Mean-weighted case specific random forests for estimating causal effects

by

Linus Addae

B.S, University of Cape Coast, Cape Coast, Ghana, 2007

MSc, Youngstown State University, Ohio, 2013

AN ABSTRACT OF A DISSERTATION

submitted in partial fulfillment of the
requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Statistics
College of Arts and Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2021

Abstract

Causal inference is a branch of statistics that deals with determining how responses are affected by treatments. In this dissertation, we examine two problems in causal inference under the Neyman-Rubin causal model (NRCM): estimation of *counterfactuals*—hypothetical unobserved responses of units under different treatment conditions—and treatment effect estimation under *treatment spillover*—when the treatment status of one unit affects the response of another.

First, we extend the case specific random forest (CSRF) methodology to develop mean-weighted case specific random forests (MWCSRF) to estimate the average treatment effect for the treated (ATT). We consider a setting under which the data contains many control and very few treated units, and covariate space for the treated units is a small subspace of that for the control units. For example, treated units may be those that underwent an experimental procedure and control units may be the set of units in a national database. Our approach is as follows. First, we compute bootstrap sample weights for each treated unit to oversample control units nearby the treated unit. Then, we average these weights together to construct one set of "treated" sample weights. Next, we use random forests to estimate the prognostic score—the expected control outcome given a set of covariates—for each treated unit. Finally, we estimate the ATT by taking the average difference of the responses and the estimated prognostic scores across all treated units. We show via a simulation study that MWCSRF performs favorably compared to the standard random forest, causal forests, and genetic matching under both homogeneous and heterogeneous treatment effect settings, especially when the number of treated units is small. Additionally, we demonstrate that, when parallelization is not available, MWCSRF requires significantly less runtime than CSRF. We confirm our findings on a study on the efficacy of the National Supported Work Demonstration. Additionally, we develop an R package for MWCSRF.

Secondly, we discuss the problem of treatment spillover in the context of Fisher's Lady Tasting Tea experiment. We show that, by design, Lady Tasting Tea can violate the stable unit treatment value assumption (SUTVA), which requires the response of a unit to be only affected by the treatment status of that unit. We show that SUTVA may be violated under this model even when, for a given cup, the Lady's milk-first likelihood is always higher when that cup actually receives milk first. Moreover, we show that SUTVA holds under two conditions: one in which the Lady's likelihood for a cup is the same regardless of whether that cup was given milk first or tea first, and one in which the Lady always makes perfect guesses. These results further emphasize that SUTVA cannot be classified solely as treatment spillover problems, but can be inherent in the design of an experiment. Additionally, this result may have implications for teaching causal inference, as it may be preferable to introduce randomized experiments using examples that do not inherently violate SUTVA.

Mean-weighted case specific random forests for estimating causal effects

by

Linus Addae

B.S, University of Cape Coast, Cape Coast, Ghana, 2007

MSc, Youngstown State University, Ohio, 2013

A DISSERTATION

submitted in partial fulfillment of the
requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Statistics
College of Arts and Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2021

Approved by:

Major Professor
Michael Higgins

Copyright

Abstract

Causal inference is a branch of statistics that deals with determining how responses are affected by treatments. In this dissertation, we examine two problems in causal inference under the Neyman-Rubin causal model (NRCM): estimation of *counterfactuals*—hypothetical unobserved responses of units under different treatment conditions—and treatment effect estimation under *treatment spillover*—when the treatment status of one unit affects the response of another.

First, we extend the case specific random forest (CSRF) methodology to develop mean-weighted case specific random forests (MWCSRF) to estimate the average treatment effect for the treated (ATT). We consider a setting under which the data contains many control and very few treated units, and covariate space for the treated units is a small subspace of that for the control units. For example, treated units may be those that underwent an experimental procedure and control units may be the set of units in a national database. Our approach is as follows. First, we compute bootstrap sample weights for each treated unit to oversample control units nearby the treated unit. Then, we average these weights together to construct one set of "treated" sample weights. Next, we use random forests to estimate the prognostic score—the expected control outcome given a set of covariates—for each treated unit. Finally, we estimate the ATT by taking the average difference of the responses and the estimated prognostic scores across all treated units. We show via a simulation study that MWCSRF performs favorably compared to the standard random forest, causal forests, and genetic matching under both homogeneous and heterogeneous treatment effect settings, especially when the number of treated units is small. Additionally, we demonstrate that, when parallelization is not available, MWCSRF requires significantly less runtime than CSRF. We confirm our findings on a study on the efficacy of the National Supported Work Demonstration. Additionally, we develop an R package for MWCSRF.

Secondly, we discuss the problem of treatment spillover in the context of Fisher's Lady Tasting Tea experiment. We show that, by design, Lady Tasting Tea can violate the stable unit treatment value assumption (SUTVA), which requires the response of a unit to be only affected by the treatment status of that unit. We show that SUTVA may be violated under this model even when, for a given cup, the Lady's milk-first likelihood is always higher when that cup actually receives milk first. Moreover, we show that SUTVA holds under two conditions: one in which the Lady's likelihood for a cup is the same regardless of whether that cup was given milk first or tea first, and one in which the Lady always makes perfect guesses. These results further emphasize that SUTVA cannot be classified solely as treatment spillover problems, but can be inherent in the design of an experiment. Additionally, this result may have implications for teaching causal inference, as it may be preferable to introduce randomized experiments using examples that do not inherently violate SUTVA.

Table of Contents

List of Figures xi

List of Tables xii

1 Introduction and Overview 1

1.1 Neyman Rubin Causal Model 2

1.2 Assumptions 4

1.3 Random Forest (RF) 6

1.4 Extensions to random forests 8

1.4.1 Adjustments on sampling units 9

1.4.2 Tuning parameters: *nodesize*, *mtry*, and B 9

1.4.3 Splitting rule 11

1.4.4 Estimation using classification or regression trees 12

1.4.5 Other extensions to random forests 13

1.5 Case Specific Random Forest (CSRF) 14

1.6 Using random forests in causal inference 15

1.6.1 Causal Forest 15

1.6.2 Estimation of counterfactuals 16

1.7 Organization of dissertation 18

2 Mean-weighted case specific random forests for estimating causal effects 19

2.1 Introduction 19

2.1.1 Motivating Data Example: The National Supported Work Demonstration 21

2.2 Preliminaries . . . 22
2.2.1 Overview of Causal Inference . . . 22
2.2.2 Random Forest (RF) . . . 25
2.2.3 Case Specific Random Forests (CSRF) . . . 27
2.2.4 Estimating causal effects using random forests . . . 28
2.3 Mean Weighted Case Specific Random Forests (MWCSRF) . . . 29
2.3.1 Estimating ATT with MWCSRF . . . 31
2.4 Simulation Study . . . 31
2.4.1 Results of homogeneous and heterogeneous data simulation . . . 35
2.4.2 Randomization of treated units in the covariate space . . . 40
2.5 Results of real data illustration . . . 42
2.6 Discussion and Conclusion . . . 44

3 MWCSRF: An R package for estimating ATT using potential outcomes . . . 45
3.1 Introduction . . . 45
3.2 Neyman Rubin Causal Model . . . 46
3.3 MWCSRF Functions . . . 47
3.4 Example . . . 49

4 The Lady Tasting Tea Revisited: Insights on SUTVA Violations from a Canonical Example . . . 52
4.1 Introduction . . . 52
4.2 Model of response and SUTVA violations . . . 53
4.2.1 Lady Tasting Tea set-up . . . 53
4.2.2 Lady's model of response . . . 54
4.3 When Lady Tasting Tea Violates SUTVA . . . 55
4.3.1 Demonstration of SUTVA Violation . . . 55
4.3.2 SUTVA Violation Under Monotonicity of Likelihoods . . . 56

4.3.3 No SUTVA Violation Under Sharp Null 57
4.3.4 No SUTVA Violation Under Perfect Knowledge 57

5 Conclusion 59
5.1 Summary of dissertation 59
5.2 Future Research 61

Bibliography 62

A Summary of estimates of the models from the homogeneous and heterogeneous with true ATT as **4** when the number of predictors is varied 70

B Some results for homogeneous and heterogeneous data simulation 86

List of Figures

2.1 *Two dimensional representation of the covariates of each simulation scenario. For figure (a), N=210, N_c =200, and N_t =10. For figure (b) and (c), N=525, N_c =500, and N_t =25 and N=1025, N_c =1000, and N_t =25 respectively.* . . . 34

2.2 *Plot of empirical bias when varying the number of predictors with sample size 210 and M=0* . . . 35

2.3 *Plot of empirical bias when varying sample size with P=10 and M=0* . . . 37

2.4 *Plot of empirical bias when varying the number control units with P=10 and M=0* . . . 38

2.5 *A random distribution of treated units in the covariate space* . . . 41

2.6 *Plot of empirical bias when varying correlation for a simulation with the random distribution of treated units in the covariate space.* . . . 41

2.7 *The covariate space of earnings in 1974 against earnings in 1975 for LaLonde's NSW dataset.* . . . 42

3.1 *Example demonstrating the use of MWCSRF to estimate treatment effect for the treated(ATT)* . . . 50

List of Tables

1.1 A summary showing the difference between a unit's potential outcome under treatment and control τ_i is not estimable since a unit can receive only one treatment condition (treatment or control), but the average treatment effect τ is estimable. Δ represent unobserve potential outcomes. 4

2.1 Summary of estimates with actual ATT unknown and run time for M = 0, P = 5 and 10, N = 1010,1025, and 1050, where T.SE is the Theoretical Standard Error. 40

2.2 Summary of estimates of the models for the NSW dataset 43

3.1 Summary of functions in MWCSRF package 48

4.1 Example of likelihoods for which guesses violate SUTVA. 55

4.2 Example of likelihoods satisfying strict monotonicity for which guesses violate SUTVA 56

A.1 Summary of estimates of the models from homogeneous data simulation with true ATE as **4.0** and $N = 210$, $N_c = 200$, $N_t = 10$, and $p = 5$ 71

A.2 Summary of estimates of the models from homogeneous data simulation with true ATE as **4.0** and $N = 210$, $N_c = 200$, $N_t = 10$, and $p = 10$ 71

A.3 Summary of estimates of the models from homogeneous data simulation with true ATE as **4.0** and $N = 210$, $N_c = 200$, $N_t = 10$, and $p = 50$ 72

A.4 Summary of estimates of the models from the heterogeneous data simulation with true ATT unknown and $N = 210$, $N_c = 200$, $N_t = 10$, and $p = 5$ 72

A.5 Summary of estimates of the models from the heterogeneous data simulation with true ATT unknown and $N = 210$, $N_c = 200$, $N_t = 10$, and $p = 10$ 73
A.6 Summary of estimates of the models from the heterogeneous data simulation with true ATT unknown and $N = 210$, $N_c = 200$, $N_t = 10$, and $p = 50$ 73
A.7 Summary of estimates of the models from the homogeneous data simulation with true ATE as **4.0** and $N = 410$, $N_c = 400$, $N_t = 10$, and $p = 10$ 74
A.8 Summary of estimates of the models from the homogeneous data simulation with true ATE as **4.0** and $N = 1025$, $N_c = 1000$, $N_t = 25$, and $p = 10$ 75
A.9 Summary of estimates of the models from the homogeneous data simulation with true ATE as **4.0** and $N = 2050$, $N_c = 2000$, $N_t = 50$, and $p = 10$ 75
A.10 Summary of estimates of the models from the heterogeneous data simulation with true ATT unknown and $N = 410$, $N_c = 400$, $N_t = 10$, and $p = 10$ 76
A.11 Summary of estimates of the models from the heterogeneous data simulation with true ATT unknown and $N = 1025$, $N_c = 1000$, $N_t = 25$, and $p = 10$ 76
A.12 Summary of estimates of the models from the heterogeneous data simulation with true ATT unknown and $N = 2050$, $N_c = 2000$, $N_t = 50$, and $p = 10$ 77
A.13 Summary of estimates of the models from homogeneous data simulation with true ATE as **4.0** and $N = 210$, $N_c = 200$, $N_t = 10$, and $p = 10$ 78
A.14 Summary of estimates of the models from homogeneous data simulation with true ATE as **4.0** and $N = 410$, $N_c = 400$, $N_t = 10$, and $p = 10$ 79
A.15 Summary of estimates of the models from homogeneous data simulation with true ATE as **4.0** and $N = 525$, $N_c = 500$, $N_t = 25$, and $p = 10$ 79
A.16 Summary of estimates of the models from homogeneous data simulation with true ATE as **4.0** and $N = 1025$, $N_c = 1000$, $N_t = 25$, and $p = 10$ 80
A.17 Summary of estimates of the models from the heterogeneous data simulation with true ATT unknown and $N = 210$, $N_c = 200$, $N_t = 10$, and $p = 10$ 80

A.18 Summary of estimates of the models from the heterogeneous data simulation with true ATT unknown and $N = 410$, $N_c = 400$, $N_t = 10$, and $p = 10$ 81
A.19 Summary of estimates of the models from the heterogeneous data simulation with true ATT unknown and $N = 525$, $N_c = 500$, $N_t = 25$, and $p = 10$ 81
A.20 Summary of estimates of the models from the heterogeneous data simulation with true ATT unknown and $N = 1025$, $N_c = 1000$, $N_t = 25$, and $p = 10$ 82
A.21 Summary of estimates of the models from homogeneous data simulation when treated units are randomly distributed in the covariate space with true ATE as **4.0** and $N = 210$, $N_c = 200$, $N_t = 10$, and $p = 5$ 83
A.22 Summary of estimates of the models from homogeneous data simulation when treated units are randomly distributed in the covariate space with true ATE as **4.0** and $N = 210$, $N_c = 200$, $N_t = 10$, and $p = 10$ 84
A.23 Summary of estimates of the models from the heterogeneous data simulation when treated units are randomly distributed in the covariate space with true ATT unknown and $N = 210$, $N_c = 200$, $N_t = 10$, and $p = 5$ 84
A.24 Summary of estimates of the models from the heterogeneous data simulation when treated units are randomly distributed in the covariate space with true ATT unknown and $N = 210$, $N_c = 200$, $N_t = 10$, and $p = 10$ 85

B.1 Summary of estimates of the models from homogeneous data simulation with true ATE as **4.0** and $N = 410$, $N_c = 400$, $N_t = 10$, and $p = 5$ 87
B.2 Summary of estimates of the models from the heterogeneous data simulation with true ATT unknown and $N = 410$, $N_c = 400$, $N_t = 10$, and $p = 5$ 88
B.3 Summary of estimates of the models from homogeneous data simulation with true ATE as **4.0** and $N = 410$, $N_c = 400$, $N_t = 10$, and $p = 10$ 89
B.4 Summary of estimates of the models from the heterogeneous data simulation with true ATT unknown and $N = 410$, $N_c = 400$, $N_t = 10$, and $p = 10$ 90

B.5 Summary of estimates of the models from homogeneous data simulation with true ATE as **4.0** and $N = 410$, $N_c = 400$, $N_t = 10$, and $p = 50$ 91
B.6 Summary of estimates of the models from the heterogeneous data simulation with true ATT unknown and $N = 410$, $N_c = 400$, $N_t = 10$, and $p = 50$ 92
B.7 Summary of estimates of the models from homogeneous data simulation with true ATE as **4.0** and $N = 1025$, $N_c = 1000$, $N_t = 25$, and $p = 5$ 93
B.8 Summary of estimates of the models from the heterogeneous data simulation with true ATT unknown and $N = 1025$, $N_c = 1000$, $N_t = 25$, and $p = 5$ 94

Chapter 1

Introduction and Overview

According to Hill and Stuart (2015), causal inference is an intellectual discipline that considers the assumptions, study designs, and estimation strategies that allow researchers to draw causal conclusions based on data. Thus, causal inference can simply be referred to as the study of the impact of treatments on response. One factor that affects the meaningful interpretation of causal relationships is the presence of confounding variables. Confounders are covariates that are related to both the treatment exposure (i.e. risk factor, intervention, or treatment) and the outcome. Bias in treatment effect estimates may occur when there are differences in the distribution of the confounders between exposed and unexposed units, and these differences are not accurately incorporated in the estimates (MacMahon et al., 1981).

Randomized experiments are the gold standard for performing causal inference. In particular, randomization insures that the allocation of confounders is independent (asymptotically) of the treatment status. The probabilities associated with each unit (Rosenbaum and Rubin, 1983)—known from the design of the experiment—are all that are required to achieve unbiased average treatment effect estimates. However, most data are generated through observational studies, in which treatment is not randomized to units, and hence, such independence between treatment and confounders cannot be guaranteed. The problem of confounding effect in observational studies is often addressed by statistical matching,

stratification models, or multivariate techniques.

Two problems often studied in causal inference are the treatment interference and use of modern machine learning methods to estimate the causal effects. Interference occurs when the treatment received by an individual may affect the outcomes of other individuals. This incident of interference often occurs in settings in which the outcomes of the various individuals are obtained through social interaction (Manski, 2000). Random forest machine learning techniques, such as standard random forests, case-specific random forests (CSRF), causal forests, generalized random forests, etc are gaining popularity in the causal inference domain for estimating treatment effect.

Our work on machine learning methods is focused on settings with many control units and very few treated units to efficiently estimate the average treatment effect for the treated (ATT). Often the treated units are condensed in one area of the covariate space. Typically, the comparison group (control group) is formed from a nonexperimental national database and the treated units originate from the design experiment. A notable example of this in the literature is that of LaLonde (1986), which compared the experimental estimate of the efficacy of National Supported Work Demonstration (NSW) to that using nonexperimental data from a national database to form the comparison group. We propose mean-weighted case-specific random forests (MWCSRF), a version of the random forest that replaces the bootstrap samples with mean weighted bootstrap samples to estimate the treatment effect, to estimate treatment effects.

1.1 Neyman Rubin Causal Model

The use of Neyman-Rubin causal model framework has received attention across many fields such as statistics, economics, political science,medicine, and so on. Some of the works done in this fields include statistics (Holland, 1986; Rosenbaum, 2002, 2005; Rubin, 1974b, 2006), economics (Abadie and Imbens, 2006; Dehejia and Wahba, 2002; Heckman, 2008),

political science (Brady, 2002; Rathbun, 2008; Sekhon, 2008), and medicine (Brady, 2002; Christakis and Iwashyna, 2003; Rubin, 1997). The origin of the model framework can be traced to Neyman (1923(1990)) and his nonparametric model for finite number of treatments where there exists two potential outcomes for each unit, one if the unit receive treatment and the other if is control. Thus, a causal effect is the difference between two potential outcomes, but only one potential outcome is actually observed.

Now, suppose there are two treatment regimes: treatment and control, and let $t = 1$ and $t = 0$ denote these respectively. Let T_i be the treatment indicator: $T_i = 1$ if unit i received the treatment and $T_i = 0$ if unit i received the control. Let Y_{i1} denote the potential outcome if unit i receives the treatment and Y_{i0} be the potential outcome for unit i in the control regime. Then the treatment effect for observation i is defined as

$$\tau_i = Y_{i1} - Y_{i0} \tag{1.1}$$

and the average treatment effect(ATE) as

$$\tau = \frac{1}{n}\sum_{i=1}^{n}\tau_i = \frac{1}{n}\sum_{i=1}^{n}(Y_{i1} - Y_{i0}), \tag{1.2}$$

where n is the number of units in a finite population.

Since only one potential outcome is only truly observed and not both, the unobserved potential outcome is often referred to as counterfactual. Thus, the observed outcome for the observation i can be modeled by Neyman-Rubin Causal Model (NRCM) as

$$Y_i = Y_{i1}T_i + Y_{i0}(1 - T_i), \tag{1.3}$$

where the potential outcomes Y_{i1} and Y_{i0} are fixed quantities and not random. This model does not make any explicit distributional assumption. To mitigate the missing data problem,

we consider the average treatment effect(ATE) instead of the unit treatment effect:

$$\tau = E(Y_{i1} - Y_{i0}) \tag{1.4}$$

Equation 1.4 implies that we can obtain the ATE using the observed response for different units. To illustrate the use of equation 1.4 to find the ATE, let's consider $n = 6$. The ATE for the $n = 6$ units can be expressed as:

$$\begin{aligned} \tau &= \frac{1}{6}\sum_{i=1}^{6} Y_{i1} - Y_{i0} \\ &= \frac{1}{6}[(Y_{11} - Y_{10}) + (Y_{21} - Y_{20}) + \cdots + (Y_{61} - Y_{60})] \\ &= \frac{1}{6}[(Y_{11} + Y_{21} + \cdots + Y_{61}) - (Y_{10} + Y_{20} + \cdots + Y_{60})] \end{aligned} \tag{1.5}$$

Table 1.1: *A summary showing the difference between a unit's potential outcome under treatment and control τ_i is not estimable since a unit can receive only one treatment condition (treatment or control), but the average treatment effect τ is estimable. Δ represent unobserve potential outcomes.*

	Potential Outcome			Observed		
Unit(i)	Treat	Control	τ_i	T_i	Y_i	τ_i
1	Y_{11}	Y_{10}	$Y_{11} - Y_{10}$	1	Y_{11}	Y_{11}- Δ
2	Y_{21}	Y_{20}	$Y_{21} - Y_{20}$	1	Y_{21}	Y_{21}- Δ
3	Y_{31}	Y_{30}	$Y_{31} - Y_{30}$	1	Y_{31}	Y_{31}- Δ
4	Y_{41}	Y_{40}	$Y_{41} - Y_{40}$	0	Y_{40}	Δ - Y_{40}
5	Y_{51}	Y_{50}	$Y_{51} - Y_{50}$	0	Y_{50}	Δ - Y_{50}
6	Y_{61}	Y_{60}	$Y_{61} - Y_{60}$	0	Y_{60}	Δ - Y_{60}
ATE	$\tau = \frac{1}{6}[(Y_{11} - Y_{10}) + \cdots + (Y_{61} - Y_{60})]$			$\hat{\tau} = \frac{1}{3}[(Y_{11} + Y_{21} + Y_{31}) - (Y_{40} + Y_{50} + Y_{60})]$		

1.2 Assumptions

The potential outcomes under treatment and control cannot be observed simultaneously in causal inference. Thus, we cannot make progress in causal inference without assumptions. Notable among the assumptions is the stable unit treatment value assumption (SUTVA),

which is implicit in the Neyman-Rubin potential outcomes model Cox (1958); Rubin (1980).

Assumption 1.2.1. *Stable unit value treatment assumption (SUTVA). SUTVA states that the potential outcomes for any unit do not vary with the treatments assigned to any other units, and there are no different versions of the treatment Rubin (1986).*

This assumption ensures that, irrespective of how a unit received treatment, the outcome Y_i will be Y_{it} or Y_{ic}. This further implies that the experimental design has no influence on the potential outcomes or the causal effect. This assumption also indicates that there is no interference between units; this can be seen in the Neyman-Rubin model (1.3) since the outcome of a unit is related to only its own treatment indicator. In causal inference, SUTVA ensures that (1) there exists as many potential outcomes as the number of value the treatment can take on, and (2) under SUTVA we can observe at least one potential outcomes for each unit.

The no interference part of this assumption is a difficult condition to attain even in a randomized design experiment. Since randomized experiment units are likely to interact with each other, randomization alone cannot ensure the realization of this part of the assumption.

In observational studies, covariates $\mathbf{X}$ are often correlated with response. Hence, under these settings, when estimating treatment effects, the following two assumptions are required.

Assumption 1.2.2. *(Ignorability). Conditional on the covariates* $\mathbf{X}$*, treatment assignment is independent of potential outcomes:*

$$P(T_i|Y_{i1}, Y_{i0}, \boldsymbol{X}) = P(T_i|Y_i, \boldsymbol{X}). \tag{1.6}$$

Assumption 1.2.3. *(Common support). For all values of observable covariates* $\mathbf{X}$*, each unit has some probability of getting treatment or control:*

$$0 < P(T_i|\boldsymbol{X} = X) < 1. \tag{1.7}$$

The common support and ignorability assumptions together constitute *strong ignorability* or unconfoundedness. These assumptions guarantee the creation of comparable groups and estimate the ATE unbiasedly. This is shown below:

$$\begin{aligned} & E(E(Y_{i1}|T_i = 1, \mathbf{X}) - E(Y_{i0}|T_i = 0, \mathbf{X})) = E(E(Y_{i1}|T_i = 1, \mathbf{X})) - E(E(Y_{i0}|T_i = 0, \mathbf{X})) \\ & = E(E(Y_{i1}|\mathbf{X})) - E(E(Y_{i0}|\mathbf{X})) \\ & = E(Y_{i1}) - E(Y_{i0}) \\ & = \tau \end{aligned} \tag{1.8}$$

The first equality holds as a result of common support and the second from ignorability.

1.3 Random Forest (RF)

Random forest is a nonparametric machine learning technique for classification and regression prediction problems which was first introduced by Breiman (2001). Random forest has become popular machine learning methodology in the literature due to its prediction efficiency and flexibility of being combined with other methods such as quantile regression forest (Meinshausen, 2006), generalized random forests (GRF) (Athey et al., 2019), random survival forests (Ishwaran et al., 2008), case specific random forests (CSRF) (Xu et al., 2016), and orthogonal random forests (Oprescu et al., 2018).

The standard random forest methodology fits a decision tree to different bootstrap samples. During the tree growing process, a random sample of $m < p$ predictors are considered in each step which lead to different trees from each sample and predictions from each tree are then averaged. We illustrate the method with a succinct description of random forest that fits a model for predicting a regression type problem for continuous or binary response Y and a set of predictors X, where X is p-dimensional. Consider the training data set $\mathbf{D} = \{D_i = (X_i, Y_i), i = 1, 2, ...N^*\}$ which consists of N^* cases. We used these N^* cases to

develop a model for predicting response at a particular vector X_0 of predictors.

The algorithm for random forest for regression type problem is given below:

1. For $j = 1, ..., B$, draw independent and uniform bootstrap samples $\mathbf{D}^* = \{D_i^* = i = 1, 2, ...N^*\}$ of size N^* from the training data set.

2. Grow regression and classification trees R_j and C_j respectively to the bootstrapped data set $\mathbf{D}^*$ in a single node K by recursively repeating the steps below until minimum pre-specified $nodesize = w$ is obtained, where $nodesize$ is the minimum number of observations in the terminal node.

    i. Select $mtry$ predictors at random from all the p predictors and split on the best predictor variable.

    ii. Split the nodes into two subnodes K_1 and K_2. For regression choose among all such splits the one that minimizes the Sum of squares error (SSE)

$$\sum_{b=1}^{2} \sum_{i \in K_b} (Y_i^* - \bar{Y_b})^2, \tag{1.9}$$

    where b=1,2 and $\bar{Y_b}$ is the mean response of the training observations within the subnode b.

    For classification, select the split that minimizes the Gini Impurity/Index

$$G = \sum_{b=1}^{B} \hat{p}_{mb}(1 - \hat{p}_{mb}), \tag{1.10}$$

    where G is the measure of total variance across the B classes, and $\hat{p}_{mb}$ is the proportion of training observations in the m^{th} region that are from the b^{th} class.

3. Output the trees $\{R_j\}_1^B$ and $\{C_j\}_1^B$ to form a random forest for regression and classification respectively.

To make prediction at a given independent variable X_0 for continuous response variable, we average the prediction from each tree $\hat{R}_j(X_0)$, that is

$$\hat{R}_j(X_0) = \frac{1}{B}\sum_{j=1}^{B}\hat{R}_j(X_0), \tag{1.11}$$

Likewise, to obtain prediction for X_0 for a categorical response variable, we first obtain the class prediction of the j^{th} tree $\hat{C}_j(X_0)$ and then obtain the majority vote for

$$\hat{C}_j(X_0) = \text{majority votes}\{\hat{C}_j(X_0)\}_1^B. \tag{1.12}$$

The concept of a majority vote is such that each random forest predicts a different outcome for the same test feature. Then, taken each expected outcome into account, votes will be calculated. Thus, the majority vote for a random forest classifier is the final class label that the classification models predict most frequently. This process is repeated using the same trees to obtain predictions at different values of independent variables.

Also, note that for classification, other split rules such as classification error rate and cross-entropy can also be considered to evaluate the quality of the split during tree construction. The classification error rate is a particularly desired split rule if the aim is prediction accuracy.

1.4 Extensions to random forests

Since its introduction by Breiman (2001), random forest have seen some tweaks to the main driving parameters that affect accuracy of the original algorithm in the literature. These parameters are the number of trees grown (*ntree*), the number of variables randomly sampled as possible variables at each split (*mtry*), the minimum number of observations in the terminal node (*nodsize*), and the bootstrap samples (D^*) for tree construction.

1.4.1 Adjustments on sampling units

The adjustment on sampling units in the random forest algorithm is a strategy in sampling the units for tree construction to improve performance. A notable approach found in the literature is Dynamic Random Forest (DRF). DRF by Bernard et al. (2012) is an extension made to the random forest methodology in the literature. This method constructs the trees in the forest adaptively. According to Bernard et al. (2012), the main idea is to guide the tree induction so that each tree will complement the existing trees in the ensemble as much as possible. This is done here by resampling the training data, inspired by boosting algorithms and combined with other randomization processes used in traditional RF methods. This procedure resampled the training sample by first randomly sampling N training observation with replacement(bagging) and then reweight the data (boosting) or using the adaptive resampling procedure employed in boosting. They concluded their method improve in terms of accuracy compared to standard random forest(RF) methodology.

The case specific random forest (CSRF) is a version of random forest introduced by Xu et al. (2016) that grows decisions trees particular to a test case. The standard random forest described in Section 1.3 makes all predictions from the same set of trees, but CSRF is built specifically for a conditional value of the covariates, say X_0, where prediction is desired. The construction of CSRF is essentially the same steps outlined in Section 1.3, but the step 1 in the RF algorithm is replaced with weighted bootstrap samples. In this setting, the training data cases closer to X_0 are allocated higher weights than those farther from X_0. The significant step here is constructing a proximity measure on the subset of predictors associated with the response.

