(X_1X_2 > 0) - 1)I(T=4) \end{aligned} \tag{4.16}$$

$$\begin{aligned} c_{m3}(X,T) =& (0.2 + X_1^2 + X_2^2 - X_3^2 - X_4^2)I(T=1) + (0.2 + X_1^1 + X_2^2 - X_3^2 - X_4^2)I(T=2) + \\ & (0.2 + X_1^2 + X_4^2 - X_2^2 - X_3^2)I(T=3) + (0.2 + X_2^2 + X_3^2 - X_1^2 - X_4^2)I(T=4) \end{aligned} \tag{4.17}$$

The complexity of the modelincreases from scenario 1 to scenario 3. In scenario 1, c_{m1} has a linear association with covariates X_1, X_2, X_3 and X_4. In Scenario 2, c_{m2} corresponds to a tree-type interaction. In scenario 3, c_{m3} has a degree 2 polynomial interaction effects. Thus, there are four confounders: X_1, X_2, X_3 and X_4 for scenarios 1 and 3. And there are two confounders: X_1 and X_2 for scenarios 2.

Training data with two sample sizes $n = 500$ and $n = 1000$ are used to estimate the ITR. The performance is first assessed by the average of the mean squared errors (AMSE) for 4 different treatments.

$$AMSE = \frac{\sum_{j=1}^{4} MSE_j}{4} \tag{4.18}$$

where MSE_j is the mean squared error for treatment $T = t_j$, $MSE_j = \frac{1}{n}\sum_{i=1}^{n}[Y_i(T = t_j) - \hat{Y}_i(T = t_j)]^2$. Then we evaluate the performance of the algorithm using the classification rate in terms of correctly identifying the optimal ITR. The assessment is based on the independent simulated test sample $n_t = 500$ and $n_t = 1000$ with 100 simulation runs. For each scenario, a total of 500 trees are grown and the value of $mtry$ is set at 5.

The prediction is first evaluated by comparing the performance of CERFIT1 and CERFIT2 with two other methods: Bayesian regularized trees (BART)[29] and decision list (DL) [82, 81]. BART is a sum-of-tree model based on Bayesian regularized trees. Each consecutive tree refits the residuals that are not explained by the other trees. Fitting and inference procedures are using the iterative Bayesian backfitting MCMC algorithm [15]. The implementation of BART is performed by the R-package *BayesTree* [16] with default settings $ntree = 200$ and 1000 MCMC iterations. The optimal ITR is determined by the treatment associated with maximum treatment effects. DL can be viewed as a special case of tree method. It applies decision lists to learn the optimal ITR with a sequence of ”if” and ”then” clauses. The implementation of DL is done through the R-package *Listdtr* [80] with default settings. To evaluate the proposed algorithm’s learning ability in identifying optimal ITR, beside BART and DL, we also include the other two methods: adaptive contrast weighted learning (ACWL)[72] and general statistical framework for subgroups identification methods [14]. ACWL is a semiparametric regression contrasts with the adaptation of treatment effects. Two different contrasts are used as seen in Tao et al.[72]'s study: ACWL1 (maximizes the objective function) and ACWL2 (minimizes the expected loss).The R codes provided in the paper's appendix are used in the simulation. The general statistical framework for subgroups identification method [14] uses weighting and A-learning for subgroup identification and constructs a comparative treatment scoring system to identify the optimal ITR. The method can be implemented using R-package *personalized* [34].

Simulation results are presented in the Figure 4.1 and Figure 4.2. We can see that the CERFIT1 and CERFIT2 has competitive performance among all methods. CERFIT1's performance has a slight edge over CERFIT2. For estimation accuracy measured by the AMSE, CERFIT1 has the smallest AMSE for sample size $n = 500$ and $n = 1000$ under tree-type and polynomial interaction effects, but not under scenario 1, where linear interaction effects are simulated. In general, for all methods, the prediction accuracy is improved as sample size increases from 500 to 1000. For the purpose of finding the optimal ITR among a class of available treatments, CERFIT1 outperforms all other methods under all scenarios. Even though the AMSE of CERFIT1 is higher than DL under scenario 1, the classification rate is similar between the two methods. All methods perform the best under the tree-type interaction (scenario 2) and have the worst performance under scenario 3. And even under scenario 3, CERFIT1 still correctly identifies the optimal ITR around 50% and 61% times at sample size 500 and 1000, respectively.

The simulation results for variable importance based on minimal depth (MD) is presented in Figure 4.3. For all three scenarios, the variable importance algorithm for CERFIT1 and CERFIT2 correctly identify variables that are associated with treatment effects. The MD based variable importance ranking are consistent with the underlying truth. From Figure 4.3 we can see that under scenarios 1 and 3, covariates X_1, X_2, X_3, X_4 are identified as important covariates associated with treatment effects reflected by their small MD values.

Under scenarios 2, only X_1 and X_2 are interacted with treatment effects, and therefore these two covariates should have relatively small MD values. Even though X_3 and X_4 are associated with the treatment selection, they are not picked up as important predictors since they are not covariates associated with treatment effects. Similarly, X_5 and X_7 are only associated with the main treatment effect. They have no impact on the interaction treatment effects and therefore are identified as being equally unimportant as X_6, X_8, X_9 and X_{10}. Additionally, it is important to point out that MD values for CERFIT1 and CERFIT2 assess the predictive power of a variable regarding different treatment contrasts. For CERFIT1, the variable importance is ranked based on all possible pairs of contrasts. However, the CERFIT2's variable importance is ranked based on the treatment differences between each treatment to the reference treatment only. In this simulation, the reference group is set as $T = 1$. This explains the different patterns of variable importance ranking in Figure 4.3. Under scenario 1, X_1, X_2, X_3 and X_4 are equality important for CERFIT1, but X_2 is ranked as the most important predictor for CERFIT2 comparing with other three important predictors. With some simple calculation, it is not difficult to see that treatment effects between each treatment to the reference treatment are: $ITE_2 = -2X_2 - 2X_3$; $ITE_3 = -2X_2 - 2X_4$ and $ITE_4 = -2X_1 - 2X_2$. And X_2 is the variable with the highest predictive power in determining the size of treatment effects with respect to reference treatment.

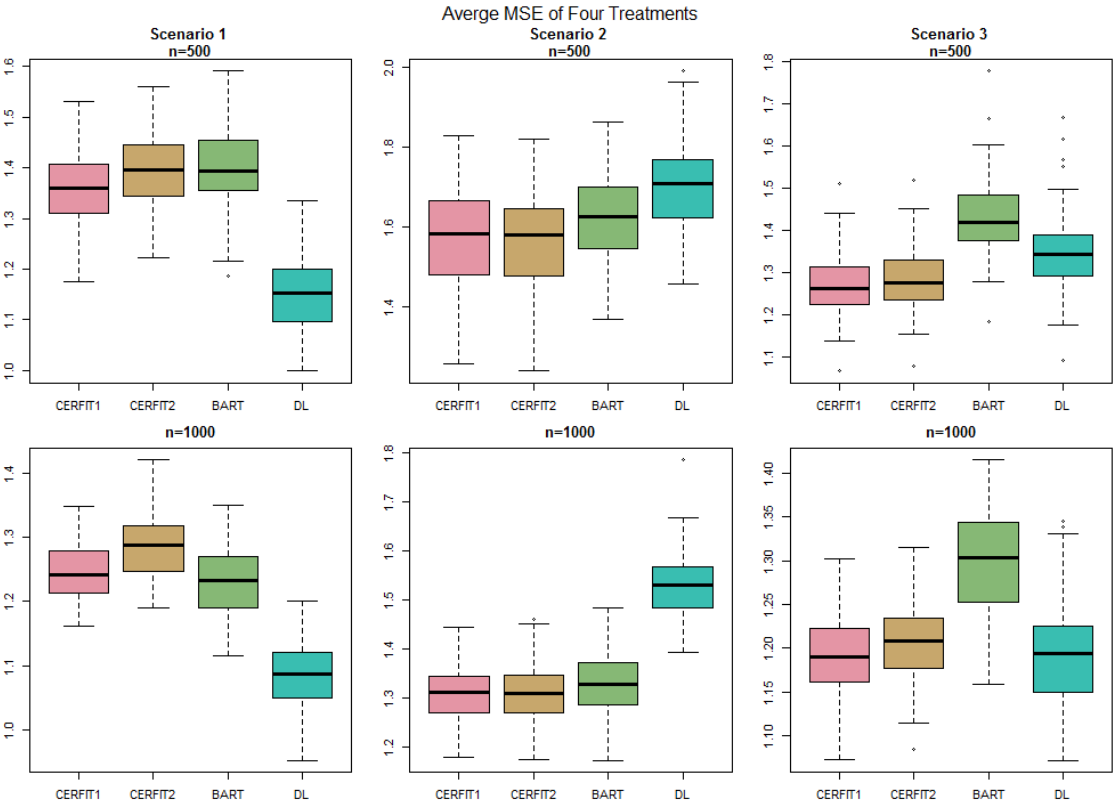

Figure 4.1. Simulation results: comparison of four methods in terms of average MSE for four treatments based on 100 simulation runs. Training and testing data with two sample size 500 and 1000. The smaller average MSE is preferable.

