ETH
Swiss Federal Institute of Technology Zurich

Seminar for Statistics

Department of Mathematics

Master Thesis Summer 2022

Alexander Gudjónsson

Nonparametric Causal Mediation Analysis with Distributional Random Forests and Generalised Propensity Scores

Submission Date: 2. August 2022

Co-Adviser Michael J. Zellinger
Adviser: Prof. Dr. Peter Bühlmann

Abstract

Causal mediation analysis is the area of statistics that attempts to carve out the details of statistical causal relationships between variables by asking whether the causal effect of a treatment on an outcome is mediated through another variable. The area is well established for parametric models. Here we extend its scope by presenting a new nonparametric method with wide generality. We review the theoretical foundations, propose the application of a recently developed method called Distributional Random Forest and discuss an enhancement, applying the concept of generalised propensity scores. We detail the methodological aspects and provide extensive simulations to demonstrate the performance and generality of our method.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The questions that drive scientific knowledge are the ones that ask *how* or *why* the phenomenons we observe in the world happen. These kind of questions concern cause-effect relationships, terms that scientific researchers ideally seek to claim from their results, but often misuse. A humorous example is the claim that eating more chocolate makes you smarter (Messerli (2012)).

The gold standard for stating a cause-effect relationship is a randomised controlled experiment, where causal results are obtainable due to the construction of the experiment and the collection of the data (Oehlert (2010)). The statistical field of causal inference extends this practice to observational data to formalise when a statistical result can be claimed as causal (Pearl (2009); Peters, Janzing, and Schlkopf (2017)).

Even with an established cause-effect relationship, the question of *how* might still be unanswered. As an example, a genome wide association study (Hung, McKay, Gaborieau, Boffetta, Hashibe, Zaridze, Mukeria, Szeszenia-Dabrowska, Lissowska, Rudnai, et al. (2008)) found genomic variations that might cause an increased risk of lung cancer. These genomic variations are within a region of genes encoding proteins that bind to a nicotine derivative and are expressed in neurons. The same genomic variations have also been found to possibly cause an increased tendency to smoke (Liu, Tozzi, Waterworth, Pillai, Muglia, Middleton, Berrettini, Knouff, Yuan, Waeber, et al. (2010)). A natural question is how these genomic variations might cause lung cancer: is it by increasing the person's amount of smoking, which then causes lung cancer, or is it a coincidence that these genomic variations associated with both increased smoking and lung cancer.

Causal mediation analysis is the area of statistical causal inference that addresses this question by asking whether the causal effect is mediated through another variable. The causal effect is dissected into a direct effect and a mediated effect, which helps to elucidate the underlying mechanism of the causal effect (Baron and Kenny (1986), Imai, Keele, and Yamamoto (2010)). Following the genomic studies of smoking and lung cancer, VanderWeele et al. (2012) performed a causal mediation analysis to test whether a tendency to smoke more mediates the effect of the genomic variations on lung cancer. They came to the conclusion that the effect was primarily not mediated through smoking tendency. That is, the direct effect was significant and explained most of the effect on lung cancer, while the mediated effect was insignificant.

Interpretation of causal mediation analysis has to be done carefully. The results of Vander-

Weele et al. (2012) do not disprove that smoking causes lung cancer. They do suggest that the genomic susceptibility to lung cancer works through some other biological pathways than just increasing the person's tendency to smoke, though those pathways are likely related to nicotine response in some way.

Mediation analysis was initially studied for linear regression models, where the procedure is to fit three different linear models and test relevant hypotheses (Baron and Kenny (1986)). This practice has been extensively studied (MacKinnon, Lockwood, Hoffman, West, and Sheets (2002)) and implemented in available software, including R (Tingley, Yamamoto, Hirose, Keele, and Imai (2014)). The assumption of a linear model simplifies the mechanism under investigation, making it possible to obtain interpretable results with low computational effort. However, in today's rapidly evolving field of statistical machine learning the increase in powerful nonparametric regression models is changing the tide (Hastie, Tibshirani, and Friedman (2009)).

Regarding causal mediation analysis, the theory has been extended to a nonparametric identification (Pearl (2001); Imai et al. (2010)), but the practical implementation of a general fully nonparametric causal mediation model has not yet caught up. The R package by Tingley et al. (2014) includes the possibility of using generalised additive models and a recent method by Huber, Hsu, Lee, and Lettry (2020), also implemented in R, offers a semiparametric model, where one has to assume normally distributed errors.

In this thesis we present a new angle of a general fully nonparametric implementation of the causal mediation model. We combine a theoretical nonparametric identification result (Imai et al. (2010)) with a recently developed statistical learning method called Distributional Random Forest (Ćevid, Michel, Näf, Meinshausen, and Bühlmann (2020)) to produce an estimator for the direct and mediated effects. The estimator supports continuous or binary variables, a multivariate mediator as well as interaction effects between the treatment and mediator on the outcome. We demonstrate the inadequacy of a naive implementation of the estimator by highlighting its bias. We then propose a bias-reducing modification that incorporates the so-called generalised propensity score. The propensity score is a well established concept in causal inference (Rosenbaum and Rubin (1983)). It is mostly studied in the context of binary treatment variables but has been extended to continuous treatments under the term generalised propensity score (Hirano and Imbens (2004)). We work out the relevant theory and demonstrate empirically that our modified estimator successfully decreases the observed bias of the initial estimator.

With an extensive simulation study we investigate the performance of our proposed estimator under various data generating models. We showcase where its strengths lie and where it breaks down. We benchmark our estimator against the method proposed by Huber et al. (2020), as it is the closest method to ours in terms of generality. The theoretical guarantees for identifying the causal direct and mediated effects rely on controversial assumptions that are untestable with real world data (Imai et al. (2010)). With simulations, we investigate the behaviour of our estimator when these assumptions are violated, in order to understand its robustness. All results in the thesis are produced in Python and R. The code to reproduce it is available through Github (`github.com/AlexanderGud/mediation_drf_gps`).

Most of the theory of causal mediation analysis was developed under the notation of potential outcomes proposed by Rubin (Rubin (2005)). However, in this thesis we will adopt the notation of directed acyclic graphs and structural equation models due to Pearl (2009). In his paper, Pearl (2014) transcribed Rubin's notation of potential outcomes and

causal mediation analysis into his graphical and structural equation notation.

We will begin in Chapter 2 by formally defining the framework of causal inference under which we will be working in this thesis. We will introduce all the necessary statistical theory and methodology needed for the rest of the thesis. In Chapter 3 we will draw on this theory and methodology to construct estimators for the mediated and direct causal effects. We will present and discuss algorithms for implementing these estimators. In Chapter 4 we will put these estimators to the test in various simulation scenarios in order to investigate where their strengths and limitations lie. Finally, in Chapter 5 we will summarise our work and discuss potential next steps in this field.

Chapter 2

Framework

We begin by laying down the causal analysis framework we will work with in this thesis. In Section 2.1 we will define the fundamental concepts of causal models and potential outcomes. In Section 2.2 we will introduce the specific causal mediation model which will be of prime interest. Then, in Section 2.3 we will introduce the necessary foundations of the statistical method called Distributional Random Forest. In the last section of this chapter, Section 2.4, we will cover the fundamentals of propensity scores. In Subsection 2.4.1 we will cover the case of a binary treatment and in Subsection 2.4.2 we will extend the theory to continuous treatments.

2.1 Causal Models and Potential Outcomes

Definition 2.1.1 (Causal Model)**.** A *causal model* is a pair $(\mathbf{Z}, \mathcal{G})$ consisting of random variables $\mathbf{Z} \in \mathbb{R}^p, \mathbf{Z} = \{Z^{(1)}, Z^{(2)}, ..., Z^{(p)}\}$ and a directed acyclic graph (DAG) $\mathcal{G}$ with p nodes which induces a set of equations

$$Z^{(i)} := f_i(Z^{(pa(i))}, \varepsilon_i), \ i = 1, 2, ..., p$$

for some real-valued functions $(f_i)_{i=1}^p$ and zero-mean random variables $(\varepsilon_i)_{i=1}^p \in \mathbb{R}$. Where $\text{pa}(i)$ denotes the set of all nodes in $\mathcal{G}$ with an edge pointing to node i, i.e. the parental set of node i. When there is a directed path from node i to node j, we say that the random variable $Z^{(i)}$ causes the random variable $Z^{(j)}$. The induced set of equations is referred to as a structural equation model (SEM). Any given SEM

$$Z^{(i)} := f_i(Z^{(1_i)}, Z^{(2_i)}, ..., Z^{(l_i)}, \varepsilon_i), \ i = 1, 2, ..., p$$

also induces a causal model $(\mathbf{Z}, \mathcal{G})$ with $\mathbf{Z} = \{Z^{(1)}, Z^{(2)}, ..., Z^{(p)}\}$ and the edges of the graph are drawn from nodes $1_i, 2_i, ..., l_i$ to node i, i.e. $\text{pa}(i) := \{1_i, 2_i, ..., l_i\}$ for each $i = 1, 2, ..., p$.

A simple example of a causal model is $X \to Y$, which would be represented with the SEM:

$$\begin{aligned} X &:= f_X(\varepsilon_X) \\ Y &:= f_Y(X, \varepsilon_Y). \end{aligned}$$

In this model we would say that X causes Y but not vice versa. The intuition is that if one would intervene on X then Y would change according to f_Y, but if one intervenes on Y then X remains unaffected.

We will discuss the concept of confounders in this thesis, therefore we define the term here.

Definition 2.1.2 (Confounder)**.** Suppose $(\mathbf{Z}, \mathcal{G})$ is a causal model and denote by X, Y, C some variables in $\mathbf{Z}$ such that X causes Y. Then the variable C is called a *confounder* for the $X \to Y$ relationship if there is a directed path from C to X and from C to Y.
A confounder is called *hidden* or *unmeasured* if we do not have realisations of it, i.e. if we can not observe it.

Here on out, let $\mathcal{T}, \mathcal{M} \subset \mathbb{R}^d$ for some $d \in \mathbb{N}$.

Definition 2.1.3 (Potential Outcomes)**.** Let $(\mathbf{Z}, \mathcal{G})$ be a causal model. Denote by X, T, Y some variables, or sets of variables, in $\mathbf{Z}$ where $\text{pa}(Y) = \{T, X\}$, and suppose $Y \in \mathbb{R}$, $T \in \mathcal{T}$. Therefore, for some function f_Y and a zero-mean random variable $\varepsilon_Y \in \mathbb{R}$, we have

$$Y := f_Y(T, X, \varepsilon_Y).$$

Then $\forall t \in \mathcal{T}$, we define the *potential outcome* of Y when $T = t$ by

$$Y(t) := f_Y(t, X, \varepsilon_Y).$$

This definition entails that the collection of random variables

$$\{Y(t)\}_{t \in \mathcal{T}} \in \mathbb{R}$$

fulfills the property that for any realisation (T_i, Y_i) we have

$$Y(T_i)_i = Y_i.$$

In words, for any realisation of the random variables (i.e. for any observed data sample), if the realisation of T is t, then the realisation of Y equals the realisation of $Y(t)$.
Counterfactually, $Y(t)$ is the random variable Y in the "parallel universe" where T was set to t. Note, that generally $P(Y(t)) \neq P(Y \mid T = t)$. A simple example is the causal model $Y \to T$, then $P(Y \mid T = t)$ depends on t while $P(Y(t))$ does not, because in this causal model $Y := f_Y(\varepsilon_Y)$ so $Y(t) := Y \ \forall t$. However in the causal model $T \to Y$, then $P(Y(t)) = P(Y \mid T = t)$. The relationship between the two distributions becomes more complicated with a more complicated causal model. From the definition, no matter the causal model, we always have $P(Y(t) \mid T = t) = P(Y \mid T = t)$.
A simple intuitive view of this is e.g. in a controlled experiment of one sample, with a treatment $T = 0, 1$ and an outcome Y. Before you conduct the experiment, you have two potential outcomes, $Y(0)$ and $Y(1)$. Say you set $T = 1$ and you measure your outcome, then you have one observation $Y = Y(1)$ and one counterfactual outcome $Y(0)$.

Generally, in any causal model, the total causal effect of one variable on another can be defined with the potential outcomes,

Definition 2.1.4 (Total Causal Effect)**.** Let $(\mathbf{Z}, \mathcal{G})$ be a causal model. Suppose $T, Y \in \mathbf{Z}$. The *total causal effect* of changing T from some t to t' on the outcome Y is defined as

$$\mathbb{E}[Y(t') - Y(t)].$$

However, the *fundamental problem of causal inference* is that with realised data we only ever observe one potential outcome per sample. In order to have any hope of estimating this quantity, one needs an additional assumption.

Definition 2.1.5 (Ignorability assumption)**.** Let $(\mathbf{Z}, \mathcal{G})$ be a causal model. Suppose $T, Y \in \mathbf{Z}$ and $X \subset \mathbf{Z}$ such that none of the variables in X are caused by T.
If the following holds,

i.) $Y(t) \perp T \mid X$

ii.) $p(T = t \mid X) > 0$

$\forall t \in \mathcal{T}$, where $p(T = t \mid X)$ denotes the conditional density of T given X,
then we say that *T is ignorable for Y given X.*
If T is multivariate, the first condition has to hold for each coordinate of T.

This assumption states that, conditional on X, there is no hidden confounder for the $T \to Y$ relationship. With this assumption we have the following theorem, which along with its proof is adapted from Pearl (2009).