1.4.2 Tuning parameters: *nodesize*, *mtry*, and B

Tuning parameters which are the main drivers of the random forest algorithm, have not been widely investigated in the literature. The default values of these parameters in the random

forest package R (randomForest) by Liaw and Wiener (2002) are based on the original Fortran code by Breiman and Cutler (2002) original Fortran code has not been extensively scrutinized by researchers. For classification and regression problems, the default value for B (tree) is 500, but *mtry* and nodesize have different values for classification and regression. For classification problems, the default value for *mtry* is $\sqrt{(p)}$ and *nodesize* is 1 and for regression, *mtry* is p/3, and *nodesize* is 5, where p is the number of predictor variables. Work focusing on these parameters include that by Díaz-Uriarte and De Andres (2006), Huang et al. (2008), and Genuer et al. (2010). A more significant number of B decreases the forest's variance, and more accurate predictions are possible to obtain. Notably, choosing a larger B does not result in overfitting problems. However, a more significant number of B results in computational burden because computations may not be accomplished in a reasonable amount of time. Thus, Díaz-Uriarte and De Andres (2006) noted the choice of B is irrelevant provided it is large enough. Díaz-Uriarte and De Andres (2006) make this observation in a prediction problem involving microarray data sets, where the objective is to classify patients according to their genetic profiles. Díaz-Uriarte and De Andres (2006) also argued that setting the *nodesize* to 1 or 5 does not have any profound effect on out-of-bag (OOB) error and the relationship between the out-of-bag(OOB) error and *mtry*. Thus, the default *nodesize*=1 should be a good choice. The influence of *mtry* on the performance of the random forest method is also extensively investigated by Díaz-Uriarte and De Andres (2006). They establish that *mtry* does not significantly impact the performance of the method, though, in some cases, increasing *mtry* results in a minimal decrease in error rate, and decreasing *mtry* often increases error rate. This observation was made specific to simulation with very few relevant genes. However, Genuer et al. (2010) concluded that the default value of *mtry* is either perfect or too small. Hence, a conventional approach may be to pick *mtry* as large as p.

1.4.3 Splitting rule

The splitting rules used in the random forest algorithm by Breiman (2001) are sum of squares error(SSE) and Gini impurity/index for regression and classification problems respectively when constructing the trees in the forest. The ranger package R by Wright and Ziegler (2017), which performs random forest for classification and regression, incorporate the same splitting rules as in random forest but has different splitting rule for survival forest called logrank.

The effect of these splitting rules on random forest has been extensively investigated in the literature by Ishwaran (2015) for regression and classification problems. They have also introduced a hybrid method that utilizes random split-point called unweighted variance. They observed that the sum of squares error(SSE) and Gini impurity/index splitting rules have end-cut preference (ECP) property (Breiman et al., 1984; Morgan and Messenger, 1973). This property generally favors splits that occurs near the edge of noisy or non-informative variables (Breiman et al., 1984). Although the ECP property has mostly been considered as unfavorable property for a splitting rule, Ishwaran (2015) claims that the ECP property is beneficial even when considering a split on an influential variable, when this important variable is in a region of space with poor signal. Ishwaran (2015) also emphases that the ECP property ensures that a split on a noisy variable occurs near the edge.

The use of Hellinger distance (HD) splitting rule for binary response (positive, negative) has been investigated in the literature by Aler et al. (2020). They claim HD is more suitable splitting rule for imbalance classification problems compare with Gini index. The extratrees is a splitting rule for growing extremely randomized trees (Geurts et al., 2006). For this splitting rule, the number of random splits are considered for each variable selected for splitting.

The maximally selected rank statistics (maxstat) splitting rule has been implemented in the literature by Wright et al. (2017). This splitting rule splits nodes in two steps (Wright

et al., 2017). In the first step, maximally selected rank statistics are computed for each covariate, and the split point with the maximal standardized test statistic is selected. This is achieved by selecting the candidate covariate with the smallest p-value under the null hypothesis of no association between the split point and the covariate. If the adjusted p-value is not less than the pre-specified level of significance α, no split is performed. Thus, the optimal split is the split point corresponding to the maximally selected rank statistics in the initial step.

Another splitting rule implemented in the literature is the beta distribution by Weinhold et al. (2019) to model bounded outcomes. In this approach, the likelihood function of the beta distribution is used to select the split during the tree construction (Weinhold et al., 2019). During the tree construction, a combination of predictor variable and split point that maximizes the likelihood function of the beta distribution, with parameter estimates obtained from the nodes of currently built tree.

The log-rank test splitting rule has been implemented in the literature by Nasejje et al. (2017). For this splitting rule, at each node, a randomly selected $\sqrt{p}$ among p predictors are selected as candidate variables for splitting the node into two subnodes. The log-rank statistic is computed at a node for the subnodes formed by all possible splits on all covariates used for splitting at that node (Nasejje et al., 2017). The covariate with the largest significant log-rank statistic computed from one of the subnode created by the split is chosen. The node is then partition into two subnodes based on the values of the covariates obtained from the split with the largest log-rank statistic.

1.4.4 Estimation using classification or regression trees

The quantile regression forest (Meinshausen, 2006) is another extension made to the Breiman (2001) random forest in the literature for estimating conditional quantiles (conditional median or other quantiles), especially in high dimensional settings. The quantile regression forest package (quantregForest) is based on the standard random forest package. Through

numerical examples, they concluded that the algorithm compares favorably in terms of predictive power.

The weighted random forest(WRF) proposed by Winham et al. (2013) is an extension of the random forest (RF) to improve predictive performance. Their method focuses on weighting trees to improve predictive performance. They train on three-fourth of the original samples, and thus, each tree contains approximately half of the entire sample as in-bag to construct the trees and one-fourth was out-of-bag (OOB) used to evaluate the tree performance and calculate tree weights. Their approach incorporates tree-level weights to produce more accurate trees in prediction and calculation of variable importance (Winham et al., 2013). They found through simulation and a real data study that their method modestly improves RF, but the effect is only noticeable in high dimensional settings and when effect sizes are larger than what is realistic in genetics of complex disease. They concluded the WRF method is unlikely to result in the improvement of identifying important features in high dimensional genetic data but may be relevant in situations with larger effect sizes.

1.4.5 Other extensions to random forests

Also in the literature is the extension of the random forest to survival analysis called random survival forest (RSF) (Ishwaran et al., 2008). Often, survival data are analyzed using parametric methods such as Cox proportional hazards regression or the Kaplan-Meier estimator (Collett, 2015). These methods fail to capture the nonlinear effects and rely on some restrictive assumptions. Thus, to overcome this challenge, a nonparametric approach is proposed by Ishwaran et al. (2008) known as RSF. Their method differs from Breiman (2001) in that the sampling mechanism and the splitting rule are altered. While the default sampling scheme for Breiman (2001) is sampling with replacement, for RSF, the sampling mechanism is with or without replacement, but the default is without replacement. The splitting rule for their method is that the candidate variable that maximizes the survival difference. This is achieved by searching through all possible variables X and split values c, and selecting

the X^* and c^* that maximize the survival difference (Ishwaran et al., 2008). They conclude that their method improves prediction.

Clustering forest (CF) (Yan et al., 2013) is motivated by the original work by Breiman (2001) random forest in the context of unsupervised classification learning. CF randomly searches a high-dimensional data cloud to obtain a good local clustering and then aggregates via spectral clustering to obtain cluster assignments for the whole data set. The search for good local clustering is guided by a cluster quality measure and progressively improves each local clustering in a fashion that resembles the tree growth in a random forest.

1.5 Case Specific Random Forest (CSRF)

The case specific random forest (CSRF) is a version of random forest introduced by Xu et al. (2016) that grows decision trees specific to a test case. The standard random forest described in Section 1.3 makes all predictions from the same set of trees but CSRF is built specific for a conditional value of the covariates, say X_0, where prediction is desired. The construction of CSRF is essentially the same steps outlined in Section 1.3, but Step 1 in the RF algorithm is replaced with weighted bootstrap samples. In this setting, the training data cases that are closer to X_0 are allocated higher weights than those farther from X_0. The significant step here is constructing a proximity measure on the subset of predictors associated with the response. Xu et al. (2016) used an initial bootstrap aggregation (bagging) of trees (Breiman, 2001) to construct the measure of proximity and the corresponding resampling probability weights.

The algorithm for CSRF replacing the Step 1 in the algorithm from Section 1.3 is given below:

1. Grow B decision trees following the RF algorithm in Section 1.3 with $mtry = p$ and $nodesize = w$. Note that this RF reduces to bagging since $mtry = p$.

2. Define the training case $\mathbf{D_i} = (X_i, Y_i)$ and let E_i be the number of B trees that contain

both X_0 and X_i in the same terminal node.

3. For $i = 1, 2, \ldots, N^*$, define Z_i be the resampling probability weights of D_i relative to X_0 as

$$Z_i = \frac{E_i}{\sum_{j=1}^{N^*} E_j}. \tag{1.13}$$

Note that $(Z_1, Z_2..., Z_{N^*})$ defines a probability distribution on the observations in D (the training data set). The *nodesize* w determines how far apart the weights Z_i are ordered on the training observations $D = \{D_1, D_2, ..., D_{N^*}\}$.

Smaller values of w results in concentrating weights on training observations that are closer to X_0 whiles larger values of w produce nonzero weights on more observations among the training observations since there are more observations in the terminal nodes of each tree. CSRF reduces to RF when the *nodesize* is larger than the the size of the training data set N^* because uniform weights will be generated.

1.6 Using random forests in causal inference

There have been a few proposed methods for estimating causal quantities using random forests. One approach is by adjusting the entire random forest framework to specifically estimate causal quantities. Another approach involves directly estimating counterfactuals.

1.6.1 Causal Forest

The estimation of heterogeneous treatment effects using random forests has been studied extensively in the causal inference literature by Wager and Athey (2018). Their forest-based method known as causal forest is made up of causal trees that estimate treatment effect at the leave of the trees. The method is composed of two distinct algorithms that estimate causal

effect based on regression and classification which are double-sample trees and propensity trees. The honesty assumption ensures that, for each training sample, the response is only used to estimate the within-leaf treatment effect or decide where to place the split, but not both. This is intrinsically the foundational assumption of their method, and ensures, under certain conditions, an unbiased estimate of the treatment effect.

For double-sample trees, the honesty assumption is attained by partitioning its training subsample into two of the same size and using one to place the splits and the other for within-leaf treatment estimation. Their second algorithm for constructing honest trees used the treatment assignment (control and treated units) to train classification trees and disregard the response for splitting and used the response for estimating treatment effect. Thus, both treatment and control outcomes were used to estimate the treatment effect.

Additionally, Athey and Wager (2019) implemented causal forests using estimated propensity scores to estimate treatment effects in datasets containing clustering errors. For this approach, they fit two separate regression forests and estimated expected outcome marginalizing over treatment and propensity scores. Out-of-bag predictions were made with the two forests and used them to grow causal forests. In their final implementation strategy, they started by training initial random forest on all features and then train on only those features in the first step with convincing number of splits which enables the forest to make adequate splits on most important features.

1.6.2 Estimation of counterfactuals

The outcome in causal inference can be observed only under one, and not both, treatment conditions. If an individual i is assigned to treatment t, then Y_{ic} cannot be observed and likewise if individual i is assigned to treatment c, then Y_{it} cannot be observed. The response of individual i under the treatment condition it does not receive is known as the *counterfactual*. Since counterfactuals are not observed, they have to be estimated in causal inference settings. However, the estimation of counterfactuals present many difficulties, especially in

observational studies. Some approaches for estimating the counterfactuals were presented in the literature.

Lu et al. (2018) use random forests to estimate the individual treatment effect by directly modeling the response in counterfactual framework. They make two extensions to the virtual twins approach (Foster et al., 2011) which is an extension of random forest. These extensions are counterfactual random forest and counterfactual synthetic random forest. Virtual twins approach first uses random forest to produce individual predictions of outcomes under both treatment conditions for each trial participant by estimating the counterfactual treatment outcome (Lu et al., 2018). The counterfactual RF fits a separate forest to each treatment group CF_1 and CF_0. The counterfactual individual treatment effect are estimated by running each data point through its natural forest as well as its counterfactual forest. For the counterfactual synthetic forest, they replace Breiman RF regression with synthetic forest regression (Ishwaran and Malley, 2014) but fit separate synthetic forest to each treatment group $synCF_1$ and $synCF_0$. The counterfactual individual treatment effect are estimated as done in counterfactual RF.

Another strategy for estimating counterfactuals in the literature is shown in the LOOP Estimator (Wu and Gagnon-Bartsch, 2018). They impute the counterfactuals (potential outcomes) using decision trees and random forests. For the decision tree approach, given a single decision tree, they assign each observation i to a group which is accomplished by applying the decision tree to observation i's covariates. The potential outcome for the treated t_i for each i is imputed by utilizing the average observed outcome of the treated units within the same group without observation i (Wu and Gagnon-Bartsch, 2018). The potential outcome for the control units c_i is estimated in a similar fashion.

For the random forest approach, two random forest were fitted each to the treated units and the control units. Out-of-bag predictions for the i observation were made using trees that do not contain observation i. The counterfactuals(potential outcomes) were imputed using the out-of-bag predictions i (Wu and Gagnon-Bartsch, 2018).

1.7 Organization of dissertation

We have so far explored essential assumptions and theories relevant for estimating the average treatment effect for the treated (ATT) in settings with few treated units and large number of control units. We have extensively delve into the random forest and its extensions related to our proposed mean-weighted case specific random forest (MWCSRF). Chapter 2 presents our proposed mean-weighted case-specific random forest (MWCSRF) method for estimating the average treatment effect for the treated(ATT). Using simulated data and the National Supported Work Demonstration (NSW) dataset, and also the population Survey of Income Dynamics (PSID) and the Current Population Survey(CPS) (LaLonde, 1986), we compare our ATT estimate using our MWCSRF method to the standard random forest (RF) method and causal forest. Chapter 3 highlight the implementation of mean-weighted case specific random forest (MWCSRF) with R package and a guide on how to use it. Chapter 4 comprehensively demonstrated when the underlining SUTVA assumption of the Neyman-Rubin model would or will not hold. We established this using Fisher's lady-tasting tea randomized experiment. Chapter 5 ends the dissertation and dwells on concluding remarks and future work.

Chapter 2

Mean-weighted case specific random forests for estimating causal effects

2.1 Introduction

In this chapter, we focus attention on settings with many control units and very few treated units with the goal of efficiently estimating the average treatment effect for the treated (ATT). Typically, we are able to manipulate the control units from a nonexperimental data source but treated units originate from different data source such as national database.

Machine learning techniques have substantially grown in popularity for estimating causal estimands. Contributing to the literature, Künzel et al. (2018) use deep neural network for Conditional Average Treatment Effect (CATE) estimation. Outcome-adaptive LASSO has been suggested for selecting important covariates for propensity score models to explain confounding bias and preserve statistical efficiency (Shortreed and Ertefaie, 2017). Liu and Yang (2018) propose penalized regression adjusted (elastic net, adaptive LASSO, and Ridge regression) for estimating Average Causal Effect (ACE) in random experiments with many pre-experiment covariates. Hill (2011) recommend the use of Bayesian Additive Regression Trees(BART) to identify the causal effects in nonexperimental situations.

Random forests in particular have been one of the popular machine learning methods used in the literature for causal inference. Athey and Wager (2019) used causal forest to study the heterogeneous treatment effect estimation in observational studies using propensity scores. Suk et al. (2019) also used causal forests in estimating treatment effects in clustered or multilevel observational data using multilevel propensity score matching.

Case specific random forest (CSRF) is a type of random forest introduced by Xu et al. (2016). For this method, different random forests are constructed specific to different test case. This method uses the weighted bootstrap resamples to construct trees and assign larger weights to training cases that are nearest to the test case of interest.

We propose an extension to CSRF called mean-weighted case specific random forest (MWCSRF) to estimate causal quantities under the Neyman-Rubin Causal Model (Splawa-Neyman et al., 1990). Our method follows the same idea of Case Specific Random Forest (CSRF) but weighted-mean bootstrap resamples are use to construct trees. Our method aims to excel in settings with many control units and few treated units, and with the covariate space of the treated units comprising a small subspace of the control units.

In section 2, we discuss causal inference and some estimands that relate to homogeneous and heterogeneous setting. We describe in details the Mean Weight Case Specific Random Forest (MWCSRF) in section 3. We present simulation results to illustrate our propose method in relation to homogeneous and heterogeneous settings in section 4. In section 5, we also present results on real data set to demonstrate our method. We conclude in section 6.

Remark: Synthetic control (SC) methods may also be useful in the setting where the goal is to estimate the causal effects given a large number of control units but very few treated units. SC uses the control group to construct counterfactuals of the treated group—estimates of the treated units if the treatment had not been applied. Thus, the counterfactual of the treated group would be predicted by the control group in addition to other possible covariates in the control group. This method was first introduced by Abadie and Gardeazabal (2003) to learn the causal effect of treatment that affects a single aggregate unit that is observed during

pre-and-post-treatment periods. They use the SC method to study the effect of terrorism in the Basque Country on its GDP per capita income. They use other regions in Spain where terrorism does not occur but their GDP was observed before and after the terrorist activities in Basque Country and use them to form the control group.

2.1.1 Motivating Data Example: The National Supported Work Demonstration

To demonstrate the MWCSRF methology, we consider the National Supported Work Demonstration (NSW) dataset which was a temporary employment program designed to assist disadvantaged workers lacking basic job skills move into the labor market by giving them work experience and counseling in a sheltered environment. This program was operated by Manpower Demonstration Research Corporation (MDRC) in the mid-1970s in ten sites across the United States which lasted for 9–18 months depending on the site and target group. Of note, assessment of the efficacy of the program was originally performed using a randomized experiment—eligible individuals were randomly assigned to participate in the program (Kemper and Long, 1981; Kemper et al., 1981).

LaLonde (1986) used this dataset to determine whether econometric methods using non-randomized studies can replicate results from an experiment. Specifically, LaLonde (1986) used nonexperimental data from a large national survey to collect non-random control units (comparison group) which have similar characteristics as those treated units from the training program. The control units (comparison group) were obtained from Panel Study of Income Dynamics (PSID) and Westat's Matched Current Population Survey - Social Security Administration File (CPS-SSA). Using linear regression, LaLonde (1986) was unable to replicate the experimental benchmarks.

This dataset has been further analyzed by Dehejia and Wahba (1999), who used propensity score matching, and Smith and Todd (2005), who used cross-sectional and longitudinal propensity score matching method. Notably, the ability for methods to replicate the exper-

imental benchmark varied substantially depending on the method used to analyze the data. A significant reason for this variability is due to differing distribution of the covariates under treated and control groups. Generally, the covariate space of the treated units is a small subspace of that of the control units.

2.2 Preliminaries

2.2.1 Overview of Causal Inference

Suppose there are N units, numbered 1 through N. Each unit has a response Y_i and a p-dimensional covariate $X_i = (X_{i1}, X_{i2}, \ldots, X_{ip})$. We denote N_t to be total number of treated units, and N_c to be total number of control units.

The framework for causal inference that we employ in this paper is the Neyman-Rubin causal model (NRCM) of response (Holland, 1986; Rubin, 1974a; Splawa-Neyman et al., 1990):

$$Y_i = Y_{i1}T_i + Y_{i0}(1 - T_i) \tag{2.1}$$

where Y_{i1} is the potential outcome for unit i if the unit receives treatment, Y_{i0} is the potential outcome for unit i in the control regime, and T_i is the treatment indicator variable. The potential outcome of unit i given treatment condition t, denoted Y_{it}, is the hypothetical outcome of that unit had that unit received treatment t. The treatment effect for unit i is $\tau_i = Y_{i1} - Y_{i0}$. Since we cannot observe both outcomes at the same time, causal inference is largely referred to in the literature as a missing data problem. Finally, NRCM assumes the stable unit treatment value assumption (SUTVA) which is a foundational assumption in causal inference which assumes that every unit has two potential outcomes, one under treatment and under control, and the observed outcome depends only on the treatment received by that unit.

Consider an experiment testing the efficacy of a blood pressure medication on blood

pressure after one month of exposure to the drug. The causal effect for an individual i would be difference in the systolic blood pressures in the presence and in the absence of the drug (Hill and Stuart, 2015). Note that both outcomes cannot be observed at the same time. For this example, we have

$$T_i = \begin{cases} 1, & \text{if patient } i \text{ receives the drug,} \\ 0, & \text{if patient } i \text{ does not receive the drug.} \end{cases} \tag{2.2}$$

Then, in context of (2.1), Y_{i1} is the systolic blood pressure that would be measured for unit i during the exposure to the drug and Y_{i0} is systolic blood pressure measured during the same time for unit i in absence of exposure to the drug. The treatment effect for patient i is $\tau_i = Y_{i1} - Y_{i0}$.

In causal inference there are two distinct sources of data: randomized experiments and observational studies. We first consider randomized experiments. In this type of study, both treatment and control groups are obtained from the same population. Thus, in the context of the blood pressure example, both subjects who received the drug (treatment group) and those who did not receive the drug (control group) would be obtained from the same population. Hence, the average treatment effect (ATE) for taken the drug is $\tau = E[Y_{i1} - Y_{i0}]$. Since randomization implies that $(Y_{i1}, Y_{i0} \perp\!\!\!\perp T_i)$ (where $\perp\!\!\!\perp$ denotes independence) (Dawid, 1979), we have for $t \in \{0, 1\}$ that

$$E(Y_{it}|T_i = 1) = E(Y_{it}|T_i = 0) = E(Y_i|T_i = t) \tag{2.3}$$

and

$$\begin{aligned} \tau &= E(Y_{it}|T_i = 1) - E(Y_{it}|T_i = 0) \\ &= E(Y_i|T_i = 1) - E(Y_i|T_i = 0). \end{aligned} \tag{2.4}$$

Thus, the ATE can be directly estimated using all available treated and control observations.

However, for observational studies, we cannot estimate the average treatment effect in the same since the treatment and control groups are drawn from different populations. For our blood pressure example, the treatment group is the population who received the drug, but the control group (comparison group) would be obtained from some different population. Thus, the treatment effect under this setting would be the average treatment effect for the treated (ATT) population. This is the quantity

$$\tau|T = 1 = E(Y_{i1}|T_i = 1) - E(Y_{i0}|T_i = 1) \tag{2.5}$$

which cannot be directly estimated since we do not observed Y_{i0} for the treated. We can proceed by assuming that selection for treatment depends on observable covariates X.

Following Rosenbaum and Rubin (1983), we assume that conditional on X, treatment assignment is unconfounded. That is, $\{Y_{i0}, Y_{i1} \perp\!\!\!\perp T_i\}|X$ and $0 \leq Pr(T = 1|X) < 1$. These conditions are also jointly known as strongly ignorable treatment assignment (Imbens and Rubin, 2015). Then, by Rubin (1974a, 1977) we obtain

$$E(Y_{it}|X_i, T_i = 1) = E(Y_{it}|X_i, T_i = 0) = E(Y_i|X_i, T_i = t) \tag{2.6}$$

for $t \in \{0, 1\}$. By conditioning on the observables, X_i, treatment and control groups are balanced. This permits us to estimate the average treatment effect for the treated (ATT) as

$$\tau|(T = 1) = E[E(Y_i|X_i, T_i = 1) - E(Y_i|X_i, T_i = 0)|T_i = 1], \tag{2.7}$$

where the outer expectation is over the distribution of $X_i|(T_i = 1)$ which is the distribution of covariates X in the treated group. The ATE and ATT are the main quantities of interest estimated in causal inference to measure causal effects.

2.2.2 Random Forest (RF)

The random forest is a nonparametric machine learning technique for classification and regression prediction problems which was first introduced by Breiman (2001). Random forests have become a popular machine learning technique due to their prediction efficiency and flexibility of being combined with other methods such as quantile regression forest (Meinshausen, 2006), generalized random forests(GRF) (Athey et al., 2019), random survival forests (Ishwaran et al., 2008), case specific random forests (CSRF) (Xu et al., 2016), and orthogonal random forests (Oprescu et al., 2018).

The standard random forest methodology fits a decision tree to different bootstrap samples. During the tree growing process, a random sample of $m < p$ predictors are considered in each step which lead to different trees from each sample and predictions from each tree are then averaged. We illustrate the method with a succinct description of random forest that fits a model for predicting a regression type problem for continuous or binary response Y and a set of predictors X, where X is p-dimensional. Consider the training data set $\mathbf{D} = \{D_i = (X_i, Y_i), i = 1, 2, ...N^*\}$ which consists of N^* cases. We use these N^* cases to develop a model for predicting response at a particular vector X_0 of predictors.

The algorithm for random forest for regression type problem is given below:

1. For $j = 1, ..., B$, draw independent and uniform bootstrap samples $\mathbf{D}^* = \{D_i^* = i = 1, 2, ...N^*\}$ of size N^* from the training data set.

2. Grow regression and classification trees R_j and C_j respectively to the bootstrapped data set $\mathbf{D}^*$ in a single node K by recursively repeating the steps below until minimum pre-specified $nodesize = w$ is obtained, where $nodesize$ is the minimum number of observations in the terminal node.

    i. Select $mtry$ predictors at random from all the p predictors and split on the best predictor variable.

ii. Split the nodes into two subnodes K_1 and K_2. For regression choose among all such splits the one that minimizes the Sum of squares error (SSE)

$$\sum_{b=1}^{2} \sum_{i \in K_b} (Y_i^* - \bar{Y}_b)^2, \tag{2.8}$$

where b=1,2 and $\bar{Y}_b$ is the mean response of the training observations within the subnode b.

For classification, select the split that minimizes the Gini Impurity/Index

$$G = \sum_{b=1}^{B} \hat{p}_{mb}(1 - \hat{p}_{mb}), \tag{2.9}$$

where G is the measure of total variance across the B classes, and $\hat{p}_{mb}$ is the proportion of training observations in the m^{th} region that are from the b^{th} class.

3. Output the trees $\{R_j\}_1^B$ and $\{C_j\}_1^B$ to form a random forest for regression and classification respectively.

To make a prediction for a continuous variable at a given independent variable value X_0, we average the prediction from each tree $\hat{R}_j(X_0)$, that is

$$\hat{R}_j(X_0) = \frac{1}{B} \sum_{j=1}^{B} \hat{R}_j(X_0), \tag{2.10}$$

Likewise, to obtain prediction for X_0 for a binary response variable, we first obtain the class prediction of the j^{th} tree $\hat{C}_j(X_0)$ and then obtain the majority vote for

$$\hat{C}_j(X_0) = \text{majority votes}\{\hat{C}_j(X_0)\}_1^B. \tag{2.11}$$

The concept of majority vote is such that each random forest predicts different outcome for the same test feature. Then, taken each predicted outcome into account, votes will be calculated. Thus, majority vote for random forest classifier is the final class label that the

classification models predict most frequently. This process is repeated using the same trees to obtain predictions at different values of independent variables.

Also, note that for classification, other split rules such as classification error rate and cross-entropy can also be considered to evaluate the quality of the split during tree construction. The classification error rate is particularly desired split rule if the aim is prediction accuracy.

2.2.3 Case Specific Random Forests (CSRF)

The case specific random forest (CSRF) is a version of random forest introduced by Xu et al. (2016) that grows decision trees specific to a test case. The standard random forest described in Section 2.2.2 makes all predictions from the same set of trees, but CSRF is built specific for a conditional value of the covariates, say X_0, where prediction is desired. The construction of CSRF is essentially the same steps outlined in Section 2.2.2, but Step 1 in the RF algorithm is replaced with weighted bootstrap samples. In this setting, the training data cases that are closer to X_0 are allocated higher weights than those farther from X_0. The significant step here is constructing a proximity measure on the subset of predictors associated with the response. Xu et al. (2016) used an initial bootstrap aggregation (bagging) of trees (Breiman, 2001) to construct the measure of proximity and the corresponding resampling probability weights.

The algorithm for CSRF replacing Step 1 in the algorithm from Section 2.2.2 is given below:

1. Grow B decision trees following the RF algorithm in Section 2.2.2 with $mtry = p$ and $nodesize = w$. Note that this RF reduces to bagging since $mtry = p$.

2. Define the training case $\mathbf{D_i} = (X_i, Y_i)$ and let E_i be the number of B trees that contain both X_0 and X_i in the same terminal node.

3. For $i = 1, 2, \ldots, N^*$, define Z_i be the resampling probability weights of D_i relative to X_0 as

$$Z_i = \frac{E_i}{\sum_{j=1}^{N^*} E_j}. \tag{2.12}$$

Note that $(Z_1, Z_2..., Z_{N^*})$ defines a probability distribution on the observations in D (the training data set). The nodesize w determines how far apart the weights Z_i are ordered on the training observations $D = \{D_1, D_2, ..., D_{N^*}\}$.

Smaller values of w results in concentrating weights on training observations that are closer to X_0 whiles larger values of w produce nonzero weights on more observations among the training observations since there are more observations in the terminal nodes of each tree. CSRF reduces to RF when the *nodesize* is larger than the the size of the training data set N^* because uniform weights will be generated.

2.2.4 Estimating causal effects using random forests

There have been a few proposed methods for estimating causal quantities using random forests. We now discuss several of these methods.

The estimation of heterogeneous treatment effect using random forests has been studied by Wager and Athey (2018). Their forest-based method known as causal forest is made up of causal trees that estimate a treatment effect at the leaves of the trees. Their approach satisfies an honesty assumption which ensures that, for each training sample, the response is only used to estimate the within-leaf treatment effect or decide where to place the split but not both. They create a double-sample tree, and ensure honesty by partitioning its training subsample into two of the same sizes and using one to place the splits and the other for within-leaf treatment estimation. Both treatment and control outcomes were used to estimate the treatment effect.

Athey and Wager (2019) also consider using random forests to estimate propensity scores, and using the estimated propensity scores to estimate treatment effects. This approach is especially useful for datasets containing clustering errors. For this approach, they fit two separate regression forests and estimate expected outcome marginalizing over treatment and propensity scores.

Our proposed approach is to use mean-weighted case specific random forests trained on the control units to estimate the prognostic score for all treated units. This will allow us to obtain estimates of ATT by taking the average of the differences between the treated units and their predicted prognostic score. The prognostic score $\Phi(x_i) = E(Y_i|\mathbf{X_i} = x_i, T_i = 0)$, formalized by Hansen (2008), is the predicted outcome under the control condition. The prognostic score also summarizes how the covariates associate with the response.

We estimate the prognostic scores by fitting a model with the control units and then use that model to obtain predictions for the outcome under the control condition for all treated individuals. The outcome type—such as continuous, binary, ore categorical—determines the relevant regression model used to estimate the prognostic scores. The advantage of using prognostic score in our is that biased will be reduced since we estimate the prognostic score using only the control group. That is, we are avoiding bias by ensuring that the estimated prognostic scores are "out-of-bag" estimates from the random forest model.

2.3 Mean Weighted Case Specific Random Forests (MWCSRF)

The algorithm described in Section 2.2.2 uses one forest of trees to make all predictions. The CSRF method described in Section 2.2.3 is designed to grow different RF to specific test cases, and hence, will can only handle a single treated unit at a time. Our proposed method—mean-weighted case specific random forests (MWCSRF)—is tailored to accommodate multiple treated units.

