A Systematic Review of Machine Learning Estimators for Causal Effects

Bachelor's Thesis
by

Maximilian Luca Franz

Chair of Pervasive Computing Systems/TECO
Institute of Telematics
Head: Prof. Dr.-Ing. Michael Beigl

First Reviewer:	Prof. Dr. Michael Beigl
Supervisors	Dr. Florian Wilhelm (inovex GmbH) Ployplearn Ravivanpong, M.Sc(TECO)

Project Period: 21/06/2019 – 21/10/2019

Erklärung

Hiermit erkläre ich, dass ich die vorliegende Bachelor selbstständig verfasst und keine anderen als die angegebenen Hilfsmittel und Quellen benutzt habe, die wörtlich oder inhaltlich übernommenen Stellen als solche kenntlich gemacht und weiterhin die Richtlinien des KIT zur Sicherung guter wissenschaftlicher Praxis beachtet habe.

Karlsruhe, den 16. Oktober 2019

Acknowledgements

I want to express my uttermost gratitude for my second supervisor, Dr. Florian Wilhelm, for providing both practical and theoretical support. He initially sparked my interest in the field of causal inference and guided me all along by asking the right, challenging questions. Also, I want to thank inovex GmbH for supporting my work, granting me access to their GPU-Cluster and providing me with a productive space, both physically and mentally, to work in.

I would like to thank my university supervisor Ployplearn Ravivanpon for allowing me to work on the topic of my choice in a self-determining manner. Further, I want to thank the chair and all members of TECO for their insightful questions and discussions of the topic.

Last but not least, I would like to thank all the researcher who took the time to answer my questions. Among them Stefan Wager, Jinsung Yoon, Christos Louizos, Michael Knaus, Martin Huber, Scott Powers, Vincent Dorie and Fredrik Johansson and many others. Without their contribution the following thesis would not have been possible.

Zusammenfassung

Treatment Effect Estimation (TEE) ist nicht nur für Medizin und Politikwissenschaften von großer Wichtigkeit, sondern erhält auch in Big Data Anwendungen immer größeren Zuspruch. Der sogenannte Treatment Effect ist definiert als die Differenz zwischen dem Szenario, in dem eine Behandlung gegeben wurde und dem, in dem keine Intervention erfolgt. Foglich ist diese Größe kontrafaktisch. Das heißt, für eine spezifische Instanz, zum Beispiel einen Patienten, können wir niemals beide Fälle messen. In dieser Thesis stellen wir die zwei mathematischen Modelle vor, die uns helfen, dieses zugrundeliegende Problem mit statistischen Methoden zu erfassen. Wir nutzen diese Theorien, um eine Vielzahl möglicher Lösungen vorzustellen und diese zu kategorisieren.

Zuerst anaylsieren wir umfangreich, wie TEE-Methoden evaluiert werden und werden können, was aufgrund mangelnder kontrafaktischer Realdaten nicht trivial ist. Im Grunde muss für eine Evaluation auf synthetische Daten zurückgegriffen werden, deren Erzeugung großen Einfluss auf die Ergebnisse des Vergleichs hat. Wir stellen gängige Referenzdatensätze vor und zeigen wie diese oftmals mit mangelnder Vorsicht verwendet werden. Aufgrund dieser mangelnden Klarheit bei der Diskussion von Ergebnisse stellen wir die ersten Züge einer ausführlichen Taxonomie für Datengenerierungsprozesse (DGP) vor.

Desweiteren nutzen wir unsere eigenen parametrischen DGPs, um eine zentrale Hypothese zu untermauern: Die Evaluation von TEE-Methoden ist stark abhängig von den Parametern der zugrundeliegenden Daten. Somit ist eine allgemeingültige Behauptung bezüglich der Qualität von Methoden nicht möglich. Wir zeigen allerdings, dass die Erkenntnisse aus der zugrundeliegenden Theorie helfen können, um die richtige Methode für einen Anwendungsfall zu wählen.

Abstract

The task of treatment effect estimation (TEE) is of great importance not only in medical and social sciences, where policy decisions are concerned, but also in big data applications. The Treatment Effect is defined as the difference in some outcome between two cases: one where some abstract treatment was given and one where it was not. Thus, TEE is a counterfactual task, because we only ever observe one of the two possible outcomes for any unit. In this thesis, we introduce the two mathematical frameworks that help us to formalise the underlying problem and we use them to offer multiple solutions to the problem.

Our contribution is then twofold. First, we analyse how treatment effect estimation methods can be evaluated, which is a non-trivial task because of missing counterfactual data. Essentially, synthetically generated data is needed. We present various reference datasets and uncover the inexpertly use of them by other researchers, which leads us to propose a novel and clear taxonomy to describe the necessary data generating processes (DGPs).

Second, we use our own parametric DGPs to validate one central hypothesis: evaluation of TEE methods is highly dependent on the setting of the underlying DGP. Hence, a general claim of *the best method*, is not possible. However, we show that the theory, on which the methods are based, helps to get an intuition of what method to use where.

Contents

1. Introduction

He who confronts the paradoxical
exposes himself to reality.

Friedrich Dürrenmatt

1.1 Motivation

Causal Effect Estimation is an essential task in many disciplines, ranging from medicine over political sciences to artificial intelligence [63, 34, 5, 31]. The human mind is apt to think in terms of causation and its reasoning about the world is similar to causal models [21]. These models allows us to ask questions like: *What would my headache feel like had I not taken the aspirin?* Computers, on the other hand, using the traditional framework of statistics, only consider correlation. In order for machines to exceed the concepts of correlations, causal frameworks are needed.

Since about 30 years there has been a growing interest in causal analysis. First, social and political sciences came on board to give their studies a firm mathematical ground. For what is science concerned with if not the causal laws that govern the world? In recent years, also the machine learning community seemed to grasp the enormous potential of causal analysis. Treatment effect estimation (TEE), as a causal discipline, is applied not only where the name implies it usage (such as the analysis of drugs), but on a wide set of problems. Since the research in this field is spread across many disjoint fields like bio-medicine and econometrics, a comprehensive review of methods is lacking.

With this motivation in mind, we set out to study and compare a multitude of methods. The focus shifted soon, however, because we encountered problems in the way evaluations are done. To simplify the comparison of TEE methods for ourselves as well as for other authors, we had to design a evaluation framework that allows a just consideration of each method in varying settings.

What was meant to be a straightforward method survey turned out to be a survey of data generating processes and evaluation procedures. What was meant to be a systematic review of estimators turned out to be a systematic review of how estimators are evaluated. This, however, is not to be perceived as failing the objective. Rather, elaborating on the latter is essentially required to be able to consider the former. We must ensure the quality of our tools, before we can get to work.

1.2 What is Causation?

Before we start with the study of causal effect estimation methods, we want to use the remainder of this chapter to discuss the conceptual distinction between causation and correlation. We do this to allow the lay reader to grasp the importance of causal methods and to remove misunderstandings of the term *Causal Inference.* At the very start we should shortly consider a philosophical question that underlies all of the following work: What is causation?

While it is neither in the scope nor in the interest of this work to elaborate on the historical debate around causation, it is essential to note that there were and still are different conceptions of causation. David Hume was the first to formulate what would become the counterfactual theory of causation when he said, *"We may define a cause to be an object followed by another, and where all the objects, similar to the first, are followed by objects similar to the second. Or, in other words, where, if the first object had not been, the second never had existed."* [27]. Note however, that the first part of the quote fundamentally differs from the second, because the first formulation is vulnerable to spurious correlation. That is to say, only because event A is and always has been followed by event B, does not mean that B is caused by A. It might be, with decreasing probability, that the two just *'happened'* to occur in this order whenever we observed them. The second, counterfactual formulation, however, plays on a wholly different field. It basically argues about causation using a world that does not exist. A world that is counter-factual. David Lewis [1, Chapter 30] was the first to formalize this counterfactual theory on which most of the following work will be based. The concept of counterfactuals is important to the theoretic framework of Judea Pearl [43] as well as it is the fundamental groundwork underlying the potential outcomes framework associated with Neyman and Rubin [49].

In practical fields of study there exist various approaches to bridging the gap from mere association to causality by defining what makes something causal. Essentially, like the Bradford Hill Criteria [38] in Epidemiology, these definitions argue about when an observed association can be considered causal. The counterfactual definition we use here, however, does not rely on such arguments. It is mathematically and philosophically clean, while being practically impenetrable. Throughout this introduction we will elaborate on the essential difficulty introduced by the counterfactual definition and draw a clearer picture as to why causal inference is both important and difficult.

1.3 Causal Inference in words

The aim of a disciplined pursuit of causality is to formalise the language around causes. The broad field of causal inference can be subdivided into *Causal Discovery* and *Causal Effect Estimation. Causal Discovery* is concerned with the retrieval of causal structure from data we collect from the world. The structured pursuit of science is strongly related to *Causal Discovery*, for it amounts to understanding the causal relationsships between things. For instance, if we are trying to determine whether there is a causal effect of smoking on lung cancer, we can use the tools provided by the field of *Causal Discovery* to weed out possible pseudo-causal relationships. *Causal Effect Estimation*, on the other hand, is concerned with the estimation of the effect a specific action, often called treatment, will have on a so called unit or a population. [28, 12]. After the action, we observe some outcome on that unit, which we need to compare to the outcome had the unit not been treated. This is the special setting we aim to study in this thesis.

1.3.1 Treatment Effect Estimation

Consider the example of a medical trial for aspirin. For any given unit (i.e. individual) we can choose whether we apply the treatment (aspirin) or a so called control (placebo). After the administration of the treatment (e.g. one hour later), we measure the outcome (e.g. has headache, does not have headache). The causal effect of the action (taking or not taking aspirin) on the individual is then formulated using the counterfactual notion introduced before: ”*For a unit i that received treatment, what would the outcome have been, had this unit not received treatment*”.

Looking at this statement we see what is known as the Fundamental Problem of Causal Inference (FPCI) [26]: For any unit i, we can only ever observe the outcome either after treatment or after a placebo has been administered, but never both. Note that a unit here refers to an individual in some specified time-frame. Thus, applying a different treatment at a different time on the same individual is considered a different unit j. Dealing with this problem is the core task of causal inference. Essentially, we are trying to derive counterfactual conclusions, i.e. causes, from only looking at factual data, the observations.[55, p. 478].

1.3.2 The Ladder of Causation

To better understand the use of causal inference as compared to statistical analysis, we introduce the Ladder of Causation by Judea Pearl [46]. The basic premise is that there are three conceptually different levels of thinking about events in the world.

First, there is observation, the realm of statistics. Suppose we go out and measure data from a group of people that take aspirin. We collect, for example, their age, gender, weight, whether or not they take aspirin and what the outcome (headache or no headache) was. In other words, we go out and observe the world. Using statistical learning we can now make predictions and analysis like: *”Given a person took aspirin, what is the probability of them still having a headache one hour later?”*. Note that the formulation *”given that”* is a statistical formulation, which we will formalize in probability theory later on.

Now on the second rung of the ladder we are concerned with something different: intervention. The related query could be: *”Given a person (unit) i, what is the effect on the headache of that person if we administer aspirin?”*. This is totally different from the statistical query, because it involves our active intervention in the process. Instead of observing a person that takes aspirin, we make a person take aspirin. To elucidate the difference, consider the following scenario. The people we observe taking aspirin are usually people who take good care of themselves. Thus they also make sure to get rest and drink enough water, while the people who don't take aspirin are, on average, less mindful or caring. What is now the effect of aspirin, if we only consider observational data of the world? Hard to say, right? It could be that the people who take aspirin only get better because they also rest and drink water. This is because we introduced something called a confounder, a variable (mindfulness, in our case) that effects both the treatment (taking aspirin) and the outcome (headache). In other words, mindfulness is a common cause of treatment and outcome. We will discuss confounders more formally in Section 2.2. It's important to remark, however, that intervening in the world by forcefully applying treatment to a unit, changes the rules completely. Now the degree of mindfulness of a person does not affect treatment, because we set treatment. This idea of intervention is the basic premise behind a Randomized Control Trial (RCT), the gold-standard of experimental design [23]. We will refer to interventional data (i.e. data collected by actively applying treatment) as experimental data.

On the third and last rung of the ladder, we are concerned with counterfactuals. This is where things get tricky, because, as noted before, counterfactuals are essentially fictional quantities. It amounts to asking: *"Given a unit i that has received treatment, what would the outcome have been, had it not received treatment?"*. Or in our aspirin example: *"Person A took aspirin and does not have a headache now. How would person A feel, had she not received treatment?"*

Knowing these three rungs, we can again consider causal inference as the enterprise of using observational or, at best, experimental data to derive conclusion about counterfactuals.

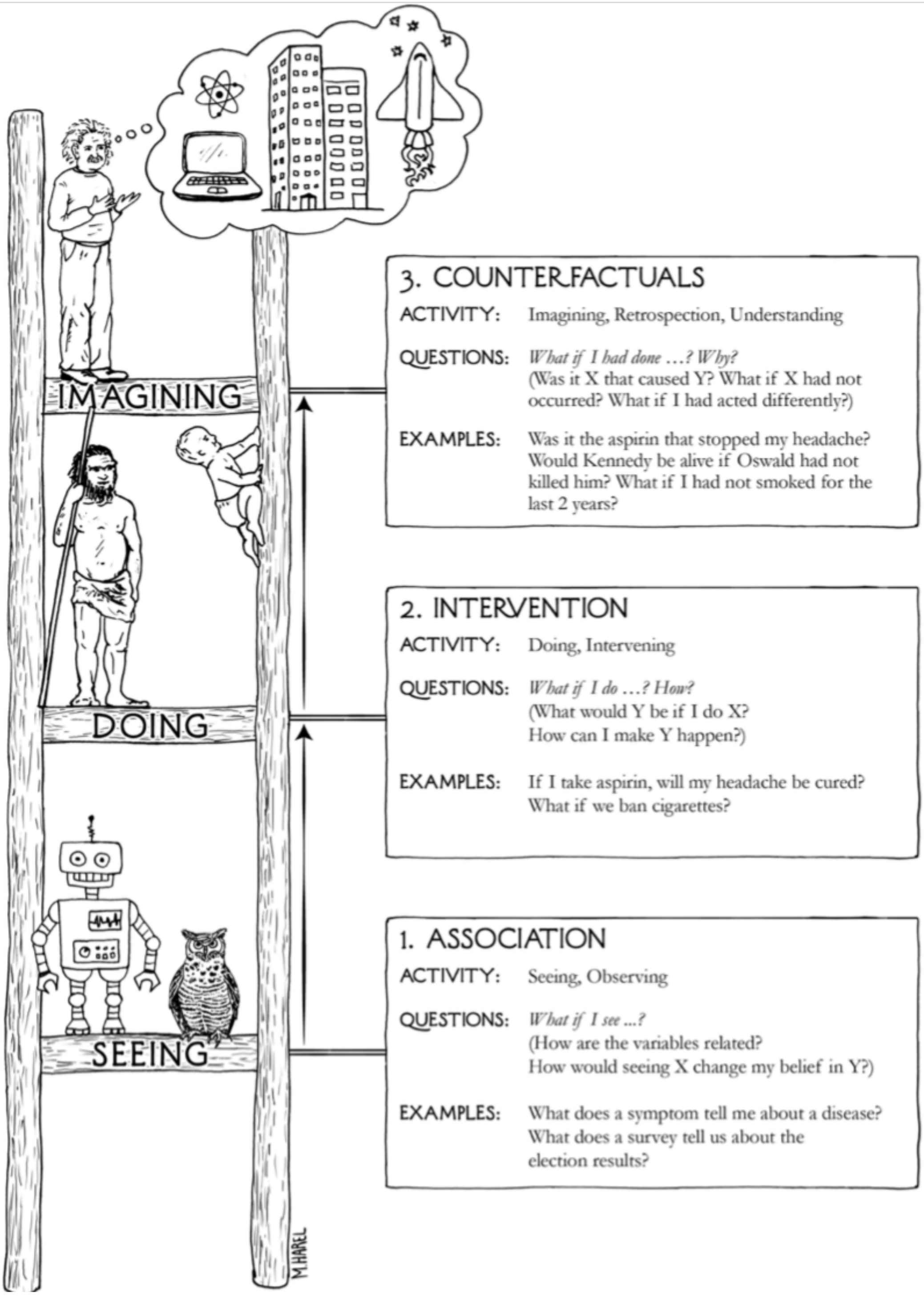
1.3.3 Why do we need Causality?

While the concept of causation is very common in the human language [21, 28], there has been a considerable reluctance of the statistical sciences to introduce the notion of cause and effect [31]. Now that we've understood the basic task of causal inference we can also consider, why we need it in the first place.

In statistics we use observational data to reveal correlations between different variables. For example, we can find a correlation between the weather (e.g. rain) and the barometer, if we were to measure the two over a period of time. Statistics however, does not tell us, whether there is a causal dependency. For another example, namely that fire causes rags to burn, C.R. Shalizi [55] put it quite well: *"For all my data shows, all the rags I burn just so happened to be on the verge of spontaneously bursting into flames anyway."*. This hints at a notion that has become prominent in statistics: *"Correlation does not imply causation"*. In fact, there are multiple scenarios, where inferring causation from correlation is illogical. For example:

- The direction of causation is reversed, because from correlation we can never know the direction of causation. For instance, the temperature and the thermometer are directly correlated, but we cannot know the direction of causation from looking at the two quantities.

- The causation is in both ways. Say we study the effect of health on happiness. Studies show that happiness influences the health, while healthy people also report being happy [24]. Two discern between the two directions is essential.



3. COUNTERFACTUALS

ACTIVITY: Imagining, Retrospection, Understanding

QUESTIONS: *What if I had done …? Why?*
(Was it X that caused Y? What if X had not occurred? What if I had acted differently?)

EXAMPLES: Was it the aspirin that stopped my headache? Would Kennedy be alive if Oswald had not killed him? What if I had not smoked for the last 2 years?

2. INTERVENTION

ACTIVITY: Doing, Intervening

QUESTIONS: *What if I do …? How?*
(What would Y be if I do X?
How can I make Y happen?)

EXAMPLES: If I take aspirin, will my headache be cured?
What if we ban cigarettes?

1. ASSOCIATION

ACTIVITY: Seeing, Observing

QUESTIONS: *What if I see ...?*
(How are the variables related?
How would seeing X change my belief in Y?)

EXAMPLES: What does a symptom tell me about a disease?
What does a survey tell us about the election results?

Figure 1.1: Figure from the Book of Why [46], illustrated by Maayan Harel (www.maayanillustration.com)

- The correlation is spurious or coincidental. E.g. the number of Nobel price laureates in a country correlates with the per capita consumption of chocolate.
- There is a common cause C of A and B. For example, the age a child causally influences its reading ability and its shoe size thus creating a non-causal correlation between shoe size and reading ability (see 2.2.2 for more).

What we mean to point at, is that the concept of correlation is insufficient to account for our thinking about the world, which often includes a causal model. E.g., if I throw a stone into the window, it will cause the window to break [12].

Machine learning as it is currently practised and applied is strongly based on statistics. Often, machine learning approaches merely constitute a form statistical learning, where we aim to estimate the distribution of some unknown target from all the data we have. The setting is observational and the language of causation is banned as the concepts of statistics don't allow for causal conclusions. Bernhard Schölkopf, Professor at the Max Planck Institute, found a practical example where this point of view on the world is insufficient. Namely, Amazon's recommendation system. When searching for a laptop rucksack the system recommends also buying a laptop to go along with it. This is reasonable from a machine learning point of view, as the two items are closely related and probably often bought together. Looking at the recommendation from a causal point of view, however, we are tempted to say that he who searches a laptop rucksack already has a laptop. Recommending a laptop to everyone who buys a laptop rucksack is like recommending to buy a house to everyone who buys a doormat.

Another requirement for the study of causal inference is related to the available data. We've seen that data from RCTs is most meaningful, because it removes potential bias through common causes of treatment and outcome. However, performing a randomized control trial is not always feasible, for several reasons. First, it might be too expensive to run a full-fledged RCT, because doing so requires significant effort and human resources. Second, it might simply be unethical or immoral to do so. "For example, it would be unethical to prevent potential students from attending college in order to study the causal effect of college attendance on labor market experiences." [5]. Causal inference comes to help in such scenarios as it tries to simulate the environment of a RCT only considering observational data. How this can be done is the content of Section 2.1 and the main focus of this thesis.

1.3.4 Simpsons Paradox

Finally, we want to direct the attention to a well known paradox of statistics, that can be better understood, using the language of causes: The Simpsons Paradox [8].

Consider the fictional data in Table 1.1 presented in [46, Chapter 6]; we follow Pearls argument closely.

We see statistics on the effect of a drug given observational data. If we look at the effect of the drug for women, we observe that the percentage of heart attacks increases from 5 to 7.5 percent. The same effect occurs if we look at the data for men, where 12 out of 40 in the control group have a heart attack (30 percent), and 8 out of 20 have a heart attack in the treatment group (40 percent). It seems to be

	Control Group		**Treatment Group**	
	Heart attack	No heart attack	Heart attack	No heart attack
Female	1	19	3	37
Male	12	28	8	12
Total	13	47	11	49

Table 1.2: Toy data to illustrate the effect of Simpsons Paradox [46, Chapter 6].

clear now that the drug must be bad. It has negative effects on women and negative effects on men. Looking at the last row, however, we are surprised to see that when we consider all patients, we observe less heart-attacks in the treatment group (18 percent) than in the control group (22 percent). In words, what this table is showing us is that the drug is bad for women, bad for men but good for people.

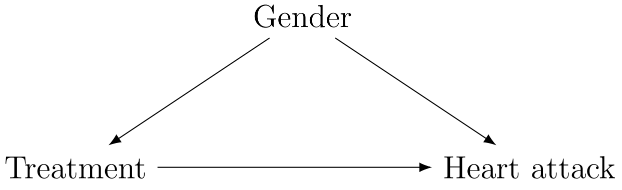


Figure 1.2: Graph G for the toy example

To solve the paradox, we can draw a causal diagram [42], where we depict our causal knowledge about the world. We'll introduce graphical models in more detail in Section 2.2. For now it ought to be enough to say that a directed edge from a node C to a node E, means there is a causal relationship from C to E. Figure 1.2 depicts the causal relationships we can observe in case of the drug experiment. Looking at the data, we can see that men in general are at greater risk of a heart attack. Thus we draw an edge from gender to heart attack. Also, we can observe that women are in general more likely to take the drug than men (e.g. stigma, ...). Thus we also draw an arrow from gender to drug and complete the diagram with the arrow form drug to heart attack, which represents the causal effect that we meant to study in the first place.

What we see now is the visual representation of what we've above called a common cause. The gender influences both the treatment and the outcome and thus introduces a dangerous bias to the data. Statistics alone can never tell us whether the drug is good or not, because we can never know whether to look at the data for each gender separately or combined. Knowing, however, that gender is a common cause, the theory of Pearl [42] tells us that we need to control for gender. That is to say, in statistical parlance, we must look at both genders separately to retrieve the causal effect of the drug on the outcome. Once we've observed the percentages of heart attacks with and without the drug for both genders we take the average and retrieve the true causal result that the drug is bad for women, bad for men and bad for people.

1.4 Goal & Outline

We used this first chapter to give a rough overview of the field of Causal Inference in common language. In the remainder of the thesis, we will now focus on *Causal Effect Estimation*. The goal is to compare a range of methods in a disjoint set of settings in order to give a better overview of the field. To do so, an evaluation framework is designed based on the problems that afflict previous evaluations.

With the foundations in place, we present a wide set of methods for treament effect estimation in Chapter 3. In order to compare these, we shed light on the evaluation procedure of other authors in Chapter 4. Also, we propose a novel evaluation framework and a rough taxonomy for data generating processes used to evaluate the methods. Subsequently, in Chapter 5, we discus the results of our experiments with regard to one specific hypothesis. Using our own evaluation framework, we analyse the methods introduced in Chapter 3 in a range of settings. Finally, Chapters 6 and 7 conclude the thesis and provide an outlook on promising research directions.

2. Foundations

> ...the most serious impediment, in my opinion, has been the fundamental gap between the vocabulary in which we cast causal questions and the traditional vocabulary in which we communicate scientific theories.
>
> Judea Pearl [46]

The gentle introduction in the previous chapter makes the fundamental problem obvious that comes with a philosophically clean definition of causality. Knowing that words and examples aren't enough, we here lay the formal foundations for the thesis. Particularly, we consider the Potential Outcomes framework [49] with its assumptions and its formalisation of the specific problem of treatment effect estimation. For sake of completeness, we also introduce the Structural Theory by Pearl[42], because this theory helps to comprehend Potential Outcomes within the bigger terrain of causal inference.

In the remainder of the thesis we use the notation described in Table 2.1 if not declared otherwise.

2.1 Potential Outcomes

When considering causal effect estimation methods, we are usually facing an experimental setting where we have collected a sample of n instances $S = \{X_i, T_i, Y_i\}_i^n$ with all d measured covariates $X_i \in \mathbb{R}^d, X_i = (X_{i1}, \dots, X_{id})$ of a unit i, its treatment vector $T_i \in \{0,1\}^k$ indicating which of the k treatments was given and its one observed outcome $Y_i \in \mathbb{R}$. Note that a unit may be an individual in a clinical trial or a measurement of some process that is clearly distinguishable from other measurements [12]. To retain generality, we continue to call these instances *units*. The treatment variable T_i defines what treatment has been applied to a given unit i. It

Syntax	**Meaning**
$P(A \mid B)$	Conditional probability of A given B
$P(A, B)$	Joint probability of A and B
$Y_i(0), Y_i(1)$	Potential outcomes of unit i
Y_i^{obs} or Y_i	Observed outcome of unit i
Y_i^{cf}	Counterfactual outcome of unit i
$T_i \in \{0, 1\}$	Observed treatment indicator for unit i
$X = (X_1, ..., X_n)$	Covariate vector as random variables
$X_i = (X_{i1}, ..., X_{in})$	Specific covariate vector for unit i. Generally, X_i alone refers to a specific unit, while distinct numbers, e.g. X_3 refer to specific features over all units.
$X = x$	Realisation of a vector-valued random variable
$\mathcal{X}, \mathcal{Y}$	Space of all covariates, space of all outcomes
$p(x) = P[T_i = 1 \| X_i = x]$	The propensity score
$\hat{y}$	Estimator of a quantity y
S	For sets of instances, e.g. $S = \{X_i, T_i, Y_i\}_i^n$
$\perp\!\!\!\perp$	Conditional independence (see Definition 2.8)
$\textsc{Bern}(p)$	Draw from the Bernoulli distribution with probability p
σ	The sigmoid function, $\sigma = \frac{1}{1+e^{-x}}$
ϵ	For error or loss measures, see Section 4.4
$\mathcal{N}, \mathcal{U}$	Normal distribution, uniform distribution
$exp(x)$	The exponential function, e^x, for longer expressions
$\mathbb{I}(C)$	Indicator function; 1 exactly if condition C is true
$\{i \mid C\}$	Set of elements i that fulfill condition C

Table 2.1: Notation and syntax used throughout the thesis.

is a unit vector where precisely one element is 1, for only one of multiple treatments is possible. Following the majority of the literature, we assume treatment to be binary, $T_i \in \{0, 1\}$ and we refer to units i with $T_i = 0$ and $T_i = 1$ as being in the control and treatment group respectively. That is to say, the control group does not receive treatment (e.g. a medication, a special ad, etc.). The outcome Y_i then is the measurement of some effect quantity after a treatment has been applied.