Theorem 2.1.6 (Identification of the Total Causal Effect)**.** *Let $(\boldsymbol{Z}, \mathcal{G})$ be a causal model. Suppose $T, Y \in \boldsymbol{Z}$ and $X \subset \boldsymbol{Z}$ such that none of the variables in X are caused by T. Assume T is ignorable for Y given X, then for any $t \in \mathcal{T}$*

$$\mathbb{E}[Y(t)] = \mathbb{E}[\mathbb{E}[Y \mid X, T = t]].$$

Proof. Let $t \in \mathcal{T}$. By the law of iterated expectation

$$\mathbb{E}[Y(t)] = \mathbb{E}[\mathbb{E}[Y(t) \mid X]]$$

Then by the ignorability assumption, this is equal to

$$\mathbb{E}[\mathbb{E}[Y(t) \mid X,\ T = t]]$$

and by the definition of the potential outcome we have

$$\mathbb{E}[\mathbb{E}[Y(t) \mid X,\ T = t]] = \mathbb{E}[\mathbb{E}[Y \mid X,\ T = t]].$$

Thus for all $t \in \mathcal{T}$,

$$\mathbb{E}[Y(t)] = \mathbb{E}[\mathbb{E}[Y \mid X,\ T = t]].$$

□

Here, by identification, we mean that the quantity of interest on the left hand side is observable, that is, it can be estimated with realised data. We will build on these results in the following section.

2.2 The Causal Mediation Model

Let $(\mathbf{Z}, \mathcal{G})$ be a causal model where $\mathbf{Z} = \{X, T, M, Y\}$, $\mathcal{G}$ is the DAG in Figure 2.1 with the corresponding structural equation model,

$$\begin{aligned} X &:= f_X(\varepsilon_X) \in \mathbb{R}^p \\ T &:= f_T(X, \varepsilon_T) \in \mathcal{T} \\ M &:= f_M(X, T, \varepsilon_M) \in \mathcal{M} \\ Y &:= f_Y(X, T, M, \varepsilon_Y) \in \mathbb{R} \end{aligned} \tag{2.2.1}$$

for some functions f_X, f_T, f_M, f_Y, zero-mean random variables $\varepsilon_X, \varepsilon_T, \varepsilon_M, \varepsilon_Y$ and subsets $\mathcal{T}, \mathcal{M} \subset \mathbb{R}^d$ for some $d \in \mathbb{N}$.

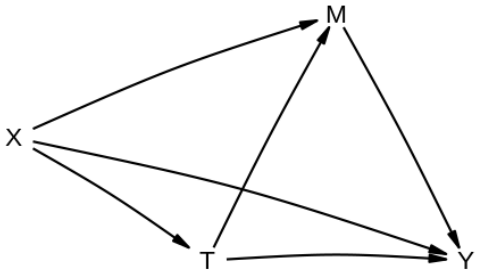

Figure 2.1: The DAG of the causal mediation model with corresponding SEM 2.2.1

We emphasize that we assume here that the causal model is correctly specified, our goal is then to estimate the causal effects, i.e. estimate the parts the functions f_M, f_Y that depend on T and M.
Henceforth we will refer to X as the pre-treatment covariates, T as the treatment, M as the mediator and Y as the outcome.
Denote by $\{\mathbf{Z}_i\}_{i=1}^n$, $\mathbf{Z}_i = \{X_i, T_i, M_i, Y_i\}$ a set of n realisations of the random variables, or data. Let $M(t)$ be the potential outcome of M when $T = t$ and $Y(t, m)$ the potential outcome of Y when $T = t$ and $M = m$.

Definition 2.2.1 (Dose-Response Function)**.** The *dose response function*, $\xi(t_d, t_m) : \mathcal{T} \times \mathcal{T} \to \mathbb{R}$, is defined as

$$\xi(t_d, t_m) := \mathbb{E}[Y(t_d, M(t_m))].$$

Definition 2.2.2 (Average Direct and Mediated Effect)**.** The *average causal direct effect* (ACDE) and *average causal mediated effect* (ACME) of the treatment on the outcome are defined as the partial derivatives of the dose-response function $\xi(t_d, t_m)$,

$$\begin{aligned} \text{ACDE}(t_d, t_m) &:= \frac{\partial \xi}{\partial t_d}(t_d, t_m) \\ \text{ACME}(t_d, t_m) &:= \frac{\partial \xi}{\partial t_m}(t_d, t_m). \end{aligned}$$

The *average total causal effect* of the treatment on the outcome is defined as the derivative of the univariate function

$$t \mapsto \xi(t, t).$$

The task of estimating this dose-response function is rather ambitious as we have a nested potential outcome and we can say that the *fundamental problem of causal mediation analysis* is that for any $t_d \neq t_m$ we never observe $Y(t_d, M(t_m))$. We thus need some ignorability assumptions similar to the one described above for the total causal effect,

Definition 2.2.3 (Sequential Ignorability Assumptions)**.** If the following holds,

i.) $\{M(t), Y(t', m)\} \perp T \mid X$

ii.) $M(t) \perp Y(t', m) \mid T, X$

iii.) $p(T = t \mid X) > 0 \;\&\; p(M = m \mid X,\ T) > 0$

for all $t, t' \in \mathcal{T},\ m \in \mathcal{M}$, where $p(T \mid X)$ and $p(M \mid X,\ T)$ denote the conditional densities of T given X and M given X and T, respectively,
then we say that *T and M are sequentially ignorable for Y given X.*
If T or M are multivariate, the conditions have to hold for each of their coordinates.

With the definition of the potential outcomes the first two assumptions can be rewritten as

i.) $\{\varepsilon_M, \varepsilon_Y\} \perp \varepsilon_T \mid X$

ii.) $\varepsilon_M \perp \varepsilon_Y \mid T, X$

Here the first assumption states that conditional on X, there is no unmeasured confounder for the $T \to M$ relationship or the $T \to Y$ relationship. The second assumption states that, conditional on X and T, there is no unmeasured confounder of the $M \to Y$ relationship. We stress that it is implicit in the definition of the causal mediation model that none of the variables in X can be caused by T, i.e. there are no post-treatment variables. Indeed, if there is a post-treatment variable confounding M and Y then the second assumption would be violated. Conditioning on such a variable would block the causal pathways from T to M and Y, leading to inaccurate estimates. In such a case, one would either have to choose a different research question and adopt a new causal model, or employ a multi-stage mediator method, which have been studied to some extent (VanderWeele and Vansteelandt (2014)).

These cross-world conditions are the spark of most of the controversy around mediation analysis, because they are untestable with realised data (Bullock, Green, and Ha (2010)). Even in a randomised controlled experiment where the treatment is randomised, the second condition may still not be satisfied. In order to accommodate, Imai et al. (2010) propose a sensitivity analysis, which allows to quantify how severely the second condition might be violated and possibly correct for it. They develop their sensitivity analysis primarily for linear models and discuss an extension to a nonparametric approach, but only for a binary mediator.

Given that the sequential ignorability assumptions hold, we have the desired result of being able to observe the dose-response function. The theorem and proof are adapted from the work of Imai et al. (2010). We begin with a small lemma.

Lemma 2.2.4. *Assume T and M are sequentially ignorable for Y given X, then*

$$Y(t', m) \perp T \mid X,\ M(t) = m$$

for all $t, t' \in \mathcal{T},\ m \in \mathcal{M}$.

Proof. Let $t, t' \in \mathcal{T},\ m \in \mathcal{M},\ x \in \mathbb{R}^p$.
We show that

$$P(Y(t', m) \mid X = x,\ M(t) = m) = P(Y(t', m) \mid T = t,\ X = x,\ M(t) = m).$$

By conditioning on T and integrating it out we get

$$\begin{aligned} &P(Y(t', m) \mid X = x,\ M(t) = m) \\ &= \int P(Y(t', m) \mid T = t_0,\ X = x,\ M(t) = m) dP(T = t_0 \mid X = x,\ M(t) = m). \end{aligned}$$

By the first ignorability assumption $P(T = t_0 \mid X = x,\ M(t) = m) = P(T = t_0 \mid X = x)$ and by the second ignorability assumption $P(Y(t', m) \mid T = t_0,\ X = x,\ M(t) = m) = P(Y(t', m) \mid T = t_0,\ X = x)$. Therefore,

$$\begin{aligned} &= \int P(Y(t', m) \mid T = t_0,\ X = x,\ M(t) = m) dP(T = t_0 \mid X = x,\ M(t) = m) \\ &= \int P(Y(t', m) \mid T = t_0,\ X = x) dP(T = t_0 \mid X = x). \end{aligned}$$

Now, by the first ignorability assumption, we can change the value of T,

$$= \int P(Y(t', m) \mid T = t,\ X = x) dP(T = t_0 \mid X = x)$$

But then the integrand no longer depends on t_0, the variable being integrated over, therefore this equals to

$$P(Y(t', m) \mid T = t,\ X = x).$$

Finally, by applying the second ingorability assumption again we obtain that this is equal to

$$P(Y(t', m) \mid T = t,\ X = x,\ M(t) = m).$$

Thus for all $t, t' \in \mathcal{T},\ m \in \mathcal{M}$

$$P(Y(t', m) \mid X = x,\ M(t) = m) = P(Y(t', m) \mid T = t,\ X = x,\ M(t) = m).$$

□

Theorem 2.2.5 (Identification of the Dose-Response Function)**.** *Assume T and M are sequentially ignorable for Y given X, then for any $t_d, t_m \in \mathcal{T}$*

$$\begin{aligned} \xi(t_d, t_m) &:= \mathbb{E}[Y(t_d, M(t_m))] \\ &= \int \int \mathbb{E}[Y \mid X = x,\ T = t_d,\ M = m] dP(M = m \mid X = x,\ T = t_m) dP(X = x). \end{aligned}$$

Proof. Let $t_d, t_m \in \mathcal{T}$. We will apply the law of iterated expectation, both of the sequential ignorability assumptions and the definition of potential outcomes.
By applying the law of iterated expectation two times we obtain,

$$\begin{aligned} &\mathbb{E}[Y(t_d, M(t_m))] \\ &= \int \mathbb{E}[Y(t_d, M(t_m)) \mid X = x] dP(X = x) \\ &= \int \int \mathbb{E}[Y(t_d, m) \mid X = x,\ M(t_m) = m] dP(M(t_m) = m \mid X = x) dP(X = x) \end{aligned}$$

We apply Lemma to condition on T in the integrand and we apply the first sequantial ignorability assumption to condition on T in the distribution of M we integrate over, to get,

$$\int \int \mathbb{E}[Y(t_d, m) \mid X = x,\ T = t_m,\ M(t_m) = m] dP(M(t_m) = m \mid X = x,\ T = t_m) dP(X = x)$$

By the definition of potential outcomes,

$$P(M(t_m) = m \mid X = x,\ T = t_m) = P(M = m \mid X = x,\ T = t_m)$$

which is observable.
We now focus on the integrand. By the second sequential ignorability assumption we can disregard the conditioning on $M(t_m)$,

$$\begin{aligned} &\mathbb{E}[Y(t_d, m) \mid X = x,\ T = t_m,\ M(t_m) = m] \\ =&\mathbb{E}[Y(t_d, m) \mid X = x,\ T = t_m] \end{aligned}$$

Then by the first ignorability assumption, $Y(t_d, m)$ is conditionally independent of T given X so we can get,

$$\mathbb{E}[Y(t_d, m) \mid X = x,\ T = t_d]$$

Again applying the second ignorability assumption we bring in $M(t_d)$,

$$\mathbb{E}[Y(t_d, m) \mid X = x,\ T = t_d,\ M(t_d) = m]$$

Now by the definition of the potential outcomes this is equal to,

$$\mathbb{E}[Y \mid X = x,\ T = t_d,\ M = m]$$

Thus, finally, for any $t_d, t_m \in \mathcal{T}$,

$$\begin{aligned} &\mathbb{E}[Y(t_d, M(t_m))] \\ &= \int \int \mathbb{E}[Y \mid X = x,\ T = t_d,\ M = m] dP(M = m \mid X = x,\ T = t_m) dP(X = x) \end{aligned}$$

□

The value in this theorem is that our object of interest, the dose-response function, can be evaluated with observed quantities like the conditional distributions of the observed variables Y and M, instead of the unobservable cross-world potential outcomes in the function definition. This is what we mean by identification. Therefore, with realised data of these random variables, we will need to empirically estimate the conditional distributions. This is where the newly proposed method called Distributional Random Forest comes in handy.

2.3 Distributional Random Forest

The Distributional Random Forest (DRF) method (Ćevid et al. (2020)) builds on the established Random Forest method (Breiman (2001)), but instead of estimating only the conditional mean of an outcome given some covariates, $\mathbb{E}[Y \mid X = x]$, it estimates the entire conditional distribution, $P(Y \mid X = x)$. This estimation is done completely non-parametrically. We will state the relevant foundations and refer to the original paper by Ćevid et al. (2020) for a thorough description of the method.

Let $Y \in \mathbb{R}^d,\ X \in \mathbb{R}^p,\ X = \{X^{(1)}, X^{(2)}, ..., X^{(p)}\}$, be random variables and $\{X_i, Y_i\}_{i=1}^n$ a set of paired realisations of those random variables, or data. We call X the covariates and Y the outcome. We are interested in estimating the conditional distribution $P(Y \mid X = x)$. The random forest consists of a set of decision trees, where each tree is built on a random subset of the data. Each tree is built recursively, with nodes representing pools of samples, a node $\mathcal{I} \subset \{1, ..., n\}$ is split in two child nodes by first taking a random subset of the covariates, then finding the one among them, $X^{(j)}$, such that for some $l \in \mathbb{R}$ the two pools $\{Y_i : i \in \mathcal{I}, X_i^{(j)} \leq l\}$ and $\{Y_i : i \in \mathcal{I}, X_i^{(j)} > l\}$ differ the most in terms of an empirical distributional metric. The distributional metric currently used is called the Maximum Mean Discrepancy (Gretton, Borgwardt, Rasch, Schölkopf, and Smola (2006)). The resulting leaf nodes will thus represent a homogeneous pool of the outcome given the covariates.