There are several benefits to mean weighting. First, the oversampling of control units that are in proximity to treated units acts as a "data trimming" step; control units that are given zero weight in the bootstrap step are removed completely from the estimation of the treated counterfactual. Second, when control units are concentrated together, mean weighting may reduce noise in selecting good bootstrap sampling weights, preventing model overfitting to possible outliers. Third, when parallelization is not available, mean-weighting offers a considerable reduction in computational cost compared to CSRF.

A slight drawback to mean weighting is that it does not allow for efficient construction of double-sampled trees. That is, it is not feasible in practice to implement MWCSRF to satisfy the honesty assumption advocated by Wager and Athey (2018). However, as we will show through our simulation results, when treated and control groups are not drawn from the same population, mean weighting is a much more effective tool than ensuring honesty when estimating treatment effects.

The MWCSRF algorithm follows the steps outlined in the standard random forest construction except that the bootstrapped samples are replaced with mean weighted bootstrap samples.

The algorithm for mean weight case specific random forest (MWCSRF)

1. Grow B decision trees following the RF algorithm in Section 2.2.2 with $mtry = p$ and $nodesize = w$. Note that this RF reduces to bagging since $mtry = p$.

2. Define the training case to be the observed potential outcomes for the control units and set of covariates D_c such that $\mathbf{D_c} = (X_c, Y_c)$. Define the testing case to be the observed potential outcome for the treated D_t and set of covariates such that $\mathbf{D_t} = (X_t, Y_t)$ and let $E_c^\star$ be the number of B trees that contain both X_t and X_c in the same terminal node.

3. For $c = 1, 2, \ldots, N_{tr}$ and $t = 1, 2, \ldots, N_{te}$, define Z_{ct} to be the resampling probability weights of D_c relative to X_t as

$$Z_{tc} = \frac{E^*_{tc}}{\sum_{j=1}^{N_c} E^*_{jc}}. \tag{2.13}$$

4. For $c = 1, 2, \ldots, N_{tr}$, $t = 1, 2, \ldots, N_{te}$, define $Z_{Ave,c}$ to be the average of the resampling probability weights Z_{tc} obtained in Step 3

$$Z_{Ave,c} = \frac{1}{N_{te}} \sum_{j=1}^{N_{te}} Z_{jc} \tag{2.14}$$

The MWCSRF reduces to CSRF when there is a single treated unit, that is $t = 1$

2.3.1 Estimating ATT with MWCSRF

To estimate the ATT with the MWCSRF, we obtain the mean weights $Z_{Ave,c}$ using the observed potential ouctcomes of the control units Y_c and their covariates X_i as the training cases following the steps in the MWCSRF algorithm. These mean weights $Z_{Ave,c}$ are used to weight bootstrap samples. We then use RF with these bootstrap sample weights to obtain predictions of the prognostic score $\hat{\Phi}$ for all treated units.

The average treatment effect for the treated (ATT) is obtained by taking the average across all treated units of the difference between the observed response and the estimated prognostic score.

$$\widehat{ATT} = \frac{1}{N_t} \sum_{i=1}^{N_t} Y_{i1} - \hat{\Phi}(X_i) \tag{2.15}$$

2.4 Simulation Study

In this section we examined the performance of MWCSRF in relation to RF, Genetic Matching, CSRF, and Causal forest for estimating the causal effect when there are large number of control units and few treated units with the treated units condensed at one point of the covariate space. We compare three simulation strategies. First, we intend to know how the

results changes across different models when we vary p and maintained the same sample sizes for both treatment and control. Secondly, we also want to scrutinize the model performance when the sample size increases, but the fraction of treated units relative to the sample sizes remained unchanged. Finally, we are interested in how the results differ across the various models when the number of control units increase, but the treated units remained the same.

We generate the covariates based on multivariate normal distribution, and do so independently for both control and treated units. For control units, $X_c \sim MultNorm(\mu_c, \Sigma)$ where the mean $\mu_c = \mathbf{0}$ and covariance matrix values $\Sigma_{ij} = M^{|i-j|}, 1 \leq i, j \leq p$ (where we define $\Sigma_{ij} = 1$ if $M = 0$ and $i = j$). For the treated units, the covariates were generated as $X_t \sim MultNorm(\mu_t, \frac{1}{16}\Sigma)$ where $\mu_t = -\mathbf{0.5}$. The variance covariance matrix Σ was varied for various choices of values M in order to induce independent and dependency in the covariance matrix. We considered these set of values for M (0, 0.05,0.25,0.5 0.9,-0.05,-0.25, -0.5 -0.9) for constructing the covariance matrix. The choice of these M values determines the strength of correlation among the covariates. For instance, if the value of M is 0, we expect all the entries in the covariance matrix to be zero except one's on the diagonal, and the covariates are independent. Likewise, if the value of M is 0.9, there would be strong dependency in the covariance matrix and the covariates will be highly correlated among themselves. We intend to investigate how our method will perform compared to other baseline methods given the strength of correlation among the covariates. We shrink the covariance matrix for the treated units to simulate the case where treated units are concentrated in one area of the covariate space.

The sample size or total number of units N, the total control units N_c, the total treated units N_t, and the number of predictors p were varied for each simulation and replicated 5000 times. We considered three different samples in the following order: (N, N_c, N_t), that is (410,400 10), (1025,1000, 25) and (2050, 2000, 50) for each simulation scenarios. We vary the number of predictors p as 5, 10, and 50 for each simulation. Even though we vary p for each simulation, the only covariates that have direct relationship with the response for each

simulation are X_1, X_2, and X_3. The choice of these parameters is to establish whether our method can perform well in instances with small sample size and large predictors and vice versa.

We generate the potential outcomes for Y_0 using the following model:

$$Y_0 = \frac{\log(|X_{i1}| + |X_{i2}|)}{0.5 + |X_{i3}|} + \epsilon, \ \epsilon \sim N(0, 1). \tag{2.16}$$

We choose this model in order to avoid a linear relationship between the response and the covariates. This is as a result of the fact that, if linear relationship exist between the response and the covariates, ordinary least square regression will be the most ideal candidate to estimate our quantity of interest (ATT).

We consider both homogeneous and heterogeneous treatment effects to determine the effectiveness of our method compare to other baseline methods when the same treatment has the same effect on different individuals and when it affects individuals differently. In the case where we are assuming homogeneous treatment effects, that is, the same treatment affects all units the same way, we set the treated potential outcome to

$$Y_1 = Y_0 + 4. \tag{2.17}$$

In the heterogeneous case, that is, when treatment affects individuals differently, we have

$$Y_1 = Y_0 + 3 + 0.5X_{i1} + \frac{1}{3}X_{i2}. \tag{2.18}$$

We considered two data simulation scenarios, that is heterogeneous data and homogeneous data in the causal inference framework. For the homogeneous data (2.16 and 2.17), the true ATE is 4. For the heterogeneous data (2.16 and 2.18), the true ATT will vary across simulations and will need to be computed.

We compare methods using the empirical bias (also referred to as expected bias), the

empirical standard error and the theoretical standard error. The empirical bias is

$$Bias(\widehat{ATT}) = \frac{1}{s}\sum_{i=1}^{s}\widehat{ATT}_i - ATT_i \tag{2.19}$$

and the empirical standard error is

$$EmpiricalSE = \sqrt{\frac{1}{s}\sum_{i=1}^{s}(\widehat{ATT}_i - ATT_i)^2} \tag{2.20}$$

where $\widehat{ATT}_i$ is the ATT estimate for simulation i, ATT is the true ATT for simulation i, and s is the number of simulations performed.

The theoretical standard error for a given simulation or study is estimated by

$$TheoreticalSE = \frac{1}{\sqrt{N_{te}}}\sqrt{\frac{1}{N_{te}-1}\sum_{j=1}^{N_{te}}(\hat{Y}_j - Y_j)^2} \tag{2.21}$$

where $\hat{Y}_j$ is the estimated response for unit j in the test set, Y_j is the actual response for unit j in the test set, and N_{te} is the number of units in the test set. In our simulation study, we average this across simulations.

In Figure 2.1, we demonstrate graphically the behavior of the covariate space for the treated and control units as described in the introduction of this paper. The treated units

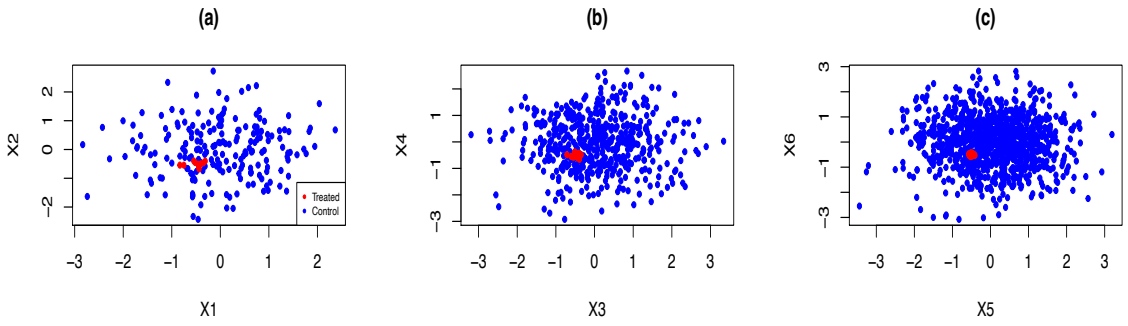

Figure 2.1: *Two dimensional representation of the covariates of each simulation scenario. For figure (a), N=210, N_c =200, and N_t =10. For figure (b) and (c), N=525, N_c =500, and N_t =25 and N=1025, N_c =1000, and N_t =25 respectively.*

are concentrated in one area of the covariate space for both homogeneous and heterogeneous simulation scenarios.

2.4.1 Results of homogeneous and heterogeneous data simulation

The result of the homogeneous data simulation scenario to estimate the average treatment effect for the treated (ATT) with the true ATT known to be 4 is presented in Table A.1, Table A.2, and Table A.3 in Appendix A. These tables compare changes in our results when the control and the treated units do not change, and p is varied.

We compare our mean weight case specific random forest (MWCSRF) method with standard random forest (RF), case specific random forest(CSRF), genetic matching, and causal forest model. Our primary focus is comparing our MWCSRF and CSRF since we made an extension to the CSRF. The first results we have considered is the case where we vary the number of covariates(p) whiles maintaining the same number of treated and control units.

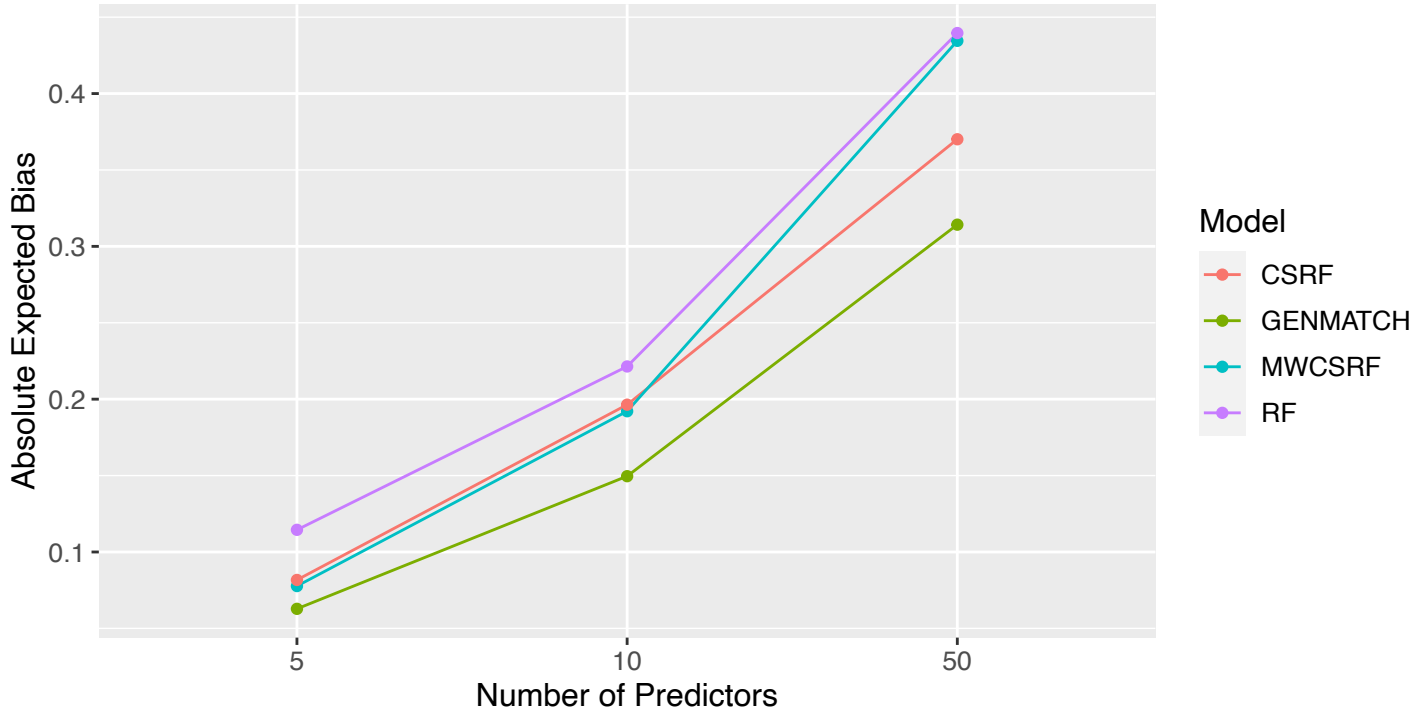

Figure 2.2: *Plot of empirical bias when varying the number of predictors with sample size 210 and M=0*

It can be inferred from Table A.1, Table A.2, and Table A.3 that our MWCSRF method appears to be more efficient for estimating treatment effect compared to other models for all the simulation scenarios considered except when there is no dependency among the predictors and number of predictors is small. When there is very weak or no dependence among the predictors, genetic matching estimates appears to be slightly closer to the true ATE compared to MWCSRF. For instance, from Table A.1 and with $M = 0$ and true ATE is 4, the genetic matching estimate of the ATE is 3.9372 with associated absolute bias and empirical standard error as 0.0628 and 0.6127, respectively. The MWCSRF ATE estimate is 3.9223 with absolute expected bias and empirical standard errors as 0.0777 and 0.3963, respectively, for the exact value of M. We observe that, even though the genetic matching estimate is slightly closer to the true ATE than the MWCSRF estimate, its empirical standard error is almost twice that of MWCSRF. From Figure 2.2, it's clear that genetic matching in terms of bias performs better than the other models. The MWCSRF model shows improvement in the estimates compared to genetic matching when there is moderate to high dependency among the covariates. There is no substantial difference between the MWCSRF estimates and that of RF and CSRF, but in terms of bias and empirical standard error, the MWCSRF model appears to have performed a bit better than RF and CSRF. This trend in performance is observed across the three simulations considered, where p is varied.

Table A.4, Table A.5, and Table A.6 in Appendix A contained the results obtained from the heterogeneous data simulation whiles altering p and preserving the sample size for treated and control units. The results show the same trend of performance as observed in the homogeneous case. The difference in the performance of MWCSRF relative to other models follows the same trend as observed in the homogeneous case.

The results obtained from homogeneous and heterogeneous models when the sample size is increased, but the fraction of treated units relative to the sample size is the same, are presented in the tables Table A.7, Table A.8, and Table A.9 in Appendix A.

The results from Table A.7, Table A.8, and Table A.9 indicate the MWCSRF model

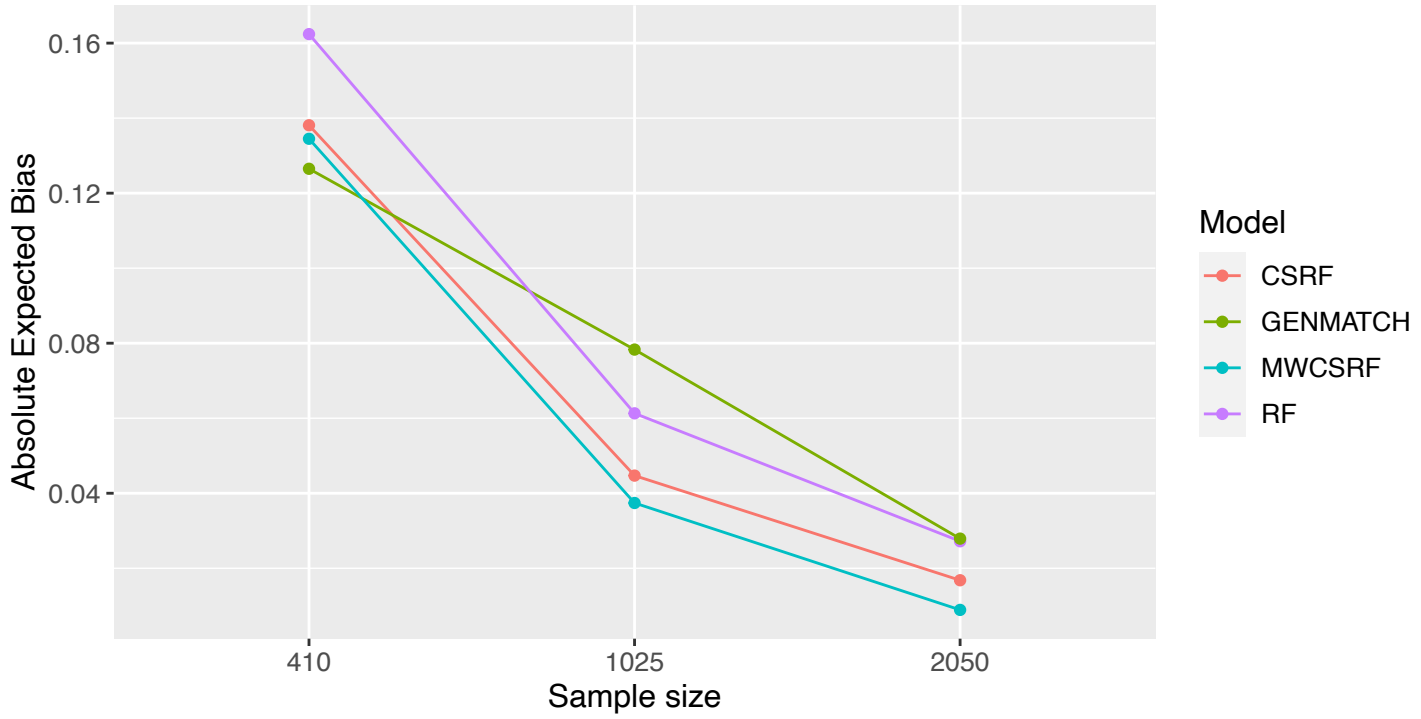

Figure 2.3: *Plot of empirical bias when varying sample size with P=10 and M=0*

performed better than the other models. From Figure 2.3, we observed that the MWCSRF model, in terms of bias, outperformed all other models.

The outcomes achieved from the heterogeneous model under the simulation where the sample size systematically increased, and the number of the treated unit chosen accordingly to attain the same fraction are in Table A.10, Table A.11, and Table A.12 in the Appendix A. The results across all three simulation setups suggest a close disparity among MWCSRF, RF, and CSRF estimates. Genetic matching estimates are generally low with high bias and empirical standard error. In terms of bias, MWCSRF comparatively performs better than the other models.

The results obtained from the heterogeneous model when the number of the control unit increased, and the number of treated units preserved are in Table A.17, Table A.18, Table A.19, and Table A.20.

The results from the heterogeneous model, whiles increasing the number of control units and retaining the number of treated units, suggest the MWCSRF model performed better

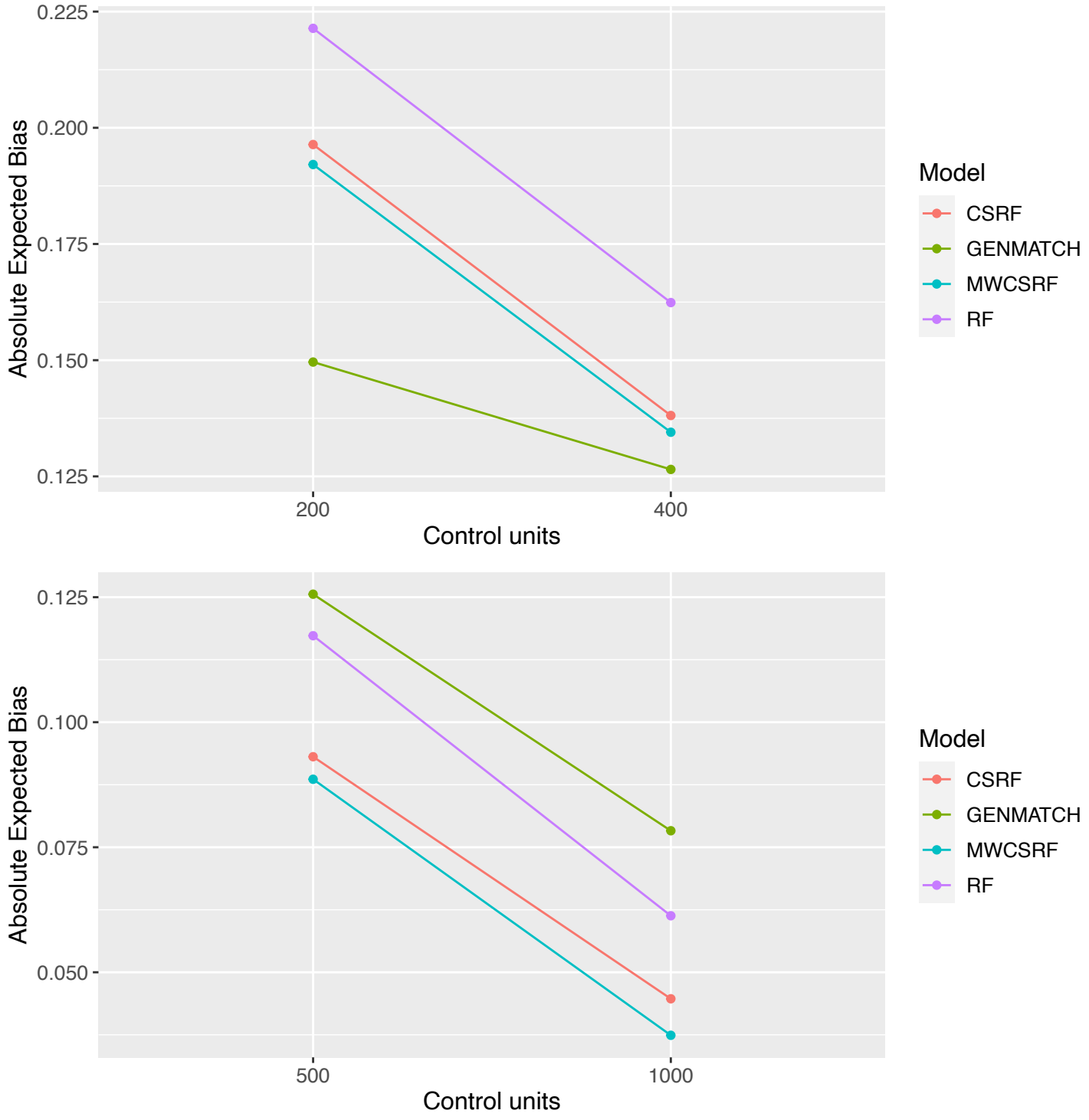

Figure 2.4: *Plot of empirical bias when varying the number control units with P=10 and M=0*

than RF and CSRF except for genetic matching when the control units increased from 200 to 400. From Figure 2.4, in terms of bias, the genetic matching model performs better when control units increased from 200 to 400, but when the control units increased from 500 to 1000, MWCSRF had better performance than other models. These results are most revealing when the number of the treated unit is small and the number of the control unit is large. For instance, in Table A.20 where N =1025, N_c = 1000, and N_t=25, there is a consistent performance of MWCSRF compared to other models in terms of ATT estimate, bias, and empirical standard error for all categories of dependency considered. These results have lent credence to our setting, where we expect the MWCSRF to perform favorably when there are many control units and small treated units. The results from the homogeneous model in Tables A.13–A.16 exhibit the same performance trend as observed in the heterogeneous case.

Generally, causal forest estimates across all simulation scenarios appear to have large bias. Ostensibly, these biases stemmed from the inability of causal forest to accurately identify the covariate space when treated units do not randomly distribute but concentrated at one point.

Results from Table 2.1 show that, given an unparallelized implementation of CSRF, the computational cost of CSRF as the sample size increases in most cases is almost five times that of MWCSRF. That is, though the ATT estimates, bias, and empirical standard errors are almost at par, the run time for CSRF is computationally arduous. For instance, for a sample size of 1050, the number of control units 1000, and treated units 50 with an actual ATT estimate of 2.560, the ATT estimate for MWCSRF is 2.5151, expected bias (-0.0452), and empirical standard error (0.3684). For the CSRF model, the ATT estimate is 2.5067, expected bias(-0.0536), and empirical standard error(0.3695). The run time for MWCSRF estimates was 380.420 seconds, whiles that of CSRF was 2579.999 seconds. Computationally CSRF model will cost six times the time required for the MWCSRF model for these estimates. Thus, the MWCSRF model provides computationally efficient cost reduction compared to

Table 2.1: *Summary of estimates with actual ATT unknown and run time for M = 0, P = 5 and 10, N = 1010,1025, and 1050, where T.SE is the Theoretical Standard Error.*

Model	M	Size(n)	P	Cont	Trt	ATT	Estimates	Expected Bias	Empirical SE	T.SE	Runtime(Sec)
Mean weight	0	1010	5	1000	10	2.6124	2.6213	0.0089	0.6959	0.6512	199.468
Random Forest	0	1010	5	1000	10	2.6124	2.6000	-0.0125	0.6903	0.6495	97.205
CSRF	0	1010	5	1000	10	2.6124	2.6114	-0.0010	0.6881	0.6524	492.861
Mean weight	0	1025	5	1000	25	2.5925	2.5911	-0.0014	0.4482	0.3990	204.278
Random Forest	0	1025	5	1000	25	2.5925	2.5732	-0.0193	0.4453	0.3987	104.133
CSRF	0	1025	5	1000	25	2.5925	2.5905	-0.0020	0.4470	0.3995	1045.470
Mean weight	0	1050	5	1000	50	2.5728	2.5462	-0.0266	0.3011	0.2871	350.473
Random Forest	0	1050	5	1000	50	2.5728	2.5281	-0.0447	0.2995	0.2866	221.624
CSRF	0	1050	5	1000	50	2.5728	2.5418	-0.0310	0.3016	0.2876	2106.153
Mean weight	0	1010	10	1000	10	2.6798	2.7171	0.0374	0.6278	0.6005	280.962
Random Forest	0	1010	10	1000	10	2.6798	2.6987	0.0190	0.6232	0.6007	209.865
CSRF	0	1010	10	1000	10	2.6798	2.7092	0.0294	0.6280	0.6009	653.887
Mean weight	0	1025	10	1000	25	2.5870	2.5922	0.0053	0.5025	0.3896	334.565
Random Forest	0	1025	10	1000	25	2.5870	2.5699	-0.0171	0.5041	0.3892	232.002
CSRF	0	1025	10	1000	25	2.5870	2.5808	-0.0062	0.5021	0.3895	1347.677
Mean weight	0	1050	10	1000	50	2.5603	2.5151	-0.0452	0.3684	0.2883	380.420
Random Forest	0	1050	10	1000	50	2.5603	2.4895	-0.0709	0.3668	0.2879	235.933
CSRF	0	1050	10	1000	50	2.5603	2.5067	-0.0536	0.3695	0.2887	2579.999

CSRF when parallelization is not available.

2.4.2 Randomization of treated units in the covariate space

The performance of causal forest in our setting, where treated units condensed in one area of the covariate space, has been generally poor. The estimates have been highly biased and inconsistent for all the simulation scenarios considered for both homogeneous and heterogeneous models. The causal forest model has not accurately identified the covariate space when treated units concentrate in one area of the covariate space, and hence, creates bias and discrepancy in its estimates. We now consider a simulation scenario where treated units are randomly distributed in the covariate space to ascertain whether the causal forest is ideal for estimating the treatment effect in homogeneous and heterogeneous data.

Figure 2.5 illustrates the random distribution of treated units in the covariate space. The results obtained are presented in Tables A.21–A.24 in Appendix A. The results show that causal forests do indeed reduce bias for both homogeneous and heterogeneous models under this setting, especially when covariates are negatively correlated.