2.1.1 Formulating a Super Distribution

An essential step in the pursuit to formalize causal effects is to make room for counterfactuals. That is to say, we must somehow capture the fact, that there are two counterfactual worlds. One in which unit i did receive the treatment and one in which it did not. The so called Neyman-Rubin potential outcomes framework (sometimes also Rubin-Neyman causal model) [49], [28] does exactly that. It is the most widely used framework in various fields ranging from social sciences over medicine to economics [9]. It is strongly based on the setting of an RCT [42] and has helped to justify and formalize causal claims in studies.

Mapping this notation on our example from Section 1.3.2 we observe a vector of covariates X_i of each individual (e.g. age, gender, ...), the treatment T_i (received aspirin vs. placebo) and the potential outcomes $Y_i(0), Y_i(1)$ if unit i had / had not received treatment. Of the two potential outcomes $Y_i(T_i), T_i \in \{0, 1\}$, we observe only one, which we denote by Y_i instead of Y_i^{obs} [28] for simplicity.

An important characteristic of the potential outcomes framework is that "the analysis itself is conducted almost entirely within the axiomatic framework of probability theory" [42, p. 127]. To do so, a so called "super" distribution P^* is introduced that contains both the factual as well as the counterfactual events. Formally we write $P^*(X, T, Y(0), Y(1))$ for the super ditribution, while the observed distribution $P(X, T, Y)$, from which the sample S is assumed to be drawn, is seen as the marginal distribution over the missing outcomes, the counterfactuals. To make this formally viable, at least one consistency constrained is introduced that ensures the correct "behaviour" of the fictional random variables $Y(1)$ and $Y(0)$:

$$T_i = t \Rightarrow Y_i(t) = Y_i, \tag{2.1}$$

which, in the binary treatment case we consider, is equivalent to

$$Y_i = T_i Y_i(1) + (1 - T_i)(Y_i(0).$$

We will see this assumption again in Section 2.1.5 as the Stable Unit Treatment Value Assumption (SUTVA). It requires the observed outcome for a given treatment is the same as the potential outcome for the treatment from the super distribution.

2.1.2 Targeting Treatment Effects

Having set the foundations we can now consider the task of treatment effect estimation again. This time from a more formal point of view. We are aiming to find causal effects for either a unit, a subgroup or the whole population. Statistically, we assume all the experimental data S we collect to stem from the, usually unknown, distribution $S \sim P(X, T, Y)$. We can then define different causal effects depending on the population we are interested in either on the distribution or empirically based on the finite sample.

Definition 2.1 (Individual Treatment Effect (ITE) [28]). *For a given unit i in a population, we define the individual treatment effect (ITE) as*

$$\tau_i := Y_i(0) - Y_i(1)$$

Definition 2.2 (Average Treatment Effect (ATE) [33]). *The average treatment effect over a distribution is defined as*

$$\tau := \mathbb{E}[Y(1) - Y(0)]$$

and for a finite sample population of n units as

$$\tau := \frac{1}{n}\sum_{i}^{n}(Y_i(1) - Y_i(0)).$$

Another formulation that is useful to consider the treatment effect of a subpopulation (e.g. only females, only people of age 34, ...) is the Conditional Average Treatment Effect (CATE).

Definition 2.3 (Conditional Average Treatment Effect (CATE) [33]). *The conditional average treatment effect of a subsample $\{(Y_i(1), Y_i(0), X_i, T_i) \mid X_i = x\}$ or the corresponding distribution $P(Y(0), Y(1), T|X = x)$ is defined as*

$$\tau(x) := \mathbb{E}[\tau|X = x] = \mathbb{E}[Y(1) - Y(0)|X = x].$$

Note that x does not have to specify all covariates in $X \in R^d$. If $X = x_i$ identifies one unit, then $\tau_i = \tau(x_i)$. In [33, 32] the finest conditioning level using all available covariates is called individualized average treatment effect (IATE).

Notice first, that essentially none of these definitions is graspable given only the observable distribution $P(X, T, Y)$, but rather using the super-distribution P^*. Notice also, the significant difference between considering average treatment effects or individual treatment effects. This is so, because of a concept called treatment effect heterogeneity. Treatment heterogeneity is the presence of subgroups in the population that react differently to the same treatment [44]. For example, for some patients a given drug might be more effective due to genetic effects. Especially in social-, political sciences and econometrics we can rarely assume treatment effect homogeneity [5]. Choosing to consider the Average Treatment Effect (ATE) amounts to ignoring the presence of varying subgroups. More importantly though, it might lead to inconclusive or plain wrong results. Consider a drug that is wildly effective for some individuals, but in general has small negative effects on the desired outcome (e.g. health). Considering the ATE would yield a positive causal effect and would thus result in the recommendation to prescribe the drug as often as possible. It becomes clear, that for an individual medical recommendation it is reasonable to consider Individual Treatment Effect (ITE) or at least a Conditional Average Treatment Effect (CATE) on a specific stratum.

2.1.3 Assignment Mechanisms

An essential factor in determining whether or not a treatment effect is tractable given data is the assignment mechanism [28, Chapter 3]. The assignment mechanism is a function of the covariates X_i and the potential outcomes $Y_i(1), Y_i(0)$ and determines the probability of treatment. We quickly define the terms in order to discern between different use-cases later on. Beware that we introduce the same concepts with different names, picking up the terminology in the literature. We then clarify the terminology used throughout the thesis in Section 2.1.5.

Definition 2.4 (Assignment Mechanism). *An assignment mechanism is a function* $P(T_i \mid X_i, Y_i(0), Y_i(1))$ *mapping features and potential outcomes to the probability of treatment* $[0, 1]$

Definition 2.5 (Individualistic Assignment). *An assignment mechanism is individualistic if the probability that a unit is assigned treatment does not depend on the covariates or potential outcomes of other units.*

Definition 2.6 (Probabilistic Assignment). *An assignment mechanism is probabilistic if the probability that a unit i is assigned treatment is strictly between zero and one:*

$$0 < P(T_i \mid X_i, Y_i(0), Y_i(1)) < 1, \text{ for all } i = 1, \ldots, n$$

Definition 2.7 (Unconfounded Assignment). *An assignment mechanism is unconfounded if it does not depend on the potential outcomes. That is to say,* $P(T_i \mid X_i, Y_i(0), Y_i(1)) = P(T_i \mid X_i)$

2.1.4 Different Experimental Settings

Roughly, we can discern between the different settings for causal effect estimation using two dimensions. The underlying assumption and the goal of estimation. The underlying assumptions are broadly split into *Unconfoundedness* and *Hidden Confounders*. When assuming *Unconfoundedness*, our belief in the property either stems from the way the experiment was designed (e.g. RCT or experimental data in general) or from the underlying distribution of the data we observed by using expert knowledge about the domain.

The goal of causal effect estimation can be subdivided into average treatment effects over a population and individual or conditional treatment effects of sub-populations. The Individual treatment effect can be understood as a conditional treatment effect where the condition directly identifies one individual.

Knowing these properties we can discern between randomised data (from a RCT) and observational data. As Rosenbaum and Rubin [48] assert, randomised trials differ from non-randomised ones in two ways:

- First, the assignment mechanism is known, because it is usually chosen by the researcher and the treatment assignment is *probabilistic* (as in Definition 2.6).

- Second, the treatment assignment and the potential outcomes are conditionally independent given the covariates. Formally, $(Y(1), Y(0)) \perp\!\!\!\perp T \mid X$. In other words, the assignment of treatment only depends on observed covariates. We say, treatment assignment is *ignorable* (as in Definition 2.7).

For completeness, let us define the conditional independence used above.

Definition 2.8. *Conditional Independence. Given random variables X, Y, Z, we say 'Y is independent of X conditioned on Z', written $X \perp\!\!\!\perp Y \mid Z$ when*

$$P(X = x, Y = y \mid Z = z) = P(X = x \mid Z = z) \cdot P(Y = y \mid Z = z)$$

or equivalently:

$$P(Y = y \mid X = x, Z = z) = P(Y = y \mid Z = z) \,\text{or}\, P(X \mid Z) = 1$$

Following Rosenbaum and Rubin [48], treatment assignment is strongly ignorable, if it is probabilistic and ignorable.

Definition 2.9 (Strong Ignorability [48]). *Treatment Assignment is strongly ignorable if*

$$(Y(1), Y(0)) \perp\!\!\!\perp T \mid X \text{ and } 0 < P(T = 1 \mid X) < 1.$$

Under these assumptions treatment effects can be identified. That is to say, we can use the presence of multiple units to make up for the missing information about the unobserved potential outcome.

In the following we will continue to use the assumptions defined in the next section to refer to the various treatment assignment mechanisms and corresponding assumptions underlying the data.

2.1.5 Assumptions to identify causes

As we've elaborated in the introduction, individual causal effects cannot be observed. It is the task of causal inference to estimate these and to do so, we have to resort to some simplifying assumptions. The task of estimating the ITE from data is often preceded by the task of identification [42, Section 3]. Identification asserts whether a given counterfactual quantity (e.g. a treatment effect) can be calculated given only factual data.

In the simplified setting of the potential outcome framework, assumptions are made to enable identification. These assumptions are reasonable in RCT, but remain questionable in observational settings. However, most of the methods we will consider in Chapter 3 require the following three.

Assumption 2.1 (Stable Unit Treatment Value Assumption (SUTVA))**.** *The SUTVA combines two aspect together, that make working with causal data easier. Namely,*

- *no interference between units (Similar to Definition 2.5),*
- *well defined treatment levels (i.e. there is no half-treatment or treatment with decreased efficacy).*

Formally, we can write $Y_i = T_i Y_i(1) + (1 - T_i) Y_i(0)$ *to capture these assumptions.*

Assumption 2.2 (Unconfoundedness)**.** *Unconfoundedness is identical to ignorability:* $(Y(1), Y(0)) \perp\!\!\!\perp T \mid X$*.*
In words, the potential outcomes are independent of the treatment given the covariates.

Assumption 2.3 (Overlap)**.** *The overlap assumption makes up the second part of strong ignorability. It states that for all instances the probability of treatment must be strictly between zero and one. Formally,* $0 < P(T = 1 \mid X) < 1$*.*
This is equivalent to the assignment mechanism in Definition 2.6.

2.1.5.1 Discussion of Assumptions

When is the *Unconfoundedness Assumption* questionable? Consider an example where the data is collected in a hospital setting. We are considering the administration of a given treatment, but the treatment assignment to individuals is not random (i.e. we are merely observing the operational processes in the hospital). Thus the assignment of the treatment on a patient might depend on some insight or knowledge the doctor has about the patient, which we do not observe in the covariate vector X_i. In this case, the treatment assignment is confounded and our methods will not work properly. If, however, we were to measure whatever feature the doctor was considering for his decision, then the treatment would be unconfounded conditional on the covariate vector.

When is the *Stable Unit Treatment Value Assumption* questionable? Consider for example a setting in which the individuals in both treatment and control group are in contact with each other. In this case the positive effect of treatment might *spill over* to other individuals in the control group (e.g. the improved mood of the treated affects the mood of the control group).

When is the *Overlap Assumption* questionable? In a case where there is a prohibitive factor for receiving a treatment, the overlap assumption is questionable. For example, if pregnant women are not allowed to take a drug, the probability of assignment is strictly zero and the overlap assumption does not hold. In such a case, given the prohibitive factors are known, the data can be trimmed in order to enforce the overlap assumption. That would mean we do not consider pregnant women at all in our study.

2.1.6 Causal Inference as a Missing Data Problem

We have so far introduced causal effect estimation in the common setting of treatment on a given unit. This is rightly so for its origin in this domain. However, through a more abstract consideration we can also see the problem we aim to solve as one of missing data. Peng Ding and Fan Li [13] discuss how the formalisation of missing data problems and the corresponding inference methods closely match the causal inference notation and methods introduced in the *Potential Outcomes* framework. Specifically, we are looking to reconstruct two distributions from incomplete data. The goal is to know the distribution of $Y(1)$ and $Y(0)$ over all covariates that describe a unit. The *Fundamental Problem of Causal Inference* now makes this problem essentially a missing data problem. For any individual, we only ever observe either $Y_i(1)$ or $Y_i(0)$, thus missing the respective counterpart. The treatment indicator in the treatment effect case tells us which of the two distributions we observe. Analogously it tells us, which data is missing. And the distribution of this *missingness* is what we aim to constrain with our assumptions. Having ignorable treatment assignment (Assumption 2.2) means that the distribution of the *missingness* does not depend on outcomes when we know all the covariates. Without the *Unconfoundedness Assumptions*, the *missingness* might be dependent on the value of the outcome such that higher outcomes are more likely to be missing.

2.2 Structural Models

Closely following the elaborations and notation in [7, 42, 43, 58] we shortly introduce Structural Causal Models (SCMs) in order to extend our understanding of the field with a formal modelling language. This language is more graspable than the statistical formulation of the potential outcomes framework. We will mainly use the graphical approach as a way to model and communicate our assumptions. We use the graphs for didactic purposes in our specific pursuit to estimate causal effects. This is not to say by any means that the mathematical tools provided by SCMs are inferior or useless. They are, if anything, too sophisticated for the simplified setting in Section 2.1, which we are concerned with. The theory around SCMs is, as Pearl writes, not a contender for the potential outcomes framework, but rather a generalizing theory that subsumes the formulations of the potential outcomes framework [42].

2.2.1 Structural Causal Models

A structural model M consists of two sets of variables $\mathcal{U}$ and $\mathcal{V}$ and a set of functions $\mathcal{F}$ that determine how the values of the endogenous variables $V_i \in \mathcal{V}$ are assigned their values. Essentially, we differentiate between *exogeneous* and *endogeneous* variables as the ones that are set externally with respect to our model and those that are causally dependent of others through functions $f_i \in \mathcal{F}$. The assignments of variables in $\mathcal{V}$ usually represent causal physical processes. Formally, we can write $M = \langle \mathcal{U}, \mathcal{V}, \mathcal{F} \rangle$. For any $V_i \in \mathcal{V}$ and $U_i \in \mathcal{U}$ we can write exemplary $V_i = f_i(V, U)$. The relations expressed in a model M can also be captured in a graph $G(M)$, or just G if the corresponding model is clear. To avoid confusion regarding notation, we will continue to name variables in the SCMs by their meaning in the potential outcomes framework (i.e. T for treatment, X for covariates, etc.). It is important to note, though, that SCMs are not constrained to the limited treatment effect estimation setting. The graphs can be used to model any causal dependence and complex system.

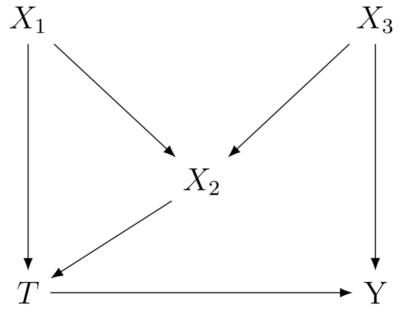


Figure 2.1: Graph G of the model M for the asthma study.

To make the use of SCMs clearer, let us consider the following example from [63] used also in [46, Chapter 4]. The authors aim to study the effects of smoking on asthma and have modelled the dependencies they know in a structural causal model M, with the corresponding graph shown in Figure 2.1. Following the notation in Section 2.1, $T \in \{0, 1\}$ stands for the treatment, in this case whether or not a person smokes. Y stands for the outcome, which in this case is the severity of asthma. The covariates X_1 through X_3 stand for (1) whether a parent of the individual smoked, (2) whether the person had asthma in her childhood and (3) for an underlying predisposition towards asthma. The variables X_1 and X_3 are *exogeneous*, meaning that we don't know or don't encode the mechanism by which they are assigned in the model. Formally, $X_1, X_3 \in \mathcal{U}$. For all other variables we write

$$\begin{aligned} T &= f_T(X_1, X_2), \\ X_2 &= f_{X_2}(X_1, X_3), \\ Y &= f_Y(T, X_3). \end{aligned}$$

With this model M we can determine the values of the endogenous variables X_2, T and Y in that order, whenever we observe an instantiation of the exogeneous variables in $\mathcal{U}$, i.e. $X_1 = x_1, X_3 = x_3$. Mostly, however, we don't know the exact functions $f \in \mathcal{F}$. In this case the graphical model can still be used to draw important conclusions about the data, as we'll see later.

Controlling for a Variable

As a quick aside, we clarify the term *"to control for"*, which is used regularly throughout the following chapter. In statistical parlance, we say that we control for a variable A, if we either try to hold it constant by design of the study or by include it in a regression model on the outcome. In either way, the goal is to separate the effect of the variable A from other variables in the study. The simplest way to consider the effect of controlling in an observational study is to look at the tabular data and only consider rows where the variable A takes a specific value. In other words, we look at the single strata of A.

In a probability distribution $P(A, B, C)$ over A, B and C, we say we control for A, if we look at the conditional distribution $P(B, C \mid A = a)$.

2.2.2 Common Typologies and the flow of information

Before we continue with our analysis of the graph in Figure 2.1, we want to introduce the fundamental typologies underlying causal graphs. There a three main component of which any causal graph is made up and they carry specific characteristics. Namely, we consider a causal chain, a confounder and a collider. For simple causal relations we use the inline notation $A \rightarrow B$, if A is required to determine B or, in other words, if A is a cause of B.

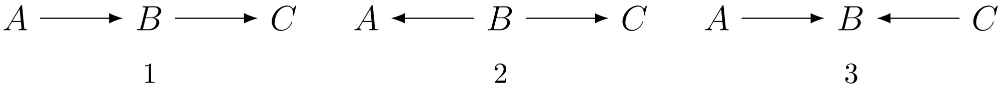


Figure 2.2: The three basic typologies from left to right: chain, confounder, collider.

The causal chain (1) can be exemplified by the causal relationship of fire and the fire alarm. A is fire which causes B, smoke, which results in C, the fire alarm going of. We say that B mediates the effect from A on C. Namely, fire does not cause the fire alarm to go off directly, it does so by means of causing smoke.

The second kind of junction is called a confounder, or more explicitly, B is a confounder of A and C. In other words, B is a common cause of A and C. An example that David Freedman proposed is ”Shoe Size” ← ”Age of Child” → ”Reading Ability”. Within in this simple topology we can observe something interesting. If we were to measure data on A and C, namely shoe sizes of children and their reading ability, without knowing about their ages, we observe that they are correlated. With increasing shoe size, reading ability increases. We could come to the wrong conclusion that there is a causal relationship between the two. If we now also measure B, the age, we can control for it and will see that within a stratum of B there is no correlation between A and C. For example, looking only at the population of 14 year old children, there is no relationship between shoe size and reading ability.

Finally, there is the collider. This topology results in a vary strange behaviour called the ”explain-away-effect” by Pearl [46, Chapter 3]. Consider this causal model: ”Talent” → ”Celebrity” ← ”Beauty”, which formalises that both beauty and talent contribute to the celebrity of an actor or actress (Example from [16]). If we now collect data on A and C independently we will, as expected by common sense, not observe any correlation between beauty and talent. If we collect B, namely celebrity by some measure, and control for it, than we do observe a negative correlation between A and C. That is to say, looking only at famous actors and actresses, we observe the general relationship that they are either beautiful or talented. In other words, if we look at a famous actor and see that he is not very beautiful, we assume he is very talented, otherwise he wouldn’t be famous. This correlation was only created by us looking at a specific stratum of B, namely famous actors.

Having considered these basic typologies we can discern between two types of information flow in a causal graph [55, p.481]. There is the causal information on the one hand side, which, figuratively speaking, flows along the edges of the directed graph. On the other hand side there is statistical information, which, as we’ve seen, flows in both directions of the edge, because correlation is inherently non-directional (See also Section 1.3.3). These types of information flow become crucially important for our analysis of causal effects. Essentially, we only want to measure causal information flowing directly from treatment to outcome while ignoring all the statistical information flowing through indirect paths contaminating our results.

2.2.3 The Do-Operator

The do operator introduced by Pearl (e.g. [42]) extends the language of statistics, located on the first rung of the ladder of causation, with a tool to model interventions, residing on the second rung. The do-operator does exactly what his name implies. Given a structural causal model (e.g. Figure 2.1), we can formulate a query $P(Y \mid do(T = t))$, which, in plain English equals the the question: *"What is the distribution of the outcomes, given we set the treatment to t / given we do treatment".* Notice that we "set" the treatment, which is different to observing a specific treatment (See Section 1.3.2).

For the purpose of studying causal effects, $P(Y \mid do(T = 1)) - P(Y \mid do(T = 0))$, the difference between intervening on treatment and not intervening treatment, is exactly what we want. However, as we've discussed before, the experimental setting in which we can actually *do* something is rarely the reality we face. The quest of causal effect estimation, from the structural point of view, is thus to determine whether we camulate the quantity $P(Y \mid do(T = t))$ only using observational data (i.e. probability distributions without *do* operator). The next section introduces a graphical method for doing exactly that.

2.2.4 Causal Effect Estimation and the Back-Door-Criterion

Having build an intuitive understanding of causal graphs and their interpretation, we now want to use the theory of structural models to look at the task of causal effect estimation from a different point of view. Essentially we are trying to block all indirect, non-causal paths between treatment T and the outcome Y.

To do so we use the tools provided by Pearl [42, 43]. First we define what it means to block a path p.

Definition 2.10 (d-separation [42]). *A set of nodes S is said to block a path p if either*

- *p contains at least one arrow-emitting node that is in S or*
- *p contains at least one collision node that is outside S and has no descendant in S.*

If S blocks all paths form X to Y, it is said to "d-separate X and Y", and then, X and Y are conditionally independent given S, for which we write $X \perp\!\!\!\perp Y \mid S$. Note that a path does not have to be directed, since statistical information flows either way [55].

To illustrate the definition, consider the three archetypical topologies. In the causal chain $A \to B \to C$, S = {B} d-separates A and C. In the confounding constellation

$A \leftarrow B \rightarrow C$, B also blocks the path from A to C. Only in the collider case we have to be careful. Here, A and C are already d-separated by the empty set $S = \{\}$. Controlling for B, as we've seen, breaks this separation and results in a correlation of A and C.

We now come back to the example introduced in the beginning of this section. We redraw the graph here for convenience.

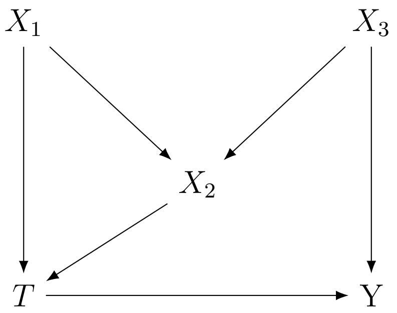


Figure 2.3: Graph G of the model M for the asthma study, redrawn from Figure 2.1.

To retrieve the unbiased effect of treatment $P(Y \mid do(T = t))$, we have to consider closing all indirect paths from T to Y, namely $(T \leftarrow X_2 \leftarrow X_3 \rightarrow Y)$ and $(T \leftarrow X_1 \rightarrow X_2 \leftarrow X_3 \rightarrow Y)$. We follow Pearl and call these indirect paths "back-door" paths [43]. A set of nodes S that fulfills this requirement is called an *admissible set* and we can define this using Pearls *back-door criterion*:

Definition 2.11 (Admissible Set [42])**.** *A set of nodes S is admissible for adjustment regarding a causal effect of T on Y if two conditions hold:*

- *No element of S is a descendant of X*
- *The elements of S "block" (Definition 2.10) all "back-door" paths from T to Y, namely all paths that end with an arrow pointing to T.*

Finding such an admissible set for a fully specified causal model can be done systematically and there are polynomial time algorithms solving the problem for any graph and query [42].

For our specific example, we know which paths to consider. We also know that we must not add X_2 to S, for this would open the backdoor path $(T \leftarrow X_1 \rightarrow X_2 \leftarrow X_3 \rightarrow Y)$, but we must close the path $(T \leftarrow X_2 \leftarrow X_3 \rightarrow Y)$, for it carries statistical information from T to Y. We can do so by controlling for X_3. Using that now

$Y \perp\!\!\!\perp T \mid X_3$ and the consistency assumption $Y = Y(1)T + Y(0)(1-T)$, we can write

$$\begin{aligned} P(Y \mid do(T=t)) &= \sum_{x_3} P(Y|do(T=t), X_3 = x_3)P(x_3) \\ &= \sum_{x_3} P(Y|do(T=t), T=t, X_3 = x_3)P(x_3) \text{ (using Asm 2.2)} \\ &= \sum_{x_3} P(Y|T=t, X_3 = x_3)P(x_3) \text{ (using Asm 2.1).} \end{aligned}$$

Thus, $S = \{X_3\}$ is an admissible set and we have the equality

$$\begin{aligned} &P(Y \mid do(T=1), S=s) - P(Y \mid do(T=0), S=s) \\ &= P(Y \mid T=1, S=s) - P(Y \mid T=0, S=s), \end{aligned}$$

where the right hand side can be evaluated from the data using statistical tools alone [42, Eq. 40], because it does not contain the *do*-operator.

2.2.5 Collider Bias - Or Why we Need Structural Models

In order to show the power of graphical modelling and the danger waiting in its absence, we now run through a little synthetic working example.

Imagine we have collected observational data from a medical trial. We aim to study the effect of treatment T (some medication) on the outcome Y (recovery time) and we collect four covariates (age, sex, severity of disease and a collider) in X. The variables X_{sex} and X_{sev} are confounders that influence both treatment and outcome, age only influences the outcome and the imaginary collider X_C is a common effect of treatment and outcome. Figure 2.4 shows the causal structure of the Data Generating Process (DGP). We also assume for now that we didn't know the causal structure and only collected the data in a format $\mathcal{S} = \{T_i, Y_i, \mathbf{X_i}\}_i^n$ for $n = 50.000$ individuals.

