Counterfactual Prediction Methods for Causal Inference in Observational Studies with Continuous Treatments

Joel Persson

June 13, 2019

STAN40: Master's thesis in Statistics (15 ECTS)

Department of Statistics
Lund University

Supervisor: Krzysztof Podgórski

Abstract

We develop methods for estimation, inference and optimization of causal effects from observational data with continuous treatments. We present a counterfactual prediction method based on the potential outcomes framework that estimates the expected value of a potential outcome given a treatment level and confounders. We show that the method identifies the average generalized treatment effect (AGTE) and the average dose-response function (ADRF) and propose estimators of these functional causal estimands. Our estimators work under high-dimensional confounding and when the treatment takes many distinct values. Under multiple treatments, the method identifies the effect of a single treatment and the joint effect of multiple treatments. Treatment effects can further be estimated from unobserved treatment levels. We provide nonparametric and computationally efficient parametric estimation procedures of uncertainty intervals of the ADRF and AGTE and develop algorithms for implementation of the estimators. Finally, we show that the counterfactual prediction method can be used to estimate the treatment level that maximizes the expected individual and population outcome.

Keywords: Causal Inference; Observational Study; Treatment Effects; Continuous Treatments; Dose-Response Function; Optimization

Acknowledgements

I would first and foremost like to thank my advisor Krzysztof Podgórski for the guidance in writing this thesis. Your questions in the initial stage challenged my understanding of the field and improved the quality of this work. I very much appreciate your immediate and hands-on feedback and for helping me narrow the scope of this thesis. Björn Holmquist, thank you for the suggestions and corrections at the last stage of writing the thesis. On a more general note, I would like to thank Antonio Marañon for all the support I have received during the last two years and the opportunities that I have been available to me through you. I cannot emphasize enough how important this has been for my education. I also want to express my gratitude towards Peter Gustafsson for always being helpful and the interesting discussions we have had. Finally, I want to thank my parents for their continuous support.

Contents

List of Abbreviations

RCT	Randomized controlled trial
ATE	Average treatment effect
ATET	Average treatment effect for the treated
SUTVA	Stable unit treatment value assumption
CATE	Conditional average treatment effect
AGTE	Average generalized treatment effect
AGTET	Average generalized treatment effect for the treated
CAGTE	Conditional average generalized treatment effect
DRF	Dose-response function
ADRF	Average dose-response function

1 Introduction

Background. Many questions in science are causal rather than descriptive or predictive. In the social and biomedical sciences, interest is often of the effect of a treatment on individuals, groups, or another unit of measurement. The word *treatment* is here meant to be interpreted generally and refers to a change in the environment of the units. In economics, a treatment can be a policy that affects a group of the population. In medicine, a treatment may be a drug that some individuals receive. Such questions are causal since we are interested in the effect of the treatment on the outcome while controlling for alternative causes of the outcome.

Randomized controlled trials (RCT's) (Fisher, 1935) are the gold standard for causal inference. The reason is that randomization ensures that alternative factors that may contaminate the estimate of the treatment effect cancel in expectation. In many areas of science RCT's are not feasible due to ethical or political reasons, the complexity involved, or the financial cost. For instance, macroeconomic policies cannot be tested experimentally since it is politically impossible to have a control group and we cannot go back in time to compare the outcome to what would have happened had the policy not been implemented. Still, if a policy is implemented, a before-after difference in outcomes will yield biased estimates of the policy's effect if other factors that affect the outcome have changed over time. To determine the effect of smoking on health, we cannot just look at the differences in health measures between those who smoke and those who don't since smoking is likely correlated with other lifestyle factors that have negative health effects. This means that methods used to estimate treatment effects in RCT's do not work in observational studies.

Following Rosenbaum (2010), we define an observational study as an empirical study of effects caused by treatments when randomized experimentation is unethical or infeasible. Similarly, we define observational data as any data not generated in an administered experimental setting. Examples of observational data are population registers collected by census bureaus, purchasing data from e-commerce businesses, and most of so called big data that is continuously generated. Recent advances in measurement and data collection has increased the amount of observational data available for researchers. The ability to analyse phenomena as it occurs in its natural habitat is attractive for empirical research and has led to the development of statistical methods for causal inference with observational data.

Literature review. Most methods for causal inference consider the treatment to be binary. The justification is that either the individual received the treatment or not. In reality, many treatments are continuous, meaning that they can take a range of values. The response to continuous treatments is often dose-dependent and non-linear. For instance, drugs often have a recommended dosage. An insufficient dose will fail to cause an effect and an excessive dose can have a negative effect. Given this intuitive fact, it is surprising that causal inference for observational data with continuous treatments has only received attention recently (Fong et al., 2018; Hirano et al., 2003; Hirano & Imbens, 2004; Hernán et al., 2000; Zhu et al., 2014; Zhang et al., 2016; Kennedy et al., 2016, 2017). Dose-response modelling has been used in the experimental life-sciences longer (see for instance Altshuler (1981) or Wang (2015)), but those methods are not applicable to observational data since they do not account for the non-randomized treatment assignment.

The causal inference literature has historically focused on average treatment effects. While averages can be useful, they have limited practical value under heterogenous responses. Heterogenous treatment effects is an active stream of research in econometrics (Chernozhukov et al., 2017, 2018; Wager & Athey, 2018; Athey & Imbens, 2016; Nie & Wager, 2017; Hitsch & Misra, 2018), (bio)statistics (Powers et al., 2018; Oprescu et al., 2018), political science (Grimmer et al., 2017; Imai & Ratkovic, 2013) and computer science (Künzel et al., 2017; Oprescu et al., 2018). However, practically all methods are developed for binary treatments or RCT's. Knowledge of the treatment effect heterogeneity enables individualization of treatments for optimization of outcomes. Applications range from personalized medicine to targeted advertisements. An early and influential series of papers in this area are Manski (2000, 2002, 2004) and Hirano and Porter (2009), which based on the theory of statistical decision functions (Wald, 1949) consider *planners* that must assign individuals to either treatment or control. Given information about individuals, the goal is to allocate the treatment among individuals such that the population outcome is maximized. By making causal inference a problem of decision-making under uncertainty, this literature directly targets the normative problem relevant to policy makers - how to assign treatments optimally.

Contribution. In this thesis, we develop methods for estimation, inference and optimization of causal effects from observational data with continuous treatments. We present a *counterfactual prediction method* based on the potential outcomes framework (Rubin, 1974; Holland, 1986) and the regression adjustment method for binary treatments. The method involves defining a dose-response function (DRF) that returns the expected outcome of a treatment level conditional on confounders of the treatment effect. The DRF thereby naturally estimates heterogenous treatment effects and can be used to estimate the average dose-response function and the average effect of one treatment level relative another level. Assuming the DRF is true, the estimator identifies the causal treatment effect and is a generalization of the binary treatment case. The method has two main benefits relative standard approaches: First, under exposure to several treatments it can identify the effect of a single treatment and the joint effect of several of the treatments. The ability to pool and separate effects from different treatments has received limited attention in the literature but is important for applications, especially in observational studies where individuals may self-select into exposure to many treatments simultaneously. Secondly, our method works under high-dimensional confounding and when the treatment and confounders take many distinct values. We develop methods for estimating confidence and prediction intervals of the treatment effect and provide algorithms for implementation. Finally, we show that the DRF identifies the treatment level that maps to the maximum mean individual or population outcome.

Structure of thesis. The thesis is structured as follows: Section 2 introduces the potential outcomes framework, the causal estimands and relevant assumptions. Section 3 presents existing theory and methods for identification and estimation of treatment effects in observational data with binary treatments. Section 4 contains the main contributions of the thesis. We first present the continuous treatment setting, introduce the adapted estimands, and explain why methods for binary treatments do not work well in this case. This is followed by the presentation of the dose-response function and then identification, estimation, implementation and inference of the estimands for continuous treatments. We then show that the DRF identifies and estimates the optimal treatment level. Section 5 concludes the thesis with a summary and a discussion of the contributions in relation to the literature.

A note on the notation: Uppercase regular type face is used for random variables and lowercase regular type face for their values. Uppercase boldface letters are matrices or random vectors. Which of these we refer to in a particular case will be apparent in the context. Lowercase boldface are realized vectors from the sample. Script letters denote sets.

2 Potential Outcomes Framework

Example 2.1. Imagine that you want to estimate the effect of a new supplement that is supposed to increase cardiovascular fitness. You go to the local gym and survey some individuals if they take the supplement and their best one mile running time. Unknown to you, supplement use is correlated with age and training history; those who take the supplement are on average younger and have trained longer than those who do not take it. Clearly, looking at the average difference in one mile running time between those who take the supplement and those who do not will yield a biased estimate of the supplement's effect since age and training history affects cardiovascular fitness. The difference in cardiovascular fitness between those who take the supplement and those who do not is partially explained by pre-treatment differences in age and training history.

We translate the concepts from the example to a statistical framework. Let Y, Z, and $\boldsymbol{X}$ be observed random variables from some population $\mathcal{P}$ of size N. Here, $Y \in \mathbb{R}$ is a scalar outcome, $Z = z \in \{0, 1\}$ is an indicator variable of exposure to treatment measured prior to Y, and $\boldsymbol{X} = (X_1, X_2, \ldots, X_p)'$ is a p-dimensional vector of pre-treatment covariates, or alternatively, post-treatment if unaffected by Z, defined on the bounded subset $\mathcal{X} \subset \mathbb{R}^p$ where $p \in \mathbb{N}^* = \{1, 2, 3, \ldots\}$ may be large. The word *treatment* refers to any action applied to individual i, which may for instance be a single person, household, or geographical region. Our target τ is the causal effect of the treatment Z on the outcome Y in $\mathcal{P}$. A more formal definition of the causal effect is defined in what follows.

Let $I \in \mathbb{N}^*$ denote a random individual uniformly sampled from $\mathcal{P}$ with individual characteristics $\boldsymbol{X}$ and exposure to treatment denoted by Z. Here, Z is a random variable whose value in the simplest case tells whether the individual received the treatment ($Z = 1$) or not ($Z = 0$). Hence $(I, Z, \boldsymbol{X})$ is a multivariate random variable where I is a uniformly sampled individual from the population, Z is the treatment status of the individual, and $\boldsymbol{X}$ is the individual's p random characteristics. Thus, Z may depend on I and $\boldsymbol{X}$. In this framework, $Y = Y(I, Z, \boldsymbol{X})$ is the observed outcome for a random individual from the population, with three sources of variability; the variability related to the sampling of individuals, the variability related to how the treatment is assigned to individuals, and the variability in individual-level characteristics.

Definition 2.1 (Potential Outcomes)**.** The *potential outcomes* (Rubin, 1974; Holland, 1986) given a treatment $Z = z \in \{0, 1\}$ is the set $\mathcal{Y} \in \{Y(I, 1, \boldsymbol{X}), Y(I, 0, \boldsymbol{X})\}$, $\mathcal{Y} \subset \mathbb{R}$, where $Y(I, 1, \boldsymbol{X}) = Y(1)$ is the potential outcome that would be observed under exposure to treatment and $Y(I, 0, \boldsymbol{X}) = Y(0)$ is the potential outcome that would be observed under exposure to control (no treatment). The observed outcome Y is the potential outcome seen under the received treatment, that is

$$Y(I, Z, \boldsymbol{X}) = Y(I, 1, \boldsymbol{X})Z + Y(I, 0, \boldsymbol{X})(1 - Z) \tag{2.1}$$

Remark. In most instances we will use the shorter notation $Y(Z)$ for $Y(I, Z, \boldsymbol{X})$ although the potential outcome also depend on I and $\boldsymbol{X}$.

We now define the causal estimand of interest, the average treatment effect.

Definition 2.2 (Average Treatment Effect)**.** Define $P_{Y(1)}$ to be the probability distribution of the outcome if all individuals received treatment and define $P_{Y(0)}$ to be the probability distribution of the outcome if all individuals received control, both defined on the set $\mathcal{Y}$. The Rubin Causal Model (Rubin, 1974; Holland, 1986) states that the population *average treatment effect* (ATE) is given by

$$\tau_{ATE} \coloneqq \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)] = \mu_1 - \mu_0. \tag{2.2}$$

We revisit the introductory example of the supplement that increases cardiovascular fitness but now formalize the problem in the potential outcomes framework.

Example 2.2. Let $Z = 1$ denote that an individual took the supplement and $Z = 0$ that an individual did not. We want to estimate the ATE of the supplement on the one mile running time, $\mathbb{E}[Y(1)] - \mathbb{E}[Y(0)]$. Under standard assumptions about sampling, $\mathbb{E}[Y(1)]$ would be the average running time if the entire population took the supplement and $\mathbb{E}[Y(0)]$ would be the average running time if none took it. If we sample individuals uniformly from $\mathcal{P}$, assign the supplement at random to those individuals, and estimate the difference in averages among individuals of each treatment status, we get an unbiased estimate of the ATE. However, the assignment of Z is based on the covariates $\boldsymbol{X}$, which in this case is information about age and training history that itself explains running time Y. Also, whether a random individual I takes the supplement or not depends on who we sample. Thus the value of Z depends on $\boldsymbol{X}$ and I. Instead of obtaining an unbiased estimate of the difference $\mathbb{E}[Y(1)] - \mathbb{E}[Y(0)]$, a difference in means of the groups will give us an unbiased estimate of $\mathbb{E}[Y(1)|Z = 1] - \mathbb{E}[Y(0)|Z = 0]$. Here, $\mathbb{E}[Y(1)|Z = 1]$ is the average running time for the young individuals with long training history (that took the supplement) and $\mathbb{E}[Y(0)|Z = 0]$ is the average outcome for the older individuals with less training experience (who did not take the supplement). Since the groups' one mile running time is different even without the supplement, this difference in averages does not yield an unbiased estimate of our target $\mathbb{E}[Y(1)] - \mathbb{E}[Y(0)]$.

The fundamental problem of causal inference. Suppose we observe the data $\mathcal{D} = \{(y_i, z_i, \boldsymbol{x}_i)\}_{i=1}^n$, which are values of Y, Z, and $\boldsymbol{X} = (X_1, X_2, \ldots, X_p)'$ for n individuals sampled without replacement from $\mathcal{P}$. Let $y_i(z_i)$ be the observed potential outcome for individual i. Using expression (2.1), we see that the realized outcome we observe in the data is

$$y_i = y_i(1)z_i + y_i(0)(1 - z_i). \tag{2.3}$$

Since it is only possible to observe one of the potential outcomes for each individual, the *individual treatment effect* $\tau_i \coloneqq y_i(1) - y_i(0)$ is unobservable. Hence causal inference is a missing data problem. Table 1 shows that for any observed outcome y_i, the *counterfactual* potential outcome $y_i^{(cf)}$ required to estimate the treatment effect is missing. This is known as the *fundamental problem of causal inference* (Imbens & Rubin, 2015).

Definition 2.3 (Counterfactual Potential Outcomes)**.** The *counterfactual potential outcome* $Y^{(cf)}$ is the unrealized potential outcome, given by the unobserved pair of random variables $(Y(I, z, \boldsymbol{X}), Z = 1 - z)$ for $z \in \{0, 1\}$. The counterfactual potential outcome for individual i is $Y_i^{(cf)} = (Y_i(z_i),\ Z = 1 - z_i)$.

Remark. The word *counterfactual* refer to something that is contrary to the facts. Hence counterfactuals are only defined ex post the treatment assignment. Ex ante the treatment assignment, all outcomes are unrealized potential outcomes.

The fundamental problem of causal inference implies that the joint distribution $P_{(Y_i(1), Y_i(0))}$ of the potential outcomes cannot be inferred from the data. Similarly, the probability distributions $P_{Y(1)}$ and $P_{Y(0)}$ cannot be observed, but just $P_{Y(1)|Z=1}$ and $P_{Y(0)|Z=0}$. Neither the distribution of the individual-level treatment effects, $P_{Y(1)-Y(0)}$, nor the ATE are estimable unless we make further assumptions.

Table 1: The Fundamental Problem of Causal Inference

Individual	$y_i(1)$	$y_i(0)$	z_i	τ_i
1	✓	?	1	?
2	✓	?	1	?
3	?	✓	0	?
4	✓	?	1	?
⋮	⋮	⋮	⋮	⋮
n	✓	?	1	?

Notes: ✓(observed), ? (unobserved)
This table demonstrates the fundamental problem of causal inference. For each individual, only one of the potential outcomes is realized. Thereby the individual treatment effect τ_i is always unobserved.

Rct's vs observational studies. In randomized controlled trials (rct's), the treatment is randomly assigned to individuals with no consideration of their potential response. This implies that the probability that the outcome Y takes a specific value in its range is the same regardless of the value of Z. Thus,

$$\{Y(0), Y(1)\} \perp Z. \tag{2.4}$$

This means that the treatment Z is independent of the potential outcomes. Using (2.3), we have that $\mathbb{E}[Y|Z=z] = \mathbb{E}[Y(z)]$ for any value z. Thus

$$\begin{aligned}\mathbb{E}[Y|Z=1] &= \mathbb{E}[Y(1)Z + Y(0)(1-Z)|Z=1] \\ &= \mathbb{E}[Y(1)|Z=1] = \mathbb{E}[Y(1)] \end{aligned} \tag{2.5}$$

and

$$\begin{aligned}\mathbb{E}[Y|Z=0] &= \mathbb{E}[Y(1)Z + Y(0)(1-Z)|Z=0] \\ &= \mathbb{E}[Y(0)|Z=0] = \mathbb{E}[Y(0)]. \end{aligned} \tag{2.6}$$

In other words, an rct ensures that the distribution of the potential outcome conditional on exposure is equal to the unconditional distribution of the potential outcome. Hence the inability to observe both outcomes for the whole population is not a problem for estimating the ATE. Let $n_1 = \sum_{i=1}^{n} z_i$ and $n_0 = \sum_{i=1}^{n}(1 - z_i)$ be the number of treated and non-treated observations, respectively, so that $n = n_1 + n_0$. Define $\bar{Y}(1) := n_1^{-1} \sum_{i=1}^{n} z_i y_i$ to be the average outcome for the treated and $\bar{Y}(0) := n_0^{-1} \sum_{i=1}^{n}(1 - z_i) y_i$ to be the average outcome for the non-treated. An unbiased estimator of the population ATE in a RCT is

$$\widehat{\tau}_{ATE} = \bar{Y}(1) - \bar{Y}(0). \tag{2.7}$$

In observational studies the treatment exposure is not random. Instead, it is likely that the characteristics that determine exposure also determine potential outcome. It may be the case that individuals self-select into treatment (that is, choose their treatment) or that an administrator assign treatments to individuals with the highest expected gain from treatment. Consequently, Z is dependent on I and $\boldsymbol{X}$ so (2.4) does not hold. Then $\mathbb{E}[Y(I, Z, \boldsymbol{X})|Z = 1] \neq \mathbb{E}[Y(I, 1, \boldsymbol{X})]$ and $\mathbb{E}[Y(I, Z, \boldsymbol{X})|Z = 0] \neq \mathbb{E}[Y(I, 0, \boldsymbol{X})]$. That is, $\mathbb{E}[Y|Z = z]$ is not equal to $\mathbb{E}[Y(z)]$. This means that the individuals that received treatment are not comparable to those who did not. Thereby equation (2.7) yields a biased estimate of the ATE.