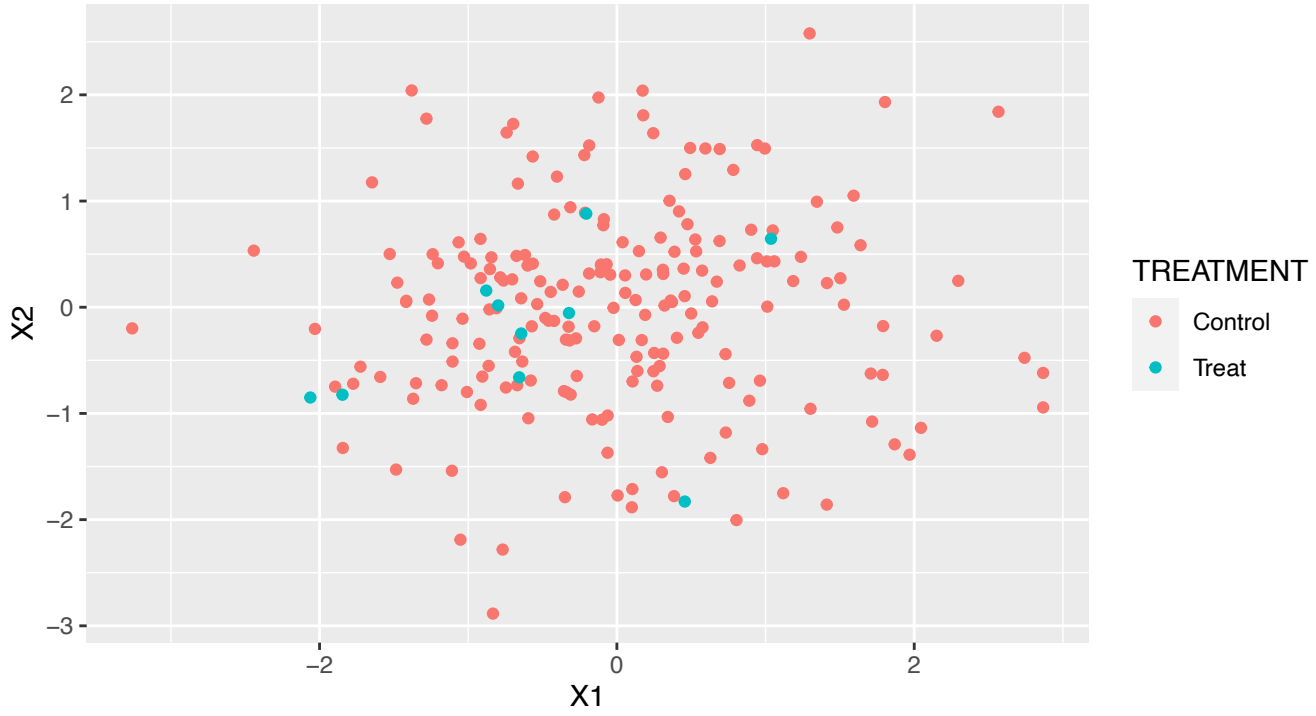

Figure 2.5: *A random distribution of treated units in the covariate space*

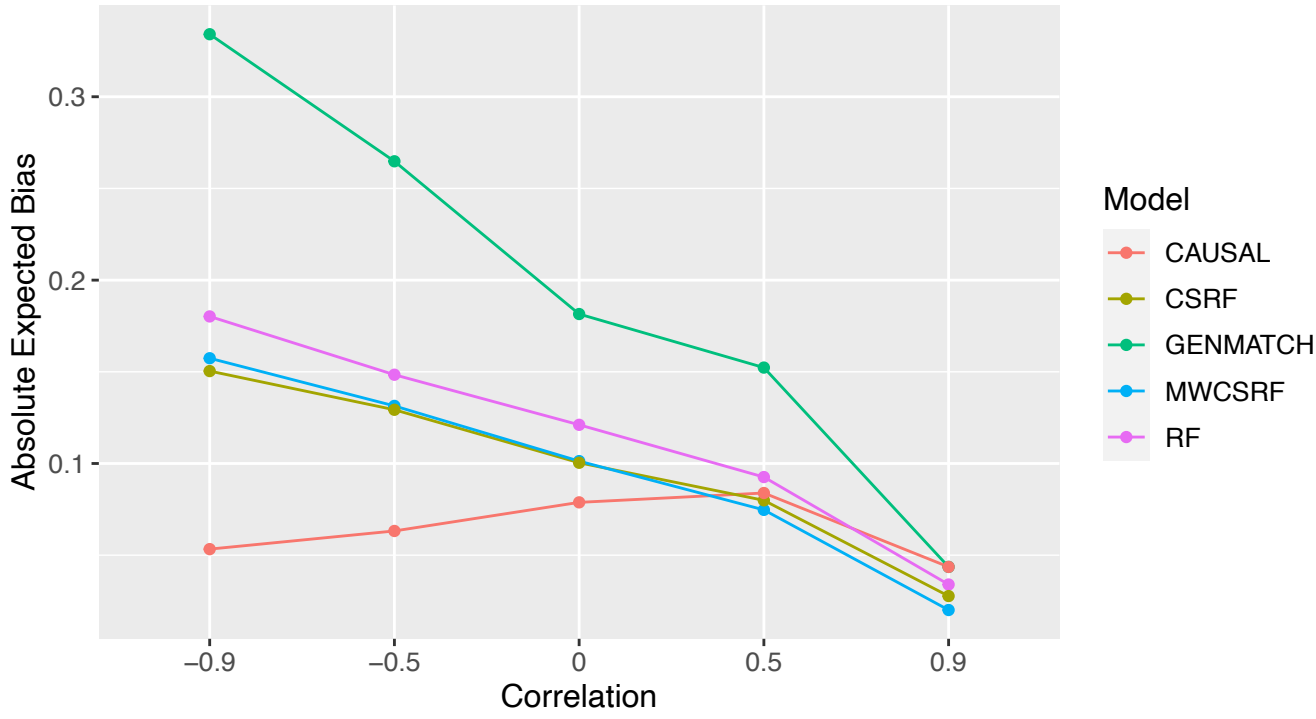

Figure 2.6: *Plot of empirical bias when varying correlation for a simulation with the random distribution of treated units in the covariate space.*

2.5 Results of real data illustration

We now apply the MWCSRF method on the dataset (LaLonde, 1986) described in Section 2.1.1. The treated groups were drawn from the NSW study, but the comparison groups(control groups) are constructed by LaLonde (1986) from a non-experimental dataset, the Population Survey of Income Dynamics (PSID) and the Current Population Survey (CPS). The treated group has 185 observations, and the comparison groups, PSID and CPS, have 2490 and 15,992 observations, respectively, with ten variables.

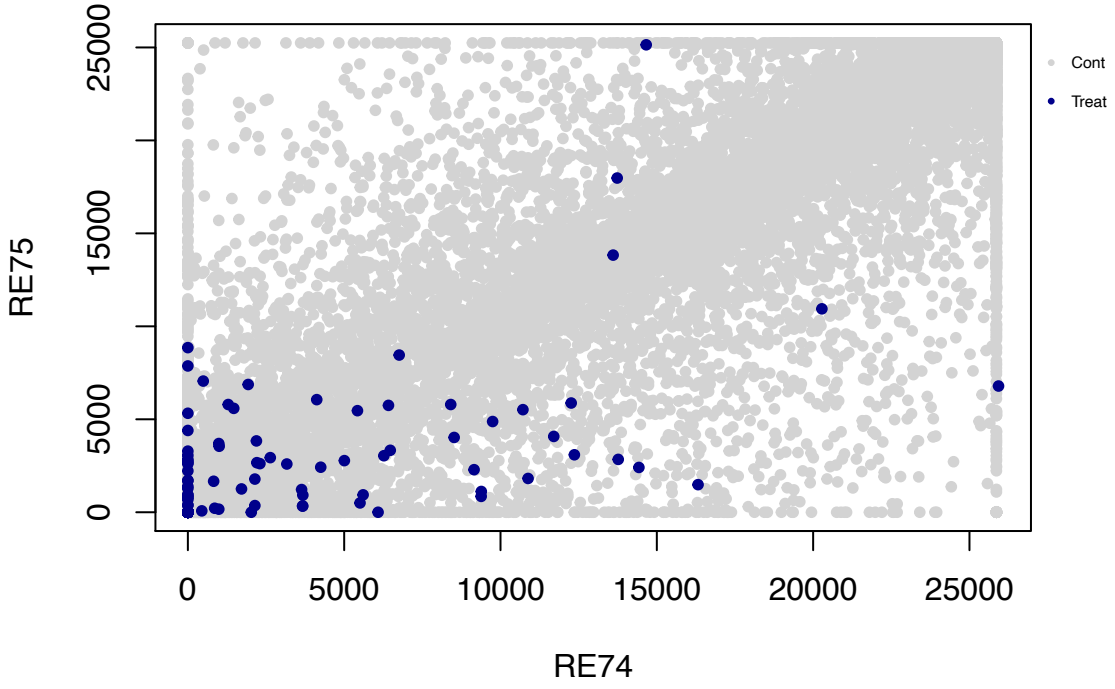

Figure 2.7: *The covariate space of earnings in 1974 against earnings in 1975 for LaLonde's NSW dataset.*

The variables are treatment indicator (1 if treated, 0 if not treated), age, education, Black (1 if black, 0 otherwise), Hispanic (1 if Hispanic, 0 otherwise), married (1 if married, 0 otherwise), no degree (1 if no degree, 0 otherwise), RE74 (earnings in 1974), RE75 (earn-

ings in 1975), and RE78 (earnings in 1978). For our model, the training and the testing observations are the comparison groups and the treated group, respectively.

Table 2.2: *Summary of estimates of the models for the NSW dataset*

Model	Size(N)	P	N_c	N_t	ATT Est	RMSE/Standard Error
Mean weight(PSID)	2675	9	2490	185	846.85	738.969
Random Forest(PSID)	2675	9	2490	185	798.23	738.320
Causal Forest(PSID)	2675	9	2490	185	5620.03	598.25
Mean weight(CPS)	16177	9	15992	185	1265.25	520.309
Random Forest(CPS)	16177	9	15992	185	1205.70	518.855
Causal Forest(CPS)	16177	9	15992	185	4306.30	544.08

Many studies have considered observational data as an alternative to forming the comparison group to determine a possible estimation of the experimental benchmark result. The benchmark estimate of the treatment effect from the randomized experiment according to Dehejia and Wahba (2002) and LaLonde (1986) is \$1,794. We used the Population Survey of Income Dynamics (PSID) and Current Population Survey (CPS) non-experimental data to form the comparison group and estimate the treatment effect using MWCSRF, standard random forest(RF), and causal forest. The results from Table 2.2 show our MWCSRF method estimates the treatment effect better than the random forest and causal forest methods. For the PSID dataset with N=2675, N_t = 185 and N_c=2490, the MWCSRF estimates of the treatment effect is \$846.85 with standard error (SE) 738.969 and that of standard random forest \$798.23 with SE of 738.320. Causal forest estimate PSID is \$5620.03 with a SE of 598.25. For the CPS non-experimental dataset with N=16177, N_t = 185 and N_c=15992, the estimate of the treatment effect and their standard errors with the MWCSRF method is \$1265.25 and Random , respectively. The treatment effect estimate for the standard random forest and causal forest methods are \$1205.70 and \$4306.30, respectively, with a corresponding standard error of 518.855 and 544.08. We see that MWCSRF method estimates are closer to the benchmark experimental estimates than random forest estimates. The CPS non-experimental dataset used as a control group provides a better estimate of the average treatment effect for both methods than the PSID as a control group. The causal forest estimates are not consistent because the treated units are not randomly distributed in the

covariate space, as illustrated in Figure 2.7. Thus, any comparison made between causal forest and MWCSRF and RF will be deceptive.

2.6 Discussion and Conclusion

This paper aims to explore the utilization of our proposed mean-weighted case specific random forest (MWCSRF) method for estimating the average treatment effect for the treated (ATT) for data with many control and few treated units and with the treated units concentrated at one area of the covariate space. The results obtained with our method are compared with the baseline standard random forest method. We obtained results from simulation on homogeneous and heterogeneous data settings as well as from real data.

The results from the simulation for both homogeneous and heterogeneous data settings indicated our method (MWCSRF) outperform the baseline standard random forest except for the instances with many treated units. We would have expected the strength of our model to reduce significantly with an increasing number of treated units, but this was not the case. For instance, for a large sample size with corresponding sizable treated units, for both homogeneous and heterogeneous data models, the estimated ATT for our method (MWCSRF) is slightly higher than that of the baseline random forest method. The root mean-square error (RMSE) estimate also shows the marginal difference for both methods. Even though the standard random forest performs better in the above cases, comparatively, the strength of our model does not appear to be considerably reduced since the ATT estimates and RMSE are very close. The MWCSRF method estimates from the real dataset are closer to the experimental benchmark estimates than the baseline standard random forest estimates.

Chapter 3

MWCSRF: An R package for estimating ATT using potential outcomes

3.1 Introduction

The use of machine learning methods for estimating causal effect is fast gaining prominence in the causal inference literature, chiefly among education (Wager and Athey, 2018), medicine (McConnell and Lindner, 2019), and economics (Shalit et al., 2017). With such interest in machine learning methods for estimating treatment effect, we tailored our attention to developing a version of random forest for estimating the average treatment effect for the treated when there are many control units with minimally treated units. Thus, we developed the R package MWCSRF, an extension of CSRF under the Neyman-Rubin causal model for estimating treatment effect in settings with enormous control units and minimally treated units, which mainly concentrate on the minute area of the covariate space. To our awareness, there is presently only one R package, grf (Athey et al., 2019), that extends the random forest methodology to estimate heterogeneous treatment effects under the potential

outcome framework.

This chapter focuses on exhibiting the MWCSRF package's functionality for estimating the average treatment effect for the treated(ATT) in a setting with a large number of control units and small treated units and treated units are concentrated in one area of the covariate space.

3.2 Neyman Rubin Causal Model

The use of Neyman-Rubin causal model framework has received attention across many fields such as statistics, economics, political science,medicine, and so on. Some of the works done in this fields include statistics (Holland, 1986; Rosenbaum, 2002, 2005; Rubin, 1974b, 2006); economics (Abadie and Imbens, 2006; Dehejia and Wahba, 2002; Heckman, 2008), political science (Brady, 2002; Rathbun, 2008; Sekhon, 2008), and medicine (Brady, 2002; Christakis and Iwashyna, 2003; Rubin, 1997)). The origin of the model framework can be traced to (Neyman, 1923(1990))) and his nonparametric model for finite number of treatments where there exists two potential outcomes for each unit, one if the unit receive treatment and the other if is control. Thus, a causal effect is the difference between two potential outcomes, but only one potential outcome is actually observed.

Now, suppose there are two treatment regimes: treatment and control, and let $t = 1$ and $t = 0$ denote these respectively. Let T_i be the treatment indicator: $T_i = 1$ if unit i received the treatment and $T_i = 0$ if unit i received the control. Let Y_{i1} denote the potential outcome if unit i receives the treatment and Y_{i0} be the potential outcome for unit i in the control regime. Then the treatment effect for observation i is defined as

$$\tau_i = Y_{i1} - Y_{i0} \tag{3.1}$$

and the average treatment effect (ATE) as

$$\tau = \frac{1}{n}\sum_{i=1}^{n}\tau_i = \frac{1}{n}\sum_{i=1}^{n}(Y_{i1} - Y_{i0}), \tag{3.2}$$

where n is the number of units in a finite population.

Since only one potential outcome is truly observed and not both, the unobserved potential outcome is often referred to as counterfactual. Thus, the observed outcome for the observation i can be modeled by Neyman-Rubin Causal Model (NRCM) as

$$Y_i = Y_{i1}T_i + Y_{i0}(1 - T_i), \tag{3.3}$$

where the potential outcomes Y_{i1} and Y_{i0} are fixed quantities and not random. This model does not make any explicit distributional assumptions.

3.3 MWCSRF Functions

The MWCSRF has two main functions, namely csrfweights and MWCSRF. The csrfweights function enables us to compute the weight of all training observations. The MWCSRF functions facilitate the computation of mean weight for all the weights obtained in csrfweights function. The MWCSRF functions also permit for each test observation (treated units), a mean weighted random forest is grown on the training data (control units). Table 3.1 lists the two functions and provides a description and arguments of each.

The csrfweights and MWCSRF functions call the ranger function in the ranger package developed by Wright and Ziegler (2015) for speedy implementation of random forest in high dimensional data. The params1 and params2 in csrfweights and MWCSRF functions contain a list of some parameters which are called from the ranger package that can be tuned in random forest model implementation as described in Subsection 1.4.2. The params1 comprises num.trees and mtry with the default value for num.trees as 500. The default

Table 3.1: *Summary of functions in MWCSRF package*

Function	Description and Arguments
csrfweights	For each observation of interest (treated units), compute the weight of all training observations (control units) by tallying the number of trees in which both treat treated and control units appear in the same terminal node. • First, grow a random forest (RF) on the training data (control units) with the parameters response, Training.data (treated units), and params1 • Get terminal nodes for training data (control units) and test data (treated units) using the RF above to predict terminal nodes for both. • Compute weights by counting the number of trees in which both training data (control units) and test data (treated units) of interest appeared in the same terminal node and divide by their total
MWCSRF	This function computes the mean weight of the weights obtained in the csrfweights function. It uses all control units as training data to fit the mean-weighted case-specific random forest (MWCSRF) model that predicts prognostic scores for all treated units. • MWCSRF.model outputs the weights created in the csrfweights function • matrix.MWCSRF generate a matrix of weights obtained in MWCSRF.model with columns equal to the length of MWCSRF.model • MWCSRF.weights get the mean weight of weights generated in the matrix.MWCSRF by finding the mean of each column in the matrix.MWCSRF. • The function uses parameters such as response, Training.data (control units), case.weights, and params2 to fit the mean-weighted case-specific random forest(MWCSRF) model for predictions of a prognostic score for all treated units.

values for mtry is the square root of the number of predictor variables (p) rounded down for classification problems and $p/3$ for regression problems. The params2 have many parameters that can be called from the ranger package, such as num.tree, importance, splitrule, etc. Thus the implementation of the MWCSRF package for estimating the treatment effect for the treated is demonstrated in Section 3.4

3.4 Example

We demonstrate using the MWCSRF package to estimate the treatment effect for the treated using the SUPPORT data on Right Heart Catheterization (RHC). The RHC is a diagnostic procedure performed for critically ill patients. The SUPPORT study has data on adult patients hospitalized at five medical centers in the USA. The data set contained 5735 individuals admitted or transferred to an ICU in the first 24 hours after joining the study. There were 2184 treated, that is, patients who received RHC within 24 hours after entering the study and 3551 controls (patients who do not receive an RHC). The response variable is survival at 30 days (death) post-admission. We included 56 covariates to demonstrate the use of the MWCSRF package. Further information and details of this study can be found in Connors et al. (1996).

The R code in Figure 3.1 illustrates the use of MWCSRF to estimate the average treatment effect for the treated. Line five in the code orders the data set according to the treatment received by a patient, RHC or No RHC. The data set is partitioned into treated and control groups by lines six and seven in the code. All patients who received No RHC form the control group, and those for whom RHC admitted constitute the treated group. Thus, the training data set is the control group, and the treated group composed the testing data set. In line eight of the code, we fit the mean-weighted case-specific random forest(MWCSRF) model to the training dataset. Users should note that the testing data set appeared in line eight of the codes to enable us to compute the mean weights described in Section 3.3, but is unused in MWCSRF model estimation. We obtain estimates of the prognostic scores as in line nine of the codes, the predicted outcome under control (No RHC) condition for all treated units(patients who received RHC) as described in Section 2.2.4. We calculate the average treatment effect for the treated(ATT) in line thirteen as the average difference between the treated units' potential outcomes and prognostic scores. Thus, running the code, we found the ATT as 0.0069 with an estimated prediction error of 0.3040. In addition,

```
> set.seed(101)
> library(MWCSRF)
> library(ranger)
> rhc2 <- read.csv("~/Desktop/STAT999/Dessertation codes/rhc2.csv")
> rhc3<-colnames(rhc2)[apply(rhc2 , 2, anyNA)]
> rhc4<-subset(rhc2,select=-c(X,cat2,dthdte,adld3p,urin1,dschdte))
> attach(rhc4)
> rhc5<-rhc4[order(swang1),]
> rhc.training=rhc5[1:3551,]
> rhc.test=rhc5[3552:5735,]
> model.rhc=MWCSRF(as.factor(death)~.,Training.data =rhc.training,
+ Test.data  =  rhc.test, params1 = list(num.trees = 2000, mtry=7),
+ params2 = list(num.trees = 5,importance="impurity"))
> Prognosticscore= predict(model.rhc, rhc.test,type="response")$predictions
> HJ=table(rhc.test$death,Prognosticscore)
> Accuracy=sum(diag(HJ))/sum(HJ)
> A=mean(Prognosticscore==rhc.test$death)
> ATT=mean(ifelse(rhc.test[,5] == "Yes",1,0)-ifelse(Prognosticscore == "Yes",1,0))
> Accuracy

[1] 0.7596154

> ATT

[1] 0.006868132

> model.rhc$prediction.error

[1] 0.3040128
```

1

Figure 3.1: *Example demonstrating the use of MWCSRF to estimate treatment effect for the treated(ATT)*

results of various simulation scenarios for both homogeneous and heterogeneous models are in Appendices A and A which exhibit better performance in most cases of the MWCSRF model compare to other models studied in this dissertation.

Chapter 4

The Lady Tasting Tea Revisited: Insights on SUTVA Violations from a Canonical Example

4.1 Introduction

A topic in causal inference under considerable recent study involves designing and analyzing experiments under treatment interference (i.e. treatment spillover). Treatment interference is a common setting for data which exhibit network structure and units interact with each other, for example, epidemiology and social media networks.

Difficulties of analysis under treatment interference stem from a violation of a foundational assumption in traditional causal inference: the stable unit treatment value assumption (SUTVA). SUTVA assumes that every unit has two potential outcomes—one under treatment and under control—and the observed outcome depends only on the treatment given to that unit. and that

The Stable Unit Treatment Value Assumption (SUTVA) was proposed in Rubin (1980), but has earlier been discussed by Cox (1958),where they assumed that no interference be-

tween units exists. SUTVA states that "The potential outcomes for any unit do not vary with the treatments assigned to any other units, and there are no different versions of the treatment." (Rubin, 1986). In causal inference, SUTVA ensures that (1) there exists as many potential outcomes as the number of value the treatment can take on, and (2) under SUTVA we can observe at least one potential outcomes for each unit.

In this paper we focus attention on Fisher's famous experiment Fisher (1935), where a lady proposed she can discern whether a tea is prepared by adding milk first or tea first by tasting it. We are interested in the design setting where there are ten cups of tea and suppose that five cups are given milk first and these cups are randomized across the ten cups. Assuming for each cup the lady estimate a likelihood scores between 0 and 1 for the cups that received milk first and then guesses the milk first cups according to the largest likelihood score. Under this setting it is possible that SUTVA may be violated. Our goal is to explore various scenarios under this design setting and formulate a hypothesis that will result in violation of SUTVA.

4.2 Model of response and SUTVA violations

To begin, we set up Fisher's Lady Tasting Tea experiment and describe our model for how the Lady makes a determination about which cups are given milk first.

4.2.1 Lady Tasting Tea set-up

The Lady Tasting Tea experiment is constructed as follows. There are n cups, numbered 1 through n. All cups are filled with tea and milk; of the n cups, n_m are filled with milk first, and then tea is added; the other $n_t = n - n_m$ cups are filled with tea first, and then milk is added. We assume that n and n_m are pre-specified and are known to both the experimenter *and the Lady*. Additionally the milk-first or tea-first condition is completely randomized across cups. Let T_i denote a treatment indicator for whether cup i is given milk first: $T_i = 1$

if cup i is given milk first, and $T_i = 0$ if cup i is given tea first.

The Lady tastes all n cups and guesses n_m cups to be given milk first. Let G_i denote the response for the lady: $G_i = 1$ if the lady guesses milk first for cup i and $G_i = 0$ if the lady guesses tea first for cup i. A correct guess occurs when $T_i = G_i$. Note that, by design, $\sum_{i=1}^{n} G_i = \sum_{i=1}^{n} T_i = n_m$.

4.2.2 Lady's model of response

For each cup i, the Lady derives a *likelihood score* L_i for that cup being given milk first. We assume $0 \leq L_i \leq 1$, where larger L_i implies a greater belief in the Lady that cup i was given milk first. We assume that the likelihood score satisfies the Neyman-Rubin potential outcomes model of response Holland (1986); Rubin (1974a); Splawa-Neyman et al. (1990):

$$L_i = \ell_{i1} T_i + \ell_{i0}(1 - T_i) \tag{4.1}$$

where ℓ_{i1} and ℓ_{i0} are the potential likelihoods for whether cup i was given milk first or tea first respectively. That are ℓ_i are assumed to be non-random, and thus, the randomness in the likelihood L_i for cup i is isolated to the random milk-first assignment given to cup i. Of particular note, the likelihood score satisfies SUTVA. For simplicity, we will assume L_i is unique for each cup i.

If the Lady has no knowledge of how many cups n_m were given milk first, she may simply guess that a cup is given milk first if $L_i > 0.5$. However, because the Lady knows that exactly n_m cups are given milk first, she will only guess milk first for the cups with the n_m largest likelihood scores. That is, the responses G_i satisfy:

$$G_i = \begin{cases} 1, & L_i \geq L_{(n_m)}, \\ 0, & \text{otherwise.} \end{cases} \tag{4.2}$$

where $L_{(n_m)}$ is the n_mth largest value of the likelihood. By the uniqueness assumption of

the L_i, we are guaranteed $\sum_{i=1}^{n} G_i = n_m$.

4.3 When Lady Tasting Tea Violates SUTVA

SUTVA requires that the response only depends on the treatment status of the given cup and is not affected by the treatment status of any other cup. Thus, the Lady's guesses will violate SUTVA if T_i is unchanged for two randomizations of milk-first cups but G_i changes across randomizations. We now detail cases under which SUTVA is and is not violated.

4.3.1 Demonstration of SUTVA Violation

For the following two examples, we will assume that there are $n = 10$ cups, with $n_m = n_t = 5$. Table 4.1 gives one example of potential likelihoods in which the Lady's guesses violate SUTVA. Note that, for this example, Cup 6 does not change treatment assignment, yet the Lady guesses this cup to be milk-first for the left randomization and guesses it to be tea-first for the right randomization.

Table 4.1: *Example of likelihoods for which guesses violate SUTVA.*

Cup i	T_i	L_i	ℓ_{i1}	ℓ_{i0}	G_i	Cup i	T_i	L_i	ℓ_{i1}	ℓ_{i0}	G_i
1	1	0.05	0.05	0.90	0	1	0	0.90	0.05	0.90	1
2	1	0.15	0.15	0.00	0	2	1	0.15	0.15	0.00	0
3	1	0.25	0.25	0.80	0	3	0	0.80	0.25	0.80	1
4	0	0.50	0.35	0.50	1	4	0	0.50	0.35	0.50	0
5	0	0.20	0.45	0.20	0	5	0	0.20	0.45	0.20	0
6	1	0.55	0.55	0.40	1	6	1	0.55	0.55	0.40	0
7	0	0.30	0.65	0.30	0	7	1	0.65	0.65	0.30	0
8	0	0.70	0.75	0.70	1	8	0	0.70	0.75	0.70	1
9	1	0.85	0.85	0.10	1	9	1	0.85	0.85	0.10	1
10	0	0.60	0.95	0.60	1	10	1	0.95	0.95	0.60	1

4.3.2 SUTVA Violation Under Monotonicity of Likelihoods

The previous example in Table 4.1 may be a seem a bit contrived. For instance, the Lady will incorrectly place a high likelihood on Cup 1 being milk-first when that cup is given tea-first and vice versa. However, a more plausible assumption may be *monotonicity of likelihoods*—that the Lady's milk-first likelihood for cup i is always as high or higher when i is given milk first. Expressed in terms of potential likelihoods, this assumption is:

$$\ell_{i1} \geq \ell_{i0}, \quad i = 1, 2, \ldots, n. \tag{4.3}$$

This monotonicity assumption is *strict* if the inequality (4.3) is strict.

Table 4.2 gives an example of potential likelihoods that satisfy strict monotonicity, but yet, the Lady's guesses violate SUTVA. For this example, Cup 4 does not change treatment assignment, yet the Lady guesses this cup to be milk-first for the left randomization and guesses it to be tea-first for the right randomization.

Table 4.2: *Example of likelihoods satisfying strict monotonicity for which guesses violate SUTVA*

Cup i	T_i	L_i	ℓ_{i1}	ℓ_{i0}	G_i	Cup i	T_i	L_i	ℓ_{i1}	ℓ_{i0}	G_i
1	1	0.25	0.25	0.00	0	1	0	0.00	0.25	0.00	0
2	1	0.40	0.40	0.05	0	2	0	0.05	0.40	0.05	0
3	0	0.10	0.45	0.10	0	3	0	0.10	0.45	0.10	0
4	1	0.50	0.50	0.15	1	4	1	0.50	0.50	0.15	0
5	0	0.20	0.55	0.20	0	5	1	0.55	0.50	0.20	1
6	0	0.35	0.65	0.35	0	6	0	0.35	0.65	0.35	0
7	1	0.75	0.75	0.55	1	7	1	0.75	0.75	0.55	1
8	1	0.85	0.85	0.60	1	8	0	0.60	0.85	0.60	1
9	0	0.70	0.90	0.70	1	9	1	0.90	0.90	0.70	1
10	0	0.80	0.95	0.80	1	10	1	0.95	0.95	0.80	1

4.3.3 No SUTVA Violation Under Sharp Null

We now detail two cases under which SUTVA is satisfied for Lady Tasting Tea. First, we show that SUTVA is satisfied under the sharp null hypothesis; that is, that a cup given milk first yields no information to the Lady about whether the cup is given milk first. Expressed in terms of likelihoods, this amounts to the following hypothesis:

$$H_0 : \ell_{i1} = \ell_{i0}, \quad i = 1, 2, \ldots, n. \tag{4.4}$$

Define $G_i(\mathbf{T})$ as the Lady's guess for cup i given treatment allocation $\mathbf{T} = (T_1, T_2, \ldots, T_n)$. To see that SUTVA holds under (4.4), it suffices to note that, under the sharp null, the likelihood score L_i will be the same regardless of that the treatment assignment for each cup i. Thus the ordering of the likelihoods, and as a consequence, the guesses, will be the same as well. In particular, for any two treatment allocations $\mathbf{T}$, $\mathbf{T}'$, with $T_i = T_i'$, $G_i(\mathbf{T}) = G_i(\mathbf{T}')$, and hence, SUTVA is satisfied.

4.3.4 No SUTVA Violation Under Perfect Knowledge

The second case for which SUTVA is satisfied is when the Lady has perfect knowledge about which cups are given milk first and which cups are given tea first. Specifically, the setting of perfect information occurs when

$$\min_i \ell_{i1} > \max_i \ell_{i0}. \tag{4.5}$$

Under this setting, the Lady will always guess the milk-first cups correctly, and hence, the Lady's guess for cup i only depends on the treatment status given to that cup.

To see this, note that, for all cups i given milk first, $L_i \geq min_i \ell_{i1}$, and for all cups j given

tea-first, $L_j \leq max_j \ell_{j0}$. That is, for any cups i, j where $T_i = 1$ and $T_j = 0$,

$$L_{(i)} \geq L_{(n_m)} \geq \min_i \ell_{i1} > \max_i \ell_{i0} \geq L_{(n_m+1)} \geq L_{(j)}. \tag{4.6}$$

Since i and j were arbitrary, it follows from (4.2) that all milk-first cups will have $G_i = 1$ and all tea-first cups j will have $G_j = 0$. In other words, for all treatment allocations $\mathbf{T}$, $G_i(T) = T_i$; the guess for cup i only depends on the treatment status given to cup i. And so, SUTVA holds under perfect information.

