ASSESSING TREATMENT EFFECT HETEROGENEITY: PREDICTIVE COVARIATE SELECTION AND SUBGROUP IDENTIFICATION

A thesis submitted to The University of Manchester
for the degree of Doctor of Philosophy
in the Faculty of Science and Engineering

2021

Konstantinos Papangelou

School of Engineering
Department of Computer Science

Contents

Word Count: 36006

List of Tables

List of Figures

Abbreviations

AE	Absolute Error
ATC	Average Treatment effect on the Control
ATE	Average Treatment Effect
ATM	Average Treatment effect among the evenly Matchable
ATO	Average Treatment effect on the Overlap population
ATT	Average Treatment effect in the Treated
BMI	Body Mass Index
CART	Classification And Regression Trees
CATE	Conditional Average Treatment Effect
CBPS	Covariate Balancing Propensity Score
CF	Causal Forest
CMI	Conditional Mutual Information
CSATE	Conditional Sample Average Treatment Effect
DR	Doubly Robust
EBAL	Entropy Balancing
EGFR	Epidermal Growth Factor Receptor
FNR	False Negative Rate
GP	Gaussian Process
IPW	Inverse Propensity Weighting
IT	Interaction Trees
ITE	Individual Treatment Effect
JMI	Joint Mutual Information
KOM	Kernel Optimal Matching
KRR	Kernel Ridge Regression
LASSO	Least Absolute Shrinkage and Selection Operator
MACE	Major Adverse Cardiovascular Event
MAE	Mean Absolute Error

MCR	Modified Covariates Regression
MIM	Mutual Information Maximisation
ML	Maximum Likelihood
MSE	Mean Squared Error
NFD	Number of False Discoveries
NSCLC	Non-Small-Cell Lung Carcinoma
PCT	Proportion of Correct Trees
PRIM	Patient Rule Induction Method
QUINT	Qualitative Interaction Trees
RA	Regression Adjustment
RBF	Radial Basis Function
RF	Random Forest
RHC	Right Heart Catheterization
RKHS	Reproducing Kernel Hilbert Space
SATC	Sample Average Treatment Effect in the Control
SATE	Sample Average Treatment Effect
SATT	Sample Average Treatment Effect in the Treated
SE	Standard Error
SIDES	Subgroup Identification using Differential Effect Search
SUTVA	Stable Unit Treatment Value Assumption
SVM	Support Vector Machine
SW	Stable Weighting
TPR	True Positive Rate
VT	Virtual Twins
WATE	Weighted Average Treatment Effect

Notation

$\mathcal{D}$	dataset
$\boldsymbol{X}$	covariate set or joint random variable depending on the context
$\boldsymbol{x}$	a realisation of $\boldsymbol{X}$
X	random variable describing a single covariate
x	a realisation of X
T	treatment variable
t	a realisation of the treatment variable
Y	outcome variable
y	a realisation of the outcome variable
$Y(t)$	potential outcome under some value of the treatment
$y(t)$	a realisation of the potential outcome
$e(\boldsymbol{x})$	propensity score of the observation $\boldsymbol{x}$
$m_t(\boldsymbol{x})$	assumed model for the potential outcome of an observation
$\boldsymbol{X}_\theta$	selected covariates (set or random variable)
$\boldsymbol{X}_{\tilde{\theta}}$	unselected covariates (set or random variable)
$\boldsymbol{w}$	vector of weights
w	a single weight
$\boldsymbol{X}_{pred}$	predictive covariates (set or random variable)
$\boldsymbol{X}_{prog}$	prognostic covariates (set or random variable)
$\boldsymbol{X}_{irr}$	irrelevant covariates (set or random variable)
γ	real-valued number describing the confounding strength
$\mathcal{S}$	set of examples $\boldsymbol{x}$ belonging in a subgroup
$S(\boldsymbol{x})$	indicator of whether an example $\boldsymbol{x}$ belongs to the subgroup
$\mathcal{I}_T$	set of internal nodes for the tree T
ρ	parameter controlling the complexity of a tree
G	splitting criterion for a tree
K	a kernel matrix

ϕ	a feature map
σ^2	variance
Q	subgroup quality
$\mathcal{M}$	membership matrix
p_f	success probability for the $f-$th column of the membership matrix
M_f^b	value of the f-th column and b-th row in a membership matrix
s_f^2	sample variance of the f-th column in a membership matrix
H_0	null model

Abstract

Assessing Treatment Effect Heterogeneity: Predictive Covariate Selection and Subgroup Identification

Konstantinos Papangelou

A thesis submitted to The University of Manchester for the degree of Doctor of Philosophy, 2021

A key objective in an interventional study, such as a randomised clinical trial, is the evaluation of heterogeneity of treatment effect in the population. This allows us to identify the most promising intervention for a given observation. In this thesis we approach this by targeting two tightly coupled sub-problems. The first concerns the identification of covariates and the second the identification of subgroups associated with treatment effect heterogeneity.