The result of this procedure is a weighting function, which, given any test point $x \in \mathbb{R}^p$, will weigh the input data in such a way that the weighted sample estimates the conditional distribution $P(Y \mid X = x)$. Specifically, we drop the point x down each of our N trees, let $\mathcal{L}_k(x)$ denote the leaf node of tree k where x belongs. The weighting function is defined as

$$w_x(X_i) = \frac{1}{N} \sum_{k=1}^{N} \frac{\mathbb{1}\{X_i \in \mathcal{L}_k(x)\}}{|\mathcal{L}_k(x)|}.$$

With this the conditional distribution can be estimated empirically by

$$\widehat{P}(Y \mid X = x) := \sum_{i=i}^{n} w_x(X_i) \cdot \delta_{Y_i}.$$

Where δ_{Y_i} denotes the point mass at Y_i.

In Ćevid et al. (2020) this estimator is proven to be asymptotically consistent under certain regularity assumptions. We will apply this method in order to empirically estimate the conditional distribution, $P(M = \cdot \mid X, T)$, present in Theorem 2.2.5.

2.4 Generalised Propensity Scores

The last of the preliminaries we will cover is the fundamentals of propensity scores. We will build on these results later in the thesis. Propensity scores were initially constructed for binary treatments by Rosenbaum and Rubin (1983), and are most extensively used in that scenario. We cover that case in Subsection 2.4.1. As before mentioned, we are interested in applications for continuous treatment variables, thus we will cover in Subsection 2.4.2 how propensity scores can be generalised to that scenario. The theorems and proofs in this subsection are adapted from the work of Rosenbaum and Rubin (1983) and Hirano and Imbens (2004).

2.4.1 Binary Treatment

In this subsection let $(\{X, T, Y\}, \mathcal{G})$ be a causal model with SEM,

$$\begin{aligned} X &= f_X(\varepsilon_X) \in \mathbb{R}^p \\ T &= f_T(X, \varepsilon_T) \in \{0, 1\} \\ Y &= f_Y(X, T, \varepsilon_Y) \in \mathbb{R} \end{aligned}$$

and let $\{X_i, T_i, Y_i\}_{i=1}^n$ be realisations of them, or data. Let $Y(0), Y(1)$ denote the potential outcomes of Y under treatment assignments $T = 0, 1$, defined as in Definition 2.1.3. Here we are interested in estimating the total causal effect

$$\mathbb{E}[Y(1) - Y(0)].$$

And we assume that T is ignorable for Y given X, so this could well be done by the identification theorem for the total causal effect, described in Theorem 2.1.6. Here we describe an alternative approach.

Definition 2.4.1 (Propensity Score)**.** For all $x \in \mathbb{R}^p$ the *propensity score* of the treatment T given $X = x$ is defined as

$$r(x) := P(T = 1 \mid X = x).$$

For each sample X_i, the score $r(X_i)$, describes how likely the sample was to receive the treatment $T = 1$, depending on the sample's pre-treatment variables.

The relevance of the propensity score derives from the following results.

Theorem 2.4.2 (Balancing Property of the Propensity Score)**.** *The propensity score is a balancing function, meaning,*

$$T \perp X \mid r(X).$$

Theorem 2.4.3 (Ignorability with the Propensity Score)**.** *Assume T is ignorable for Y given X. Then T is ignorable for Y given $r(X)$, meaning*

$$\{Y(0), Y(1)\} \perp T \mid r(X).$$

Theorem 2.4.4 (Identification of the Total Causal Effect with Propensity Scores)**.** *Assume T is ignorable for Y given X. Then for $t = 0, 1$ we have*

$$\mathbb{E}[Y(t)] = \mathbb{E}[\mathbb{E}[Y \mid r(X), T = t]].$$

We will prove these results in the more general case of a continuous treatment below. Theoretically, this result provides no advantage over the previous identification strategy presented in Theorem 2.1.6, but as the saying goes, in theory there is no difference between theory and practice, but in practice there is. In situations where p, the dimensionality of X, is rather high the propensity score is a way to reduce all the relevant information in X down to one scalar.

2.4.2 Continuous Treatment

Now, we will generalise these results to a continuous treatment. In this subsection let $(\{X, T, Y\}, \mathcal{G})$ be a causal model with SEM

$$\begin{aligned} X &= f_X(\varepsilon_X) \in \mathbb{R}^p \\ T &= f_T(X, \varepsilon_T) \in \mathcal{T} \\ Y &= f_Y(X, T, \varepsilon_Y) \in \mathbb{R}, \end{aligned}$$

and let $\{X_i, T_i, Y_i\}_{i=1}^n$ be realisations of this SEM. Let $\{Y(t)\}_{t\in\mathcal{T}}$ denote the potential outcomes of Y under any treatment assignment $T = t$, defined as in Definition 2.1.3. Here we are interested in estimating the function

$$t \mapsto \mathbb{E}[Y(t)],$$

from which we can estimate the total causal effect. We assume that T is ignorable for Y given X.

Definition 2.4.5 (Generalised Propensity Function)**.** Denote the conditional density of T given X by

$$\begin{aligned} &p(T = \cdot \mid X) : \mathcal{T} \times \mathbb{R}^p \to \mathbb{R}_+ \\ &(t, x) \mapsto p(T = t \mid X = x) \end{aligned}$$

The generalised propensity function is defined as,

$$r(t, x) := p(T = t \mid X = x).$$

For each sample X_i, the function $t \to r(t, X_i)$ describes how likely the sample was to receive any given treatment $T = t$, depending on the sample's pre-treatment variables. The value $r(T_i, X_i)$ describes how likely the sample was to receive the treatment it actually received.

We have analogous results for the generalised propensity score as with the binary case.

Theorem 2.4.6 (Balancing Property of the Generalised Propensity Function)**.** *The generalised propensity function is a balancing function, meaning,*

$$p(T = t \mid X,\ r(t, X)) = p(T = t \mid r(t, X))\ \forall t \in \mathcal{T}.$$

Proof. Let $t \in \mathcal{T}$. Firstly, as $r(t, X)$ is a function of X we have

$$p(T = t \mid X,\ r(t, X)) = p(T = t \mid X) = r(t, X).$$

Secondly, by the law of iterated expectation,

$$\begin{aligned} p(T = t \mid r(t, X)) &= \int p(T = t \mid X = x,\ r(t, X)) dP(X = x \mid r(t, X)) \\ &= \int p(T = t \mid X = x) dP(X = x \mid r(t, X)) \\ &= \int r(t, X) dP(X = x \mid r(t, X)) \\ &= r(t, X) \end{aligned}$$

Thus, for any $t \in \mathcal{T}$

$$p(T = t \mid X,\ r(t, X)) = r(t, X) = p(T = t \mid r(t, X)).$$

□

Theorem 2.4.7 (Ignorability with the Generalised Propensity Function)**.** *Suppose T is ignorable for Y given X, then,*

$$p(Y(t) \mid T = t,\ r(t, X)) = p(Y(t) \mid r(t, X))\ \forall t \in \mathcal{T}.$$

Proof. Let $t \in \mathcal{T}$. We begin by showing that

$$p(T = t \mid Y(t),\ r(t, X)) = p(T = t \mid r(t, X)),$$

then we will apply Bayes' rule to obtain the result.
By the previous theorem it suffices to show

$$p(T = t \mid Y(t),\ r(t, X)) = r(t, X).$$

By the law of iterated expectation we have for any $r \in \mathbb{R}$

$$\begin{aligned} &p(T = t \mid Y(t),\ r(t, X) = r) \\ &= \int p(T = t \mid X = x,\ Y(t),\ r(t, X) = r) dP(X = x \mid Y(t),\ r(t, X) = r). \end{aligned}$$

Then by the assumption that T is ignorable for Y given X, we have

$$\begin{aligned} p(T = t \mid X = x,\ Y(t),\ r(t, X) = r) &= p(T = t \mid X = x) \\ &= r(t, x)\mathbb{1}\{r(t, x) = r\} \end{aligned}$$

Therefore

$$\begin{aligned} p(T = t \mid Y(t),\ r(t, X) = r) &= \int r(t, x)\mathbb{1}\{r(t, x) = r\} dP(X = x \mid Y(t),\ r(t, X) = r) \\ &= r. \end{aligned}$$

Thus for any $t \in \mathcal{T}$

$$p(T = t \mid Y(t),\ r(t, X)) = p(T = t \mid r(t, X)).$$

Now by Bayes' rule, we obtain for any $t \in \mathcal{T}$

$$\begin{aligned} p(Y(t) \mid T = t,\ r(t, X)) &= \frac{p(T = t \mid Y(t),\ r(t, X))\ p(Y(t) \mid r(t, X))}{p(T = t \mid r(t, X))} \\ &= p(Y(t) \mid r(t, X)) \end{aligned}$$

□

Theorem 2.4.8 (Identification of the Total Causal Effect with the Generalised Propensity Function)**.** *Suppose T is ignorable for Y given X. Then for any $t \in \mathcal{T}$ we have*

$$\mathbb{E}[Y(t)] = \mathbb{E}[\mathbb{E}[Y \mid r(t, X), T = t]].$$

Proof. Let $t \in \mathcal{T}$. By the law of iterated expectation, the previous theorem and the definition of potential outcomes, respectively, we have

$$\begin{aligned} \mathbb{E}[Y(t)] &= \mathbb{E}[\mathbb{E}[Y(t) \mid r(t, X)]] \\ &= \mathbb{E}[\mathbb{E}[Y(t) \mid r(t, X),\ T = t]] \\ &= \mathbb{E}[\mathbb{E}[Y \mid r(t, X),\ T = t]] \end{aligned}$$

□

In the next chapter we will build on all the theory presented in the current chapter and apply it to build empirical estimators of causal mediation effects for continuous treatments.

Chapter 3

Estimation of the Causal Effects

In this chapter we will empirically estimate the dose-response function defined in Definition 2.2.1. In Section 3.1 we will apply the equation presented in Theorem 2.2.5 directly to construct an algorithm. In Subsection 3.1.1 we will discuss how to quantify the uncertainty of the estimation. In Subsection 3.1.2 we will discuss the choice of a regressor and in subsection 3.1.3 we will discuss a potential drawback of the algorithm. In Section 3.2 we will address this drawback, propose an adjustment and present a new algorithm. Throughout this chapter we assume we have a causal mediation model and realised data as described in Section 2.2.

3.1 Covariate-Adjustment Estimator

If we analyse the parts that make up the equation of Theorem 2.2.5, we find that we can perform a Monte Carlo integration strategy to evaluate the double integral present in the equation. The general procedure of a Monte Carlo estimator is to sample random variables from the relevant distributions, evaluate the integrand with the samples and then average over the results. The two distributions we need to integrate over are $P(M \mid X, T)$ and $P(X)$, respectively. For the conditional distribution $P(M \mid X, T)$ we suggest here to apply the Distributional Random Forest method. Our integrand in question is the expected value of Y given X, T and M. This task can for example be solved with any type of regressor with inputs $\{X, T, M\}$ and output Y. We will discuss the choice of the regressor in more detail below. The pseudo-code for this Monte Carlo estimator is presented here.

Algorithm 3.1.1 (Estimating the Dose-Response Function by Covariate Adjustment)**.**

i.) Estimate the conditional distribution of M given X and T, $P(M \mid X,\ T)$, with a Distributional Random Forest with inputs X, T and output M. Denote the estimate by $\widehat{P}(M \mid X,\ T)$.

ii.) Train a regressor function $G(X, T, M)$ to predict Y.

iii.) Estimate $\xi(t_d, t_m)$ by $\widehat{\xi}(t_d, t_m) := \frac{1}{k}\frac{1}{l}\sum_{i=1}^{k}\sum_{j=1}^{l} G(\tilde{x}_i, t_d, \tilde{m}_{i,j})$, where the $\tilde{x}_i$ are sampled from $X_1, ..., X_n$ and the $\tilde{m}_{i,j}$ are sampled from $\widehat{P}(M \mid X = \tilde{x}_i,\ T = t_m)$ (both with replacement).

We can say this algorithm has three distinct parts, corresponding to the three steps, the training of the Distributional Random Forest, the training of the regressor and then the estimator itself, which is just the empirical version of the equation in Theorem 2.2.5.

Once you have trained, you can estimate the dose-response function at all points of interest (t_d, t_m). Estimating the average causal direct or mediated effects as defined by the partial derivatives in Definition 2.2.2, can then be done by taking the appropriate differences between function estimates,

$$\widehat{ACDE}_h(t_d, t_m) := \frac{\widehat{\xi}(t_d + h, t_m) - \widehat{\xi}(t_d, t_m)}{h}$$
$$\widehat{ACME}_h(t_d, t_m) := \frac{\widehat{\xi}(t_d, t_m + h) - \widehat{\xi}(t_d, t_m)}{h}.$$

We note that we could also evaluate the inner integral without sampling, by directly applying the weights obtained from the Distributional Random Forest. Recall from section 2.3 that the estimated conditional distribution is

$$\widehat{P}_M(m \mid X = x,\ T = t_m) := \sum_{i=i}^{n} w_{\{x,t_m\}}(X_i, T_i) \cdot \delta_{M_i}.$$

By substituting $P(M = m \mid X = x,\ T = t_m)$ in the equation in Theorem 2.2.5 with this expression and then sampling X for the outer integral, we get

$$\widehat{\xi}(t_d, t_m) := \frac{1}{k} \sum_{i=1}^{k} \sum_{j=1}^{n} G(\tilde{x}_i, t_d, M_j) \cdot w_{\{\tilde{x}_i, t_m\}}(X_j, T_j) \tag{3.1.1}$$

where $\tilde{x}_i$ are sampled from $X_1, ..., X_n$.