Example 2.3. The government wants to test a job market program on a group of unemployed individuals prior to implementing the program nation-wide. The individuals that apply for the trial of the program are likely part of the subgroup of the unemployed individuals that have sufficient ability and drive to be accepted to such programs. Since drive and ability are positively associated with job prospects, a trial of the program on the self-selected individuals with a comparison to those who did not participate will overestimate the effectiveness of the job market program. In other words, the distributions of the potential outcomes of the unemployed individuals that did not participate in the program are not equal to the distributions of the potential outcomes for the unemployed people that participated in the program.

A solution to the problem of self-selection into treatment is to estimate the treatment effect only for the treated individuals.

Definition 2.4 (AVERAGE TREATMENT EFFECT FOR THE TREATED)**.** The population *average treatment effect for the treated* (ATET) is

$$\tau_{ATET} := \mathbb{E}[Y(1) - Y(0)|Z = 1]. \tag{2.8}$$

The ATE and ATET are mathematically closely related. By adding and subtracting the unobserved expectation $\mathbb{E}[Y(0)|Z=1]$ to the ATE and rearranging terms, we see that the ATE is equal to the ATET plus the *selection bias*:

$$\begin{aligned}\tau_{ATE} &= \underbrace{\Big(\mathbb{E}[Y(1)|Z=1] - \mathbb{E}[Y(0)|Z=0]\Big)}_{=\tau_{ATE}} + \underbrace{\Big(\mathbb{E}[Y(0)|Z=1] - \mathbb{E}[Y(0)|Z=1]\Big)}_{=0} \\ &= \mathbb{E}[Y(1)|Z=1] - \mathbb{E}[Y(0)|Z=0] + \mathbb{E}[Y(0)|Z=1] - \mathbb{E}[Y(0)|Z=1] \\ &= \underbrace{\Big(\mathbb{E}[Y(1)|Z=1] - \mathbb{E}[Y(0)|Z=1]\Big)}_{=\tau_{ATET}} + \underbrace{\Big(\mathbb{E}[Y(0)|Z=1] - \mathbb{E}[Y(0)|Z=0]\Big)}_{\textit{selection bias}}.\end{aligned} \tag{2.9}$$

Definition 2.5 (Selection Bias)**.** *Selection bias* refers to the bias in estimates introduced by the self-selection of individuals into exposure such that the treatment assignment is not random. The selection bias is given by

$$\mathbb{E}[Y(0)|Z=1] - \mathbb{E}[Y(0)|Z=0]. \tag{2.10}$$

Under independence of potential outcomes and exposure, $\mathbb{E}[Y(0)|Z=1]$ equals $\mathbb{E}[Y(0)|Z=0]$ and $\mathbb{E}[Y(1)|Z=1]$ equals $\mathbb{E}[Y(1)|Z=0]$. Hence, in RCT's the selection bias is zero so that the ATE equals ATET. The reason is that random treatment assignment ensures that, on average, pre-treatment outcomes are equal for the treated and non-treated. In layman terms; treated individuals would have the same outcomes as the individuals in the control group had they not received treatment, and vice versa. This is a strong assumption that is not likely to hold in observational data due to self-selection into treatment. The following example shows that the ATET is also appropriate in cases when the ATE is ill-defined.

Example 2.4. A retailer want to see if offering a discount for orders that exceed a threshold value makes customers place orders above the threshold. High-value customers are indifferent to the discount since their orders exceed the threshold. Low-value customers place orders below the threshold and find the threshold too high to warrant the discount and do not qualify for it. In other words, the pre-treatment outcome (order value) for the groups are not equal. Since some customers never qualify for the discount, the ATE is not interesting. The interesting estimand is the effect for the mid-value customers that due to the discount increase their order values to above the threshold. This is what the ATET would estimate.

The problem with the ATET is that $\mathbb{E}[Y(0)|Z=1]$ cannot be observed. Just as with the ATE, the individual treatment effect for the treated is always unobserved. The solution to this problem is to identify covariates related to both treatment status and the pre-treatment outcome.

Definition 2.6 (CONFOUNDER)**.** A *confounder* is a covariate that jointly affects pre-treatment outcome and treatment status, causing spurious relationships.

Remark. Throughout this thesis we assume that $\boldsymbol{X}$ is all confounders. In the job market program example, drive and ability are confounders since they are correlated with both career outcome and exposure to the program.

Assumptions. The existence of confounders and the fact that all potential outcomes cannot be observed require us to make a set of assumptions about the *treatment assignment mechanism*, that is, how treatments are assigned, to be able to get an unbiased estimate the treatment effect from observational data.

Assumption 2.1 (STABLE UNIT TREATMENT VALUE ASSUMPTION)**.** The *Stable Unit Treatment Value Assumption* (SUTVA) consists of two components:

1. *No interference*: the treatment status of an individual is unaffected by the potential outcome of any other individual, meaning $\mathbb{P}(Z) = \mathbb{P}(Z|Y(0), Y(1))$.
2. *No hidden variations in treatment*: For each treatment there exists only a single form of that treatment. Thereby $\mathbb{E}[Y(z)] = \mathbb{E}[Y|Z=z]$.

Remark. No interference is equivalent to the assumption that observations are independent and identically distributed (i.i.d.) samples from $\mathcal{P}$. It implies that a potential outcome of one individual is not affected by a potential outcome of another individual. No hidden variations in treatment means that the variation in the treatment is known. Thus, if an individual is assigned $Z = z$ we observe $Y(z)$. The variation can be with respect to the level of the treatment or the type of treatment. For example, if we are interested in the effect of smoking on lung cancer, the treated individuals (those who smoke) should only consists of individuals that smoke one form of cigarettes. Otherwise we will be averaging over individuals that have been exposed to different treatments (cigarettes) in the estimation of the effect. If either SUTVA component is not satisfied the potential outcomes are not uniquely defined (Imbens & Rubin, 2015).

Assumption 2.2 (Strong Unconfoundedness)**.** Let $\mathcal{X} \subset \mathbb{R}^p$ be the support of the joint distribution function $P_{\boldsymbol{X}}$ of $\boldsymbol{X}$. For the ATE, the treatment assignment mechanism is *strongly unconfounded* if for all $\boldsymbol{x} \in \mathcal{X}$ and $Z = z \in \{0, 1\}$,

$$\{Y(1), Y(0)\} \perp Z | \boldsymbol{X} = \boldsymbol{x}, \tag{2.11}$$

whereas for the ATET it suffices that for $\boldsymbol{x} \in \mathcal{X}$ and $Z = z \in \{0, 1\}$

$$Y(0) \perp Z | \boldsymbol{X} = \boldsymbol{x}. \tag{2.12}$$

Remark. Strong unconfoundedness can also be stated as that for $Z = z \in \{0, 1\}$

$$\mathbb{E}[Y(z)|Z = z, \boldsymbol{X}] = \mathbb{E}[Y(z)|\boldsymbol{X}]. \tag{2.13}$$

Unconfoundedness[1] means that conditional on confounders, the treatment assignment is random. Then the potential outcomes are jointly independent with respect to $\boldsymbol{X}$ so the data behaves as if coming from an RCT. The weaker assumption is sufficient for the ATET due to that the ATET is estimated as the difference between $\mathbb{E}[Y(1)|Z = 1]$ and $\mathbb{E}[Y(0)|Z = 1]$ where $\mathbb{E}[Y(1)|Z = 1]$ is observable. Thereby only the non-treated observations need to be unconfounded.

Unconfoundedness is a statement about the conditional distribution $P_{\mathcal{Y}|\boldsymbol{X}}$. Since only one of $Y_i(1)$ or $Y_i(0)$ is observable for each individual the assumption is empirically unverifiable. In an RCT, the treatment assignment mechanism is controlled by the researcher. Thereby the assumption can hold by design. In observational studies, the assumptions may not hold since the treatment assignment is not controlled.

Unconfoundedness is sometimes called *selection on observables* since the assumption is that the confounders that determine selection into treatment are observable. This is important since under unobserved confounding, the distribution of the observed data is consistent with multiple possibly contradictory explanations indistinguishable from the data (D'Amour, 2019). With unobserved confounding, there is no way to determine the causal mechanism that has generated the observed data. Then it is not possible to obtain the true value of the causal estimand no matter the amount of data. The causal estimand is then said to be *unidentifiable.* We discuss identification in further detail at the beginning of Section 3.

[1] We will for the remainder of the thesis use the shorter term unconfoundedness to refer to strong unconfoundedness, and similarly, write that the treatment assignment is unconfounded when we mean strongly unconfounded, unless otherwise stated.

Under unconfoundness, we can estimate the causal effect by adjusting for the confounders among treated and non-treated. This leads to the next assumption about the conditional probability distribution of receiving the treatment.

Assumption 2.3 (Common Support)**.** The *common support* assumption for the ATE is that for all $\boldsymbol{x} \in \mathcal{X}$ and $z \in \{0, 1\}$,

$$0 < \mathbb{P}(Z = z|\boldsymbol{X} = \boldsymbol{x}) < 1. \tag{2.14}$$

For ATET the weaker assumption of *partial common support* is that for all $\boldsymbol{x} \in \mathcal{X}$,

$$0 < \mathbb{P}(\boldsymbol{X} = \boldsymbol{x}) < 1. \tag{2.15}$$

Remark. The implication of common support is that for $z \in \{0, 1\}$

$$\mathbb{E}[Y|Z = z, \boldsymbol{X}] = \mathbb{E}[Y(z)|Z = z, \boldsymbol{X}]. \tag{2.16}$$

Remark. The assumption states that the conditional probability distributions $P_{\boldsymbol{X}|Z=1}$ and $P_{\boldsymbol{X}|Z=0}$ have common support $\mathcal{X}$. It means that each individual has a positive probability of receiving the treatment. Partial common support only requires that there is a positive probability that $\boldsymbol{X}$ takes the value $\boldsymbol{x}$.

If there for each value $\boldsymbol{x}$ in the data exists both treated and non-treated individuals, then common support holds by definition. However, a failure of this ad-hoc test does not reject the common support assumption. That each individual has a positive probability to receive the treatment does not rule out the possibility that only similar individuals received the treatment.

A violation of common support may be a statistical problem (the data is not a random sample from the population) or structural (some individuals in the population never receive the treatment). The structural violation cannot be solved in observational studies.

Under unconfoundedness and common support, the treatment assignment mechanism is said to be *strongly ignorable* (Rosenbaum & Rubin, 1983), meaning that how treatments are assignment can be ignored since it is independent of the potential outcomes. Then the observed treated and non-treated observations are comparable conditional on the confounder.

3 Binary Treatments

This section deals with identification and estimation of treatment effects in observational studies with binary treatments.

Definition 3.1 (IDENTIFIABILITY)**.** A parameter is said to be *identifiable* if it is a function of the observed probability distribution.

The definition implies that only for identifiable parameters is it possible to obtain the true parameter value. Identification is about how much we can learn from an estimand if we had an infinite amount of data. Statistical inference, on the other hand, is about how much we can learn about an estimand from a finite sample. The distinction is important due to the fundamental problem of causal inference; no amount of data will help us identify the treatment effect unless we impose causal assumptions. Hence identification of τ does not depend on the sample size but on the assumptions of the treatment assignment mechanism. For observational studies with binary treatments, SUTVA and strong ignorability are sufficient. Only for identifiable estimands are the statistical issues of estimation and inference relevant.

There are three standard approaches to identification and estimation of binary treatment effects under SUTVA and strong ignorability; 1) matching, 2) propensity score methods, and 3) regression adjustment. All methods are based on adjustment for confounding but do so in different ways (see Imbens and Rubin (2015) or Hernán and Robins (2019) for a survey of the methods).

Matching methods. Matching estimators (Rubin, 1973; Heckman et al., 1997) work as follows: The number of treated and non-treated observations are counted. Observations in the smaller set are matched with observations with opposite treatment status that have similar values on pre-treatment confounders. The similarity is estimated with a statistical distance measure such as the Malahanobis distance to obtain a distance value d. Matching done with a classification algorithm, commonly k-nearest neighbour (Rubin, 1973) or genetic search (Diamond & Sekhon, 2013). The result is paired observations $\{y_i(1), y_j(0)\}$, $i \neq j$, for $i, j = 1, \dots, n_{min}$ where $n_{min} = min(n_1, n_0)$ that are sufficiently similar in terms of d. In that sense, matching can be viewed as imputing the missing potential outcomes. Unmatched observations are discarded. The ATE is estimated as the difference in the average values of the treated and non-treated matched observations. Since $(Y(1),\ Z=1)$ is observed, identification of the ATET only requires that the counterfactual $(Y(0),\ Z=1)$ is identified and imputed for each treated individual with $Y(1)$.

Propensity score methods. The propensity score method (Rosenbaum & Rubin, 1983) was developed as a solution to the problem that arise with matching if $\boldsymbol{X}$ is high-dimensional or take many distinct values. The method is based on taking differences in outcomes between individuals with similar probability of receiving the treatment. For each individual the *propensity score* $e(\boldsymbol{x}_i) = \mathbb{P}(Z|\boldsymbol{X} = \boldsymbol{x}_i)$ is estimated, commonly with a binary (for instance logistic) regression model. Thereby the treatment assignment mechanism is modelled. Since $\boldsymbol{X}$ are the covariates that determine selection into treatment, the propensity score is the probability of receiving the treatment. The propensity score is a sufficient statistic for strong ignorability. This implies that observations of opposite treatment status can be matched on the scalar $e(\boldsymbol{x})$ instead of on all p components of $\boldsymbol{X}$. The ATE is then estimated as the difference in average outcomes among the propensity score matched observations. The propensity score can be viewed as a one-dimensional measure of confounding. Thus, taking the difference between outcomes matched on the propensity score reduces bias in the estimate of the ATE due to confounding.

Regression adjustment. *Regression adjustment* does not involve matching or modelling the treatment assignment mechanism. Instead, it is based on the following idea: Under strong ignorability, Z can be considered exogenous and a regression function can be used to adjust the estimate of the treatment effect for confounding and omitted variable bias. Imbens and Wooldridge (2009) show this by defining the treatment effect τ conditional on $\boldsymbol{X} = \boldsymbol{x}$:

Lemma 3.1. *Define $\mu_z(\boldsymbol{x}) \coloneqq \mathbb{E}[Y(z)|\boldsymbol{X} = \boldsymbol{x}]$, $z \in \{0,1\}$, as regression functions of the potential outcomes conditional on $\boldsymbol{x}$. We then have that*

$$\begin{aligned}\tau(\boldsymbol{x}) &\coloneqq \mathbb{E}[Y(1) - Y(0)|\boldsymbol{X} = \boldsymbol{x}]\\ &= \mathbb{E}[Y(1)|\boldsymbol{X} = \boldsymbol{x}] - \mathbb{E}[Y(0)|\boldsymbol{X} = \boldsymbol{x}]\\ &= \mathbb{E}[Y(1)|Z = 1, \boldsymbol{X} = \boldsymbol{x}] - \mathbb{E}[Y(0)|Z = 0, \boldsymbol{X} = \boldsymbol{x}]\\ &= \mathbb{E}[Y|Z = 1, \boldsymbol{X} = \boldsymbol{x}] - \mathbb{E}[Y|Z = 0, \boldsymbol{X} = \boldsymbol{x}]\\ &= \mu_1(\boldsymbol{x}) - \mu_0(\boldsymbol{x}), \qquad (3.1)\end{aligned}$$

where the third equality holds by unconfoundedness and the fourth by common support.

Remark. The function $\tau(\boldsymbol{x})$, known as the *conditional average treatment effect* (CATE), is often of interest in itself since it provides an estimate of the ATE conditional on individuals with $\boldsymbol{X} = \boldsymbol{x}$. It estimates the individual treatment effect for an individual with characteristics $\boldsymbol{x}$. The estimand of the CATE is $\mathbb{E}[Y(1) - Y(0)|\boldsymbol{X}]$.

Lemma 3.2. *By the law of iterated expectations, the expected outcome given $Z = z$ averaged over all $\boldsymbol{X}$ is*

$$\begin{aligned}\mathbb{E}[Y(z)] = \mathbb{E}\{\mathbb{E}[Y(z)|\boldsymbol{X}]\} &= \mathbb{E}\{\mathbb{E}[Y(z)|Z = z, \boldsymbol{X}]\} \\ &= \mathbb{E}\{\mathbb{E}[Y|Z = z, \ \boldsymbol{X}]\} \\ &= \mathbb{E}[\mu_z(\boldsymbol{X})]. \qquad (3.2)\end{aligned}$$

Here, $\mathbb{E}[\mu_z(\boldsymbol{X})]$ is a partial mean function (Newey, 1994), which is an average of a regression function over conditioning variables, here $\boldsymbol{X}$, while holding others fixed, in this case z.