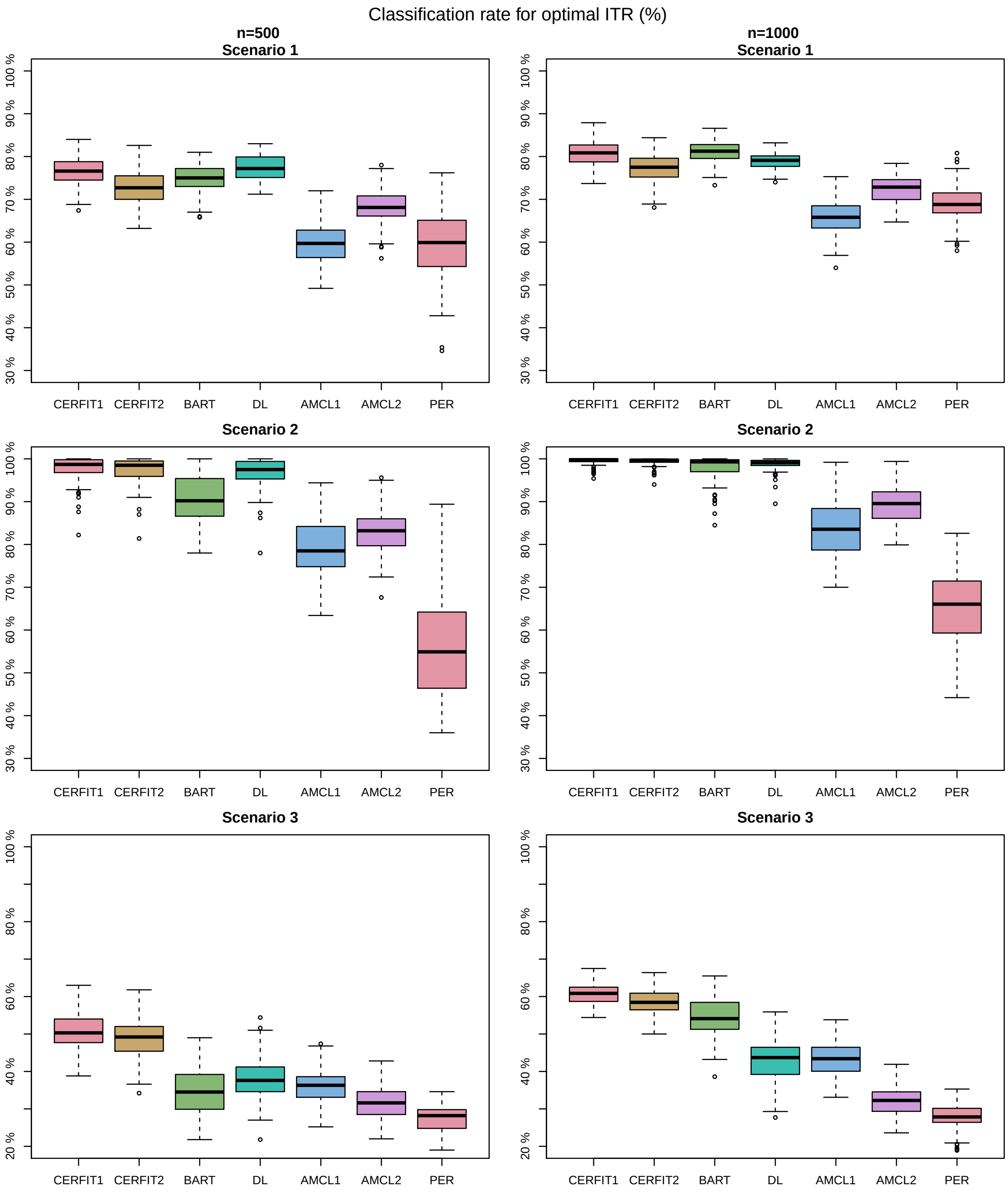

Figure 4.2. Simulation results: comparison of seven methods in terms of classification rate for correctly identifying optimal treatment regimes based on 100 simulations runs. Training and testing data with two sample size 500 and 1000. The higher classification rate is preferable.

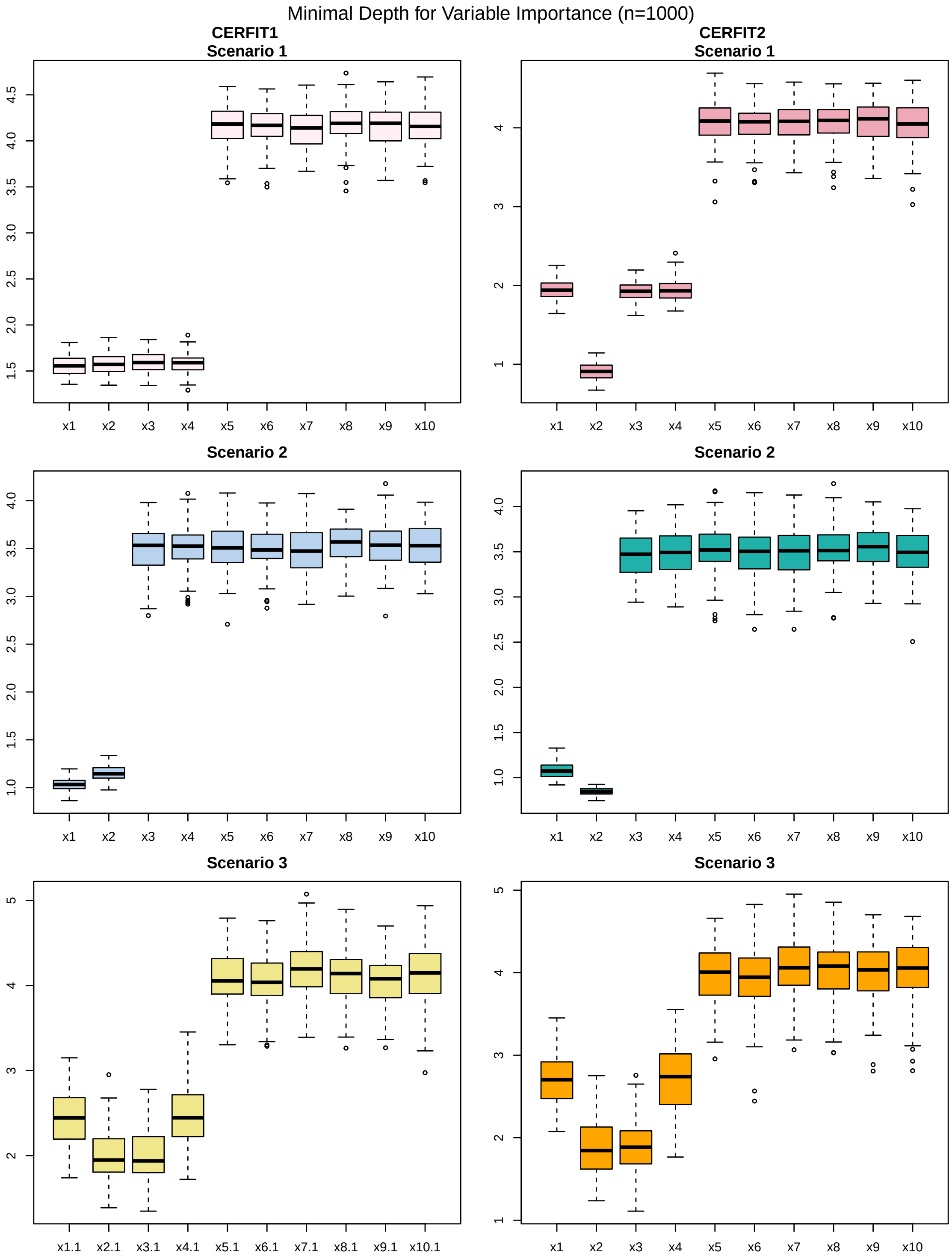

Figure 4.3. Simulation results: minimal depth for each covariate. The smaller minimal depth corresponding to a higher variable importance ranking. Results based 100 simulation runs with sample size n=1000.

4.5 Simulation Studies for Continuous Treatment

The generation of observed (Y, T, X) for continuous treatment are as follow. We generate 10 baseline covariates X_j $(j = 1, ..., 10)$ from uniform $(0, 1)$ and the treatment assignment using the framework slightly modified from Zhu, Coffman and Ghosh [84]. Treatment assignment is generated using a linear additive model associated with 4 covariates: X_1, X_2, X_3 and X_4. Then the treatment is normalized between $(0, 1)$ before simulating responses.

$$T = 0.5 + 0.3X_1 + 0.65X_2 - 0.35X_3 - 0.4X_4 + \varepsilon; \tag{4.19}$$

The continuous outcomes are simulated based on the same model ($Y = h_c(X) + c_c(X, T) + \varepsilon$) as seen for multiple treatments. The main effect $h_c(X)$ has no contribution to define the true ITR. The second term $c_c(X, T)$ defines the optimal ITR for each of observations.

Similar as Zhu et al. [84], the $h_c(X)$ is simulated as follows:

$$h_c(X) = 3.85 + 0.3X_1 + 0.36X_2 + 0.73X_3 - 0.2X_4 \tag{4.20}$$

The $c_c(X, T)$ is simulated similarly as the three models proposed by Laber et al. [44]. Let ϕ and Φ denote the density and cumulative distribution of a standard normal random variable. The three forms of $c_c(X, T)$ are presented below. In each scenario the positive proportionality constant is chosen so that $var\{c_c(X, T)\} = 1$

$$\begin{aligned} c_{c1}(X, T) \propto\, & I\{X_1 \geq 0.7\}\, \phi[3(\Phi^{-1}(T) + \Phi^{-1}(0.75))] \\ & + I\{X_1 < 0.7, X_2 > 0.5\}\, \phi[3\Phi^{-1}(T)] \\ & + I\{X_1 < 0.7, X_2 \leq 0.5\}\, \phi[\Phi^{-1}(T) + \Phi^{-1}(0.25)]; \end{aligned} \tag{4.21}$$

Where the optimal regime is treatment $t = 0.25$ when $X_1 \geq 0.7$, $t = 0.5$ when $X_1 < 0.7$ and $X_2 > 0.5$, and $t = 0.75$ otherwise.

$$\begin{aligned} c_{c2}(X,T) \propto (1 - \frac{\mid T - 0.20 \mid}{0.80})I\{X_1 > 0.5, X_3 > 0.5\} \\ + (1 - \frac{\mid T - 0.40 \mid}{0.80})I\{X_1 > 0.5, X_3 \leq 0.5\} \\ (1 - \frac{\mid T - 0.60 \mid}{0.80})I\{X_1 \leq 0.5, X_2 > 0.25\} \\ + (1 - \frac{\mid T - 0.80 \mid}{0.80})I\{X_1 \leq 0.5, X_2 \leq 0.25\}; \end{aligned} \tag{4.22}$$

The optimal regime is treatment $t = 0.2$ when $X_1 > 0.5$ and $X_3 > 0.5$, $t = 0.4$ when $X_1 > 0.5$ and $X_3 \leq 0.5$, $t = 0.60$, when $X_1 \leq 0.5$ and $X_2 > 0.25$, and $t = 0.80$ otherwise.

$$c_{c3}(X,T) \propto \frac{1}{1 + 10(2T - X_1 - X_2)^2} \tag{4.23}$$

For the last scenario, the optimal regime is determined by $t = (X_1 + X_2)/2$. And the tree-based decision rule is misspecified.

The performance of the proposed algorithm is assessed in a similar process as discussed under a multiple treatments setting. First, we evaluate the accuracy in identifying optimal ITR. In Figure 4.4 and Figure 4.5, we visualize the true and estimated optimal treatment regimes based on ITR model $c_{c1}(X,T)$ and $c_{c3}(X,T)$ defined as a function of the covariates X_1 and X_2 for each testing observations $(t_n = 10,000)$ averaged over 100 repetitions. As shown in the two figures, the predicted optimal treatment rules are roughly similar as the true underlying structure. DL's [82] performance greatly deteriorated as number of treatment levels increased, therefore the results are not presented here.

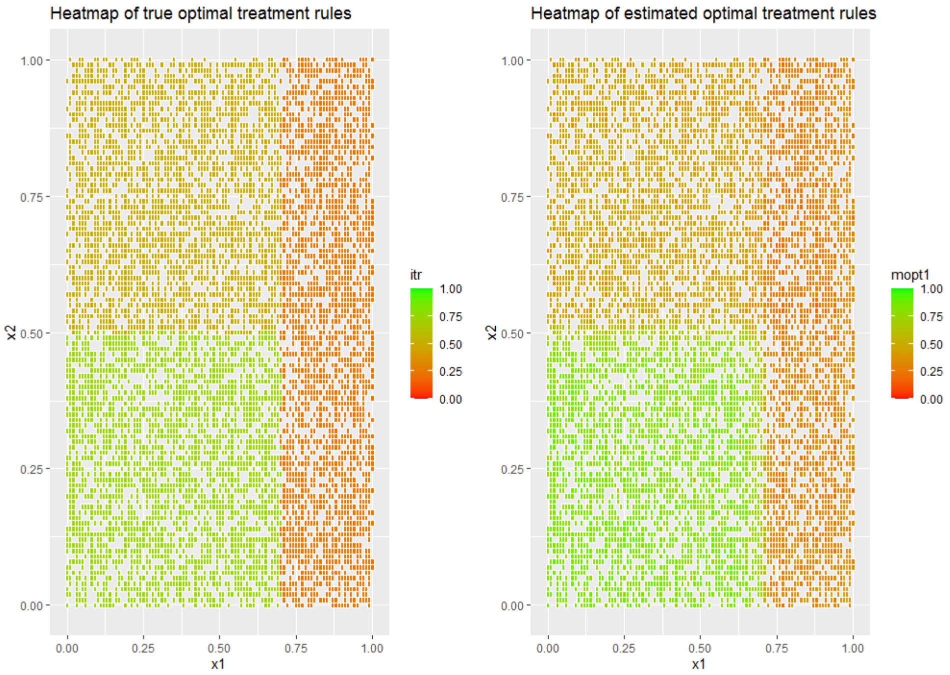

Figure 4.4. Simulation results: heatmaps of true and estimated optimal treatment regimes as a function of X_1 and X_2 as defined in the ITR model $c_{c1}(X, T)$.