Regarding the first problem we study an information theoretic approach. This can be motivated by phrasing the predictive covariate selection problem in log-likelihood terms. We study the properties of this approach in the case of randomised studies and evaluate low-dimensional approximations that are better suited for small-sample and/or high-dimensional studies. We identify some limitations and propose extensions based on propensity score weighting and stratification that extend this criterion in scenarios when the treatment assignment depends on the covariates.

Regarding the second problem, we discuss recursive partitioning approaches coupled with weighting methods for treatment effect estimation. The purpose of these methods is to tackle the problem of subgroup identification in the presence of confounders in the data. Finally, studying the literature of subgroup identification we identify a significant number of approaches. Given such a large number of methods to choose from, an important question is how to select the best for a given task. We introduce a framework that uses the subgroup stability as a measure to capture the variations in the identified subgroups due to small changes in the data.

Declaration

No portion of the work referred to in this thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Copyright

Acknowledgements

Firstly, I would like to thank my supervisor Prof. Gavin Brown. I am extremely fortunate to have met him and receive his guidance throughout these years. He taught me not only how to become a researcher but he also provided significant academic and emotional support throughout the process. He has been an inspiration to me and I hope he will continue to guide more students with the same energy.

My deepest thanks to my friend Kostas Sechidis who was always there for me and always willing to provide me the necessary motivation and support to move forward. Additionally, I would like to thank my friends and colleagues from MLO and APT for all the things I learnt from them and for making those years unforgettable. Thanos, Nikos, Idoia, Paris, Smaragda, Sarah, Ainur, Mustafa, Henry, Charlie, Danny, Georgiana, Cameron, Serhat, Andrew, Panagiotis, Eleni, Ioanna thank you for everything. Special thanks to the Centre for Doctoral Training (CDT) in Computer Science, funded by the Engineering and Physical Sciences Research Council (EPSRC) grant [EP/1038099/1], for supporting my studies.

Lastly, I would like to thank my family for their love and the much needed support they provided all those years, specifically: my parents Tania and Zafeiris and my brother Alexandros. I would also like to thank Giannis, Ariadni and Kostas for helping take my mind off when I most needed it.

Chapter 1

Introduction

Over the last decades, researchers in diverse fields, such as statistics, healthcare, sociology, political and economic science, have given great attention to the improvement of methodological tools for assessing causation. At the same time, the increasing number of observational studies that can be used to support the scarce and often unavailable randomised studies has resulted in much attention on the development of methodologies for analysing such data. The primary question for many methodological and empirical studies is estimating the average effect of an intervention (and other supporting analyses), hence answering the question of whether one variable (the treatment[1]) affects another (the outcome). This is known as the *average causal effect* of the treatment/intervention or simply the *average treatment effect*.

Estimation of this quantity may not reveal a causal relationship, and even if it does, the researchers or study's sponsors may be interested in identifying which treatment would be better suited for certain observations. A prominent example is personalised (or precision) medicine, where the selection of the treatment for a patient can be guided by their characteristics. In this thesis, we focus on the question of *how* the treatment affects the outcome of interest, and we focus on assessing the heterogeneity of the effect in subsets of the data. We will explore this problem by deriving a set of novel methodologies, and while we do so, we will explore diverse but interconnected areas of modern causal inference. Overall this thesis discusses a wide range of methodologies regarding the causal inference

[1]Following the terminology adopted by the greatest part of the literature we will use the term "treatment" to refer to an intervention and "treatment effect" to refer to the causal effect of the intervention. These are not necessarily medical treatments, and when they are, it will be clear from the context.

process in the presence of a binary intervention, from identifying covariates of interest to identifying subgroups with desirable characteristics.