This version of the estimator does eliminate the Monte Carlo sampling error for the inner integral and thus should result in a lower variance of the estimation. In practice, when n, the sample size, is large this version runs slower and takes more resources than the Monte Carlo version presented in Algorithm 3.1.1, which has tunable sampling sizes k and l to trade off accuracy and speed. In our simulations in Chapter 4, we use the Monte Carlo version with a sampling size $k = l = 1000$, which performs congruently with the weighted estimator in Equation 3.1.1.

3.1.1 Uncertainty Quantification

This estimator only delivers point estimates without any quantification of the statistical uncertainty. However, uncertainty quantification can be obtained via a bootstrap procedure.

Algorithm 3.1.2 (Bootstrapping the Dose-Response Estimator)**.**
Repeat the following steps B times (typically $B = 100$):

i.) Sample $\{\tilde{Z}_1, \tilde{Z}_2, .., \tilde{Z}_k\}$ from $\{Z_1, Z_2, ..., Z_n\}$ with replacement.

ii.) Train the Distributional Random Forest and regressor (steps i and ii above) on the drawn sample.

iii.) Evaluate $\widehat{\xi}(t_d, t_m)$ *at all points of interest.*

At each point of interest, we have B *estimates, from which we can quantify the uncertainty, for example by taking the empirical 0.025% and 97.5% quantiles, to obtain a 95% bootstrap confidence interval, assuming the bootstrap distribution is symmetric.*

3.1.2 The Regressor

As the algorithm stands, any type of regressor can be plugged in, however due to the nature of the problem, some regressors might be more beneficial than others.

Firstly, the main idea of this estimator is to perform mediation analysis nonparametrically. Therefore, a parametric regressor like a linear model or a generalised linear model would defeat the purpose. Nonparametric regressors have been substantially studied in the recent years in the field of supervised machine learning. Two distinct methods can be called the state-of-the-art today, those are neural networks and tree-based ensemble learning, like Random Forest and Boosting (Hastie et al. (2009)).

One prime difference between these two classes is that neural networks produce a prediction function that extrapolates into unknown territory beyond the observed inputs, much like a linear model does, while tree-based methods bin the observed inputs and the extrapolation beyond them is simply flattened out and cast into the closest bin. In any standard machine learning problem, this is not a big issue, as prediction outside of the observed range is often not the prime interest.

However, if we think about our scenario and recall the definition of the dose-response function, we know that it includes this "cross-world" nested potential outcome that is never observed. Even though Theorem 2.2.5 says that the expression is observable, we sample the M given $T = t_m$ and then predict Y with that M and a different $T = t_d$, which can possibly be far away from our observed inputs, into unknown territory, especially when t_m and t_d are far apart. This is the prime reason we suggest that a neural network is better suited for the job in estimating the dose-response function than tree-based ensemble methods.

We should note that prediction far away from the observed inputs can never be made with the same certainty as predictions within the range of observed inputs. This is another reason our task is rather ambitious. One should keep this in mind while employing the proposed estimator. Estimations of points with t_d far from t_m will generally be less certain.

3.1.3 A Potential Source of Bias

As stated in the framework chapter, the Distributional Random Forest estimate of the conditional distribution is proven to be asymptotically consistent under certain regularity assumptions. Thus, theoretically, using a sufficiently deep neural network (Cybenko (1989)), our estimator of the dose-response function, proposed in Algorithm 3.1.1, is asymptotically consistent.

However, as demonstrated with the simulations in Figure 4.1 in Chapter 4, we do observe an underestimation in the mediated effect, i.e. in the absolute value of the partial derivative of the dose-response function with respect to t_m. The Distributional Random Forest may account for this phenomenon.

The source of the causal mediated effect in our estimator consists of two parts. Initially

it comes from the influence T has on the estimated conditional distribution $\widehat{P}(M \mid T, X)$, derived from the Distributional Random Forest. Subsequently, the regressor for Y depends on M.

Suppose the pre-treatment variables X are of relatively high dimension and have a similar or larger effect on M than T. Then, by the nature of the Random Forest algorithm, how it randomly considers a subset of the covariates for node splitting and then chooses the optimal one among them, it is likely that the algorithm will rarely choose T as the splitting variable. This might ultimately result in an underestimation in the effect of T on M. Since the estimated mediation effect depends on the effect T has on M, this might lead to an underestimation in the mediated effect.

In the following section we will address this underestimation problem by incorporating propensity scores into the estimator.

3.2 Propensity Score Estimator

As stated in the framework chapter, the propensity score is another way of achieving ignorability of the treatment, as opposed to adjusting for all the covariates. One can view it as a dimensionality reduction scheme that summarises all the covariates into one scalar. We propose here to incorporate the propensity score method into the estimation of the dose-response function. Specifically, when estimating the conditional distribution of M given $T = t_m$, we adjust for the generalised propensity score $r(t_m, X)$ instead of all the covariates X.

Implementing this strategy requires an empirical estimation of the generalised propensity function. This can quite easily be done with the Distributional Random Forest.

Algorithm 3.2.1 (Estimating Generalised Propensity Functions with DRF)**.**

i.) Estimate $P(T \mid X)$ with a Distributional Random Forest with input $\{X_i\}_{i=0}^n$ and output $\{T_i\}_{i=0}^n$. Denote the estimate by $\widehat{P}(T \mid X)$.

ii.) For each sample $i = 1, ..., n$:

(a) Sample $\{t_j\}_{j=1}^k \sim \widehat{P}(T \mid X_i)$.

(b) Estimate the density function $p(T = t \mid X_i)$ with a kernel density estimator of the sample $\{t_j\}_{j=1}^k$. Denote the estimate by $\widehat{p}_T(t \mid X_i)$

(c) Set $\widehat{r}(t, X_i) := \widehat{p}_T(t \mid X_i)$

We present a new identification theorem incorporating the generalised propensity function.

Theorem 3.2.2 (Identification of the Dose-Response Function with Propensity Scores)**.** *Assume T and M are sequentially ignorable for Y given X. Denote by $r(t, x)$ the generalised*

propensity function of T given X. Then for any $t_d, t_m \in \mathcal{T}$

$$\begin{aligned}\xi(t_d, t_m) :&= \mathbb{E}[Y(t_d, M(t_m))] \\ &= \int\int\int \mathbb{E}[Y \mid X = x, T = t_d, M = m] \\ &\quad dP(X = x \mid M(t_m) = m,\ r(t_m, X) = r) \\ &\quad\quad dP(M = m \mid T = t_m,\ r(t_m, X) = r) \\ &\quad\quad\quad dP(r(t_m, X) = r)\end{aligned}$$

Proof. The proof is similar to that of Theorem 2.2.5.
Let $t_d, t_m \in \mathcal{T}$. We will apply the law of iterated expectation, both of the sequential ignorability assumptions, the ignorability result of the generalised propensity function and the definition of potential outcomes.
By applying the law of iterated expectation three times we obtain,

$$\begin{aligned}\mathbb{E}[Y(t_d, M(t_m))] &= \int \mathbb{E}[Y(t_d, M(t_m)) \mid r(t_m, X) = r] dP(r(t_m, X) = r) \\ &= \int\int \mathbb{E}[Y(t_d, m) \mid M(t_m) = m,\ r(t_m, X) = r] \\ &\quad dP(M(t) = m \mid r(t_m, X) = r) \\ &\quad\quad dP(r(t_m, X) = r) \\ &= \int\int\int \mathbb{E}[Y(t_d, m) \mid X = x,\ M(t_m) = m,\ r(t_m, X) = r] \\ &\quad dP(X = x \mid M(t_m) = m,\ r(t_m, X) = r) \\ &\quad\quad dP(M(t) = m \mid r(t_m, X) = r) \\ &\quad\quad\quad dP(r(t_m, X) = r)\end{aligned}$$

We apply Lemma to condition on T in the integrand and we apply the ignorability results of the generalised propensity function from Theorem 2.4.7 to condition on T in the distribution of M we integrate over to get,

$$\begin{aligned}&\int\int\int \mathbb{E}[Y(t_d, m) \mid X = x,\ T = t_m,\ M(t_m) = m,\ r(t_m, X) = r] \\ &\quad dP(X = x \mid M(t_m) = m,\ r(t_m, X) = r) \\ &\quad\quad dP(M(t) = m \mid T = t_m,\ r(t_m, X) = r) \\ &\quad\quad\quad dP(r(t_m, X) = r)\end{aligned}$$

By the definition of potential outcomes,

$$P(M(t) = m \mid T = t_m,\ r(t_m, X) = r) = P(M = m \mid T = t_m,\ r(t_m, X) = r)$$

which is observable.
We now focus on the integrand. By the second sequential ignorability assumption we can disregard the conditioning on $M(t_m)$,

$$\begin{aligned}&\mathbb{E}[Y(t_d, m) \mid X = x,\ T = t_m,\ M(t_m) = m,\ r(t_m, X) = r] \\ =&\mathbb{E}[Y(t_d, m) \mid X = x,\ T = t_m,\ r(t_m, X) = r]\end{aligned}$$

Now, since $r(t_m, X)$ is a deterministic function of X, it is conditionally independent of Y given X, thus we get,

$$\mathbb{E}[Y(t_d, m) \mid X = x,\ T = t_m].$$

Then by the first ignorability assumption, $Y(t_d, m)$ is conditionally independent of T given X so we can get,

$$\mathbb{E}[Y(t_d, m) \mid X = x,\ T = t_d]$$

Again applying the second ignorability assumption we bring in $M(t_d)$,

$$\mathbb{E}[Y(t_d, m) \mid X = x,\ T = t_d,\ M(t_d) = m]$$

Now by the definition of the potential outcomes this is equal to,

$$\mathbb{E}[Y \mid X = x,\ T = t_d,\ M = m]$$

Thus, finally for any $t_d, t_m \in \mathcal{T}$,

$$\begin{aligned}\mathbb{E}[Y(t_d, M(t_m))] = \int\int\int \mathbb{E}[Y \mid X = x,\ T = t_d,\ M = m] \\ dP(X = x \mid M(t_m) = m,\ r(t_m, X) = r) \\ dP(M = m \mid T = t_m,\ r(t_m, X) = r) \\ dP(r(t_m, X) = r)\end{aligned}$$

□

One concern arises immediately after this result. The right hand side of the equation presented is not exactly observable as promised, namely the distribution $P(X = x \mid M(t_m) = m, r(t_m, X) = r)$ still contains the potential outcome $M(t_m)$. However, in the empirical version of this equation, presented in Algorithm 3.2.3, we make a certain simplification. To sample r from the distribution of $r(t_m, X)$ we can sample $X = x$, and then plug in $r(t_m, x)$. Therefore, to sample r in the outermost integral, we draw one X_i and set $r = r(t_m, X_i)$. Then when we should sample x in the innermost integral, we simply take the one X_i we already drew. Effectively, we are estimating a double integral over the distributions $dP(M = m \mid T = t_m,\ r(t_m, X) = r(t_m, x))dP(X = x)$ instead of the one presented in Theorem 3.2.2. With this simplification, we do not have strict theoretical guarantees that our proposed method should be asymptotically consistent. However, our empirical results are excellent, as we show in Chapter 4.

Since the whole purpose of Theorem 3.2.2 was to condition of the generalised propensity score instead of the covariates, one might be wondering why we are conditioning on X in the expectation of $Y(t_d, m)$. The conditioning on $r(t_d, X)$ alone would not have been sufficient to finish the proof because we do not have the certainty of $Y(t_d, m) \perp M(t_m) \mid r(t_d, X)$ like the second ignorability assumption.

We now present a new algorithm for estimating the dose-response function which is the empirical version of the equation presented in Theorem 3.2.2.

Algorithm 3.2.3 (Estimating the Dose-Response Function with Propensity Scores)**.**

i.) Estimate the generalised propensity function $r(t, x)$ with Algorithm 3.2.1. Denote the estimate by $\widehat{r}(t, x)$.

ii.) Estimate the conditional distribution $P(M \mid r(T, X),\ T)$, with a Distributional Random Forest with inputs $\{\widehat{r}(T_i, X_i), T_i\}_{i=0}^n$ and output $\{M_i\}_{i=0}^n$. Denote the estimate by $\widehat{P}(M \mid r(T, X),\ T)$.

iii.) Train a regressor function $G(X, T, M, \widehat{r}(T, X))$ *to predict* Y*.*

iv.) Estimate $\xi(t_d, t_m)$ *by* $\widehat{\xi}(t_d, t_m) := \frac{1}{k}\frac{1}{l}\sum_{i=1}^{k}\sum_{j=1}^{l} G(\tilde{x}_i, t_d, \tilde{m}_{i,j}, \widehat{r}(t_d, \tilde{x}_i))$*, where the* $\tilde{x}_i$ *are sampled from* $X_1, ..., X_n$ *and the* $\tilde{m}_{i,j}$ *are sampled from* $\widehat{P}(M \mid r(T, X) = r(t_m, \tilde{x}_i),\ T = t_m)$ *(both with replacement).*

The major difference between this algorithm and the covariate adjustment Algorithm 3.1.1 is the training of the Distributional Random Forest. Previously it was done with $\{X_i, T_i\}_{i=0}^n$ as input, which is of dimension $p + 1$, but now only with the two dimensional input $\{r(T_i, X_i), T_i\}_{i=0}^n$. The dimensionality of the problem has not been erased, but just moved to the estimation of the generalised propensity function. This way the estimated conditional distribution $\widehat{P}_M(\cdot \mid r(T, X), T)$ is forced to consider T for branch splitting more often than before.