Remark. Let $P_{\boldsymbol{X}}$ and $P_{\boldsymbol{X}|Z}$ be the probability distributions of $\boldsymbol{X}$ and of $\boldsymbol{X}$ conditional on Z, respectively. The outer expectation in Lemma 3.2 is with respect to the marginal distribution of $\boldsymbol{X}$, that is[2]

$$\mathbb{E}\{\mathbb{E}[Y(z)|\boldsymbol{X}]\} = \int_{\mathcal{X}} \mathbb{E}[Y(z)|\boldsymbol{X} = \boldsymbol{x}]dP_{\boldsymbol{X}}(\boldsymbol{x}).$$

Since the confounder is present, this expectation is not equal to

$$\mathbb{E}[Y|Z = z] = \int_{\mathcal{X}} \mathbb{E}[Y(z)|\boldsymbol{X} = \boldsymbol{x}]dP_{\boldsymbol{X}|Z}(\boldsymbol{x}|z).$$

Proposition 3.3. *By Lemma 3.1 and 3.2,*

$$\begin{aligned}\tau_{ATE} \coloneqq \mathbb{E}[Y(1) - Y(0)] &= \mathbb{E}\{\mathbb{E}[Y(1) - Y(0)|\boldsymbol{X}]\} \\ &= \mathbb{E}[\mu_1(\boldsymbol{X}) - \mu_0(\boldsymbol{X})], \qquad (3.3)\end{aligned}$$

[2] We use the abuse of notation $\mathbb{E}\{\mathbb{E}[Y(z)|Z = z, \boldsymbol{X}]\}$ for $\mathbb{E}_{\boldsymbol{X}}\{\mathbb{E}[Y(z)|Z = z, \boldsymbol{X}]\}$ throughout, that is, for outer expectations with respect to $\boldsymbol{X}$. Also, Riemann-Stieltjes integrals are used throughout where the notation $dP_{\boldsymbol{X}}(\boldsymbol{x})$ denotes integrating with respect to the distribution function of $\boldsymbol{X}$.

$$\begin{aligned}\tau_{ATET} &\coloneqq \mathbb{E}[Y(1) - Y(0)|Z = 1] = \mathbb{E}\{\mathbb{E}[Y(1) - Y(0)|Z = 1, \boldsymbol{X}]\}\\ &= \mathbb{E}[\mu_1(\boldsymbol{X}) - \mu_0(\boldsymbol{X})|Z = 1].\end{aligned} \tag{3.4}$$

It follows that the ATE and ATET can be estimated if $\mu(1, \cdot)$ and $\mu(0, \cdot)$ are identified over $\mathcal{X}$, since conditional on $\boldsymbol{X}$, the sample is random and we can average over the distribution $P_{Y(z)|\boldsymbol{X}}$. Now, under common support,

$$\begin{aligned}\mathbb{E}[Y|Z, \boldsymbol{X}] &= \mathbb{E}[Y(1)|Z, \boldsymbol{X}]Z + \mathbb{E}[Y(0)|Z, \boldsymbol{X}](1 - Z)\\ &= \mathbb{E}[Y(1)|\boldsymbol{X}]Z + \mathbb{E}[Y(0)|\boldsymbol{X}](1 - Z)\\ &\coloneqq \mu_1(\boldsymbol{X})Z + \mu_0(\boldsymbol{X})(1 - Z).\end{aligned} \tag{3.5}$$

That is, $\mu_z(\boldsymbol{X}) = \mathbb{E}[Y|Z, \boldsymbol{X}]$ is identified for all $\boldsymbol{x} \in \mathcal{X}$ since under common support the regression function only consists of observable quantities from the data.

Regression adjustment estimators use result (3.5) together with Proposition 3.3. Obtain $\widehat{\mu}_1(\boldsymbol{X})$ by regressing $\boldsymbol{X}$ on Y in the treated subsample and obtain $\widehat{\mu}_0(\boldsymbol{X})$ by regressing $\boldsymbol{X}$ on Y in the non-treated subsample. With each response function, predict the outcomes over all realized values of $\boldsymbol{X}$ in the full sample. The regression function should be chosen by theory and data and need not be the same for the subsamples. Given that $\mu_1(\cdot)$ and $\mu_0(\cdot)$ are consistent estimators,

$$\widehat{\tau}_{ATE} = \frac{1}{n}\sum_{i=1}^{n}\Big(\widehat{\mu}_1(\boldsymbol{x}_i) - \widehat{\mu}_0(\boldsymbol{x}_i)\Big), \tag{3.6}$$

$$\widehat{\tau}_{ATET} = \frac{1}{\sum_{i=1}^{n} z_i}\sum_{i:z_i=1}\Big(\widehat{\mu}_1(\boldsymbol{x}_i) - \widehat{\mu}_0(\boldsymbol{x}_i)\Big), \tag{3.7}$$

will be consistent estimates of the respective estimands.

Regression adjustment estimates the treatment effect conditional on the realized values of the confounders. If the population consists of individuals with different values of $\boldsymbol{X}$ than in the sample, the estimate will not be representative of the population. If the sample of individuals is random, then the sample values of $\boldsymbol{X}$ will on average be representative of the population. Since observational samples are seldom random, these estimators are often used for sample inference. Only under certain circumstances do they generalize to the population (Rosenbaum, 2010). This is not a drawback, but rather a feature of working with observational data.

4 Continuous Treatments

Many treatments take continuous values. Examples include drugs with a recommended dose or price promotions of varying depth. Discretising such treatments neglects the relationship between dose and response. Then $Z \in \mathcal{Z} = (z_{min}, z_{max})$, where $\mathcal{Z} \subset \mathbb{R}$ is a finite real interval. The potential outcomes consists of the set of real-valued outcomes $\mathcal{Y} = \{Y(z) : z \in \mathcal{Z}\}$ where $Y(z)$ is the potential outcome that would be observed under exposure at level z. For a given individual, $Y_i(z)$ is a potential outcome path indexed by $z \in \mathcal{Z}$. By the fundamental problem of causal inference, the potential outcome path is not observable. We can only observe the potential outcome $Y_i(Z)$ for the realized level of Z for individual i. We call the $n \times 1$ vector $\boldsymbol{z} = (z_1, z_2, \ldots, z_n)'$ containing a treatment level for each individual a *treatment regime*. A treatment regime need not be set by an administrator as in an RCT; a treatment regime can also consist self-selected treatment levels.

Let $z_{0,i} \in \mathcal{Z}$ denote a treatment level for individual i different from z_i. The unobserved counterfactuals are $y_i^{(cf)} = (y_i(z_{0,i}),\ z_i \neq z_{0,i})$ for all $z_{0,i} \in \mathcal{Z}$. In this setting, the *individual generalized treatment effect* is given by difference between $y_i(z_i)$ and whichever counterfactual $(y_i(z_{0,i}), z_i)$, $z_i \neq z_{0,i}$, that we want to compare the outcome $y_i(z_i)$ to. Since there are now more than two potential outcomes, Definition 2.2 of the ATE is not informative. We thereby introduce the average generalized treatment effect as the average of the potential outcomes of one treatment level less the average of the potential outcomes of another treatment level.

Definition 4.1 (Average Generalized Treatment Effect)**.** The population *average generalized treatment effect* (AGTE) of treatment level z relative z_0 is given by

$$\tau_{AGTE}(z, z_0) \coloneqq \mathbb{E}[Y(z) - Y(z_0)]. \tag{4.1}$$

Remark. The corresponding population *conditional average generalized treatment effect* (CAGTE) given $\boldsymbol{X} = \boldsymbol{x}$ is

$$\tau_{CAGTE}(z, z_0, \boldsymbol{x}) \coloneqq \mathbb{E}[Y(z) - Y(z_0)|\boldsymbol{X} = \boldsymbol{x}]. \tag{4.2}$$

In the definition of the AGTE and CAGTE the treatment levels are fixed across individuals. This may not be the case in observational data since it is unlikely that all individuals would self-select the same level of a continuous treatment. It is more likely that the status quo (observed) treatment regime consists of treatment levels that vary over individuals. However, since Z is continuous, it has infinitely many distinct values, and consequently, there exists infinitely many possible treatment regimes. It is not obvious which two treatment regimes should be compared when estimating the AGTE. In some applications it may be most useful to estimate the effect of the status quo treatment regime for the treated individuals relative if they received no treatment. The following example demonstrates when this is the interesting estimand.

Example 4.1. A retailer wants to know the effect of a promotional campaign. Implicitly, the retailer then asks what the effect has been of that specific campaign in comparison to if the campaign would not had taken place. The retailer is not interested in the effect of promotions in general. Further, the retailer does not ask what the effect would have been if the promotions had been assigned differently. If the campaign was targeted to a group of customers, the retailer only wants to know effect for the targeted customers, not what the effect would have been if the promotions had been assigned to all of their customers. Thereby, the goal is to make inferences of the effect of the observed treatments levels on the treated individuals.

Definition 4.2 (Average Generalized Treatment Effect of the Treated)**.** The *average generalized treatment effect for the treated* (AGTET) of the observed treatment levels relative no treatment is given by

$$\tau_{AGTET}(Z, 0) \coloneqq \mathbb{E}[Y(Z) - Y(0)|Z > 0]. \tag{4.3}$$

Assumptions. The assumptions required for identification of the estimands are different from in the binary setting. Hirano and Imbens (2004) introduced the concept of weak unconfoundedness for continuous treatments:

Assumption 4.1 (Weak Unconfoundedness)**.** The treatment assignment mechanism is *weakly unconfounded* if for all $\boldsymbol{x} \in \mathcal{X}$ and each $z \in \mathcal{Z}$

$$Y(z) \perp Z|\boldsymbol{X} = \boldsymbol{x}. \tag{4.4}$$

Remark. Weak unconfoundness implies that for all $\boldsymbol{x} \in \mathcal{X}$ and each $z \in \mathcal{Z}$,

$$\mathbb{E}[Y(z)|Z = z, \boldsymbol{X} = \boldsymbol{x}] = \mathbb{E}[Y(z)|\boldsymbol{X} = \boldsymbol{x}]. \tag{4.5}$$

This assumption is weaker than strong unconfoundedness since it does not assume joint independence of the potential outcomes $\mathcal{Y}$, but only conditional independence of the potential outcome $Y(z)$ given the treatment level. We further assume that $Y(z)$ is continuous in z and that the triplets $(Y_i, Z_i, \boldsymbol{X}_i)$ are i.i.d. for all individuals in the sample and population. The common support for continuous Z is that $0 < \mathbb{P}(Z \in \mathcal{Z}|\boldsymbol{X} = \boldsymbol{x}) < 1$ for all $\boldsymbol{x} \in \mathcal{X}$ in the population of interest with the implication that $\mathbb{E}[Y|Z = z, \boldsymbol{X}] = \mathbb{E}[Y(z)|Z = z, \boldsymbol{X}]$ for all $z \in \mathcal{Z}$. This means that every member of the population has some possibility to receive any dose of the treatment. SUTVA still applies, but with a continuous treatment the assumption of no interference is that $\mathbb{P}(z_i) = \mathbb{P}(z_i|Y(z_i), Y(z_j))$ for each individual $i \neq j$.

Infeasibility of standard binary treatment methods. The methods for estimating treatment effects with binary treatments cannot be used for continuous treatments. Matching becomes infeasible as the number of unique values that Z take increase, since for each observed $z \in (z_{min}, z_{max})$ the observed outcomes must be matched on all p components of $\boldsymbol{X} = (X_1, X_2, \ldots, X_p)'$. Thus if Z take many distinct values in $\mathcal{Z}$, the data at or close to any value z in the space will be sparse. Stated in this way, we see that matching also becomes harder as the dimension of $\boldsymbol{X}$ increase or if $\boldsymbol{X}$ has continuous components. The likelihood of finding matches in this situation is low. Further problems arise if we want to estimate the joint and individual treatment effects when there are multiple treatments. Then the number of required matches increases further. Also, matching estimators cannot identify the joint effect of treatments if individuals in the sample are not exposed to them simultaneously. Propensity score matching alleviate these problems somewhat by matching on the scalar $e(\boldsymbol{x})$ instead of on the vector $\boldsymbol{x}$. Still, there is no guarantee that there exists identical values of $e(\boldsymbol{X})$ for each treatment level z. The regression adjustment method of fitting one regression model per treatment level is infeasible since it for a continuous treatment implies an infinite number of models. Dichotomising the treatment and fitting one model to the treated and another to the non-treated neglects the dose-response relationship inherit to continuous treatments. It also makes it impossible to estimate the AGTE for two non-zero levels of the treatment. In the next section we show that adapting the regression adjustment method to the continuous treatment setting solves these problems.

4.1 The Dose-Response Function

Definition 4.3 (Dose Response Function)**.** The *dose-response function* (DRF) is a function $f : \mathcal{Z} \times \mathcal{X} \to \mathcal{Y}$ that belongs to a convex class of functions $\mathcal{F}$ such that $Y(Z) = Y(I, Z, \boldsymbol{X}) = f(Z, \boldsymbol{X}) + \varepsilon$ where ε is assumed to be approximately distributed[3] as $\mathcal{N}(0, \sigma^2)$.

The DRF[4] takes as arguments a treatment level $z \in \mathcal{Z} \subset \mathbb{R}^k$ and covariates $\boldsymbol{x} \in \mathcal{X} \subset \mathbb{R}^p$ and returns the expected value of the potential outcome $Y(z) \in \mathcal{Y} \subset \mathbb{R}$ associated with that treatment level conditional on $\boldsymbol{x}$. Here, $k > 1$ corresponds to the multiple treatment case. In this section we only consider a single treatment to not let technical details distract from the concepts. We return to the case $k > 1$ in the remaining subsections.

Although $Y(Z) = Y(I, Z, \boldsymbol{X})$ depend on the random individual I, the confounders $\boldsymbol{X}$ contain all information about individuals. Thereby the DRF does not need to take I as input. Under appropriate regularity conditions and assumptions of continuity and smoothness of $f \in \mathcal{F}$, we have that

$$\begin{aligned} \mathbb{E}[Y|Z = z_i, \boldsymbol{X} = \boldsymbol{x}_i] &= \mathbb{E}[Y(z_i)|Z = z_i, \boldsymbol{X} = \boldsymbol{x}_i] \\ &= \mathbb{E}[Y(z_i)|\boldsymbol{X} = \boldsymbol{x}_i] \\ &= \mathbb{E}[f(z_i, \boldsymbol{X}) + \varepsilon_i|\boldsymbol{X} = \boldsymbol{x}_i] \\ &= f(z_i, \boldsymbol{x}_i), \end{aligned} \tag{4.6}$$

where the first equality follows from common support, the second by weak unconfoundedness, the third by the definition of the DRF, and the last from the linearity property of expectation together with $\mathbb{E}[\varepsilon_i] = 0$ and that $f(z, \boldsymbol{x})$ is not random. This shows that $f(z_i, \boldsymbol{x}_i)$ is an individual's expected outcome conditional on a treatment level and confounders. The DRF is analogous to the regression functions $\mu_z(\cdot)$ for binary treatments but instead of one function per treatment level, a single function is used for all treatment levels.

A related function is the average of the DRF over all individuals in the population.

[3] Since the data is observational Z and $\boldsymbol{X}$ are random. Hence the error term only follow this distribution conditional on Z and $\boldsymbol{X}$. For simpler notation we will by abuse of notation not condition on these covariates in relevant derivations although that is implicitly understood.

[4] The life-sciences has a long history of modelling responses to treatment doses with DRF's, see for instance Altshuler (1981) or Wang (2015). Armstrong and Kolesár (2018); Hill (2011); Zhang et al. (2016); Zhu et al. (2014) consider a similar method to us, although they do not call it a DRF.

Definition 4.4 (Average Dose Response Function)**.** The *average dose-response function* (ADRF) is a partial mean function $\mu(\cdot) : \mathbb{R} \to \mathbb{R}$ that maps a scalar treatment level z to its associated population mean potential outcome $\mathbb{E}[Y(z)]$. Using result (4.6), the ADRF is given by

$$\begin{aligned} \mathbb{E}[Y(z)] = \mathbb{E}\{\mathbb{E}[Y(z)|\boldsymbol{X}]\} &= \mathbb{E}[f(z, \boldsymbol{X})] \\ &= \int_{\mathcal{X}} f(z, \boldsymbol{x}) dP_{\boldsymbol{X}}(\boldsymbol{x}) \\ &= \mu(z). \end{aligned} \tag{4.7}$$

Assuming all individuals have identical DRF's, evaluating the ADRF for all $z \in \mathcal{Z}$ gives the curve of the mean potential outcomes over individuals as the treatment takes all of its possible values (Hill, 2011). If the sample consists of random draws from the population, the ADRF can also be viewed as a function that returns the expected potential outcomes for a randomly sampled individual.

Averaging the ADRF over Z gives the population mean potential outcome,

$$\mathbb{E}[\mu(Z)] = \int_{z_{min}}^{z_{max}} \int_{\mathcal{X}} f(z, \boldsymbol{x}) dP_{\boldsymbol{X}}(\boldsymbol{x}) dP_{Z|\boldsymbol{X}}(z|\boldsymbol{x}) \tag{4.8}$$

where $P_{Z|\boldsymbol{X}}$ is the conditional distribution of a treatment level given the confounders. Similarly, we can obtain the mean potential outcome over individuals with fixed covariate values $\boldsymbol{x}$ by averaging the conditional potential outcomes over Z,

$$\mathbb{E}[Y(Z)|\boldsymbol{X} = \boldsymbol{x}] = \mathbb{E}[f(Z, \boldsymbol{x})] = \int_{z_{min}}^{z_{max}} f(z, \boldsymbol{x}) dP_{Z|\boldsymbol{X}}(z|\boldsymbol{x}). \tag{4.9}$$

Expression (4.9) gives the mean response to the treatment for individuals with the same values on the confounders. We now formulate one of the thesis' main ideas.

Proposition 4.2 (Counterfactual prediction)**.** *The DRF $f(Z, \boldsymbol{X})$ returns the conditional expected potential outcome $\mathbb{E}[Y(Z)|\boldsymbol{X}] = \mathbb{E}[Y(I, Z, \boldsymbol{X})|\boldsymbol{X}]$ for any individual I in the population given a treatment level Z and covariates $\boldsymbol{X}$.*

Proof. The proof follows directly from Definition 4.3 and result (4.6). □

The proposition implies that if we can learn f, then by changing Z while holding $\boldsymbol{X}$ fixed we can estimate all potential outcomes for a given individual. If we fix Z and vary $\boldsymbol{X}$, then we can estimate how treatment responses vary with individual-level characteristics.

Model selection. In real life f may be complex and nonlinear. In principle it can be any supervised learning algorithm or regression function that belongs to $\mathcal{F}$. Guidance in model selection is thereby important. In machine learning, the performance of a prediction model is most commonly assessed on its expected prediction risk when applied to new data according to some loss L. Perhaps the most common loss function in this instance is the squared error loss, which here corresponds to the mean squared error. Given that the DRF captures the functional relationship between treatment and response, f should be chosen such that it minimizes

$$\begin{aligned}\mathbb{E}[L(Y, f(Z, \boldsymbol{X}))] &= \mathbb{E}[(Y - f(Z, \boldsymbol{X}))^2] \\ &= \int_{\mathcal{X}} \int_{z_{min}}^{z_{max}} \int_{\mathcal{Y}} \Big(y - f(z, \boldsymbol{x})\Big)^2 dP_{Y|Z,\boldsymbol{X}}(y|z, \boldsymbol{x}) dP_{Z|\boldsymbol{X}}(z|\boldsymbol{x}) dP_{\boldsymbol{X}}(\boldsymbol{x}).\end{aligned} \tag{4.10}$$

Data-splitting is routinely used in machine learning to not overfit a prediction model to sample data. Randomly split observations in the data $\mathcal{D}$ into a training sample $\mathcal{T}$ that the models are fitted on and a validation sample $\mathcal{V}$, where $\mathcal{D} = \mathcal{T} \cup \mathcal{V}$ and $|\mathcal{V}| + |\mathcal{T}| = |\mathcal{D}| = n$. Here, $|\mathcal{A}|$ denotes the cardinality of the set $\mathcal{A}$, that is, the number of elements (observations) in the set. An estimator of (4.10) is then the estimated f-risk (Schuler et al., 2018), given by

$$\widehat{f\text{-risk}} = \frac{1}{|\mathcal{V}|} \sum_{i \in \mathcal{V}}^{|\mathcal{V}|} \Big(y_i - f(z_i, \boldsymbol{x}_i)\Big)^2, \tag{4.11}$$

Thus, $\arg\min_{f \in \mathcal{F}} \widehat{f\text{-risk}}$ implies that f is chosen such that the risk of using a suboptimal DRF in the prediction sense is minimized.