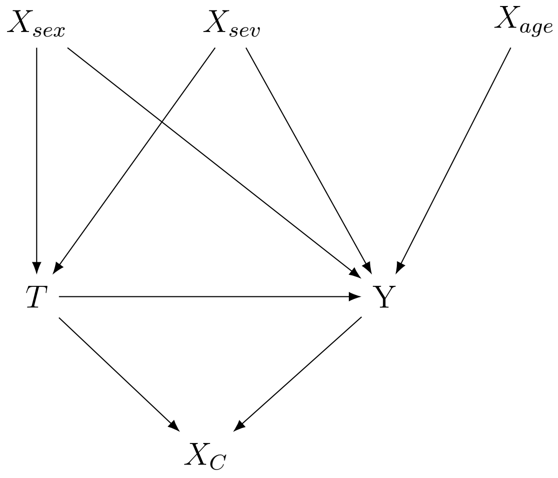


Figure 2.4: Graph G_C for the collider toy example.

Following the approach that Pearl criticizes in [42, p.130] we would be tempted to include all covariates X in our analysis using, for example, regression adjustment.

However, if we compare the results for the average treatment effect, we observe that including the collider into our analysis totally skews the results. Interesting from a machine learning point of view is that the Mean Squared Error (MSE) of the target prediction (outcome in the case of regression adjustment) is significantly lower when we include the collider variable. This is so, because as Pearl says, X_C still carries statistical information that helps to improve the prediction. We are, however, as discussed above, not concerned with a merely predictive task. Instead we want to find the average coefficient of the edge $T \rightarrow Y$ and thus we have to block all back-door paths carrying statistical information. Controlling for X_C by including it in the analysis opens the backdoor path $T \rightarrow X_C \leftarrow Y$ and thus yields wrong results. Instead an admissible set for the identification of $P(Y \mid do(T=1)) - P(Y \mid do(T=0))$ is $\{X_{sex}, X_{sev}\}$, as that set closes the two back-door paths through the confounders. Also, we include X_{age} into the feature set for our outcome regressor, because it influences the outcome variable and leaving it out would amount to ignoring the influence of age on the recovery time, which is in our example unreasonable.

Note that this is a simplified example to illustrate the devastating effect of a collider. Realistically, statisticians often only consider pre-treatment variables as confounders. In this case, that would mean they don't consider X_C, because it cannot be a pre-treatment variable. After all, the outcome Y is its cause. Figure 2.1, however, also includes a collider that any well-meaning statistician would include as a confounder [46, 42]. Namely, X_2 is associated with T and Y and could thus be considered a confounder. Controlling for it, though, will open the collider-path $T \leftarrow X_1 \rightarrow X_2 \leftarrow X_3 \rightarrow Y$ and pollute the results. The constellation present in Figure 2.1 is, due to it's iconic shape, called M-bias. We've seen what Pearl sum up by saying: "M-bias puts a finger on what is wrong with the traditional approach" [46, Chapter 4.]. Specifically, the common statistical procedure without Structural Models does not contain appropriate tools to make arguments about ignorability. That is to say that statistician using the Potential Outcomes framework alone often make the unconfoundedness assumption without a theoretically sound argument for its validity. Joffe et al. [29] highlight that, *"such assumptions [i.e. unconfoundedness] are usually made casually, largely because they justify the use of available statistical methods and not because they are truly believed".* Pearl and other advocates of the structural approach point at the fact that still today, a concept as essential as a confounder, is often misunderstood and not handled with the appropriate care. To illustrate, look at this opening definition from a 2016 paper: *"all pre-treatment variables that predict both treatment and outcome, also known as confounding covariates"* [15]. What sounds intuitively right from a statistical perspective fails to discern the M-bias example from a true confounder. For the variable X_2 does predict both treatment and outcome as it is associated with both. It does not, however, yield the expected result to control for X_2 as if it were a confounder.

Notice that while the structural theory aims to make clean mathematical statements, the actual estimation of effects given a sample is still prone to sampling errors. Thus, knowing that an admissible set closes a backdoor paths theoretically, does not guarantee a perfect estimation of the causal effect given a finite sample.

The collaborative approach using both frameworks outlined here is used in Microsofts *DoWhy*[1] package by default. Essentially, they split the task into three steps: mod-

[1] More information: https://microsoft.github.io/dowhy/

elling the scenario as a graph, identifying the effect, estimating the effect. A collider, like in our example, would by design not be included in the estimation procedure to ensure that the true estimation is possible.

2.3 When to use what?

The approaches taken by the potential outcome practitioners and the advocates of structural theory differ fundamentally. We note some of the differences we have encountered and elaborate on the use of each of the methods. Most importantly, it ought to be stated that the Potential Outcome Framework (Section 2.1) and the Structural Models are not competing alternatives. Rather, they both have their merits and requirements. Generally, as Pearl notes [42], the Structural Theory comprises the potential outcomes framework and we ought to use the best of both worlds to achieve our desired results.

Structural Causal Models for Modelling Assumptions. Using graphical models, i.e. causal graphs, to encode assumptions about an underlying data distribution is much more convenient and transparent than formulating model blind conditional independence assumptions in the potential outcomes framework. [42, 45, 7]. For example it is easy to assert via the graph $A \rightarrow B \rightarrow C$ that $C \perp\!\!\!\perp A \mid B$. The potential outcome analyst, however, would write $B_a \perp\!\!\!\perp \{C_b, A\}$ instead of the graph and read: "the value that B would obtain had A been a is independent of the value that C would obtain had B been b jointly with the value of A. In theory, the same conditional independence $C \perp\!\!\!\perp A \mid B$ can be derived, but the way the knowledge is modeled is far removed from our own understanding. Thus, it is easier to put our knowledge about the world we observe in the form of a causal graph. Studies have been done that show that the causal graph represents closely the way humans think about the relation of events in the world [21].

Structural Causal Models for Identification. It is often said that Pearls structural models are more for the qualitative analysis of a problem. We found this to be a confusing statement, for the SCMs also allow the quantitative estimation of an effect from data once it has been identified, using normal regression approaches. However, SCMs are far more capable in identifying the causal effect in the first place. In fact, they are the only model to do so, because in the potential outcomes framework we simply assume the necessary condition (i.e. unconfoundedness). Using a graph on the other hand we can determine rigorously whether a desired query is answerable given the data.

Structural Causal Models for Infinite Data. Within the structural framework causal effects can be estimated by controlling for the right adjustment set and then looking at the respective conditional distribution, which can be observed from the data. If, however, we have many confounders and thus must control for a large set of covariates, the sample size of the observed data becomes a bottleneck. Thus, the estimation implied by SCMs is ideal for the rare cases with small adjustment sets or infinite data [62].

Potential Outcomes for quantitive approximation. In case we don't know the causal structure of all covariates, we often ressort to the potential outcomes framework and the methods that are build on top of it. This, while being logically questionable at times (see Section 2.2.5), at least allows us to get some approximation without a tedious theoretical analysis of the causal structures. Instead, we often introduce the unconfoundedness assumption and then continue to work with the statistical methods of the Potential Outcome framework.

3. Treatment Effect Estimation Methods

> Blessed is he who has been able to win knowledge of the causes of things.
>
> Virgil (29 BC)

Having set the foundations for the study of causal effect estimators we now dive into the various methods, we want use in the performance evaluation in Chapter 5. We have chosen the following methods, because they each stand for a peculiar paradigm and all of them use largely different approaches to the problem.

3.1 Generic Meta-learners

The following baseline methods are firmly based on the theoretical underpinning of causal effect estimation and thus are introduced first. Other than the more elaborate methods later on, these have been widely used in the respective literature for almost fourty years. They are generic in the sense that all of them employ a regression method without specifying which regression to use. Thus we can plug in any supervised learning method or statistical regression of our choice.

3.1.1 Propensity Score Weighting

Being one of the most straightforward and oldest techniques in the field, inverse propensity score weighting [48, 20] still has its right for existence. We first introduce what the propensity score is and then show how it can be used to estimate the average treatment effect.

Definition 3.1 (Propensity Score [48]). *The propensity score is the probability of treatment given the covariates of a unit. Formally,*

$$p(x) = P(T = 1 \mid X = x) = \mathbb{E}[T \mid X = x].$$

Using this propensity score estimate, the idea is to create a pseudo-population by weighting each sample with the *inverse propensity score* (IPS-Weighting, or IPSW). According to Rosenbaum and Rubin [48], the treatment assignment is independent of the observed covariates given the propensity score, $X \perp\!\!\!\perp T \mid p(x)$, if unconfoundedness (Assumption 2.2) holds. This Theorem is also known as the *Sufficiency of the Propensity Score.* Essentially it says that if the covariates are sufficient for adjusment then so is the propensity score.

We can use any generic supervised machine learning method or regression algorithm to estimate $p(x)$. Assuming that our propensity score estimate $\hat{p}(x)$ equals the true propensity score $p(x)$, we get

$$\begin{aligned}
\mathbb{E}\left[\frac{TY}{p(x)}\right] &= \mathbb{E}\left[\frac{TY(1)}{p(x)}\right] \quad \text{using Assumption 2.1} \\
&= \mathbb{E}\left[\mathbb{E}\left[\frac{TY(1)}{p(x)} \mid Y(1), X\right]\right] \\
&= \mathbb{E}\left[\frac{Y(1)}{p(x)}\,\mathbb{E}[T \mid Y(1), X]\right] \\
&= \mathbb{E}\left[\frac{Y(1)}{p(x)}\,\mathbb{E}[T \mid X]\right] \quad \text{using Assumption 2.2} \\
&= \mathbb{E}\left[\frac{Y(1)}{p(x)} p(x)]\right] = \mathbb{E}[Y(1)] \quad \text{using } p(x) = \mathbb{E}[T \mid X].
\end{aligned}$$

We can follow the same steps to show that $\mathbb{E}[\frac{(1-T)Y}{(1-p(x))}] = \mathbb{E}[Y(0)]$. Then we can again simply estimate the average causal effect $\tau = \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)]$. For a finite sample S, with $|S| = n$, this can be written, using propensity score estimate $\hat{p}$, as

$$\hat{\tau}_{IPW} = n^{-1} \sum_{i=1}^{n} \frac{T_i Y_i}{\hat{p}(x)} - n^{-1} \sum_{i=1}^{n} \frac{(1 - T_i) Y_i}{1 - \hat{p}(x)}.$$

Essentially, this gives a higher weight to instances that are underrepresented. This means that in the first summand, treated instances whose predicted probability of treatment is very low, are weighted very high. The second summand achieves the same for control instances, in reverse logic. Notice that we are dependent on the quality of our propensity estimate for this method to work. The propensity score model has to approximate the true assignment mechanism for the weighting to be meaningful. Thus, only if the chosen regression model equals the true regression, our propensity score weighted estimate recovers the true average treatment effect.

Despite this straightforward weighting method, we can use the propensity score to partition our data into strata, in which we can assume covariate-similarity or we can use 1:1 matching based on the propensity score, using greedy algorithms [36, 20]. In our evaluation we only consider propensity score weighting with generic base regressors as a baseline.

3.1.2 Conditional Mean Regression

Another straightforward method for estimating causal effects is conditional mean regression, also called *regression adjustment* [20]. Following [33], we discern between

S- and T-Learners. S- or Single-Learners use a *single* supervised machine learning technique or regression algorithm to estimate the combined response function

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

Let $\hat{\mu}$ be the estimator of μ. We can then estimate the conditional treatment effect as

$$\tau(x) = \hat{\mu}(x,1) - \hat{\mu}(x,0).$$

T-Learners on the other hand use *two* estimators to perform the same task. Essentially we split the dataset into treated and control and learn a outcome regression on each of the subsets. Thus, we learn

$$\mu_0(x) = \mathbb{E}[Y \mid X = x, T = 1],$$

using observations in the treated group. Similarly we learn

$$\mu_1(x) = \mathbb{E}[Y \mid X = x, T = 0],$$

using observations in the control group. Finally, we use the estimates $\hat{\mu_1}$ and $\hat{\mu_0}$ to get

$$2\hat{\tau}(x) = \hat{\mu_1}(x) - \hat{\mu_0}(x).$$

A key advantage of T-learners is that we can use different methods to estimate μ_1 and μ_0 and thus adapt to the structure of the data if necessary. Also, a notable side effect of using a S-Learner with Linear Regression is that it cannot capture heterogeneous effects because the coefficient α_T remains fixed for any predicted instance (see the results in Section 5.3.1 for more details).

Why does it work?

Consider the simple example where we use a linear regressor with the S-learner method and estimate

$$\mathbb{E}[Y \mid X, T] \approx \alpha_0 + \alpha_T + X^T \alpha_X.$$

Then, intuitively $\hat{\mu}(x,1) - \hat{\mu}(x,0)$ is equal to the coefficient α_T of the treatment indicator. This is to say that α_T is the average treatment effect. More formally and generalized, the efficacy of regression adjustment is based on the following derivation:

$$\begin{aligned}\mathbb{E}\left[\mathbb{E}[Y \mid X, T = 1]\right] &= \mathbb{E}\left[\mathbb{E}[Y(1) \mid X, T = 1]\right] \text{ using Assumption 2.1} \\ &= \mathbb{E}\left[\mathbb{E}[Y(1) \mid X]\right] = \mathbb{E}[Y(1)] \text{ using Assumption 2.2}\end{aligned}$$

However, that is only true if the linear regression is actually the true regression of the outcome on covariates and treatment. Also, as always, unconfoundedness and consistency constraints from Assumptions 2.2 and 2.1 must hold.

It's important to note that, just like propenstiy score weighting, outcome regression is dependent on the regression model we choose.

3.1.3 Doubly Robust Estimators

Having introduced *Inverse Propensity Score Weighting* and *Regression Adjustment*, the next natural step is to look at *Doubly Robust Estimators*. We've seen that both methods are dependent on the choice of the model. Thus, if we select the wrong model for our regression, we get wrong results. The core idea behind Doubly Robust Estimators (DRE) is to combine IPSW and T-Learners in a smart way, such that only one of two regression models has to be chosen correctly. Formally, the double robust estimator for finite samples is defined as

$$\begin{aligned}\widehat{\tau}_{DR} =& n^{-1}\sum_{i=1}^{n}\left[\frac{T_iY_i}{p(X_i)} - \frac{T_i - p(X_i)}{p(X_i)}\mu_1(X_i)\right]\\ &- n^{-1}\sum_{i=1}^{n}\left[\frac{(1-T_i)Y_i}{1-p(X_i)} + \frac{T_i - p(X_i)}{1-p(X_i)}\mu_0(X_i)\right].\end{aligned}$$

Following [11], we can split this term into an estimation for $\mathbb{E}[Y(1)]$ and $\mathbb{E}[Y(0)]$. The two seperate forms can then be written as the quantity they estimate plus a residual which must be zero for the estimator to work perfectly. We can derive

$$\mu_{DR,1} \approx \mathbb{E}[Y(1)] + \mathbb{E}\left[\frac{T-p(X)}{p(X)} \cdot (Y(1) - \mu_1(X)\right],$$

where $E[Y(1)]$ is the target quantity. For the second term to be zero, either the propensity score or the outcome regression has to match the true regression. As a user of a doubly robust estimator, we have to choose one model for the propensity and one for the outcomes. Only one of the two, however, must be specified correctly for the estimation to work. That is what is known as the doubly robust property, which is desirable in real world applications, as we can never be sure of the true regression and having two guesses open, increases our chances.

3.1.4 X-Learner

The last meta-learner we want to introduce here is the X-learner [33]. The approach taken is similar to that of the T-learner, but it is enhanced through reweighting. The authors themselves claim that the X-learner is particularly strong in cases where treatment and control group are of significantly different size. This is so, they argue, because the X-learner *"can use information from the control group to derive better estimators for the treatment group and vice versa."* [33].

Formally, we can again use any supervised machine learning method to estimate μ_0 and μ_1. The second step is then to impute the treatment effects for units in the treated and control group separately as

$$\begin{aligned}D_i^1 &:= Y_i^1 - \hat{\mu}_0(X_i^1) \text{ and}\\ D_i^0 &:= \hat{\mu}_1(X_i^1) - Y_i^0.\end{aligned}$$

We then use another pair of base learners to learn estimators $\hat{\tau}_1$ and $\hat{\tau}_0$ by using D_i^1 and D_i^0 as response variables. Finally, these estimators are combined in a weighted manner to yield

$$\hat{\tau}(x) = g(x)\hat{\tau}_0(x) + (1-g(x))\hat{\tau}_1(x),$$

with $g(x) \in [0,1]$ as a weight function. The authors themselves recommend using $g(x) = \hat{p}(x)$, a propensity score estimator. The X-Learner scores by means of a simple algorithm that is flexible enough to be used with all base learners but powerful enough to yield competitive results on many problems.

We use the implementation provided by S. Wager and X. Nie [41] with linear lasso as a base learner.[1]

3.1.5 R-Learner

The R-learner is not directly related to concepts of T- and S-learners. Rather, it approaches the task by using a neat reformulation of the problem statement that is better graspable for machine learning estimators. A detailed discussion and comparison is found in the proposal of X. Nie and S. Wager [41]. Here, we only briefly outline the novelty.

The essential background behind the R-learner is what they call *Robinsons Transformation*, which also gives name to the R-Learner [41]. Given the perfect response surfaces $\mu_0(x)$ and $\mu_1(x)$ and the true propensity score $p(x)$, they write

$$\mathbb{E}\left[\varepsilon_i\left(T_i\right)\left|X_i, T_i\right]=0, \text{ where } \varepsilon_i(t) := Y_i(t) - \left(\mu_0\left(X_i\right) + t\tau\left(X_i\right)\right).$$

The authors argue that, given this setup, they can and should rewrite the quantity of interest $\tau(x)$ in a new equation as

$$Y_i - m\left(X_i\right) = \left(T_i - p\left(X_i\right)\right)\tau\left(X_i\right) + \varepsilon_i(T_i),$$

where

$$m(x) = \mathbb{E}[Y|X=x] = \mu_0\left(X_i\right) + p\left(X_i\right)\tau\left(X_i\right).$$

Note that the quantities $m(x)$ a $p(x)$ are not known. Were they known, however, the treatment effect could be estimated by means of an empirical loss minimisation on a finite sample via

$$\hat{\tau}(\cdot) = \operatorname{argmin}_\tau \left\{\frac{1}{n}\sum_{i=1}^{n}\left(\left(Y_i - m\left(X_i\right)\right) - \left(T_i - p\left(X_i\right)\right)\tau\left(X_i\right)\right)^2 + \Lambda_n(\tau(\cdot))\right\}, \quad (3.1)$$

where Λ_n is a regularisation term on the complexity of the CATE function $\tau(x)$.

Thus, the estimation of a heterogeneous treatment effect functions comes down to a two step procedure, which they study in more detail in the paper:

1. Learn $\hat{m}(x)$ and $\hat{p}(x)$ via regression tuned for predictive performance.

2. Estimate treatment effects via the Equation 3.1

They summarise, *"the first step learns an approximation for the oracle objective, and the second step optimizes it."* [41].

[1] The code can be found here: https://github.com/xnie/rlearner.

3.2 Causal Forests

Causal Forests, as the name implies, are a recent adaption of causal trees [4] using the ideas from random forests [10]. Similar to nearest-neighbor methods, causal forests speficially and random forests in general aim to make predictions using information about *nearby* instances [60]. Other than classical distance metrics, however, random forests [10] can be understood as an adaptive neighborhood metric [64]. That is to say, they can weight the influence of dimensions depending on their significance to the target. As Wager and Athey [60] state, this is an important property in environments with many covariates or complex interactions among covariates. However, this adaptive consideration of features in the training data also bears a problem. Namely, trees tend to overfit the data they are constructed on, as spurious extreme values in the target are considered, while they may not be present in the general population [4].

Single decision trees alone are known to be unstable and exhibit a high variance due also to the fact that more splits with a normal objective of minimizing the MSE are usually better, but lead to overfitting. Using forests based on a large number of decorrelated trees solves this problem [32, p.14]. Also, Athey and Imbens [4], derive a splitting criterion that encourages low variance (bigger leaves), while the bias is reduced through a concept called *honesty*.

The main difficulty that causal forests have to capture compared to regression forests is what we have encountered before as the *fundamental problem of causal inference* [26]. Since usually training off-the-shelf regression models requires a known ground-truth (e.g. a class-label) we cannot use them directly for causal inference tasks. Athey and Imbens address this problem by constructing the trees with the objective to have heterogeneous leaves with regard to the treatment effect τ. The following construction works under the common assumptions (SUTVA, unconfoundedness, consistency).

The core idea is to split the covariate-space into similarity leaves minimizing the squared error loss $\sum_i (\hat{\mu}(X_i) - Y_i)^2$, which is equivalent to maximizing the variance of $\hat{\mu}(X_i)$ [60]. Within these leaves, the instances are assumed to behave as if coming from a randomized experiment [4].

3.2.1 Honest Trees for Treatment Effects

Before attempting the more difficult task of estimating causal effects, Athey and Imbens first consider the more common task of predicting population averages. This is easier as we have a ground truth form the training data. For sake of brevity, we skip this elaboration here and continue right away with the estimation of treatment effects.

Also skipping the detailed derivation of the splitting objective in [4], we present the formulas underlying the training of honest trees. First, let us consider a tree as a partition of the feature space $\mathcal{X}$ and write it as a set of leaves $\Pi = \{l_1, \dots, l_{|(\Pi)|}\}$. We denote by $l(x, \Pi)$ the leaf $l \in \Pi$ such that $x \in l$.

Given a tree Π the estimation of the population average outcome is

$$\mu(w, x, \Pi) := \mathbb{E}[Y_i(w) \mid X_i \in l(x, \Pi)],$$

and can be estimated over a sample $\mathcal{S}$ as

$$\hat{\mu}(w, x, \mathcal{S}, \Pi) := \frac{1}{|\{i \in \mathcal{S}_w | X_i \in l(x, \Pi)\}|} \sum_{i \in \mathcal{S}_w | X_i \in l(x, \Pi)} Y_i,$$

from which we can directly derive the treatment effect estimate

$$\hat{\tau}(x, \mathcal{S}, \Pi) := \hat{\mu}(1, x, \mathcal{S}, \Pi) - \hat{\mu}(0, x, \mathcal{S}, \Pi),$$

as a comparison of instances within a leaf. Now we introduce a goodness-of-fit metric, which is the mean squared error of the target τ, subtracted by the square of the target τ^2. This subtraction is used in the derivation of unbiased estimators of this metric. It does not affect the way the metric ranks estimators, but merely constitutes a mathematical trick to simplify the derivation. Formally,

$$\mathrm{MSE}_\tau \left(\mathcal{S}^{\mathrm{te}}, \mathcal{S}^{\mathrm{est}}, \Pi\right) := \frac{1}{\#\left(\mathcal{S}^{\mathrm{te}}\right)} \sum_{i \in \mathcal{S}^{\mathrm{te}}} \left\{ \left(\tau_i - \hat{\tau}\left(X_i; \mathcal{S}^{\mathrm{est}}, \Pi\right)\right)^2 - \tau_i^2 \right\},$$

where $\mathcal{S}^{te}$ and $\mathcal{S}^{est}$ are test and evaluation samples, respectively. $\mathrm{EMSE}_\tau(\Pi)$ is the expectation over the estimation and test samples defined as $\mathrm{EMSE}_\tau(\Pi) := \mathbb{E}_{\mathcal{S}^{\mathrm{te}}, \mathcal{S}^{\mathrm{est}}} \left[\mathrm{MSE}_\tau \left(\mathcal{S}^{\mathrm{te}}, \mathcal{S}^{\mathrm{est}}, \Pi\right)\right]$. This is then used as a criterion to find the best tree.

The central difficulty we encounter here is that we do not know τ_i and thus the MSE function remains infeasible. Using the unbiased estimator $\hat{\tau}$ and the fact that the estimator $\hat{\mu}$ is constant within any specific leaf, Athey and Imbens propose an estimator of EMSE_τ as

$$\begin{aligned} -\widehat{\mathrm{EMSE}}_\tau \left(\mathcal{S}^{\mathrm{tr}}, N^{\mathrm{est}}, \Pi\right) := & \frac{1}{N^{\mathrm{tr}}} \sum_{i \in \mathcal{S}^{\mathrm{tr}}} \hat{\tau}^2 \left(X_i; \mathcal{S}^{\mathrm{tr}}, \Pi\right) \\ & -\left(\frac{1}{N^{tr}} + \frac{1}{N^{est}}\right) \cdot \sum_{l \in \Pi} \left(\frac{S^2_{\mathcal{S}_1^{tr}}(l)}{p} + \frac{S^2_{\mathcal{S}_0^{tr}}(l)}{1-p} \right), \end{aligned}$$

where $S^2_{\mathcal{S}_1^{tr}}(l)$ is the within-leaf variance using the treatment group in the training sample $\mathcal{S}_1^{tr}$. The first term reflects the heterogeneity of treatment effects across the leaves and the second reflects the variance within leafs. The latter can also be understood as the uncertainty about the in-leaf treatment effects [39].

The conventional CART methods, on the other hand, use a so called *in-sample* goodness-of-fit criterion $\widehat{\mathrm{MSE}}_\tau \left(\mathcal{S}^{\mathrm{te}}, \mathcal{S}^{\mathrm{tr}}, \Pi\right) = \frac{1}{N^{\mathrm{tr}}} \sum_{i \in \mathcal{S}^{\mathrm{tr}}} \hat{\mu}^2 \left(X_i; \mathcal{S}^{\mathrm{tr}}, \Pi\right)$, which can also be approximated, but does not use a separate sample for validation. Also, it does not penalise for high variance [4].

The essential difference of honest causal trees to the CART approach is that different samples $\mathcal{S}^{tr}$ and $\mathcal{S}^{est}$ are used for constructing and evaluating a partition by its conditional mean [4]. This property is called honesty and ensures unbiased prediction.