Another modification is that we also include the generalised propensity function, $r(t_d, X)$ as a feature in the regressor. As stated in the proof of Theorem 3.2.2, the function $r(t_d, X)$ is deterministic in X, so theoretically this should not be of any advantage, but in practice it is.

As demonstrated with simulations in Figure 4.1 in 4, this estimator performs better than the one using covariate-adjustment, successfully decreasing the negative bias of the mediated effect.

Chapter 4

Simulations

We now benchmark our proposed methods with simulation models to formally investigate how the algorithms behave in terms of their bias and variance, under various models. In Section 4.1 we will describe the simulation setup and the data generating models. In Section 4.2 we will compare the two proposed algorithms from Chapter 3. In Section 4.3 we will benchmark our estimator against a competing method in various scenarios. In section 4.4 we will simulate models with interaction effects. In Section 4.5 we will investigate the estimator's behaviour when the sequential ignorability assumptions are violated by introducing confounders. The results in this chapter were produced in Python and R.

4.1 Setup

We implemented in Python the two mediation analysis algorithms, using covariate adjustment (Algorithm 3.1.1) and propensity scores (Algorithm 3.2.3), as well as the generalised propensity function estimator (Algorithm 3.2.1).

For the simulations we will use a Distributional Random Forest with the default parameters (500 trees and a minimum node size of 5). We investigated various combinations of possibilities, but found this setting to be the optimal one for the simulations (results not shown). For example, decreasing the number of trees resulted in a larger variance as expected, but increasing the number above 500 did not improve the performance enough relative to the increased computation time.

We will use a neural network regressor from the Python package `scikit-learn` (Pedregosa et al. (2011)). We use a fully connected feed-forward network with two hidden layers. The first hidden layer has $2p$ nodes and the second has p nodes, where p is the dimension of X. We fix the activation functions throughout the network to be rectified linear units (ReLU). We will vary the L2 regularisation coefficient α within the range $\{10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}\}$ and perform a 5-fold cross validation to choose the optimal value for each model. We investigated varying the hidden layer structure and activation functions as well, but found no significant change in the model (results not shown). Thus, for simplicity and ease of computation time we fix the structure.

The Monte Carlo sampling sizes for both integrals, k and l, are set to 1000.

Our data generating mechanism works much like a structural equation model described in

Chapter 2,

$$\begin{aligned} X &:= f_X(\varepsilon_X) \in \mathbb{R}^p \\ T &:= f_T(X, \varepsilon_T) \in \mathbb{R} \\ M &:= f_M(X, T, \varepsilon_M) \in \mathbb{R} \\ Y &:= f_Y(X, T, M, \varepsilon_Y) \in \mathbb{R}. \end{aligned} \tag{4.1.1}$$

For each model, we specify the functions f_X, f_T, f_M, f_Y, the distribution of the errors $\varepsilon_X, \varepsilon_T, \varepsilon_M \varepsilon_M$, the sample size n, and the dimensionality p. As long as the data generating mechanism follows this structure and the errors are generated independently, then T and M are sequentially ignorable given Y, and so the algorithm is valid in estimating the true dose-response function.

We will estimate the dose-response function at the nine deciles of the treatment variable T while keeping the alternative treatment fixed at the median value, if not specified otherwise. For example if the deciles are $\{q_1, q_2, ..., q_9\}$ then the dose-response function is estimated at $(t_d, t_m) = (q_i, q_5)$, (q_5, q_i) and (q_i, q_i) for $i = 1, 2, ..., 9$. We obtain three dose-response curves, whose slope correspond to the direct, mediated and total causal effects, respectively.

For each model we run 100 simulations, generating new data sets in each run. From these simulations we obtained empirical estimates of the bias, variance, mean-squared-error and a 95% confidence interval of the function estimates.

It is possible to obtain other performance measurements, such as an empirical measurement of the accuracy of the slope of the effect, by keeping track of the slope of the partial derivatives in each individual simulation.

4.2 Covariate-Adjustment vs. Propensity Scores

We begin by comparing the covariate-adjustment algorithm and the propensity score algorithm. We will apply both algorithms to three data generating models: a linear model, a nonlinear model with monotone effects and a model with more complicated non-monotone effects. The exact data generating models are

Model 1.

$$\begin{aligned} X^{(i)} &:= \varepsilon_{X^{(i)}}, \ i = 1, 2, ..., p \\ T &:= \beta_1 X + \varepsilon_T \\ M &:= \tfrac{1}{2} T + \beta_2 X + \varepsilon_M \\ Y &:= 3M + T + \beta_3 X + \varepsilon_Y \end{aligned}$$

Model 2.

$$\begin{aligned} X^{(i)} &:= \varepsilon_{X^{(i)}}, \ i = 1, 2, ..., p \\ T &:= (3 + \beta_1 X + \varepsilon_T)_+ \\ M &:= 3 + \sqrt{T} + \beta_2 X + \varepsilon_M \\ Y &:= M + \tanh(T) + \beta_3 X + \varepsilon_Y \end{aligned}$$

Model 3.

$$\begin{aligned} X^{(i)} &:= \varepsilon_{X^{(i)}}, \ i = 1, 2, ..., p \\ T &:= \beta_1 X + \varepsilon_T \\ M &:= 2 + \cos(T) + \beta_2 X + \varepsilon_M \\ Y &:= 3\sqrt{|M|} - \arctan(T) + \beta_3 X + \varepsilon_Y \end{aligned}$$

Where $\beta_1, \beta_2, \beta_3$ are vectors of coefficients depending on p. The notation x_+ in Model 2 denotes $x_+ := \max(0, x)$. In all three models the errors, $\varepsilon_{X^{(i)}}, \varepsilon_T, \varepsilon_M, \varepsilon_Y$, are independently $\mathcal{N}(0, 1)$ distributed, we take a sample of size $n = 5,000$ and let $p = 4$. The results are displayed in Figure 4.1.

From Figure 4.1 we can see that the propensity score estimator performs substantially better than the covariate adjustment estimator. For all models, the bias is decreased at the cost of a slightly higher variance, always resulting in a smaller mean squared error.

The results shown are only from a limited set of models, however they are representative of the advantage of the propensity score estimator. In the following sections of this chapter we will thus focus on the propensity score estimator and investigate its behaviour under more general models.

4.3 Benchmark

In this section we will benchmark the performance of the propensity score algorithm while varying the sample size, n, the dimensionality p, and the error distribution, for the linear model, Model 1 and the non-linear model, Model 3. As a competing method, we will apply the estimator proposed by Huber et al. (2020) to the same set of models and compare the performance.

With an increased dimensionality p, we change the structural equation functions such that each of the p covariates has a causal effect on T, M or Y. For an alternative error distribution we will use the t_3 distribution, i.e. the t distribution with 3 degrees of freedom. It is a heavy tailed distribution, potentially producing values far from its mean.

The estimator proposed by Huber is available in R. We estimate the dose response function at the same set of points and evaluate its performance in the same manner as our estimator. We will use the mean squared error (MSE) of the mediated effect and the direct effect, normalised by the range of the outcome Y, as our quantities of interest while comparing the performance.

The complete results from this benchmark study can be found in table 4.1. A few chosen examples have been visualised in more depth in Figure 4.2.

The results of the benchmark simulation analysis, apparent in Table 4.1, reveal several findings. Firstly, our proposed estimator outperforms the competing one in almost all respects. Secondly, for our method we do observe a decreased performance with increased p and heavy tailed errors, though not as substantially as for the competing method. Thirdly, increasing n, the sample size, has the expected effect of reducing the mean squared error. Fourthly, our method estimates the mediated effect and direct effect roughly equally well.

In almost all model configurations, our method outperforms the competing one, especially in models with large p or heavy tailed errors. We note that Huber's estimator is only

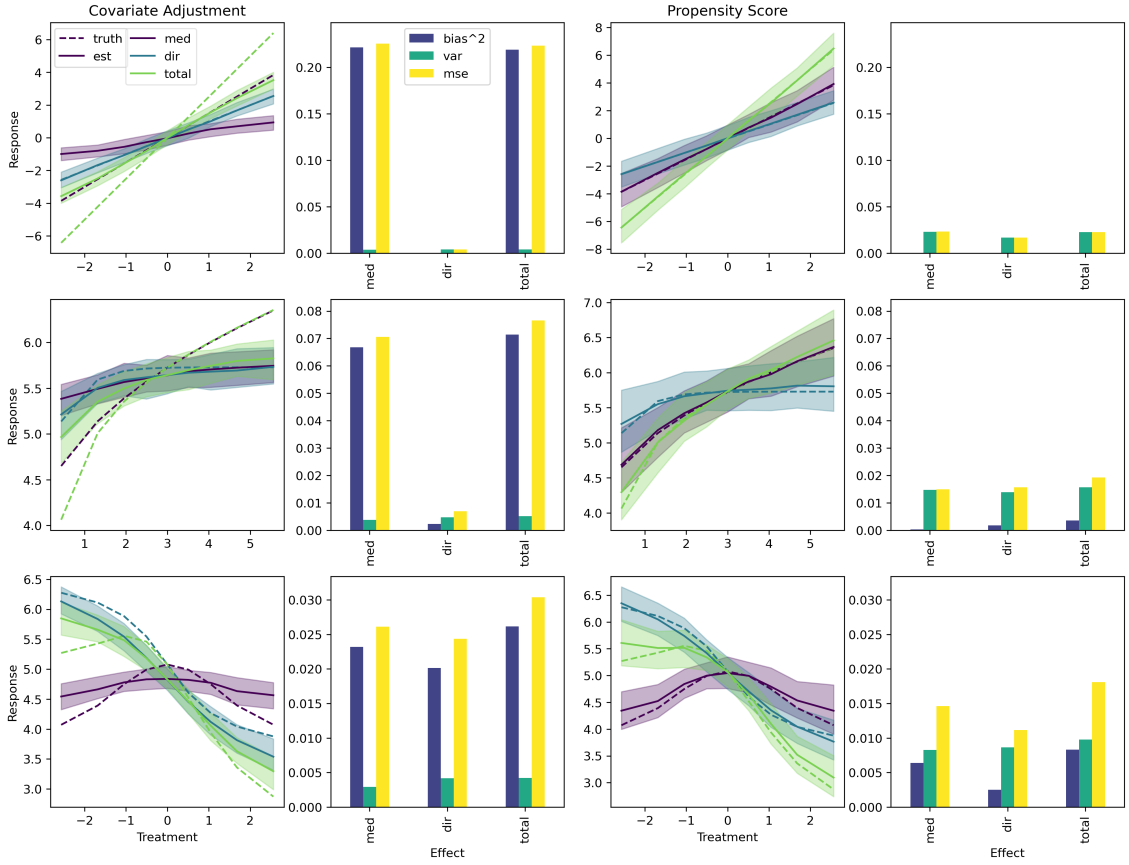

Figure 4.1: A comparison of the performance of the covariate adjustment estimator and the propensity score estimator for simulated data from three different models. The first row corresponds to Model 1, the second row to Model 2, the third row to Model 3. The first two columns correspond to the covariate adjustment estimator. The third and fourth column correspond to the propensity score estimator. The line graphs demonstrate the estimation of the dose-response function. The dashed lines show the true values, the solid lines are the average estimates from the 100 runs and the shaded areas demonstrate the empirical 95% confidence intervals. The purple line, corresponding to the mediated effect, demonstrates the response of altering t_m while t_d is fixed at the median of the treatment range, vice versa for the blue line, corresponding the direct effect. The green line, corresponding to the total effect, demonstrates the response while t_m and t_d are altered jointly, i.e. $t_m = t_d$. The bar plots visualise the quantification of the squared bias, the variance and the mean squared error, averaged over all nine points of estimation. These values are normalised by the range of the response.

semi-parametric and assumes that the error distribution is normal, therefore the drop in performance with heavy tailed errors can be expected. For the linear model with normally distributed errors, our method always outperforms by a factor of at least 3. We also note that Huber's estimator can not handle high dimensional models, i.e. where $p > n$, while ours can. The competing method performs better on estimating the nonlinear mediation effects with low n and moderate p.