A test set $\mathcal{S}$ can be used to test the performance of the choice of f that minimizes the f-risk. Note that $\mathcal{S}$ should be independent of $\mathcal{D}$ but have the same probability distribution. This can be accomplished by holding out a share of the total n observations for $\mathcal{S}$, which is known as the holdout method. Cross-validation may also be used for model selection (with or without the test set), where the training-validation process is repeated for a pre-specified number of rounds of different data splits. The risk is estimated for each round and then averaged over all rounds.

A complete discussion of the choice of f is beyond the scope of this thesis. We can nonetheless mention some suggestions from the literature for binary treatments and randomized treatment assignments. Some non-parametric suggestion are bayesian additive regression trees (Hill, 2011), variants of random forests (Athey & Imbens, 2016; Wager & Athey, 2018; Athey et al., 2019; Oprescu et al., 2018), deep neural networks (Hartford et al., 2017) and kernel smoothing functions (Kennedy et al., 2017). Parametric suggestions are an appropriately specified generalized linear model, generalized additive model (Zhang et al., 2016) or a model based on generalized estimating equations (also known as M-estimators). If $p+k>n$, then LASSO or regularization may be used. We recommend the researcher to let subject-matter expertise and relevant domain theory guide their choice of f since the assumptions of a specific model may not be justified in a given application. Despite the attractiveness of supervised learning algorithms, we warn against blindly applying such automatic data-driven models, since causal identification is different from prediction in that $f \in \mathcal{F}$ must be unbiased, consistent, include relevant confounders, and have correct functional specification of the relation between treatment and response.

Having chosen a DRF, the problem is to decide which pretreatment covariates to include as confounders. This will have a large influence on the estimate of the treatment effect since unbiased estimation assumes that the treatment assignment is unconfounded. This means that all relevant confounders, but not intermediate outcomes, should be included. Then the estimate of the treatment effect is conditioned on all alternative causes. Correct selection of these covariates typically require knowledge of the causal structure of the data generating process. Such knowledge is possible to have in RCT's where the researcher sample individuals, chooses and assigns treatments, and then controls the values of the confounders. This is not the case in observational studies due to that the data is produced in an environment beyond the control of the researcher. The problem of covariate selection in observational studies is especially difficult if the dimension of $\boldsymbol{X}$ is large. Then, algorithms for data-driven selection of confounders can be used. We refer to the algorithms by Luna et al. (2011); Schnitzer et al. (2016); Persson et al. (2017); Häggström (2018) and their respective references for details and implementation.

4.2 Identification

In this section we show that the DRF point identifies the causal estimands. It is assumed throughout that f is the true DRF and that SUTVA, weak unconfoundedness and common support for continuous treatments hold.

Theorem 4.3. *The* DRF *identifies the* CAGTE*,* AGTE *and the* AGTET *of the observed treatment levels relative no treatment.*

Proof. Suppose z and z_0 are the treatment levels of interest. We begin with the CAGTE. Assuming that f is the true DRF, by (4.6),

$$\begin{aligned}\tau_{CAGTE}(z, z_0, \boldsymbol{x}) &\coloneqq \mathbb{E}[Y(z) - Y(z_0)|\boldsymbol{X} = \boldsymbol{x}] \\ &\coloneqq \mathbb{E}[(f(z, \boldsymbol{x}) + \varepsilon) - (f(z_0, \boldsymbol{x}) + \varepsilon_0)] \\ &= f(z, \boldsymbol{x}) - f(z_0, \boldsymbol{x}), \end{aligned} \tag{4.12}$$

which we were supposed to show. Here, $\varepsilon \neq \varepsilon_0$ but $\mathbb{E}[\varepsilon] = \mathbb{E}[\varepsilon_0] = 0$. The AGTE is obtained by averaging the CAGTE over $\mathcal{X}$. By (4.7),

$$\begin{aligned}\tau_{AGTE}(z, z_0) \coloneqq \mathbb{E}[Y(z) - Y(z_0)] &= \mathbb{E}\{\mathbb{E}[Y(z) - Y(z_0)|\boldsymbol{X}]\} \\ &= \mathbb{E}\{\mathbb{E}[Y(z)|\boldsymbol{X}] - \mathbb{E}[Y(z_0)|\boldsymbol{X}]\} \\ &= \mathbb{E}[f(z, \boldsymbol{X}) - f(z_0, \boldsymbol{X})] \\ &= \mu(z) - \mu(z_0). \end{aligned} \tag{4.13}$$

Finally, the AGTET of the observed treatment levels relative no treatment is

$$\begin{aligned}\tau_{AGTET}(Z, 0) &\coloneqq \mathbb{E}[Y(Z) - Y(0)|Z > 0] \\ &= \mathbb{E}\{\mathbb{E}[Y(Z) - Y(0)|Z > 0, \boldsymbol{X}]\} \\ &= \mathbb{E}\{\mathbb{E}[Y(Z)|Z > 0, \boldsymbol{X}] - \mathbb{E}[Y(0)|Z > 0, \boldsymbol{X}]\} \\ &= \mathbb{E}[f(Z, \boldsymbol{X}) - f(0, \boldsymbol{X})|Z > 0]. \end{aligned} \tag{4.14}$$

Hence we have shown that the DRF identifies the estimands of interest. □

Expressions (4.12) and (4.13) show that the same estimated potential outcomes can be used for the AGTE as the CAGTE. The only difference between the estimands is that the AGTE averages out the variation in the CAGTE that arises by conditioning on different values of $\boldsymbol{X}$. Expression (4.13) further shows that the AGTE is equal to the difference in the ADRF evaluated at the levels for which we want to estimate the AGTE. Note that the ADRF, AGTE and the AGTET are functionals since they take as argument the function f evaluated over the whole support $\mathcal{X}$.

Several easy to derive results follow from Theorem 4.3.

Corollary 4.4. *Let $Z = z \in \{0, 1\}$ be an indicator variable of treatment exposure and let the true* DRF *be a linear regression function. Then the* AGTE *is equal to the* ATE *and is given by the parameter τ for Z.*

Proof. Assume that $f(Z, \boldsymbol{X}) = \alpha + \tau Z + \boldsymbol{X}\boldsymbol{\beta}$ is the true DRF. Since $Z = z \in \{0, 1\}$, there are only two potential outcomes $Y(1)$ and $Y(0)$. Under strong ignorability, the regression parameters are interpreted as the effect of a one unit increase in the associated covariate holding the other covariates fixed. Thus, the proof is complete if the AGTE is equal to τ. We have that

$$\begin{aligned}\tau_{AGTE}(z, z_0) &= \mathbb{E}[Y(1) - Y(0)] \\ &= \mathbb{E}[f(1, \boldsymbol{X}) + \varepsilon - (f(0, \boldsymbol{X}) + \varepsilon_0)] \\ &= \mathbb{E}[(\alpha + \tau \times 1 + \boldsymbol{X}\boldsymbol{\beta} + \varepsilon) - (\alpha + \tau \times 0 + \boldsymbol{X}\boldsymbol{\beta} + \varepsilon_0)] \\ &= \mathbb{E}[\tau] = \tau, \end{aligned} \tag{4.15}$$

which was to be shown. □

Corollary 4.5. *If the true* DRF *is a linear regression function without treatment interactions, then for treatment levels z and z_0, the* AGTE *of z relative z_0 is equal to $\tau(z - z_0)$.*

Proof. Assume that $f(z, \boldsymbol{X}) = \alpha + \tau z + \boldsymbol{X}\boldsymbol{\beta}$ is the true DRF. Then

$$\begin{aligned}\tau_{AGTE}(z, z_0) &= \mathbb{E}[Y(z) - Y(z_0)] \\ &= \mathbb{E}[(\alpha + \tau z + \boldsymbol{X}\boldsymbol{\beta} + \varepsilon) - (\alpha + \tau z_0 + \boldsymbol{X}\boldsymbol{\beta} + \varepsilon_0)] \\ &= \mathbb{E}[\tau z - \tau z_0] = \tau(z - z_0), \end{aligned} \tag{4.16}$$

which was to be shown. □

Theorem 4.3 implies that by supplying the DRF individual-specific treatment levels z_i and $z_{0,i}$ and confounders $\boldsymbol{x}_i$, the DRF identifies the CAGTE for individual i. The following corollaries use this result.

Corollary 4.6. *If the true* DRF *is a linear model with parameter τ^h for the treatment-covariate interaction, then the expected heterogeneity in response to treatment level z between individuals i and j with characteristics $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ is given by $(z\boldsymbol{x}_i' - z\boldsymbol{x}_j')\tau^h$.*

Proof. Assume that $f(z, \boldsymbol{x}_i) = \alpha + \tau z + \boldsymbol{x}_i'\boldsymbol{\beta} + z\boldsymbol{x}_j'\tau^h$ is the true DRF. Then

$$\begin{aligned}
\tau_{CAGTE}(z, \boldsymbol{x}_i, \boldsymbol{x}_j) &= \mathbb{E}[Y_i(z)|\boldsymbol{X} = \boldsymbol{x}_i] - \mathbb{E}[Y_j(z)|\boldsymbol{X} = \boldsymbol{x}_j] \\
&= f(z, \boldsymbol{x}_i) - f(z, \boldsymbol{x}_j) \\
&= (\alpha + \tau z + \boldsymbol{x}_i'\boldsymbol{\beta} + z\boldsymbol{x}_i'\tau^h) - (\alpha + \tau z + \boldsymbol{x}_j'\boldsymbol{\beta} + z\boldsymbol{x}_j'\tau^h) \\
&= z\boldsymbol{x}_i'\tau^h - z\boldsymbol{x}_j'\tau^h = (z\boldsymbol{x}_i' - z\boldsymbol{x}_j')\tau^h, \qquad (4.17)
\end{aligned}$$

which was to be shown. □

The above corollaries assume that the DRF is a linear regression function. Although unrealistic, it simplifies the derivations so that the focus is on the intuition. Corollary 4.4 shows that the AGTE is equal to the ATE when Z is binary. Thereby the AGTE is a generalization of the ATE. Corollary 4.5 shows that for a linear f, the AGTE is a linear function of τ and the treatment levels for the estimand. Corollary 4.6 shows that the expected heterogeneity in the treatment effect between two different individuals is also a linear function if f is linear.

The counterfactual prediction method allows for estimation of effects from multiple treatments. The value of the evaluated DRF is a scalar no matter the number of treatments and confounders. Thereby estimation of the AGTE and CAGTE for multiple treatments is similar to the single treatment case. The following corollary shows that the counterfactual prediction method identifies the individual treatment or joint treatment CAGTE from several individual-specific treatment levels.

Corollary 4.7. *Let $\{\boldsymbol{z}_i, \boldsymbol{z}_{0,i}\} \in \mathbb{R}^k$ be values of k continuous treatments for individual i from the $n \times k$ multivariate treatment regimes $\boldsymbol{Z}$ and $\boldsymbol{Z}_0$. Suppose that the true* DRF *is a linear regression function without treatment interactions. If $z_{il} \neq z_{0,il}$ for all k treatments, then the* CAGTE *of $\boldsymbol{z}_i$ relative $\boldsymbol{z}_{0,i}$ is given by $\sum_{l=1}^{k} \tau_l(z_{il} - z_{0,il}) = \boldsymbol{\tau}'(\boldsymbol{z}_i - \boldsymbol{z}_{0,i})$. If only treatment l differ between the regimes, that is,*

$$\boldsymbol{Z} = \begin{bmatrix} z_{11} & z_{12} & \cdots & z_{1l} & \cdots & z_{1k} \\ z_{21} & z_{22} & \cdots & z_{1l} & \cdots & z_{2k} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ z_{n1} & z_{n2} & \cdots & z_{nl} & \cdots & z_{nk} \end{bmatrix}, \quad \boldsymbol{Z}_0 = \begin{bmatrix} z_{11} & z_{12} & \cdots & z_{0,1l} & \cdots & z_{1k} \\ z_{21} & z_{22} & \cdots & z_{0,1l} & \cdots & z_{2k} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ z_{n1} & z_{n2} & \cdots & z_{0,nl} & \cdots & z_{nk} \end{bmatrix},$$

then the CAGTE *of treatment l is $\tau_l(z_{il} - z_{0,il})$.*

Proof. Assume that $f(\boldsymbol{z}_i, \boldsymbol{x}_i) = \alpha + \boldsymbol{z}_i'\boldsymbol{\tau} + \boldsymbol{x}_i'\boldsymbol{\beta}$ is true. For $z_{il} \neq z_{0,il}$ for $l = 1, \ldots, k$, we have that

$$\begin{aligned}
\tau_{CAGTE}(\boldsymbol{Z}, \boldsymbol{Z}_0, \boldsymbol{x}_i) &= \mathbb{E}[Y_i(\boldsymbol{z}_i) - Y_i(\boldsymbol{z}_{0,i}) | \boldsymbol{X} = \boldsymbol{x}_i] \\
&= (\alpha + \boldsymbol{z}_i'\boldsymbol{\tau} + \boldsymbol{x}_i'\boldsymbol{\beta}) - (\alpha + \boldsymbol{z}_{0,i}'\boldsymbol{\tau} + \boldsymbol{x}_i'\boldsymbol{\beta}) \\
&= \boldsymbol{z}_i'\boldsymbol{\tau} - \boldsymbol{z}_{0,i}'\boldsymbol{\tau} \\
&= \boldsymbol{z}_i'\boldsymbol{\tau} - \boldsymbol{z}_{0,i}'\boldsymbol{\tau} = \boldsymbol{\tau}'(\boldsymbol{z}_i - \boldsymbol{z}_{0,i}), \qquad (4.18)
\end{aligned}$$

which was to be shown. We now consider the case that only treatment l differ between $\boldsymbol{Z}$ and $\boldsymbol{Z}_0$. If we decompose $\sum_{l=1}^{k} \tau_l z_{il}$ into $\tau_l z_{il} + \sum_{s \neq l}^{k} \tau_s z_{is}$ then

$$\begin{aligned}
\tau_{CAGTE}^{l}(\boldsymbol{Z}, \boldsymbol{Z}_0, \boldsymbol{x}_i) &= \mathbb{E}[Y_i(\boldsymbol{z}_i) - Y_i(\boldsymbol{z}_{0,i}) | \boldsymbol{X} = \boldsymbol{x}_i] \\
&= (\alpha + \sum_{l=1}^{k} \tau_l z_{il} + \boldsymbol{x}_i'\boldsymbol{\beta}) - (\alpha + \tau_l z_{0,il} + \sum_{s \neq l}^{k} \tau_s z_{is} + \boldsymbol{x}_i'\boldsymbol{\beta}) \\
&= \sum_{l=1}^{k} \tau_l z_{il} - \tau_l z_{0,il} - \sum_{s \neq l}^{k} \tau_s z_{is} \\
&= \tau_l z_{il} + \sum_{s \neq l}^{k} \tau_s z_{is} - \tau_l z_{0,il} - \sum_{s \neq l}^{k} \tau_s z_{is} \\
&= \tau_l z_{il} - \tau_l z_{0,il} = \tau_l (z_{il} - z_{0,il}), \qquad (4.19)
\end{aligned}$$

which was also to be shown. □

Remark. The joint CAGTET for k treatments relative no treatment on any of the treatments, meaning $\boldsymbol{z}_{0,i} = \boldsymbol{0}$, is $\boldsymbol{z}_i'\boldsymbol{\tau}$. The CAGTET of treatment l relative no treatment on treatment l, meaning $z_{0,il} = 0$, is $\tau_l z_{il}$.

All of the above derivations show that the CAGTE simplifies to a linear function of τ when f is a linear regression function. Such simplifications do not arise for a non-linear DRF. For instance, even in the simple non-linear case that $f(z_i, \boldsymbol{x}_i) = \exp(\alpha + \tau z_i + \boldsymbol{x}_i'\boldsymbol{\beta})$ with $z \in \{0, 1\}$, we get that

$$\tau_{CAGTE}(z, z_0, \boldsymbol{x}_i) = \mathbb{E}[\exp(\alpha + \tau + \boldsymbol{x}_i'\boldsymbol{\beta} + \varepsilon_i) - \exp(\alpha + \boldsymbol{x}_i'\boldsymbol{\beta} + \varepsilon_{0,i})],$$

which cannot be simplified further.

4.3 Estimation

Estimating the ADRF. A problem in estimating the ADRF is that only one potential outcome is observed for each individual. Since individuals may self-select the treatment level, the regression curve $f(Z, \boldsymbol{X})$ does not have a causal interpretation and is in general not equal to the causal ADRF given by $\mathbb{E}[f(Z, \boldsymbol{X})]$. However, if Z would be independent of $Y(Z)$ as in an RCT, then the individuals at each value z would be a random draw from the population and the regression curve would coincide with the ADRF. Although we can state a mathematical expression of the ADRF, it is not obvious how to estimate it in observational studies. Our solution is that by Proposion 4.2, the DRF can predict the potential outcomes for the treatment levels that each individual in the sample did not receive. Then the potential outcome paths can be estimated for all individuals and we can average the predicted potential outcomes $\widehat{y}_i(z) = f(z, \boldsymbol{x}_i)$ for all $z \in \mathcal{Z}$ over the empirical distribution of $\boldsymbol{X}$ to obtain the ADRF[5]. The result will be the estimated sample ADRF.

Definition 4.5 (Counterfactual Prediction ADRF Estimator)**.** The counterfactual prediction estimator of the ADRF at treatment level z is

$$\widehat{\mu}_f(z) \coloneqq \frac{1}{n} \sum_{i=1}^{n} f(z, \boldsymbol{x}_i). \tag{4.20}$$

It is clear that predicting $y_i(z)$ for a larger number of values of Z when using this estimator results in a smoother curve for the estimated ADRF.

Estimating the AGTE. The estimator of AGTE is obtained by evaluating the ADRF at the two levels for which want to estimate the AGTE and taking their difference.

Definition 4.6 (Counterfactual Prediction AGTE Estimator)**.** The counterfactual prediction estimator of the AGTE of treatment level z relative z_0 is

$$\widehat{\tau}_{AGTE}(z, z_0) \coloneqq \frac{1}{n} \sum_{i=1}^{n} \Big(f(z, \boldsymbol{x}_i) - f(z_0, \boldsymbol{x}_i) \Big). \tag{4.21}$$

[5] We will in some contexts write $\widehat{y}_i(z)$ for $f(z, \boldsymbol{x}_i)$ to emphasize that the expected value of a potential outcome as given by the DRF is a predicted value.