Definition 3.2 (Honest Tree [60]). *A tree is honest if and only if, for each training sample i, it only uses the response Y_i to either estimate the within-leaf treatment* *effect τ or to decide where to place the splits, but never both. That is to say, each sample i is either in $\mathcal{S}^{est}$ to place the splits or in $\mathcal{S}^{te}$ to validate the split within $-\widehat{\text{EMSE}}_\tau$.*

Using the training sample for both construction and evaluation leads to over-fitting. The tree will use spurious extreme values (in our case variances) in the training data to construct the splits. These extreme relationships are, however, not present in the overall population from which the sample is drawn and thus the estimate is biased. Honest trees circumvent this problem by evaluating the constructed splits on a separate sample. This reflects a tradeoff between variance and bias, as we accept a higher variance due to smaller sample sizes. Athey and Imbens argue that the drawbacks are far outweighted by the benefits.

3.2.2 Inferring causal effect estimates

Given the tree Π, we expand the causal effect estimate $\hat{\tau}$ as

$$\begin{aligned}\hat{\tau}(x) = & \frac{1}{|\{i|T_i = 1, X_i \in l(x,\Pi)\}|} \sum_{\{i|T_i=1,X_i \in l(x,\Pi)\}} Y_i \\ - & \frac{1}{|\{i|T_i = 0, X_i \in l(x,\Pi)\}|} \sum_{\{i|T_i=0,X_i \in l(x,\Pi)\}} Y_i\end{aligned}$$

This implies that we think of the leaves as small enough to assume that the instances in one leaf are similar, which allows us to think of the instances within a leaf as if coming from a randomized experiment, for their covariates are similar while only their treatment varies.

3.2.3 From Trees to Forests

The natural next step, once the causal trees were established was for the authors to incorporate these trees into Causal Forests [60]. These bring two advantages: 1) they use the power of multiple trees to improve prediction 2) they remove the trade-off of using smaller samples. The first property is closely related to the motivation of classic random forests [10]. The second property is due to the fact that each individual tree must be honest, that is to say, it can only *use* half of the data for training. In an ensemble of trees however, each tree can consider a different split of the whole sample $\mathcal{S}$ such that eventually all data is used for constructing the forest.

Inferring the causal effect using a forest merely amounts to averaging the estimations of every single tree. Considering an ensemble of B trees with estimates $\hat{\tau}_b(x)$ we infer

$$\hat{\tau}(x) = B^{-1} \sum_{b=1}^{B} \hat{\tau}_b(x).$$

3.2.4 Overcoming the Unconfoundedness Assumption

Causal Forests are good with heterogeneous effects, because of the special criterion presented above. However, as the authors note, they struggle with observed confounding. To deal with this, Athey et al. [6] introduced a method to make causal forests more robust to confounding: *local centering.* Knaus et al. [32] find that Causal forests with *local centering* indeed perform better under confounding.

Local centering is a preprocessing step. To perform it, we write $y(x) = \mathbb{E}[Y_i|X = x]$ and $t(x) = \mathbb{E}[T_i|X = x] = p(x)$ for the conditional expectations of Y_i and T_i and $\hat{y}^{-i}(X_i)$, $\hat{t}^{-i}(X_i)$ are leave-one-out estimates of those expectations. We then compute $\tilde{Y}_i = Y_i - \hat{y}^{-i}(X_i)$ and $\tilde{T}_i = T_i - \hat{t}^{-i}(X_i)$ and run the causal forest procedure on $S = \{X_i\tilde{Y}_i, \tilde{T}_i\}_{i=1}^n$ instead of the original sample $S = \{X_i, Y_i, T_i\}_{i=1}^n$.

3.3 GANITE

The basic idea behind Generative Adversarial Nets for Inference of Individual Treatment Effects (GANITE) is to use Generative Adversarial Networks [19] to simulate the counterfactual distribution. More specifically, the authors see the counterfactual outcome as a missing label and they model the Generative Adversarial Nets (GAN) such that the descriminator has to decide which of the two given outcomes is the factual one. The following construction again works using Assumptions 2.2 and 2.3.

We denote in this section, for sake of readability, by φ the joint super-distribution in the potential outcomes framework (see Section 2.1). That is, φ is the joint distribution on $X \times \mathbf{T} \times (Y(1), Y(0)$, where $\mathbf{T} = \{0, 1\}^2$ is a unit vector that is one at precisely one index, denoting whether treatment or control were given. E.g. $\mathbf{T}_i = (0, 1)$ means that treatment has been given, because $\mathbf{T}_i$ is one at index one. $\mathbf{Y} = (Y(1), Y(0))$ is the outcome vector of which only one is observed and the other is counterfactual. The construction of $\mathbf{T}$ and $\mathbf{Y}$ might seem convoluted for the simple binary treatment setting, but is motivated by the fact that the mathematical derivation works similarly for a setting with multiple treatments. Finally, we denote by φ_X the marginal distribution of X and by $\varphi_{\mathbf{Y}}(x)$ the conditional distribution of $\mathbf{Y}$ given $X = x$. Notice that we slightly change our notation compared to Section 2.1 to allow better handling in the formulas below.

The final goal of GANITE is to estimate an individual treatment effect τ_i for any given instance. The performance of this estimation $\mathbf{I}(x)$ is measured using criteria introduced in [25], which is the expected Precision in Estimation of Heterogeneous Effects

Definition 3.3. *(Expected PEHE [25]) The Expected PEHE error ϵ_{PEHE} is defined over the true distribution as*

$$\epsilon_{PEHE} = \mathbb{E}_{\mathbf{x}\sim\varphi_{\mathbf{X}}}\left[\left(\mathbb{E}_{\mathbf{Y}\sim\varphi_{\mathbf{Y}(\mathbf{x})}}\left[Y(1) - Y(0)\right] - \mathbb{E}_{\hat{\mathbf{Y}}\sim\mathbf{I}(\mathbf{x})}\left[\hat{Y}(1) - \hat{Y}(0)\right]\right)^2\right]. \tag{3.2}$$

which is approximated in finite samples S, with $|S| = n$, as

$$\epsilon_{PEHE} = n^{-1}\sum_{i=1}^{n}\left([Y_i(1) - Y_i(0)] - \left[\hat{Y}_i(1) - \hat{Y}_i(0)\right]\right)^2.$$

3.3.1 Two generators and two discriminators

The first step of the problem, as the authors divided it, is to generate the counterfactual samples $\tilde{y}_{cf}$ from distribution $\varphi_{Y(cf)}(x, \mathbf{t}, Y(f))$, where $Y(f)$ and $Y(cf)$

denote factual and counterfactual outcomes respectively. These counterfactual samples are then combined with the observed data to obtain a complete dataset $\tilde{\mathcal{S}} = \{X_i, \mathbf{T}_i, \mathbf{Y}_i\}_1^n$, which is used to train a second generator on the individual treatment effects. The layout is depicted in Figure 3.1.

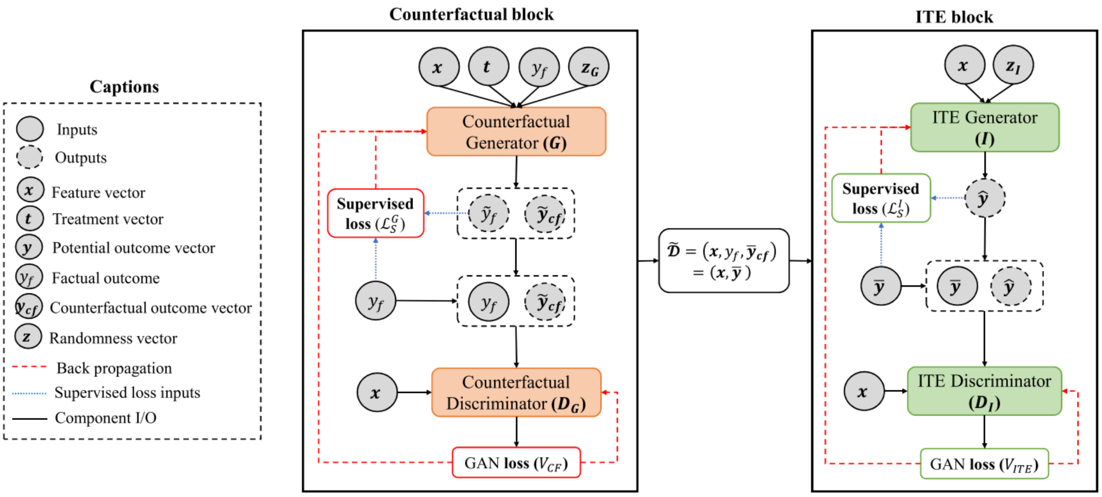

Captions
Inputs
Outputs
Feature vector
Treatment vector
Potential outcome vector
Factual outcome
Counterfactual outcome vector
Randomness vector
Back propagation
Supervised loss inputs
Component I/O
Counterfactual block
Counterfactual Generator (G)
Supervised loss ($\mathcal{L}_S^G$)
Counterfactual Discriminator (D_G)
GAN loss (V_{CF})
$\tilde{\mathcal{D}} = (x, y_f, \bar{y}_{cf}) = (x, \bar{y})$
ITE block
ITE Generator (I)
Supervised loss ($\mathcal{L}_S^I$)
ITE Discriminator (D_I)
GAN loss (V_{ITE})

Figure 3.1: Diagram showing the layout of the GANITE method, taken from [65]. We see the two blocks $\mathbf{G}$ and $\mathbf{I}$, the combined loss functions with V_{CF} and $\mathcal{L}_S^G$ as well as V_{ITE} and L_S^I.

The first generator $\mathbf{G}$ is defined as random variable depending on observed covariates x, treatment indictation vector $\mathbf{t} \in \mathbf{T}$ and factual outcome $Y(f)$. The goal is then to find a function g such that $G(x, \mathbf{t}, Y(f)) = g(x, \mathbf{t}, Y(f), \mathbf{z_G}) \sim \varphi_{\mathbf{Y}}(x, \mathbf{t}, Y(f))$, where $\mathbf{z_G} \sim \mathcal{U}(-1, 1)$ is uniform random variable. We denote, following the authors, by $\tilde{\mathbf{y}}$ a sample from $\mathbf{G}$ and by $\bar{\mathbf{y}}$ the outcome vector when we replace the generated factual outcome with the observed factual outcome in $\tilde{\mathbf{y}}$.

The corresponding discriminator $\mathbf{D_G}$ is modelled differently than in the usual GAN framework. There, the discriminator is given a sample and has to decide from which of two distributions (the true and the generated) it came from. Instead $\mathbf{D_G}$ maps pairs $(\mathbf{x}, \bar{\mathbf{y}})$ to vectors in $[0, 1]^2$, denoting the probability that the first or second component in $\bar{\mathbf{y}}$ is the factual outcome. Note that the construction in [65] works for and is in fact designed for multiple treatments. For our purposes, however, we stick with binary treatment.

Following the general idea of GAN[19], the discriminator is trained to maximize the probability of correctly identifying the factual outcome in the vector y while the generator is trained to minimize that probability. This idea is captured in the classic minimax game in the formula:

$$\min_{\mathbf{G}} \max_{\mathbf{D_G}} \mathbb{E}_{(x,\mathbf{t},Y(f))} \left[\mathbb{E}_{\mathbf{z_G}} \left[\mathbf{t}^T \log \mathbf{D_G}(x, \bar{\mathbf{y}}) + (\mathbf{1} - \mathbf{t})^T \log (\mathbf{1} - \mathbf{D_G}(x, \bar{\mathbf{y}})) \right] \right], \tag{3.3}$$

where log is perfomed element-wise and $\varphi_{Y(f)}$ is the factual distribution on which the generator works. Practically, this objective is optimized iteratively using mini-batches (See [65, Section 4] for details).

Using the generator $\mathbf{G}$ we generate the partially synthetic dataset $\tilde{S}$ by augmenting the observed outcomes with generated counterfactuals. This is then passed to the second block with generator $\mathbf{I}$ and discriminator $\mathbf{D_I}$.

The ITE generator $\mathbf{I}$ uses only the feature vector $X = x$ to generate a potential outcome vector $\hat{\mathbf{y}}$, similar to the method used for $\mathbf{G}$, we search for a function h such that $I(x) = h(\mathbf{x}, \mathbf{z_I}) \sim \varphi_\mathbf{Y}(x)$, where $\mathbf{z_I}$ induces randomness. Note that we consider $\mathbf{G}$ and $\mathbf{I}$ to be random variables.

The corresponding discriminator $\mathbf{D_I}$ now works like the standard conditional GAN discriminator, because we have access to our *complete* dataset $\tilde{S}$. It takes a pair $(x, \mathbf{y}^*)$ and returns the probability of $\mathbf{y}^*$ being drawn from the dataset $\tilde{S}$ instead of from the generator $\mathbf{I}$. The stereotypical optimization game is

$$\min_{\mathbf{I}} \max_{\mathbf{D_I}} \mathbb{E}_{\mathbf{x} \sim \varphi \mathbf{X}} \left[\mathbb{E}_{\mathbf{y}^* \sim \varphi_\mathbf{Y}(\mathbf{x})} \left[\log \mathbf{D_I} \left(\mathbf{x}, \mathbf{y}^* \right) \right] + \mathbb{E}_{\mathbf{y}^* \sim \mathbf{I}(\mathbf{x})} \left[\log \left(\mathbf{1} - \mathbf{D_I} \left(\mathbf{x}, \mathbf{y}^* \right) \right) \right] \right] \tag{3.4}$$

3.3.2 Incorporating two loss functions

Having set the scenario, the authors in [65] introduce the loss functions used in the minibatch optimization. Namely, for each generator $\mathbf{G}$ and $\mathbf{I}$, they introduce two loss functions. One of which models a supervised loss and one of which models the loss implied by the respective descriminator. Formally, we define

$$V_{CF}(\mathbf{x}, \mathbf{t}, \overline{\mathbf{y}}) = \mathbf{t}^T \log \big(\mathbf{D_G}(\mathbf{x}, \overline{\mathbf{y}}) \big) + (\mathbf{1} - \mathbf{t})^T \log \left(\mathbf{1} - \mathbf{D_G}(\mathbf{x}, \overline{\mathbf{y}}) \right)$$

as the objective from the minimax problem in Equation 3.3. Aditionally we define

$$\mathcal{L}_S^G \left(Y(f), \tilde{\mathbf{y}}(f) \right) = \left(Y(f) - \tilde{\mathbf{y}}(f) \right)^2,$$

which aims to enforce that the generated outcome sampled from $\mathbf{G}$, for the observed treatment equals the observed outcome. This is important as $\mathbf{G}$ generates the whole potential outcome vector $\tilde{\mathbf{y}}$ not only the counterfactual outcome. Although we eventually replace the generated factual outcome with the observed factual outcome in $\bar{\mathbf{y}}$, we want the generator to take into account the correct distribution of factual outcomes. Together this yields the two objectives for the counterfactual block:

$$\begin{aligned} &\min_{\mathbf{D_G}} - \sum_{k_C} V_{CF}(\mathbf{x}, \mathbf{t}, \overline{\mathbf{y}}), \\ &\min_{\mathbf{G}} \sum_{k_G} \left[V_{CF}(\mathbf{x}, \mathbf{t}, \overline{\mathbf{y}}) + \alpha \mathcal{L}_S^G \left(Y(f), \tilde{\mathbf{y}}(f) \right) \right], \end{aligned}$$

where $\alpha \geq 0$ is a hyper-parameter. Similarly we define for the ITE block the objective from the minimax competition as

$$V_{ITE}(\mathbf{x}, \overline{\mathbf{y}}, \hat{\mathbf{y}}) = \log \left(\mathbf{D_I}(\mathbf{x}, \mathbf{y}) \right) + \log \left(1 - \mathbf{D_I}(\mathbf{x}, \hat{\mathbf{y}}) \right).$$

And likewise the authors add a supervised loss to ensure generated potential outcomes from $\mathbf{I}$, $\hat{\mathbf{y}}$ closely match the augmented potential outcomes $\overline{\mathbf{y}}$ with respect to their ITE:

$$\mathcal{L}_S^I(\overline{\mathbf{y}}, \hat{\mathbf{y}}) = \left((\overline{\mathbf{y}}(1) - \overline{\mathbf{y}}(0)) - (\hat{\mathbf{y}}(1) - \hat{\mathbf{y}}(0)) \right)^2.$$

Combined we get the objective functions for optimization:

$$\min_{\mathbf{D_I}} - \sum_{k_I} V_{ITE}(\mathbf{x}, \overline{\mathbf{y}}, \hat{\mathbf{y}}),$$
$$\min_{\mathbf{I}} \sum_{k_G} \left[V_{ITE}(\mathbf{x}, \overline{\mathbf{y}}, \hat{\mathbf{y}}) + \beta \mathcal{L}_S^I(\overline{\mathbf{y}}, \hat{\mathbf{y}}) \right].$$

For the implementation we use the source code provided by [53], which is build using TensorFlow. We use 200 epochs for training both **G** and **I**. While the training is stable, the results returned by the model are not expected and in fact, unusable. Since the authors refuse to provide their own project code and reimplementation is beyond the scope of this work, we are bound to leave GANITE out of the evaluation.

3.4 Neural Representation Learning

Lastly, we want to introduce another modern machine learning approach to the problem of causal effect estimation. We briefly outline the idea and underlying thought processes before introducing a specific method that uses them: DragonNet.

3.4.1 Balancing Representations

The idea of a representation learning approach to counterfactual inference, i.e. causal effect estimation, was first introduced in [30]. The authors argue that the *factual distribution* P^F from which we draw our samples $\{X_i, T_i, Y_i\}$ might not be the same as the *counterfactual distribution* P^{CF}, which would contain the samples $\{X_i, 1 - T_i, Y_i^{cf}\}$ and is not observable. If we transfer this situation to machine learning terms, one could say that the training sample, in our case drawn from P^F, has a different distribution than the test sample P^{CF}. Johansson et al. point out that this is a form of *covariate shift*.

In order to tackle this problem, they propose to preprocess the covariates by learning a representation $\Phi : \mathcal{X} \to \mathbb{R}^d$, which is then plugged into a function $h : \mathbb{R}^d \times \{0, 1\} \to \mathbb{R}$, to predict outcomes. In words, the learned representation trades off three objectives:

1. enabling low-error prediction of the observed outcome,
2. enabling low-error prediction of unobserved counterfactual outcomes under consideration of relevant factual outcomes,
3. ensuring a similar or balanced distribution of both treatment groups.

Formally, these objectives are expressed in

$$B_{\alpha,\gamma}(\Phi, h) = \frac{1}{n} \sum_{i=1}^{n} \left| h\left(\Phi\left(x_i\right), t_i\right) - y_i^F \right| + \tag{3.5}$$

$$\alpha \operatorname{disc}\left(\hat{P}_\Phi^F, \hat{P}_\Phi^{CF}\right) + \frac{\gamma}{n} \sum_{i=1}^{n} \left| h\left(\Phi\left(x_i\right), 1 - t_i\right) - y_{j(i)}^F \right| \tag{3.6}$$

where $\alpha, \gamma > 0$ are hyperparameters to *"control the strength of the imbalance penalties"* [30]. The operator disc refers to a measure of disrepancy that is elaborated on in the paper. The second objective is represented in the term $\left|h\left(\Phi\left(x_i\right), 1-t_i\right)-y_{j(i)}^F\right|$ where $y_{j(i)}^F$ stands for the outcome of the nearest counterfactual neighbour of unit i in the metric space $\mathcal{X}$.

Uri Shalit et al.[54] now use a very similar approach in their generalizating work. While in [30], the function h gets both treatment and representation as input, they here use two different networks h_1 and h_0 to represent the two potential outcomes. They argue that when the dimension d of the representation is high, the influence of treatment might get lost. The resulting objective they minimize by means of end-to-end training is very similar to the one from Johansson et al [30]. It contains a measure of imbalance in the representation space as well as a measure of predictive accuracy. The network architecture that results from their considerations is shown in Figure 3.2.

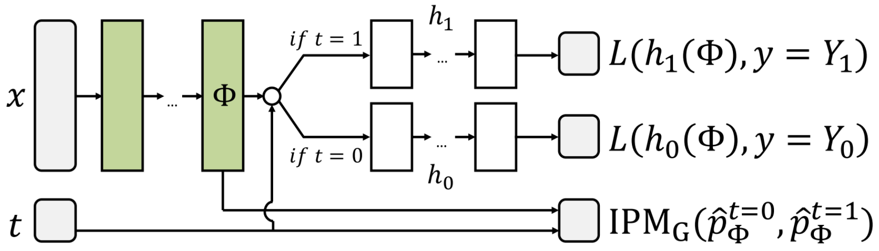


Figure 3.2: The network architecture introduced in [54]. IPM_G refers to a distance metric for probability distributions. In this case, the distribution between treated and control group are compared in the representation space. h_1 and h_0 are the two separate heads for the two potential outcomes. During training, only one of each heads is updated for each unit i, depending on t_i. Φ refers to the representation learned by the neural network on the left.

3.4.2 DragonNet

DragonNet [56], which is a conceptually simpler approach, is similar in architecture to the neural network from [54], but theoretically based on another consideration. Namely, the *Sufficiency of the Propensity Score.* This theorem says that if the treatment effect τ is identifiable by adjusting for X (Assumptions 2.2 and 2.3) then it is sufficient to adjust for the true propensity score $p(X)$ (see [56] and [48, Theorem 3]).

The objective of DragonNet is to learn a representation $Z(X)$ that is good for prediction of the outcomes but also good for prediction of the propensity score. The intuition is that the close link to the propensity score objective will *convince* the representation network to use mostly features that have an influence on treatment assignment, i.e. confounders. Considering features that only influence the outcome, the authors argue, hurts the estimation of the causal effect.

In practice, this is achieved by training a neural network with multiple heads. Namely we have a propensity prediction head $g(X)$ and two outcome prediction

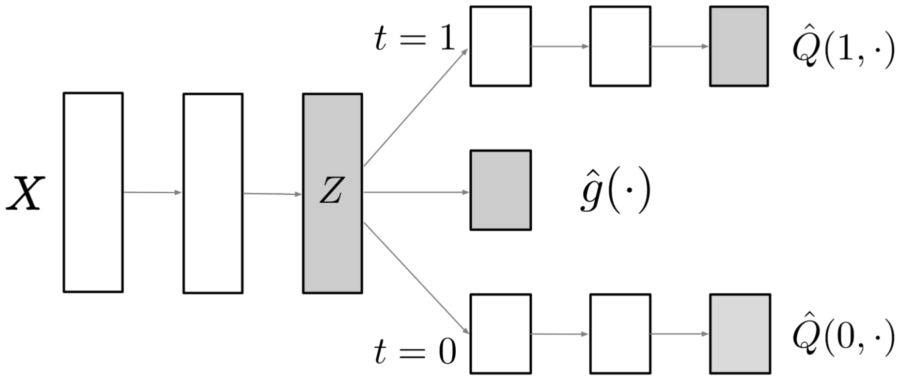


Figure 3.3: The multi-headed DragonNet architecture from [56]. Similar to Figure 3.2, we have two heads for the potential outcomes, here denoted by $\hat{Q}(1, \cdot)$ and $\hat{Q}(0, \cdot)$. Again, only one head is updated for each unit. Additionally, the propensity score head $\hat{g}(\cdot)$ is added to enforce a representation that allows good treatment prediction.

heads $Q(1, X)$ and $Q(0, X)$, that were refered to in [54] as h_1 and h_0. Figure 3.3 shows the multi-headed architecture of the DragonNet.

We train these estimators using a combined loss function on a finite sample of n instances

$$\arg\min_{Q,g} = n^{-1} \sum_i \left((Q(T_i, X_i) - Y_i)^2 - \texttt{CrossEntropy}(g(X_i), t_i) \right).$$

Similar to the T-learner in Section 3.1.2 we can then estimate the treatment effect as

$$\hat{\tau}(x) = Q(1, x) - Q(0, x).$$

The rationale behind the architecture is to have representation network with a high capacity that can learn a meaningful representation of the covariates based on the objectives of outcome and propensity estimation. The authors argue and show empirically that the strong incorporation of the propensity score leads the network to *ignore* elements of X that are not predictive of the treatment. This is counter-intuitive for the prediction of $Y(1)$ and $Y(0)$, but aids the true prediction of the treatment effect as we have seen in Section 2.1.

We train DragonNet over 100 epochs using the default configuration proposed by the authors[2]. Specifically, they use three fully connected layers with 200 units for the representation network. The function $g(x)$ is implemented as a single linear layer with a sigmoid activation and the two potential outcome heads have three fully connected layers with 100 units each. If not specified otherwise, an Exponential Linear Unit (ELU) is used for activation.

[2] The code can be found here https://github.com/claudiashi57/dragonnet

3.5 Other Methods

The following are a few methods which we do not introduce in detail, but still use in the evaluation in Chapter 5. We shortly introduce their respective ideas and refer to the original papers for a comprehensive discussion.

3.5.1 Pollinated Transformed Outcome Forests

The first method proposed in [47] is based on simple off-the-shelf regression forests which are trained on a propensity weighted outcome (referred to as the transformed outcome). To alliviate the burden of high variance that comes with this propensity adjustment, the authors introduce the method of *'pollination'*. It entails that they replace the estimates of the tranformed outcome in each leaf with the difference in conditional means between treated and untreated instances in that leaf.

3.5.2 Causal Boosting

Causal Boosting [47] uses *Causal Trees*, which we've introduced in Section 3.2. Instead of combining them into a Causal Forest, however, gradient boosting for least squares is used [17].

3.5.3 Causal MARS

Finally, the the authors of [47] argue that tree based methods always come with potentially high bias, due to their reliance on the within-leaf ATE estimates. Subsequently they propose the *Causal MARS* method, based on Multivariate Adaptive Regression Splines (MARS), which are said to reduce the inherent bias problem of tree-based methods.

4. Design of an Evaluation Framework

> Fundamentally, there is no way to know what we don't know before we know it.
>
> Alejandro Schuler [50, 52, 47] upon the question, what he thinks about evaluation of causal effect estimators.

With the foundations in place, we have introduced a wide spectrum of applicable methods for causal effect estimation in Chapter 3. Naturally, the question arises, how these methods fare against each other. Or, in other words, how can we select the best model for our task? In order to make such statements, we first have to consider some obstacles. Namely, we want to shed light on the inherent problem in evaluating causal effect estimators that is derived from the FPCI. Subsequently, we will look at existing benchmark datasets with their underlying assumptions. Finally, we introduce interesting dimensions of adjustment for a synthetic data generation process (DGP) based on the existing work. We use these dimensions to design experiments for specific hypothesis in Chapter 5.

4.1 The Problem of Evaluation

We have seen in the introduction that the Fundamental Problem of Causal Inference [26] makes causal inference without further assumptions an impenetrable task. For any instance we only ever observe one of two potential outcomes. The effect we are trying to estimate, however, is exactly the difference between the two potential outcomes. Thus, evaluating an estimator requires us to know a quantity that is never observable for real data. To put it differently, we will never have a ground truth. At best, we have a RCT of the same phenomenon to give us a good approximation of the average treatment effect [51, 47].