The performance of our method does also decrease with increased p or heavy tailed errors. In our method, increasing p means increasing the dimensionality of the input features for

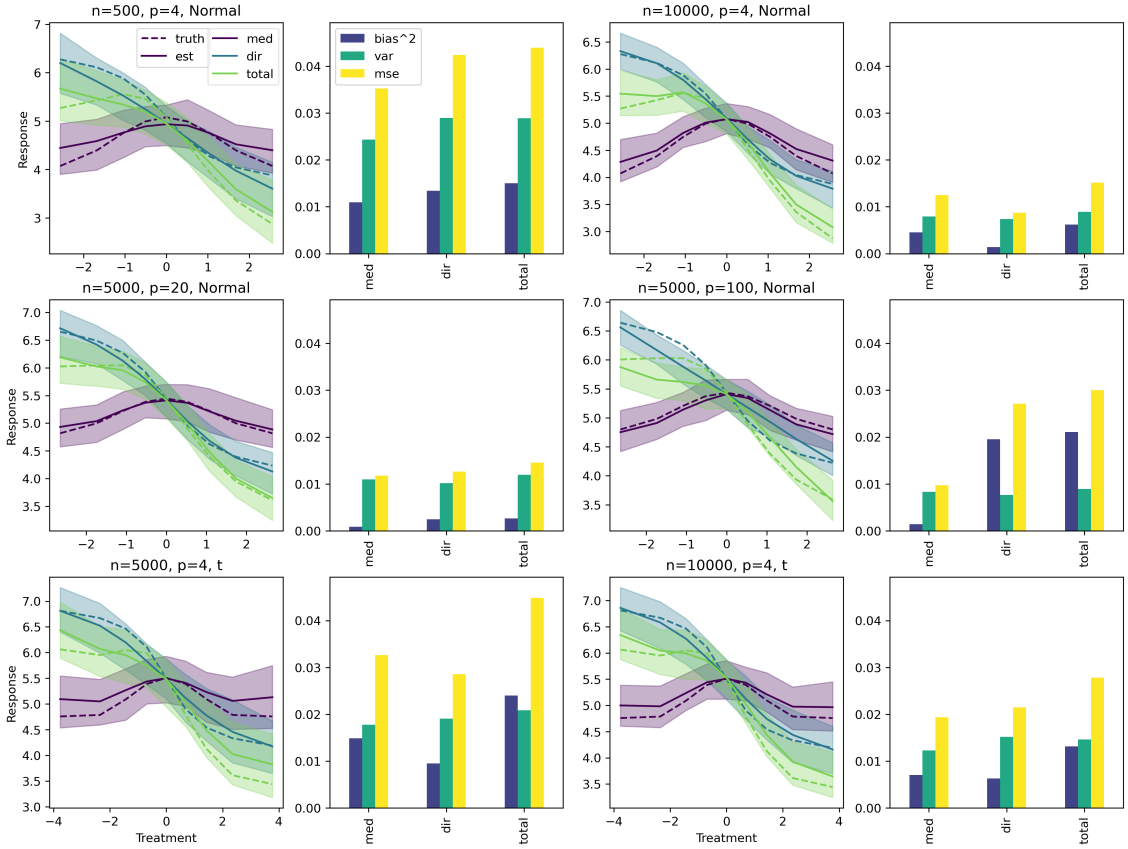

Figure 4.2: The propensity score algorithm tested against various data generating models. The graphs represent the same information as in Figure 4.1. The title above each line graph states the aspects of the model.

both the neural network and the Distributional Random Forest we use to estimate the generalised propensity score. From Figure 4.2, second and third rows, we notice that with larger p, the increased MSE is mainly due to an increased bias, and with the heavy tailed error the increased MSE can be explained by both an increased variance and bias.

The effect of altering n, the sample size, is visualised in Figure 4.2, top row. It is apparent that the MSE decreases with increased n, due to both a decreased bias and a decreased variance. Increasing n also reduces the MSE in models with large p. An example, consider the linear model with normal errors, $n = 5,000$ and $p = 20$. In this setting, the MSE for the mediated effect is 0.033. Increasing p from 20 to 200 raises the MSE to 0.042. However, when we increase the n to $n = 10,000$, the MSE decreases back down to 0.034.

From Table 4.1 we can also conclude that with our method, the MSE of the estimation of the mediated effect and the direct effect are generally on the same scale. The estimator seems to focus equally on both effects. This is not the case for the competing method, for which the MSE of the direct effect is substantially larger than of the mediated effect.

We also note that our method performs congruently for both the linear and non-linear model.

From the graphs in Figure 4.2 we can see the phenomenon discussed in Subsection 3.1.2, regarding the regressor, that the further t_d and t_m are apart, the worse the estimation

becomes.

Table 4.1: Simulation benchmark results. The number in each cell is the mean squared error for the corresponding method applied to the corresponding model, errors, effect, n and p as denoted by the columns and rows. Hierarchical columns indicate first the model type and error distribution, then mediating or direct effect and lastly the currently proposed method (DRF) or the competing method (Huber). Each row corresponds to a combination of values for n, the sample size, and p, the dimensionality of X, indicated by the two leftmost columns. An empty cell indicates that the corresponding method was unable to produce an estimation. In our case, the competing method does not support high-dimensional cases with $p > n$.

		Linear, $\mathcal{N}$-err				Linear, t_3-err				Non-linear, $\mathcal{N}$-err				Non-linear, t_3-err			
		Mediating		Direct		Mediating		Direct		Mediating		Direct		Mediating		Direct	
n	p	DRF	Huber	DRF	Huber	DRF	Huber	DRF	Huber	DRF	Huber	DRF	Huber	DRF	Huber	DRF	Huber
100	4	**.252**	.908	**.136**	3.59	**.437**	3.21	**.262**	7.69	.19	**.076**	**.263**	4.42	.397	**.0546**	**.412**	8.41
100	20	**.311**	2.5	**.15**	8.95	**.549**	5.95	**.294**	18.9	.0688	**.0551**	**.128**	2.79	.182	**.145**	**.265**	6.9
100	40	**.315**	3.05	**.127**	11.8	**.531**	7.62	**.288**	20	.0678	**.056**	**.14**	3.42	**.146**	.222	**.268**	7.26
100	100	**.358**	–	**.149**	–	**.51**	–	**.271**	–	**.0738**	–	**.197**	–	**.158**	–	**.361**	–
100	200	**.331**	–	**.145**	–	**.49**	–	**.182**	–	**.0702**	–	**.255**	–	**.14**	–	**.409**	–
500	4	**.0831**	.277	**.0595**	1.39	**.186**	30.3	**.141**	19.7	**.0353**	.0433	**.0424**	1.39	.114	**.0361**	**.0755**	6.05
500	20	**.0911**	.485	**.0599**	3.03	**.177**	16	**.0965**	16	.0302	**.0295**	**.0429**	1.11	.0965	**.0745**	**.12**	3.48
500	40	**.102**	.74	**.0638**	2.9	**.167**	10.3	**.0983**	8.95	**.0233**	.0291	**.0461**	1.04	**.0603**	.0809	**.129**	2.46
500	100	**.108**	.83	**.0652**	4.45	**.158**	9.67	**.0894**	16	**.0274**	.0309	**.0814**	1.35	**.0554**	.0879	**.165**	4.31
500	200	**.114**	1.67	**.0566**	6.85	**.164**	9.68	**.087**	12.3	**.0222**	.0324	**.099**	2.34	**.0524**	.0963	**.193**	3.42
1,000	4	**.0459**	.252	**.0319**	1	**.124**	41	**.0722**	30.2	**.0232**	.0418	**.023**	.969	.101	**.0338**	**.0502**	4.54
1,000	20	**.0596**	.359	**.0397**	2.33	**.128**	23.1	**.0766**	16.1	**.0229**	.0289	**.0308**	.64	**.0678**	.0747	**.0995**	2.21
1,000	40	**.0696**	.561	**.0392**	2.41	**.124**	30.3	**.0609**	15.2	**.0196**	.0288	**.0344**	.776	**.0461**	.0803	**.0922**	2.97
1,000	100	**.0824**	.406	**.0433**	2.46	**.125**	31.5	**.0664**	28.7	**.0205**	.0285	**.0565**	.755	**.0477**	.0857	**.143**	3.2
1,000	200	**.0814**	.71	**.0358**	3.36	**.124**	39.4	**.0662**	29.8	**.0195**	.0289	**.0786**	1.12	**.0417**	.0928	**.153**	3.37
5,000	4	**.0234**	.116	**.0169**	.659	**.0407**	120	**.0305**	78.9	**.0146**	.0411	**.0111**	.496	**.0326**	.0334	**.0286**	9.96
5,000	20	**.033**	.196	**.0249**	1.15	**.0722**	65.3	**.0603**	37.4	**.0118**	.0287	**.0126**	.305	**.0436**	.074	**.0412**	6.02
5,000	40	**.0377**	.0915	**.0261**	1.11	**.0836**	84.3	**.0505**	98	**.0104**	.0283	**.0166**	.327	**.0323**	.0793	**.059**	5.86
5,000	100	**.0416**	.191	**.0238**	1.19	**.0824**	77.4	**.0518**	48	**.00972**	.0283	**.0272**	.328	**.0349**	.0856	**.112**	4.46
5,000	200	**.0419**	.215	**.0288**	1.17	**.0835**	100	**.0514**	70.3	**.0172**	.0284	**.0437**	.348	**.0487**	.0908	**.144**	5.09
10,000	4	**.0194**	.0882	**.011**	.552	**.0307**	263	**.0233**	114	**.0125**	.041	**.00876**	.273	**.0194**	.0333	**.0215**	13
10,000	20	**.0263**	.0862	**.0221**	.708	**.0625**	103	**.0489**	83.1	**.0102**	.0287	**.0083**	.196	**.051**	.0739	**.0256**	8.02
10,000	40	**.0299**	.121	**.02**	.834	**.066**	107	**.0489**	96.8	**.0105**	.0283	**.0134**	.223	**.0359**	.0792	**.0426**	6.2
10,000	100	**.0359**	.105	**.0267**	.714	**.0736**	91.2	**.0356**	64.4	**.00798**	.0283	**.0212**	.244	**.0266**	.0855	**.0839**	5.1
10,000	200	**.0336**	.117	**.0196**	1.11	**.0729**	256	**.053**	140	**.0124**	.0284	**.0304**	.429	**.0427**	.0908	**.135**	5.21

4.4 Multivariate Mediators and Interaction Effects

We will now investigate how our method deals with having multiple mediators or interaction effects. In the latter case, we will study situations where the outcome Y depends on an interaction term involving T and M. We introduce four new data generating models.

Model 4.

$$\begin{aligned} X^{(i)} &:= \varepsilon_{X^{(i)}},\ i = 1, 2, ..., p \\ T &:= 2 + \beta_1 X + \varepsilon_T \\ M^{(1)} &:= 2 + \sqrt{T_+} + \beta_2 X + \varepsilon_{M^{(1)}} \\ M^{(2)} &:= \tfrac{1}{2} T + \beta_3 X + \varepsilon_{M^{(2)}} \\ Y &:= 3M^{(1)} + \arctan(M^{(2)}) - \cos(T) + \beta_4 X + \varepsilon_Y \end{aligned}$$

Model 5.

$$\begin{aligned} X^{(i)} &:= \varepsilon_{X^{(i)}},\ i = 1, 2, ..., p \\ T &:= 2 + \beta_1 X + \varepsilon_T \\ M^{(1)} &:= T + \beta_2 X + \varepsilon_{M^{(1)}} \\ M^{(2)} &:= -\sqrt{T_+} + \beta_3 X + \varepsilon_{M^{(2)}} \\ M^{(3)} &:= \arctan(T) + \beta_4 X + \varepsilon_{M^{(3)}} \\ Y &:= \cos(M^{(1)} + M^{(2)}) + M^{(3)} - \sqrt{T_+} + \beta_5 X + \varepsilon_Y \end{aligned}$$

These models include a two and three dimensional mediator, respectively. Both mediators fulfill the sequential ignorability assumption.

Model 6.

$$\begin{aligned} X^{(i)} &:= \varepsilon_{X^{(i)}},\ i = 1, 2, ..., p \\ T &:= 2 + \beta_1 X + \varepsilon_T \\ M &:= T - \beta_2 X + \varepsilon_M \\ Y &:= M \cdot \arctan(T) + \beta_3 X + \varepsilon_Y \end{aligned}$$

Model 7.

$$\begin{aligned} X^{(i)} &:= \varepsilon_{X^{(i)}},\ i = 1, 2, ..., p \\ T &:= 2 + \beta_1 X + \varepsilon_T \\ M &:= 2 + \sqrt{T_+} + \beta_2 X + \varepsilon_M \\ Y &:= M \cdot \cos(T) + \beta_3 X + \varepsilon_Y \end{aligned}$$

In these data generating models, the strength of the mediated effect depends on the value of t_d, and the strength of the direct effect depends on the value of t_m. To demonstrate this we perform a simulation as before but instead of fixing the base treatment value at the median of the treatment range, we fix it at the third and seventh deciles. That is, if the deciles of the treatment variable are $\{q_1, q_2, ..., q_9\}$ then we estimated the dose-response function at the values $(t_d, t_m) = (q_i, q_3), (q_i, q_7), (q_3, q_i), (q_7, q_i)$ for $i = 1, 2, ..., 9$.

With these four new data generating models, we benchmarked our estimator against the competing method proposed by Huber et al. We fix the error distribution to be $\mathcal{N}(0, 1)$,

but vary the sample size and dimension of X. Full results can be found in Table 4.2. Selective examples were visualised in Figures 4.3 and 4.4.

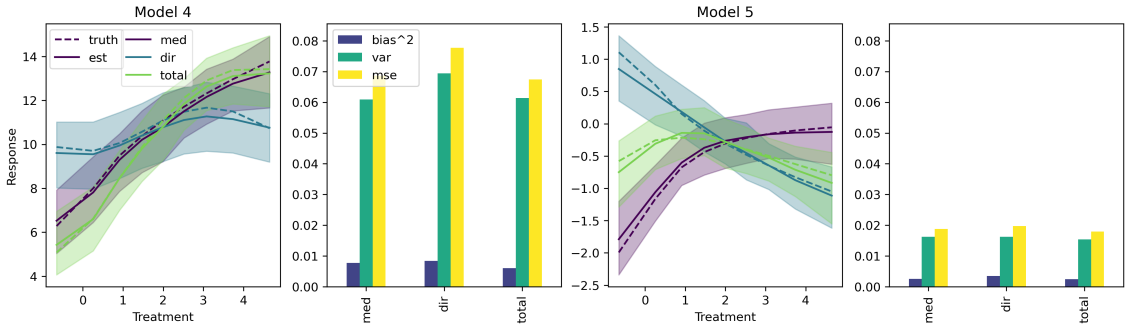

Figure 4.3: Simulation results of models with multidimensional mediators. Left two columns corresponds to Model 4, right two columns to Model 5, both with $n = 5000$ and $p = 20$.