Chapter 5

Conclusion

5.1 Summary of dissertation

Causal inference, an aspect of statistics, deals with defining the effects of treatment on the response. In this dissertation, we studied two problems in causal inference under the Neyman-Rubin causal model of potential outcomes. Under this model, we estimate counterfactuals — hypothetical unobserved responses of units under different treatment conditions—treatment effect estimation under treatment spillover— when the treatment status of one unit affects the response of another

In Chapter 2, we developed the mean-weighted case-specific random forests (MWCSRF) to estimate the average treatment effect for the treated (ATT). Our study focused on settings where the data comprise many control units and very few treated units and treated units occupy one area of the covariate space. For instance, for COVID-19 data, treated units would be those that participated in clinical trials for vaccines efficacy and side effects, and control units may originate from the national COVID-19 database. We made an extension to the case-specific random forest(CSRF), a method designed to improve prediction on individual observations by taking bootstrap samples from data set that assigned more weight to cases in proximity to where prediction is of utmost importance. Our mean-weighted case-

specific random forest(MWCSRF), an extension of CSRF, selects bootstrap samples using mean weights. Thus, the main distinction between CSRF and MWCSRF is selecting bootstrap samples for random forest model estimation and prediction. We used all the control units as the training data set to develop the MWCSRF model. Using the treated units as the testing data set, we predict prognostic scores—the expected control outcome given a set of covariates. We obtained the ATT as the average difference between the treated units' potential outcome and prognostic score. We compare the performance of the MWCSRF method with random forest(RF), CSRF, genetic matching, and causal forest. Simulation results suggest encouraging performance of the MWCSRF method compared to other methods under both homogeneous and heterogeneous treatment effect settings, especially when the number of treated units is small and treated units concentrated in one area of the covariate space. We then apply our method to a study on the efficacy of the National Supported Work Demonstration program.

In Chapter 3, we present a direction to our R package MWCSRF to facilitate the implementation of the MWCSRF method. This package is a supportive accompaniment to the work completed in this dissertation. It comprises functions that will estimate the MWCSRF model for predicting prognostic scores and average treatment effect for the treated(ATT) estimation. Hopefully, researchers will find this package helpful in calculating the average treatment effect for the treated(ATT), especially in settings with small treated units and substantial control units, where treated units in the covariate space not randomly distributed.

Finally, in Chapter 4, we discuss the issue of treatment spillover in the framework of Fisher's Lady Tasting Tea experiment. Randomized controlled experiments are the "gold standard" in Causal Inference, and Fisher's Lady Tasting Tea is a foundational example in experimental design. We illustrate that Lady Tasting Tea can violate the stable unit treatment value(SUTVA) assumption under certain conditions. (SUTVA) entails the response of a unit to be only affected by the treatment status of that unit. SUTVA is a generally utilized assumption in causal inference, and its violation usually occurred in studies showing

some type of treatment spillover. We consider the following model of response for the Lady Tasting Tea experiment. The Lady knows that half of all cups receive tea first and the other half receive milk first and that this assignment is completely randomized. For each cup, the Lady obtains a "likelihood score" of that cup receiving milk first, which may differ depending on whether that cup is indeed given milk first. The Lady guesses the cups with the most significant likelihood scores receive milk first after tasting tea from all cups. We demonstrate that SUTVA may be violated under this model even when, for a given cup, the Lady's milk-first likelihood score is always higher when that cup actually receives milk first. Furthermore, we illustrate that SUTVA holds under two conditions: one in which the Lady's likelihood score for a cup is the same irrespective of whether that cup was given milk first or tea first, and one in which the Lady always makes perfect guesses.

5.2 Future Research

The work in this dissertation focuses exclusively on model estimation and prediction, but in our future research, we will provide theoretical background and inferential analysis. This includes asymptotic derivations, confidence intervals, and hypothesis testing for our estimates. Generally, it's pretty challenging to give asymptotic derivation to a random forest model, which will lead to the construction of confidence interval and hypothesis testing for random forest estimates. There are few results on the theoretical behavior of random forests (Mentch and Hooker, 2016; Wager and Athey, 2018). Such theoretical papers help enable confidence interval construction and hypothesis testing for the estimates. For future research, we also intend to extend our MWCSRF method to quantile regression and compare our estimates with MWCSRF, where the regression option in the random forest is the technique for obtaining the estimates.

Bibliography

Alberto Abadie and Javier Gardeazabal. The economic costs of conflict: A case study of the basque country. *American economic review*, 93(1):113–132, 2003.

Alberto Abadie and Guido W Imbens. Large sample properties of matching estimators for average treatment effects. *econometrica*, 74(1):235–267, 2006.

Ricardo Aler, José M Valls, and Henrik Boström. Study of hellinger distance as a splitting metric for random forests in balanced and imbalanced classification datasets. *Expert Systems with Applications*, 149:113264, 2020.

S Athey and S Wager. Estimating treatment effects with causal forests: An application, forthcoming in. *Observational Studies*, 2019.

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

Simon Bernard, Sébastien Adam, and Laurent Heutte. Dynamic random forests. *Pattern Recognition Letters*, 33(12):1580–1586, 2012.

Henry E Brady. Models of causal inference: Going beyond the neyman-rubin-holland theory. In *Annual Meetings of the Political Methodology Group*, 2002.

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

Leo Breiman and Adele Cutler. The original fortran code of the randfor. *Statistics Department University of California Berkeley, CA, USA*, 2002.

Leo Breiman, Jerome H Friedman, Richard A Olshen, and Charles J Stone. Classification and regression trees. belmont, ca: Wadsworth. *International Group*, 432:151–166, 1984.

Nicholas A Christakis and Theodore J Iwashyna. The health impact of health care on families: a matched cohort study of hospice use by decedents and mortality outcomes in surviving, widowed spouses. *Social science & medicine*, 57(3):465–475, 2003.

David Collett. *Modelling survival data in medical research*. CRC press, 2015.

Alfred F Connors, Theodore Speroff, Neal V Dawson, Charles Thomas, Frank E Harrell, Douglas Wagner, Norman Desbiens, Lee Goldman, Albert W Wu, Robert M Califf, et al. The effectiveness of right heart catheterization in the initial care of critically iii patients. *Jama*, 276(11):889–897, 1996.

D. R. Cox. The Planning of Experiments. *Operations Research*, 1958.

A Philip Dawid. Conditional independence in statistical theory. *Journal of the Royal Statistical Society: Series B (Methodological)*, 41(1):1–15, 1979.

Rajeev H Dehejia and Sadek Wahba. Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American statistical Association*, 94 (448):1053–1062, 1999.

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

Ramón Díaz-Uriarte and Sara Alvarez De Andres. Gene selection and classification of microarray data using random forest. *BMC bioinformatics*, 7(1):3, 2006.

Ronald A. Fisher. The Design of Experiments. *Edinburgh; Oliver and Boyd*, 1935.

Jared C Foster, Jeremy MG Taylor, and Stephen J Ruberg. Subgroup identification from randomized clinical trial data. *Statistics in medicine*, 30(24):2867–2880, 2011.

Robin Genuer, Jean-Michel Poggi, and Christine Tuleau-Malot. Variable selection using random forests. *Pattern recognition letters*, 31(14):2225–2236, 2010.

Pierre Geurts, Damien Ernst, and Louis Wehenkel. Extremely randomized trees. *Machine learning*, 63(1):3–42, 2006.

Ben B Hansen. The prognostic analogue of the propensity score. *Biometrika*, 95(2):481–488, 2008.

James J Heckman. Econometric causality. *International statistical review*, 76(1):1–27, 2008.

Jennifer Hill and Elizabeth A Stuart. Causal inference: Overview. In *International Encyclopedia of the Social & Behavioral Sciences: Second Edition*, pages 255–260. Elsevier Inc., 2015.

Jennifer L Hill. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011.

Paul W Holland. Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960, 1986.

De-Shuang Huang, Donald C Wunsch, Daniel S Levine, and Kang-Hyun Jo. *Advanced Intelligent Computing Theories and Applications. With Aspects of Artificial Intelligence: Fourth International Conference on Intelligent Computing, ICIC 2008 Shanghai, China, September 15-18, 2008, Proceedings*, volume 5227. Springer, 2008.

Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

Hemant Ishwaran. The effect of splitting on random forests. *Machine Learning*, 99(1):75–118, 2015.

Hemant Ishwaran and James D Malley. Synthetic learning machines. *BioData mining*, 7(1): 28, 2014.

Hemant Ishwaran, Udaya B Kogalur, Eugene H Blackstone, Michael S Lauer, et al. Random survival forests. *The annals of applied statistics*, 2(3):841–860, 2008.

Peter Kemper and David Long. *The Supported Work Evaluation: Technical Report on Value of In-program Output and Costs*. Manpower Demonstration Research Corporation New York, 1981.

Peter Kemper, David A Long, Craig VD Thornton, and Robinson G Hollister. The supported work evaluation: Final benefit-cost analysis. 1981.

Sören R Künzel, Bradly C Stadie, Nikita Vemuri, Varsha Ramakrishnan, Jasjeet S Sekhon, and Pieter Abbeel. Transfer learning for estimating causal effects using neural networks. *arXiv preprint arXiv:1808.07804*, 2018.

Robert J LaLonde. Evaluating the econometric evaluations of training programs with experimental data. *The American economic review*, pages 604–620, 1986.

Andy Liaw and Matthew Wiener. Classification and regression by randomforest. *R News*, 2 (3):18–22, 2002. URL https://CRAN.R-project.org/doc/Rnews/.

Hanzhong Liu and Yuehan Yang. Penalized regression adjusted causal effect estimates in high dimensional randomized experiments. *arXiv preprint arXiv:1809.08732*, 2018.

Min Lu, Saad Sadiq, Daniel J Feaster, and Hemant Ishwaran. Estimating individual treatment effect in observational data using random forest methods. *Journal of Computational and Graphical Statistics*, 27(1):209–219, 2018.

Brian MacMahon, Stella Yen, Dimitrios Trichopoulos, Kenneth Warren, and George Nardi. Coffee and cancer of the pancreas. *New England Journal of Medicine*, 304(11):630–633, 1981.

Charles F Manski. Economic analysis of social interactions. *Journal of economic perspectives*, 14(3):115–136, 2000.

K John McConnell and Stephan Lindner. Estimating treatment effects with machine learning. *Health services research*, 54(6):1273–1282, 2019.

Nicolai Meinshausen. Quantile regression forests. *Journal of Machine Learning Research*, 7 (Jun):983–999, 2006.

Lucas Mentch and Giles Hooker. Quantifying uncertainty in random forests via confidence intervals and hypothesis tests. *The Journal of Machine Learning Research*, 17(1):841–881, 2016.

James N Morgan and Robert C Messenger. Thaid, a sequential analysis program for the analysis of nominal scale dependent variables. 1973.

Justine B Nasejje, Henry Mwambi, Keertan Dheda, and Maia Lesosky. A comparison of the conditional inference survival forest model to random survival forests based on a simulation study as well as on two applications with time-to-event data. *BMC medical research methodology*, 17(1):115, 2017.

J. Neyman. On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. *Statistical Science*, 5:465–472, 1923(1990).

Miruna Oprescu, Vasilis Syrgkanis, and Zhiwei Steven Wu. Orthogonal random forest for causal inference. *arXiv preprint arXiv:1806.03467*, 2018.

Brian Christopher Rathbun. Interviewing and qualitative field methods: pragmatism and practicalities. 2008.

Paul R Rosenbaum. Overt bias in observational studies. In *Observational studies*, pages 71–104. Springer, 2002.

Paul R Rosenbaum. Heterogeneity and causality: Unit heterogeneity and design sensitivity in observational studies. *The American Statistician*, 59(2):147–152, 2005.

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

Donald B. Rubin. Estimating Causal Effects of Treatment in Randomized and Nonrandomized Studies. *Journal of Educational Psychology*, 66:688–701, 1974a.

Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974b.

Donald B Rubin. Assignment to treatment group on the basis of a covariate. *Journal of educational Statistics*, 2(1):1–26, 1977.

Donald B. Rubin. Comment on: Randomization analysis of experimental data in the fisher randomization test by D. Basu. *American Statistical Association*, 75:591–593, 1980.

Donald B. Rubin. Which if have causal answers? Comment on: Statistics and causal inference by P. Holland. *American Statistical Association*, 81:961–962, 1986.

Donald B Rubin. Estimating causal effects from large data sets using propensity scores. *Annals of internal medicine*, 127(8_Part_2):757–763, 1997.

Donald B Rubin. *Matched sampling for causal effects*. Cambridge University Press, 2006.

Jasjeet S Sekhon. The neyman-rubin model of causal inference and estimation via matching methods. *The Oxford handbook of political methodology*, 2:1–32, 2008.

Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *International Conference on Machine Learning*, pages 3076–3085. PMLR, 2017.

Susan M Shortreed and Ashkan Ertefaie. Outcome-adaptive lasso: Variable selection for causal inference. *Biometrics*, 73(4):1111–1122, 2017.

Jeffrey A Smith and Petra E Todd. Does matching overcome lalonde's critique of nonexperimental estimators? *Journal of econometrics*, 125(1-2):305–353, 2005.

Jerzy Splawa-Neyman, Dorota M Dabrowska, and TP Speed. On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science*, pages 465–472, 1990.

Youmi Suk, Hyunseung Kang, and Jee-Seon Kim. Random forests approach for causal inference with clustered observational data. 2019.

Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.

Leonie Weinhold, Matthias Schmid, Marvin N Wright, and Moritz Berger. A random forest approach for modeling bounded outcomes. *arXiv preprint arXiv:1901.06211*, 2019.

Stacey J Winham, Robert R Freimuth, and Joanna M Biernacka. A weighted random forests approach to improve predictive performance. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 6(6):496–505, 2013.

Marvin N Wright and Andreas Ziegler. ranger: A fast implementation of random forests for high dimensional data in c++ and r. *arXiv preprint arXiv:1508.04409*, 2015.

Marvin N. Wright and Andreas Ziegler. ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, 77(1):1–17, 2017. doi: 10.18637/jss.v077.i01.

Marvin N Wright, Theresa Dankowski, and Andreas Ziegler. Unbiased split variable selection for random survival forests using maximally selected rank statistics. *Statistics in medicine*, 36(8):1272–1284, 2017.

Edward Wu and Johann A Gagnon-Bartsch. The loop estimator: Adjusting for covariates in randomized experiments. *Evaluation review*, 42(4):458–488, 2018.

Ruo Xu, Dan Nettleton, and Daniel J Nordman. Case-specific random forests. *Journal of Computational and Graphical Statistics*, 25(1):49–65, 2016.

Donghui Yan, Aiyou Chen, and Michael I Jordan. Cluster forests. *Computational Statistics & Data Analysis*, 66:178–192, 2013.

Appendix A

Summary of estimates of the models from the homogeneous and heterogeneous with true ATT as 4 when the number of predictors is varied

Table A.1: *Summary of estimates of the models from homogeneous data simulation with true ATE as* ***4.0*** *and* $N = 210$, $N_c = 200$, $N_t = 10$, *and* $p = 5$

Model	M	Size(n)	P	Cont	Trt	Estimate	Expected Bias	Empirical SE	Theoretical SE
Mean weight	0	210	5	200	10	3.9223	-0.0777	0.3963	0.3201
Random Forest	0	210	5	200	10	3.8855	-0.1145	0.3855	0.3188
Genetic Matching	0	210	5	200	10	3.9372	-0.0628	0.6127	0.5303
CSRF	0	210	5	200	10	3.9183	-0.0817	0.3966	0.3205
Causal Forest	0	210	5	200	10	3.1940	-0.8060	0.3293	0.0032
Mean weight	0.5	210	5	200	10	3.9489	-0.0511	0.3986	0.3254
Random Forest	0.5	210	5	200	10	3.9159	-0.0841	0.3902	0.3242
Genetic Matching	0.5	210	5	200	10	3.9526	-0.0474	0.5569	0.5097
CSRF	0.5	210	5	200	10	3.9434	-0.0566	0.3983	0.3255
Causal Forest	0.5	210	5	200	10	3.2855	-0.7145	0.3398	0.0039
Mean weight	-0.5	210	5	200	10	4.0361	0.0361	0.3987	0.3171
Random Forest	-0.5	210	5	200	10	4.0068	0.0068	0.3870	0.3158
Genetic Matching	-0.5	210	5	200	10	4.1036	0.1036	0.6308	0.5371
CSRF	-0.5	210	5	200	10	4.0324	0.0324	0.3988	0.3176
Causal Forest	-0.5	210	5	200	10	3.2810	-0.7190	0.3179	0.0029
Mean weight	0.9	210	5	200	10	4.0434	0.0434	0.3881	0.3266
Random Forest	0.9	210	5	200	10	4.0247	0.0247	0.3818	0.3253
Genetic Matching	0.9	210	5	200	10	4.0010	0.0010	0.5487	0.4793
CSRF	0.9	210	5	200	10	4.0430	0.0430	0.3893	0.3277
Causal Forest	0.9	210	5	200	10	3.6308	-0.3692	0.3385	0.0035
Mean weight	-0.9	210	5	200	10	4.2654	0.2654	0.4107	0.3198
Random Forest	-0.9	210	5	200	10	4.2536	0.2536	0.3968	0.3182
Genetic Matching	-0.9	210	5	200	10	4.5053	0.5053	0.6781	0.5698
CSRF	-0.9	210	5	200	10	4.2690	0.2690	0.4101	0.3204
Causal Forest	-0.9	210	5	200	10	3.5393	-0.4607	0.3346	0.0020

Table A.2: *Summary of estimates of the models from homogeneous data simulation with true ATE as* ***4.0*** *and* $N = 210$, $N_c = 200$, $N_t = 10$, *and* $p = 10$

Model	M	Size(n)	P	Cont	Trt	Estimate	Expected.Bias	Empirical.SE	Theoretical.SE
Mean weight	0	210	10	200	10	3.8079	-0.1921	0.3993	0.3259
Random Forest	0	210	10	200	10	3.7786	-0.2214	0.3921	0.3253
Genetic Matching	0	210	10	200	10	3.8504	-0.1496	0.6309	0.5529
CSRF	0	210	10	200	10	3.8036	-0.1964	0.3987	0.3258
Causal Forest	0	210	10	200	10	3.0799	-0.9201	0.3228	0.0028
Mean weight	0.5	210	10	200	10	3.8429	-0.1571	0.3919	0.3363
Random Forest	0.5	210	10	200	10	3.8131	-0.1869	0.3857	0.3355
Genetic Matching	0.5	210	10	200	10	3.8080	-0.1920	0.5871	0.5497
CSRF	0.5	210	10	200	10	3.8374	-0.1626	0.3912	0.3361
Causal Forest	0.5	210	10	200	10	3.1432	-0.8568	0.3362	0.0036
Mean weight	-0.5	210	10	200	10	3.9392	-0.0608	0.3837	0.3197
Random Forest	-0.5	210	10	200	10	3.9168	-0.0832	0.3765	0.3190
Genetic Matching	-0.5	210	10	200	10	4.0100	0.0100	0.6534	0.5529
CSRF	-0.5	210	10	200	10	3.9381	-0.0619	0.3829	0.3198
Causal Forest	-0.5	210	10	200	10	3.1416	-0.8584	0.3117	0.0027
Mean weight	0.9	210	10	200	10	4.0298	0.0298	0.3845	0.3314
Random Forest	0.9	210	10	200	10	4.0088	0.0088	0.3795	0.3306
Genetic Matching	0.9	210	10	200	10	3.9573	-0.0427	0.5282	0.5065
CSRF	0.9	210	10	200	10	4.0231	0.0231	0.3854	0.3319
Causal Forest	0.9	210	10	200	10	3.5176	-0.4824	0.3354	0.0045
Mean weight	-0.9	210	10	200	10	4.2365	0.2365	0.3763	0.3198
Random Forest	-0.9	210	10	200	10	4.2295	0.2295	0.3680	0.3193
Genetic Matching	-0.9	210	10	200	10	4.5225	0.5225	0.6665	0.5626
CSRF	-0.9	210	10	200	10	4.2370	0.2370	0.3758	0.3202
Causal Forest	-0.9	210	10	200	10	3.4232	-0.5768	0.3146	0.0028

Table A.3: *Summary of estimates of the models from homogeneous data simulation with true ATE as* ***4.0*** *and* $N = 210$, $N_c = 200$, $N_t = 10$, *and* $p = 50$

Model	M	Size(n)	P	Cont	Trt	Estimate	Expected.Bias	Empirical.SE	Theoretical.SE
Mean weight	0	210	50	200	10	3.5655	-0.4345	0.3836	0.3309
Random Forest	0	210	50	200	10	3.5604	-0.4396	0.3810	0.3306
Genetic Matching	0	210	50	200	10	3.6858	-0.3142	0.6452	0.5546
CSRF	0	210	50	200	10	3.5656	-0.4344	0.3823	0.3311
Causal Forest	0	210	50	200	10	3.0072	-0.9928	0.3313	0.0027
Mean weight	0.5	210	50	200	10	3.6318	-0.3682	0.3931	0.3498
Random Forest	0.5	210	50	200	10	3.6260	-0.3740	0.3898	0.3488
Genetic Matching	0.5	210	50	200	10	3.6263	-0.3737	0.6590	0.5798
CSRF	0.5	210	50	200	10	3.6299	-0.3701	0.3924	0.3499
Causal Forest	0.5	210	50	200	10	3.0308	-0.9692	0.3395	0.0031
Mean weight	-0.5	210	50	200	10	3.7055	-0.2945	0.3610	0.3206
Random Forest	-0.5	210	50	200	10	3.6996	-0.3004	0.3585	0.3204
Genetic Matching	-0.5	210	50	200	10	3.8177	-0.1823	0.6330	0.5461
CSRF	-0.5	210	50	200	10	3.7036	-0.2964	0.3623	0.3207
Causal Forest	-0.5	210	50	200	10	3.0557	-0.9443	0.3112	0.0026
Mean weight	0.9	210	50	200	10	3.9703	-0.0297	0.4289	0.3479
Random Forest	0.9	210	50	200	10	3.9657	-0.0343	0.4196	0.3464
Genetic Matching	0.9	210	50	200	10	3.8740	-0.1260	0.7363	0.5963
CSRF	0.9	210	50	200	10	3.9650	-0.0350	0.4284	0.3481
Causal Forest	0.9	210	50	200	10	3.3164	-0.6836	0.3549	0.0040
Mean weight	-0.9	210	50	200	10	4.1351	0.1351	0.3624	0.3191
Random Forest	-0.9	210	50	200	10	4.1327	0.1327	0.3586	0.3190
Genetic Matching	-0.9	210	50	200	10	4.3158	0.3158	0.6774	0.5530
CSRF	-0.9	210	50	200	10	4.1330	0.1330	0.3617	0.3191
Causal Forest	-0.9	210	50	200	10	3.2689	-0.7311	0.2997	0.0029

Table A.4: *Summary of estimates of the models from the heterogeneous data simulation with true ATT unknown and* $N = 210$, $N_c = 200$, $N_t = 10$, *and* $p = 5$

Model	M	Size(n)	P	Cont	Trt	ATT	Estimates	Expected Bias	Empirical SE	Theoretical SE
Mean weight	0	210	5	200	10	2.5924	2.5148	-0.0777	0.6784	0.6168
Random Forest	0	210	5	200	10	2.5924	2.4779	-0.1145	0.6713	0.6161
Genetic Matching	0	210	5	200	10	2.5924	2.5297	-0.0628	0.8276	0.8223
CSRF	0	210	5	200	10	2.5924	2.5107	-0.0817	0.6786	0.6171
Causal Forest	0	210	5	200	10	2.5924	1.9422	-0.6503	0.5830	0.0060
Mean weight	0.5	210	5	200	10	2.5928	2.5413	-0.0511	0.6777	0.6186
Random Forest	0.5	210	5	200	10	2.5928	2.5083	-0.0841	0.6724	0.6178
Genetic Matching	0.5	210	5	200	10	2.5928	2.5450	-0.0474	0.7912	0.7881
CSRF	0.5	210	5	200	10	2.5928	2.5358	-0.0566	0.6775	0.6189
Causal Forest	0.5	210	5	200	10	2.5928	2.0132	-0.5792	0.5924	0.0070
Mean weight	-0.5	210	5	200	10	2.5916	2.6276	0.0361	0.6850	0.6164
Random Forest	-0.5	210	5	200	10	2.5916	2.5983	0.0068	0.6777	0.6158
Genetic Matching	-0.5	210	5	200	10	2.5916	2.6952	0.1036	0.8443	0.8441
CSRF	-0.5	210	5	200	10	2.5916	2.6240	0.0324	0.6850	0.6166
Causal Forest	-0.5	210	5	200	10	2.5916	2.0468	-0.5447	0.5751	0.0055
Mean weight	0.9	210	5	200	10	2.5933	2.6367	0.0434	0.6728	0.6216
Random Forest	0.9	210	5	200	10	2.5933	2.6179	0.0247	0.6682	0.6207
Genetic Matching	0.9	210	5	200	10	2.5933	2.5942	0.0010	0.7857	0.7513
CSRF	0.9	210	5	200	10	2.5933	2.6362	0.0430	0.6734	0.6227
Causal Forest	0.9	210	5	200	10	2.5933	2.3127	-0.2805	0.6090	0.0062
Mean weight	-0.9	210	5	200	10	2.5909	2.8563	0.2654	0.6944	0.6185
Random Forest	-0.9	210	5	200	10	2.5909	2.8445	0.2536	0.6846	0.6176
Genetic Matching	-0.9	210	5	200	10	2.5909	3.0962	0.5053	0.8752	0.9043
CSRF	-0.9	210	5	200	10	2.5909	2.8598	0.2690	0.6935	0.6188
Causal Forest	-0.9	210	5	200	10	2.5909	2.3396	-0.2513	0.5858	0.0035

Table A.5: *Summary of estimates of the models from the heterogeneous data simulation with true ATT unknown and* $N = 210$, $N_c = 200$, $N_t = 10$, *and* $p = 10$

Model	M	Size(n)	P	Cont	Trt	ATT	Estimates	Expected Bias	Empirical SE	Theoretical SE
Mean weight	0	210	10	200	10	2.5771	2.3850	-0.1921	0.6747	0.6261
Random Forest	0	210	10	200	10	2.5771	2.3557	-0.2214	0.6709	0.6258
Genetic Matching	0	210	10	200	10	2.5771	2.4275	-0.1496	0.8299	0.8399
CSRF	0	210	10	200	10	2.5771	2.3808	-0.1964	0.6742	0.6260
Causal Forest	0	210	10	200	10	2.5771	1.8319	-0.7452	0.5671	0.0053
Mean weight	0.5	210	10	200	10	2.5808	2.4235	-0.1573	0.6650	0.6293
Random Forest	0.5	210	10	200	10	2.5808	2.3940	-0.1869	0.6613	0.6290
Genetic Matching	0.5	210	10	200	10	2.5808	2.3888	-0.1920	0.7929	0.8226
CSRF	0.5	210	10	200	10	2.5808	2.4183	-0.1625	0.6643	0.6297
Causal Forest	0.5	210	10	200	10	2.5808	1.8815	-0.6994	0.5769	0.0063
Mean weight	-0.5	210	10	200	10	2.5775	2.5167	-0.0608	0.6700	0.6253
Random Forest	-0.5	210	10	200	10	2.5775	2.4943	-0.0832	0.6660	0.6249
Genetic Matching	-0.5	210	10	200	10	2.5775	2.5876	0.0100	0.8602	0.8519
CSRF	-0.5	210	10	200	10	2.5775	2.5156	-0.0619	0.6691	0.6253
Causal Forest	-0.5	210	10	200	10	2.5775	1.9170	-0.6605	0.5612	0.0051
Mean weight	0.9	210	10	200	10	2.5806	2.6107	0.0301	0.6687	0.6309
Random Forest	0.9	210	10	200	10	2.5806	2.5905	0.0098	0.6653	0.6299
Genetic Matching	0.9	210	10	200	10	2.5806	2.5496	-0.0310	0.7683	0.7862
CSRF	0.9	210	10	200	10	2.5806	2.6057	0.0251	0.6696	0.6313
Causal Forest	0.9	210	10	200	10	2.5806	2.2113	-0.3693	0.5968	0.0080
Mean weight	-0.9	210	10	200	10	2.5777	2.8141	0.2365	0.6600	0.6246
Random Forest	-0.9	210	10	200	10	2.5777	2.8071	0.2295	0.6548	0.6244
Genetic Matching	-0.9	210	10	200	10	2.5777	3.1001	0.5225	0.8650	0.8705
CSRF	-0.9	210	10	200	10	2.5777	2.8146	0.2370	0.6592	0.6247
Causal Forest	-0.9	210	10	200	10	2.5777	2.2313	-0.3463	0.5638	0.0050

Table A.6: *Summary of estimates of the models from the heterogeneous data simulation with true ATT unknown and* $N = 210$, $N_c = 200$, $N_t = 10$, *and* $p = 50$

Model	M	Size(n)	P	Cont	Trt	ATT	Estimates	Expected Bias	Empirical SE	Theoretical SE
Mean weight	0	210	10	200	10	2.5771	2.3850	-0.1921	0.6747	0.6261
Random Forest	0	210	10	200	10	2.5771	2.3557	-0.2214	0.6709	0.6258
Genetic Matching	0	210	10	200	10	2.5771	2.4275	-0.1496	0.8299	0.8399
CSRF	0	210	10	200	10	2.5771	2.3808	-0.1964	0.6742	0.6260
Causal Forest	0	210	10	200	10	2.5771	1.8319	-0.7452	0.5671	0.0053
Mean weight	0.5	210	10	200	10	2.5808	2.4235	-0.1573	0.6650	0.6293
Random Forest	0.5	210	10	200	10	2.5808	2.3940	-0.1869	0.6613	0.6290
Genetic Matching	0.5	210	10	200	10	2.5808	2.3888	-0.1920	0.7929	0.8226
CSRF	0.5	210	10	200	10	2.5808	2.4183	-0.1625	0.6643	0.6297
Causal Forest	0.5	210	10	200	10	2.5808	1.8815	-0.6994	0.5769	0.0063
Mean weight	-0.5	210	10	200	10	2.5775	2.5167	-0.0608	0.6700	0.6253
Random Forest	-0.5	210	10	200	10	2.5775	2.4943	-0.0832	0.6660	0.6249
Genetic Matching	-0.5	210	10	200	10	2.5775	2.5876	0.0100	0.8602	0.8519
CSRF	-0.5	210	10	200	10	2.5775	2.5156	-0.0619	0.6691	0.6253
Causal Forest	-0.5	210	10	200	10	2.5775	1.9170	-0.6605	0.5612	0.0051
Mean weight	0.9	210	10	200	10	2.5806	2.6107	0.0301	0.6687	0.6309
Random Forest	0.9	210	10	200	10	2.5806	2.5905	0.0098	0.6653	0.6299
Genetic Matching	0.9	210	10	200	10	2.5806	2.5496	-0.0310	0.7683	0.7862
CSRF	0.9	210	10	200	10	2.5806	2.6057	0.0251	0.6696	0.6313
Causal Forest	0.9	210	10	200	10	2.5806	2.2113	-0.3693	0.5968	0.0080
Mean weight	-0.9	210	10	200	10	2.5777	2.8141	0.2365	0.6600	0.6246
Random Forest	-0.9	210	10	200	10	2.5777	2.8071	0.2295	0.6548	0.6244
Genetic Matching	-0.9	210	10	200	10	2.5777	3.1001	0.5225	0.8650	0.8705
CSRF	-0.9	210	10	200	10	2.5777	2.8146	0.2370	0.6592	0.6247
Causal Forest	-0.9	210	10	200	10	2.5777	2.2313	-0.3463	0.5638	0.0050

Summary of estimates of the models from homogeneous and heterogeneous with true ATT as 4 when the sample size is increased but fraction of treated unit remain the same.