Remark. The estimator of the CAGTE at $\boldsymbol{X} = \boldsymbol{x}$ is obtained by fixing $\boldsymbol{X}$ at $\boldsymbol{x}$:

$$\widehat{\tau}_{CAGTE}(z, z_0, \boldsymbol{x}) := \frac{1}{n} \sum_{i=1}^{n} \Big(f(z, \boldsymbol{x}) - f(z_0, \boldsymbol{x}) \Big). \tag{4.22}$$

Heterogeneity in treatment effects is estimated by taking the difference between $f(z, \boldsymbol{x}_i)$ and $f(z, \boldsymbol{x}_j)$, that is, the potential outcomes at the same treatment level for different individuals with a DRF that captures the confounder-treatment interaction.

Estimating the AGTET. To estimate (4.14), compute the average difference between the fitted values of the realized outcomes $\widehat{y}_i$ and $f(0, \boldsymbol{x}_i)$ over the individuals that received treatment. However, since the effect is estimated for the observed treatments for which we know the true values, it might make more sense to use the realized potential outcomes rather than their predicted values. Thus, an alternative but equivalent estimator weights the predicted counterfactuals with the residual. Specifically, we have that $y_i(z_i) = f(z_i, \boldsymbol{x}_i) + e_i$ and $\widehat{y}_i(z_i) = f(z_i, \boldsymbol{x}_i)$ since $\mathbb{E}[\varepsilon_i] = 0$. Because $y_i(z_i) - \widehat{y}_i(z_i) = e_i \neq 0$, the prediction $f(0, \boldsymbol{x}_i)$ should be weighted with e_i if we wish to compare it to the observed outcome y_i under the status quo treatment.

Definition 4.7 (COUNTERFACTUAL PREDICTION AGTET ESTIMATOR)**.** The counterfactual prediction estimator of the AGTET of the observed treatments relative no treatment is given by

$$\widehat{\tau}^{(adj)}_{AGTET}(Z, 0) := \frac{1}{\sum_{i=1}^{n} \mathbb{1}(z_i > 0)} \sum_{i:z_i>0} \Big(y_i - f^{(adj)}(0, \boldsymbol{x}_i) \Big). \tag{4.23}$$

Remark. Let $\{z_i\}_{i=1}^{n}$ be the status quo treatments. The estimator is derived as:

$$\begin{aligned}
\widehat{\tau}^{(adj)}_{AGTET}(Z, 0) &:= \frac{1}{\sum_{i=1}^{n} \mathbb{1}(z_i > 0)} \sum_{i:z_i>0} \left(\Big(f(z_i, \boldsymbol{x}_i) - f(0, \boldsymbol{x}_i) \Big) \times \frac{y_i}{f(z_i, \boldsymbol{x}_i)} \right) \\
&= \frac{1}{\sum_{i=1}^{n} \mathbb{1}(z_i > 0)} \sum_{i:z_i>0} \left(y_i - f(0, \boldsymbol{x}_i) \times \frac{y_i}{f(z_i, \boldsymbol{x}_i)} \right) \\
&= \frac{1}{\sum_{i=1}^{n} \mathbb{1}(z_i > 0)} \sum_{i:z_i>0} \Big(y_i - f^{(adj)}(0, \boldsymbol{x}_i) \Big).
\end{aligned} \tag{4.24}$$

Here, $y_i / f(z, \boldsymbol{x}_i)$ is the error adjustment factor that weights the predicted counterfactual potential outcome $\widehat{y}_i(0) = f(0, \boldsymbol{x}_i)$ with the prediction error e_i and $f^{(adj)}(\cdot, \boldsymbol{X}) = f(\cdot, \boldsymbol{X}) \times (Y / f(Z, \boldsymbol{X}))$ is the error adjusted DRF.

Since the average is taken over the scalars $\{f(z_i, \boldsymbol{x}_i)\}_{i=1}^{n}$ these estimators works for arbitrarily high-dimensional confounding as long as $f(z, \boldsymbol{x}_i)$ is estimable.

External validity. The obtained estimates may not approximate the population quantities well if the values of Z and $\boldsymbol{X}$ in the sample are different from those in the population. The precision of a predicted potential outcome $f(z, \boldsymbol{x})$ for values $(z, \boldsymbol{x})$ not in the data depends on if there exists an individual i in the data with values $(z_i, \boldsymbol{x}_i)$ that are similar to $(z, \boldsymbol{x})$. In other words, the true values of predicted potential outcomes for values of Z and $\boldsymbol{X}$ outside the space of the data are unknown. Thereby the errors of such predictions are also unknown. This problem is not specific to causal inference, but arises in any out-of-sample prediction. The estimators may then be used only for sample inference or the population can be redefined to a subgroup that the data cover. If interest is still in the population effect, the user may generate synthetic data $\tilde{\mathcal{D}} = \{(\tilde{z}_j, \tilde{\boldsymbol{x}}_j)\}_{j=1}^{\tilde{n}}$, $j \neq i \in \mathcal{D}$, for $\tilde{n}$ hypothetical individuals with values of Z and $\boldsymbol{X}$ that approximate the values of the individuals in population that are missing in the sample, and then estimate the quantity of interest on the redefined data given by $\mathcal{D} \cup \tilde{\mathcal{D}}$.

Estimand precision. We now turn to validation of the causal estimands. Treatment effects most often varies with individuals. Hence, the optimal DRF is the one that best captures the treatment effect heterogeneity, as measured by the variance in CAGTE over $\boldsymbol{X}$. More formally, the optimal DRF is the one that for some $z, z_0 \in \mathcal{Z}$ minimizes the τ-risk (Hill, 2011; Schuler et al., 2018) of f given by

$$\tau\text{-risk} := \int_{\mathcal{X}} \Big(\tau_{CAGTE}(z, z_0, \boldsymbol{x}) - \widehat{\tau}_{CAGTE}(z, z_0, \boldsymbol{x})\Big)^2 dP_{\boldsymbol{X}}(\boldsymbol{x}). \tag{4.25}$$

By the fundamental problem of causal inference the true $\tau_{CAGTE}(z, z_0, \boldsymbol{x})$ cannot be observed. Hence the τ-risk is inestimable. A feasible compromise is to validate the estimated causal quantities on the basis of f-risk shown in expression (4.11). The justification for substituting the τ-risk with the f-risk is that a model that perfectly predicts potential outcomes also perfectly models the CATGE. However, since the f-risk metric is based on the DRF it suffers from that it cannot measure the precision of predictions of potential outcomes not in the data. As mentioned, this means that an estimated CATGE for values of Z and $\boldsymbol{X}$ not in the data has unknown error. We refer to Schuler et al. (2018) for alternative methods that alleviate this problem.

4.4 Implementation

Using the estimators requires counterfactual potential outcomes. Algorithm 1 shows a procedure for predicting counterfactual potential outcomes for any treatment level. Here, $\boldsymbol{y}$ denotes the $n \times 1$ vector of realized outcomes and $\boldsymbol{y}(z)$ denotes a vector of the predicted outcomes.

Algorithm 1: Counterfactual potential outcome prediction

```
Input  : A DRF $f$, data $\mathcal{D} = \{(y_i, z_i, \boldsymbol{x}_i)\}_{i=1}^n$ and a finite set $\tilde{\mathcal{Z}} \subset \mathcal{Z}$ of
         treatment levels of a single treatment.
Output: Predicted counterfactual outcomes $\widehat{\boldsymbol{y}}(\boldsymbol{z}^{(cf)})$.

Fit $f(\boldsymbol{z}, \boldsymbol{X})$ to the observed outcomes $\boldsymbol{y} \in \mathcal{D}$.
foreach $z \in \tilde{\mathcal{Z}}$ do
    Set $\mathcal{D}^{(cf)} \leftarrow \mathcal{D}$
    foreach $z_i^{(cf)} \in \mathcal{D}^{(cf)}, i = 1, \ldots, n$ do
        Set $z_i^{(cf)} \leftarrow z$
    end
    return $\widehat{\boldsymbol{y}}(z^{(cf)}) = f(z^{(cf)}, \boldsymbol{X})$ by predicting $\boldsymbol{y}^{(cf)}$ given data $\mathcal{D}^{(cf)}$
end
```

Algorithm 1 consists of two stages: In the first step, the DRF is fitted to the observed data $\mathcal{D}$. If f is parametric, the parameters of f are estimated. In the second stage, the counterfactual potential outcomes are estimated. The user has supplied a set $\tilde{\mathcal{Z}}$ of $|\tilde{\mathcal{Z}}|$ treatment levels for which potential outcomes should be estimated. To estimate the AGTE of two counterfactual treatment levels, the user supplies z_1 and z_2 so that $|\tilde{\mathcal{Z}}| = 2$. To estimate the AGTET relative no treatment, set $z = 0$. To estimate the ADRF, supply a sufficient amount of treatment levels to make the estimated ADRF smooth. For each level $z \in \tilde{\mathcal{Z}}$, the observed data $\mathcal{D}$ are copied to the counterfactual data $\mathcal{D}^{(cf)}$. The dose z is assigned to $\mathcal{D}^{(cf)}$ and the predicted potential outcomes $\widehat{y}_i(z) = f(z, \boldsymbol{x}_i)$ are obtained by estimating the responses using $\mathcal{D}^{(cf)}$. The process is repeated for the next z in $\tilde{\mathcal{Z}}$.

Algorithm 2 estimates potential outcomes from multiple treatments. We consider estimation of two types of potential outcomes in the multiple treatment case. The first is the potential outcome corresponding to when one treatment has a counterfactual level but the remaining treatments are as observed. The second is the potential outcome corresponding to that all treatments are set to counterfactual levels.

It is difficult to imagine an ADRF in the multiple treatment setting. Even if a multiple treatment ADRF would be of practical use, it would require a very large number of predictions, since for each treatment the user must estimate the potential outcomes from all elements of $\tilde{\mathcal{Z}}$. The multiple treatment setting also creates the possibility of different combinations of counterfactuals across treatments, which further increases the number of required predictions. Thereby Algorithm 2 is meant for estimation of the AGTE or CAGTE, not the ADRF.

Algorithm 2: Multi-treatment counterfactual potential outcome prediction

```
Input : DRF $f$, data $\mathcal{D} = \{(y_i, \boldsymbol{z}_i, \boldsymbol{x}_i)\}_{i=1}^n$, and a $k$-dimensional vector of
        treatments $\boldsymbol{z}_0$ containing doses $\{z_{0,l}\}_{l=1}^k$.
Output: Predicted counterfactual outcomes $\widehat{\boldsymbol{y}}(\boldsymbol{z}_l^{(cf)})$ for each treatment
        $l = 1, \ldots, k$ and predicted counterfactual outcomes $\boldsymbol{y}(\boldsymbol{z}^{(cf)})$ of
        simultaneous counterfactual treatments on all $k$ treatments.

Fit $f(\boldsymbol{z}, \boldsymbol{X})$ to the observed outcomes $\boldsymbol{y} \in \mathcal{D}$.
foreach $z_l \in \boldsymbol{z}, l = 1, \ldots, k$ do
    Set $\mathcal{D}_l^{(cf)} \leftarrow \mathcal{D}$
    foreach $z_{il}^{(cf)} \in \mathcal{D}_l^{(cf)}, i = 1, \ldots, n$ do
        Set $z_{il}^{(cf)} \leftarrow z_{0,l}$
    end
    return $\widehat{\boldsymbol{y}}(\boldsymbol{z}^{(cf)}) = f(\boldsymbol{z}^{(cf)}, \boldsymbol{X})$ by predicting $\boldsymbol{y}^{(cf)}$ given data $\mathcal{D}_l^{(cf)}$
end
Set $\mathcal{D}^{(cf)} \leftarrow \mathcal{D}$
foreach $z_l \in \boldsymbol{z}, l = 1, \ldots, k$ do
    foreach $z_{il}^{(cf)} \in \mathcal{D}^{(cf)}, i = 1, \ldots, n$ do
        Set $z_{il}^{(cf)} \leftarrow z_{0,l}$
    end
end
return $\widehat{\boldsymbol{y}}(\boldsymbol{z}_0) = f(\boldsymbol{z}^{(cf)}, \boldsymbol{X})$ by predicting $\boldsymbol{y}^{(cf)}$ given data $\mathcal{D}^{(cf)}$
```

Algorithm 2 consists of three stages: In the first stage, f is fitted to the observed outcomes $\boldsymbol{y}(\boldsymbol{z})$. Here, $\boldsymbol{y}(\boldsymbol{z})$ denotes the $n \times 1$ vector of potential outcomes realized under the k treatments the individual are exposed to. In the algorithm's second stage, counterfactual potential outcomes are estimated for the case that treatment l would have had level $z_{0,l}$ instead of the observed z_l but the remaining $k-1$ treatments would have had their respective observed treatment levels. First, the copy data $\mathcal{D}_l^{(cf)}$ of $\mathcal{D}$ are assigned the supplied counterfactual level $z_{0,l}$ for treatment l. The counterfactual potential outcomes $\widehat{\boldsymbol{y}}(z^{(cf)})$ are estimated using f on $\mathcal{D}_l^{(cf)}$. Stage two is repeated for each of the k treatments. In the third stage, the counterfactual potential outcomes are estimated for the case that all k treatments would jointly receive their respective counterfactual treatment levels. A copy dataset $\mathcal{D}^{(cf)}$ of the observed data $\mathcal{D}$ is created and all observed k treatment levels are replaced with their respective counterfactual treatment levels. After all treatment levels have been replaced, the predicted potential outcomes $\widehat{\boldsymbol{y}}(\boldsymbol{z}^{(cf)})$ are obtained with by predicting the responses using the counterfactual data $\mathcal{D}^{(cf)}$. If $k = 1$, meaning that there is only one treatment, then stage two and three will perform the same procedure.

For error adjusted estimation of the AGTET of the status quo treatment regime relative no treatment, the returned expected counterfactual potential outcomes $\widehat{\boldsymbol{y}}(\boldsymbol{z}_0) = f(\boldsymbol{z}^{(cf)}, \boldsymbol{X})$ should be multiplied with $\boldsymbol{y}(\boldsymbol{z})/\widehat{\boldsymbol{y}}(\boldsymbol{z}) = \boldsymbol{y}(\boldsymbol{z})/f(\boldsymbol{z}, \boldsymbol{X})$, that is, the observed outcomes divided by the fitted values of the observed outcomes.

Data-splitting approach. Data-splitting can be used in the prediction of the counterfactual outcomes and evaluation of the causal estimands. The user supplies training data $\mathcal{T}$ and validation data $\mathcal{V}$ as inputs to the algorithms instead of the full sample of observed data $\mathcal{D}$. In the first stage of the algorithms, f is fitted to $\mathcal{T}$. In the second stage, all rows and columns of $\mathcal{V}$ are assigned to $\mathcal{V}^{(cf)}$, or $\mathcal{V}_l^{(cf)}$ in the case of stage 2 of the Algorithm 2. Then the counterfactual levels of the treatments are assigned to these counterfactual validation datasets from which the counterfactual potential outcomes are predicted. The ADRF, AGTE or AGTET is then estimated on the observations in the validation data.

If the data-splitting approach is used, the inferential methods explained in the next section should also be estimated on the validation data. This is done simply by evaluating the estimators over the observations in the validation set. Since the principle of the estimators are the same but the notation becomes more complex with data-splitting, all methods are presented as if data-splitting is not used.

4.5 Inference

The estimated ADRF or AGTE is a random variable. Thus, statements about the population value of the estimand requires an uncertainty interval[6]. Using asymptotic theory, we can obtain normal-based asymptotically valid uncertainty intervals. Formally, let τ denote the true value of any of the causal estimands and let $\lambda_{\alpha/2}$ be the $\alpha/2$ quantile of the t-distribution. We wish to estimate an interval such that

$$\mathbb{P}\Big(\tau \in \Big[\widehat{\tau} \pm \lambda_{\alpha/2} \times \mathbb{S}[\widehat{\tau}]\Big]\Big) = 1 - \alpha, \tag{4.26}$$

where $\mathbb{S}[\widehat{\tau}] \coloneqq \textsc{se}_{\widehat{\tau}}$ is the standard error of the estimate. We first consider the ADRF.

Let z be the dose of a single treatment. Given that the observed outcomes $\{Y_i\}_{i=1}^n$ are i.i.d. and f is $\sqrt{n}$-consistent, then by the law of large numbers (LLN)[7],

$$\widehat{\mu}_f(z) \coloneqq \frac{1}{n}\sum_{i=1}^n f(z, \boldsymbol{x}_i) \xrightarrow{a.s.} \mathbb{E}[f(z, \boldsymbol{X})] \coloneqq \mu(z), \quad \text{as } n \to \infty, \tag{4.27}$$

and by the central limit theorem (CLT),

$$\sqrt{n}\big(\widehat{\mu}_f(z) - \mu(z)\big) \xrightarrow{d.} \mathcal{N}\big(0, \mathbb{V}[\widehat{\mu}_f(z)]\big), \quad \text{as } n \to \infty, \tag{4.28}$$

where $\mathbb{V}[\widehat{\mu}_f(z)]$ is the asymptotic variance for the estimated ADRF evaluated at the treatment level z.

Proposition 4.8. *A two-sided* $(1-\alpha)\times 100$ *percent confidence interval for the* ADRF *at* $Z = z$ *is given by*

$$\mathcal{I}_{1-\alpha}^{\mu(z)} \coloneqq \widehat{\mu}_f(z) \pm \lambda_{\alpha/2} \times \mathbb{S}[\widehat{\mu}_f(z)], \tag{4.29}$$

where $\lambda_{\alpha/2} \coloneqq t_{n-(p+k)}(\alpha/2)$ *is the* $\alpha/2$ *quantile of the* t*-distribution with* $n-(p+k)$ *degrees of freedom and* $\mathbb{S}[\widehat{\mu}_f(z)]$ *is the asymptotic standard error of* $\widehat{\mu}_f(z)$*.*

The sample ADRF at z is computed by averaging $\{f(z, \boldsymbol{x}_i)\}_{i=1}^n$ over the empirical distribution of $\boldsymbol{X}$. This means that the standard error of the estimated sample ADRF at z is equal to the standard error of $f(z, \boldsymbol{x}_i) = \widehat{y}_i(z)$, where the variability stems from the sample variation in $\boldsymbol{X}$. We will return how to estimate this.

[6] We use the term uncertainty interval to refer to any type of interval estimate that reflects the variability in the estimate around the population value, here confidence and prediction intervals.

[7] We use the notation $\xrightarrow{a.s.}$ for almost sure convergence and $\xrightarrow{d.}$ for convergence in distribution.