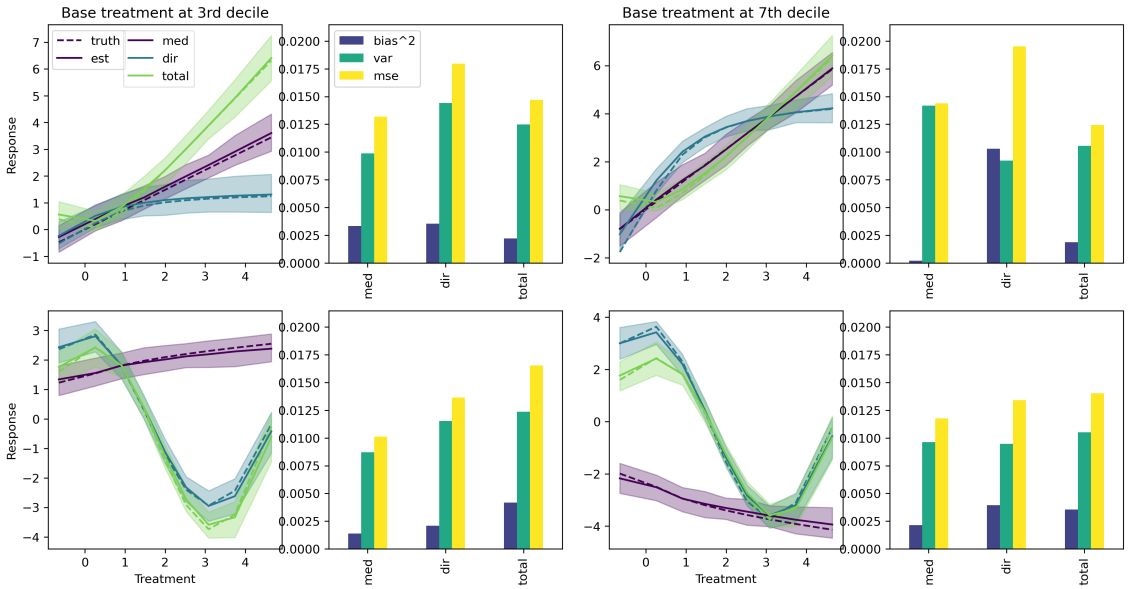

Figure 4.4: Simulation results of models with interaction effects between T and M. Top row corresponds to Model 6, bottom row to Model 7, both with $n = 5000$ and $p = 4$. On the left the alternative treatment was fixed at the third decile, on the right it is fixed at the seventh decile.

From the simulation results in Table 4.2, we find several conclusions. Firstly, our estimator seems to handle these two multidimensional mediator models as well as these two interaction models quite well. Secondly, we again outperform the competing method on almost all settings. Thirdly, our method successfully captures the change in effect curves depending on interaction effects.

For all of the four new models, the MSE is roughly on the same scale as for the previous models that were tested in Table 4.1. From Figures 4.3 and 4.4 we can see that the bias remains relatively low for all models. We only test here the presence of a two and three dimensional mediator. A more thorough analysis is needed to investigate the behaviour when the dimension of M increases on a greater scale.

Comparing to the competing method, our estimator performs overall better on all models. The difference is most substantial in the interaction models, for example in Model 7 with $n = 10,000$ and $p = 100$ the competing method has an MSE of 0.473 while ours is at 0.00478.

From Figure 4.4, where we plot the estimated function values for two different base treatment values, we see that we successfully estimate the change in the effect curves depending on the base treatment value.

Table 4.2: Simulation benchmark results for models 4,5,6 and 7. The number in each cell is the mean squared error for the corresponding method applied to the corresponding model, errors, effect, n and p as denoted by the columns and rows. Hierarchical columns indicate first the model, then mediating or direct effect and lastly the currently proposed method (DRF) or the competing method (Huber). Each row corresponds to a combination of values for n, the sample size, and p, the dimensionality of X, indicated by the two leftmost columns. An empty cell indicates that the corresponding method was unable to produce an estimation. In our case, the competing method does not support high-dimensional cases with $p > n$.

		Model 4				Model 5				Model 6				Model 7			
		Mediating		Direct		Mediating		Direct		Mediating		Direct		Mediating		Direct	
n	p	DRF	Huber	DRF	Huber	DRF	Huber	DRF	Huber	DRF	Huber	DRF	Huber	DRF	Huber	DRF	Huber
100	4	**1.6**	2.94	**1.14**	10.2	**.191**	1.5	**.274**	2.68	**.336**	1.7	**.178**	3.59	**.113**	.567	**.32**	3.31
100	20	**.446**	4.9	**.32**	13.6	**.113**	1.64	**.146**	2.46	**.189**	2.27	**.173**	3.66	**.0851**	.611	**.579**	3.88
100	40	**.394**	5.72	**.213**	17.6	**.151**	1.79	**.168**	2.95	**.209**	2.6	**.208**	4.87	**.132**	.687	**.649**	4.4
100	100	**.438**	–	**.266**	–	**.153**	–	**.188**	–	**.262**	–	**.272**	–	**.167**	–	**.729**	–
100	200	**.556**	–	**.347**	–	**.179**	–	**.201**	–	**.3**	–	**.297**	–	**.151**	–	**.789**	–
500	4	**1.01**	1.15	**.735**	4.46	**.0757**	.989	**.108**	1.45	**.0909**	.841	**.0646**	1.47	**.0441**	.629	**.0897**	2.21
500	20	**.182**	1.48	**.131**	5.52	**.0409**	.55	**.0508**	.783	**.038**	.774	**.0868**	1.56	**.0217**	.647	**.154**	1.91
500	40	**.175**	1.65	**.149**	6.14	**.0385**	.661	**.0501**	1.26	**.0429**	.88	**.12**	1.39	**.0518**	.64	**.51**	1.83
500	100	**.167**	2.24	**.103**	7	**.0612**	.821	**.0987**	1.21	**.0647**	1.24	**.155**	2.2	**.107**	.663	**.63**	2.39
500	200	**.166**	4.04	**.0988**	12.2	**.101**	.981	**.128**	1.78	**.103**	1.38	**.222**	2.48	**.131**	.944	**.675**	3.08
1,000	4	**.454**	1.05	**.249**	4.2	**.0535**	.683	**.0831**	1.04	**.0501**	.498	**.0438**	1.2	**.0443**	.541	**.0596**	1.3
1,000	20	**.121**	.848	**.0974**	4.27	**.0445**	.42	**.0548**	.731	**.0239**	.559	**.0643**	1.2	**.0221**	.578	**.0538**	1.19
1,000	40	**.132**	1.11	**.123**	3.58	**.0241**	.412	**.0322**	.92	**.029**	.653	**.102**	1.16	**.0263**	.613	**.309**	1.3
1,000	100	**.137**	1.51	**.0923**	5.35	**.0366**	.438	**.0598**	.662	**.0454**	.618	**.124**	1.21	**.078**	.585	**.597**	1.33
1,000	200	**.131**	1.51	**.0906**	5.89	**.064**	.577	**.0939**	.802	**.0657**	.914	**.176**	1.43	**.11**	.673	**.648**	1.74
5,000	4	**.12**	.584	**.0523**	2.38	**.0341**	.45	**.0518**	.545	**.0182**	.246	**.02**	.508	**.0246**	.517	**.0299**	.528
5,000	20	**.0686**	.333	**.0778**	1.65	**.0188**	.301	**.0198**	.394	**.0128**	.338	**.016**	.556	**.00941**	.488	**.013**	.331
5,000	40	**.0795**	.446	**.0648**	1.9	**.0195**	.246	**.0214**	.343	**.0146**	.23	**.0322**	.525	**.00931**	.491	**.0196**	.342
5,000	100	**.0959**	.433	**.0786**	2.01	**.0149**	.203	**.0172**	.315	**.0174**	.327	**.0847**	.53	**.00882**	.485	**.223**	.429
5,000	200	**.0882**	.498	**.0741**	2.36	**.0195**	.187	**.0294**	.316	**.0238**	.313	**.111**	.617	**.047**	.506	.591	**.41**
10,000	4	**.0853**	.355	**.039**	1.32	**.0285**	.392	**.0468**	.552	**.014**	.22	**.0136**	.461	**.0191**	.491	**.0214**	.33
10,000	20	**.0557**	.278	**.0507**	1.46	**.0126**	.153	**.0123**	.271	**.0102**	.194	**.0109**	.353	**.00754**	.484	**.00988**	.187
10,000	40	**.0569**	.346	**.0597**	1.63	**.0191**	.146	**.0175**	.272	**.0112**	.197	**.0151**	.351	**.00753**	.476	**.00821**	.283
10,000	100	**.0889**	.303	**.0843**	1.44	**.014**	.142	**.0161**	.239	**.013**	.204	**.0541**	.394	**.00478**	.473	**.0318**	.196
10,000	200	**.074**	.339	**.076**	1.44	**.0159**	.187	**.0232**	.254	**.0187**	.213	**.093**	.482	**.017**	.48	**.276**	.321

4.5 Hidden Confounders

We now turn to a different class of simulation models, in which the sequential ignorability assumption is not satisfied. We are interested in exploring to what degree the assumption can be violated while still retaining a meaningful estimation. We will do so by introducing hidden confounders into the simulation model. Figure 4.5 visualises the four different confounders we will consider.

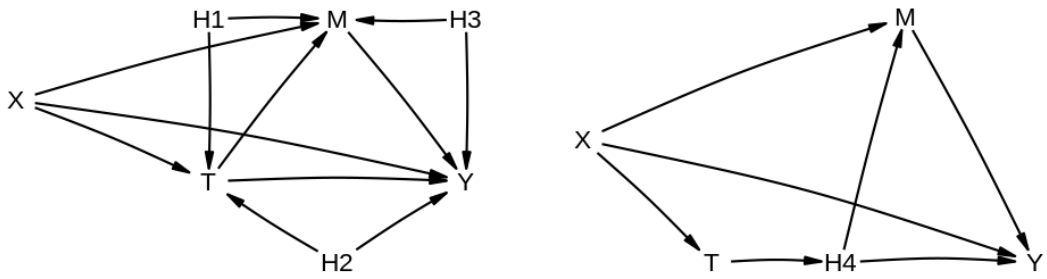

Figure 4.5: DAGs showcasing the various confounders we will consider in our simulation studies. The DAG on the left includes three possible pre-treatment confounders. The DAG on the right includes a post-treatment confounder.

We will use Model 3 as a base model to which we will add confounders. For all simulations we will set $n = 5,000$, $p = 4$ and have $\mathcal{N}(0,1)$ distributed errors for the observed (non-confounder) variables. In each data generating model, we will include one of the four confounders present in Figure 4.5.

The confounders H_1 and H_2 violate the first condition in the sequential ignorability assumption in Definition 2.2.3. In this case the potential outcomes $\{M(t_m), Y(t_d, m)\}$ and the treatment T are no longer conditionally independent given X. The confounders H_3 and H_4 violate the second condition. In this case the potential outcomes $M(t_m)$ and $Y(t_d, m)$ are no longer conditionally independent given T and X.

The confounders H_1, H_2 and H_3 are called pre-treatment confounders as they are not caused by the treatment. They will enter the data generating model linearly. The confounder H_4 is called a post-treatment confounder as it is causally affected by the treatment variable. In the data generating models it is defined as $H_4 := T + \gamma\varepsilon_{H_4}$ where ε_{H_4} is a $\mathcal{N}(0,1)$ distributed random variable and γ is a parameter controlling the strength of the confounding effect, and then H_4 replaces T in the generation of M and Y. Therefore, in a linear model the post-treatment confounder H_4 is essentially the same as H_3. Note that in the training of the forest and regression models, we still use the variable T as input, treating H_4 as an unobserved variable.

We will investigate how the estimator responds to increasing strengths of the confounding effects. The strength of the confounding effects is determined by a parameter γ. For the pre-treatment confounding variables, H_1, H_2, H_3, the parameter γ represents the standard deviation. For the post-treatment confounder, H_4, the parameter γ enters the model as stated above.

The strength of the confounding effect is always relative to the variables it affects. In order to give an idea of the relative strength of the confounding effects in our simulations,

Table 4.3: Simulation results from models with hidden confounders. The number in each cell is the mean squared error in the presence of the corresponding confounder, as denoted by the columns, with the corresponding standard deviation, as denoted by the rows.