Table A.7: *Summary of estimates of the models from the homogeneous data simulation with true ATE as* ***4.0*** *and* $N = 410$, $N_c = 400$, $N_t = 10$, *and* $p = 10$

Model	M	Size(n)	P	Cont	Trt	Estimates	Expected Bias	Empirical SE	Theoretical SE
Mean weight	0	410	10	400	10	3.8655	-0.1345	0.3777	0.3204
Random Forest	0	410	10	400	10	3.8376	-0.1624	0.3695	0.3196
Genetic Matching	0	410	10	400	10	3.8735	-0.1265	0.6226	0.5409
CSRF	0	410	10	400	10	3.8619	-0.1381	0.3761	0.3205
Causal Forest	0	410	10	400	10	3.1937	-0.8063	0.3143	0.0066
Mean weight	0.5	410	10	400	10	3.9111	-0.0889	0.3799	0.3253
Random Forest	0.5	410	10	400	10	3.8831	-0.1169	0.3719	0.3245
Genetic Matching	0.5	410	10	400	10	3.8788	-0.1212	0.5855	0.5172
CSRF	0.5	410	10	400	10	3.9052	-0.0948	0.3772	0.3253
Causal Forest	0.5	410	10	400	10	3.2905	-0.7095	0.3298	0.0077
Mean weight	-0.5	410	10	400	10	3.9892	-0.0108	0.3537	0.3157
Random Forest	-0.5	410	10	400	10	3.9672	-0.0328	0.3451	0.3150
Genetic Matching	-0.5	410	10	400	10	4.0688	0.0688	0.6179	0.5436
CSRF	-0.5	410	10	400	10	3.9880	-0.0120	0.3519	0.3164
Causal Forest	-0.5	410	10	400	10	3.2749	-0.7251	0.2946	0.0060
Mean weight	0.9	410	10	400	10	4.0210	0.0210	0.3694	0.3246
Random Forest	0.9	410	10	400	10	4.0011	0.0011	0.3634	0.3234
Genetic Matching	0.9	410	10	400	10	3.9456	-0.0544	0.5427	0.4904
CSRF	0.9	410	10	400	10	4.0172	0.0172	0.3699	0.3252
Causal Forest	0.9	410	10	400	10	3.6045	-0.3955	0.3236	0.0102
Mean weight	-0.9	410	10	400	10	4.2279	0.2279	0.3546	0.3168
Random Forest	-0.9	410	10	400	10	4.2193	0.2193	0.3503	0.3163
Genetic Matching	-0.9	410	10	400	10	4.5415	0.5415	0.6270	0.5445
CSRF	-0.9	410	10	400	10	4.2294	0.2294	0.3550	0.3175
Causal Forest	-0.9	410	10	400	10	3.4802	-0.5198	0.3151	0.0064

Table A.8: *Summary of estimates of the models from the homogeneous data simulation with true ATE as* ***4.0*** *and* $N = 1025$, $N_c = 1000$, $N_t = 25$, *and* $p = 10$

Model	M	Size(n)	P	Cont	Trt	Estimates	Expected Bias	Empirical SE	Theoretical SE
Mean weight	0	1025	10	1000	25	3.9626	-0.0374	0.2343	0.2041
Random Forest	0	1025	10	1000	25	3.9387	-0.0613	0.2282	0.2036
Genetic Matching	0	1025	10	1000	25	3.9217	-0.0783	0.4759	0.4281
CSRF	0	1025	10	1000	25	3.9553	-0.0447	0.2329	0.2043
Causal Forest	0	1025	10	1000	25	3.2402	-0.7598	0.2035	0.0015
Mean weight	0.5	1025	10	1000	25	3.9692	-0.0308	0.2360	0.2059
Random Forest	0.5	1025	10	1000	25	3.9492	-0.0508	0.2308	0.2053
Genetic Matching	0.5	1025	10	1000	25	3.9104	-0.0896	0.4137	0.4019
CSRF	0.5	1025	10	1000	25	3.9628	-0.0372	0.2365	0.2061
Causal Forest	0.5	1025	10	1000	25	3.3301	-0.6699	0.2062	0.0019
Mean weight	-0.5	1025	10	1000	25	4.0384	0.0384	0.2442	0.2023
Random Forest	-0.5	1025	10	1000	25	4.0209	0.0209	0.2362	0.2017
Genetic Matching	-0.5	1025	10	1000	25	4.0711	0.0711	0.4915	0.4339
CSRF	-0.5	1025	10	1000	25	4.0347	0.0347	0.2430	0.2026
Causal Forest	-0.5	1025	10	1000	25	3.4436	-0.5564	0.2186	0.0028
Mean weight	0.9	1025	10	1000	25	4.0242	0.0242	0.2391	0.2055
Random Forest	0.9	1025	10	1000	25	4.0059	0.0059	0.2342	0.2050
Genetic Matching	0.9	1025	10	1000	25	3.9751	-0.0249	0.3752	0.3604
CSRF	0.9	1025	10	1000	25	4.0227	0.0227	0.2388	0.2059
Causal Forest	0.9	1025	10	1000	25	3.6035	-0.3965	0.2068	0.0036
Mean weight	-0.9	1025	10	1000	25	4.2355	0.2355	0.2388	0.2032
Random Forest	-0.9	1025	10	1000	25	4.2305	0.2305	0.2344	0.2028
Genetic Matching	-0.9	1025	10	1000	25	4.5706	0.5706	0.5358	0.4704
CSRF	-0.9	1025	10	1000	25	4.2382	0.2382	0.2400	0.2034
Causal Forest	-0.9	1025	10	1000	25	3.6332	-0.3668	0.2635	0.0073

Table A.9: *Summary of estimates of the models from the homogeneous data simulation with true ATE as* ***4.0*** *and* $N = 2050$, $N_c = 2000$, $N_t = 50$, *and* $p = 10$

Model	M	Size(n)	P	Cont	Trt	Estimates	Expected Bias	Empirical SE	Theoretical SE
Mean weight	0	2050	10	2000	50	3.9911	-0.0089	0.1741	0.1435
Random Forest	0	2050	10	2000	50	3.9728	-0.0272	0.1669	0.1431
Genetic Matching	0	2050	10	2000	50	3.9721	-0.0279	0.3933	0.3635
CSRF	0	2050	10	2000	50	3.9832	-0.0168	0.1727	0.1439
Causal Forest	0	2050	10	2000	50	3.6032	-0.3968	0.1803	0.0089
Mean weight	0.5	2050	10	2000	50	4.0024	0.0024	0.1721	0.1451
Random Forest	0.5	2050	10	2000	50	3.9858	-0.0142	0.1668	0.1446
Genetic Matching	0.5	2050	10	2000	50	3.9382	-0.0618	0.3356	0.3351
CSRF	0.5	2050	10	2000	50	3.9945	-0.0055	0.1715	0.1454
Causal Forest	0.5	2050	10	2000	50	3.6392	-0.3608	0.1710	0.0096
Mean weight	-0.5	2050	10	2000	50	4.0629	0.0629	0.1801	0.1431
Random Forest	-0.5	2050	10	2000	50	4.0482	0.0482	0.1724	0.1427
Genetic Matching	-0.5	2050	10	2000	50	4.1230	0.1230	0.4257	0.3852
CSRF	-0.5	2050	10	2000	50	4.0586	0.0586	0.1791	0.1433
Causal Forest	-0.5	2050	10	2000	50	3.8982	-0.1018	0.2108	0.0137
Mean weight	0.9	2050	10	2000	50	4.0156	0.0156	0.1650	0.1448
Random Forest	0.9	2050	10	2000	50	4.0009	0.0009	0.1615	0.1445
Genetic Matching	0.9	2050	10	2000	50	3.9771	-0.0229	0.2752	0.2846
CSRF	0.9	2050	10	2000	50	4.0140	0.0140	0.1652	0.1452
Causal Forest	0.9	2050	10	2000	50	3.7682	-0.2318	0.1594	0.0142
Mean weight	-0.9	2050	10	2000	50	4.2245	0.2245	0.1839	0.1436
Random Forest	-0.9	2050	10	2000	50	4.2246	0.2246	0.1797	0.1433
Genetic Matching	-0.9	2050	10	2000	50	4.5673	0.5673	0.4662	0.4165
CSRF	-0.9	2050	10	2000	50	4.2277	0.2277	0.1846	0.1439
Causal Forest	-0.9	2050	10	2000	50	3.5566	-0.4434	0.2380	0.0347

Table A.10: *Summary of estimates of the models from the heterogeneous data simulation with true ATT unknown and* $N = 410$, $N_c = 400$, $N_t = 10$, *and* $p = 10$

Model	M	Size(n)	P	Cont	Trt	ATT	Estimates	Expected Bias	Empirical SE	Theoretical SE
Mean weight	0	410	10	400	10	2.5677	2.4331	-0.1345	0.6685	0.6185
Random Forest	0	410	10	400	10	2.5677	2.4052	-0.1624	0.6629	0.6181
Genetic Matching	0	410	10	400	10	2.5677	2.4411	-0.1265	0.8355	0.8172
CSRF	0	410	10	400	10	2.5677	2.4295	-0.1381	0.6666	0.6186
Causal Forest	0	410	10	400	10	2.5677	1.9290	-0.6386	0.5683	0.0126
Mean weight	0.5	410	10	400	10	2.5678	2.4793	-0.0885	0.6611	0.6190
Random Forest	0.5	410	10	400	10	2.5678	2.4509	-0.1169	0.6564	0.6186
Genetic Matching	0.5	410	10	400	10	2.5678	2.4466	-0.1212	0.8036	0.7810
CSRF	0.5	410	10	400	10	2.5678	2.4731	-0.0947	0.6584	0.6190
Causal Forest	0.5	410	10	400	10	2.5678	2.0097	-0.5580	0.5809	0.0141
Mean weight	-0.5	410	10	400	10	2.5681	2.5573	-0.0108	0.6482	0.6174
Random Forest	-0.5	410	10	400	10	2.5681	2.5353	-0.0328	0.6433	0.6171
Genetic Matching	-0.5	410	10	400	10	2.5681	2.6369	0.0688	0.8226	0.8292
CSRF	-0.5	410	10	400	10	2.5681	2.5561	-0.0120	0.6468	0.6180
Causal Forest	-0.5	410	10	400	10	2.5681	2.0262	-0.5419	0.5514	0.0117
Mean weight	0.9	410	10	400	10	2.5678	2.5838	0.0160	0.6498	0.6211
Random Forest	0.9	410	10	400	10	2.5678	2.5650	-0.0028	0.6484	0.6207
Genetic Matching	0.9	410	10	400	10	2.5678	2.5333	-0.0345	0.7657	0.7540
CSRF	0.9	410	10	400	10	2.5678	2.5820	0.0142	0.6500	0.6219
Causal Forest	0.9	410	10	400	10	2.5678	2.2774	-0.2903	0.5947	0.0185
Mean weight	-0.9	410	10	400	10	2.5692	2.7971	0.2279	0.6471	0.6175
Random Forest	-0.9	410	10	400	10	2.5692	2.7885	0.2193	0.6456	0.6173
Genetic Matching	-0.9	410	10	400	10	2.5692	3.1106	0.5415	0.8222	0.8387
CSRF	-0.9	410	10	400	10	2.5692	2.7986	0.2294	0.6476	0.6179
Causal Forest	-0.9	410	10	400	10	2.5692	2.2736	-0.2956	0.5500	0.0112

Table A.11: *Summary of estimates of the models from the heterogeneous data simulation with true ATT unknown and* $N = 1025$, $N_c = 1000$, $N_t = 25$, *and* $p = 10$

Model	M	Size(n)	P	Cont	Trt	ATT	Estimates	Expected Bias	Empirical SE	Theoretical SE
Mean weight	0	1025	10	1000	25	2.5729	2.5355	-0.0374	0.4115	0.3980
Random Forest	0	1025	10	1000	25	2.5729	2.5116	-0.0613	0.4285	0.3977
Genetic Matching	0	1025	10	1000	25	2.5729	2.4946	-0.0783	0.5871	0.6495
CSRF	0	1025	10	1000	25	2.5729	2.5282	-0.0447	0.4205	0.3982
Causal Forest	0	1025	10	1000	25	2.5729	2.0163	-0.5566	0.3569	0.0028
Mean weight	0.5	1025	10	1000	25	2.5753	2.5451	-0.0301	0.4119	0.3981
Random Forest	0.5	1025	10	1000	25	2.5753	2.5244	-0.0509	0.4291	0.3978
Genetic Matching	0.5	1025	10	1000	25	2.5753	2.4856	-0.0896	0.5251	0.6089
CSRF	0.5	1025	10	1000	25	2.5753	2.5385	-0.0367	0.4223	0.3984
Causal Forest	0.5	1025	10	1000	25	2.5753	2.0741	-0.5012	0.3591	0.0036
Mean weight	-0.5	1025	10	1000	25	2.5761	2.6146	0.0384	0.4217	0.3979
Random Forest	-0.5	1025	10	1000	25	2.5761	2.5971	0.0209	0.4367	0.3975
Genetic Matching	-0.5	1025	10	1000	25	2.5761	2.6472	0.0711	0.6049	0.6688
CSRF	-0.5	1025	10	1000	25	2.5761	2.6108	0.0347	0.4285	0.3981
Causal Forest	-0.5	1025	10	1000	25	2.5761	2.1993	-0.3768	0.3789	0.0054
Mean weight	0.9	1025	10	1000	25	2.5759	2.5937	0.0178	0.4096	0.3994
Random Forest	0.9	1025	10	1000	25	2.5759	2.5768	0.0008	0.4176	0.3991
Genetic Matching	0.9	1025	10	1000	25	2.5759	2.5474	-0.0285	0.4925	0.5533
CSRF	0.9	1025	10	1000	25	2.5759	2.5924	0.0165	0.4101	0.3998
Causal Forest	0.9	1025	10	1000	25	2.5759	2.2788	-0.2972	0.3714	0.0070
Mean weight	-0.9	1025	10	1000	25	2.5738	2.8093	0.2355	0.4158	0.3977
Random Forest	-0.9	1025	10	1000	25	2.5738	2.8043	0.2305	0.4241	0.3975
Genetic Matching	-0.9	1025	10	1000	25	2.5738	3.1444	0.5706	0.6389	0.7238
CSRF	-0.9	1025	10	1000	25	2.5738	2.8120	0.2382	0.4174	0.3978
Causal Forest	-0.9	1025	10	1000	25	2.5738	2.4344	-0.1394	0.4198	0.0114

Table A.12: *Summary of estimates of the models from the heterogeneous data simulation with true ATT unknown and* $N = 2050$, $N_c = 2000$, $N_t = 50$, *and* $p = 10$

Model	M	Size(n)	P	Cont	Trt	ATT	Estimates	Expected Bias	Empirical SE	Theoretical SE
Mean weight	0	2050	10	2000	50	2.5793	2.5704	-0.0089	0.2979	0.2818
Random Forest	0	2050	10	2000	50	2.5793	2.5521	-0.0272	0.2933	0.2816
Genetic Matching	0	2050	10	2000	50	2.5793	2.5513	-0.0279	0.4565	0.5547
CSRF	0	2050	10	2000	50	2.5793	2.5625	-0.0168	0.2965	0.2820
Causal Forest	0	2050	10	2000	50	2.5793	2.3095	-0.2698	0.2942	0.0167
Mean weight	0.5	2050	10	2000	50	2.5804	2.5827	0.0024	0.2942	0.2824
Random Forest	0.5	2050	10	2000	50	2.5804	2.5661	-0.0142	0.2920	0.2822
Genetic Matching	0.5	2050	10	2000	50	2.5804	2.5186	-0.0618	0.4037	0.5083
CSRF	0.5	2050	10	2000	50	2.5804	2.5749	-0.0055	0.2937	0.2827
Causal Forest	0.5	2050	10	2000	50	2.5804	2.3141	-0.2663	0.2856	0.0181
Mean weight	-0.5	2050	10	2000	50	2.5782	2.6411	0.0629	0.3032	0.2818
Random Forest	-0.5	2050	10	2000	50	2.5782	2.6264	0.0482	0.2984	0.2816
Genetic Matching	-0.5	2050	10	2000	50	2.5782	2.7012	0.1230	0.4931	0.5926
CSRF	-0.5	2050	10	2000	50	2.5782	2.6368	0.0586	0.3024	0.2819
Causal Forest	-0.5	2050	10	2000	50	2.5782	2.5618	-0.0164	0.3360	0.0237
Mean weight	0.9	2050	10	2000	50	2.5807	2.6001	0.0194	0.2938	0.2832
Random Forest	0.9	2050	10	2000	50	2.5807	2.5856	0.0049	0.2930	0.2830
Genetic Matching	0.9	2050	10	2000	50	2.5807	2.5780	-0.0027	0.3668	0.4402
CSRF	0.9	2050	10	2000	50	2.5807	2.5988	0.0181	0.2944	0.2834
Causal Forest	0.9	2050	10	2000	50	2.5807	2.4069	-0.1738	0.2837	0.0265
Mean weight	-0.9	2050	10	2000	50	2.5790	2.8059	0.2268	0.3045	0.2815
Random Forest	-0.9	2050	10	2000	50	2.5790	2.8062	0.2272	0.3008	0.2813
Genetic Matching	-0.9	2050	10	2000	50	2.5790	3.1352	0.5562	0.5039	0.6457
CSRF	-0.9	2050	10	2000	50	2.5790	2.8099	0.2309	0.3045	0.2815
Causal Forest	-0.9	2050	10	2000	50	2.5790	2.4557	-0.1233	0.3665	0.0459

Summary of estimates of the models from the heterogeneous with true ATT as 4 when the number of control unit is increase and treated unit remain the same.

Table A.13: *Summary of estimates of the models from homogeneous data simulation with true ATE as* ***4.0*** *and* $N = 210$, $N_c = 200$, $N_t = 10$, *and* $p = 10$

Model	M	Size(n)	P	Cont	Trt	Estimates	Expected Bias	Empirical SE	Theoretical SE
Mean weight	0	210	10	200	10	3.8079	-0.1921	0.3993	0.3259
Random Forest	0	210	10	200	10	3.7786	-0.2214	0.3921	0.3253
Genetic Matching	0	210	10	200	10	3.8504	-0.1496	0.6309	0.5529
CSRF	0	210	10	200	10	3.8036	-0.1964	0.3987	0.3258
Causal Forest	0	210	10	200	10	3.0799	-0.9201	0.3228	0.0028
Mean weight	0.5	210	10	200	10	3.8429	-0.1571	0.3919	0.3363
Random Forest	0.5	210	10	200	10	3.8131	-0.1869	0.3857	0.3355
Genetic Matching	0.5	210	10	200	10	3.8080	-0.1920	0.5871	0.5497
CSRF	0.5	210	10	200	10	3.8374	-0.1626	0.3912	0.3361
Causal Forest	0.5	210	10	200	10	3.1432	-0.8568	0.3362	0.0036
Mean weight	-0.5	210	10	200	10	3.9392	-0.0608	0.3837	0.3197
Random Forest	-0.5	210	10	200	10	3.9168	-0.0832	0.3765	0.3190
Genetic Matching	-0.5	210	10	200	10	4.0100	0.0100	0.6534	0.5529
CSRF	-0.5	210	10	200	10	3.9381	-0.0619	0.3829	0.3198
Causal Forest	-0.5	210	10	200	10	3.1416	-0.8584	0.3117	0.0027
Mean weight	0.9	210	10	200	10	4.0298	0.0298	0.3845	0.3314
Random Forest	0.9	210	10	200	10	4.0088	0.0088	0.3795	0.3306
Genetic Matching	0.9	210	10	200	10	3.9573	-0.0427	0.5282	0.5065
CSRF	0.9	210	10	200	10	4.0231	0.0231	0.3854	0.3319
Causal Forest	0.9	210	10	200	10	3.5176	-0.4824	0.3354	0.0045
Mean weight	-0.9	210	10	200	10	4.2365	0.2365	0.3763	0.3198
Random Forest	-0.9	210	10	200	10	4.2295	0.2295	0.3680	0.3193
Genetic Matching	-0.9	210	10	200	10	4.5225	0.5225	0.6665	0.5626
CSRF	-0.9	210	10	200	10	4.2370	0.2370	0.3758	0.3202
Causal Forest	-0.9	210	10	200	10	3.4232	-0.5768	0.3146	0.0028

Table A.14: *Summary of estimates of the models from homogeneous data simulation with true ATE as* ***4.0*** *and* $N = 410$, $N_c = 400$, $N_t = 10$, *and* $p = 10$

Model	M	Size(n)	P	Cont	Trt	Estimate	Expected Bias	Empirical SE	Theoretical SE
Mean weight	0	410	10	400	10	3.8655	-0.1345	0.3777	0.3204
Random Forest	0	410	10	400	10	3.8376	-0.1624	0.3695	0.3196
Genetic Matching	0	410	10	400	10	3.8735	-0.1265	0.6226	0.5409
CSRF	0	410	10	400	10	3.8619	-0.1381	0.3761	0.3205
Causal Forest	0	410	10	400	10	3.1937	-0.8063	0.3143	0.0066
Mean weight	0.5	410	10	400	10	3.9111	-0.0889	0.3799	0.3253
Random Forest	0.5	410	10	400	10	3.8831	-0.1169	0.3719	0.3245
Genetic Matching	0.5	410	10	400	10	3.8788	-0.1212	0.5855	0.5172
CSRF	0.5	410	10	400	10	3.9052	-0.0948	0.3772	0.3253
Causal Forest	0.5	410	10	400	10	3.2905	-0.7095	0.3298	0.0077
Mean weight	-0.5	410	10	400	10	3.9892	-0.0108	0.3537	0.3157
Random Forest	-0.5	410	10	400	10	3.9672	-0.0328	0.3451	0.3150
Genetic Matching	-0.5	410	10	400	10	4.0688	0.0688	0.6179	0.5436
CSRF	-0.5	410	10	400	10	3.9880	-0.0120	0.3519	0.3164
Causal Forest	-0.5	410	10	400	10	3.2749	-0.7251	0.2946	0.0060
Mean weight	0.9	410	10	400	10	4.0210	0.0210	0.3694	0.3246
Random Forest	0.9	410	10	400	10	4.0011	0.0011	0.3634	0.3234
Genetic Matching	0.9	410	10	400	10	3.9456	-0.0544	0.5427	0.4904
CSRF	0.9	410	10	400	10	4.0172	0.0172	0.3699	0.3252
Causal Forest	0.9	410	10	400	10	3.6045	-0.3955	0.3236	0.0102
Mean weight	-0.9	410	10	400	10	4.2279	0.2279	0.3546	0.3168
Random Forest	-0.9	410	10	400	10	4.2193	0.2193	0.3503	0.3163
Genetic Matching	-0.9	410	10	400	10	4.5415	0.5415	0.6270	0.5445
CSRF	-0.9	410	10	400	10	4.2294	0.2294	0.3550	0.3175
Causal Forest	-0.9	410	10	400	10	3.4802	-0.5198	0.3151	0.0064

Table A.15: *Summary of estimates of the models from homogeneous data simulation with true ATE as* ***4.0*** *and* $N = 525$, $N_c = 500$, $N_t = 25$, *and* $p = 10$

Model	M	Size(n)	P	Cont	Trt	Estimate	Expected Bias	Empirical SE	Theoretical SE
Mean weight	0	525	10	500	25	3.9114	-0.0886	0.2578	0.2052
Random Forest	0	525	10	500	25	3.8827	-0.1173	0.2521	0.2048
Genetic Matching	0	525	10	500	25	3.8744	-0.1256	0.4898	0.4499
CSRF	0	525	10	500	25	3.9069	-0.0931	0.2574	0.2054
Causal Forest	0	525	10	500	25	3.1000	-0.9000	0.2122	0.0017
Mean weight	0.5	525	10	500	25	3.9325	-0.0675	0.2582	0.2095
Random Forest	0.5	525	10	500	25	3.9077	-0.0923	0.2508	0.2092
Genetic Matching	0.5	525	10	500	25	3.8824	-0.1176	0.4403	0.4262
CSRF	0.5	525	10	500	25	3.9255	-0.0745	0.2587	0.2097
Causal Forest	0.5	525	10	500	25	3.2120	-0.7880	0.2240	0.0023
Mean weight	-0.5	525	10	500	25	4.0140	0.0140	0.2679	0.2032
Random Forest	-0.5	525	10	500	25	3.9940	-0.0060	0.2583	0.2027
Genetic Matching	-0.5	525	10	500	25	4.0877	0.0877	0.5152	0.4602
CSRF	-0.5	525	10	500	25	4.0110	0.0110	0.2663	0.2034
Causal Forest	-0.5	525	10	500	25	3.2845	-0.7155	0.2325	0.0023
Mean weight	0.9	525	10	500	25	4.0338	0.0338	0.2595	0.2090
Random Forest	0.9	525	10	500	25	4.0145	0.0145	0.2542	0.2084
Genetic Matching	0.9	525	10	500	25	3.9901	-0.0099	0.3883	0.3801
CSRF	0.9	525	10	500	25	4.0312	0.0312	0.2595	0.2094
Causal Forest	0.9	525	10	500	25	3.5817	-0.4183	0.2298	0.0039
Mean weight	-0.9	525	10	500	25	4.2464	0.2464	0.2717	0.2045
Random Forest	-0.9	525	10	500	25	4.2390	0.2390	0.2631	0.2042
Genetic Matching	-0.9	525	10	500	25	4.5651	0.5651	0.5731	0.4867
CSRF	-0.9	525	10	500	25	4.2464	0.2464	0.2704	0.2048
Causal Forest	-0.9	525	10	500	25	3.5845	-0.4155	0.2788	0.0043

Table A.16: *Summary of estimates of the models from homogeneous data simulation with true ATE as* ***4.0*** *and* $N = 1025$, $N_c = 1000$, $N_t = 25$, *and* $p = 10$

Model	M	Size(n)	P	Cont	Trt	Estimate	Expected Bias	Empirical SE	Theoretical SE
Mean weight	0	1025	10	1000	25	3.9626	-0.0374	0.2343	0.2041
Random Forest	0	1025	10	1000	25	3.9387	-0.0613	0.2282	0.2036
Genetic Matching	0	1025	10	1000	25	3.9217	-0.0783	0.4759	0.4281
CSRF	0	1025	10	1000	25	3.9553	-0.0447	0.2329	0.2043
Causal Forest	0	1025	10	1000	25	3.2402	-0.7598	0.2035	0.0015
Mean weight	0.5	1025	10	1000	25	3.9692	-0.0308	0.2360	0.2059
Random Forest	0.5	1025	10	1000	25	3.9492	-0.0508	0.2308	0.2053
Genetic Matching	0.5	1025	10	1000	25	3.9104	-0.0896	0.4137	0.4019
CSRF	0.5	1025	10	1000	25	3.9628	-0.0372	0.2365	0.2061
Causal Forest	0.5	1025	10	1000	25	3.3301	-0.6699	0.2062	0.0019
Mean weight	-0.5	1025	10	1000	25	4.0384	0.0384	0.2442	0.2023
Random Forest	-0.5	1025	10	1000	25	4.0209	0.0209	0.2362	0.2017
Genetic Matching	-0.5	1025	10	1000	25	4.0711	0.0711	0.4915	0.4339
CSRF	-0.5	1025	10	1000	25	4.0347	0.0347	0.2430	0.2026
Causal Forest	-0.5	1025	10	1000	25	3.4436	-0.5564	0.2186	0.0028
Mean weight	0.9	1025	10	1000	25	4.0242	0.0242	0.2391	0.2055
Random Forest	0.9	1025	10	1000	25	4.0059	0.0059	0.2342	0.2050
Genetic Matching	0.9	1025	10	1000	25	3.9751	-0.0249	0.3752	0.3604
CSRF	0.9	1025	10	1000	25	4.0227	0.0227	0.2388	0.2059
Causal Forest	0.9	1025	10	1000	25	3.6035	-0.3965	0.2068	0.0036
Mean weight	-0.9	1025	10	1000	25	4.2355	0.2355	0.2388	0.2032
Random Forest	-0.9	1025	10	1000	25	4.2305	0.2305	0.2344	0.2028
Genetic Matching	-0.9	1025	10	1000	25	4.5706	0.5706	0.5358	0.4704
CSRF	-0.9	1025	10	1000	25	4.2382	0.2382	0.2400	0.2034
Causal Forest	-0.9	1025	10	1000	25	3.6332	-0.3668	0.2635	0.0073

Table A.17: *Summary of estimates of the models from the heterogeneous data simulation with true ATT unknown and* $N = 210$, $N_c = 200$, $N_t = 10$, *and* $p = 10$

Model	M	Size(n)	P	Cont	Trt	ATT	Estimates	Expected Bias	Empirical SE	Theoretical SE
Mean weight	0	210	10	200	10	2.5771	2.3850	-0.1921	0.6747	0.6261
Random Forest	0	210	10	200	10	2.5771	2.3557	-0.2214	0.6709	0.6258
Genetic Matching	0	210	10	200	10	2.5771	2.4275	-0.1496	0.8299	0.8399
CSRF	0	210	10	200	10	2.5771	2.3808	-0.1964	0.6742	0.6260
Causal Forest	0	210	10	200	10	2.5771	1.8319	-0.7452	0.5671	0.0053
Mean weight	0.5	210	10	200	10	2.5808	2.4235	-0.1573	0.6650	0.6293
Random Forest	0.5	210	10	200	10	2.5808	2.3940	-0.1869	0.6613	0.6290
Genetic Matching	0.5	210	10	200	10	2.5808	2.3888	-0.1920	0.7929	0.8226
CSRF	0.5	210	10	200	10	2.5808	2.4183	-0.1625	0.6643	0.6297
Causal Forest	0.5	210	10	200	10	2.5808	1.8815	-0.6994	0.5769	0.0063
Mean weight	-0.5	210	10	200	10	2.5775	2.5167	-0.0608	0.6700	0.6253
Random Forest	-0.5	210	10	200	10	2.5775	2.4943	-0.0832	0.6660	0.6249
Genetic Matching	-0.5	210	10	200	10	2.5775	2.5876	0.0100	0.8602	0.8519
CSRF	-0.5	210	10	200	10	2.5775	2.5156	-0.0619	0.6691	0.6253
Causal Forest	-0.5	210	10	200	10	2.5775	1.9170	-0.6605	0.5612	0.0051
Mean weight	0.9	210	10	200	10	2.5806	2.6107	0.0301	0.6687	0.6309
Random Forest	0.9	210	10	200	10	2.5806	2.5905	0.0098	0.6653	0.6299
Genetic Matching	0.9	210	10	200	10	2.5806	2.5496	-0.0310	0.7683	0.7862
CSRF	0.9	210	10	200	10	2.5806	2.6057	0.0251	0.6696	0.6313
Causal Forest	0.9	210	10	200	10	2.5806	2.2113	-0.3693	0.5968	0.0080
Mean weight	-0.9	210	10	200	10	2.5777	2.8141	0.2365	0.6600	0.6246
Random Forest	-0.9	210	10	200	10	2.5777	2.8071	0.2295	0.6548	0.6244
Genetic Matching	-0.9	210	10	200	10	2.5777	3.1001	0.5225	0.8650	0.8705
CSRF	-0.9	210	10	200	10	2.5777	2.8146	0.2370	0.6592	0.6247
Causal Forest	-0.9	210	10	200	10	2.5777	2.2313	-0.3463	0.5638	0.0050