1.1 Motivation

In healthcare, the field of personalised medicine studies approaches for identifying the right treatment for each patient. In advertising, we may be interested in the impact of ad exposure which can vary between different groups of customers. In public policy evaluation, a new program may be better suited to certain individuals than others. In these cases, we are interested in understanding the mechanisms by which the treatment affects the outcome of interest and possibly identify subpopulations where we could recommend the most suitable option. In this thesis, we will approach this by exploring the heterogeneity of the effect of a treatment in the population. In particular, we will tackle the following three challenges:

- *Identifying covariates that cause treatment effect heterogeneity.*
- *Identifying subgroups in our data that exhibit treatment effect heterogeneity.*
- *Evaluating subgroup identification algorithms.*

When we measure the overall causal effect of an intervention in some given sample, it is likely the case that this is not going to be the same across all observations. There might exist subsets where there is a much larger effect compared to the average effect or the opposite. If we have collected a set of covariates describing the characteristics of each observation (such as demographics, biomarkers, etc.), we can try to identify which of these covariates are responsible for the observed heterogeneity of the causal effect. We can then try to identify subsets of the data where there is substantial heterogeneity. In the context of clinical trials, these two challenges are interconnected, and they are at the heart of personalised medicine (Lipkovich et al., 2017a). As we will see, there is a large number of methods for identifying subgroups, particularly in randomised studies. Given such a large pool of methods to choose from, the third challenge deals with the problem of evaluating them. This is a non-trivial task since the quantities we are normally interested in (e.g. treatment effects, subgroups, and covariates that define them) are not observed. Let us provide an overview of these challenges.

1.1.1 Identifying Predictive Covariates

Suppose we perform a randomised clinical trial to assess the effectiveness of a novel drug against the standard medical treatment for patients with some type of cancer. We find that patients who got the novel treatment had longer progression-free survival compared to standard care. In hope of identifying potential subsets where the novel drug performs even better, we study how the values of certain biomarkers may change the effect of the treatment. Interestingly, we find that patients who had a specific value for a biomarker (e.g. some gene mutation) and received the novel drug had a longer progression-free survival compared to standard care and this effect is larger than what we observed in the whole sample. Even though this is an unplanned analysis, we can hypothesise that this biomarker may result in heterogeneity of the treatment effect, which can be evaluated by additional studies in the future. The task of identifying subsets of the data that exhibit heterogeneity of the effect of a novel treatment or other desirable characteristics (such as enhanced effects) is called *subgroup identification*. The covariates that interact with the treatment causing the observed heterogeneity are called *predictive*, or *treatment moderators* (Chen et al., 2017).

In the context of clinical trials, researchers are often concerned with distinguishing between *prognostic* and *predictive* biomarkers[2]. In the literature we can identify various definitions of the two types of covariates (e.g. (Simon, 2010; Lipkovich et al., 2017b; Dunn et al., 2013; Ruberg and Shen, 2015)). In the context of the thesis, (solely) prognostic will be a covariate that provides information for predicting the outcome irrespective of the applied treatment (it is not an irrelevant covariate) and does not exhibit an interaction with the treatment. In contrast, a predictive covariate will interact with the treatment. A predictive covariate can also be prognostic, in which case we may discuss about different degrees of predictive and prognostic values or strengths. We notice that even though the definitions of predictive and prognostic covariates appear primarily in a clinical trial context, they can be rather general, and the identification of predictive covariates can be important irrespective of the domain of study. The distinction between the two types of covariates is depicted graphically in figure 1.1.

Subgroup identification is a task that is commonly performed in late-stage

[2]The term "biomarker" is often used in the context of clinical trials. Here, for consistency we will use the term "covariate" to refer to pre-treatment variables irrespective of the context.