We now consider confidence intervals of the AGTE and AGTET in the general case that there are k treatments[8]. Assuming again that $\{Y_i\}_{i=1}^n$ are i.i.d. and f is $\sqrt{n}$-consistent, then by the LLN and Slutsky's theorem,

$$\widehat{\tau}_{AGTET}(\boldsymbol{Z}, \boldsymbol{z}_0) := \frac{1}{n}\sum_{i=1}^{n}\Big(\widehat{y}_i(\boldsymbol{z}_i) - \widehat{y}_i(\boldsymbol{z}_0)\Big) \xrightarrow{a.s.} \tau_{AGTET}(\boldsymbol{Z}, \boldsymbol{z}_0), \quad \text{as } n \to \infty. \tag{4.30}$$

Here, $\tau_{AGTET}(\boldsymbol{Z}, \boldsymbol{z}_0)$ denotes the AGTET of the status quo multivariate treatment regime $\boldsymbol{Z}$ of k treatments relative k constant counterfactual levels $\boldsymbol{z}_0$. By the CLT,

$$\sqrt{n}(\widehat{\tau}_{AGTET}(\boldsymbol{Z}, \boldsymbol{z}_0) - \tau_{AGTET}(\boldsymbol{Z}, \boldsymbol{z}_0)) \xrightarrow{d.} \mathcal{N}(0, \mathbb{V}[\widehat{\tau}_{AGTET}(\boldsymbol{Z}, \boldsymbol{z}_0)]), \quad \text{as } n \to \infty \tag{4.31}$$

where $\mathbb{V}[\widehat{\tau}_{AGTET}(\boldsymbol{Z}, \boldsymbol{z}_0)]$ is the asymptotic variance of the estimated AGTET.

The AGTE and AGTET are given by the difference between the averages of two vectors of potential outcomes. Thus, the uncertainty of the estimands is given by the uncertainty of the average of the predicted potential outcomes. Let $\mathbf{1}_n = n^{-1}\mathbf{1}$ be an n-dimensional vector of ones divided by n so that $n^{-1}\sum_{i=1}^{n}\widehat{y}_i(\boldsymbol{z}_0) = \mathbf{1}_n'\widehat{\boldsymbol{y}}(\boldsymbol{z}_0)$ is the average predicted potential outcome given the k treatment levels $\boldsymbol{z}_0$.

Proposition 4.9. *A two-sided* $(1-\alpha)\times 100$ *percent confidence interval of the AGTET of the observed treatment regime* $\boldsymbol{Z}$ *relative counterfactual levels* $\boldsymbol{z}_0$ *is given by*

$$\mathcal{I}_{1-\alpha}^{\tau_{AGTET}(\boldsymbol{Z},\boldsymbol{z}_0)} := \widehat{\tau}_{AGTET}(\boldsymbol{Z}, \boldsymbol{z}_0) \pm \lambda_{\alpha/2} \times \mathbb{S}[\mathbf{1}_n'\widehat{\boldsymbol{y}}(\boldsymbol{z}_0)], \tag{4.32}$$

where $\mathbb{S}[\mathbf{1}_n'\widehat{\boldsymbol{y}}(\boldsymbol{z}_0)]$ *is the asymptotic standard error of* $\mathbf{1}_n'\widehat{\boldsymbol{y}}(\boldsymbol{z}_0)$.

Proposition 4.9 states that if we estimate the AGTET of the status quo multivariate treatment regime $\boldsymbol{Z}$ relative counterfactual treatment levels $\boldsymbol{z}_0$, where $\boldsymbol{z}_0 = \mathbf{0}$ to get the AGTET relative no treatment, then the variance of the estimated AGTET is given by the variance of $\mathbf{1}_n'\widehat{\boldsymbol{y}}(\boldsymbol{z}_0)$. The reason is that if $\boldsymbol{Z}$ is observed then $\boldsymbol{y}(\boldsymbol{Z})$, and hence $\mathbf{1}_n'\boldsymbol{y}(\boldsymbol{Z})$, is known and contains no uncertainty. This is seen by writing

$$\begin{aligned}\mathbf{1}_n'(\boldsymbol{y}(\boldsymbol{Z}) - \widehat{\boldsymbol{y}}(\boldsymbol{Z})) &= \frac{1}{n}\sum_{i=1}^{n} y_i(\boldsymbol{z}_i) - \frac{1}{n}\sum_{i=1}^{n}\widehat{y}_i(\boldsymbol{z}_i) \\ &= \frac{1}{n}\sum_{i=1}^{n}\Big(f(\boldsymbol{z}_i, \boldsymbol{x}_i) + e_i\Big) - \frac{1}{n}\sum_{i=1}^{n} f(\boldsymbol{z}_i, \boldsymbol{x}_i) = \frac{1}{n}\sum_{i=1}^{n} e_i = 0,\end{aligned}$$

[8] The asymptotic results for the AGTET presented in this section apply to the AGTE by replacing the observed outcomes y_i in (4.30) and (4.31) with the predicted outcomes $\widehat{y}_i(\boldsymbol{z})$.

where the last equality follows from that the sum of the residuals is zero. If we on the other hand want to estimate the AGTE of two treatments levels $\boldsymbol{z}$ and $\boldsymbol{z}_0$ constant across individuals, then both $\boldsymbol{y}(\boldsymbol{z})$ and $\boldsymbol{y}(\boldsymbol{z}_0)$ are unobserved and needs to be predicted. Now both components of the AGTE are uncertain. Hence the variance of the AGTE is on average wider than the variance of the AGTET.

Proposition 4.10. *A confidence interval for the AGTE of two treatment levels is on average wider than a confidence interval of the AGTET of the observed treatment regime relative counterfactual treatment levels.*

Proof. The AGTET of $\boldsymbol{Z}$ relative counterfactual treatment levels $\boldsymbol{z}_0$ can be written as the AGTE of $\boldsymbol{Z}$ relative $\boldsymbol{z}_0$ conditional on $\boldsymbol{Z}$. By the law of total variance,

$$\mathbb{V}[\widehat{\tau}_{AGTE}(\boldsymbol{Z}, \boldsymbol{z}_0)] = \mathbb{E}\{\mathbb{V}[\widehat{\tau}_{AGTE}(\boldsymbol{Z}, \boldsymbol{z}_0)|\boldsymbol{Z}]\} + \mathbb{V}\{\mathbb{E}[\widehat{\tau}_{AGTE}(\boldsymbol{Z}, \boldsymbol{z}_0)|\boldsymbol{Z}]\}. \tag{4.33}$$

Both terms on the right hand side are always positive and the first term is expected variance of the estimated AGTE when $\boldsymbol{Z}$ is the observed treatment regime. Hence

$$\mathbb{V}[\widehat{\tau}_{AGTE}(\boldsymbol{Z}, \boldsymbol{z}_0)] > \mathbb{E}\{\mathbb{V}[\widehat{\tau}_{AGTE}(\boldsymbol{Z}, \boldsymbol{z}_0)|\boldsymbol{Z}]\}, \tag{4.34}$$

which we were supposed to show. □

This proves that a confidence interval for the AGTE of two treatment levels whose outcomes for all individuals are unknown is on average wider than the AGTET of the observed treatment regime whose outcomes are known. In (4.33), the second term on the right hand side gives the size of the reduction in variance by observing $\boldsymbol{Z}$, which in practical applications will be of interest since it provides a measure of the precision of the estimated AGTET.

Estimating the standard error. We now present how to estimate the unknown standard errors $\mathbb{S}[\widehat{\mu}_f(z)]$ and $\mathbb{S}[\mathbf{1}_n'\widehat{\boldsymbol{y}}(\boldsymbol{z})]$, which can be used to estimate an uncertainty interval for the ADRF, AGTE or AGTET.

Proposition 4.11. *The prediction variance of the average of the predicted potential outcomes given k treatment levels $\boldsymbol{z}$ is given by*

$$\widehat{\mathbb{V}}\left[\frac{1}{n}\sum_{i=1}^{n}\widehat{y}_i(\boldsymbol{z})\right] = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\widehat{\sigma}_{ij}. \tag{4.35}$$

Proof. Let $\widehat{\boldsymbol{\Sigma}}$ with elements $\widehat{\sigma}_{ij}$ be the estimated variance-covariance matrix of the n-dimensional vector $\widehat{\boldsymbol{y}}(\boldsymbol{z})$ of predicted potential outcomes. The rules for the covariance of a linear combination give that

$$\begin{aligned}\widehat{\mathbb{V}}\left[\frac{1}{n}\sum_{i=1}^{n}\widehat{y}_i(\boldsymbol{z})\right] &= \widehat{\mathbb{V}}[\mathbf{1}_n'\widehat{\boldsymbol{y}}(\boldsymbol{z})] = \mathbf{1}_n'\widehat{\mathbb{V}}[\widehat{\boldsymbol{y}}(\boldsymbol{z})]\mathbf{1}_n = \mathbf{1}_n'\widehat{\mathbb{V}}\{\mathbb{E}[\boldsymbol{y}(\boldsymbol{z})|\boldsymbol{X}]\}\mathbf{1}_n \\ &= \mathbf{1}_n'\widehat{\boldsymbol{\Sigma}}\mathbf{1}_n = \frac{1}{n^2}\sum_{i=1}^{n}\widehat{\sigma}_i^2 + \frac{2}{n^2}\sum_{i,j:i<j}\widehat{\sigma}_{ij} = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\widehat{\sigma}_{ij}, \qquad (4.36)\end{aligned}$$

where the third equality follows from that $\widehat{\boldsymbol{y}}(\boldsymbol{z}) = f(\boldsymbol{z}, \boldsymbol{X}) = \mathbb{E}[\boldsymbol{y}(\boldsymbol{z})|\boldsymbol{X}]$, the fourth from that $\mathbb{E}[\boldsymbol{y}(\boldsymbol{z})|\boldsymbol{X}] = \big(\mathbb{E}[y_1(\boldsymbol{z})|\boldsymbol{x}_1], \mathbb{E}[y_2(\boldsymbol{z})|\boldsymbol{x}_2], \ldots, \mathbb{E}[y_n(\boldsymbol{z})|\boldsymbol{x}_n]\big)'$ is the $n\times 1$ vector of all individuals' expected potential outcomes given $\boldsymbol{z}$ and $\boldsymbol{X}$, and where in the last equality $\widehat{\sigma}_{ij} = \widehat{\sigma}_i^2$ for all $i = j \leq n$. □

Remark. Expression (4.35) shows that the prediction variance is the sum of all elements in the estimated covariance matrix divided by n^2. The covariances are only equal to zero if $\widehat{y}_i(\boldsymbol{z})$ and $\widehat{y}_j(\boldsymbol{z})$, $i \neq j$, are independent, which in general they are not even when the observed potential outcomes $y_i(\boldsymbol{z})$ and $y_j(\boldsymbol{z})$ are independent since the predicted values are generated from the same DRF. For instance, if the DRF is parametric, then the predicted values share parameter estimates. Although $\boldsymbol{z}$ is a k-dimensional vector, the method also applies to estimation of $\mathbb{S}[\mu_f(z)]$ since in that case k equals one. This also holds for the remaining methods in this section.

Taking the square root of expression (4.35) yields an estimate of the standard error of the average of the potential outcomes given by the levels of the k treatments. That is, $\mathbb{S}[\mathbf{1}_n'\widehat{\boldsymbol{y}}(\boldsymbol{z})] \approx \sqrt{\mathbf{1}_n'\widehat{\boldsymbol{\Sigma}}\mathbf{1}_n}$. The estimated standard error will only be an approximation of the true standard error since the true standard error is unknown and its estimate depends on 1) the size of the sample, 2) the sampling of individuals and their values on the confounders, and 3) the choice of f.

Note that this estimator of the standard error makes no assumptions about Y. Hence by Proposition 4.11, a point and interval estimate of the estimands is non-parametrically estimable without any distributional assumptions of Y. Standard bootstrap procedures are also applicable for non-parametric interval estimation.

The problem with this estimator of the standard error is that for large n, the $n \times n$ covariance matrix $\boldsymbol{\Sigma}$ of the potential outcomes may be impossible to estimate. More efficient methods that does not involve $\boldsymbol{\Sigma}$ are possible for parametric DRF's. We illustrate the first method for a gaussian linear DRF.

Example 4.2. We wish to obtain an uncertainty interval of the estimated AGTET of the k-dimensional varying treatment regime $\boldsymbol{Z}$ relative k constant counterfactual treatment levels $\boldsymbol{z}_0$. Assume for simplicity that the true f is a linear model estimated with OLS. Then

$$\boldsymbol{y}(\boldsymbol{Z}) = f(\boldsymbol{Z}, \boldsymbol{X}) + \boldsymbol{\varepsilon} = \boldsymbol{X\beta} + \boldsymbol{Z\tau} + \boldsymbol{\varepsilon}$$

and

$$\boldsymbol{y}(\boldsymbol{z}_0) = f(\boldsymbol{z}_0, \boldsymbol{X}) + \boldsymbol{\varepsilon} = \boldsymbol{X\beta} + \boldsymbol{Z}_0\boldsymbol{\tau} + \boldsymbol{\varepsilon}$$

with $\boldsymbol{\varepsilon} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I})$ due to that $\{Y_i\}_{i=1}^n$ are i.i.d.[9]. Let $\boldsymbol{U} = [\boldsymbol{Z}_{(n\times k)}\ \boldsymbol{X}_{(n\times p)}]$ be the observed $n \times (p+k)$ design matrix of all p regressors in f from $\boldsymbol{X}$ and the n levels of all k treatments in $\boldsymbol{Z}$. Similarly, let $\boldsymbol{U}_0 = [\boldsymbol{Z}_{0(n\times k)}\ \boldsymbol{X}_{(n\times p)}]$ be the counterfactual design matrix that contains $\boldsymbol{X}$ and the counterfactual treatment levels $\boldsymbol{z}_0$ that are the same for all n individuals. Let $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \mathbb{R}^{p+k}$ be the matrix of parameters for $\boldsymbol{U}$ and $\boldsymbol{U}_0$ with the estimate $\widehat{\boldsymbol{\theta}} = (\boldsymbol{U}'\boldsymbol{U})^{-1}\boldsymbol{U}'\boldsymbol{y}$. Then $\boldsymbol{y}(\boldsymbol{Z}) = \boldsymbol{U\theta} + \boldsymbol{\varepsilon}$ and $\boldsymbol{y}(\boldsymbol{z}_0) = \boldsymbol{U}_0\boldsymbol{\theta} + \boldsymbol{\varepsilon}$, and the vector of predicted counterfactual outcomes is

$$\widehat{\boldsymbol{y}}(\boldsymbol{z}_0) = \boldsymbol{U}_0\widehat{\boldsymbol{\theta}} = \boldsymbol{U}_0(\boldsymbol{U}'\boldsymbol{U})^{-1}\boldsymbol{U}'\boldsymbol{y}(\boldsymbol{z}_0) = \boldsymbol{U}_0(\boldsymbol{U}'\boldsymbol{U})^{-1}\boldsymbol{U}'(\boldsymbol{U}_0\boldsymbol{\theta} + \boldsymbol{\varepsilon}).$$

Define $\sigma^2 = (n-(p+k))^{-1}\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}$ to be the error variance. Since $\boldsymbol{\theta}$ is a population constant with zero variance,

$$\begin{aligned}\mathbb{V}[\widehat{\boldsymbol{\theta}}] = \mathbb{V}[(\boldsymbol{U}'\boldsymbol{U})^{-1}\boldsymbol{U}'\boldsymbol{y}(\boldsymbol{z})] &= \mathbb{V}[(\boldsymbol{U}'\boldsymbol{U})^{-1}\boldsymbol{U}'(\boldsymbol{U\theta} + \boldsymbol{\varepsilon})] \\ &= (\boldsymbol{U}'\boldsymbol{U})^{-1}\boldsymbol{U}'\mathbb{V}[\varepsilon]\boldsymbol{U}(\boldsymbol{U}'\boldsymbol{U})^{-1} \\ &= \sigma^2(\boldsymbol{U}'\boldsymbol{U})^{-1}\boldsymbol{U}'\boldsymbol{U}(\boldsymbol{U}'\boldsymbol{U})^{-1} \\ &= \sigma^2(\boldsymbol{U}'\boldsymbol{U})^{-1},\end{aligned}$$

and the variance-covariance matrix of $\widehat{\boldsymbol{y}}(\boldsymbol{z}_0)$ is

$$\mathbb{V}[(\widehat{\boldsymbol{y}}(\boldsymbol{z}_0)] = \mathbb{V}[\boldsymbol{U}_0\widehat{\boldsymbol{\theta}}] = \boldsymbol{U}_0\mathbb{V}[\widehat{\boldsymbol{\theta}}]\boldsymbol{U}_0' = \sigma^2\boldsymbol{U}_0(\boldsymbol{U}'\boldsymbol{U})^{-1}\boldsymbol{U}_0' = \sigma^2\boldsymbol{P}_0, \tag{4.37}$$

where $\boldsymbol{P}_0 = \boldsymbol{U}_0(\boldsymbol{U}'\boldsymbol{U})^{-1}\boldsymbol{U}_0'$ is the projection matrix for the counterfactual potential outcomes. This gives the estimated prediction variance $\boldsymbol{1}_n'(\widehat{\sigma}^2\boldsymbol{P}_0)\boldsymbol{1}_n$ of the average

[9] It is not strictly required that $\{Y_i\}_{i=1}^n$ are i.i.d. For instance, if the errors are heteroscedastic or are dependent over individuals and/or time, one can simply replace (4.37) with the formula for an appropriate choice of robust standard errors, or alternatively, specify a dynamic or spatial model, or model the error dependence directly, for instance with a model based on generalized estimating equations (M-estimator).

of the predicted counterfactual outcomes. We thereby have that

$$\widehat{\tau}_{AGTET}(\boldsymbol{Z}, \boldsymbol{z}_0) = \mathbf{1}_n'(\widehat{\boldsymbol{y}}(\boldsymbol{Z}) - \widehat{\boldsymbol{y}}(\boldsymbol{z}_0)) \sim \mathcal{N}(\tau_{AGTET}(\boldsymbol{Z}, \boldsymbol{z}_0), \mathbf{1}_n'(\sigma^2 \boldsymbol{P}_0)\mathbf{1}_n) \tag{4.38}$$

It follows that $(\widehat{\tau}_{AGTET}(\boldsymbol{Z}, \boldsymbol{z}_0) - \tau_{AGTET}(\boldsymbol{Z}, \boldsymbol{z}_0))/\sqrt{\mathbf{1}_n'(\widehat{\sigma}^2 \boldsymbol{P}_0)\mathbf{1}_n} \sim t_{n-(p+k)}$, so we can obtain a confidence interval for the AGTET based on the t-statistic. A two-sided $(1-\alpha) \times 100$ percent confidence interval estimate of the AGTET is given by

$$\mathcal{I}_{1-\alpha}^{\tau_{AGTET}(\boldsymbol{Z},\boldsymbol{z}_0)} := \widehat{\tau}_{AGTET}(\boldsymbol{Z}, \boldsymbol{z}_0) \pm t_{n-(p+k)}(\alpha/2)\sqrt{\mathbf{1}_n'(\widehat{\sigma}^2 \boldsymbol{P}_0)\mathbf{1}_n}. \tag{4.39}$$

A confidence interval is appropriate if we estimate the AGTET as the difference in the averages of the predictions of the observed outcomes and their predicted counterfactuals. A prediction interval is suitable if we wish to estimate the AGTET as the difference between the observed potential outcomes and counterfactuals that are predicted, as with the estimator shown in expression (4.24). Now, the value of an observed potential outcome is given by both the value of f and the value of the error. Hence the variance of the counterfactual outcomes' error terms must be included in the prediction interval of the counterfactual potential outcomes. Since the variance is always positive, this implies that prediction intervals are wider than confidence intervals. Assuming that $\{y_i\}_{i=1}^n$ are i.i.d., we have that $\mathbb{V}[\boldsymbol{\varepsilon}] = \sigma^2 \boldsymbol{I}$. Using result (4.37), we get that

$$\mathbb{V}[\boldsymbol{y}(\boldsymbol{z}_0)] = \mathbb{V}[\boldsymbol{U}_0\widehat{\boldsymbol{\theta}} + \boldsymbol{\varepsilon}] = \sigma^2 \boldsymbol{P}_0 + \sigma^2 \boldsymbol{I} = \sigma^2(\boldsymbol{P}_0 + \boldsymbol{I}). \tag{4.40}$$

Thus for inference about the AGTET of the realized outcomes from the k-dimensional treatment regime $\boldsymbol{Z}$ in relation to some counterfactual levels $\boldsymbol{z}_0$ with the estimator shown in (4.24), simply use $\mathbf{1}_n'(\boldsymbol{y} - f^{(adj)}(\boldsymbol{z}_0, \boldsymbol{X}))$ to obtain the point estimate and plug in the square root of $\mathbf{1}_n'(\widehat{\sigma}^2(\boldsymbol{P}_0 + \boldsymbol{I}))\mathbf{1}_n$ as the estimated standard error into the expression of the interval to get a prediction interval.