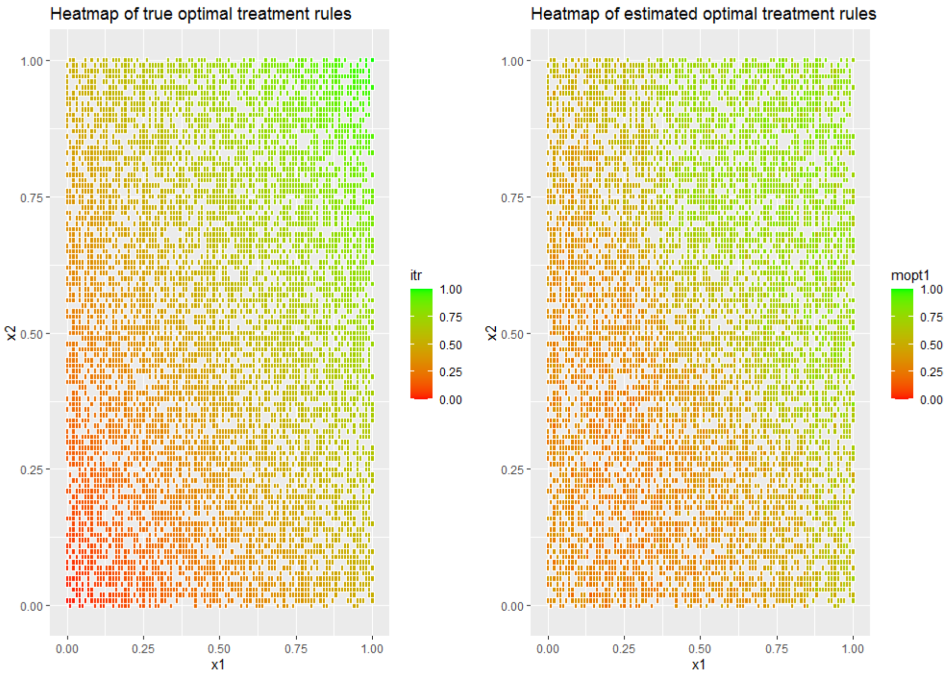

Figure 4.5. Simulation results: heatmaps of true and estimated optimal treatment regimes as a function of X_1 and X_2 as defined in the ITR model $c_{c3}(X, T)$.

Secondly, we compare our algorithm's prediction accuracy in terms of treatment outcomes $Y_t|X$ with the BART algorithm using AMSE at the optimized treatment level. Specifically, we compared the AMSE with treatment values at $0.25, 0.50$ and 0.75 for $c_{c1}(X,T)$; $0.2, 0.4, 0.6$ and 0.8 for $c_{c2}(X,T)$. For $c_{c3}(X,T)$, the AMSE is calculated using 10 treatment levels $(0.1, 0.2, 0.3, \cdots, 1)$. Results from 100 simulation runs are presented in Figure 4.6. The CERFIT algorithm outperforms BART under three model settings in predicting the treatment outcomes at the optimal treatment levels specified above.

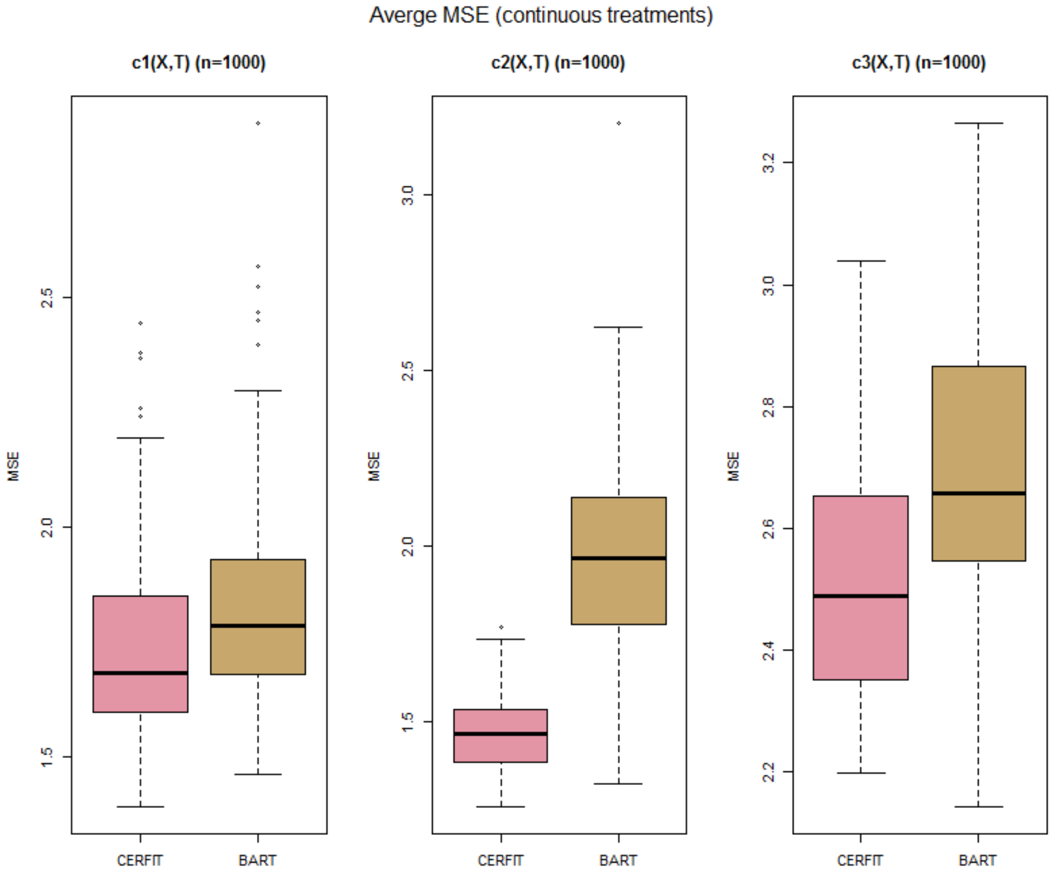

Figure 4.6. Simulation Results: comparison between CERFIT and BART regarding prediction accuracy using MSE based on 100 simulation runs for three proposed models under continuous treatments setting.

Lastly, we also evaluate the variable importance using MD. The results are presented in Figure 4.7. The smaller value of MD represents a stronger impact on the treatment effect, and therefore, indicates a higher rank of variable importance. In all three models, the proposed algorithm correctly identifies the important variables that interact with treatment effects. For both case 1 $c_{c1}(X, T)$ and case 3 $c_{c3}(X, T)$, only covariates X_1 and X_2 associated with the treatment effects. X_1 and X_2 are equally important in case 3, and X_2 is less important than X_1 for case 1. For case 2 $c_{c2}(X, T)$, X_1, X_2 and X_3 are the important covariates that directly interact with the treatment. Among which X_1 is the most important factor since X_2 and X_3's association with the treatment effects are both conditional on the value of X_1.

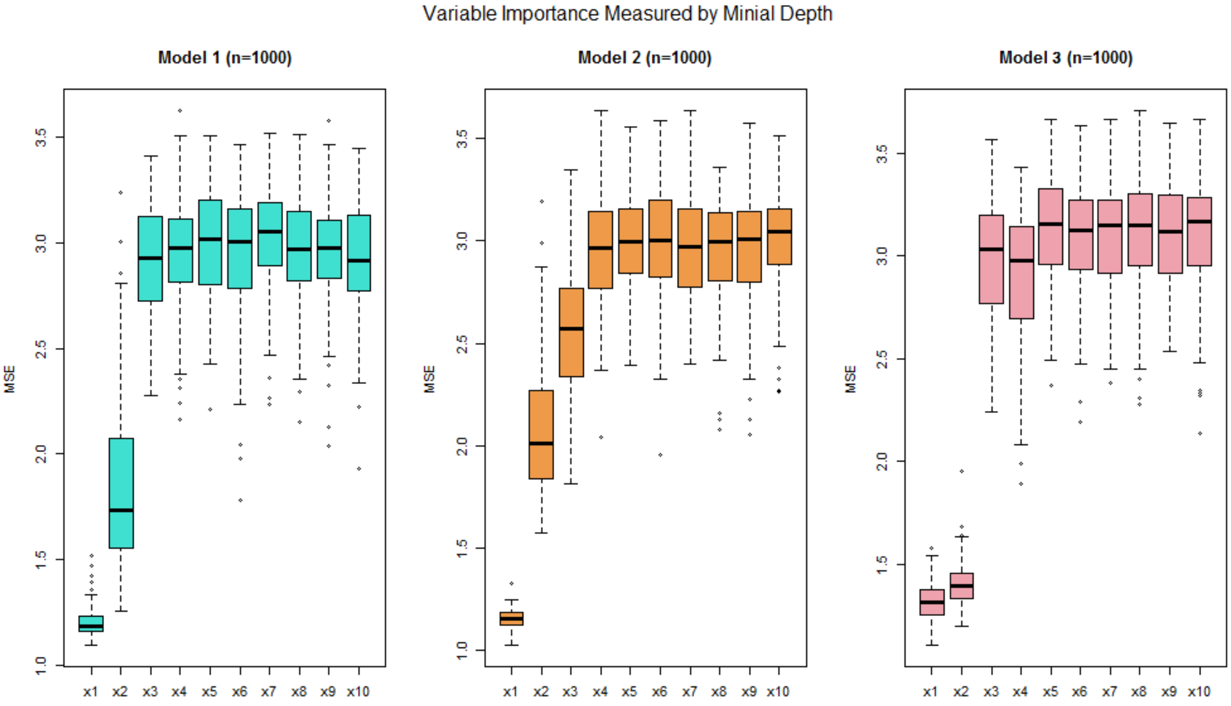

Figure 4.7. Simulation results: variable importance evaluated by minimal depth for three models under continuous treatment setting based on 100 simulation runs and sample size 1000.