Table A.18: *Summary of estimates of the models from the heterogeneous data simulation with true ATT unknown and* $N = 410$, $N_c = 400$, $N_t = 10$, *and* $p = 10$

Model	M	Size(n)	P	Cont	Trt	ATT	Estimates	Expected Bias	Empirical SE	Theoretical SE
Mean weight	0	410	10	400	10	2.5677	2.4331	-0.1345	0.6685	0.6185
Random Forest	0	410	10	400	10	2.5677	2.4052	-0.1624	0.6629	0.6181
Genetic Matching	0	410	10	400	10	2.5677	2.4411	-0.1265	0.8355	0.8172
CSRF	0	410	10	400	10	2.5677	2.4295	-0.1381	0.6666	0.6186
Causal Forest	0	410	10	400	10	2.5677	1.9290	-0.6386	0.5683	0.0126
Mean weight	0.5	410	10	400	10	2.5678	2.4793	-0.0885	0.6611	0.6190
Random Forest	0.5	410	10	400	10	2.5678	2.4509	-0.1169	0.6564	0.6186
Genetic Matching	0.5	410	10	400	10	2.5678	2.4466	-0.1212	0.8036	0.7810
CSRF	0.5	410	10	400	10	2.5678	2.4731	-0.0947	0.6584	0.6190
Causal Forest	0.5	410	10	400	10	2.5678	2.0097	-0.5580	0.5809	0.0141
Mean weight	-0.5	410	10	400	10	2.5681	2.5573	-0.0108	0.6482	0.6174
Random Forest	-0.5	410	10	400	10	2.5681	2.5353	-0.0328	0.6433	0.6171
Genetic Matching	-0.5	410	10	400	10	2.5681	2.6369	0.0688	0.8226	0.8292
CSRF	-0.5	410	10	400	10	2.5681	2.5561	-0.0120	0.6468	0.6180
Causal Forest	-0.5	410	10	400	10	2.5681	2.0262	-0.5419	0.5514	0.0117
Mean weight	0.9	410	10	400	10	2.5678	2.5838	0.0160	0.6498	0.6211
Random Forest	0.9	410	10	400	10	2.5678	2.5650	-0.0028	0.6484	0.6207
Genetic Matching	0.9	410	10	400	10	2.5678	2.5333	-0.0345	0.7657	0.7540
CSRF	0.9	410	10	400	10	2.5678	2.5820	0.0142	0.6500	0.6219
Causal Forest	0.9	410	10	400	10	2.5678	2.2774	-0.2903	0.5947	0.0185
Mean weight	-0.9	410	10	400	10	2.5692	2.7971	0.2279	0.6471	0.6175
Random Forest	-0.9	410	10	400	10	2.5692	2.7885	0.2193	0.6456	0.6173
Genetic Matching	-0.9	410	10	400	10	2.5692	3.1106	0.5415	0.8222	0.8387
CSRF	-0.9	410	10	400	10	2.5692	2.7986	0.2294	0.6476	0.6179
Causal Forest	-0.9	410	10	400	10	2.5692	2.2736	-0.2956	0.5500	0.0112

Table A.19: *Summary of estimates of the models from the heterogeneous data simulation with true ATT unknown and* $N = 525$, $N_c = 500$, $N_t = 25$, *and* $p = 10$

Model	M	Size(n)	P	Cont	Trt	ATT	Estimates	Expected Bias	Empirical SE	Theoretical SE
Mean weight	0	525	10	500	25	2.5813	2.4928	-0.0886	0.4394	0.3981
Random Forest	0	525	10	500	25	2.5813	2.4640	-0.1173	0.4373	0.3980
Genetic Matching	0	525	10	500	25	2.5813	2.4557	-0.1256	0.6091	0.6833
CSRF	0	525	10	500	25	2.5813	2.4883	-0.0931	0.4396	0.3983
Causal Forest	0	525	10	500	25	2.5813	1.9161	-0.6652	0.3767	0.0032
Mean weight	0.5	525	10	500	25	2.5834	2.5160	-0.0673	0.4391	0.3998
Random Forest	0.5	525	10	500	25	2.5834	2.4910	-0.0924	0.4354	0.3996
Genetic Matching	0.5	525	10	500	25	2.5834	2.4657	-0.1176	0.5661	0.6435
CSRF	0.5	525	10	500	25	2.5834	2.5089	-0.0745	0.4400	0.4001
Causal Forest	0.5	525	10	500	25	2.5834	1.9807	-0.6027	0.3831	0.0041
Mean weight	-0.5	525	10	500	25	2.5840	2.5985	0.0145	0.4509	0.3982
Random Forest	-0.5	525	10	500	25	2.5840	2.5780	-0.0060	0.4445	0.3980
Genetic Matching	-0.5	525	10	500	25	2.5840	2.6717	0.0877	0.6356	0.7051
CSRF	-0.5	525	10	500	25	2.5840	2.5952	0.0112	0.4496	0.3983
Causal Forest	-0.5	525	10	500	25	2.5840	2.0964	-0.4877	0.3978	0.0043
Mean weight	0.9	525	10	500	25	2.5830	2.6100	0.0269	0.4369	0.4010
Random Forest	0.9	525	10	500	25	2.5830	2.5901	0.0070	0.4344	0.4007
Genetic Matching	0.9	525	10	500	25	2.5830	2.5801	-0.0029	0.5398	0.5858
CSRF	0.9	525	10	500	25	2.5830	2.6077	0.0246	0.4374	0.4014
Causal Forest	0.9	525	10	500	25	2.5830	2.2640	-0.3190	0.3989	0.0065
Mean weight	-0.9	525	10	500	25	2.5829	2.8293	0.2464	0.4541	0.3988
Random Forest	-0.9	525	10	500	25	2.5829	2.8218	0.2390	0.4488	0.3987
Genetic Matching	-0.9	525	10	500	25	2.5829	3.1480	0.5651	0.6857	0.7496
CSRF	-0.9	525	10	500	25	2.5829	2.8293	0.2464	0.4534	0.3989
Causal Forest	-0.9	525	10	500	25	2.5829	2.4421	-0.1408	0.4578	0.0064

Table A.20: *Summary of estimates of the models from the heterogeneous data simulation with true ATT unknown and* $N = 1025$, $N_c = 1000$, $N_t = 25$, *and* $p = 10$

Model	M	Size(n)	P	Cont	Trt	ATT	Estimates	Expected Bias	Empirical SE	Theoretical SE
Mean weight	0	1025	10	1000	25	2.5729	2.5355	-0.0374	0.4115	0.3980
Random Forest	0	1025	10	1000	25	2.5729	2.5116	-0.0613	0.4285	0.3977
Genetic Matching	0	1025	10	1000	25	2.5729	2.4946	-0.0783	0.5871	0.6495
CSRF	0	1025	10	1000	25	2.5729	2.5282	-0.0447	0.4205	0.3982
Causal Forest	0	1025	10	1000	25	2.5729	2.0163	-0.5566	0.3569	0.0028
Mean weight	0.5	1025	10	1000	25	2.5753	2.5451	-0.0301	0.4119	0.3981
Random Forest	0.5	1025	10	1000	25	2.5753	2.5244	-0.0509	0.4291	0.3978
Genetic Matching	0.5	1025	10	1000	25	2.5753	2.4856	-0.0896	0.5251	0.6089
CSRF	0.5	1025	10	1000	25	2.5753	2.5385	-0.0367	0.4223	0.3984
Causal Forest	0.5	1025	10	1000	25	2.5753	2.0741	-0.5012	0.3591	0.0036
Mean weight	-0.5	1025	10	1000	25	2.5761	2.6146	0.0384	0.4217	0.3979
Random Forest	-0.5	1025	10	1000	25	2.5761	2.5971	0.0209	0.4367	0.3975
Genetic Matching	-0.5	1025	10	1000	25	2.5761	2.6472	0.0711	0.6049	0.6688
CSRF	-0.5	1025	10	1000	25	2.5761	2.6108	0.0347	0.4285	0.3981
Causal Forest	-0.5	1025	10	1000	25	2.5761	2.1993	-0.3768	0.3789	0.0054
Mean weight	0.9	1025	10	1000	25	2.5759	2.5937	0.0178	0.4096	0.3994
Random Forest	0.9	1025	10	1000	25	2.5759	2.5768	0.0008	0.4176	0.3991
Genetic Matching	0.9	1025	10	1000	25	2.5759	2.5474	-0.0285	0.4925	0.5533
CSRF	0.9	1025	10	1000	25	2.5759	2.5924	0.0165	0.4101	0.3998
Causal Forest	0.9	1025	10	1000	25	2.5759	2.2788	-0.2972	0.3714	0.0070
Mean weight	-0.9	1025	10	1000	25	2.5738	2.8093	0.2355	0.4158	0.3977
Random Forest	-0.9	1025	10	1000	25	2.5738	2.8043	0.2305	0.4241	0.3975
Genetic Matching	-0.9	1025	10	1000	25	2.5738	3.1444	0.5706	0.6389	0.7238
CSRF	-0.9	1025	10	1000	25	2.5738	2.8120	0.2382	0.4174	0.3978
Causal Forest	-0.9	1025	10	1000	25	2.5738	2.4344	-0.1394	0.4198	0.0114

Summary of estimates of the models from the homogeneous and heterogeneous when the treated units are randomly distributed in the covariate space.

Table A.21: *Summary of estimates of the models from homogeneous data simulation when treated units are randomly distributed in the covariate space with true ATE as* ***4.0*** *and* $N = 210$, $N_c = 200$, $N_t = 10$, *and* $p = 5$

Model	M	Size(n)	P	Cont	Trt	Estimate	Expected Bias	Empirical SE	Theoretical SE
Mean weight	0	210	5	200	10	4.1013	0.1013	0.3740	0.3477
Random Forest	0	210	5	200	10	4.1211	0.1211	0.3730	0.3471
Genetic Matching	0	210	5	200	10	4.1815	0.1815	0.5251	0.4733
CSRF	0	210	5	200	10	4.1004	0.1004	0.3762	0.3492
Causal Forest	0	210	5	200	10	4.0560	0.0560	0.3785	0.0053
Mean weight	0.5	210	5	200	10	4.0747	0.0747	0.3730	0.3430
Random Forest	0.5	210	5	200	10	4.0926	0.0926	0.3741	0.3428
Genetic Matching	0.5	210	5	200	10	4.1523	0.1523	0.5026	0.4645
CSRF	0.5	210	5	200	10	4.0799	0.0799	0.3777	0.3444
Causal Forest	0.5	210	5	200	10	4.0710	0.0710	0.3764	0.0064
Mean weight	-0.5	210	5	200	10	4.1314	0.1314	0.3717	0.3418
Random Forest	-0.5	210	5	200	10	4.1484	0.1484	0.3694	0.3410
Genetic Matching	-0.5	210	5	200	10	4.2648	0.2648	0.5267	0.4764
CSRF	-0.5	210	5	200	10	4.1293	0.1293	0.3740	0.3430
Causal Forest	-0.5	210	5	200	10	4.0240	0.0240	0.3691	0.0052
Mean weight	0.9	210	5	200	10	4.0201	0.0201	0.3616	0.3339
Random Forest	0.9	210	5	200	10	4.0340	0.0340	0.3603	0.3321
Genetic Matching	0.9	210	5	200	10	4.0436	0.0436	0.4946	0.4411
CSRF	0.9	210	5	200	10	4.0277	0.0277	0.3663	0.3374
Causal Forest	0.9	210	5	200	10	4.0356	0.0356	0.3618	0.0045
Mean weight	-0.9	210	5	200	10	4.1574	0.1574	0.3812	0.3370
Random Forest	-0.9	210	5	200	10	4.1802	0.1802	0.3726	0.3345
Genetic Matching	-0.9	210	5	200	10	4.3341	0.3341	0.5535	0.4913
CSRF	-0.9	210	5	200	10	4.1504	0.1504	0.3829	0.3387
Causal Forest	-0.9	210	5	200	10	3.8507	-0.1493	0.3609	0.0053

Table A.22: *Summary of estimates of the models from homogeneous data simulation when treated units are randomly distributed in the covariate space with true ATE as* ***4.0*** *and* $N = 210,\ N_c = 200,\ N_t = 10,$ *and* $p = 10$

Model	M	Size(n)	P	Cont	Trt	Estimate	Expected Bias	Empirical SE	Theoretical SE
Mean weight	0	210	10	200	10	4.1288	0.1288	0.3921	0.3650
Random Forest	0	210	10	200	10	4.1425	0.1425	0.3936	0.3648
Genetic Matching	0	210	10	200	10	4.2956	0.2956	0.5534	0.5057
CSRF	0	210	10	200	10	4.1266	0.1266	0.3937	0.3642
Causal Forest	0	210	10	200	10	4.0574	0.0574	0.3868	0.0049
Mean weight	0.5	210	10	200	10	4.0619	0.0619	0.3844	0.3594
Random Forest	0.5	210	10	200	10	4.0795	0.0795	0.3882	0.3596
Genetic Matching	0.5	210	10	200	10	4.1955	0.1955	0.5471	0.5008
CSRF	0.5	210	10	200	10	4.0614	0.0614	0.3868	0.3593
Causal Forest	0.5	210	10	200	10	4.0390	0.0390	0.3961	0.0057
Mean weight	-0.5	210	10	200	10	4.1479	0.1479	0.3833	0.3570
Random Forest	-0.5	210	10	200	10	4.1600	0.1600	0.3845	0.3566
Genetic Matching	-0.5	210	10	200	10	4.3384	0.3384	0.5531	0.5051
CSRF	-0.5	210	10	200	10	4.1425	0.1425	0.3842	0.3560
Causal Forest	-0.5	210	10	200	10	4.0233	0.0233	0.3877	0.0049
Mean weight	0.9	210	10	200	10	4.0177	0.0177	0.3575	0.3446
Random Forest	0.9	210	10	200	10	4.0304	0.0304	0.3579	0.3439
Genetic Matching	0.9	210	10	200	10	4.0747	0.0747	0.4873	0.4677
CSRF	0.9	210	10	200	10	4.0217	0.0217	0.3583	0.3464
Causal Forest	0.9	210	10	200	10	4.0252	0.0252	0.3646	0.0066
Mean weight	-0.9	210	10	200	10	4.1545	0.1545	0.3692	0.3409
Random Forest	-0.9	210	10	200	10	4.1754	0.1754	0.3663	0.3406
Genetic Matching	-0.9	210	10	200	10	4.3722	0.3722	0.5263	0.4935
CSRF	-0.9	210	10	200	10	4.1496	0.1496	0.3696	0.3417
Causal Forest	-0.9	210	10	200	10	3.8540	-0.1460	0.3702	0.0059

Table A.23: *Summary of estimates of the models from the heterogeneous data simulation when treated units are randomly distributed in the covariate space with true ATT unknown and* $N = 210,\ N_c = 200,\ N_t = 10,$ *and* $p = 5$

Model	M	Size(n)	P	Cont	Trt	ATT	Estimates	Expected Bias	Empirical SE	Theoretical SE
Mean weight	0	210	5	200	10	2.5948	2.6961	0.1013	0.6846	0.6476
Random Forest	0	210	5	200	10	2.5948	2.7159	0.1211	0.6809	0.6467
Genetic Matching	0	210	5	200	10	2.5948	2.7763	0.1815	0.7901	0.7157
CSRF	0	210	5	200	10	2.5948	2.6951	0.1004	0.6845	0.6496
Causal Forest	0	210	5	200	10	2.5948	2.6736	0.0788	0.6792	0.0096
Mean weight	0.5	210	5	200	10	2.5946	2.6693	0.0747	0.6924	0.6571
Random Forest	0.5	210	5	200	10	2.5946	2.6872	0.0926	0.6889	0.6557
Genetic Matching	0.5	210	5	200	10	2.5946	2.7468	0.1523	0.7758	0.721
CSRF	0.5	210	5	200	10	2.5946	2.6745	0.0799	0.6936	0.6586
Causal Forest	0.5	210	5	200	10	2.5946	2.6784	0.0839	0.6860	0.0119
Mean weight	-0.5	210	5	200	10	2.5911	2.7225	0.1314	0.6766	0.6372
Random Forest	-0.5	210	5	200	10	2.5911	2.7396	0.1484	0.6743	0.6368
Genetic Matching	-0.5	210	5	200	10	2.5911	2.8560	0.2648	0.7817	0.7142
CSRF	-0.5	210	5	200	10	2.5911	2.7205	0.1293	0.6772	0.6382
Causal Forest	-0.5	210	5	200	10	2.5911	2.6544	0.0632	0.6604	0.0093
Mean weight	0.9	210	5	200	10	2.5980	2.6181	0.0201	0.6997	0.6694
Random Forest	0.9	210	5	200	10	2.5980	2.6320	0.0340	0.6954	0.6669
Genetic Matching	0.9	210	5	200	10	2.5980	2.6416	0.0436	0.7823	0.7163
CSRF	0.9	210	5	200	10	2.5980	2.6257	0.0277	0.7017	0.6719
Causal Forest	0.9	210	5	200	10	2.5980	2.6416	0.0436	0.6924	0.0101
Mean weight	-0.9	210	5	200	10	2.5885	2.7459	0.1574	0.6693	0.6260
Random Forest	-0.9	210	5	200	10	2.5885	2.7686	0.1802	0.6634	0.6244
Genetic Matching	-0.9	210	5	200	10	2.5885	2.9225	0.3341	0.7771	0.7313
CSRF	-0.9	210	5	200	10	2.5885	2.7388	0.1504	0.6703	0.6274
Causal Forest	-0.9	210	5	200	10	2.5885	2.5352	-0.0533	0.6168	0.0084

Table A.24: *Summary of estimates of the models from the heterogeneous data simulation when treated units are randomly distributed in the covariate space with true ATT unknown and* $N = 210$, $N_c = 200$, $N_t = 10$, *and* $p = 10$

Model	M	Size(n)	P	Cont	Trt	ATT	Estimates	Expected Bias	Empirical SE	Theoretical SE
Mean weight	0	210	10	200	10	2.5755	2.7180	0.1425	0.6822	0.6613
Random Forest	0	210	10	200	10	2.5755	2.8711	0.2956	0.7982	0.7440
Genetic Matching	0	210	10	200	10	2.5755	2.8711	0.2956	0.7982	0.7440
CSRF	0	210	10	200	10	2.5755	2.7021	0.1266	0.6833	0.6620
Causal Forest	0	210	10	200	10	2.5755	2.6602	0.0847	0.6695	0.0086
Mean weight	0.5	210	10	200	10	2.5904	2.6523	0.0619	0.6920	0.6664
Random Forest	0.5	210	10	200	10	2.5904	2.6699	0.0795	0.6919	0.6654
Genetic Matching	0.5	210	10	200	10	2.5904	2.7860	0.1955	0.8190	0.7446
CSRF	0.5	210	10	200	10	2.5904	2.6519	0.0614	0.6928	0.6674
Causal Forest	0.5	210	10	200	10	2.5904	2.6479	0.0574	0.6895	0.0099
Mean weight	-0.5	210	10	200	10	2.5771	2.7250	0.1479	0.6753	0.6527
Random Forest	-0.5	210	10	200	10	2.5771	2.7371	0.1600	0.6755	0.6523
Genetic Matching	-0.5	210	10	200	10	2.5771	2.9155	0.3384	0.7803	0.7363
CSRF	-0.5	210	10	200	10	2.5771	2.7197	0.1425	0.6755	0.6524
Causal Forest	-0.5	210	10	200	10	2.5771	2.6414	0.0642	0.6643	0.0084
Mean weight	0.9	210	10	200	10	2.5883	2.6059	0.0177	0.7005	0.6755
Random Forest	0.9	210	10	200	10	2.5883	2.6187	0.0304	0.6978	0.6739
Genetic Matching	0.9	210	10	200	10	2.5883	2.6630	0.0747	0.7819	0.7314
CSRF	0.9	210	10	200	10	2.5883	2.6100	0.0217	0.6991	0.6773
Causal Forest	0.9	210	10	200	10	2.5883	2.6190	0.0308	0.6988	0.0140
Mean weight	-0.9	210	10	200	10	2.5777	2.7322	0.1545	0.6574	0.6349
Random Forest	-0.9	210	10	200	10	2.5777	2.7531	0.1754	0.6548	0.6345
Genetic Matching	-0.9	210	10	200	10	2.5777	2.9499	0.3722	0.7547	0.7243
CSRF	-0.9	210	10	200	10	2.5777	2.7273	0.1496	0.6575	0.6354
Causal Forest	-0.9	210	10	200	10	2.5777	2.5216	-0.0561	0.6212	0.0095

Appendix B

Some results for homogeneous and heterogeneous data simulation

Table B.1: *Summary of estimates of the models from homogeneous data simulation with true ATE as* ***4.0*** *and* $N = 410$, $N_c = 400$, $N_t = 10$, *and* $p = 5$

Model	M	Size(n)	P	Cont	Trt	Estimate	Expected Bias	Empirical SE	Theoretical SE
Mean weight	0	410	5	400	10	3.9874	-0.0126	0.3752	0.3224
Random Forest	0	410	5	400	10	3.9550	-0.0500	0.3672	0.3213
Genetic Matching	0	410	5	400	10	3.9926	-0.0074	0.5685	0.5068
CSRF	0	410	5	400	10	3.9832	-0.0168	0.3747	0.3229
Causal Forest	0	410	5	400	10	3.3431	-0.6569	0.3184	0.0075
Mean weight	0.05	410	5	400	10	3.9863	-0.0137	0.3815	0.3234
Random Forest	0.05	410	5	400	10	3.9535	-0.0465	0.3723	0.3223
Genetic Matching	0.05	410	5	400	10	4.0045	0.0045	0.5750	0.5074
CSRF	0.05	410	5	400	10	3.9831	-0.0169	0.3807	0.3239
Causal Forest	0.05	410	5	400	10	3.3446	-0.6554	0.3238	0.0075
Mean weight	-0.05	410	5	400	10	3.9889	-0.0111	0.3755	0.3244
Random Forest	-0.05	410	5	400	10	3.9561	-0.0439	0.3669	0.3229
Genetic Matching	-0.05	410	5	400	10	3.9904	-0.0096	0.5596	0.5213
CSRF	-0.05	410	5	400	10	3.9858	-0.0142	0.3734	0.3248
Causal Forest	-0.05	410	5	400	10	3.3389	-0.6611	0.3178	0.0074
Mean weight	0.25	410	5	400	10	3.9755	-0.0245	0.3776	0.3262
Random Forest	0.25	410	5	400	10	3.9449	-0.0551	0.3708	0.3247
Genetic Matching	0.25	410	5	400	10	3.9601	-0.0399	0.5533	0.5041
CSRF	0.25	410	5	400	10	3.9722	-0.0278	0.3775	0.3268
Causal Forest	0.25	410	5	400	10	3.3622	-0.6378	0.3270	0.0079
Mean weight	-0.25	410	5	400	10	4.0176	0.0176	0.3774	0.3232
Random Forest	-0.25	410	5	400	10	3.9880	-0.0120	0.3687	0.3217
Genetic Matching	-0.25	410	5	400	10	4.0155	0.0155	0.5595	0.5250
CSRF	-0.25	410	5	400	10	4.0146	0.0146	0.3779	0.3237
Causal Forest	-0.25	410	5	400	10	3.3591	-0.6409	0.3169	0.0071
Mean weight	0.5	410	5	400	10	3.9961	-0.0039	0.3723	0.3278
Random Forest	0.5	410	5	400	10	3.9676	-0.0324	0.3658	0.3263
Genetic Matching	0.5	410	5	400	10	3.9916	-0.0084	0.5322	0.5004
CSRF	0.5	410	5	400	10	3.9917	-0.0083	0.3721	0.3281
Causal Forest	0.5	410	5	400	10	3.4196	-0.5804	0.3250	0.0087
Mean weight	-0.5	410	5	400	10	4.0766	0.0766	0.3704	0.3211
Random Forest	-0.5	410	5	400	10	4.0559	0.0559	0.3615	0.3197
Genetic Matching	-0.5	410	5	400	10	4.1185	0.1185	0.5822	0.5389
CSRF	-0.5	410	5	400	10	4.0756	0.0756	0.3714	0.3218
Causal Forest	-0.5	410	5	400	10	3.4282	-0.5718	0.3209	0.0065
Mean weight	0.9	410	5	400	10	4.0479	0.0479	0.3657	0.3268
Random Forest	0.9	410	5	400	10	4.0308	0.0308	0.3614	0.3255
Genetic Matching	0.9	410	5	400	10	4.0234	0.0234	0.5371	0.4675
CSRF	0.9	410	5	400	10	4.0483	0.0483	0.3668	0.3279
Causal Forest	0.9	410	5	400	10	3.7000	-0.3000	0.3301	0.0103
Mean weight	-0.9	410	5	400	10	4.2717	0.2717	0.3802	0.3222
Random Forest	-0.9	410	5	400	10	4.2635	0.2635	0.3687	0.3206
Genetic Matching	-0.9	410	5	400	10	4.5636	0.5636	0.6523	0.5717
CSRF	-0.9	410	5	400	10	4.2747	0.2747	0.3788	0.3233
Causal Forest	-0.9	410	5	400	10	3.5992	-0.4008	0.3299	0.0054

Table B.2: *Summary of estimates of the models from the heterogeneous data simulation with true ATT unknown and* $N = 410$, $N_c = 400$, $N_t = 10$, *and* $p = 5$

Model	M	Size(n)	P	Cont	Trt	ATT	Estimates	Expected Bias	Empirical SE	Theoretical SE
Mean weight	0	410	5	400	10	2.6037	2.5916	-0.0122	0.6654	0.6261
Random Forest	0	410	5	400	10	2.6037	2.5588	-0.0450	0.6603	0.6257
Genetic Matching	0	410	5	400	10	2.6037	2.5964	-0.0074	0.7949	0.7889
CSRF	0	410	5	400	10	2.6037	2.5872	-0.0166	0.6649	0.6266
Causal Forest	0	410	5	400	10	2.6037	2.0927	-0.5111	0.5797	0.0143
Mean weight	0.05	410	5	400	10	2.6022	2.5875	-0.0147	0.6705	0.6269
Random Forest	0.05	410	5	400	10	2.6022	2.5551	-0.0471	0.6659	0.6264
Genetic Matching	0.05	410	5	400	10	2.6022	2.6067	0.0045	0.8053	0.7901
CSRF	0.05	410	5	400	10	2.6022	2.5853	-0.0170	0.6700	0.6273
Causal Forest	0.05	410	5	400	10	2.6022	2.0894	-0.5128	0.5825	0.0143
Mean weight	-0.05	410	5	400	10	2.6040	2.5932	-0.0109	0.6653	0.6281
Random Forest	-0.05	410	5	400	10	2.6040	2.5601	-0.0440	0.6597	0.6272
Genetic Matching	-0.05	410	5	400	10	2.6040	2.5944	-0.0096	0.7768	0.8054
CSRF	-0.05	410	5	400	10	2.6040	2.5903	-0.0137	0.6633	0.6284
Causal Forest	-0.05	410	5	400	10	2.6040	2.0888	-0.5153	0.5784	0.0142
Mean weight	0.25	410	5	400	10	2.6020	2.5898	-0.0122	0.6697	0.6277
Random Forest	0.25	410	5	400	10	2.6020	2.5582	-0.0438	0.6651	0.6270
Genetic Matching	0.25	410	5	400	10	2.6020	2.5823	-0.0197	0.7801	0.7778
CSRF	0.25	410	5	400	10	2.6020	2.5851	-0.0169	0.6691	0.6281
Causal Forest	0.25	410	5	400	10	2.6020	2.1025	-0.4995	0.5877	0.0150
Mean weight	-0.25	410	5	400	10	2.6052	2.6228	0.0176	0.6690	0.6276
Random Forest	-0.25	410	5	400	10	2.6052	2.5931	-0.0121	0.6640	0.6268
Genetic Matching	-0.25	410	5	400	10	2.6052	2.6207	0.0155	0.7806	0.8177
CSRF	-0.25	410	5	400	10	2.6052	2.6197	0.0146	0.6696	0.6279
Causal Forest	-0.25	410	5	400	10	2.6052	2.1161	-0.4891	0.5794	0.0136
Mean weight	0.5	410	5	400	10	2.6007	2.5965	-0.0043	0.6628	0.6270
Random Forest	0.5	410	5	400	10	2.6007	2.5689	-0.0318	0.6583	0.6262
Genetic Matching	0.5	410	5	400	10	2.6007	2.5819	-0.0188	0.7768	0.7683
CSRF	0.5	410	5	400	10	2.6007	2.5920	-0.0087	0.6628	0.6276
Causal Forest	0.5	410	5	400	10	2.6007	2.1480	-0.4528	0.5874	0.0159
Mean weight	-0.5	410	5	400	10	2.6037	2.6803	0.0766	0.6591	0.6266
Random Forest	-0.5	410	5	400	10	2.6037	2.6596	0.0559	0.6544	0.6259
Genetic Matching	-0.5	410	5	400	10	2.6037	2.7222	0.1185	0.7914	0.8422
CSRF	-0.5	410	5	400	10	2.6037	2.6793	0.0756	0.6599	0.6271
Causal Forest	-0.5	410	5	400	10	2.6037	2.1891	-0.4146	0.5781	0.0125
Mean weight	0.9	410	5	400	10	2.6025	2.6522	0.0496	0.6589	0.6300
Random Forest	0.9	410	5	400	10	2.6025	2.6346	0.0320	0.6561	0.6294
Genetic Matching	0.9	410	5	400	10	2.6025	2.6196	0.0170	0.7536	0.7372
CSRF	0.9	410	5	400	10	2.6025	2.6521	0.0496	0.6606	0.6306
Causal Forest	0.9	410	5	400	10	2.6025	2.3860	-0.2165	0.6060	0.0190
Mean weight	-0.9	410	5	400	10	2.6034	2.8738	0.2705	0.6627	0.6259
Random Forest	-0.9	410	5	400	10	2.60345	2.8675	0.2641	0.6569	0.6251
Genetic Matching	-0.9	410	5	400	10	2.6034	3.1651	0.5617	0.8351	0.9019
CSRF	-0.9	410	5	400	10	2.6034	2.8788	0.2754	0.6636	0.6263
Causal Forest	-0.9	410	5	400	10	2.6034	2.4025	-0.2008	0.5672	0.0094