	T-M		T-Y		M-Y		Post-treat	
$\gamma_{rel}(\%)$	Med	Dir	Med	Dir	Med	Dir	Med	Dir
0	.0146	.0111	.0146	.0111	.0146	.0111	.0146	.0111
10	.0159	.0104	.014	.0163	.00948	.012	.0158	.0126
20	.0203	.00974	.0146	.0764	.0134	.0243	.0177	.0218
30	.0372	.0115	.0139	.197	.029	.0493	.0232	.0382
40	.063	.00958	.0173	.374	.0571	.0814	.0309	.0622
50	.0934	.0105	.0175	.579	.0881	.112	.0382	.0774

we present the value

$$\gamma_{rel} := 100 \cdot \frac{\gamma}{\max\{\text{std}(T), \text{std}(M), \text{std}(Y)\}}.$$

Where the standard deviations are calculated in absence of any confounding. We present the values as percentages. We divide by the maximum standard deviation instead of the minimum to obtain conservative estimates of robustness, as the resulting γ_{rel} is a lower bound of the relative level of noise. We note that the standard deviations of the three variables are around $\text{std}(T) = 2.0$, $\text{std}(M) = 2.7$ and $\text{std}(Y) = 3.2$.

The results of this analysis is laid out in Table 4.3. The results for $\gamma_{rel} = 30\%$ are visualised in Figure 4.6.

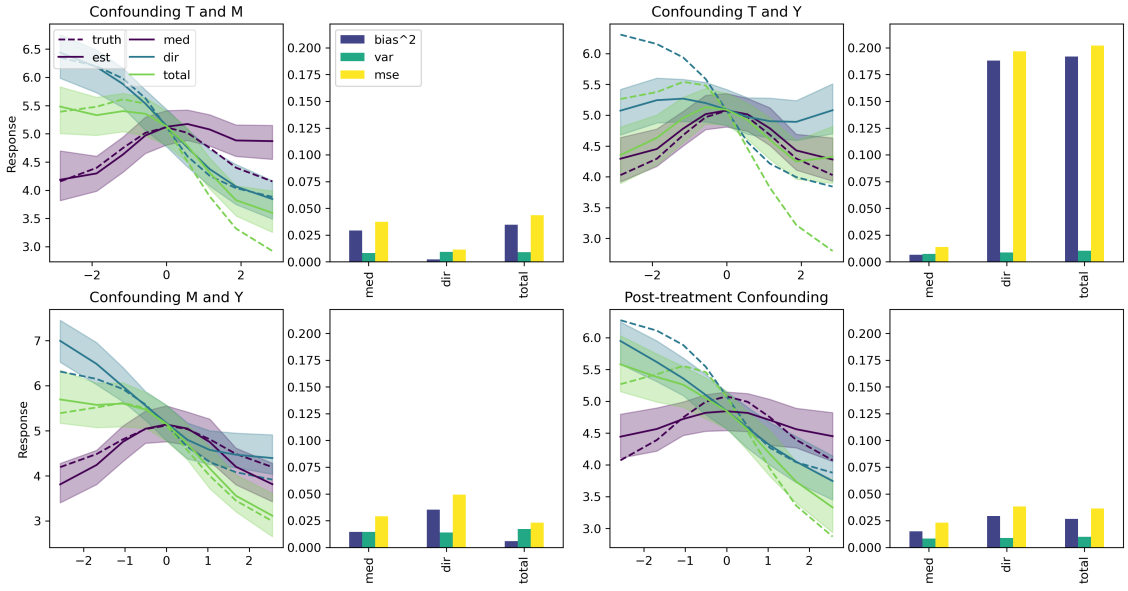

Figure 4.6: Simulation results of models with hidden confounders which violate the sequential ignorability assumptions. The title above each graph indicates which kind of confounder was introduced in the corresponding model.

From the results demonstrated in Table 4.3 and Figure 4.6, we can deduce that a $T - M$ confounder does distort the estimation of the mediated effect while it does not seem to

affect the direct effect in this scenario. Meanwhile, the $T - Y$ confounder distorts the direct affect but only induces a slight underestimation of the mediated effect. The $M - Y$ confounder distorts both effects.

This is an expected result if we consider the equation in Theorem 2.2.5 with the DAGs in Figure 4.5. Even though H_1 does also confound the relationship between T and Y, when we estimate the direct effect of T on Y we fix the value of M, effectively blocking the path from H_1 to Y. Therefore the estimated direct effect is not affected in the presence of a $T - M$ confounder like H_1.

The variable H_2 does not confound the $T - M$ relationship. It does confound the $M - Y$ relationship, however when we estimate the causal effect of M on Y as part of the mediated effect, we fix the value of T, effectively blocking the path from H_2 to M. Therefore the estimated mediated effect is not affected in the presence of a $T - Y$ confounder like H_2.

The confounder H_3, confounds the $M - Y$ relationship and thus affects the estimation of the mediated effect. It is not a confounder for the $T - Y$ relationship, but by fixing the value of M in the estimation of the direct effect, we open up the path from H_3 to T and thus H_3 also confounds the estimation of the direct effect.

In the presence of the post-treatment confounder H_4, both effect curves are a bit flattened out, indicating a weaker effect from the treatment. This is also an expected behaviour as the effect of the treatment effect is weakened because it does not fully explain the variability in H_4 which ultimately affects M and Y. The added unmeasured noise in H_4 confounds the relationship between M and Y.

We notice from Table 4.3 that, as expected, the stronger the confounding effect is, the more distorted the estimates of the causal effects become. With $\gamma_{rel} = 10\%$ the performance is still congruent with no hidden confounders ($\gamma_{rel} = 0$). With $\gamma_{rel} = 20\%$, the error starts to increase and for $\gamma_{rel} = 30\%$, which is shown in Figure 4.6, the effects are quite distorted, and for a strong confounding effect, $\gamma_{rel} = 40 - 50\%$, the errors rise drastically. In the presence of a $T - M$, M_Y or a post-treatmeant confounder, the errors rise about 7-10 times that of having no hidden confounder. In the presence of a T_Y confounder, the error for the direct causal effect is 50 times that of having no hidden confounder.

These results show that the estimator is able to retain its performance in the presence of weak hidden confounders, where with weak we mean that the standard deviation of the confounding effect is about 20% or less of the standard deviation of the observed variables. We also conclude that even with certain strong confounders present, some causal effects can still be estimated accurately. In the presence of a $T - M$ confounder the direct effect estimate retains its accuracy, and in the presence of a $T - Y$ confounder the mediated effect estimate retains its accuracy. We did not investigate the behaviour in the presence of multiple simultaneous confounders. This could be done in future work.

Chapter 5

Conclusion

In this thesis we have combined the theory of causal mediation analysis with a novel statistical learning method to create a fully nonparametric estimator with wide generality. We provided and proved a new identification theorem which incorporates the generalised propensity score and demonstrated its advantages over the analogous theorem with covariate adjustment, as proposed by Imai et al. (2010).

We investigated the performance of our proposed estimator with extensive simulation studies. We showcased its superior qualities compared to its closest competitor in terms of generality. Our method performed well in linear and non-linear models alike, in higher dimensional settings, in cases of non-normal heavy tailed errors, including multidimensional mediators, and in the presence of interaction effects. These qualities arise from the statistical learning methods applied within the estimator. Neither the distributional random forest or the neural network have any parametric restraints hindering the performance in these scenarios.

We found that our estimator does breaks down in the presence of relatively strong confounding effects which violate the assumptions we rely on for the identification of the dose-response function. However if we wish to compute only the mediating effect then the presence of a hidden confounder between the treatment and outcome does not distort our estimate. Similarly, if only the direct effect is of interest then the presence of a hidden confounder between the treatment and the mediator should not concern us. Unfortunately, a confounder affecting the mediator and outcome distorts our estimates of both effects. In application areas of mediation analysis, such a confounder is perhaps the most difficult to address.

A logical next step on the theoretical side of this work would be to develop a sensitivity analysis for this nonparametric estimator. That could improve one's confidence in the results obtained in the cases where the presence of a confounder is ambiguous. Another way forward would be to investigate alternative ways to incorporate the generalised propensity score, for example a propensity of the mediating variable could be employed in the regressor. On the methodological side, the next steps would be to improve the current implementation regarding time complexity and memory usage.

The Python implementation of the estimator as well as all the code used to generate the results in this thesis, written in R and Python, is available on Github (`github.com/AlexanderGud/mediation_drf_gps`).

In conclusion, we present a new nonparametric estimation strategy for causal mediation analysis, which, given the required ignorability assumptions, has wide generality in terms of the structure of the data and the underlying true model.

Bibliography

Baron, R. and D. Kenny (1986, 01). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology 51*, 1173–1182.

Breiman, L. (2001, Oct). Random forests. *Machine Learning 45*(1), 5–32.

Bullock, J., D. Green, and S. Ha (2010, April). Yes, but what's the mechanism? (don't expect an easy answer). *Journal of Personality and Social Psychology 98*(4), 550–558.

Cybenko, G. (1989, Dec). Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems 2*(4), 303–314.

Gretton, A., K. Borgwardt, M. Rasch, B. Schölkopf, and A. Smola (2006). A kernel method for the two-sample-problem. In B. Schölkopf, J. Platt, and T. Hoffman (Eds.), *Advances in Neural Information Processing Systems*, Volume 19. MIT Press.

Hastie, T., R. Tibshirani, and J. Friedman (2009). *The Elements of Statistical Learning, Second Edition.* Springer Series in Statistics. New York, NY, USA: Springer New York Inc.

Hirano, K. and G. W. Imbens (2004). *The Propensity Score with Continuous Treatments*, Chapter 7, pp. 73–84. John Wiley & Sons, Ltd.

Huber, M., Y.-C. Hsu, Y.-Y. Lee, and L. Lettry (2020). Direct and indirect effects of continuous treatments based on generalized propensity score weighting. *Journal of Applied Econometrics 35*(7), 814–840.

Hung, R. J., J. D. McKay, V. Gaborieau, P. Boffetta, M. Hashibe, D. Zaridze, A. Mukeria, N. Szeszenia-Dabrowska, J. Lissowska, P. Rudnai, et al. (2008). A susceptibility locus for lung cancer maps to nicotinic acetylcholine receptor subunit genes on 15q25. *Nature 452*(7187), 633–637.

Imai, K., L. Keele, and T. Yamamoto (2010). Identification, inference and sensitivity analysis for causal mediation effects. *Statistical Science 25*(1), 51–71.

Liu, J. Z., F. Tozzi, D. M. Waterworth, S. G. Pillai, P. Muglia, L. Middleton, W. Berrettini, C. W. Knouff, X. Yuan, G. Waeber, et al. (2010). Meta-analysis and imputation refines the association of 15q25 with smoking quantity. *Nature genetics 42*(5), 436–440.

MacKinnon, D. P., C. M. Lockwood, J. M. Hoffman, S. G. West, and V. Sheets (2002, Mar). A comparison of methods to test mediation and other intervening variable effects. *Psychological methods 7*(1), 83–104. 11928892[pmid].

Messerli, F. H. (2012). Chocolate consumption, cognitive function, and nobel laureates. *New England Journal of Medicine 367*(16), 1562–1564. PMID: 23050509.

Oehlert, G. W. (2010). A first course in design and analysis of experiments. Retrieved from the University of Minnesota Digital Conservancy, https://hdl.handle.net/11299/168002.

Pearl, J. (2001). Direct and indirect effects. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, UAI'01, San Francisco, CA, USA, pp. 411–420. Morgan Kaufmann Publishers Inc.

Pearl, J. (2009). Causal inference in statistics: An overview. *Statistics Surveys 3*(none), 96 – 146.

Pearl, J. (2014, 06). Interpretation and identification of causal mediation. *Psychological methods 19*.

Pedregosa, F. et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research 12*, 2825–2830.

Peters, J., D. Janzing, and B. Schlkopf (2017). *Elements of Causal Inference: Foundations and Learning Algorithms*. The MIT Press.

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

Rubin, D. B. (2005). Causal inference using potential outcomes. *Journal of the American Statistical Association 100*(469), 322–331.

Tingley, D., T. Yamamoto, K. Hirose, L. Keele, and K. Imai (2014). mediation: R package for causal mediation analysis. *Journal of Statistical Software 59*(5), 1–38.

VanderWeele, T. J. et al. (2012, 02). Genetic Variants on 15q25.1, Smoking, and Lung Cancer: An Assessment of Mediation and Interaction. *American Journal of Epidemiology 175*(10), 1013–1020.

VanderWeele, T. J. and S. Vansteelandt (2014, Jan). Mediation analysis with multiple mediators. *Epidemiologic methods 2*(1), 95–115. 25580377[pmid].

Ćevid, D., L. Michel, J. Näf, N. Meinshausen, and P. Bühlmann (2020). Distributional random forests: Heterogeneity adjustment and multivariate distributional regression.

Declaration of Originality

The signed declaration of originality is a component of every semester paper, Bachelor's thesis, Master's thesis and any other degree paper undertaken during the course of studies, including the respective electronic versions.

Lecturers may also require a declaration of originality for other written papers compiled for their courses.

I hereby confirm that I am the sole author of the written work here enclosed and that I have compiled it in my own words. Parts excepted are corrections of form and content by the supervisor .

Title of work (in block letters):

NON PARAMETRIC CAUSAL MEDIATION ANALYSIS WITH DISTRIBUTIONAL RANDOM FORESTS AND GENERALISED PROPENSITY SCORES

Authored by (in block letters):

For papers written by groups the names of all authors are required.

Name(s):	**First name(s):**
Gudjónsson	Alexander

With my signature I confirm that

- I have committed none of the forms of plagiarism described in the Citation etiquette information sheet.
- I have documented all methods, data and processes truthfully.
- I have not manipulated any data.
- I have mentioned all persons who were significant facilitators of the work .
- I am aware that the work may be screened electronically for plagiarism.
- I have understood and followed the guidelines in the document *Scientific Works in Mathematics.*

Place, date:	**Signature(s):**
28.6.22	Alexander Gj

For papers written by groups the names of all authors are required. Their signatures collectively guarantee the entire content of the written paper.