4.6 Application: Learning ITR from Multiple Education Interventions

In this section, the proposed methods are illustrated using observational data collected from the educational field. Specifically, we are looking at three intervention programs introduced at a large university aimed to help students to be more successful in the bottleneck statistics introduction course. The first one is the free tutoring program (Tutor). Students voluntarily attend any tutoring sessions at their convenience, as the tutoring program is provided five days a week. Tutors are the graduate teaching associates who lead recital courses as a support class to the bottleneck statistics course mentioned above. Students receive one-on-one or group tutoring when they attend the tutoring sessions. Tutors answer student's specific statistical concept questions, homework questions or other course related questions. The recitation course (RC) is the second support program. School funding was available to offer the RC for around 20% of students. Students voluntarily enrolled in the course and met twice per week in a small group with an active learning environment to review the topic of the week, discuss conceptual issues, and work on extra but related statistics problem sets and data analyses. The third one is the peer-lead supplemental instruction (SI) academic support program. This program employs undergraduate students who have successfully completed the course in previous years to facilitate peer-learning sessions for current enrolling students. Students voluntarily drop into the SI session to meet SI leaders in a small group to review the topic of the week and discuss conceptual issues. The three programs were provided in parallel during the study semester. We apply the proposed methods to identify the optimal treatment regimes that maximize student's success in the course, which is measured by the final scores of the student. At individual student level, actionable outcome is providing students with individualized advising on selecting optimal intervention programs. At the school level, the study results intend to assist in determining financial resources should be devoted to which intervention programs, as well as identifying the group of students who benefit most from the interventions.

Data was collected from students who enrolled in the course from two consecutive Spring semesters with the same instructor. The descriptive summary of the study variables is present in Table 4.4. Totally 16 predictors are considered covering students' demographics, university information, course specific information, as well as admission information. The response variable is the final score, which is the number of points scored out of 1030 possible points. There were $n = 1401$ total students in the study, among which 842 students (60.1%) were female. The mean age was 18.62 years old with a standard deviation 0.88 years. The majority of students are sophomores 71.4%. In addition, there are 39.8% students who were under-represented minorities and 23.1% of students were first generation students attending college. The other covariates included in the study are whether students were part of the scholarship program, in a STEM major, or living on campus. We also considered students' academic performance covariates, such as SAT scores, high school GPA, and students' college GPA at the beginning of the study semester, as well as the total number of units enrolled and failed during the study semester. We consider each intervention alone and a combination of two or three interventions. Limited by the data availability, treatments of SI and Tutor are dichotomized using at least one attendance as the cutoff for treated and untreated. This helps to retain an acceptable size of the treated group, but may result in a small effect size. We intend to use this data to illustrate the proposed methods. The results should be cautiously interpreted due to this limitation.Eight treatments are considered in the study includingno treatment, which means those students did not participate in any of the treatments ($n =$ 589, $[42\%]$). There were 15.6% ($n =$ 219), 11.4% ($n =$ 160) and 8.9% $(n = 125)$ students that took the recitation course, attended SI, or attended Tutor programs, respectively. Only a small proportion of students had participated in more than two programs, except for those taking both RC and attending tutoring ($n = 112, [8.0\%]$).

Table 4.4. Descriptive Summary for Study Variables in the Application Data.

Variable		n=1401
Semester (%)	1	695 (49.6)
	2	706 (50.4)
Age (mean (SD))		18.62 (0.88)
Gender (%)	Female	842 (60.1)
	Male	559 (39.9)
College Levels (%)	Freshman	178 (12.7)
	Sophomore	1000 (71.4)
	Junior	181 (12.9)
	senior	42 (3.0)
URM (%)	No	843 (60.2)
under-represented minorities	Yes	558 (39.8)
First Generation at College (%)	No	1078 (76.9)
	Yes	323 (23.1)
EOP (%)	No	1321 (94.3)
Educational Opportunity Program	Yes	80 (5.7)
COMPACT scholar (%)	No	1247 (89.0)
	Yes	154 (11.0)
Scholarship Program (%)	No	1334 (95.2)
	Yes	67 (4.8)
On Campus Housing (%)	No	908 (64.8)
	Yes	493 (35.2)
STEM Major(%)	No	964 (68.8)
	Yes	437 (31.2)
Campus GPA (%)	A	442 (31.5)
	B	683 (48.8)
	C	223 (15.9)
	D	52 (3.7)
	F	1 (0.1)
SAT Composite (mean (SD))		1169.30 (130.76)
High School GPA (mean (SD))		3.69 (0.31)
Total Units Enrolled (mean (SD))		15.16 (2.04)
Total Units Failed (mean (SD))		0.84 (2.37)
Treatment (Programs) (%)	None	589 (42.0)
Tutoring program	Tutor	125 (8.9)
Supplemental Instruction	SI	160 (11.4)
Recitation course	RC	219 (15.6)
	RC+Tutor	112 (8.0)
	RC+SI	55 (3.9)
	Tutor+SI	72 (5.1)
	Tutor+SI+RC	69 (4.9)
Grade (mean (SD))		775.78 (173.88)

To identify the optimal treatment levels for individual students and identify important variables that impact treatment effects, we use 100-fold cross validation procedures. The original data is randomly split into 100 equal-sized groups, with each group containing about 14 students (one group contains 15 students). We use data from 99 groups as training data to grow the CERFIT and leave out one group of data as new data to make predictions. A total of $ntree = 500$ trees are constructed for each forest and 100 forests of interaction trees are built for 100-fold cross validations. Using our default $mtry$ formula, $mtry$ is set at 8. Variable importance is measured by the average of MD for each variable across 100 forests. The propensity scores and weights are estimated using GBM [54].

The analysis results are presented from two aspects. First, from the school administration perspective, the main questions are 1) Whether and which programs provide best results; 2) Which group of students benefit most from the provided intervention programs. In Figure 4.6, we present the treatment benefits for each of interventions contrast to no intervention at all. From the results we can see that for all different intervention levels, we have a proportion of the students receiving less than 0 score benefits, which means those students would not benefit from the specific intervention. And it is not surprising to see more than 90% students would benefit from taking all three interventions. If the university's budget allows for two interventions, the combination of SI and Tutor or the combination of Tutor and RC helps boost more than 75% students' performance by taking two interventions. If only one program can be provided, more than 75% of students would benefit from attending SI at least once; therefore, it should be recommended. The tutoring program may not be recommended since more than 50% students will not benefit from taking at least one tutoring session. In addition, tutoring also has the largest variation of treatment benefits. There are a noticeable number of students who could boost around 100 points by taking tutoring programs alone.

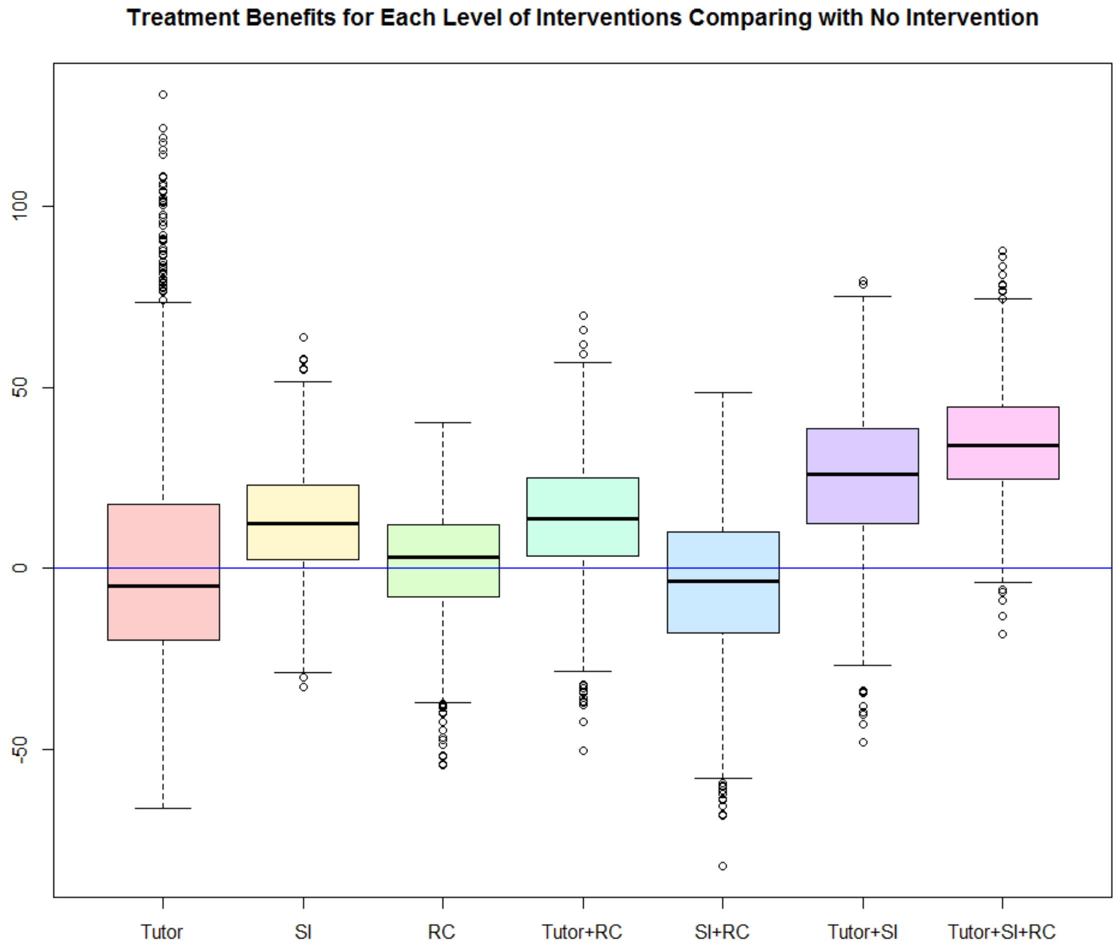

Figure 4.8. Application: treatment benefits of each interventions reflected by the differences of expected final scores students would receive when taking corresponding intervention as opposed to no intervention.

The variable importance ranking based on MD is presented in Figure 4.12. Variables with higher importance ranking are on the top of the figure. From Figure 4.12 we can see that high school GPA, SAT composite score, and total units enrolled are ranked as the most important predictors. Other important predictors include student's college GPA, age, gender, under-represented minorities (URM) status, on-campus housing, STEM major, college level and first Generation at College status. Variables with the least predictive power are whether a student is in a scholarship program or Educational Opportunity Program. The results indicate student's algebra readiness reflected by the student's high school GPA and SAT performance are top factors that impact on treatment effects. In addition, a student's socioeconomic background measured by the first generation college experience and under-represented minorities are also important predictors. Another interesting finding is that the number of units attempted also plays a role on whether a student can benefit from the intervention programs. Table 4.5 shows descriptive summarization of the top five important variables for each of treatments and treatment combinations. We transform the campus GPA from A to F to 1 to 5 and calculate the average to reflect the average campus GPA for each group. From Table 4.5 we can see that the average high school GPA and SAT composite score for Tutor group is the lowest with a mean 3.29 and 1088.03, respectively. This group also has the lowest campus GPA at the beginning of the semester. There are 92% of students in the SI group who have URM status. The SI and RC group also has a high proportion (81%) of students are URM status. The average number of units enrolled in the semester are lower than sample average (15.16 units) in the Tutor, SI and RC group, and higher than the sample average for the rest of treatment groups. There is no clear distinction observed for the group of students being recommended with all three treatments compared with the average characteristics of the sample data.