To solve the inherent problem of causal effect benchmarks, several partially synthetic and completely synthetic datasets have been introduced. The idea is simple. We use existing data and synthetically create outcomes for which we know the individual treatment effect by construction. This approach, which we will shed more light on in Section 4.3, also comes with significant drawbacks. Specifically:

- The data generating processes of fully synthetic datasets are often quite simplistic (e.g. in [37, 64, 35]) and thus a questionable representation of reality.
- The DGPs introduce assumptions as to the structure of the problem. For example, the number of confounders or the function used to determine the treatment effect based on covariates. We show in Section 5.3.1 that the performance of methods can only be judged for a specific problem setting. Thus limiting the claim of generality of any reference dataset.
- In practice, there exists a confusing array of different versions of reference datasets due to slightly different simulations. (e.g. [35] vs [64] vs. [37] for the TWINS data or [54] vs [40] for IHDP[1]). This leads some authors to copy the results instead of validating the numbers themselves.[2]
- There is no single established procedure for measuring results. Different metrics are used for the same dataset, different numbers of replicas, and so forth. A standard procedure like held out k-fold cross validation for supervised learning would be desirable.

We now elucidate the approaches introduced to circumvent the inherent problem of evaluating causal treatment effects. Also, we elaborate on the problems outlined above.

4.2 Existing Benchmark Datasets

While the respective literature is widely spread throughout many different disciplines, a few reference datasets have emerged, that are used to compare and rank causal estimation methods. Here, we show their background and their assumptions with respect to the fundamental problem explained above.

4.2.1 Infant Health Development Program

The Infant Health Development Program (IHDP) provides a dataset that has first been used by Jennifer L. Hill in [25] to evaluate causal estimators. The original study was constructed to study the effect of special child care for low birthweight, premature infants. In total, six continuous and 19 binary pretreatment variables

[1] V. Dorie points out that the script used to generate IHDP evaluation data is probably faulty: *"The only complication is that R changed how the sample function works, so you'll need to use* `RNGkind` *to set* `sample.kind` *to Rounding if you want to get identical results. [...] I should really take it down [...]"*

[2] Citing a personal communication with the lead author of [65], Jinsung Yoon: *"I also got some different results when I used their code for reproducing the results. However, there may be a chance that the reviewers complains that I did not use their reported results. Therefore, I just copied and pasted their results in their paper"*

were collected in a randomised control trial. Using the covariates of all instances in both treatment groups, the potential outcomes are generated synthetically as decsribed in Equation 4.1. Finally, Hill simulates an observational study by omitting a non-random set of samples from the treatment group. The way the subset is generated from the experimental data does not ensure complete overlap [25].

Specifically, the observational subset is created by throwing away the set of all children with nonwhite mothers from the treatment group. Thus, for computation of conditional average treatment effects on the treated (CATT), the overlap condition is met, while this not the case for the control group.

After the adaptions from Hill, we are left with 139 instances in the treated group and 608 instances in the control group. For these, Hill then proposes two different response surfaces, of which we use the one most commonly used by others, referred to in the paper as setting B[3]. The following elaboration on the data generation is taken from [25].

The potential outcomes are defined as

$$Y(0) \sim \mathcal{N}(\exp(X + W) \cdot \beta_B, 1) \text{ and} \tag{4.1}$$

$$Y(1) \sim \mathcal{N}(X\beta_B - \omega_B^s, 1), \tag{4.2}$$

where X represents the standardised covariate matrix, W is an offset matrix with the dimensions of X and all values set to 0.5 and finally ω_B^s is chosen such that the mean CATE is 4. The entries of the coefficient vector β_B are sampled from the values $(0, 0.1, 0.2, 0.3, 0.4)$ with probabilites $(0.6, 0.2, 0.1, 0.1, 0.1)$ respectively. This results in a nonlinear response with heterogeneous treatment effect.

IHDP is used in [35, 65, 54, 64, 30, 56, 53, 40].

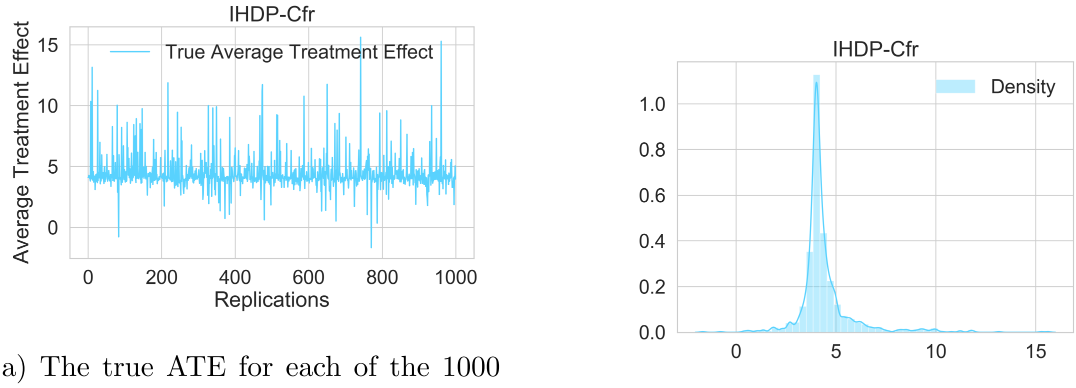
4.2.1.1 IHDP Replication

Using the instructions provided in the aforementioned papers, we generate 1000 replications of this DGP using the script provided by Vincent Dorie[4]. To make sure the data generated is equal to the ones used by other authors, we contacted Christos Louizos [37] to get their set of 1000 replications. The Figures 4.1 and 4.2 show, however, that the true ATE has a significantly different distribution over the 1000 simulations. Thus, we cannot expect to reproduce results using our own generated data. Considering the intricacy of the script used for generation and the potentially unknown changes in the `R` internal handling of random sampling, pinpointing the error is tedious if not impossible. The author himself pointed out that he cannot guarantee the results of the script to be correct as it is not maintained anymore.

For these reasons, we resort to using the replications provided by Christos Louizos, as they have been used in [30] and [37], where the same results as in [54] are reported.

[3] Fredrik D. Johansson [30] clarified via mail that the setting used in their implementation of IHDP refers to setting B in [25], while the implementation by V. Dorie refers to it as setting A. Johansson: *”The setup that we want to emulate is Jennifer Hills “B” setting in the original paper, but that is implemented as “A” in NPCI (confusingly enough). This is a typo in our paper which we should fix if we have the opportunity.”*

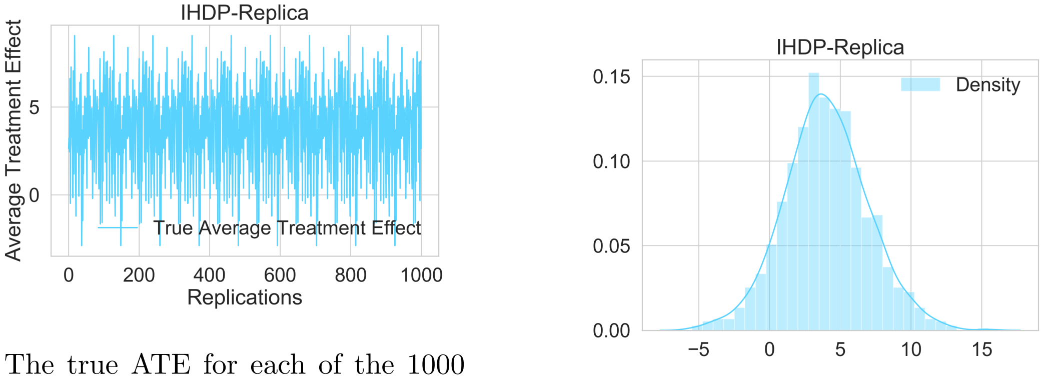
[4] The script was retrieved from GitHub (https://github.com/vdorie/npci) on August 18th 2019.



(a) The true ATE for each of the 1000 replications

(b) The true ATE distribution

Figure 4.1: 1000 replications IHDP dataset provided by Christos Louizos [37].



(a) The true ATE for each of the 1000 replications

(b) The true ATE distribution.

Figure 4.2: 1000 replications of the IHDP dataset we generated using the exact instructions from [54] and [30]

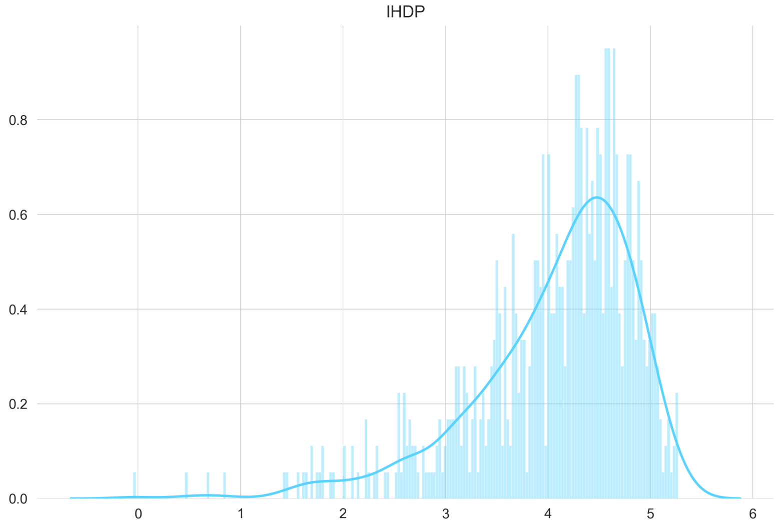


Figure 4.3: Distribution of the individual treatment effects for one replication of the IHDP dataset provided by C. Louizos. Most notably, the distribution is almost gaussian around the ATE and does not show clear heterogeneous groups.

4.2.2 Twins

The Twins dataset is derived from birth data collected in the US between 1989 and 1991. The original data is compiled and analyzed by Almond et al. [3]. From all these births, only the twins are considered, because these allow us to perform a special kind of synthetic generation. Namely, we only consider twins with a low birthweight and then define treatment $T = 1$ as being the heavier twin. In doing so, we follow other authors that have used the Twins dataset for comparisons [65, 64, 37].

This construction means that we know the outcome, mortality in this case, for both treatments. The only synthetic part in the data is the assignment of treatment. From the full data containing both potential outcomes, we want to generate data that resemble an observational study. As mentioned above, this is done via different functional relationships in different papers. For our purposes we present the process described in [65].

First, we only use low birthweight twin pairs for which all 30 features are available. That leaves us with 8215 samples. Since we know the mortality for both twins, we know the ground truth. To generate a observational study we now assign treatment by defining

$$\begin{aligned} P(T \mid X) &\sim \textsc{Bern}(\sigma(W^T X + n)), \text{ with} \\ W^T &\sim \mathcal{U}((-0.1, 0.1)^{30\times 1}) \text{ and} \\ n &\sim \mathcal{N}(0, 0.1), \end{aligned}$$

where σ is the sigmoid function, Bern refers to the Bernoulli distribution, $\mathcal{U}$ to a uniform distribution and $\mathcal{N}(0, 0.1)$ to a normal distribution with mean 0 and standard deviation of 0.1.

Twins is used in [35, 64, 65, 37].

4.2.3 Atlantic Causal Inference Challenge - ACIC

The dataset used for the Atlantic Causal Inference Challenge (ACIC)[5] 2018 is based on the Linked Births and Infant Deaths Database (LBIDD), a set of de-identified clinical records made available by The National Vital Statistics System (NVSS) and The National Center for Health Statistics (NCHS). A detailed description of the the covariates and the underlying study can be found in [14].

Similar to the procedure in IHDP the authors of the challenge [57] use the real covariates and generate synthetic outcomes. Additionally, in this case, the treatment assignment is also generated. Unfortunately, the exact DGP underlying the data is not available. This is due to the fact, that the data was part of a public causal inference challenge and thus publishing the underlying assumptions of the DGP would simplify the estimation, as we've seen in Section 3.1.1. The authors do, however, provide many different versions of the data, generated from different DGPs. For example, they modify the number of confounders or the degree of the polynomial used to link covariates and outcome or covariates and treatment. Since both treatment and outcome are generated synthetically, $Y_i(0)$ and $Y_i(1)$ are known for all instances and we can evaluate the performance of estimators.

[5] The challenge data and further explanations of the corresponding data can be found here: https://www.synapse.org/#!Synapse:syn11738767/wiki/512854

4.3 Data Generating Processes

Inspired by the rigid formalisation of a DGP in [22] and the parametric approach in [57], we now look in more detail at the generating processes behind benchmarking datasets. As mentioned above, the DGPs can be categorised in three major classes. There are real processes, partly synthetic processes and fully synthetic processes. Fully synthetic DGPs create covariate, treatment and outcome data from scratch. Partly synthetic processes, like the ones used in Ihdp, Acic and for Empirical Monte Carlo Study (EMCS) in general, create only treatment and/or outcome assignment and use real covariates. At last, real processes refer to the observational or experimental measurement of natural phenomena through a study. Such data are hard to come by for evaluation purposes because of the fundamental problem of causal inference. There is simply no way to know the causal effects of real phenomena, especially, when we are concerned with individual treatment effects.

The fully synthetic approach, on the other side of the spectrum, is rarely taken because of its dependence on parameter choices. However, Powers et al. [47] argue, synthetic data allow us to focus on a specifically constructed problem. A first step away from the simplicity of fully synthetic data is given with the idea of *Structured Design* [2]. Essentially, the idea is to create synthetic distribution which uses parameters estimated from the original, real data [2, Section 2.2]. The next step is that used for the IHDP and ACIC datasets, which has been called EMCS in [32].

4.3.1 Empirical Monte Carlo Studies

The idea of an EMCS is to use as many components of the final DGP as possible from real data [32]. Knaus et al. [32, Table 3] introduce a specific procedure for the creation of such a study based on collected real world data. Essentially, their specific approach is to omit all treated instances, such that for all remaining samples $Y_i(0)$ is known. They then generate a treatment effect $\tau(x_i)$ for all instances based on a sophisticated concatenation of a sinus function, normalization and a discrete choice based on constraints of the desired outcome. They argue that their ITE is *”highly non-linear and complicated”* [32, p. 20].

More generally, EMCS describes an approach to create evaluation studies for causal effect estimation. They can thus be used for meaningful evaluation of estimators. The specific procedure by which simulated outcomes are added to the real data varies, like we’ve seen with Ihdp and Twins.

At the same time, Advani et al. argue in [2] that these procedures have to be considered carefully. In their study they theoretically show that for both *Placebo Design* and *Structured Design*, the evaluation can be designed in a way that it does not yield meaningful results. They show that the simulation can not help to choose the best estimator for a task. A problem that Alejandro Schuler [50] considered in detail.

Still, with our current understanding, the evaluation of an EMCS best matches real life performance compared to fully synthetic data. Accordingly, we continue to use the approach taken by Knaus et al. for most of our evaluation. In Section 7.2, we give a brief outlook on methods proposed to improve the real life applicability of evaluation studies.

4.3.2 Design Dimensions of a DGP

In the wider context of Empirical Monte Carlo Studies there are many possible dimensions to consider when designing the distributions of treatment and outcome. Here, we introduce some of the degrees of freedom that one has and show how they possibly affect the difficulty of the DGP. In the discussion of our results in Chapter 5 we come back to these and show their effects on the performance of estimators.

During the research for this thesis, we've discovered ambiguous descriptions of data used in evaluations. This concerns the degree of detail with which the used data is described and also the terminology for describing it. Specifically, the IHDP and TWINS datasets are often used with a lack of insight as to how their covariates, treatment and outcome are structured. It is this observation that motivated us to further elaborate on the possible dimensions one can tweak when generating synthetic and partially synthetic data. As further discussed in Chapter 7, we aim to point towards the more thoughtful description and discussion of DGPs in causal effect estimation, as thoughtfully exercised by the authors of the ACIC challenges [22].

Refreshing the notation from Section 2.1, we consider how our covariates, X, outcomes $Y(1), Y(0)$ and Y_{obs} and treatment T are structured and how they depend on each other. Following the formalisation in [22] and [32], we aim to identify an, yet incomplete, set of parameters to tweak the following dimensions. Simplifying notation, we assume that any unit i can be identified by it's set of covariates X_i, thus $\tau_i = \tau(X_i)$ refers to the treatment effect of unit i identified by X_i.

If not stated otherwise we assume the potential outcomes to follow a certain structure introduced in [22] or, much more precisely, to be expressed in a certain form:

$$\mathbb{E}[Y_i \mid X_i = x_i, T_i] = \mu(x_i) + \tau(x_i) \cdot T_i \tag{4.3}$$

That is to say, the expected value of the outcome is defined by the base outcome plus the treatment effect in case $T_i = 1$. For the individual case we can generally write

$$Y_i = \mu(x_i) + \tau(x_i) \cdot T_i + \epsilon_i, \tag{4.4}$$

including an unspecified error term ϵ_i that is centred around zero.

The following dimensions are summarised in Table 4.1.

Covariates

The starting point for most DGPs is a set of covariates. As we've seen above, these are often real, taken from a observational or experimental study. The two major parameters describing the set of covariates are dimensionality, $k = |X_i|$, and the correlation. The number of covariates influences the quality of estimators and usually, more covariates, i.e. a high-dimensional setting [47, 61, 39], make the estimations worse, if the sample size remains fixed. This is intuitive for any machine learning method, because overfitting becomes more of a problem when k is high with respect to the number of samples n. Furthermore, strong correlation or even multicollinearity between covariates make it hard for regressors to choose the right features and result in a fluctuation of feature importance.

Heterogeneity

With the homogeneity and heterogeneity of a DGP we refer to the structure of the treatment effect. A homogeneous treatment effect means, that the ITE for all instances is the same, i.e. $\tau(X_i) = \tau, i = 1, \ldots, n$. When we use a homogeneous treatment effect, we also remove all dependencies of the covariates on the treatment effect and thus remove confounding by design. A heterogeneous treatment effect, as we'll show, is harder to estimate for most methods, because it introduces complexity in the function to estimate and also introduces possible confounding. The strength of heterogeneity can be expressed by the variance of the effect minus the variance of the error $\text{Var}(\tau(x)) - \text{Var}(\epsilon_i)$.

Distribution of Treatment Effects

Depending on the covariates we use to determine the treatment effect, we can observe different distributions for it. In general, many covariates follow a normal distribution, when they refer to natural phenomena, but other distributions are also likely, especially for categorical features. The distribution of $\tau(x)$ can be easily plotted as done in Chapter 5. It has a significant influence on the difficulty of the problem, especially for linear methods like the S-Learner with linear regression, that are not able to capture heterogeneous treatment effects. Is the distribution widely spread on the range of possible treatment values or even a multi-modal distribution, these methods will fail predictably, because the ATE is not a good approximation of the ITE. The distribution can be adjusted by either using a synthetic covariate with the desired distribution or by choosing the right covariate out of real collected data.

Size of Treatment Effect

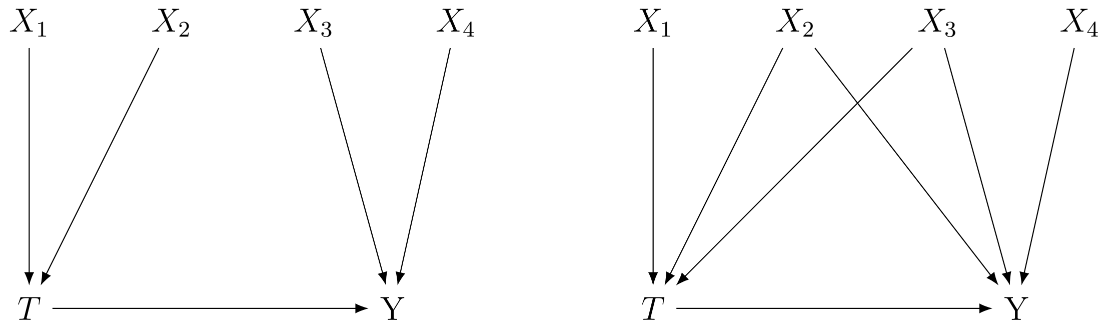
The size of the treatment effect is a parameter that can be used to challenge the sensitivity of estimation methods. We can vary on a scale from no treatment effect to a very strong relative treatment effect, defined as

$$\tau_{rel}(x_i) = \frac{\tau(x_i)}{Y_i(0)}.$$

A small relative treatment effect is harder to grasp, because it is potentially difficult to differentiate between a causal effect and noise in Equation 4.4.

Confounding

The most essential design choice in any DGP for treatment effect estimation is that of confounding. As we've shown in the introduction and in Section 2.1, confounding introduces a bias that skews our estimates of the treatment effect. However, we cannot directly adjust the degree of confounding. In practice, we can only design treatment assignment and outcome assignment. By choosing the same covariates to determine treatment and outcome, we introduce confounding, as is shown exemplary in Figure 4.4. We can then choose the number of confounders and their respective strength, both increasing the difficulty of the resulting problem. Another variation, not considered in our work, is the use of so called hidden confounders [18, 44, 40]. That is to say, covariates that influence both treatment and outcome, but are not in the data we provide to the methods. See [40] for a discussion and one method covering the use of hidden confounders.



Graph without confounding.

(a) Graph with confounders X_2 and X_3.

Figure 4.4: Visual representation of a confounding and no-confounding scenario. Using the theory from Section 2.2, we can observe that there is no back-door path in the left graph, while $T \leftarrow X_2 \rightarrow Y$ is an open back-door path in the right graph.

Treatment Assignment

The treatment assignment is just as essential and, as we've shown, closely related to the design of confounding. What separates a randomised trial form a observational study is the treatment assignment mechanism (Section 2.1.3), which makes this parameter so important. With the treatment assignment we also control the overlap condition. That is to say, we control whether the treatment and control groups are similar or completely different. Also, we control the respective sizes of the two groups, thus also heavily influencing the difficulty of our DGP. In general, we can say that less overlap makes the problem more difficult. A common setting in medical observational studies is that treatment is assigned to instances where the treatment effect is biggest, thus creating an inherent selection bias in the sample [47]. We denote the treatment assignment as a function of the covariates that maps to the probability of the instances described by the covariates receiving treatment. In other words, we manually fabricate the propensity score $p(x)$.

Outcome Assignment

Just as we assign treatment based on the covariates, we have to assign our outcomes in the setting of Equation 4.4 by means of the functions μ and τ. We call μ the base function and τ the treatment effect function.

Functional Form

For the relationship of covariates on treatment, outcome and effect, we have the freedom to choose any functional form. That is to say, we can map the features using linear functions, polynomials, exponential functions, logits or any other form we can think of.

Noise

Lastly, we can determine the distribution of the error term ϵ. From a simple additive normal to interactive measurment errors, biasing the results, there is a wide range of possibilites.

Sample Size

The number of instances we provide heavily influences the performances of methods that rely on larger datasets to learn. For any DGP we can evaluate the estimator perfomance on different sub-samples and thus get an idea of how the methods perform when we scale the number of instances. In reality, we are often confronted with limited data for medical or political settings, as even observational studies are costly. In a big data scenario, on the other hand, collecting more samples is often only a matter of time.

4.4 Benchmarking Metrics

To score the results of the methods presented, different metrics are proposed to measure error, bias and confidence on individual treatment effects as well as average treatment effects.

4.4.1 Average Treatment Effect Error

For the evaluation of the estimation of an average treatment effect, we use a an absolute error, calculated as

$$\epsilon_{ATE} = |\tau - \hat{\tau}|,$$

where $\hat{\tau}$ is often retrieved as the average of individual effect estimations, $\hat{\tau} = n^{-1}\sum_i^n \hat{\tau}(x_i)$.

4.4.2 Precision in Estimation of Heterogeneous Effects

The most relevant score for our scenarios, where the heterogeneous use-case is more interesting, is the Precision in Estimation of Heterogeneous Effects (PEHE). First proposed in [25], the PEHE score is actually just a MSE on the individual treatment effects, comparing the estimates with the true effects. Formally, ϵ_{PEHE} is defined for finite samples as

$$\begin{aligned}\epsilon_{PEHE} &= n^{-1}\sum_{i=1}^{n}\left(\left[Y_i(1) - Y_i(0)\right] - \left[\hat{Y}_i(1) - \hat{Y}_i(0)\right]\right)^2 \\ &= n^{-1}\sum_{i=1}^{n}\left(\hat{\tau}(x_i) - \tau(x_i)\right)^2.\end{aligned}$$

Often, instead of reporting ϵ_{PEHE} directly, the root $\sqrt{\epsilon_{PEHE}}$ is listed, thus making the PEHE score a Root Means Squared Error (RMSE) on individual treatment effects.

4.4.3 Bias

Since the absolute error on ATE is agnostic to the *direction* in which the method errs, the bias is also an interesting metric to consider [57]. It allows us to see if methods tend to over- or under-estimate the treatment effect and is calculated as

$$\epsilon_{BIAS} = n^{-1}\sum_{i}^{n}(\hat{\tau}(x_i) - \tau(x_i)).$$

Dimension	**Possible Considerations**
Covariates	Real or synthetic Number of covariates Correlation between covariates
Treatment Effect	Homogeneous vs. heterogeneous Small vs. large relative effect Distribution of effects
Confounding	Number of confounders Strength of the confounding relation Hidden confounders
Treatment Assignment	Random vs. confounded Overlap condition Treatment/Control size ratio
Sample Size	Number of instances used for training
Noise	Signal to noise ratio Distribution of noise terms
Homoscedasity	Different variance of covariates or not

Table 4.1: Overview of possible dimensions and their description

4.4.4 Effect Normalized Root Means Squared Error

The Effect Normalized Root-Mean-Squared-Error (ENoRMSE) is, to our knowledge, first proposed in [57] with the intent to report an error score that is independent of the absolute treatment effect size. Naturally, it considers the relative treatment effect error, $\frac{\hat{\tau}(x_i)}{\tau(x_i)}$, instead of the absolute one. To penalize wrong estimation this is incorporated into a RMSE via

$$\epsilon_{ENORMSE} = \sqrt{n^{-1}\sum_i^n (1 - \frac{\hat{\tau}(x_i)}{\tau(x_i)})^2}.$$

This score, however, has a problem with treatment effects close to zero and might overpenalize such effects due to a lack of computational precision [57].