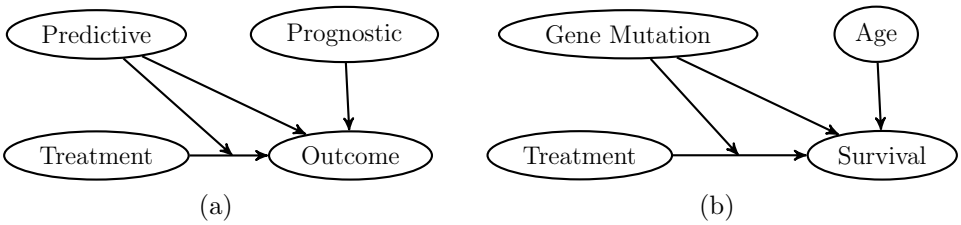


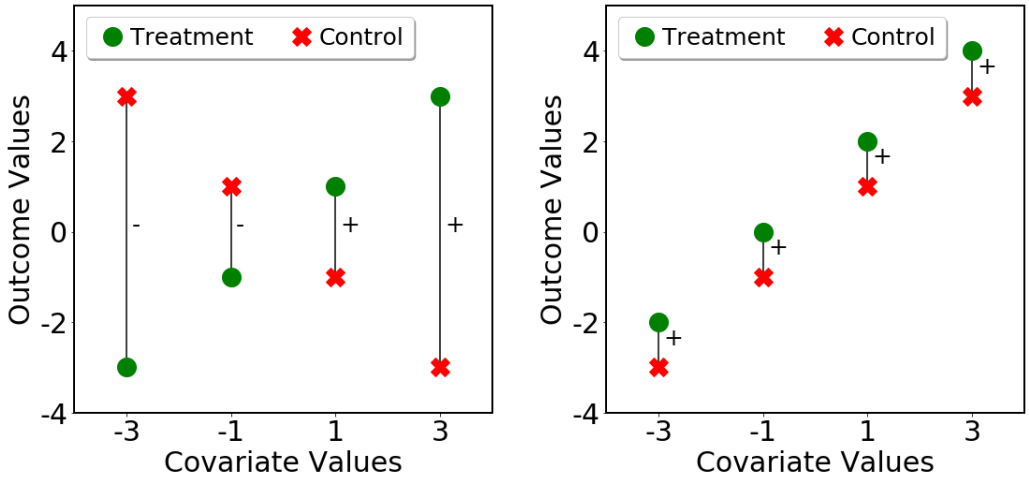
Figure 1.1: Identifying predictive covariates is an important task for designing personalised solutions as they indicate the presence of differential treatment effects (notation adopted from (Dunn et al., 2013)). In this toy example the existence of a particular gene mutation affects the survival of patients treated with the novel drug. On the other hand, age affects the survival of a patient irrespective of whether she gets the novel drug or standard care.

trials due to the presence of a larger sample size. In early-stage trials, higher emphasis can be placed on the selection of potentially predictive covariates (Lipkovich et al., 2017b). Additionally, in the presence of a large number of covariates, we may first try to reduce the dimensionality before performing subgroup identification in order to reduce computational complexity and potentially the number of false discoveries. Even though it is an important question, the identification of predictive covariates has not received much attention. In this thesis, we study this problem in detail, describing existing methodologies (most of which have not been introduced for this task) and suggesting new ones.

1.1.2 Identifying Subgroups of Heterogeneous Effects

As we described, the effect of a treatment is likely to show some variations in the sample. Identifying subgroups with desirable characteristics, such as enhanced effects, allows us to focus future studies on members of the population who are more likely to benefit from a particular treatment and ultimately design tailored solutions, such as tailored therapies (figure 1.2). In this thesis, we focus on *exploratory* subgroup identification; that is, we generate hypotheses rather than testing pre-defined ones. The latter is described as *confirmatory* subgroup analysis (Lipkovich et al., 2017a). Therefore any discoveries we might make by using a subgroup identification algorithm will need to be examined by domain experts and verified in a statistically rigorous manner by performing a confirmatory analysis.

With the increasing interest in salvaging failed studies and providing personalised treatments, there has been a significant effort to develop methodologies



(a) (b)

Figure 1.2: Subgroup identification is the task of identifying subsets with desirable characteristics. In (a) a single covariate defines which observations will benefit from the treatment (indicated with a '+' sign) and which will not benefit (indicated with a '-' sign). On the other hand in (b) all observations benefit equally from the treatment compared to the control, therefore this covariate does not define a subgroup. As we will see later in the thesis the covariate is predictive in (a) and prognostic but not predictive in (b).

for subgroup identification. A detailed review of the problem and an analysis of existing approaches is given by Lipkovich et al. (2017a), while a more recent empirical comparison of 13 algorithms is performed by Loh et al. (2019). Most existing algorithms focus on this problem in the context of randomised studies, where there is no selection bias, i.e. the treatment assigned to each observation is independent of its covariates. We will explore how this can be problematic when applying some existing methods to observational studies. Additionally, as we will see, most algorithms require estimation of treatment effects within subgroups, a problem that is often ignored.

Regarding the latter, studies have shown that modelling the outcome under each treatment arm may increase efficiency and reduce the variance of the estimated effect (Zhang et al., 2008; Bloniarz et al., 2016; Steingrimsson et al., 2017), a method that has also been adopted for subgroup identification (Steingrimsson and Yang, 2019). In observational studies, where there might exist significant selection bias, estimation of treatment effects becomes even more imperative since simple comparisons of average outcomes between the treatment groups will likely

result in biased estimates. We study the problem of subgroup identification in observational studies with no hidden confounders by combining two well-established methodologies: recursive partitioning for identifying subsets of the sample and weighting methods for unbiased estimation of the treatment effect.