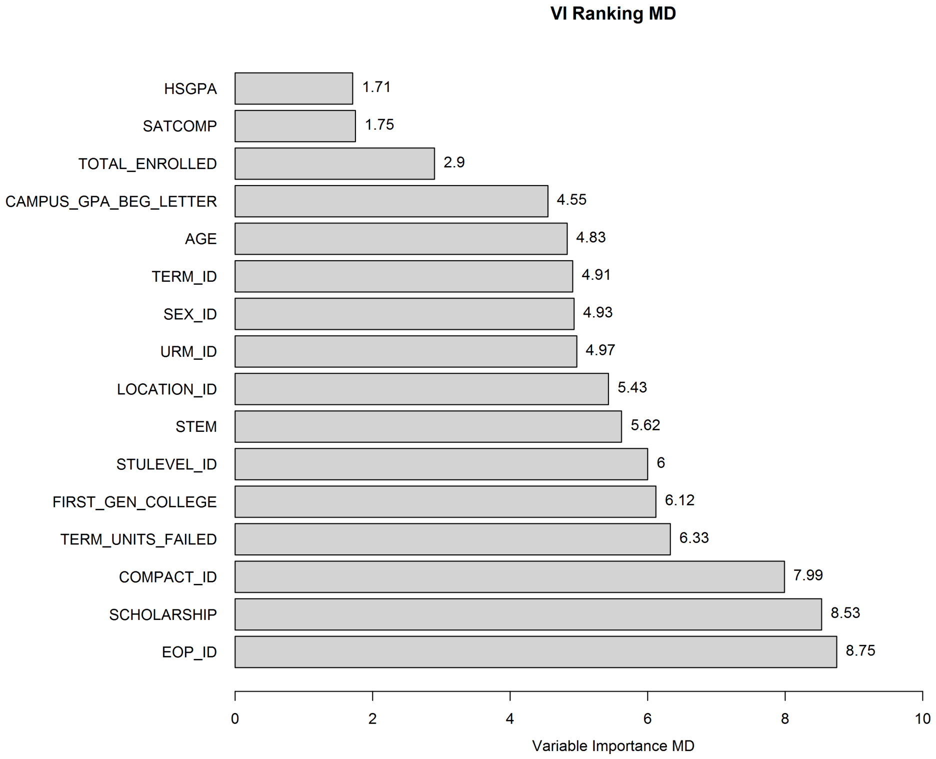

Figure 4.9. Application: summary of identified important predictors for the education interventions.

Table 4.5. Summary of Important Variables for the Education Interventions.

	HSGPA	SATCOMP	♯ENR	CAMGPA	AGE	URM	1stGEN
Tutor	3.29	1088.03	14.08	2.81	18.69	55%	30%
SI	3.59	1113.00	13.70	2.05	19.35	92%	52%
RC	3.80	1140.00	13.00	2.00	19.50	50%	50%
Tutor+RC	3.72	1157.35	15.32	2.09	18.78	24%	19%
SI+RC	3.85	1128.75	16.62	1.75	18.56	81%	50%
Tutor+SI	3.71	1246.94	15.37	1.76	18.56	40%	11%
All 3	3.76	1142.28	15.31	1.80	18.59	34%	27%

Next, from the individual student's perspective, we demonstrate how the results can be applied in personalized academic advising of optimal ITR for individual students. We randomly select 7 students, one from each recommended optimal treatment group. Table 4.6 presents the seven students' profile. The overview of student's benefits of taking each treatment is presented in Figure 4.6. The first student (ID=7) was a 19 years old under-represented minority. The student had a high school GPA at 3.56 and SAT composite score 990. The student enrolled 12 units and had a campus GPA of C at the beginning of the semester. The optimal treatment for the student is Tutor with an estimated score gain at 83.12. The suboptimal treatment for this student is the combination of Tutor and SI. Other treatments, especially the treatments including RC are not recommended. The student (ID=11) has slightly higher high school GPA and SAT scores, but same campus GPA and same numbers of units enrolled. The most distinct differences between the two students among these five important predictors are the age and whether the student has a URM status. The optimal treatment recommended to the student (ID=11) is SI with a benefit score at 39.49. The suboptimal recommendation is Tutor and SI, which is the same as student (ID=7). The student (ID=230) and the student (ID=19) has similar high school GPA, same SAT scores, same age and both are first generation at college, but student ID230 has URM status and only enrolled 13 units, while student ID19 was not an URM and enrolled 16 units for the semester. Their optimal treatments are RC (benefit score $= 34.32$) and both Tutor and RC (benefit score $= 44.18$), respectively. Interestingly, their suboptimal treatments are both the combination of Tutor and SI. Student (ID=148) had a good high school GPA and campus GPA. This student enrolled in the highest numbers of units (19). Even though the optimal treatment recommended is the combination of Tutor and SI, estimated benefits are less than 5 points comparing the suboptimal recommendation, which is taking SI alone. Student (ID=3) has a similar profile as student(ID=19), besides a slightly higher SAT score and a better campus GPA. This student is recommended with Tutor and SI, but with the lowest benefits among all 7 students. And it is noteworthy that the suboptimal treatment, which includes all three interventions has almost the same benefits as the optimal one. For the last student (ID=1), she has the highest GPA both from high school and college. Although she would receive the highest benefits by taking all three interventions, the suboptimal treatment (SI alone) would provide comparable benefits (18.93). As a summary, each program's treatment effects vary greatly among students, which confirms the heterogeneity of ITE. In general, the students with lower academic status seem to benefit most from the intervention. In addition, with multiple choices of treatments, the benefit differences between the optimal and the suboptimal treatment could be subtle, especially when considering the estimation natures. Recommendation with a proper consideration of financial constraints, time consuming and other tangible costs should be considered as well.

Table 4.6. Profile of Students: For Each Intervention, Random Selected One Student Received Optimal Treatment Recommendation with the Corresponding Intervention.

ID	HSGPA	SAT	♯ ENR	CAMGPA	AGE	URM	1GEN	TREAT	BENs	Grade
7	3.56	990	12	C	19	1	0	Tutor	83.12	463.00
11	3.63	1070	12	C	21	0	0	SI	39.49	800.82
230	3.84	1140	13	B	19	1	1	RC	34.32	601.49
19	3.82	1140	16	C	19	0	1	Tutor+RC	44.18	838.12
148	3.97	1100	19	B	18	1	0	SI+RC	25.17	767.27
3	3.82	1180	16	B	18	0	1	Tutor+SI	19.21	753.33
1	4.14	1210	15	A	18	1	0	All 3	22.31	965.59

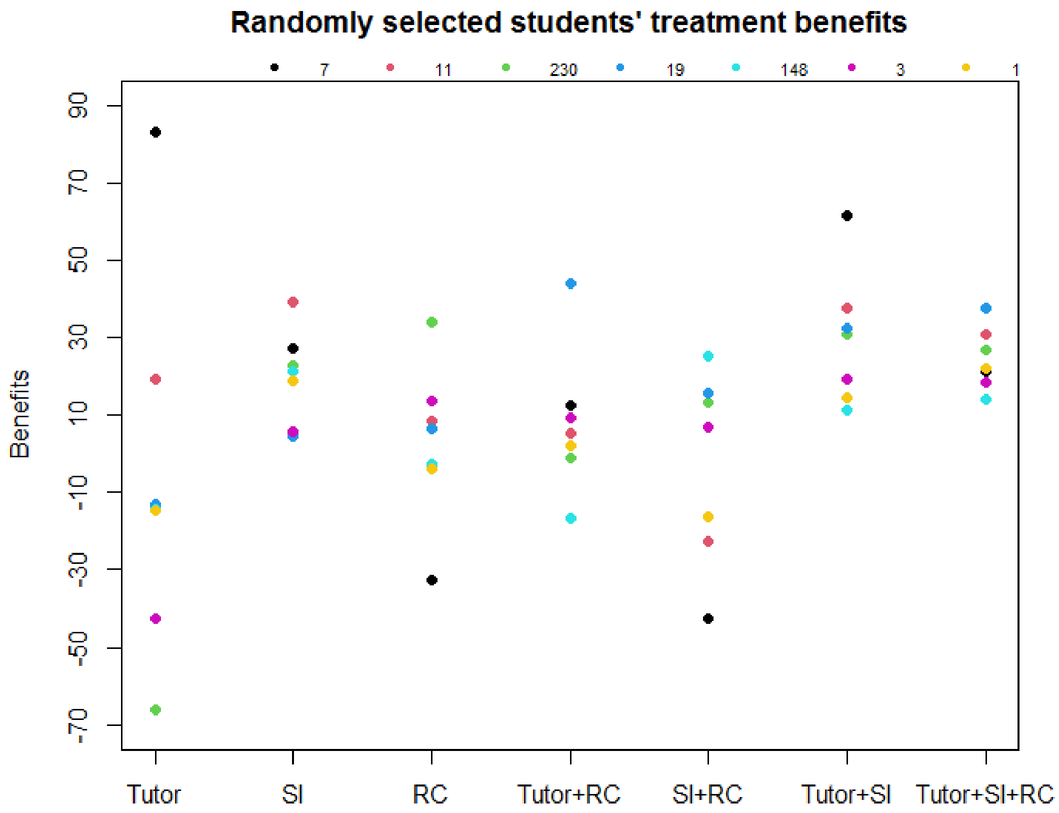

Figure 4.10. Application: dot plot of randomly selected individual student's treatment benefits for each of interventions. With each color corresponding to a unique student id.

Now, considering the situation that the university's budget could only support one intervention program, SI would be picked since the above analysis suggests that more than 75% students would benefit by attending the SI session at least once. The treatment of SI is considered as continuous treatment measured by the number of SI session students attended. Within the same data, we have 25% $(n = 356)$ students attended the SI at least once, with the highest number of SI attendance at 20. This dataset is not ideal since only a small proportion of students attended SI more than 12 times (1%) and the log transformation does not significantly improve the skewness of the data. With this in mind, we intend to briefly demonstrate how the proposed methods can be applied to identify the individualized optimal numbers of SI attendance. Taking the same 100-fold cross validation procedure, we grow 100

CERFIT forests, 500 trees for each forest. Because students can be involved with multiple interventions, RC indication and number of tutoring attendance are used as the extra covariates in the analysis. In Figure 4.6, we present treatment benefits of the 6 random selected students. We also included the students' detailed profile in the Table 4.7 for a reference. From Figure 4.6 we can see that the optimal ITR for each student varied. Student ID10, ID55 and ID15 could all benefit significantly from attending SI. However, the optimal numbers of attendances vary. Generally, it seems that at least 5 attendances should be recommended. Student (ID=5) does not benefit from SI. The benefits of attending the SI for Student ID201 and ID602 are trivial. The variable importance ranking for treatment effects of attending SI is presented in Figure 4.6. The top three important variables are SAT score, total units enrolled and high school GPA. The number of units failed in the semester also ranked within the top 5 predictors. The SI treatment also interacts with student's Tutor attendance (MSLC_visits) and college levels.

Table 4.7. Profile of Students: Randomly Selected Students with Different ITR Recommendations for Attending Supplemental Instruction Intervention.