4.5 Towards a Benchmarking Framework

In the the process of designing the experiments we wanted to run, it became clear that to simplify the workflow we had to design and implement a small framework. Essentially, the motivation was to be able to quickly compare a multitude of methods on a set of (potentially parametric) DGPs or datasets using different metrics. Furthermore, results should be easy to analyse, stored persistently and categorised according to their settings. Finally, as we've recognised the large dependence of performance on the specifics of the data, we wanted to add tools to analyse reference datasets visually using plots.

The result of this process is a preliminary benchmarking framework[6] that abstracts `CausalMethods` and `DataProviders` so that new methods and new datasets can be added efficiently. The experiments are tracked using the python package `sacred`[7], which makes the workflow a lot easier and faster. All experiments are stored with their parameters and results in a database that can be searched using the corresponding dashboard `omniboard`[8]. More information about the module structure and usage examples can be found in the repository.

[6] The framework can be found at https://github.com/inovex/justcause. For the experiment in this work, the version v0.1 was used. Since then, we've restructured the package completely to provide it as a library.

[7] More information: https://github.com/IDSIA/sacred

[8] More information: https://github.com/vivekratnavel/omniboard

5. Results of the Evaluation

> The short story is that there is no single optimal way to estimate treatment effects – the performance of different methods depends on problem-specific questions.
>
> *Stefan Wager [60, 4, 6] in response to the questions, why Causal Forests are outperformed by simple methods.*

Already during the design of the experiment in the previous chapter, the inherent problems of treatment effect estimator evaluation became clear. We have specified a set of possible dimensions which can be tweaked to create unique challenges. Here, we now show the results we observed for a small subset of possible experiments. Specifically, we validate one major claim with experimental results using the framework introduced in Section 4.5. We further try to replicate the results other authors have observed on Ihdp, with limited success. Finally, we discuss some of the properties of specific methods that we observed during the experiments and try to give an explanation for their behaviour.

5.1 What We Evaluate

To clarify which methods are used, we briefly introduce the abbreviations we used occasionally for formatting reasons. In general, there are meta-learners like the S-, T-, R- and X-learners, that can be implemented using any statistical regression model like linear regression or random forests. To distinguish between different implementations, we list the meta-learners always in conjunction with the used base-learner. All abbreviations are listed in Table 5.1.

GANITE, presented in Section 3.3, is not evaluated because the implementation provided by a third-party [53], does not yield meaningful results. While training seems to be stable, the method performs worse than the simplest baseline. The authors themselves [65] refused to provide their project code for replication. Because re-implementation with hyperparameter search is beyond the scope of this thesis, we had to leave GANITE out of the evaluation.

Full Name	**Abbreviation**	**Origin**	**Implementation**
Causal Forest	CF	[60, 6, 4]	`grf`[1]
Random Forest	RF	[10]	`sklearn`[2]
Linear Regression	LR	-	`sklearn`
Logistic Regression	LogR	-	`sklearn`
CausalBoost	CB	[47]	`causalLearning`[3]
CausalMARS	CM	[47]	`causalLearning`
PTOForest	PTO	[47]	`causalLearning`
Propensity Score Weighting	PSW	[48]	FW[4]
Doubly Robust Estimator	DR	[11]	FW
S-Learner	SL	[33]	FW
T-Learner	TL	[33]	FW
GANITE	GANITE	[65]	`perfect-match`[5]
DragonNet	DN	[56]	`dragonnet`[6]
X-Learner	XL	[33]	`r-learner`[7]
R-Learner	RL	[41]	`r-learner`

Table 5.1: Overview of methods under comparison, their abbreviations, original paper and the implementation used.

5.2 Evaluation Procedure

In Section 4.3.2 we've elucidated the multitude of dimensions one can adjust during evaluation. What's more, the parameters of the methods can also be tweaked to see how they change the performance in a specific setting. The sheer number of possibilities for experimentation led us to follow only a specific hypothesis. We leave the detailed empirical analysis of methods open for future research, but remark on general directions and intuitions about specific methods and their strengths.

The major claim we aim to show with examples in Section 5.3 is the following:

Hypothesis: The performance of methods heavily depends on the DGP settings used. Thus, different DGPs will lead to a different ordering of methods with respect to a metric.

In general, the aim was to design a data generating process (DGP) that isolates a specific component, so that we can rule out any other influence on the results. We implement all DGPs in our benchmarking framework and run the desired methods on them. Since most DGPs are based on random variable distributions for either

[1] https://github.com/grf-labs/grf

[2] http://scikit-learn.github.io/stable

[3] https://github.com/saberpowers/causalLearning

[4] Refers to our own framework implementation; more details in Section 4.5.

[5] https://github.com/d909b/perfect_match

[6] https://github.com/claudiashi57/dragonnet

[7] https://github.com/xnie/rlearner

their covariates, treatment or outcome, we evaluate the methods on a number of replications to rule out sampling specific problems. That is to say, we sample from the DGP a set of instances, run the method, score the results and then re-sample a new dataset. This is the procedure proposed in [54] for IHDP. It ensures that sampling variance does not skew results.

For all experiments, we report every metric introduced in Chapter 4, but we mostly list the $\sqrt{\epsilon_{PEHE}}$ score as a measure of accuracy. We do this, because ϵ_{ATE} is not relevant for the heterogeneous settings, ϵ_{BIAS} is relevant mostly for practical applications but not for theoretical comparison and $\epsilon_{ENoRMSE}$ has rounding problems with small effects. Also, the $\sqrt{\epsilon_{PEHE}}$ score is the most commonly used reference score.

5.2.1 A Note on Plots

Throughout this section we use plots to visually report results and communicate an intuition about them. Specifically, we'll use a point cloud plot to get a sense of how a dataset looks like. Figure 5.1 shows such a plot. The left subplot shows the difference between treated and control instances, colored black and cyan respectively. The right shows the difference between observed and unobserved instances, colored blue and red respectively. The z dimension refers to the outcomes, x and y are the two dimensions of a t-SNE (t-Stochastic Nearest Neighbor Embedding) [59] of the covariates. Without going into details here, t-SNE tries to find an embedding that keeps the t nearest neighbours intact. Thus, for real covariates, t-SNE will cluster similar instances together, while a gaussian distribution, for example, results in a more even cloud.

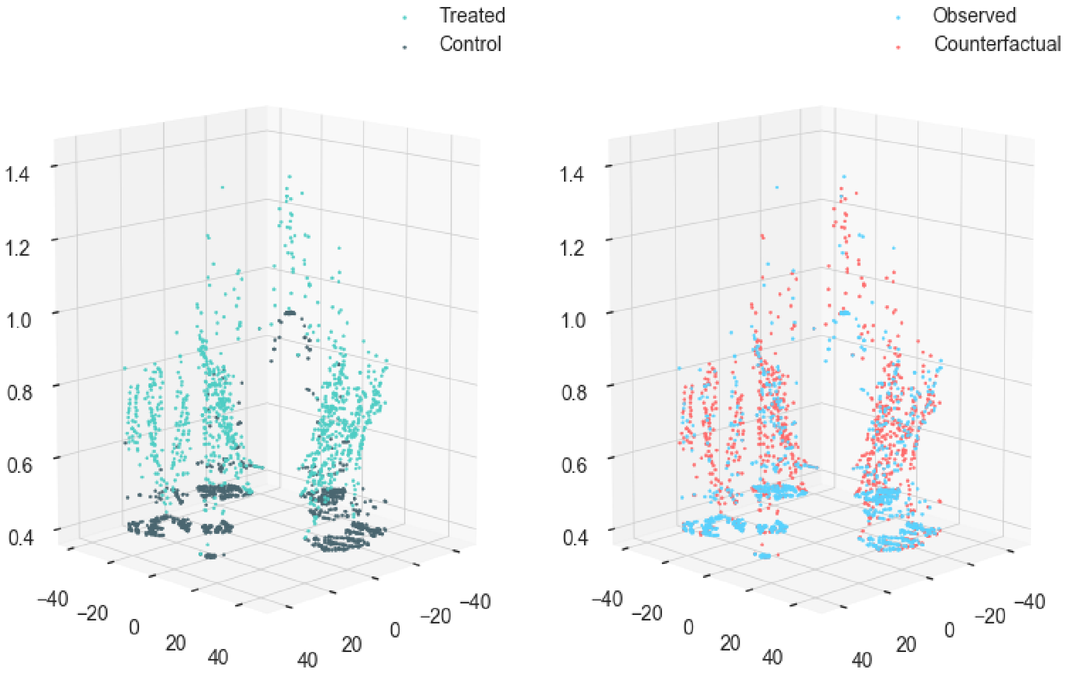


Figure 5.1: An example t-SNE plot.

5.3 Method perfomance varies between settings

Others have hinted at the fact, that method evaluation is highly dependend on the setting we evaluate in [51, 50, 47]. In this Section we present two experiments that make this point clearer and show how difficult it is to get a trustworthy performance estimation for any method.

5.3.1 Multi-Modal versus Exponential

We designed two simple DGPs based on the covariates from the IHDP dataset. One has an explicitly multi-modal effect distribution, denoted by S_M, and the other has a rather simple uni-modal effect function, denoted by S_E. This is to say, in setting S_M we have two groups of units in the treated group with distinctly different treatment effects, that can be distinguished by one covariate. The differences in treatment effects in setting S_E are less clear. This distinction becomes clear in Figure 5.2. We use the 25 covariates from the underlying IHDP study as our covariates, $X = (X_1, \ldots, X_{25})$. Covariates are each normalized to zero mean and standard deviation. The treatment and outcome assignment for the the multi-modal and exponential case are then defined as follows. We set

$$\begin{aligned} \tau_i &= \exp\left(1 + \sigma(X_{i8})\right), \\ T_i &= \textsc{Bern}\left(\sigma(X_{i8})\right), \\ Y_i(0) &\sim \mathcal{N}(X_i^1 + 1, \frac{1}{2}) \text{ and} \\ Y_i(1) &= Y_i(0) + \tau_i, \end{aligned}$$

for the setting S_E. And for the multi-modal setting, S_M, we change τ_i to

$$\gamma_i = \mathbb{I}\left(\sigma(X_{i8}) > \frac{1}{2}\right), \tag{5.1}$$

$$\tau_i \sim \mathcal{N}(3\gamma_i + 1(1 - \gamma_i), 0.1). \tag{5.2}$$

The treatment assignment remains the same. Thus, for both settings, X_8 is the confounder. Figure 5.2 gives a visual insight into the two distributions of treatment effects dependent on the confounder X_8. Covariate X_8 refers to the birthweight of the units in the original IHDP study and is chosen because of its clean normal distribution. The threshold for the sigmoid function is chosen to ensure a 50/50 split between the two treatment effects. The coefficients for τ_i in seeting S_M are chosen to present a clear split and no overlap due to a small deviation of 0.1. Confounding is similar in strength in both scenarios, thus the difference in predictions stems mainly from the structure of the treatment effect. The t-SNE plots in Figures 5.3 and 5.4 show the resulting data distributions.

We run a set of 13 methods over 100 replications of the data and measure the PEHE scores. Firstly, our intuition from visual analysis, that the multi-modal setting is more difficult to grasp, is vindicated by the ENoRMSE score. Namely, the average ENoRMSE score for the exponential setting over all methods is 0.14 versus 0.89 for the multi-modal setting.

Secondly, the results in Table 5.2 show what we claim with our Hypothesis. There is no one best method and performance is highly dependent on the structure of the problem.

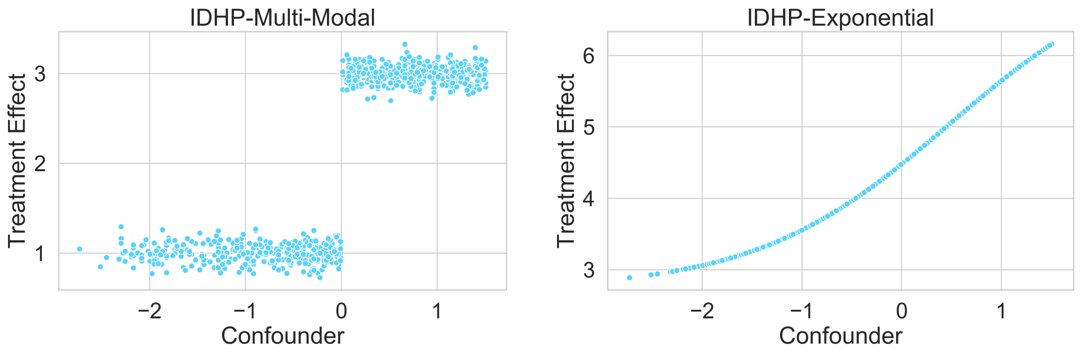


Figure 5.2: Scatter plot of the treatment effect dependent on the confounder for the two settings S_M on the left and S_E on the right.

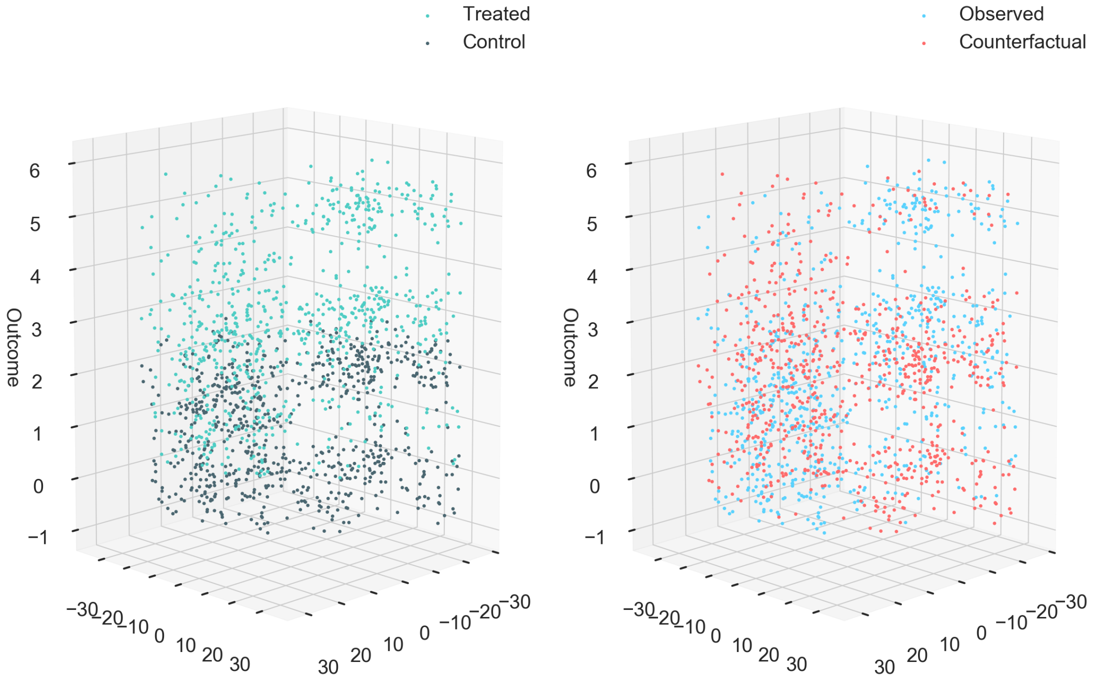


Figure 5.3: Data distribution of the multi-modal setting in the t-SNE plot. Because part of the effects is small and thus hard to detect, the multi-modal setting is considered more difficult.

In general, however, we see that linear methods don't perform well on setting S_M, because they have difficulties grasping the two groups of treatment effects. Instead, they tend to approximate them by choosing the mean, somewhere around 2, resulting in an root mean squared error close to 1. Figures 5.5 (f) and (h) on the next page clarify this intuition. We plotted the true effect against the estimated effect to see how they are approximated. Especially the S-Learner with linear regression and the Propensity Score Weighted methods struggle, because they can only estimate the ATE, not the ITE. Thus, the ATE estimate $\hat{\tau}$ is broadcast to all units yielding the distinctive plots (f) and (h) in Figure 5.5. For the exponential setting, S_E, on the other hand, the lasso based X- and R-learner perform best besides the

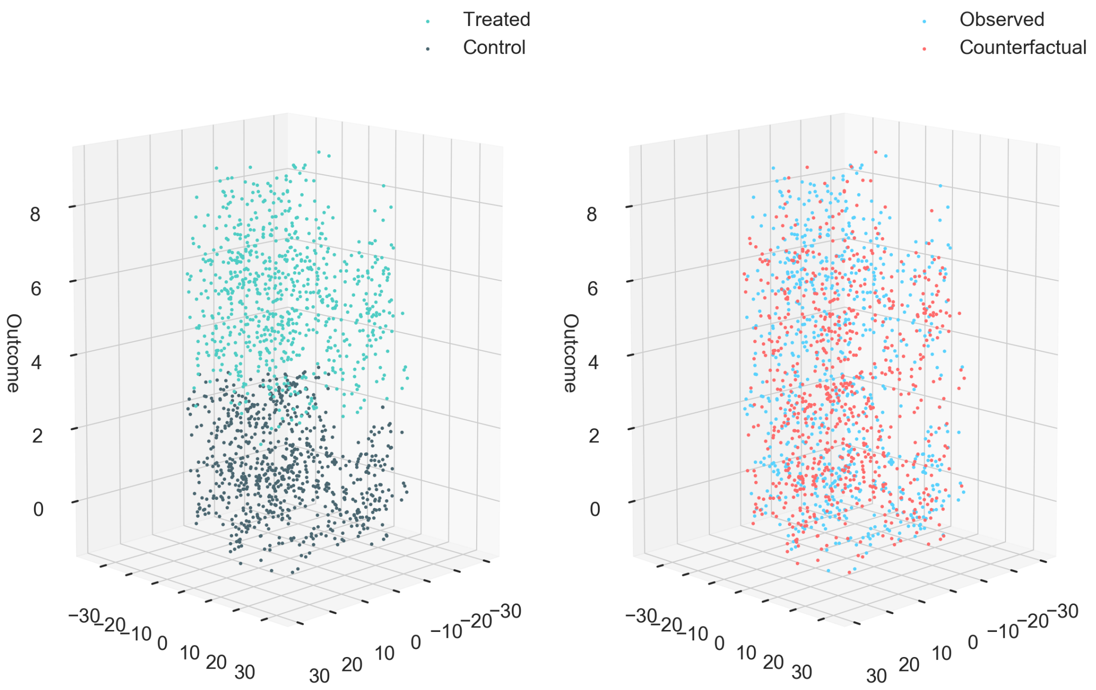
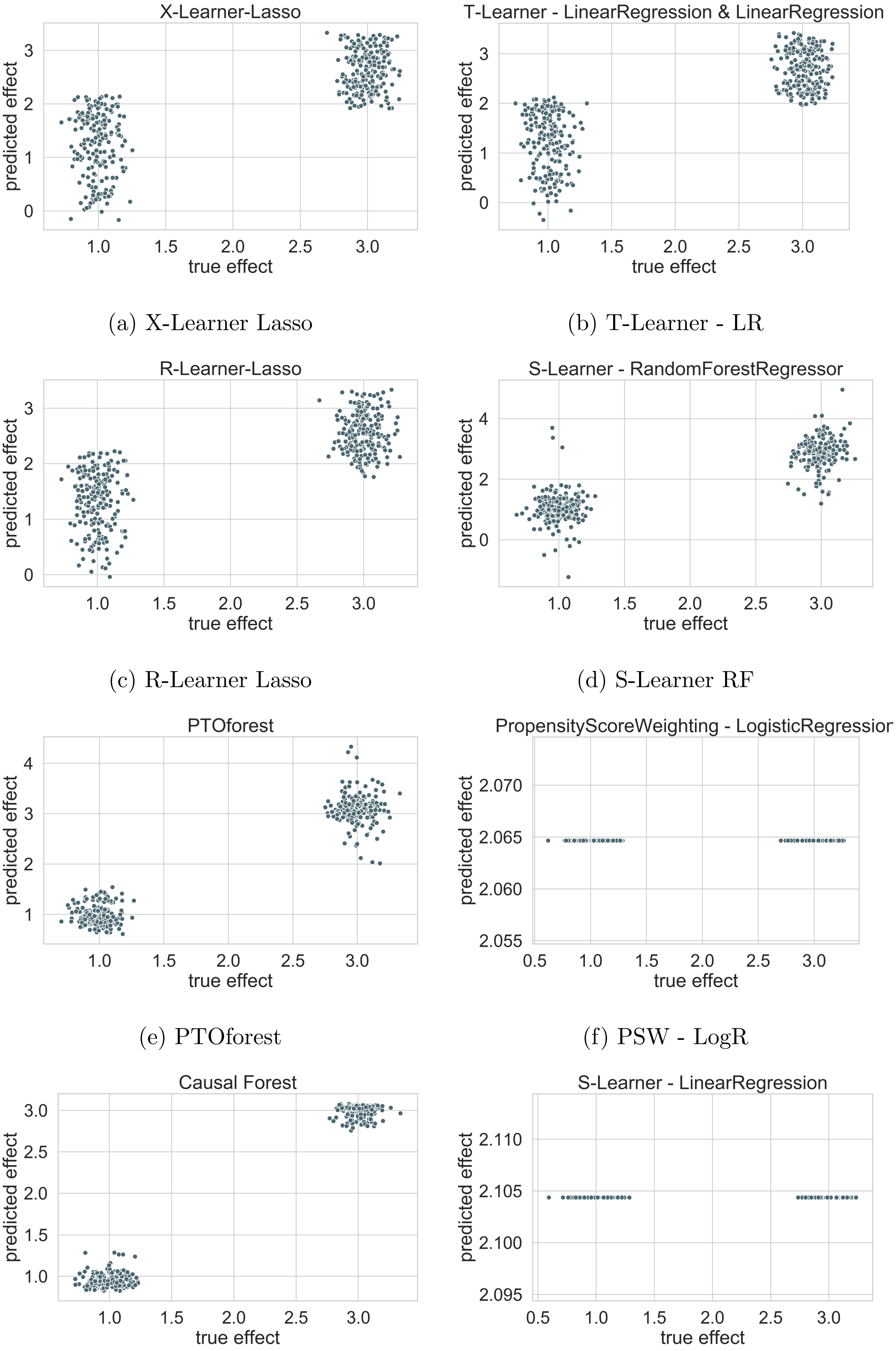


Figure 5.4: Data distribution of the exponential setting in the t-SNE plot.

surprisingly good PTOforest. These results for linear learners fit the theory, because the treatment effect curve shown in Figure 5.2 for S_E is almost a linear function of the confounder X_8, due to the normalisation with the sigmoid function. Only in the lower effect range can we observe the typical 'knee' of an exponential function. As expected, this is also where the T- and X-learner consistently underestimate the effect according to the subplots (c) and (d) in Figure 5.6, because they approach the problem with a linear regressor. Tree based learners like the PTOforest or the Causal Forest struggle with the exponential setting, because the limited number of training samples (747 instances) is not enough for them to learn a smooth representation of the curve.

Surprising is also the below average performance of DragonNet in both settings, which we ascribe to the lack of training data. A hypothesis we consider in more detail in the next section.



(a) X-Learner Lasso

(b) T-Learner - LR

(c) R-Learner Lasso

(d) S-Learner RF

(e) PTOforest

(f) PSW - LogR

(g) Causal Forest

(h) S-Learner - LR

Figure 5.5: Predicted treatment effect plotted against the true treatment effect in the multi-modal setting S_M for various learners.

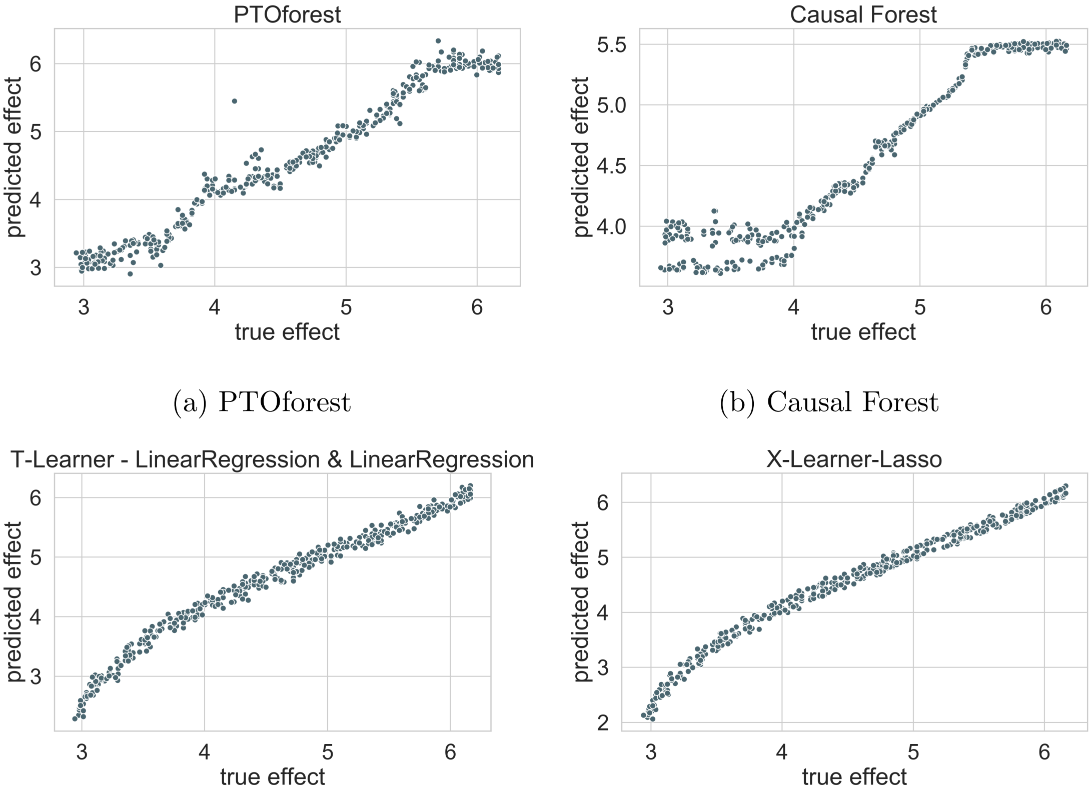
Method	$\sqrt{\epsilon_{PEHE}}$
Causal Forest	0.1648 ±0.01
PTOforest	0.3864 ±0.03
T-Learner - RF	0.4158 ±0.02
S-Learner - RF	0.4179 ±0.03
CausalBoosting	0.5315 ±0.02
X-Learner-Lasso	0.5601 ±0.00
T-Learner - LR	0.5837 ±0.01
R-Learner-Lasso	0.5838 ±0.00
DragonNet100	0.6153 ±0.02
CausalMARS	0.8705 ±0.22
PSW - LogR	0.9960 ±0.01
S-Learner - LR	1.0094 ±0.00
PSW - RF	1.0094 ±0.02

Setting S_M

Method	$\sqrt{\epsilon_{PEHE}}$
PTOforest	0.1801 ±0.01
R-Learner-Lasso	0.1996 ±0.02
X-Learner-Lasso	0.2071 ±0.01
Causal Forest	0.2518 ±0.03
T-Learner - LR	0.2641 ±0.02
CausalMARS	0.3448 ±0.06
T-Learner - RF	0.3607 ±0.02
S-Learner - RF	0.3713 ±0.02
DragonNet100	0.5524 ±0.05
S-Learner - LR	0.9515 ±0.00
CausalBoosting	0.9522 ±0.02
PSW - RF	1.0711 ±0.05
PSW - LR	1.1264 ±0.06

Setting S_E

Table 5.2: Method performance on multi modal setting (left) and exponential setting (right) compared. The score is averaged over 100 replications on a held-out test sample. The different orderings among methods support our Hypothesis.