Note that $\widehat{\mathbb{V}}[(\widehat{y}_i(\boldsymbol{z}_{0,i})] = \widehat{\sigma}^2 \boldsymbol{P}_{0,i}$ from expression (4.37) is equal to what $\widehat{\sigma}_i^2$ from Proposition 4.11 would be for $\widehat{y}_i(\boldsymbol{z}_{0,i})$. Thereby, given that same DRF is used, the non-parametric procedure outlined in Proposition 4.11 gives the same estimate of the standard error, and thus the same interval estimate, as the parametric approach explained in this example.

QR-decomposition based estimation. We can generalize the interval estimation procedure for an arbitrary parametric regression model. Make a QR-decomposition of the observed design matrix $\boldsymbol{U}$ such that

$$\boldsymbol{U} = \boldsymbol{Q}\boldsymbol{R} = \begin{bmatrix}\boldsymbol{Q}_1 & \boldsymbol{Q}_2\end{bmatrix}\begin{bmatrix}\boldsymbol{R}_1 \\ \mathbf{0}\end{bmatrix} = \boldsymbol{Q}_1\boldsymbol{R}_1 \tag{4.41}$$

where $\boldsymbol{R}_1$ is $(p+k)\times(p+k)$ upper triangular, $\boldsymbol{Q}_1$ is the first $p+k$ columns of the orthogonal $n\times n$ matrix $\boldsymbol{Q}$ and $\boldsymbol{Q}_2$ are the last $n-(p+k)$ columns of $\boldsymbol{Q}$. If $\operatorname{rank}(\boldsymbol{U}) = r < p+k$ then the first r columns of $\boldsymbol{Q}$ constitute an orthogonal basis for the column span of $\boldsymbol{U}\boldsymbol{P}$ where $\boldsymbol{P}$ is the $(p+k)\times(p+k)$ permutation matrix. Now, $\boldsymbol{U}'\boldsymbol{U} = (\boldsymbol{Q}_1\boldsymbol{R}_1)'(\boldsymbol{Q}_1\boldsymbol{R}_1) = \boldsymbol{R}_1'\boldsymbol{Q}_1'\boldsymbol{Q}_1\boldsymbol{R}_1 = \boldsymbol{R}_1'\boldsymbol{R}_1$. Hence, an estimate of the variance-covariance matrix of the counterfactuals $\boldsymbol{y}(\boldsymbol{z}_0)$ can be obtained with

$$\widehat{\sigma}^2(\boldsymbol{U}_0\boldsymbol{P})(\boldsymbol{R}_1'\boldsymbol{R}_1)^{-1}(\boldsymbol{U}_0\boldsymbol{P})' = \widehat{\sigma}^2(\boldsymbol{U}_0\boldsymbol{P})\boldsymbol{R}_1^{-1}\boldsymbol{R}_1'^{-1}(\boldsymbol{U}_0\boldsymbol{P})' \coloneqq \widehat{\sigma}^2(\boldsymbol{B}'\boldsymbol{B}) \tag{4.42}$$

where $\boldsymbol{B} = \boldsymbol{R}_1'^{-1}(\boldsymbol{U}_0\boldsymbol{P})'$. The prediction variance for the average prediction $\mathbf{1}_n'\widehat{\boldsymbol{y}}(\boldsymbol{z}_0)$ is given by $\mathbf{1}_n'(\widehat{\sigma}^2(\boldsymbol{B}'\boldsymbol{B}))\mathbf{1}_n$ whereas for the prediction interval it is $\mathbf{1}_n'(\widehat{\sigma}^2(\boldsymbol{B}'\boldsymbol{B}+\boldsymbol{D}))\mathbf{1}_n$. Here, $\boldsymbol{D}$ is the matrix that accounts for the added error variance in prediction intervals. If the errors are homoscedastic, then $\boldsymbol{D} = \boldsymbol{I}$.

It is possible to speed up the estimation of the prediction variance by replacing the matrix multiplications with vector multiplications. Define the vector $\boldsymbol{w} \coloneqq \boldsymbol{B}\mathbf{1}_n = \boldsymbol{R}_1'^{-1}(\boldsymbol{U}_0\boldsymbol{P})'\mathbf{1}_n = \boldsymbol{R}_1'^{-1}\boldsymbol{P}'(\boldsymbol{U}_0'\mathbf{1}_n)'$. The prediction variance for the confidence interval of the AGTET can then be estimated with

$$\mathbf{1}_n'(\widehat{\sigma}^2(\boldsymbol{B}'\boldsymbol{B}))\mathbf{1}_n = \widehat{\sigma}^2(\boldsymbol{B}\mathbf{1}_n)'(\boldsymbol{B}\mathbf{1}_n) = \widehat{\sigma}^2(\boldsymbol{w}'\boldsymbol{w}). \tag{4.43}$$

Similarly, the prediction variance for the prediction interval can in the simplifying case that $\boldsymbol{D} = \boldsymbol{I}$ be estimated with

$$\begin{aligned}\mathbf{1}_n'(\widehat{\sigma}^2(\boldsymbol{B}'\boldsymbol{B}+\boldsymbol{I}))\mathbf{1}_n &= \mathbf{1}_n'(\widehat{\sigma}^2(\boldsymbol{B}'\boldsymbol{B}))\mathbf{1}_n + \widehat{\sigma}^2(\mathbf{1}_n'\mathbf{1}_n)\\ &= \widehat{\sigma}^2(\boldsymbol{w}'\boldsymbol{w}) + \widehat{\sigma}^2(\mathbf{1}_n'\mathbf{1}_n)\end{aligned} \tag{4.44}$$

Here, the matrix multiplication $\boldsymbol{B}'\boldsymbol{B}$ have been replaced with the more efficient vector multiplication $\boldsymbol{w}'\boldsymbol{w}$. Hence in estimating the prediction variance there is no need for memory storage of the matrix $\boldsymbol{B}$, but only of the vectors $\boldsymbol{w}$ and $\mathbf{1}_n$.

Taking the square root of the prediction variance estimated with any of the methods explained in this section yields the estimated standard error of the average of the predicted potential outcomes. It can be used as a plug-in estimator for the unknown population standard error in an uncertainty interval of the ADRF or AGTET, and the AGTE if correction is made for that both averages are uncertain.

Hypothesis testing. A natural question is whether the average treatment effect in the population is different from zero. Formally, we then want to test the hypothesis $\mathcal{H}_0 : \tau_{AGTE}(\boldsymbol{z}, \boldsymbol{z}_0) = 0$ against $\mathcal{H}_A : \tau_{AGTE}(\boldsymbol{z}, \boldsymbol{z}_0) \neq 0$.

Proposition 4.12. *By* (4.31)*, we have that under* $\mathcal{H}_0$*,*

$$\frac{\mathbf{1}_n'(\widehat{\boldsymbol{y}}(\boldsymbol{z}) - \widehat{\boldsymbol{y}}(\boldsymbol{z}_0))}{\mathbb{S}[\widehat{\tau}_{AGTE}(\boldsymbol{z}, \boldsymbol{z}_0)]} = \frac{\widehat{\tau}_{AGTE}(\boldsymbol{z}, \boldsymbol{z}_0)}{\mathbb{S}[\widehat{\tau}_{AGTE}(\boldsymbol{z}, \boldsymbol{z}_0)]} \sim t_{n-(p+k)}. \tag{4.45}$$

If the test statistic $|t| = |\widehat{\tau}_{AGTE}(\boldsymbol{z}, \boldsymbol{z}_0)/\mathbb{S}[\widehat{\tau}_{AGTE}(\boldsymbol{z}, \boldsymbol{z}_0)]| > t_{n-(p+k)}(\alpha/2)$*, then* $\mathcal{H}_0$ *is rejected at the level of significance* $1 - \alpha$*.*

Remark. We may wish test the hypothesis for the ADRF that $\mathcal{H}_0 : \mu(z) = 0$ against $\mathcal{H}_A : \mu(z) \neq 0$. Since the ADRF evaluated at z is also a mean,

$$\frac{\widehat{\mu}_f(z)}{\mathbb{S}[\widehat{\mu}_f(z)]} \sim t_{n-(p+k)}, \tag{4.46}$$

and $\mathcal{H}_0$ is rejected if $|\widehat{\mu}_f(z)/\mathbb{S}[\widehat{\mu}_f(z)]| > t_{n-(p+k)}(\alpha/2)$.

Here, $\mathbb{S}[\widehat{\tau}_{AGTE}(\boldsymbol{z}, \boldsymbol{z}_0)]$ and $\mathbb{S}[\widehat{\mu}_f(z)]$ are the standard error of the estimated AGTE and ADRF, estimated with any of the aforementioned methods.

We may have evidence or theory indicating that the AGTE should be different from zero and close to some known population value $\tau^0_{AGTE}(\boldsymbol{z}, \boldsymbol{z}_0)$. We then reject the null hypothesis $\mathcal{H}_0 : \tau_{AGTE}(\boldsymbol{z}, \boldsymbol{z}_0) = \tau^0_{AGTE}(\boldsymbol{z}, \boldsymbol{z}_0)$ if

$$\left| \frac{\widehat{\tau}_{AGTE}(\boldsymbol{z}, \boldsymbol{z}_0) - \tau^0_{AGTE}(\boldsymbol{z}, \boldsymbol{z}_0)}{\mathbb{S}[\widehat{\tau}_{AGTE}(\boldsymbol{z}, \boldsymbol{z}_0)]} \right| > t_{n-(p+k)}(\alpha/2). \tag{4.47}$$

These tests will be valid no matter if the AGTE is estimated for a single treatment or multiple treatments, since the AGTE is a scalar mean in either way. The difference between an AGTE for a single and multiple treatments is solely in the estimation of the counterfactual potential outcomes, which is shown in Algorithm 1 and 2.

4.6 Optimization

Summary statistics such as the ATE or AGTE are important for understanding the impact of treatments on outcomes. However, even with perfect knowledge of the treatment effect, the normative problem of causal inference remains; how to optimally assign treatments. In this section we build on the theory of treatment choice by Manski (2000, 2002, 2004), Qian and Murphy (2011) and Armstrong and Shen (2015) for binary treatments and extend it for continuous treatments. We show that the methods presented in the previous sections can be used to estimate the treatment level that map to the maximum expected outcome.

The treatment assignment problem. A *planner* has the goal of assigning a treatment to each individual of a heterogenous population such that the outcome is maximized. A planner can for instance be a government official, judge, doctor or advertising manager. The planner can observe the characteristics of each individual in the population but not how they will respond to the treatment. The planner does however know that the treatment effect vary between individuals. To her aid, the planner has access to a previous study on a sample from the same population where the treatment assignment and treatment response for each individual is known.

We define the problem in more formal terms. Assume without loss of generality that larger values of the outcome are preferred. The planner wants to assign a dose z of Z from its support $\mathcal{Z}$ to individual $i = 1, \dots, n$ of $\mathcal{P}$ such that $\mathbb{E}[Y(Z)]$ is maximized. The planner observes the covariates $\boldsymbol{X}$ that contain information of each individual. If the planner knew the conditional distributions $P_{Y(Z)|\boldsymbol{X}}$ for all $\boldsymbol{x} \in \mathcal{X}$, the planner would simply choose z such that the outcome given the covariates is maximized. There are two problems facing the planner. By the fundamental problem of causal inference, $P_{Y(Z)|\boldsymbol{X}=\boldsymbol{x}}$ cannot be observed in the data from the previous study. That is, even though the planner can observe $\boldsymbol{x}_i$ for all individuals in the new data, the response $y(\cdot) = y(i, \cdot, \boldsymbol{x}_i)$ is unknown for all levels in $\mathcal{Z}$ that was not assigned to each individual in the study.

To achieve the goal, the planner wants to assign z_i given $\boldsymbol{x}_i$. Define the *individual treatment rule* (Qian & Murphy, 2011) given by $r(\cdot) : \mathcal{X} \to \mathcal{Z}$. The individual treatment rule $r \in \mathcal{R} \subset \mathbb{R}$ is a deterministic decision rule that maps individual-specific information $\boldsymbol{x}_i$ to a real-valued treatment level z_i. The potential outcome for individual i under rule $r(\boldsymbol{x}_i)$ is $y_i[r(\boldsymbol{x}_i)]$. This shows that the treatment an individual receives only depends on her individual-specific covariates and that the individual's potential outcomes only depends on the individual's own treatments. Thereby the no interference assumption of SUTVA required for causal inference holds. We now present the decision rule that assigns treatments optimally.

Definition 4.8 (OPTIMAL INDIVIDUAL TREATMENT RULE)**.** The *optimal individual treatment rule* is given by

$$r^{(opt)}(\boldsymbol{x}_i) := \underset{r(\cdot)\in\mathcal{R}}{\arg\max}\, \mathbb{E}\{y_i[r(\boldsymbol{x}_i)]\}. \tag{4.48}$$

The outcome $y_i[r^{(opt)}(\boldsymbol{x}_i)]$ is the *optimal individual treatment outcome.*

The definition states that an individual treatment rule is optimal if it assigns an individual the treatment that maximizes the expected value of the individual's potential outcome. The solution to this problem is obtained by assigning each individual the treatment level that conditional on the individual's covariates maximizes the expected response to the treatment. Using the law of iterated expectations, Manski (2002) show that for each $r(\cdot) \in \mathcal{R}$, we have that

$$\begin{aligned}\mathbb{E}\{y_i[r(\boldsymbol{x}_i)]\} &= \mathbb{E}\{y_i[r(\boldsymbol{x}_i)]|\boldsymbol{X} = \boldsymbol{x}_i\}\\ &= \mathbb{E}\Big\{\sum_{z_i\in\mathcal{Z}} \mathbb{E}[y_i(z_i)|\boldsymbol{X} = \boldsymbol{x}_i]\mathbb{1}[r(\boldsymbol{x}_i) = z_i]\Big\}\\ &= \int_{\mathcal{X}} \sum_{z_i\in\mathcal{Z}} \mathbb{E}[y_i(z_i)|\boldsymbol{X} = \boldsymbol{x}_i]\mathbb{1}[r(\boldsymbol{x}_i) = z]dP_{\boldsymbol{X}}(\boldsymbol{x}_i). \end{aligned}\tag{4.49}$$

That is, for each $\boldsymbol{x}_i \in \mathcal{X}$, the integrand $\sum_{z_i\in\mathcal{Z}} \mathbb{E}[y_i(z_i)|\boldsymbol{X} = \boldsymbol{x}_i]\mathbb{1}[r(\boldsymbol{x}_i) = z_i]$ is maximized by choosing the assignment $r(\boldsymbol{x}_i)$ that maximizes $\mathbb{E}[y_i(z_i)|\boldsymbol{X} = \boldsymbol{x}_i]$ on $z_i \in \mathcal{Z}$. Hence, the optimal treatment rule $r^{(opt)}(\boldsymbol{x}_i)$ is the rule that for all $\boldsymbol{x}_i \in \mathcal{X}$ solves

$$\underset{z_i\in\mathcal{Z}}{\arg\max}\, \mathbb{E}[y_i(z_i)|\boldsymbol{X} = \boldsymbol{x}_i]. \tag{4.50}$$

The mean optimal treatment level $z^{(opt)}$ is obtained by averaging the solutions to the individual-level optima over $\mathcal{X}$,

$$\begin{aligned}z^{(opt)} = \mathbb{E}[\underset{z\in\mathcal{Z}}{\arg\max}\, \mathbb{E}\{Y(z)|\boldsymbol{X}\}] &= \int_{\mathcal{X}} \underset{z\in\mathcal{Z}}{\arg\max}\, \mathbb{E}[Y(z)|\boldsymbol{X} = \boldsymbol{x}]dP_{Z|\boldsymbol{X}}(z|\boldsymbol{x})\\ &= \underset{z\in\mathcal{Z}}{\arg\max}\, \mathbb{E}[Y(z)].\end{aligned}\tag{4.51}$$

The counterfactual prediction solution. Since $r(\boldsymbol{x}) \mapsto z$, expression (4.51) implies that finding $r^{(opt)}(\cdot) \in \mathcal{R}$ that maximizes $\mathbb{E}\{Y[r(\boldsymbol{X})]$ is equivalent to finding $z \in \mathcal{Z}$ that maximizes $\mathbb{E}[Y(z)]$. Hence, knowledge of the individual treatment rule is not required for obtaining an optimal outcome. Using this result, the counterfactual prediction method enables estimation of optimal outcomes. The key is to note that

$$\mathbb{E}[y_i(z_i)|\boldsymbol{X} = \boldsymbol{x}_i] = \mathbb{E}[f(z_i, \boldsymbol{x}_i) + \varepsilon_i] = f(z_i, \boldsymbol{x}_i).$$

and

$$\mathbb{E}[Y(z)] = \mathbb{E}\{\mathbb{E}[Y(z)|\boldsymbol{X}]\} = \mathbb{E}[f(z, \boldsymbol{X})] = \mu(z).$$

Thereby, the expected optimal individual treatment outcome is

$$\mathbb{E}[Y_i^{(opt)}] = \max_{z_i \in \mathcal{Z}} f(z_i, \boldsymbol{x}_i) \tag{4.52}$$

and the optimal population mean outcome is

$$\mathbb{E}[Y^{(opt)}] = \max_{z \in \mathcal{Z}} \mu(z). \tag{4.53}$$

To estimate the optimal outcome, use data splitting on the previous study's data and choose the DRF trained on the test data that minimizes the f-risk on the validation data. Then with Algorithm 1, evaluate the DRF for different values of Z and the values of $\boldsymbol{X}$ in the new data. Finally, evaluate the ADRF with the estimator in Proposition 4.5. The maximum of the curve given by the ADRF is the estimated optimized population mean outcome $\widehat{\mu}_f^{(opt)}$. That is,

$$\widehat{\mu}_f^{(opt)} \coloneqq \max_{z \in \mathcal{Z}} \frac{1}{n} \sum_{i=1}^{n} f(z, \boldsymbol{x}_i) = \max_{z \in \mathcal{Z}} \widehat{\mu}_f(z). \tag{4.54}$$

The estimated optimal treatment level $\widehat{z}^{(opt)}$ is the level $z \in \mathcal{Z}$ that maps to $\widehat{\mu}_f^{(opt)}$. To evaluate the gain of $\widehat{\mu}_f^{(opt)}$ relative a suboptimal outcome, the planner may compute

$$\widehat{\tau}_{AGTE}(\widehat{z}^{(opt)}, z_0) = \widehat{\mu}_f(\widehat{z}^{(opt)}) - \widehat{\mu}_f(z_0). \tag{4.55}$$

where $z_0 \neq z^{(opt)}$. Methods from Section 4.5 can be used for inference of the estimated maximum of the ADRF, the optimal treatment outcome.