ID	10	201	55	602	15	5
GRADE	919	898	748	756	336	849
GENDER	Female	Male	Male	Female	Female	Male
SEMESTER	2	1	2	2	1	1
COLLEGE LEV	Sophomore	Sophomore	Sophomore	Junior	Junior	Junior
AGE	19	18	18	19	19	19
URM	Yes	No	Yes	No	Yes	No
SAT COMP	1280	1210	1100	1240	1280	1180
HSGPA	4.00	3.74	3.71	3.97	3.76	3.94
FIRST GEN COLLEGE	No	No	No	Yes	No	No
DORM	No	No	No	No	Yes	No
STEM	No	Yes	Yes	Yes	No	Yes
CAMPUS GPA	A	A	B	B	B	C
TOTAL ENROLLED	16	14	15	16	15	14
UNITS FAILED	0	0	0	0	6	0
TUTOR	0	0	0	0	0	0
RC	Yes	No	No	No	No	Yes

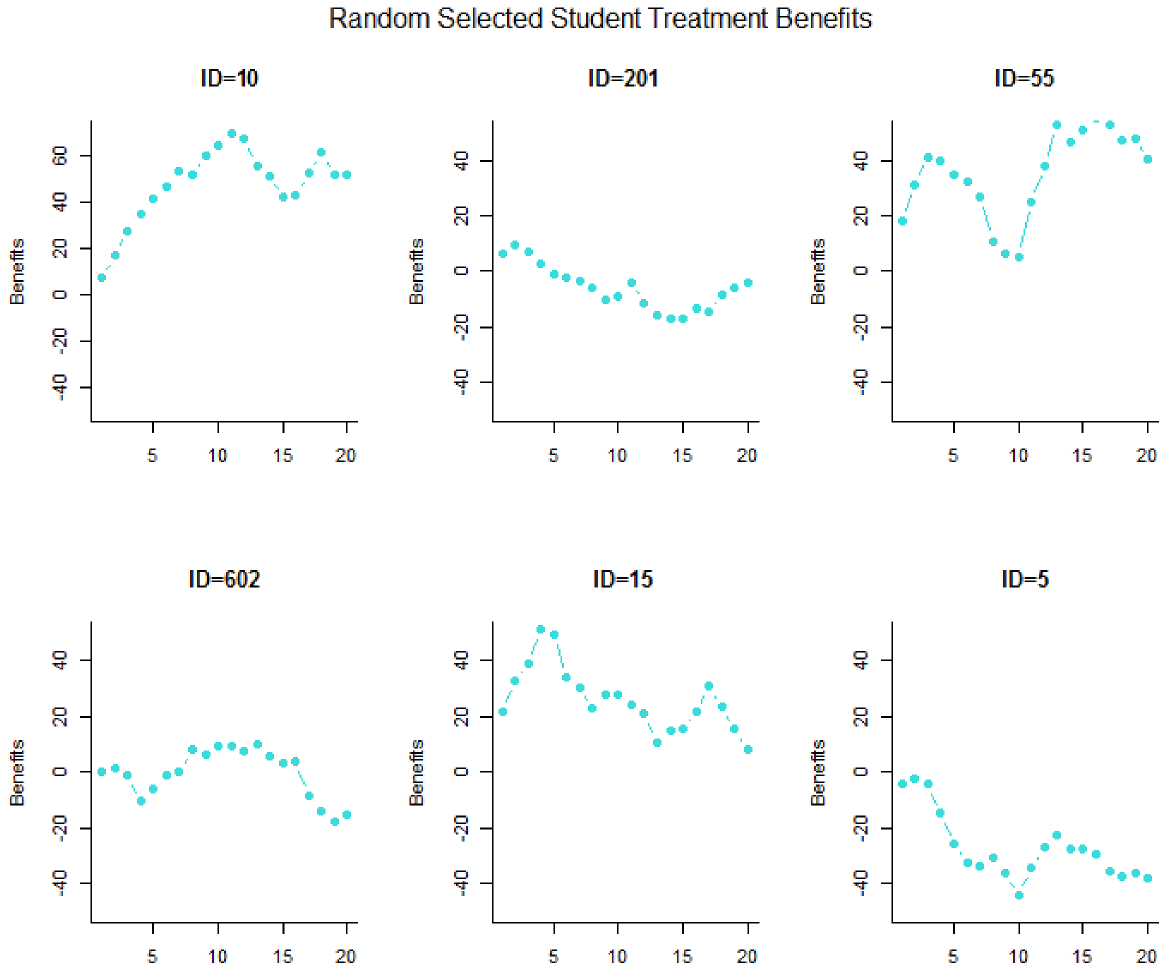

Figure 4.11. Application: sample plot of optimal ITR for randomly selected students attending supplemental instruction.

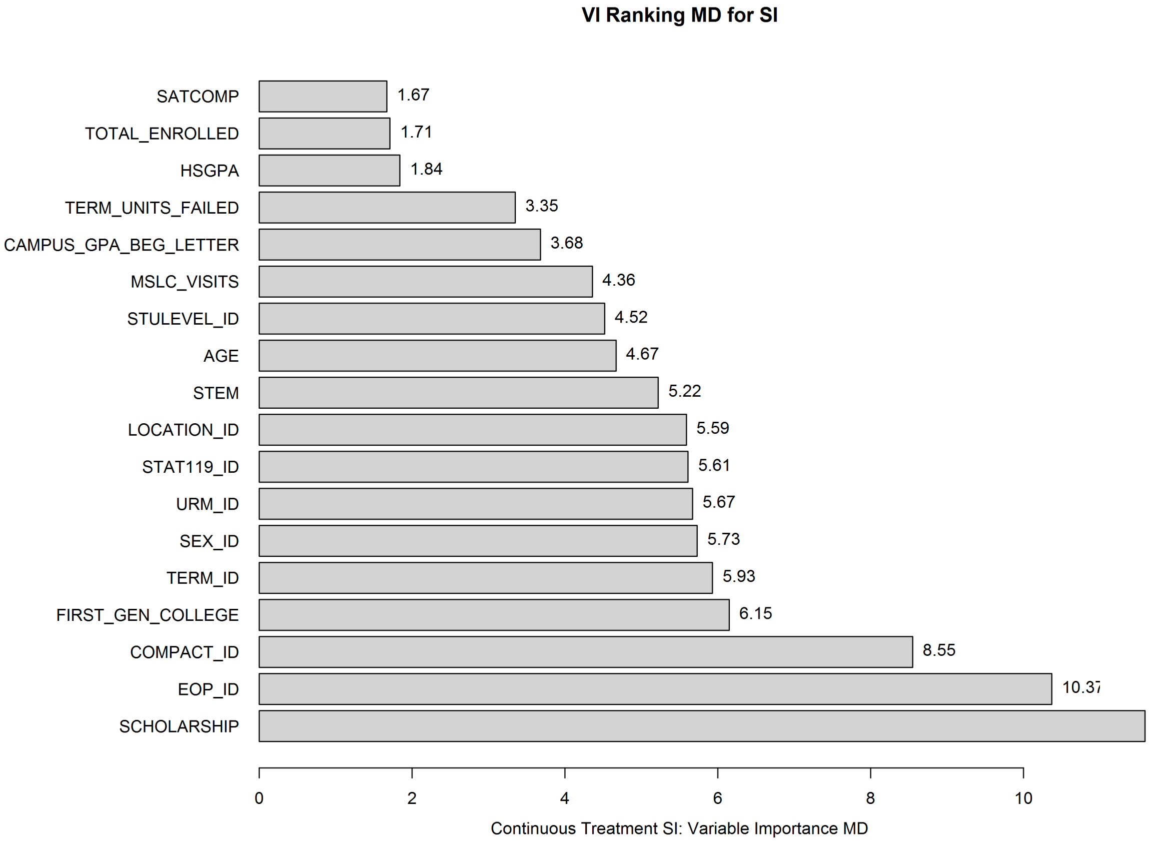

Figure 4.12. Application: variables importance ranking regarding treatment effects of attending supplemental instruction.

4.7 Discussion

Estimation of individualized treatment regimes using observational data holds great interest in various research fields. In this Chapter, we propose interaction tree based methods to estimate the optimal ITRs in multiple treatments and continuous treatment settings. By incorporating the general propensity score in the tree growth process, the proposed method can be applied to both random controlled and observational study data. The estimation accuracy and stability are further improved by replacing original responses with residuals estimated from the linear regression model. Simulation results show that CERFIT has competitive performance among all comparing methods with respect to correctly identifying the optimal ITR. The CERFIT algorithm also produces accurate variable importance ranking.

Several possible extensions can be explored for future study. By using Wald test of the interaction terms from the logistic regression model or in the Cox regression model for survival analysis in the splitting rule, the proposed method has the potential to be extended to binary outcomes or time to event survival data. In addition, the linear regression based residuals are used as responses in the current study. It might be worthwhile to compare the performance by using residuals estimated by other methods.

CHAPTER 5
CERFIT R PROGRAMMING

5.1 Overview

Proposed methods are all programmed and implemented using R software. Since there is no existing R Package available for building interaction trees, we develop a series of R functions to implement all the analysis in the study. All functions for implement CERFIT can be found at GitHub repository: `https://github.com/ll120/CERFIT`.

We utilize the R base function **lm** and **vcovHC** from the *sandwich* R package [79] for calculating robust Wald test statistics. We use the *partykit* R package [33] for growing a tree structure. The construction of a tree adopts a modified $CART$ and $RandomForest$ procedure. Each tree is grown using a weighted bootstrapping sample. A tree is grown by recursively partitioning the data into subgroups by exhaustive search of a best split. A tree stops growing when one of the following predetermined rules is reached: 1) the number of observations required to continue splitting goes below the predetermined $minsplit$; or 2) the number of observations required in each child node goes below the $minbuket$; 3) the tree depth reaches the $maxdepth.$ A total of $ntree$ trees form the final forest.

5.2 Constructing CERFIT

The **CERFIT** function constructs the CERFIT forest. It is a wrapper function that calls for functions to split, partition and grow trees. The usage of the function is as follows:

```
CERFIT (formula,
        data,
        ntrees,
        subset=NULL,
        method=c("RCT","observation"),
        PropForm=c("randomForest","CBPS","GBM"),
        mtry=NULL,
        nsplit=NULL,
        nsplit.random=TRUE,
        minsplit=30,
        minbucket=round(minsplit/3),
        maxdepth=30,
        sampleMethod=c("bootstrap","subsample"),
        useRes=TRUE)
```

The main arguments in the function include:

- formula: Formula to build CERFIT. Categorical predictors must be listed as a factor. e.g., $Y \sim x1 + x2 + x3 \mid treatment$
- data: Data use to grow a tree.
- ntrees: Number of trees grown.
- subset: Subset of data use to grow a tree.
- mtry: Number of variables randomly considered at each split.
- method: For observational study, method="observation"; for randomized study, method="RCT".
- PropForm: Methods used to generate propensity scores. Options are "randomForest","CBPS" or "GBM".
- nsplit: Number of outpoints selected for "exhaustive" search
- nsplit.random: Logical. indicates if process to select cutpts are random for "exhaustive" search.
- minsplit: Number of observations required to continue growing tree.
- minbucket: Number of observations required in each child node.
- sampleMethod: Method to sample learning sample. Options are"bootstrap" or "subsample".

- useRes: Logical, indicates whether growing a tree using linear regression residuals instead of original responses.

All trees are of object class $constparty$, which allows for using the print, plot functionality or extract elements from the tree provided by the *partykit*.

5.3 MAKING PREDICTION

The **predict.CERFIT** function can be used to make predictions of new test data using a CERFIT subject. The usage of the function is as follows:

```
predict.CERFIT (cerfit,
                data,
                newdata,
                gridval=NULL,
                prediction=c("overall","by iter"),
                type=c("response","ITE","node","opT"),
                alpha=1,
                useRes=TRUE)
```

The main arguments in the function include:

- cerfit: CERFIT forest subject
- data: Data use to grow a tree.
- newdata: New test data use to make prediction.
- gridval: For continuous treatment only. Specify grid values to make prediction.
- prediction: Method to return prediction using all trees or using first i trees. Options are “overall” and “by iter”.
- type: Prediction type returned. Options are “response”,“ITE”,“node” and “opT”. ”response” returns predicted response at each treatment level. “ITE” returns the treatment effects contrast to the first level. “node” returns node numbers. “opT” only works for true continuous treatment.