Weighting methods are becoming increasingly popular both from an application and a methodological perspective (e.g. (Austin and Stuart, 2015; Kallus, 2020b)). They are easy to implement, often non-parametric approaches that apply a weight to each observation such that on expectation the re-weighted data satisfy some pre-defined properties. When working with observational data, the presence of confounders may result in treatment groups that are not directly comparable. For example, imagine we have a dataset where patients with a specific value for some biomarker receive a novel cancer treatment more often than standard care, while the opposite holds for those that do not have this value. At the same time, the former patients have shorter progression-free survival compared to the latter. A simple comparison of the outcomes of patients under the two values of the treatment might reveal that the novel treatment is not effective. However, this might be due to the fact that we observe more patients who received the novel treatment and were experiencing worse outcomes (irrespective of the applied treatment) compared to the others. In order to estimate the effect of the novel treatment we need to compare groups that are very close in their pre-treatment covariates. Weighting methods try to achieve exactly this by re-weighting the data such that the two treatment groups are *balanced* or *matched*. Weighting is going to be a core subject of this thesis and we will explore a variety of methodologies.

1.1.3 Evaluation of Subgroup Identification Algorithms

In the following chapters we will explore subgroup identification algorithms with diverse characteristics. These are only a few representative examples from a literature that includes a large number of algorithms. Therefore, a natural question that arises is how to evaluate and compare different algorithms. This can be very challenging in an exploratory analysis given that we may not know the true subgroups and the predictive covariates that define them and we never observe the true treatment effect each observation. The latter is due to the fact that we can never observe the outcome of an observation under all possible values of a treatment since we only intervene once.

In the literature of subgroup identification a measure that has been adopted is the *quality* of a subgroup (Foster et al., 2011). If we are interested in identifying subgroups of enhanced effect, then this would quantify the excess treatment effect within a subgroup compared to the whole sample. In the literature of treatment effect estimation, the error of the (unobserved) treatment effect is often quantified by assuming some approximation of the ground truth (Schuler et al., 2018). In addition to these measures, a key concern is also the *reproducibility* of the identified subgroups. In other words, small changes in the data should not affect the subgroup definition. We quantify the *stability* of an algorithm using concepts from the literature of feature selection stability (Nogueira et al., 2017). Combining these measures we show how tasks such as hyper-parameter selection and algorithm comparison can be performed in a multi-objective framework, where different objectives capture different aspects of the algorithm. By navigating in the space of solutions, a practitioner may choose an algorithm that sacrifices e.g. quality of the subgroup if it results in a highly stable result and vice-versa.

1.2 Research Questions

Subgroup identification is one of the primary objectives when analysing the heterogeneity of treatment effects, particularly when the sample size allows that. In some cases performing this task directly may be very challenging either due to the small sample size or because we have a large number of covariates. In these scenarios identifying few covariates that interact with the treatment can help us generate useful hypotheses. We find that this task is usually performed by training some model to infer unknown quantities such as the outcome under each value of the treatment and the treatment effect. In the literature of feature selection, information theoretic criteria (Brown et al., 2012) are a promising alternative that do not require performing inference, tuning hyper-parameters, while also being computationally efficient. Therefore, our first question is:

Q1 : "*How should we adapt information theoretic criteria to identify predictive covariates?*"

From a more practical point of view we explore how existing methods for subgroup identification and treatment effect estimation compare with information theoretic criteria on the task of identifying predictive covariates. In particular, we examine:

Q2 : "*How do information theoretic criteria compare to subgroup identification approaches on the task of predictive covariate selection in marginally randomised studies and how do they perform when the treatment assignment depends on the covariates?*"

In the presence of larger sample sizes and/or a suitable number of covariates we may wish to directly identify subgroups of interest. Given the increasing availability of observational data and focusing on the easy-to-interpret recursive partitioning methods we study:

Q3 : "*How can we modify existing recursive partitioning approaches for subgroup identification in order to account for the presence of confounders in the data?*"

In order to answer this we suggest a methodology based on weighting estimators for estimation of average treatment effects (Kallus, 2020b; Kallus and Santacatterina, 2019b; Kallus et al., 2021). These approaches have some interesting properties – e.g. they do not require a correctly specified parametric model for the treatment and are computationally efficient. From a more practical perspective we examine:

Q4 : "*What are the benefits from using weighting estimators in the context of subgroup identification?*"

In order to answer the above questions we will describe some representative examples of subgroup identification algorithms from a literature that is abundant with methods. Given such a large number of algorithms we then focus on how we can evaluate them. In particular we ask:

Q5 : "*How should we evaluate subgroup identification algorithms in order to account their robustness to small changes?*"

We introduce the concept of stability in this setting and propose a multi-objective framework that captures various desirable aspects of an algorithm.