5 Discussion

In this thesis we have developed counterfactual prediction methods for estimation, inference and optimization of causal effects from observational data with continuous treatments. Our approach build on the regression adjustment method for binary treatments and matching and propensity score methods for continuous treatments (Fong et al., 2018; Kennedy et al., 2017; Hirano & Imbens, 2004; Zhang et al., 2016; Zhu et al., 2014) of which some are developed for experimental data. Several researchers (Armstrong & Kolesár, 2018; Hill, 2011; Zhang et al., 2016; Zhu et al., 2014) also propose estimating potential outcomes of a continuous treatment with a dose-dependent function, but do not consider using the function to estimate a generalized treatment effect or the methods we present for inference and optimization.

Summary. The counterfactual prediction method involves using a regression model fitted to the observed response surface to predict the counterfactual potential outcomes. No model of the treatment assignment mechanism is required. It can be viewed as regression adjustment generalized to the continuous treatment setting. Under strong ignorability, SUTVA, and that the DRF is a consistent, the causal effect of the treatment is identifiable and estimable. Given that the DRF is estimable, the method works under arbitrarily high-dimensional confounding. Several papers (see for instance Hill (2011); Wager and Athey (2018); Powers et al. (2018)) show that data-driven regression algorithms often outperform estimators that require manual specification, even when the true regression is a simple linear function. Thereby the counterfactual prediction method require minimal assumptions about the treatment assignment mechanism and the true functional form.

We consider interval estimation and hypothesis testing for inference of the treatment effect. Three methods are presented for estimating the standard error of the estimand of interest. The first method is to sum all elements of the covariance matrix of the predicted counterfactual potential outcomes and divide by the squared number of observations. This method assumes no form of the model or distribution of the outcome. Since the DRF can be non-parametric, this implies that a point and interval estimate of the estimand is non-parametrically estimable. The second method use that for a parametric DRF, the prediction variance is a function of the DRF's error variance and projection and design matrices. This method is analogous to standard analytical methods for inference of a prediction. The third method use QR-decomposition of the design matrix and basic properties of vectors and matrices to obtain an alternative expression of the prediction variance. This method works for any parametric DRF and is computationally efficient.

Finally, we showed that the counterfactual prediction method estimates the treatment level that maximizes the expected individual and population mean outcome. The existing literature has considered estimation of the decision rule that maps covariates into a treatment assignment. The solution with the counterfactual prediction method is simpler; just estimate DRF and the ADRF on data from a previous study and look at the treatment level that maps to the maximum of the corresponding function. Hence, the counterfactual prediction method is not only useful for estimation and inference of summary statistics such as the average treatment effect but also for the normative question of how to assign treatments.

Comparison to standard methods. Both the existing standard regression adjustment and the counterfactual prediction method are based on the idea of predicting the counterfactual potential outcomes. The main difference is that standard regression adjustment involves fitting one regression for each treatment status whereas our method only involves fitting one regression to all observations in the sample. Standard regression adjustment does not work with continuous treatments since that would imply the need for as many response functions as realized treatment levels. We thereby only compare the two methods for the binary treatment case. The following paragraph explains why the counterfactual prediction method still work well in this case.

If the common support of the covariate distributions among the treated and non-treated is small or nonexistent, then counterfactual prediction better adjust for the effect of the confounders than standard regression adjustment. Since the potential outcomes are predicted from a single model fitted to the full sample of both treated and non-treated observations, the model learns the effect of the confounders over their full range of realized values. The subsample-specific regressions of standard regression adjustment only learns the effect of the confounders on its subsample. Hence if the treated and non-treated observations are not similar on the covariates, meaning that there is a lack of common support, then the counterfactual predictions of the subsample specific response functions cannot identify the true counterfactuals for the individuals of opposite treatment status. As an illustrative example, if it happens that only men are treated and women are controls, then a response function trained on men shall predict the counterfactual treated outcomes for women, and a response function trained on women shall predict the counterfactual control outcomes for men. The counterfactual prediction method would in this example instead fit the same regression function to both men and females and then predict the respective counterfactuals for each gender with this function. Still, the counterfactual prediction method is not fully robust against a lack of common support. The predicted counterfactual potential outcomes will have larger prediction errors the larger the lack of overlap in the distributions of the confounders. This will increase the uncertainty in the estimate of the causal effect.

The counterfactual prediction estimator makes a stronger assumption about functional form. In standard regression adjustment, only the relationship between the confounders and the outcome needs to be correctly specified since the treatment status is not used as a predictor; the regression is estimated on the subsample restricted to the treatment level. In the counterfactual prediction method the model is fitted to the outcomes using both the treatment and the confounders as a predictors.

A benefit of the counterfactual prediction method is that it handles missing data well. Simply fit the model to the complete cases and make predictions for the missing observations. No matter how the pattern of the missing data look or whether the number of missing observations vary over treatment levels, the result will be complete series of estimated potential outcomes for all observations. For the standard regression adjustment, missing data poses a more difficult problem. Since standard regression adjustment involves fitting one regression per treatment level, the uncertainty in the predicted outcomes for the missing observations depend on the number of complete cases at each treatment level.

Future research. The counterfactual prediction method is easily extended to other popular causal estimands, for instance the local average treatment effect. Instrumental variables can be used if the treatment is endogenous. A promising research direction would be to apply the framework of functional data analysis. This would solve problems that arise if the data are high-dimensional. It also naturally fits into the potential outcomes framework since the causal estimands are in fact functions and the potential outcome $Y(Z) = Y(I, Z, \boldsymbol{X})$ may be viewed as a function. Then every observation is a function, which is precisely the idea of functional data analysis.

References

Altshuler, B. (1981). Modeling of Dose-Response Relationships. *Environmental Health Perspectives*, *42*(December), 185–195. doi: 10.1289/ehp.814223

Armstrong, T., & Kolesár, M. (2018). Finite-Sample Optimal Estimation and Inference on Average Treatment Effects Under Unconfoundedness. *arXiv preprint arXiv:1712.04594*. Retrieved from `https://arxiv.org/abs/1712.04594`

Armstrong, T., & Shen, S. (2015). Inference on Optimal Treatment Assignments (April 9, 2015). *Cowles Foundation Discussion Paper No. 1927RR*, 46. doi: 10.2139/ssrn.2592479

Athey, S., & Imbens, G. (2016). Recursive Partitioning for Heterogeneous Causal Effects. *Proceedings of the National Academy of Sciences*, *113*(27), 7353–7360. doi: 10.1073/pnas.1510489113

Athey, S., Tibshirani, J., & Wager, S. (2019). Generalized Random Forests. *Annals of Statistics*, *47*(2), 1179–1203. doi: 10.1214/18-AOS1709

Chernozhukov, V., Demirer, M., Duflo, E., & Fernandez-Val, I. (2017). Generic Machine Learning Inference on Heterogenous Treatment Effects in Randomized Experiments. *arXiv preprint arXiv:1712.04802*. Retrieved from `http://arxiv.org/abs/1712.04802`

Chernozhukov, V., Fernandez-Val, I., & Luo, Y. (2018). The Sorted Effects Method: Discovering Heterogeneous Effects Beyond Their Averages. *Econometrica*, *86*(6), 1911–1938. doi: 10.3982/ECTA14415

D'Amour, A. (2019). On Multi-Cause Causal Inference with Unobserved Confounding: Counterexamples, Impossibility, and Alternatives. *arXiv preprint arXiv:1902.10286*. Retrieved from `http://arxiv.org/abs/1902.10286`

Diamond, A., & Sekhon, J. S. (2013). Genetic Matching for Estimating Causal Effects: Balance in Observational Studies. *The Review of Economics and Statistics*, *95*(July), 932–945. doi: 10.1162/REST_a_00318

Fisher, R. A. (1935). *The Design of Experiments* (1st ed.). Oxford: Oliver & Boyd, England.

Fong, C., Hazlett, C., & Imai, K. (2018). Covariate Balancing Propensity Score for a Continuous Treatment: Application to the Efficacy of Political Advertisements. *Annals of Applied Statistics*, *12*(1), 156–177. doi: 10.1214/17-AOAS1101

Grimmer, J., Westwood, S. J., & Messing, S. (2017). Estimating Heterogeneous Treatment Effects and the Effects of Heterogeneous Treatments with Ensemble Methods. *Political Analysis*, *25*(4), 413–434. doi: 10.1017/pan.2017.15.

Häggström, J. (2018). Data-driven Confounder Selection via Markov and Bayesian Networks. *Biometrics*, *74*(2), 389–398. doi: 10.1111/biom.12788

Hartford, J., Lewis, G., Leyton-Brown, K., & Taddy, M. (2017). Deep IV: A Flexible Approach for Counterfactual Prediction. In *Proceedings of machine learning research* (Vol. 70, pp. 1414–1423). Retrieved from `http://proceedings.mlr.press/v70/hartford17a.html`

Heckman, J. J., Ichimura, H., & Todd, P. (1997). Matching As An Econometric Evaluation Estimator. *Review of Economic Studies*, *65*(4), 605–654. Retrieved from `http://www.jstor.org/stable/2971733` doi: 10.2307/2971733

Hernán, M., Brumback, B., & Robins, J. (2000). Marginal Structural Models and Causal Inference in Epidemiology. *Epidemiology*, *11*(5), 550–560.

Hernán, M., & Robins, J. (2019). *Causal Inference* (1st ed.). Boca Raton: Chapman & Hall/CRC, forthcoming.

Hill, J. L. (2011). Bayesian Nonparametric Modeling for Causal Inference. *Journal of Computational and Graphical Statistics*, *20*(1), 217–240. doi: 10.1198/jcgs.2010.08162

Hirano, K., & Imbens, G. W. (2004). The Propensity Score with Continuous Treatments. In *Applied bayesian modeling and causal inference from incomplete-data perspectives: An essential journey with donald rubin's statistical family* (Vol. 0226164, pp. 73–84). doi: 10.1002/0470090456.ch7

Hirano, K., Imbens, G. W., & Ridder, G. (2003). Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score. *Econometrica*, *71*(4), 1161–1189.

Hirano, K., & Porter, J. R. (2009). Asymptotics for Statistical Treatment Rules. *Econometrica*, *77*(5), 1683–1701. Retrieved from `http://www.jstor.org.ludwig.lub.lu.se/stable/25621374` doi: 10.3982/ECTA6630

Hitsch, G. J., & Misra, S. (2018). Heterogeneous Treatment Effects and Optimal Targeting Policy Evaluation. *Available at SSRN*, 1–64. Retrieved from `https://ssrn.com/abstract=3111957` doi: 10.2139/ssrn.3111957

Holland, P. (1986). Statistics and Causal Inference. *Journal of the American Statistical Association*, *81*(396), 945–960. doi: 10.2307:2289064

Imai, K., & Ratkovic, M. (2013). Estimating Treatment Effect Heterogeneity in Randomized Program Evaluation. *Annals of Applied Statistics*, *7*(1), 443–470. doi: 10.1214/12-AOAS593

Imbens, G. W., & Rubin, D. B. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences* (1st ed.). Cambridge: Cambridge University Press.

Imbens, G. W., & Wooldridge, J. M. (2009). Recent Developments in the Econometrics of Program Evaluation. *Journal of Economic Literature*, *47*(1), 5–86. doi: 10.1257/jel.47.1.5

Kennedy, E. H., Lorch, S. A., & Small, D. S. (2016). Robust Causal Inference with Continuous Instruments Using the Local Instrumental Variable Curve. *arXiv preprint arXiv:1607.02566*. Retrieved from `http://arxiv.org/abs/1607.02566`

Kennedy, E. H., Ma, Z., Mchugh, M. D., & Small, D. S. (2017). Nonparametric Methods for Doubly Robust Estimation of Continuous Treatment Effects. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *79*(4), 1229–1245. doi: 10.1111/rssb.12212

Künzel, S. R., Sekhon, J. S., Bickel, P. J., & Yu, B. (2017). Meta-learners for Estimating Heterogeneous Treatment Effects using Machine Learning. *arXiv preprint arXiv:1706.03461*.

Luna, X. D., Waernbaum, I., & Richardson, T. S. (2011). Covariate Selection for the Nonparametric Estimation of an Average Treatment Effect. *Biometrika*, *98*(4), 861–875. doi: 10.1093/biomet/asr041

Manski, C. F. (2000). Identification Problems and Decisions under Ambiguity: Empirical Analysis of Treatment Response and Normative Analysis of Treatment Choice. *Journal of Econometrics*, *95*(2), 415–442. doi: 10.1016/S0304-4076(99)00045-7

Manski, C. F. (2002). Treatment Choice under Ambiguity Induced by Inferential Problems. *Journal of Statistical Planning and Inference*, *105*(1), 67–82. doi: 10.1016/S0378-3758(01)00204-X

Manski, C. F. (2004). Statistical Treatment Rules for Heterogeneous Populations. *Econometrica*, *72*(4), 1221–1246. Retrieved from `http://www.jstor.org/stable/3598783`

Newey, W. K. (1994). Kernel Estimation of Partial Means and a General Variance Estimator. *Econometric Theory*, *10*(2), 233–253. Retrieved from `https://www.jstor.org/stable/3532868`

Nie, X., & Wager, S. (2017). Quasi-Oracle Estimation of Heterogeneous Treatment Effects. *arXiv preprint arXiv:1712.04912*. Retrieved from `http://arxiv.org/abs/1712.04912`

Oprescu, M., Syrgkanis, V., & Wu, Z. S. (2018). Orthogonal Random Forest for Causal Inference. *arXiv preprint arXiv:1806.03467*. Retrieved from `https://arxiv.org/abs/1806.03467`

Persson, E., Häggström, J., Waernbaum, I., & de Luna, X. (2017). Data-driven Algorithms for Dimension Reduction in Causal Inference. *Computational Statistics and Data Analysis*, *105*, 280–292. doi: 10.1016/j.csda.2016.08.012

Powers, S., Qian, J., Jung, K., Schuler, A., Shah, N. H., Hastie, T., & Tibshirani, R. (2018). Some Methods for Heterogeneous Treatment Effect Estimation in High Dimensions. *Statistics in Medicine*, *37*(11), 1767–1787. doi: 10.1002/sim.7623

Qian, M., & Murphy, S. A. (2011). Performance Guarantees for Individualized Treatment Rules. *The Annals of Statistics*, *39*(2), 1180–1210. doi: 10.1214/10-aos864

Rosenbaum, P. R. (2010). *Design of Observational Studies* (Vol. 27) (No. 2). New York: Springer. doi: 10.1007/978-1-4419-1213-8

Rosenbaum, P. R., & Rubin, D. B. (1983). The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika*, *70*(1), 41–55. doi: 10.1093/biomet/70.1.41

Rubin, D. B. (1973). Matching to Remove Bias in Observational Studies. *Biometrics*, *29*(1), 62–80. doi: 10.2307/2529684

Rubin, D. B. (1974). Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. *Journal Of Educational Psychology*, *66*(5), 688–701. doi: 10.1037/h0037350

Schnitzer, M. E., Lok, J. J., & Gruber, S. (2016). Variable Selection for Confounder Control, Flexible Modeling and Collaborative Targeted Minimum Loss-based Estimation in Causal Inference. *International Journal of Biostatistics*, *12*(1), 97–115. doi: 10.1515/ijb-2015-0017

Schuler, A., Baiocchi, M., Tibshirani, R., & Shah, N. (2018). A Comparison of Methods for Model Selection when Estimating Individual Treatment Effects. *arXiv preprint arXiv:1804.05146*. Retrieved from `http://arxiv.org/abs/1804.05146`

Wager, S., & Athey, S. (2018). Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. *Journal of the American Statistical Association*, *113*(523), 1228–1242. doi: 10.1080/01621459.2017.1319839

Wald, A. (1949). Statistical Decision Functions. *The Annals of Mathematical Statistics*, *20*(2), 165–205.

Wang, J. (2015). *Exposure-Response Modeling: Methods and Practical Implementation*. Boca Raton: Chapman & Hall/CRC. doi: 10.1201/b18717

Zhang, Z., Zhou, J., Cao, W., & Zhang, J. (2016). Causal Inference with a Quantitative Exposure. *Statistical Methods in Medical Research*, *25*(1), 315–335. doi: 10.1177/0962280212452333

Zhu, Y., Coffman, D., & Ghosh, D. (2014). A Boosting Algorithm for Estimating Generalized Propensity Scores with Continuous Treatments. *Journal of Causal Inference*, *3*(1), 25–40. doi: 10.1515/jci-2014-0022