- alpha: The elastic-net mixing parameter α, with range $\alpha \in [0, 1]$. $\alpha = 1$ is the lasso (default) and $\alpha = 0$ is the ridge.
- useRes: Logical. indicates whether the prediction is based on trees growing using residuals from linear regression model.

5.4 OTHERS

The variable importance can be produced using **MinDepth** function, with a CERFIT subject. Fitting hundreds of trees involves calculating thousands of splitting statistics. Computation is intense and time consuming. Software programs such as $C++$ would improve the computation efficiency and will be implemented in future works. In the meantime, parallel computing is recommended with current R codes. A **CERFITparallel** function is provided in the current R program with limited options to adjust number of processes. More flexible parallel computing can be implemented using other existing R packages with high performance parallel functions, such as *snow*.

CHAPTER 6
CONCLUSION

6.1 Summary and Future Work

Estimation of individualized treatment regimes (ITR) using observational data, holds great interest in various research fields. ITR can be defined as a mapping between individual characteristics to a treatment assignment. The optimal ITR is the treatment assignment that maximizes expected individual treatment effects. However, treatment effects estimation under the counterfactual framework usually requires the assumption of strong ignorability. Without properly addressing issues of confounding when using observational study data, the estimation of treatment effects can be biased and unreliable. In addition, multiple treatments are common in many fields. Assigning treatment with optimal treatment effects among several or even continuous treatment options is important but challenging. In this study, we were interested in addressing these issues with proposed algorithms.

In Chapter 2, we considered a binary treatment setting. After giving an introduction of random forest of interaction trees (RFIT) and propensity score methods, we presented the new causal effect RFIT (CERFIT) algorithm. By integrating the propensity scores into the tree growing process, we extended the application of RFIT to the observational study context. Simulation studies demonstrated that CERFIT has superior performance with respect to prediction accuracy and variable importance ranking. We also illustrated the CERFIT algorithm for binary treatment through the assessment of a supplemental instruction course at a large public university. In Chapter 3, we proposed a residual based CERFIT and provided the proof of improved numerical stability by replacing original responses with residuals estimated from the linear regression model in the CERFIT algorithm. We conducted various simulation studies and the results demonstrated a significant improvement of prediction accuracy of residual based CERFIT. In Chapter 4, we proposed new CERFIT algorithms to transform non-binary treatment effects learning into a binary type learning task. We also proposed using general propensity score methods for non-binary treatment settings. Moreover, we conducted extensive simulation studies to assess the CERFIT's performance. We also illustrated the CERFIT method through learning optimal ITR among multiple education interventions. In Chapter 5, we introduced the main R functions we developed to implement the proposed methods. CERFIT demonstrates competitive performance among all competing methods in simulation studies for both binary and non-binary treatment settings. CERFIT's learning target is set as the treatment effects themselves. This allows the CERFIT algorithm to produce accurate variable importance ranking in terms of treatment effects. Even though we demonstrated the application of CERFIT with study data from the field of education, CERFIT can be applied to various areas where the estimation of individualized treatment regimes using observational data is of interest.

Since CERFIT depends on IPTW, its performance can be greatly impacted by the accuracy of the propensity score estimation. The selection of a proper prediction model for the propensity score is critical. We recommend machine learning methods such as boosting [53] because of their superior prediction accuracy. In addition, just as random forest variable importance measures are in favor of variables with many possible splits [67], CERFIT's variable importance measures may suffer from the same bias. For studies dealing with variables of different types, approaches suggested by [67] and [11] can be combined with the current CERFIT's variable importance algorithms to achieve unbiased variable selection.

Due to the intensive computations, the current R functions could be time consuming in growing a large forest. Prediction based on lasso regression with 10-fold cross validation for continuous treatment setting also slows down the program. The next step is improving the program efficiency by coding the splitting algorithm using $C++$ software program. There are also several other possible extensions that can be explored in future studies. By using Wald test of the interaction terms from the logistic regression model or in the Cox regression model for survival analysis in the splitting rule, the proposed method has the great potential to be extended to binary outcomes or time to event survival data. In addition, the linear regression based residuals are used as responses in the current study. It might be worthwhile to compare the performance by using residuals estimated by other methods.

BIBLIOGRAPHY

[1] D. Alemayehu, Y. Chen, and M. Markatou, *A comparative study of subgroup identification methods for differential treatment effect: performance metrics and recommendations*, Statistical Methods in Medical Research, 27 (2018), pp. 3658–3678.

[2] C. Allegra, J. Jessup, M. Somerfield, S. Hamilton, E. Hammond, D. Hayes, P. McAllister, R. Morton, and R. Schilsky, *American society of clinical oncology provisional clinical opinion: testing for kras gene mutations in patients with metastatic colorectal carcinoma to predict response to anti-epidermal growth factor receptor monoclonal antibody therapy*, Journal of Clinical Oncology, 27 (2009), pp. 2091–2096.

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

[4] P. C. Austin and E. A. Stuart, *The performance of inverse probability of treatment weighting and full matching on the propensity score in the presence of model misspecification when estimating the effect of treatment on survival outcomes*, Statistical Methods in Medical Research, 26 (2015), pp. 1654–1670.

[5] T. Bowles and J. Jones, *An analysis of the effectiveness of supplemental instruction: the problem of selection bias and limited dependent variables*, Journal of College Student Retention, 5 (2003), pp. 235–243.

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

[7] L. Breiman, A. Cutler, A. Liaw, and M. Wiener, *RandomForest: Breiman and Cutler's Random Forests for Classification and Regression*, 2018. R package version 4.6-14.

[8] L. Breiman, J. Friedman, R. Olshen, and C. Stone, *Classification and regression trees*, Wadsworth, Belmont,CA, 1984.

[9] L. Breimen, *Random forest*, Machine Learning, 45 (2001), pp. 5–32.

[10] R. P. Brent, *Algorithms for Minimization Without Derivatives*, Prentice-Hall, Englewood Cliffs, NJ, 1973.

[11] G. Cafri, P. Calhoun, and J. Fan, *High dimensional variable selection with clustered data: an application of random multivariate survival forests for detection of outlier medical device components*, Journal of Statistical Computation and Simulation, 89 (2019), pp. 1410–1422.

[12] R. Caruana, N. Karampatziakis, and A. Yessenalina, *An empirical evaluation of supervised learning in high dimensions*, JInternational Conference on Machine Learning, 08 (2008), pp. 96–103.

[13] R. Caruana and A. Niculescu-Mizil, *An empirical comparison of supervised learning algorithms.*, International Conference on Machine Learning, 06 (2006), pp. 161–168.

[14] S. Chen, L. Tian, T. Cai, and M. Yu, *A general statistical framework for subgroup identification and comparative treatment scoring*, Biometrics, 73 (2017), pp. 1199–1209.

[15] H. Chipman, E. George, and R. McCulloch, *Bart*: *bayesian additive regression trees*, The Annals of Applied Statistics, 4 (2010), pp. 266–298.

[16] H. Chipman and R. McCulloch, *BayesTree: Bayesian Additive Regression Trees*, 2016. R package version 0.3-1.4.

[17] S. R. Cole and M. A. Hernan, *Constructing inverse probability weights for marginal structural models*, American Journal of Epidemiology, 168 (2008), pp. 656–664.

[18] P. Dawson, J. van der Meer, J. Skalicky, and K. Cowley, *On the effectiveness of supplemental instruction: A systematic review of supplemental instruction and peer-assisted study sessions literature between 2001 and 2010*, Review of Educational Research, 84 (2014), pp. 609–639.

[19] E. Dusseldorp and I. Van Mechelen, *Qualitative interaction trees: a tool to identify qualitative treatmentsubgroup interactions*, Statistics in Medicine, 33 (2014), pp. 219–237.

[20] P. Feng, X. H. Zhou, Q. M. Zou, M. Y. Fan, and X. S. Li, *Generalized propensity score for estimating the average treatment effect of multiple treatments*, Statistics in Medicine, 31 (2012), pp. 681–697.

[21] C. Fong, C. Hazlett, and K. Imai, *Covariate balancing propensity score for a continuous treatment*: *application to the efficacy of political advertisements*, Annals of Applied Statistics, 12 (2018), pp. 156–177.

[22] C. Fong and K. Imai, *Covariate balancing propensity score for general treatment regimes*, 2014.

[23] J. C. Foster, J. M. Taylor, and S. J. Ruberg, *Subgroup identification from randomized clinical trial data*, Statistics in Medicine, 30 (2011), pp. 2867–2880.

[24] H. Fu, J. Zhou, and D. E. Faries, *Estimating optimal treatment regimes via subgroup identification in randomized control trials and observational studies*, Statistics in Medicine, 35 (2016), pp. 3285–3302.

[25] I. G., *The role of the propensity score in estimating dose-response functions*, Biometrika, 87 (2000), pp. 706–710.

[26] M. HAMBURG AND F. S. COLLINS, *The path to personalized medicine*, New England Journal of Medicine, 363 (2010), pp. 301–304.

[27] F. HARRELL, R. CALIFF, D. PRYOR, K. LEE, AND R. ROSATI, *Evaluating the yield of medical tests*, JAMA, 247 (1982), pp. 2543–2546.

[28] M. HERNN, B. BRUMBACK, AND J. ROBINS, *Marginal structural models to estimate the causal effect of zidovudine on the survival of hiv-positive men*, Epidemiology, 11 (2000), pp. 561–570.

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

[30] K. HIRANO AND G. W. IMBENS, *The Propensity Score with Continuous Treatments*, John Wiley and Sons, Ltd, 2004, ch. 7, pp. 73–84.

[31] T. HO, *Random decision forest*, in Proceedings of the Third International Conference on Document Analysis and Recognition, G. V. Avery, ed., vol. 1, ICDAR, 1995, pp. 278–282.

[32] P. W. HOLLAND AND D. B. RUBIN, *Causal inference in retrospective studies*, Evaluation Review, 12 (1988), pp. 203–231.

[33] T. HOTHORN AND A. ZEILEIS, *partykit: a modular toolkit for recursive partytioning in r*, Journal of Machine Learning Research, (2015), pp. 1–7.

[34] J. D. HULING AND M. YU, *Subgroup identification using the personalized package*, (2018).

[35] K. IMAI AND M. RATKOVIC, *Covariate balancing propensity score*, Journal of the Royal Statistical Society, 76 (2014), p. 243263.

[36] K. IMAI AND D. A. VAN DYK, *Causal inference with general treatment regimes*, Journal of the American Statistical Association, 99 (2004), pp. 854–866.

[37] H. ISHWARAN, T. GERDS, U. KOGALUR, R. MOORE, S. GANGE, AND B. LAU, *Random survival forests for competing risks*, Biostatistics, 15 (2004), pp. 757–773.

[38] H. ISHWARAN AND U. KOGALUR, *Fast Unified Random Forests for Survival, Regression, and Classification (RF-SRC)*, 2018. R package version 2.5.1.

[39] H. ISHWARAN, U. KOGALUR, A. GORODESKI, AND M. MINN, *High-dimensional variable selection for survival data*, JASA., 105 (2010), pp. 205–217.

[40] H. ISHWARAN AND J. D. MALLEY, *Synthetic learning machines*, BioData mining, 7 (2014), p. 28.