Table B.3: *Summary of estimates of the models from homogeneous data simulation with true ATE as* ***4.0*** *and* $N = 410$, $N_c = 400$, $N_t = 10$, *and* $p = 10$

Model	M	Size(n)	P	Cont	Trt	Estimate	Expected Bias	Empirical SE	Theoretical SE
Mean weight	0	410	10	400	10	3.8655	-0.1345	0.3777	0.3204
Random Forest	0	410	10	400	10	3.8376	-0.1624	0.3695	0.3196
Genetic Matching	0	410	10	400	10	3.8735	-0.1265	0.6226	0.5409
CSRF	0	410	10	400	10	3.8619	-0.1381	0.3761	0.3205
Causal Forest	0	410	10	400	10	3.1937	-0.8063	0.3143	0.0066
Mean weight	0.05	410	10	400	10	3.8746	-0.1254	0.3651	0.3212
Random Forest	0.05	410	10	400	10	3.8450	-0.1550	0.3584	0.3206
Genetic Matching	0.05	410	10	400	10	3.8691	-0.1309	0.6186	0.5328
CSRF	0.05	410	10	400	10	3.8698	-0.1302	0.3665	0.3214
Causal Forest	0.05	410	10	400	10	3.2041	-0.7959	0.3072	0.0067
Mean weight	-0.05	410	10	400	10	3.8702	-0.1298	0.3628	0.3198
Random Forest	-0.05	410	10	400	10	3.8408	-0.1592	0.3565	0.3190
Genetic Matching	-0.05	410	10	400	10	3.8767	-0.1233	0.6044	0.5318
CSRF	-0.05	410	10	400	10	3.8660	-0.1340	0.3630	0.3197
Causal Forest	-0.05	410	10	400	10	3.1939	-0.8061	0.3076	0.0065
Mean weight	0.25	410	10	400	10	3.8727	-0.1273	0.3815	0.3245
Random Forest	0.25	410	10	400	10	3.8426	-0.1574	0.3744	0.3235
Genetic Matching	0.25	410	10	400	10	3.8446	-0.1554	0.6151	0.5250
CSRF	0.25	410	10	400	10	3.8683	-0.1317	0.3810	0.3246
Causal Forest	0.25	410	10	400	10	3.2171	-0.7829	0.3206	0.0071
Mean weight	-0.25	410	10	400	10	3.9163	-0.0837	0.3693	0.3172
Random Forest	-0.25	410	10	400	10	3.8885	-0.1115	0.3635	0.3163
Genetic Matching	-0.25	410	10	400	10	3.9373	-0.0627	0.6037	0.5378
CSRF	-0.25	410	10	400	10	3.9127	-0.0873	0.3692	0.3176
Causal Forest	-0.25	410	10	400	10	3.2170	-0.7830	0.3053	0.0062
Mean weight	0.5	410	10	400	10	3.9111	-0.0889	0.3799	0.3253
Random Forest	0.5	410	10	400	10	3.8831	-0.1169	0.3719	0.3245
Genetic Matching	0.5	410	10	400	10	3.8788	-0.1212	0.5855	0.5172
CSRF	0.5	410	10	400	10	3.9052	-0.0948	0.3772	0.3253
Causal Forest	0.5	410	10	400	10	3.2905	-0.7095	0.3298	0.0077
Mean weight	-0.5	410	10	400	10	3.9892	-0.0108	0.3537	0.3157
Random Forest	-0.5	410	10	400	10	3.9672	-0.0328	0.3451	0.3150
Genetic Matching	-0.5	410	10	400	10	4.0688	0.0688	0.6179	0.5436
CSRF	-0.5	410	10	400	10	3.9880	-0.0120	0.3519	0.3164
Causal Forest	-0.5	410	10	400	10	3.2749	-0.7251	0.2946	0.0060
Mean weight	0.9	410	10	400	10	4.0210	0.0210	0.3694	0.3246
Random Forest	0.9	410	10	400	10	4.0011	0.0011	0.3634	0.3234
Genetic Matching	0.9	410	10	400	10	3.9456	-0.0544	0.5427	0.4904
CSRF	0.9	410	10	400	10	4.0172	0.0172	0.3699	0.3252
Causal Forest	0.9	410	10	400	10	3.6045	-0.3955	0.3236	0.0102
Mean weight	-0.9	410	10	400	10	4.2279	0.2279	0.3546	0.3168
Random Forest	-0.9	410	10	400	10	4.2193	0.2193	0.3503	0.3163
Genetic Matching	-0.9	410	10	400	10	4.5415	0.5415	0.6270	0.5445
CSRF	-0.9	410	10	400	10	4.2294	0.2294	0.3550	0.3175
Causal Forest	-0.9	410	10	400	10	3.4802	-0.5198	0.3151	0.0064

Table B.4: *Summary of estimates of the models from the heterogeneous data simulation with true ATT unknown and* $N = 410$, $N_c = 400$, $N_t = 10$, *and* $p = 10$

Model	M	Size(n)	P	Cont	Trt	ATT	Estimates	Expected Bias	Empirical SE	Theoretical SE
Mean weight	0	410	10	400	10	2.5677	2.4331	-0.1345	0.6685	0.6185
Random Forest	0	410	10	400	10	2.5677	2.4052	-0.1624	0.6629	0.6181
Genetic Matching	0	410	10	400	10	2.5677	2.4411	-0.1265	0.8355	0.8172
CSRF	0	410	10	400	10	2.5677	2.4295	-0.1381	0.6666	0.6186
Causal Forest	0	410	10	400	10	2.5677	1.9290	-0.6386	0.5683	0.0126
Mean weight	0.05	410	10	400	10	2.5654	2.4400	-0.1254	0.6584	0.6193
Random Forest	0.05	410	10	400	10	2.5654	2.4104	-0.1550	0.6544	0.6190
Genetic Matching	0.05	410	10	400	10	2.5654	2.4345	-0.1309	0.8330	0.8064
CSRF	0.05	410	10	400	10	2.5654	2.4353	-0.1302	0.6598	0.6194
Causal Forest	0.05	410	10	400	10	2.5654	1.9353	-0.6301	0.5650	0.0127
Mean weight	-0.05	410	10	400	10	2.5682	2.4384	-0.1298	0.6550	0.6184
Random Forest	-0.05	410	10	400	10	2.5682	2.4091	-0.1592	0.6524	0.6181
Genetic Matching	-0.05	410	10	400	10	2.5682	2.4449	-0.1233	0.8061	0.8087
CSRF	-0.05	410	10	400	10	2.5682	2.4342	-0.1340	0.6552	0.6185
Causal Forest	-0.05	410	10	400	10	2.5682	1.9295	-0.6388	0.5648	0.0124
Mean weight	0.25	410	10	400	10	2.5681	2.4407	-0.1273	0.6711	0.6200
Random Forest	0.25	410	10	400	10	2.5681	2.4107	-0.1574	0.6672	0.6193
Genetic Matching	0.25	410	10	400	10	2.5681	2.4126	-0.1554	0.8293	0.7920
CSRF	0.25	410	10	400	10	2.5681	2.4363	-0.1317	0.6707	0.6202
Causal Forest	0.25	410	10	400	10	2.5681	1.9431	-0.6250	0.5760	0.0131
Mean weight	-0.25	410	10	400	10	2.5664	2.4830	-0.0834	0.6621	0.6179
Random Forest	-0.25	410	10	400	10	2.5664	2.4549	-0.1115	0.6590	0.6173
Genetic Matching	-0.25	410	10	400	10	2.5664	2.5037	-0.0627	0.8078	0.8158
CSRF	-0.25	410	10	400	10	2.5664	2.4791	-0.0873	0.6618	0.6181
Causal Forest	-0.25	410	10	400	10	2.5664	1.9601	-0.6063	0.5622	0.0121
Mean weight	0.5	410	10	400	10	2.5678	2.4793	-0.0885	0.6611	0.6190
Random Forest	0.5	410	10	400	10	2.5678	2.4509	-0.1169	0.6564	0.6186
Genetic Matching	0.5	410	10	400	10	2.5678	2.4466	-0.1212	0.8036	0.7810
CSRF	0.5	410	10	400	10	2.5678	2.4731	-0.0947	0.6584	0.6190
Causal Forest	0.5	410	10	400	10	2.5678	2.0097	-0.5580	0.5809	0.0141
Mean weight	-0.5	410	10	400	10	2.5681	2.5573	-0.0108	0.6482	0.6174
Random Forest	-0.5	410	10	400	10	2.5681	2.5353	-0.0328	0.6433	0.6171
Genetic Matching	-0.5	410	10	400	10	2.5681	2.6369	0.0688	0.8226	0.8292
CSRF	-0.5	410	10	400	10	2.5681	2.5561	-0.0120	0.6468	0.6180
Causal Forest	-0.5	410	10	400	10	2.5681	2.0262	-0.5419	0.5514	0.0117
Mean weight	0.9	410	10	400	10	2.5678	2.5838	0.0160	0.6498	0.6211
Random Forest	0.9	410	10	400	10	2.5678	2.5650	-0.0028	0.6484	0.6207
Genetic Matching	0.9	410	10	400	10	2.5678	2.5333	-0.0345	0.7657	0.7540
CSRF	0.9	410	10	400	10	2.5678	2.5820	0.0142	0.6500	0.6219
Causal Forest	0.9	410	10	400	10	2.5678	2.2774	-0.2903	0.5947	0.0185
Mean weight	-0.9	410	10	400	10	2.5692	2.7971	0.2279	0.6471	0.6175
Random Forest	-0.9	410	10	400	10	2.5692	2.7885	0.2193	0.6456	0.6173
Genetic Matching	-0.9	410	10	400	10	2.5692	3.1106	0.5415	0.8222	0.8387
CSRF	-0.9	410	10	400	10	2.5692	2.7986	0.2294	0.6476	0.6179
Causal Forest	-0.9	410	10	400	10	2.5692	2.2736	-0.2956	0.5500	0.0112

Table B.5: *Summary of estimates of the models from homogeneous data simulation with true ATE as* ***4.0*** *and* $N = 410$, $N_c = 400$, $N_t = 10$, *and* $p = 50$

Model	M	Size(n)	P	Cont	Trt	Estimate	Expected Bias	Empirical SE	Theoretical SE
Mean weight	0	410	50	400	10	3.6404	-0.3596	0.3484	0.3311
Random Forest	0	410	50	400	10	3.6368	-0.3632	0.3466	0.3304
Genetic Matching	0	410	50	400	10	3.6535	-0.3465	0.6027	0.5503
CSRF	0	410	50	400	10	3.6390	-0.3610	0.3486	0.3310
Causal Forest	0	410	50	400	10	3.0822	-0.9178	0.3034	0.0065
Mean weight	0.05	410	50	400	10	3.6324	-0.3676	0.3488	0.3315
Random Forest	0.05	410	50	400	10	3.6277	-0.3723	0.3457	0.3310
Genetic Matching	0.05	410	50	400	10	3.6408	-0.3592	0.6128	0.5448
CSRF	0.05	410	50	400	10	3.6390	-0.3610	0.3486	0.3310
Causal Forest	0.05	410	50	400	10	3.0765	-0.9235	0.3027	0.0065
Mean weight	-0.05	410	50	400	10	3.6401	-0.3599	0.3518	0.3303
Random Forest	-0.05	410	50	400	10	3.6365	-0.3635	0.3502	0.3299
Genetic Matching	-0.05	410	50	400	10	3.6839	-0.3161	0.6079	0.5453
CSRF	-0.05	410	50	400	10	3.6376	-0.3624	0.3510	0.3306
Causal Forest	-0.05	410	50	400	10	3.0779	-0.9221	0.3074	0.0065
Mean weight	0.25	410	50	400	10	3.6523	-0.3477	0.3596	0.3350
Random Forest	0.25	410	50	400	10	3.6470	-0.3530	0.3563	0.3341
Genetic Matching	0.25	410	50	400	10	3.6882	-0.3118	0.6216	0.5508
CSRF	0.25	410	50	400	10	3.6490	-0.3510	0.3589	0.3352
Causal Forest	0.25	410	50	400	10	3.0948	-0.9052	0.3099	0.0067
Mean weight	-0.25	410	50	400	10	3.6878	-0.3122	0.3394	0.3239
Random Forest	-0.25	410	50	400	10	3.6843	-0.3157	0.3373	0.3235
Genetic Matching	-0.25	410	50	400	10	3.7395	-0.2605	0.5879	0.5363
CSRF	-0.25	410	50	400	10	3.6845	-0.3155	0.3403	0.3238
Causal Forest	-0.25	410	50	400	10	3.1007	-0.8993	0.2958	0.0062
Mean weight	0.5	410	50	400	10	3.6387	-0.3613	0.3603	0.3295
Random Forest	0.5	410	50	400	10	3.6335	-0.3665	0.3565	0.3290
Genetic Matching	0.5	410	50	400	10	3.6905	-0.3095	0.6306	0.5436
CSRF	0.5	410	50	400	10	3.6364	-0.3636	0.3604	0.3297
Causal Forest	0.5	410	50	400	10	3.0755	-0.9245	0.3131	0.0065
Mean weight	-0.5	410	50	400	10	3.6378	-0.3622	0.3491	0.3288
Random Forest	-0.5	410	50	400	10	3.6333	-0.3667	0.3473	0.3282
Genetic Matching	-0.5	410	50	400	10	3.6770	-0.3230	0.5950	0.5436
CSRF	-0.5	410	50	400	10	3.6343	-0.3657	0.3482	0.3290
Causal Forest	-0.5	410	50	400	10	3.0766	-0.9234	0.3040	0.0064
Mean weight	0.9	410	50	400	10	4.0042	0.0042	0.3814	0.3447
Random Forest	0.9	410	50	400	10	3.9971	-0.0029	0.3762	0.3425
Genetic Matching	0.9	410	50	400	10	3.8857	-0.1143	0.6322	0.5777
CSRF	0.9	410	50	400	10	3.9976	-0.0024	0.3820	0.3450
Causal Forest	0.9	410	50	400	10	3.4839	-0.5161	0.3280	0.0086
Mean weight	-0.9	410	50	400	10	4.1855	0.1855	0.3384	0.3177
Random Forest	-0.9	410	50	400	10	4.1838	0.1838	0.3361	0.3175
Genetic Matching	-0.9	410	50	400	10	4.2911	0.2911	0.6003	0.5348
CSRF	-0.9	410	50	400	10	4.1828	0.1828	0.3390	0.3178
Causal Forest	-0.9	410	50	400	10	3.4138	-0.5862	0.2971	0.0064

Table B.6: *Summary of estimates of the models from the heterogeneous data simulation with true ATT unknown and* $N = 410$, $N_c = 400$, $N_t = 10$, *and* $p = 50$

Model	M	Size(n)	P	Cont	Trt	ATT	Estimates	Expected Bias	Empirical SE	Theoretical SE
Mean weight	0	410	50	400	10	2.5940	2.2334	-0.3605	0.6315	0.6260
Random Forest	0	410	50	400	10	2.5940	2.2310	-0.3630	0.6317	0.6257
Genetic Matching	0	410	50	400	10	2.5940	2.2475	-0.3465	0.7904	0.8043
CSRF	0	410	50	400	10	2.5940	2.2325	-0.3615	0.6325	0.6260
Causal Forest	0	410	50	400	10	2.5940	1.8495	-0.7445	0.5471	0.0121
Mean weight	0.05	410	50	400	10	2.5957	2.2283	-0.3674	0.6328	0.6257
Random Forest	0.05	410	50	400	10	2.5957	2.2233	-0.3725	0.6313	0.6256
Genetic Matching	0.05	410	50	400	10	2.5957	2.2365	-0.3592	0.8096	0.8013
CSRF	0.05	410	50	400	10	2.5957	2.2260	-0.3697	0.6337	0.6256
Causal Forest	0.05	410	50	400	10	2.5957	1.8438	-0.7519	0.5469	0.0122
Mean weight	-0.05	410	50	400	10	2.5962	2.2436	-0.3526	0.6375	0.6256
Random Forest	-0.05	410	50	400	10	2.5962	2.2396	-0.3566	0.6364	0.6254
Genetic Matching	-0.05	410	50	400	10	2.5962	2.2912	-0.3050	0.8329	0.8040
CSRF	-0.05	410	50	400	10	2.5962	2.2406	-0.3556	0.6375	0.6255
Causal Forest	-0.05	410	50	400	10	2.5962	1.8493	-0.7469	0.5481	0.0121
Mean weight	0.25	410	50	400	10	2.5924	2.2447	-0.3477	0.6400	0.6252
Random Forest	0.25	410	50	400	10	2.5924	2.2394	-0.3530	0.6384	0.6250
Genetic Matching	0.25	410	50	400	10	2.5924	2.2806	-0.3118	0.8063	0.8050
CSRF	0.25	410	50	400	10	2.5924	2.2414	-0.3510	0.6395	0.6254
Causal Forest	0.25	410	50	400	10	2.5924	1.8589	-0.7335	0.5522	0.0123
Mean weight	-0.25	410	50	400	10	2.5957	2.2829	-0.3128	0.6285	0.6229
Random Forest	-0.25	410	50	400	10	2.5957	2.2788	-0.3169	0.6285	0.6228
Genetic Matching	-0.25	410	50	400	10	2.5957	2.3460	-0.2497	0.7992	0.8011
CSRF	-0.25	410	50	400	10	2.5957	2.2813	-0.3144	0.6304	0.6229
Causal Forest	-0.25	410	50	400	10	2.5957	1.8708	-0.7249	0.5428	0.0121
Mean weight	0.5	410	50	400	10	2.5956	2.2957	-0.2999	0.6407	0.6286
Random Forest	0.5	410	50	400	10	2.5956	2.2917	-0.3039	0.6385	0.6283
Genetic Matching	0.5	410	50	400	10	2.5956	2.2345	-0.3611	0.8193	0.8123
CSRF	0.5	410	50	400	10	2.5956	2.2928	-0.3029	0.6390	0.6289
Causal Forest	0.5	410	50	400	10	2.5956	1.9038	-0.6918	0.5541	0.0128
Mean weight	-0.5	410	50	400	10	2.5955	2.3753	-0.2202	0.6353	0.6216
Random Forest	-0.5	410	50	400	10	2.5955	2.3731	-0.2224	0.6350	0.6217
Genetic Matching	-0.5	410	50	400	10	2.5955	2.4481	-0.1474	0.8003	0.7964
CSRF	-0.5	410	50	400	10	2.5955	2.3731	-0.2223	0.6360	0.6217
Causal Forest	-0.5	410	50	400	10	2.5955	1.9341	-0.6614	0.5438	0.0118
Mean weight	0.9	410	50	400	10	2.5985	2.6026	0.0042	0.6484	0.6303
Random Forest	0.9	410	50	400	10	2.5985	2.5956	-0.0029	0.6458	0.6296
Genetic Matching	0.9	410	50	400	10	2.5985	2.4841	-0.1143	0.8097	0.8236
CSRF	0.9	410	50	400	10	2.5985	2.5961	-0.0024	0.6489	0.6307
Causal Forest	0.9	410	50	400	10	2.5985	2.2267	-0.3718	0.5690	0.0155
Mean weight	-0.9	410	50	400	10	2.5964	2.7816	0.1853	0.6304	0.6220
Random Forest	-0.9	410	50	400	10	2.5964	2.7799	0.1835	0.6296	0.6220
Genetic Matching	-0.9	410	50	400	10	2.5964	2.8874	0.2911	0.7999	0.7923
CSRF	-0.9	410	50	400	10	2.5964	2.7791	0.1827	0.6307	0.6222
Causal Forest	-0.9	410	50	400	10	2.5964	2.2496	-0.3467	0.5365	0.0123

Table B.7: *Summary of estimates of the models from homogeneous data simulation with true ATE as* ***4.0*** *and* $N = 1025$, $N_c = 1000$, $N_t = 25$, *and* $p = 5$

Model	M	Size(n)	P	Cont	Trt	Estimate	Expected Bias	Empirical SE	Theoretical SE
Mean weight	0	1025	5	1000	25	3.9919	-0.0081	0.2397	0.2051
Random Forest	0	1025	5	1000	25	3.9706	-0.0294	0.2347	0.2041
Genetic Matching	0	1025	5	1000	25	3.9801	-0.0199	0.3941	0.3791
CSRF	0	1025	5	1000	25	3.9896	-0.0104	0.2389	0.2056
Causal Forest	0	1025	5	1000	25	3.4317	-0.5683	0.2104	0.0044
Mean weight	0.05	1025	5	1000	25	4.0027	0.0027	0.2473	0.2055
Random Forest	0.05	1025	5	1000	25	3.9804	-0.0196	0.2395	0.2044
Genetic Matching	0.05	1025	5	1000	25	4.0031	0.0031	0.3957	0.3779
CSRF	0.05	1025	5	1000	25	4.0007	0.0007	0.2460	0.2059
Causal Forest	0.05	1025	5	1000	25	3.4313	-0.5687	0.2115	0.0045
Mean weight	-0.05	1025	5	1000	25	4.0072	0.0072	0.2430	0.2050
Random Forest	-0.05	1025	5	1000	25	3.9860	-0.0140	0.2366	0.2041
Genetic Matching	-0.05	1025	5	1000	25	3.9969	-0.0031	0.3759	0.3815
CSRF	-0.05	1025	5	1000	25	4.0033	0.0033	0.2417	0.2054
Causal Forest	-0.05	1025	5	1000	25	3.4469	-0.5531	0.2112	0.0045
Mean weight	0.25	1025	5	1000	25	4.0003	0.0003	0.2460	0.2061
Random Forest	0.25	1025	5	1000	25	3.9780	-0.0220	0.2397	0.2051
Genetic Matching	0.25	1025	5	1000	25	3.9951	-0.0049	0.3801	0.3687
CSRF	0.25	1025	5	1000	25	3.9944	-0.0056	0.2460	0.2067
Causal Forest	0.25	1025	5	1000	25	3.4417	-0.5583	0.2111	0.0045
Mean weight	-0.25	1025	5	1000	25	4.0191	0.0191	0.2347	0.2049
Random Forest	-0.25	1025	5	1000	25	3.9982	-0.0018	0.2277	0.2038
Genetic Matching	-0.25	1025	5	1000	25	4.0308	0.0308	0.4021	0.3913
CSRF	-0.25	1025	5	1000	25	4.0146	0.0146	0.2327	0.2056
Causal Forest	-0.25	1025	5	1000	25	3.5087	-0.4913	0.2117	0.0052
Mean weight	0.5	1025	5	1000	25	4.0059	0.0059	0.2354	0.2068
Random Forest	0.5	1025	5	1000	25	3.9860	-0.0140	0.2296	0.2058
Genetic Matching	0.5	1025	5	1000	25	3.9737	-0.0263	0.3597	0.3588
CSRF	0.5	1025	5	1000	25	4.0027	0.0027	0.2345	0.2070
Causal Forest	0.5	1025	5	1000	25	3.4876	-0.5124	0.2053	0.0047
Mean weight	-0.5	1025	5	1000	25	4.0572	0.0572	0.2465	0.2040
Random Forest	-0.5	1025	5	1000	25	4.0405	0.0405	0.2390	0.2030
Genetic Matching	-0.5	1025	5	1000	25	4.0876	0.0876	0.4439	0.4103
CSRF	-0.5	1025	5	1000	25	4.0550	0.0550	0.2470	0.2046
Causal Forest	-0.5	1025	5	1000	25	3.6422	-0.3578	0.2298	0.0062
Mean weight	0.9	1025	5	1000	25	4.0198	0.0198	0.2264	0.2065
Random Forest	0.9	1025	5	1000	25	4.0058	0.0058	0.2235	0.2058
Genetic Matching	0.9	1025	5	1000	25	3.9790	-0.0210	0.3344	0.3332
CSRF	0.9	1025	5	1000	25	4.0213	0.0213	0.2272	0.2071
Causal Forest	0.9	1025	5	1000	25	3.6782	-0.3218	0.2039	0.0048
Mean weight	-0.9	1025	5	1000	25	4.2125	0.2125	0.2624	0.2040
Random Forest	-0.9	1025	5	1000	25	4.2164	0.2164	0.2489	0.2032
Genetic Matching	-0.9	1025	5	1000	25	4.4782	0.4782	0.5564	0.4843
CSRF	-0.9	1025	5	1000	25	4.2177	0.2177	0.2614	0.2044
Causal Forest	-0.9	1025	5	1000	25	3.7146	-0.2854	0.2744	0.0114

Table B.8: *Summary of estimates of the models from the heterogeneous data simulation with true ATT unknown and* $N = 1025$, $N_c = 1000$, $N_t = 25$, *and* $p = 5$

Model	M	Size(n)	P	Cont	Trt	ATT	Estimates	Expected Bias	Empirical SE	Theoretical SE
Mean weight	0	1025	5	1000	25	2.5833	2.5752	-0.0081	0.4158	0.4000
Random Forest	0	1025	5	1000	25	2.5833	2.5539	-0.0294	0.4135	0.3994
Genetic Matching	0	1025	5	1000	25	2.5833	2.5635	-0.0199	0.5212	0.5925
CSRF	0	1025	5	1000	25	2.5833	2.5729	-0.0104	0.4153	0.4003
Causal Forest	0	1025	5	1000	25	2.5833	2.1534	-0.4300	0.3706	0.0085
Mean weight	0.05	1025	5	1000	25	2.5830	2.5854	0.0024	0.4232	0.4000
Random Forest	0.05	1025	5	1000	25	2.5830	2.5634	-0.0196	0.4180	0.3995
Genetic Matching	0.05	1025	5	1000	25	2.5830	2.5862	0.0031	0.5198	0.5887
CSRF	0.05	1025	5	1000	25	2.5830	2.5829	-0.0001	0.4218	0.4003
Causal Forest	0.05	1025	5	1000	25	2.5830	2.1526	-0.4304	0.3741	0.0085
Mean weight	-0.05	1025	5	1000	25	2.5822	2.5896	0.0075	0.4151	0.4000
Random Forest	-0.05	1025	5	1000	25	2.5822	2.5681	-0.0141	0.4113	0.3996
Genetic Matching	-0.05	1025	5	1000	25	2.5822	2.5791	-0.0031	0.5021	0.5975
CSRF	-0.05	1025	5	1000	25	2.5822	2.5857	0.0036	0.4142	0.4002
Causal Forest	-0.05	1025	5	1000	25	2.5822	2.1672	-0.4149	0.3720	0.0087
Mean weight	0.25	1025	5	1000	25	2.5830	2.5834	0.0003	0.4228	0.4003
Random Forest	0.25	1025	5	1000	25	2.5830	2.5610	-0.0220	0.4187	0.3997
Genetic Matching	0.25	1025	5	1000	25	2.5830	2.5781	-0.0049	0.5156	0.5743
CSRF	0.25	1025	5	1000	25	2.5830	2.5775	-0.0056	0.4230	0.4007
Causal Forest	0.25	1025	5	1000	25	2.5830	2.1575	-0.4255	0.3731	0.0085
Mean weight	-0.25	1025	5	1000	25	2.5821	2.6012	0.0191	0.4057	0.4003
Random Forest	-0.25	1025	5	1000	25	2.5821	2.5804	-0.0018	0.4017	0.3997
Genetic Matching	-0.25	1025	5	1000	25	2.5821	2.6129	0.0308	0.5225	0.6129
CSRF	-0.25	1025	5	1000	25	2.5821	2.5967	0.0146	0.4041	0.4007
Causal Forest	-0.25	1025	5	1000	25	2.5821	2.2191	-0.3631	0.3731	0.0101
Mean weight	0.5	1025	5	1000	25	2.5822	2.5881	0.0059	0.4128	0.4006
Random Forest	0.5	1025	5	1000	25	2.5822	2.5682	-0.0140	0.4095	0.4000
Genetic Matching	0.5	1025	5	1000	25	2.5822	2.5559	-0.0263	0.4980	0.5613
CSRF	0.5	1025	5	1000	25	2.5822	2.5850	0.0027	0.4116	0.4008
Causal Forest	0.5	1025	5	1000	25	2.5822	2.1917	-0.3905	0.3693	0.0088
Mean weight	-0.5	1025	5	1000	25	2.5829	2.6401	0.0572	0.4181	0.3996
Random Forest	-0.5	1025	5	1000	25	2.5829	2.6234	0.0405	0.4141	0.3991
Genetic Matching	-0.5	1025	5	1000	25	2.5829	2.6705	0.0876	0.5575	0.6475
CSRF	-0.5	1025	5	1000	25	2.5829	2.6379	0.0550	0.4191	0.4000
Causal Forest	-0.5	1025	5	1000	25	2.5829	2.3420	-0.2409	0.3965	0.0118
Mean weight	0.9	1025	5	1000	25	2.5823	2.6021	0.0198	0.4082	0.4013
Random Forest	0.9	1025	5	1000	25	2.5823	2.5880	0.0058	0.4065	0.4008
Genetic Matching	0.9	1025	5	1000	25	2.5823	2.5613	-0.0210	0.4762	0.5237
CSRF	0.9	1025	5	1000	25	2.5823	2.6035	0.0213	0.4085	0.4017
Causal Forest	0.9	1025	5	1000	25	2.5823	2.3453	-0.2370	0.3771	0.0092
Mean weight	-0.9	1025	5	1000	25	2.5827	2.7979	0.2152	0.4260	0.3989
Random Forest	-0.9	1025	5	1000	25	2.5827	2.8014	0.2187	0.4195	0.3984
Genetic Matching	-0.9	1025	5	1000	25	2.5827	3.0602	0.4775	0.6484	0.7711
CSRF	-0.9	1025	5	1000	25	2.5827	2.8032	0.2205	0.4262	0.3991
Causal Forest	-0.9	1025	5	1000	25	2.5827	2.4974	-0.0853	0.4516	0.0160