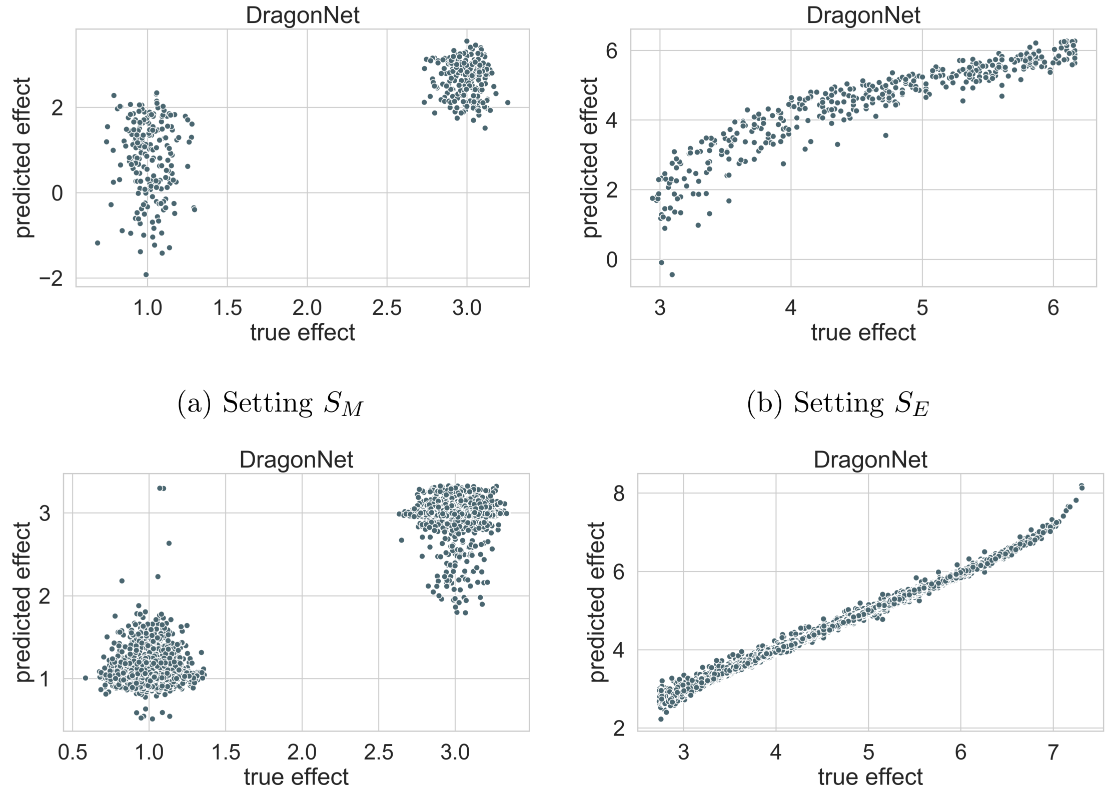
(a) PTOforest

(b) Causal Forest

(c) T-Learner - LR

(d) X-learner

Figure 5.6: Predicted treatment effect plotted against the true treatment effect in the exponential setting S_E for linear learners and tree-based learners. We see that tree-based learners struggle to approximate the smooth treatment effect function. Note that a perfect approximation would result in straight diagonal line in this plot.



(a) Setting S_M

(b) Setting S_E

(c) Setting $\bar{S}_M$ with 20000 training samples

(d) Setting $\bar{S}_E$ with 20000 training samples

Figure 5.7: Treatment scatter plots of DragonNet estimations for small and big training sets. Surprisingly, the performance of DragonNet on the two settings S_M and S_E is in the lower segment (see Table 5.2). The treatment scatter plots show relatively high variance that hints at a lack of training data when the original 747 samples are used. Figures (c) and (d) show the approximation with 20000 samples, where the estimation is much better and more robust.

5.3.2 Scaling Effects

The previous section introduced the multi-modal setting S_M as well as the exponential setting S_E. We observed, among other things, that tree based methods struggle to approximate the treatment effect for setting S_E, because of its smooth and linear structure and that the deep neural network of DragonNet performed below average. Naturally, we thought that the limited training data that the IHDP covariates provide might explain these observations. Thus, we created two very similar settings $\bar{S}_M$ and $\bar{S}_E$ based on ACIC covariates, so that we can evaluate scaling effects. The ACIC datasets provides full covariates for 100.000 samples, of which we used 20.000 for our specific experiment.

As the confounding variable for both settings as described in the section above we chose birth weight, because of its clean normal distribution. In total, 7 covariates are included, all other settings remain the same as in the section above. Results are compiled in Table 5.3

Method	$\sqrt{\epsilon_{PEHE}}$
Causal Forest	0.2073
T-Learner - LR	0.2886
X-Learner-Lasso	0.3003
DragonNet100	0.3173
S-Learner - RF	0.3727
R-Learner-Lasso	0.3745
CausalMARS	0.3962
T-Learner - RF	0.4152
DR - LogR & RF	0.8542
S-Learner - LR	0.8576
DR - RF & RF	0.8582
PSW - RF	0.8652
PSW - LogR	0.8858

Setting $\bar{S}_E$ on 1000 samples

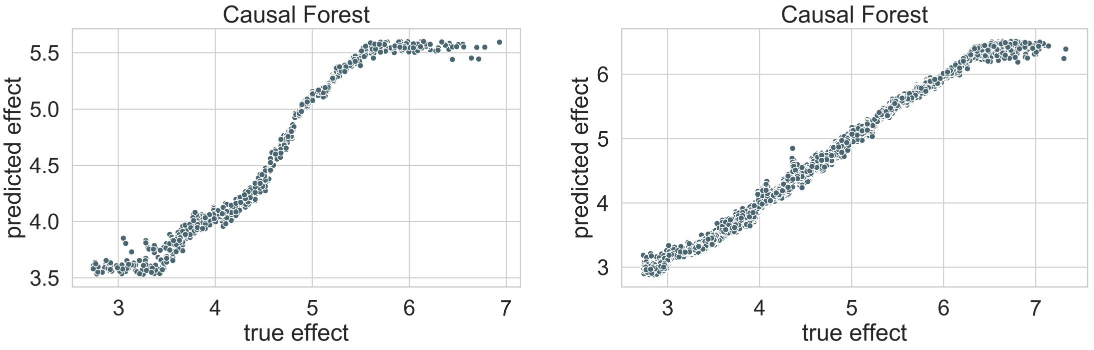
Method	$\sqrt{\epsilon_{PEHE}}$
DragonNet100	0.0606
Causal Forest	0.0851
R-Learner-Lasso	0.2627
X-Learner-Lasso	0.2871
T-Learner - LR	0.2955
T-Learner - RF	0.3499
S-Learner - RF	0.3638
CausalMARS	0.7462
DR - LogR & RF	0.8542
S-Learner - LR	0.8549
PSW - LogR	0.8566
DR - RF & RF	0.8582
PSW - RF	0.8802

Setting $\bar{S}_E$ on 20000 samples

Table 5.3: Method performance on the new exponential setting $\bar{S}_E$ trained on 1000 and 20000 samples

While the results are not comparable directly with the results above due to the different covariates, we can still make interesting observations. Firstly, we see that some methods benefit more than others from the higher number of training samples. Among the beneficiaries are Causal Forests, DragonNet, R-Learner-Lasso and the T-Learner with random forests. This matches our intuition that tree based methods as well as sophisitcated deep models require more training data. For the DragonNet the improvement is most significant. This can also be seen visually by comparing the plots (b) and (d) in Figure 5.7. We see that for setting $\bar{S}_E$, the variance of estimation is far less. The same observations counts for Causal Forests in Figure 5.8, where we see that Linear Models like the T- or S-Learner with linear regression, don't improve their performance with increasing training data.

For the multi-modal setting $\bar{S}_M$, we observe that the number of training instances is less relevant. We argue that this is due to the fact that the distribution of

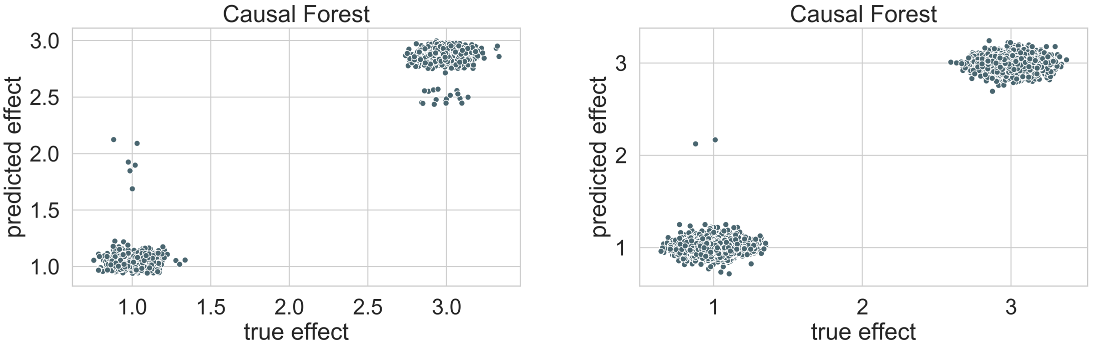


(a) Setting $\bar{S}_E$ with 1000 samples

(b) Setting $\bar{S}_E$ with 20000 samples

Figure 5.8: Causal Forest evaluated on the exponential setting with 1000 and 20000 samples.

treatment effects is *simpler*. That is to say, in order to learn the two clusters of treatment effects, the number of instances in each cluster does not matter. To learn the smooth distribution of the exponential setting, however, more data is always helpful. This intuition is visualised in Figure 5.9, where we see that the clusters are already recovered almost perfectly by the Causal Forest with 1000 training samples.



(a) Setting $\bar{S}_M$ with 1000 samples

(b) Setting $\bar{S}_M$ with 20000 samples

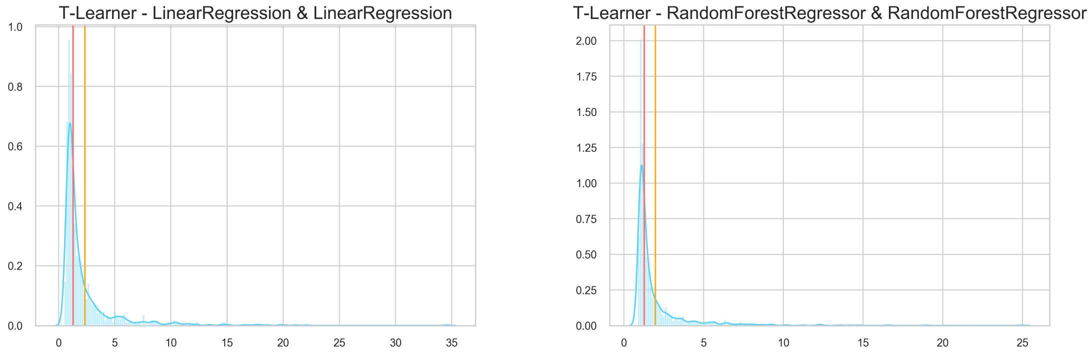
Figure 5.9: Causal Forest evaluated on the multi-modal setting with 1000 and 20000 samples. The effect of more samples is less pronounced than for the exponential setting, as shown in Figure 5.8.

5.4 Experiments on IHDP

The Ihdp dataset we introduced in Section 4.2.1 is among the most commonly used reference datasets for new causal effect estimation methods. Thus, we aim to reproduce the results here. Besides the issues we elucidated in the design chapter, Ihdp suffers from another curse, that is not considered in the evaluation of other authors. Namely, the error distributions of almost all methods on the 1000 replications of Ihdp are heavily skewed.

5.4.1 The IHDP Error Distribution

We use the 1000 replications provided by Christos Louizos [40] and Frederik Johansson [30], because of the problems we had replicating the same distribution with



(a) T-learner - LR

(b) T-learner - RF

Figure 5.10: Distribution of ϵ_{pehe} two T-Learners over the 1000 replications of Ihdp. The red line marks the median, the yellow line marks the mean over all replications. For other methods, the long-tail effect is even worse.

the provided script. In Figure 5.10 we've plotted the distribution of $\sqrt{\epsilon_{PEHE}}$ scores on the 1000 replications for a simple T-learner with a linear base learner. We observe that the scores are not normally distributed, but highly skewed towards bigger values. To make things worse, results are usually reported as the average of all replications [54, 65, 35, 37, 40]. Thus, the reported results are highly dependent on a few of the 1000 replications leading to significantly different scores when running experiments with fewer replications due to time constraints. Because of this observation we propose reporting the median instead of the mean, since mean is sensitive to extreme values. The median is located closer to the highest density of the distribution as can be seen in Figure 5.10. It is thus less dependent on the chosen subset of replications and accordingly more robust.

When it comes to understanding why the 1000 replications of Ihdp have such different difficulty levels, we assume the cause to be the random sampling of coefficients in β_B, as outlined in Section 4.2.1. Because treatment assignment is fixed, depending on the sampled coefficients, the problem is has different levels of confounding, ranging from none to strong. Also treatment effect sizes change significantly among the 1000 replications, as was reported in Figure 4.1. Because the $\sqrt{\epsilon_{PEHE}}$ score is not effect-size-normalised, the few replications with high treatment effects have a greater influence on the resulting mean score.

In Table 5.4 we can observe that reporting the median instead of the mean yields the same ordering of methods that perfectly matches with our theoretical expectations[8]. However, the results for T-learner with random forest vs. linear forest lay closer together when the median is reported. This is a result of the different long-tails that can be observed in Figure 5.10. Namely, the T-learner with linear regression is less robust and results in an average $\sqrt{\epsilon_{PEHE}}$ score that is substantially worse than that of the T-learner with random forest. The same significant difference is not present when the median is reported. Essentially, the T-Learner-Linear performs really bad on some problems, thus nudging the results towards a higher mean score.

[8] To be specific, we expect the non-parametric T-Learner with random forests to best learn the heterogeneous distributions, while S-Learner variants or T-Learner variants with linear regression fall behind.

Method	$\sqrt{\epsilon_{PEHE}}$
T-Learner - RF	1.9749
T-Learner - LR	2.3448
S-Learner - RF	3.5076
S-Learner - LR	5.7849
DR - LR & RF	5.8168
PSW - LR	10.4242

Results as Mean

Method	$\sqrt{\epsilon_{PEHE}}$
T-Learner - RF	1.2634
T-Learner - LR	1.2849
S-Learner - RF	1.458
S-Learner - LR	3.0779
DR - LR & RF	3.1146
PSW - LR	5.2498

Results as Median

Table 5.4: Comparing Ihdp results reported as mean (left) and median (right). We only report these base-methods, because their training and inference time is low. Computational limitations make it infeasible to run a large number of experiments with sophisticated methods on 1000 replications of the Ihdp set.

5.4.2 The Lack of Overlap

We've shortly pointed out in Section 4.2.1, that the overlap condition (Assumption 2.3) is not met for the Ihdp dataset. Specifically, the way an observational subset was created from the experimental data yields a set of instances, where overlap cannot be guaranteed for the control group. Because all units with a value of the feature `momrace` (mother's race) other than `white` are removed from the treated group, a theoretically sound evaluation of treatment effects on non-white instances in the control group is not possible. The probability of treatment for these instances is strictly zero and thus $0 < P[T \mid X] < 1$ (Assumption 2.3) does not hold.

The fact that the overlap condition is not met in both directions is often overlooked. Most authors use the IHDP reference dataset without referring to this abnormality. Theoretically, the evaluation of methods on IHDP with respect to ITE estimation is not possible, as the methods require the overlap condition to be met. An estimation of individual treatment effects, and thus its evaluation, is only valid for the subset of treated instances. Indeed, evaluation of a T- and S-Learner on the individualised Conditional Average Treatment Effect on the Treated (CATT) yields better results than on the whole set of instances. To show this, we trained the two estimators on the whole set of 747 instances, but evaluated the ITE predictions on only the treated or control instances respectively. The results in Table 5.5 hint at the fact that evaluation on the whole set does not do justice to the true performance of the estimators.

Method	$\sqrt{\epsilon_{PEHE}}$
T-Learner - LR	2.2406 ±0.1
S-Learner - LR	5.3863 ±0.2

Only on treated

Method	$\sqrt{\epsilon_{PEHE}}$
T-Learner - LR	2.6203 ±0.1
S-Learner - LR	5.8576 ±0.3

Only on control

Table 5.5: Comparing ϵ_{pehe} scores on the subset of treated instances and on the whole dataset.

What's more, the size of the treated group is far smaller than the control group with 139 versus 608 instances. Results reported as the mean of ITE predictions over all instances are thus dominated by the results on the control group, where the overlap condition is not met.

5.4.3 Reproducing IHDP Results

We run all the methods presented in Chapter 3 on the IHDP to see which results we can reproduce. In Table 5.6, we compiled the results with mean and standard errors for the $\sqrt{\epsilon_{PEHE}}$ score on a held-out test set. Comparing the results with [54], we are able to perfectly reproduce the scores of the base methods T-learner and S-learner with linear regression. Causal Forests, however, perform worse in our experiment than was reported by others [54, 65, 64].

The best performance of the methods compared has the DragonNet, because it uses both theoretical underpinning and the capacity of modern deep neural networks. However, the result cannot be compared with what the authors themselves report, because they report the mean absolute error instead of the $\sqrt{\epsilon_{PEHE}}$ score, obfuscating their results unnecessarily in comparison to other evaluations. At least they mention the limits of IHDP and note specifically, that these results alone are meaningless [56].

The order of propensity score weighting implementations is interesting, because it verifies the intuition that default random forests over-fit on the training set and thus don't work as well as logistic regression or a limited random forest, where the number of splits is artificially constrained.

Method	$\sqrt{\epsilon_{PEHE}}$
DragonNet-100	2.0350 ± 0.1
X-Learner - Lasso	2.4173 ± 0.1
T-Learner - LR	2.5180 ± 0.1
R-Learner - Lasso	2.7188 ± 0.1
T-Learner - RF	3.0884 ± 0.1
CausalMARS	3.4904 ± 0.1
S-Learner - RF	3.7921 ± 0.2
CausalBoosting	3.8348 ± 0.2
CausalForest	4.3146 ± 0.2
S-Learner - LR	5.7920 ± 0.3
DoubleRobustEstimator - LogR & RF	5.7401 ± 0.3
DoubleRobustEstimator - RF & RF	5.8021 ± 0.4
PropensityScoreWeighting - RF limited[9]	7.6800 ± 0.3
PropensityScoreWeighting - LogR	8.3973 ± 0.3
PropensityScoreWeighting - RF std	10.4296 ± 0.4

Table 5.6: Full results on 1000 replications of IHDP. The evaluation is in the *within-sample* setting proposed in [54], where treatment assignment and the factual outcome can be used for treatment effect estimation.

[9] To check the effect of a random forest with less capacity on the propensity regression, we limit the depth of each tree in the forest to 2. That means, each tree can only make two splits.

5.4.4 IHDP as Benchmark Data

Having reported the difficulty to reproduce the exact Ihdp DGP in Section 4.2.1, we showed in this section that the way results are reported is questionable.

First, the 1000 replications contain vastly different data yielding vastly different estimation results with a skewed distribution. Secondly, the results are reported as mean, making them dependent on the few really difficult replications. Lastly, results are reported on the whole sample of 747 instances, where the overlap condition cannot be guaranteed by design. To make things worse, all this is done without acknowledgement of the problems that can occur.

This leads us to the conclusion that using Ihdp as a reference dataset should be done with great care. Authors should provide a clear reference to the specific problem that the DGP models as well as the underlying problems of the evaluation. Ideally, multiple evaluations on different, opposing problem settings are performed to show where the method performs well and where it struggles.

5.5 The Speciality of Causal Forests

Right from the start of our research, Causal Forests played an essential role. Not least, because of their popularity in the field and the large reputation the authors enjoy among causal inference researchers. The average performance on Causal Forests on Ihdp, however, sparked the question, why Causal Forests are so popular after all.

5.5.1 Causal Forests for Small Treatment Effects

Upon the question why Causal Forests are outperformed by a simple T-learner with Linear regression on the Ihdp reference data, Stefan Wager responded with a little toy example to demonstrate the strong-ground of Causal Forests. He claimed that Causal Forests are particularly strong in settings with small and confounded heterogeneous treatment effects. This section aims to analyse the hypothesis along with the example given by Wager.

The demonstration uses a very simplistic DGP, with all 10 covariates sampled from a standard normal $\mathcal{N}(0, 1)$. Further, as Stefan Wager proposed,

$$\begin{aligned} Y(0) &= \sigma(X_1) + \mathcal{N}(0, 1), \\ \tau &= X_2 + X_3, \\ Y(1) &= Y(0) + \tau, \\ T &= \textsc{Bern}(0.5), \end{aligned}$$

for the simple setting. For this settings, the hypothesis is, that both Causal Forests and T-learner perform well, because treatment is randomised and treatment effects are large.

The hard setting S_H as defined by Wager is slightly different, with

$$\begin{aligned} Y(0) &= \sigma(X_1) + \mathcal{N}(0, 1), \\ \tau &= \sigma(X_2 + X_3)/2, \\ Y(1) &= Y(0) + \tau, \\ T &= \textsc{Bern}(\sigma(X_1)). \end{aligned}$$

We've found this experimental setup to be confusing, because it introduces variation in the two DGPs beyond the dimension we are concerned with. Specifically, applying the sigmoid function σ onto the treatment effect not only makes it smaller, which is the desired focus of the comparison, but also changes it's range to $[0, 1]$. In effect, this produces are clearer split between treated and control group with respect to outcome, because in the simple setting, the normal distribution of X_2 and X_3 results in a blended distribution of treated and control.

Figures 5.11 and 5.12 show that the two original DGPs do not only differ in the size of the treatment effect. For these proposed DGPs the results we observe match those Stefan Wager predicted. Namely, T-Learner and Causal Forest both perform well on the simple setting. In the hard setting, however, the Causal Forest far outperforms the T-learner.

To make the claim more meaningful, we slightly adjust the proposed DGP. Because we are only interested in the size of the treatment effect we leave everything else equal. Thus, for the simple setting, we change τ to

$$\tau = \sigma(X_2 + X_3) \cdot 3,$$

and denote this adjusted simple setting by S_S and the hard setting by S_H. It follows, that now the two DGPs are equal with the only difference being the multiplicative coefficient of the treatment effect and the assignment mechanism. In Table 5.7 we compiled the results on our adjusted challenge for the simple and hard settings. Given this adjusted DGP, Stefan Wagers claim does not hold. The T-Learner with linear regression outperforms the Causal Forest in both settings. That is to say, the mere size of treatment effects does not qualify the Causal Forest. Rather, it is the detection of strictly heterogeneous treatment effects like the ones in setting S_M in Section 5.3.1.

Method	$\sqrt{\epsilon_{PEHE}}$
X-Learner - Lasso	0.1408
R-Learner - Lasso	0.1431
T-Learner - LR	0.1553
CausalMARS	0.1700
S-Learner - RF	0.1859
T-Learner - RF	0.1955
Causal Forest	0.1956
PTOforest	0.2517
CausalBoosting	0.4040
S-Learner - LR	0.7880

Setting S_H

Method	$\sqrt{\epsilon_{PEHE}}$
X-Learner - Lasso	0.0240
R-Learner - Lasso	0.0258
CausalMARS	0.0272
T-Learner - LR	0.0281
Causal Forest	0.0337
S-Learner - RF	0.0641
T-Learner - RF	0.0680
PTOforest	0.0794
CausalBoosting	0.0817
S-Learner - LR	0.1314

Setting S_S

Table 5.7: Method performance on hard (left) and simple setting (right) compared. The score is averaged over 100 replications.

5.5.2 Explaining the Performance of Causal Forests

We asked other authors how they explain the average performance of the renowned Causal Forests on problems like the IHDP dataset. Jinsung Yoon, author of the

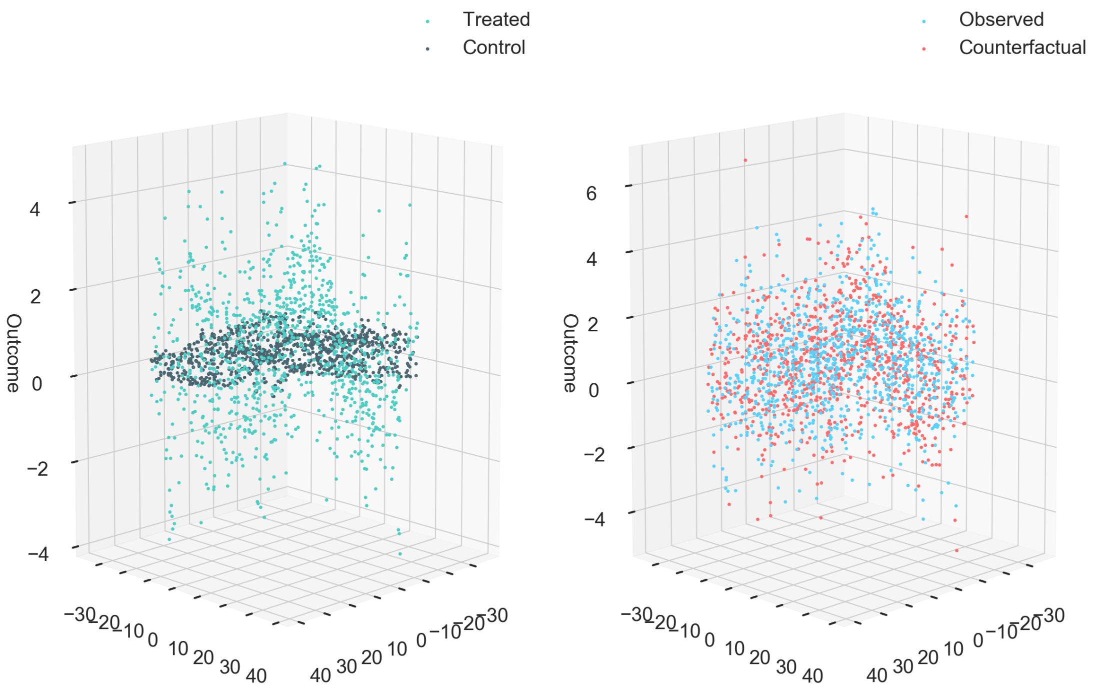


Figure 5.11: Data distribution of the simple setting proposed by Wager in the t-SNE plot.