[41] M. JOFFE, T. TEN HAVE, H. FELDMAN, AND S. KIMMEL, *Model selection, confounder control, and marginal structural models: review and new applications*, The American Statistician, 58 (2004), p. 272279.

[42] A. P. JONES, F. G. HAPP, F. GILBERT, S. BURNETT, AND E. VIDING, *Feeling, caring, knowing: different types of empathy deficit in boys with psychopathic tendencies and autism spectrum disorder*, Journal of Child Psychology and Psychiatry, 51 (2010), pp. 1188–1197.

[43] J. KANG, X. SU, L. LIU, AND M. L. DAVIGLUS, *Causal inference of interaction effects with inverse propensity weighting, g−computation and tree based standardization*, Statistical Analysis and Data Mining, 7 (2014), pp. 323–336.

[44] E. B. LABER AND Y. Q. ZHAO, *Tree-based methods for individualized treatment regimes*, Biometrika, 102 (2015), pp. 501–514.

[45] M. LECHNER, *Program heterogeneity and propensity score matchin An application to the evaluation of active labor market policies*, Review of Economics and Statistics, 84 (2002), p. 205220.

[46] B. K. LEE, J. LESSLER, AND E. A. STUART, *Improving propensity score weighting using machine learning*, Statistics in Medicine, 29 (2010), pp. 337–346.

[47] B. K. LEE, J. LESSLER, AND E. A. STUART, *Weight trimming and propensity score weighting*, PLoS One, 6 (2011), p. 18174.

[48] M. LOPEZ AND R. GUTMAN, *Estimation of causal effects with multiple treatments*: *a review and new ideas*, Statistical Science, 32 (2017), pp. 432–454.

[49] M. LU, S. SADIQ, D. J. FEASTER, AND H. ISHWARAN, *Estimating individual treatment effect in observational data using random forest methods*, Journal of Computational and Graphical Statistics, 27 (2018), pp. 209–219.

[50] J. LUDWIG AND J. WEINSTEIN, *Biomarkers in cancer staging, prognosis and treatment selection*, Nature Review Cancer, 5 (2005), p. 845856.

[51] J. K. LUNCEFORD AND M. DAVIDIAN, *Stratification and weighting via the propensity score in estimation of causal treatment effects*: *a comparative study*, Statistics in Medicine, 23 (2004), pp. 2937–2960.

[52] L. M., *Identification and estimation of causal effects of multiple treatments under the conditional independence assumption*, vol. 13, 2001, pp. 43–58.

[53] D. MCCAFFREY, G. RIDGEWAY, AND A. MORRAL, *Propensity score estimation with*

boosted regression for evaluating causal effects in observational studies, Psychological Methods, 9 (2004), pp. 403–425.

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

[55] J. Pearl, *Causes of effects and effects of causes*, Sociological Methods & Research, 44 (2015), p. 2149164.

[56] K. Pelaez, R. Levine, J. Fan, M. Guarcello, and M. Laumakis, *Using a latent class forest to identify at risk students in higher education*, Journal of Educational Data Mining, 11 (2019), pp. 18–46.

[57] A. Peterffeund, . Rath, S. Xenos, and F. Bayliss, *The impact of supplemental instruction on students in stem courses*: *results from san francisco state university*, Journal of College Student Retention: Research, Theory and Practice, 9 (2008), pp. 487–503.

[58] Z. Qi, D. Liu, H. Fu, and Y. Liu, *Multi-armed angle-based direct learning for estimating optimal individualized treatment rules with various outcomes*, Journal of the American Statistical Association, 115 (2018), pp. 1–35.

[59] X. Qiu and Y. Wang, *Composite interaction tree for simultaneous learning of optimal individualized treatment rules and subgroups*, Statistics in Medicine, 38 (2019), pp. 2632–2651.

[60] J. Rabitoy, E.and Hoffman and D. Person, *Supplemental instruction*: *the effect of demographic and academic preparation variables on community college student academic achievement in stem-related fields*, Journal of Hispanic Higher Education, 14 (2015), pp. 244–255.

[61] . . Rath, S. P. Peterfreund, A. R.and Xenos, F. Bayliss, and N. Carnal, *Supplemental instruction in introductory biology i*: *enhancing the performance and retention of underrepresented minority students*, CBE−Life Sciences Education, 6 (2007), pp. 203–216.

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

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

[64] D. B. Rubin, *Estimating causal effects of treatments in randomized and non-randomized*, Journal of Educational Psychology, 66 (1974), pp. 668–701.

[65] D. B. RUBIN, *Causal inference using potential outcomes*, Journal of the American Statistical Association, 100 (2005), pp. 322–331.

[66] S. SETOGUCHI, S. SCHNEEWEISS, M. BROOKHART, R. GLYNN, AND E. COOK, *Evaluating uses of data mining techniques in propensity score estimation: a simulation study*, Pharmacoepidemiol Drug Safe, 17 (2008), pp. 546–555.

[67] C. STROBL, A.-L. BOULESTEIX, A. ZEILEIS, AND T. HOTHORN, *Bias in random forest variable importance measures: illustrations, sources and a solution*, BMC Bioinformatics, 8 (2007), pp. 1471–2105.

[68] X. SU, A. PEñA, L. LIU, AND R. LEVINE, *Random forests of interaction trees for estimating individualized treatment effects in randomized trials*, Statistics in Medicine, 37 (2018), pp. 2547–2560.

[69] X. SU, C.-L. TSAI, H. WANG, D. M. NICKERSON, AND B. LI, *Subgroup analysis via recursive partitioning*, Journal of Machine Learning Research, 10 (2009), pp. 141–158.

[70] C. D. SUTTON, *Classification and regression trees, bagging, and boosting*, 2005.

[71] Y. TAO, L. WANG, AND D. ALMIRALL, *Tree-based reinforcement learning for estimating optimal dynamic treatment regimes*, Annals of Applied Statistics, 12 (2018), pp. 1914–1938.

[72] Y. TAO, L. WANG, AND D. ALMIRALL, *Tree-based reinforcement learning for estimating optimal dynamic treatment regimes*, The annals of applied statistics, 12 (2018), p. 19141938.

[73] J. TIBSHIRANI, S. ATHEY, R. FRIEDBERG, V. HADAD, D. HIRSHBERG, L. MINER, E. SVERDRUP, S. WAGER, AND M. WRIGHT, *grf: Generalized Random Forests*, 2018. R package version 0.10.2.

[74] R. TIBSHIRANI, *Regression shrinkage and selection via the lasso*, ournal of the Royal Statistical Society, 58 (1996), pp. 267–288.

[75] S. VAN BUUREN, s. GROOTHUIS-OUDSHOORN, A. ROBITZSCH, G. VINK, L. DOOVE, S. JOLANI, R. SCHOUTEN, P. GAFFERT, F. MEINFELDER, AND B. GRAY, *mice: Multivariate Imputation by Chained Equations*, 2018. R package version 3.4.0.

[76] M. VAN DER LANN, E. POLLEY, AND A. HUBBARD, *Super learner*, Statistical Applications in Genetics and Molecular Biology, 6 (2006).

[77] S. WAGER AND S. ATHEY, *Estimation and inference of heterogeneous treatment effects using random forests*, Journal of the American Statistical Association, 113 (2018), pp. 1228–1242.

[78] S. XU, C. ROSS, M. A. RAEBEL, S. SHETTERLY, C. BLANCHETTE, AND D. SMITH,

Use of stabilized inverse propensity scores as weights to directly estimate relative risk and its confidence intervals, Value in Health, 13 (2010), pp. 273–277.

[79] A. ZEILEIS, S. KÖLL, AND N. GRAHAM, *Various versatile variances: An object-oriented implementation of clustered covariances in R*, Journal of Statistical Software, 95 (2020), pp. 1–36.

[80] Y. ZHANG, *listdtr: List-Based Rules for Dynamic Treatment Regimes*, 2016. R package version 1.0.

[81] Y. ZHANG, E. B. LABER, M. DAVIDIAN, AND A. A. TSIATIS, *Interpretable dynamic treatment regimes*, Journal of the American Statistical Association, 113 (2018), pp. 1541–1549.

[82] Y. ZHANG, E. B. LABER, A. TSIATIS, AND M. DAVIDIAN, *Using decision lists to construct interpretable and parsimonious treatment regimes*, Biometrics, 71 (2015), pp. 895–904.

[83] Y. ZHAO, D. ZENG, A. J. RUSH, AND M. R. KOSOROK, *Estimating individualized treatment rules using outcome weighted learning*, Journal of the American Statistical Association, 107 (2012), p. 11061118.

[84] Y. ZHU, D. COFFMAN, AND D. GHOSH, *A boosting algorithm for estimating generalized propensity scores with continuous treatments*, Journal of Causal Inference, 3 (2015), p. 2540.

ABSTRACT OF THE DISSERTATION

CAUSAL EFFECT RANDOM FOREST OF INTERACTION TREES FOR LEARNING INDIVIDUALIZED TREATMENT REGIMES IN OBSERVATIONAL STUDIES: WITH APPLICATIONS TO EDUCATION STUDY DATA

by

LUO LI

Doctor of Philosophy in Computational Science-Statistics

Claremont Graduate University and San Diego State University, 2020

Learning individualized treatment regimes (ITR) using observational data holds great interest in various fields, as treatment recommendations based on individual characteristics may improve individual treatment benefits with a reduced cost. It has long been observed that different individuals may respond to a certain treatment with significant heterogeneity. ITR can be defined as a mapping between individual characteristics to a treatment assignment. The optimal ITR is the treatment assignment that maximizes expected individual treatment effects. Rooted from personalized medicine, many studies and applications of ITR are in medical fields and clinical practice. Heterogeneous responses are also well documented in educational interventions. However, unlike the efficacy study in medical studies, educational interventions are often not randomized. Study results often suffer greatly from self-selection bias. Besides the intervention itself, the efficacy and effectiveness of interventions usually interact with a wide range of confounders.

In this study, we propose a novel algorithm to extend random forest of interaction trees to Casual Effect Random Forest of Interaction Trees (CERFIT) for learning individualized treatment effects and regimes. We first consider the study under a binary treatment setting. Each interaction tree recursively partitions the data into two subgroups with greatest heterogeneity of treatment effect. By integrating propensity score into the tree growing

process, subgroups from the proposed CERFIT not only have maximized treatment effect differences, but also similar baseline covariates. Thus it allows for the estimation of the individualized treatment effects using observational data. In addition, we also propose to use residuals from linear models instead of the original responses in the algorithm. By doing so, the numerical stability of the algorithm is greatly improved, which leads to an improved prediction accuracy. We then consider the learning problem under non-binary treatment settings. For multiple treatments, through recursively partitioning data into two subgroups with greatest treatment effects heterogeneity with respect to two randomly selected treatment groups, the algorithm transforms the multiple learning ITR into a binary task. Similarly, continuous treatment can be handled through recursively partitioning the data into subgroups with greatest homogeneity in terms of the association between the response and the treatment within a child node. For all treatment settings, the CERFIT provides variable importance ranking in terms of treatment effects. Extensive simulation studies for assessing estimation accuracy and variable importance ranking are presented. CERFIT demonstrates competitive performance among all competing methods in simulation studies. The methods are also illustrated through an assessment of a voluntary education intervention for binary treatment setting and learning optimal ITR among multiple interventions for non-binary treatments using data from a large public university.