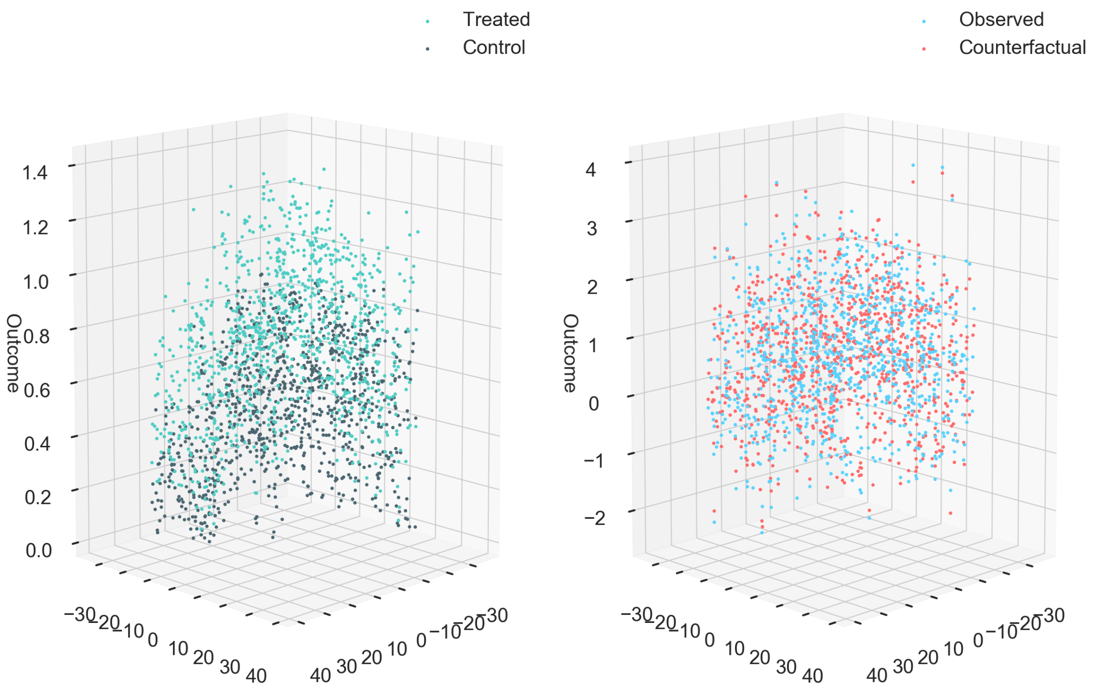


Figure 5.12: Data distribution of the proposed hard setting in the t-SNE plot.

GANITE paper [65], attributed the behaviour to over-fitting. This is possible, as a single Causal Tree splits the data and essentially only uses one half for finding the splits. This is required for the honest property (see Definition 3.2) to work. The negative effect of this split, however, is reduced if not completely alleviated by using Causal Forests, because each tree uses a different split. Hence, over all trees in a forest, all data samples are used for training. All things considered, we do not think that overfitting is the reason for the average performance, rather it is the specific setting with which Causal Forests struggle.

Through the example in Section 5.3.1 and especially the plots in Figure 5.5, we could show that the theoretical stronghold of Causal Forests, i.e. identifying heterogeneous groups, matches the practical results. Hence, the performance on Ihdp is less surprising, not because of the method but because of the structure of the problem. In Figure 4.3, we saw that there aren't any clear splits between different treatment effects on Ihdp. Thus the Causal Forest approach to identify these groups is more likely to fail than approximation of outcomes with regression, which does not explicitly search for heterogeneous groups. This is especially so, with a limited number of training instances. To put it differently: the outcome regression methods don't require distinct treatment effect groups, while the splitting objective of Causal Forests does. This claim is corroborated by the performance of the Causal Forest on setting S_E, shown in Table 5.2, where it is also trumped by linear R- and X-Learners.

Given the clear results of the section above, it is surprising that the author of the method himself gives an example that does not match the theoretical underpinnings of the approach and does not seem robust to detailed analysis. By accident we stumbled across an example that matches the theory with the multi-modal effect setting in Section 5.3.1. Considering how the objectives are chosen for tree splits in Section 3.2, it makes sense to construct a distinct treatment effect rather than a continuous normal distribution.

Another argument that speaks for Causal Forests is given by the authors. In their proposition of generalised random forests, which are a generalisation of Causal Forests, [6], Susan Athey et al. write: *"We develop a large sample theory for our method showing that our estimates are consistent and asymptotically Gaussian, and provide an estimator for their asymptotic variance that enables valid confidence intervals".* Thus, we attribute the high interest in the method they developed to the statistical properties rather than to the hard-measured performance on general settings. Especially in the original field of study of Susan Athey and Stefan Wager, econometrics, these properties are of higher importance.

6. Conclusion

> I would rather discover one causal law than be King of Persia.
>
> Democritus (460–370 B.C.)

In this thesis we provided an overview of the field of causal inference and elaborated on the two major frameworks providing tools for working with causal claims: *Potential Outcomes* and *Structural Causal Models*. We then focused on the subset of causal inference that is concerned with the estimation of treatment effects. To tackle this task we introduced a wide set of methods and clarified the underlying theories based on the theoretical foundations. In the second part of the thesis, starting from Chapter 4, we 1) analysed how causal effect estimators are evaluated, 2) uncovered a bad state of affairs regarding the use of reference datasets and 3) provided a more comprehensive taxonomy of data generating processes (DGPs) to allow for clearer communication of results. Finally, we presented experimental results for one major hypothesis concerned with the very nature of evaluations and we discussed interesting observations. In the meantime, we implemented a benchmarking framework using parametric DGPs that allows to run a variety of experiments easily, ensuring consistency, logging and visual reports.

This section concludes our work and summarises the observations.

6.1 Two Viewpoints United

After a detailed study of the theoretical foundations and assumptions underlying the *Potential Outcomes* framework and *Structural Causal Models* in Chapter 2, we came to the conclusion that the two theories stand for entirely different paradigms in the field of Causal Inference.

The *Potential Outcomes* framework is concerned with the very specific setting where some form of treatment is applied to a population or to individuals. The theory considers the problems that are derived from this setting. It's primary achievement and use is to give the formal foundations for using statistical learning methods, both old and new, to tackle the problem of treatment effect estimation in various settings.

Structural Causal Models, on the other hand, tackle a wider problem by modelling causal claims through graphs. The primary contribution is more on the logical, theoretical side than on the task of actual estimation. Namely, SCMs allow to make clear statements whether a quantity, i.e. a causal effect, can be recovered from observational data. The algorithmic machinery introduced by Pearl et al. [43] is comprehensive and powerful, but not all-mighty.

For treatment effect estimation we learned that we can unite the two paradigms and use SCMs to model our assumptions and study the feasibility of our task in theory. We then use the formulations of the PO framework to estimate the quantities of interest. In Section 2.2.5, we showed by means of an example, what such a collaboration can look like.

6.2 Different Settings, Different Results

By virtue of the experiments outlined in Section 5.3, we corroborated the hypothesis that has been hinted at before by others [47, 51, 52]. Namely, different settings for the data generating process yield different evaluation results. The gist being, we cannot make general statements as to the performance of estimators. Or, to put it differently, when we proclaim the good quality of a method, we must accompany this claim with a clear definition of the setting used.

In our specific setting we could show that the case of a clearly separated, heterogeneous treatment effect is best handled by tree based method, foremost by the specifically designed Causal Forest [4, 60]. Linear methods, on the other hand, struggled to grasp the difference in treatment effects.

Furthermore, we could show that some methods benefit more from increasing sample sizes, while others keep a steady performance or even deteriorate slightly.

6.3 Misuse of reference data

Throughout our work with reference datasets for treatment effect estimation, namely Ihdp and Twins, we encountered difficulties in reproducing DGPs as well as results. These obstacles put a big question mark behind the evaluation results of other authors, because they hinder replicability.

In Sections 4.2.1 and 5.4, we revealed that:

1. The script used to generate the Ihdp data is faulty with invisible traps. What's more, apparently different versions of the dataset are being used under the same name.

2. The fundamental overlap condition is not met for the larger part of the samples within the Ihdp data, making a theoretically sound evaluation infeasible. While this was clear to the authors of the original paper [25], later users of the dataset don't mention it.

3. The distribution of $\sqrt{\epsilon_{PEHE}}$ over the 1000 replications used for IHDP evaluation is long-tailed, making the reported mean result dependent on the performance of a method on a few very difficult replications.

4. The way IHDP is generated does not follow a specific hypothesis. It only covers one possible problem setting and does not allow a general comparison of methods, as seen in Section 5.3.

These observations led us to believe that the way by which many new treatment effect estimation methods are evaluated is flawed and, to put it bluntly, unscientific.

7. Future Work

Building on the concluding observations of Chapter 6, we want to point out a few directions of further research that we believe to be important, if not inevitable for the field.

7.1 More Hypothesis

In Section 5.3, we've elaborated on a hypothesis by designing a specific DGP. In Section 5.4 we've attempted a detailed analysis of a commonly used reference dataset. The sheer number of dimensions, methods and datasets introduced in Chapter 4, however, made it infeasible to consider these hypothesis in more detail, let alone consider more research questions.

To make up for this, we list a few potentially interesting experiments for study that can reveal specific strengths and weaknesses of methods. Thus, further illuminating the vague field of performance.

1. *Number of Observations.* The so called scaling properties first considered in [57] are of interest especially for big data applications, where huge dataset are potentially available.

2. *Number of Confounders.* In our work, we've only looked at examples with a few specific confounders. Studying the effect of more confounders and more interaction between them is of great importance for real world applications.

3. *Level of Noise.* Introducing more noise to the error terms in the synthetic DGPs can yield meaningful results for the robustness of method.

7.2 Toward Better Evaluation of Causal Effect Estimators

We've seen in our evaluation in Chapter 5 that the performance of causal effect estimators is largely dependent on the properties of the data. For the selection of an estimator for a given task, this means that we have to evaluate the performance of different methods on the dataset we target. As we've learned, though, this is not possible.

One way to tackle this is to evaluate on a large set of DGPs, the direction pointed out in the previous section. Another way to cope with this problem is to study real world validation. Alejandro Schuler et al. [52] proposed a method that aims to take into account the properties of the target dataset while synthetically providing ground-truth. Based on it's similarity to cross-validation they call their method *Synth-Validation*.

7.2.1 Synth-Validation

The goal of synth-validation [52] is to estimate a generative distribution $P^*(X, W, Y)$ of covariates X, binary treatment T and outcomes Y, that has a synthetic treatment effect $\tilde{\tau}$ while being *informed by* the real data distribution. That is to say, formally, we want a distribution that has a chosen treatment effect, while also maximising the likelihood of the observed data. Note that Synth-Validation does not estimate the original distribution $P(\bar{X}, \bar{W}, \bar{Y})$ directly, for this would mean it could calculate the true treatment effect. Instead, it estimates a distribution that comes close to the original covariate and treatment distribution $P(X, W)$ while having a specified synthetic treatment effect.

Given a synthetic treatment effect $\tilde{\tau}$, we want the estimated distribution to satisfy

$$\begin{aligned}\tilde{\tau} &= \mathbb{E}[Y \mid X, T = 1] - \mathbb{E}[Y \mid X, T = 0] \\ &= \frac{1}{n}\sum_{i}^{n} \mathbb{E}[Y \mid X = x_i, T = 1] - \frac{1}{n}\sum_{i}^{n} \mathbb{E}[Y \mid X = x_i, T = 0] \\ &= \frac{1}{n}\sum_{i}^{n} [\mu_1(x_i) - \mu_0(x_i)],\end{aligned}$$

for the finite observed sample $\mathcal{S} = \{X_i, T_i, Y_i\}_i^n$. This works because we can factorize the distribution $P(X, T, Y) = P(Y|X, T)P(X, T)$, and estimate $P(X, T)$ from the empirical data via standard *Maximum Likelihood Estimation*. To estimate $P(Y|X, T)$, we assume that

$$y_i = \mathbb{I}\{T_i = 0\} \cdot \mu_0(x_i) + \mathbb{I}\{T_i = 1\} \cdot \mu_1(x_i).$$

The essential step is now to estimate μ_0, μ_1 while also holding the constraint for the synthetic treatment effect. Thus, normal machine learning approaches to estimation do not work. Instead we retrieve the results via:

$$\mu_0, \mu_1 = \begin{cases} \arg\min_{f_0, f_1} \sum_{S_0} l(y_i, f_0(x_i)) + \sum_{S_1} l(y_i, f_1(x_i)) \\ \text{subject to } \frac{1}{n}\sum_i [f_1(x_i) - f_0(xi)] = \tilde{\tau} \end{cases}$$

How exactly this constrained optimisation can be done, is described in detail in the original paper [52].

Choosing the synthetic treatment effect $\tilde{\tau}$ is not a trivial endeavour either, because we want it to be somewhat realistic. Therefore, the authors resort to a simple heuristic. They run various inference methods on the data and capture their estimated effects. They then choose a synthetic treatment effect from the range between the maximum and minimum estimated effect.

We point toward Synth-Validation at this stage, because we believe it to be an important approach that needs further research. Especially in practice we can never be sure as to how our data looks. Thus having a oracle approach like Synth-Validation that tells us which method likely performs best, is a promising idea. Also, Advani et al. [2] point toward Synth-Validation after discarding the standard EMCS approach. Schuler et al. show empirically that while being computationally expensive and *"somewhat unorthodox"* [50, p. 40], Synth-Validation led to a better model selection on datasets where the best method was already known.

7.3 A Taxonomy for Benchmarking Datasets

Building upon the ideas outlined in Section 4.3.2, we propose that further research ought to be directed at designing a clear taxonomy of benchmarking datasets.

On the one hand, there is a significant difficulty in practice to validate the quality and origin of a dataset, as was shown with the IHDP example in Section 4.2.1. Knowing the classification of a reference dataset and it's true generating process, we could avoid using erroneous versions of the dataset[1].

On the other hand, the reference datasets are used without deliberate discussion of their respective features. This is particularly dangerous in light of the fact that the quality of causal effect estimation methods can only ever be evaluated for a specific setting. Especially more recent methods [65, 35, 37, 54, 44] resort to using the most common datasets IHDP and TWINS to show the quality of their methods. The way these benchmarks are listed, seems to claim generality. The two common benchmarks, however, only cover a very small set of possible combinations of the dimensions listed in Section 4.3.2.

Besides more research on the various possible settings described in Section 4.3.2, a clearer language is required. Along with this clear taxonomy goes a heightened awareness for the necessity of differentiating between the performance of a method in a specific field and its performance in general.

While there have been more elaborate efforts to formalize DGPs, like in [22, 47, 32], a overarching categorisation and classification is missing.

[1] We noted in Section 4.2, that some authors have defaulted to copying results instead of validating them for the lack of replicability.

Bibliography

[1] Jonathan E Adler and Lance J Rips. *R E A S O N I N G Studies of Human Inference and Its Foundations.* 2008. ISBN: 9780521883290.

[2] Arun Advani et al. *MOSTLY HARMLESS SIMULATIONS? USING MONTE CARLO STUDIES FOR ESTIMATOR SELECTION **. Tech. rep. 2019.

[3] Douglas Almond, Kenneth Y. Chay, and David S. Lee. *The costs of low birth weight.* 2005.

[4] Susan Athey and Guido Imbens. "Recursive partitioning for heterogeneous causal effects". In: *PNAS* 113 (2016).

[5] Susan Athey and Guido W Imbens. "The State of Applied Econometrics: Causality and Policy Evaluation". In: *Journal of Economic Perspectives* 31 (2017), pp. 3–32.

[6] Susan Athey, Julie Tibshirani, and Stefan Wager. "Generalized random forests". In: *Annals of Statistics* 47.2 (2019), pp. 1179–1203. ISSN: 00905364.

[7] Elias Bareinboim and Judea Pearl. "Causal inference and the data-fusion problem". In: ().

[8] Colin R Blyth. *On Simpson's Paradox and the Sure-Thing Principle.* Tech. rep. 338. 1972, p. 338.

[9] Janet Box-Steffensmeier et al. *The Neyman-Rubin Model of Causal Inference and Estimation via Matching Methods.* Tech. rep. 2007.

[10] Leo Breiman. *Random Forests.* Tech. rep. 2001, pp. 5–32.

[11] Marie Davidian. *Double Robustness in Estimation of Causal Treatment Effects.* Tech. rep. NC State University, 2007.

[12] Author A P Dawid. "Causal Inference Without Counterfactuals". In: 95.450 (2000), pp. 407–424.

[13] Peng Ding Fan Li et al. *Causal Inference: A Missing Data Perspective.* Tech. rep. 2018.

[14] Centers for Disease Control. *User Guide to the 2017 Period Linked Birth/Infant Death Public Use File 2017 Period Linked Birth/Infant Death Data Set.* Tech. rep. 2017.

[15] Vincent Dorie et al. "A flexible, interpretable framework for assessing sensitivity to unmeasured confounding". In: (2016).

[16] Felix Elwert and Christopher Winship. "Endogenous Selection Bias: The Problem of Conditioning on a Collider Variable". 2014.

[17] Jerome H Friedman. *Greedy Function Approximation: A Gradient Boosting Machine*. Tech. rep. 5. 2001, pp. 1189–1232.

[18] Minna Genbäck and Xavier De Luna. *Causal inference taking into account unobserved confounding*. Tech. rep.

[19] Ian J Goodfellow et al. *Generative Adversarial Nets*. Tech. rep.

[20] Ruocheng Guo et al. *A Survey of Learning Causality with Data: Problems and Methods*. Tech. rep. 2019, p. 36.

[21] York Hagmayer and SA Sloman. “Causal Reasoning Through Intervention”. In: *Gopnik* (2007), pp. 86–101.

[22] P. Richard Hahn, Vincent Dorie, and Jared S. Murray. “Atlantic Causal Inference Conference (ACIC) Data Analysis Challenge 2017”. In: (2019), pp. 1–13.

[23] Eduardo Hariton and Joseph J Locascio. “Randomised controlled trials-the gold standard for effectiveness research HHS Public Access”. In: *BJOG* 125.13 (2018), p. 1716.

[24] Bruce Headey and Ruud Muffels. *Two-way Causation in Life Satisfaction Research: Structural Equation Models with Granger-Causation*. Tech. rep.

[25] Jennifer L Hill. “Bayesian Nonparametric Modeling for Causal Inference”. In: *Journal of Computational and Graphical Statistics* 20.1 (2011), pp. 217–240. ISSN: 1537-2715.

[26] Paul W Holland. “Statistics and Causal Inference”. In: *Source: Journal of the American Statistical Association* 81.396 (1986), p. 45.

[27] David Hume. “An enquiry concerning human understanding”. In: *Seven Masterpieces of Philosophy*. Routledge, 2016, pp. 191–284.

[28] Guido Imbens and Donald B Rubin. *Causal Inference for Statistics , Social , and Biomedical Sciences*. 2018, pp. 11–14. ISBN: 9781139025751.

[29] Marshall M Joffe, Wei Peter Yang, and Harold I Feldman. “Selective Ignorability Assumptions in Causal Inference”. In: *The International Journal of Biostatistics CAUSAL INFERENCE M* 6.2 (2010), p. 11.

[30] Fredrik D Johansson, Uri Shalit, and David Sontag. *Learning Representations for Counterfactual Inference*. Tech. rep. 2016.

[31] Kevin Hartnett. *To Build Truly Intelligent Machines, Teach Them Cause and Effect | Quanta Magazine*.

[32] Michael C Knaus et al. *Machine Learning Estimation of Heterogeneous Causal Effects: Empirical Monte Carlo Evidence*. Tech. rep. 2018.

[33] Sören R Künzel et al. *Meta-learners for Estimating Heterogeneous Treatment Effects using Machine Learning*. Tech. rep. 2019.

[34] Robert J Lalonde. *American Economic Association Evaluating the Econometric Evaluations of Training Programs with Experimental Data*. Tech. rep. 4. 1986, pp. 604–620.

[35] Changhee Lee, Nicholas Mastronarde, and Mihaela Van Der Schaar. *Estimation of Individual Treatment Effect in Latent Confounder Models via Adversarial Learning*. Tech. rep.

[36] Michael J Lopez and Roee Gutman. *Estimation of causal effects with multiple treatments: a review and new ideas*. Tech. rep.

[37] Christos Louizos et al. *Causal Effect Inference with Deep Latent-Variable Models*. Tech. rep. 2017.

[38] Robyn M Lucas and Anthony J Mcmichael. *Association or causation: evaluating links between environment and disease*. Tech. rep. 10. 2005.

[39] Ian Lundberg. *Causal forests A tutorial in high-dimensional causal inference*. Tech. rep. 2017.

[40] David Madras et al. "Fairness through Causal Awareness: Learning Causal Latent-Variable Models for Biased Data". In: (2018).

[41] Xinkun Nie and Stefan Wager. *Quasi-Oracle Estimation of Heterogeneous Treatment Effects*. Tech. rep.

[42] Judea Pearl. "Causal inference in statistics: An overview". In: *Statistics Surveys* 3.0 (2009), pp. 96–146. ISSN: 1935-7516.

[43] Judea Pearl. *Causality*. Cambridge university press, 2009. ISBN: 9780521895606.

[44] Judea Pearl. *Detecting Latent Heterogeneity*. Tech. rep.

[45] Judea Pearl. "The seven tools of causal inference, with reflections on machine learning". In: *Communications of the ACM* 62.3 (2019), pp. 54–60. ISSN: 00010782.

[46] Makenzie D Pearl J. *The Book of Why*. Vol. 1. 2018, p. 402. ISBN: 9780465097616.

[47] Scott Powers et al. *Some methods for heterogeneous treatment effect estimation in high-dimensions*. Tech. rep. 2017.

[48] Paul R Rosenbaum and Donald B Rubin. "The central role of the propensity score in observational studies for causal effects". In: *Biometrika* 70.1 (1983), pp. 41–55. ISSN: 0006-3444.

[49] Donald B Rubin. *ESTIMATING CAUSAL EFFECTS OF TREATMENTS IN RANDOMIZED AND NONRANDOMIZED STUDIES 1*. Tech. rep. 5. 1974, pp. 688–701.

[50] Alejandro Schuler. "Compare real world performance of Causal Estimators". PhD thesis. Stanford University, 2019.

[51] Alejandro Schuler et al. *A comparison of methods for model selection when estimating individual treatment effects*. Tech. rep. 2018.

[52] Alejandro Schuler et al. *Synth-Validation: Selecting the Best Causal Inference Method for a Given Dataset*. Tech. rep. 2017.

[53] Patrick Schwab, Lorenz Linhardt, and Walter Karlen. *Perfect Match: A Simple Method for Learning Representations For Counterfactual Inference With Neural Networks*. Tech. rep.

[54] Uri Shalit, Fredrik D Johansson, and David Sontag. *Estimating individual treatment effect: generalization bounds and algorithms*. Tech. rep. 2017.

[55] Cosma Rohilla Shalizi. "Advanced data analysis from an elementary point of view". In: *Book Manuscript* (2013), p. 801.

[56] Claudia Shi, David M Blei, and Victor Veitch. *Adapting Neural Networks for the Estimation of Treatment Effects*. Tech. rep.

[57] Yishai Shimoni et al. *Benchmarking Framework for Performance-Evaluation of Causal Inference Analysis*. Tech. rep. 2018.

[58] Peter Sprites, Clark Glymour, and Richard Scheines. *Causation, Prediction, Search*. Vol. The MIT Press, 2000. ISBN: 0262194406.

[59] Laurens Van Der Maaten and Geoffrey Hinton. *Visualizing Data using t-SNE*. Tech. rep. 2008, pp. 2579–2605.

[60] Stefan Wager and Susan Athey. "Estimation and Inference of Heterogeneous Treatment Effects using Random Forests". In: *Journal of the American Statistical Association* 113.523 (2018), pp. 1228–1242. ISSN: 1537274X.

[61] T Wendling et al. "Comparing methods for estimation of heterogeneous treatment effects using observational data from health care databases". In: (2018).

[62] Priyantha Wijayatunga. *Causal Effect Estimation Methods*. Tech. rep. 2014.

[63] Elizabeth J Williamson et al. "INVITED REVIEW SERIES: MODERN STATISTICAL METHODS IN RESPIRATORY MEDICINE Introduction to causal diagrams for confounder selection". In: (2014).

[64] Liuyi Yao et al. *Representation Learning for Treatment Effect Estimation from Observational Data*. Tech. rep.

[65] Jinsung Yoon, James Jordon, and Mihaela van der Schaar. "GANITE: Estimation of Individualized Treatment Effects using Generative Adversarial Nets". In: *Iclr 2018* 2010 (2018), pp. 1–15.

Glossary

ATE Average Treatment Effect. 12, 39, 43, 48, 50, 57

CATE Conditional Average Treatment Effect. 12, 29, 43

CATT Conditional Average Treatment Effect on the Treated. 65

DGP Data Generating Process. 21, 47, 48, 56, 67, 68, 72, 77

EMCS Empirical Monte Carlo Study. 46, 77

ENoRMSE Effect Normalized Root-Mean-Squared-Error. 52, 56

FPCI Fundamental Problem of Causal Inference. 3, 41

GAN Generative Adversarial Nets. 33, 34

GANITE Generative Adversarial Nets for Inference of Individual Treatment Effects. 33

ITE Individual Treatment Effect. 12, 14, 35, 46, 48, 57, 65, 66

MSE Mean Squared Error. 22, 30, 50

PEHE Precision in Estimation of Heterogeneous Effects. 50, 56

RCT Randomized Control Trial. 4, 6, 11, 13, 14, 41

RMSE Root Means Squared Error. 50, 52

SCM Structural Causal Model. 16, 17, 72