1.3 Contributions of the Thesis

In order to answer questions Q1, Q3 and Q5 we will study existing methodologies and suggest new ones. For questions Q2 and Q4 we perform empirical studies and

explore the properties of the studied methods in different scenarios. In particular we make the following contributions in each chapter:

- We define the predictive covariate selection problem in log-likelihood terms which in turn results in an information theoretic objective. In the case of marginally randomised studies, properties of the resulting information theoretic criteria are discussed and we show that these can be influenced by the treatment assignment mechanism. Extensions are proposed that use propensity score weighting and stratification and are better suited when the treatment assignment depends on the covariates. We perform a comparison of information theoretic criteria and approaches designed for treatment effect estimation or subgroup identification at the task of predictive covariate selection. Finally, we perform an evaluation of information theoretic criteria in the presence of confounders in the data (Chapter 4).

- We discuss approaches for subgroup identification in observational studies with no hidden confounders that combine recursive partitioning and treatment effect estimation via weighting. We evaluate a recursive partitioning method and its extensions that can handle the presence of confounders in the data (Chapter 5).

- We propose a multi-objective evaluation framework for subgroup identification algorithms that uses the concept of subgroup stability (Chapter 6).

1.4 Structure of the Thesis

In Chapter 2 we present the background material on causal effect estimation. We first introduce the potential outcomes framework and discuss common causal quantities. We then focus on estimation of these quantities using weighting methods, ranging from the popular class of Inverse Propensity Weighting (IPW) estimators to more recent approaches that will be used in this thesis. We conclude the chapter with a brief description of modelling approaches for conditional average treatment effect estimation, a problem that has attracted much attention in the Machine Learning literature.

In Chapter 3 we present some common approaches for variable selection, focusing particularly on information theoretic criteria. These methods will be used

to identify predictive covariates. We then present a categorisation of the covariates that we often need to identify in a study. We conclude this chapter with a categorisation of subgroup identification methods. We emphasise that throughout this thesis we will focus on a few representative approaches.

In Chapter 4 the problem of predictive covariate selection is phrased in log-likelihood terms which results in an information theoretic objective. We discuss some properties of information theoretic criteria and perform an empirical comparison with other frameworks. We present results in three randomised studies with diverse characteristics. We show both theoretically and empirically that the studied approach can be problematic in the presence of confounders in the data and introduce new algorithms that can ameliorate the identified issues.

In Chapter 5 we turn our attention to the problem of subgroup identification. In particular we focus on the case where there might be observed confounders (i.e. observational studies with no hidden confounders). We introduce the two components of our method: a recursive partitioning method and a weighing approach for treatment effect estimation. We evaluate this approach in various scenarios and compare it with standard approaches.

Chapter 6 introduces a framework for evaluating subgroup identification algorithms. In particular, we discuss how stability, a concept particularly used in the context of feature selection, can be adopted to evaluate algorithms based on the robustness of the identified subgroups in small changes in the data. We propose a multi-objective evaluation framework combining stability with measures that capture other characteristics of a subgroup, such as the subgroup quality and the error on the estimated treatment effect.

Chapter 7 reviews the results of this thesis, discusses some limitations and suggests a number of future directions that would further improve the methodology of this thesis.

Chapter 2

Introduction to Causal Effect Estimation

This chapter gives an overview of the problem of treatment effect estimation. We describe approaches for average and conditional average treatment effect estimation in observational and randomised studies. Emphasis will be given on defining the problem and discussing methods that will be adopted in this thesis. We assume the following setting commonly used in the context of causal inference. We have a sample where each observation is described by some pre-treatment covariates and for each one we apply some treatment, which we will assume is binary. For each observation we observe an outcome after the treatment is applied, which is going to be the primary outcome of interest. In the thesis we will focus on intention-to-treat effects, i.e. effects of the assigned treatment which can perhaps be different from the actual treatment received (e.g. due to noncompliance) (Hernán and Robins, 2020).

In order to estimate treatment effects or perform other tasks that will be introduced later such as subgroup identification and variable selection, we need to adopt a mathematical framework that allows us to describe the relevant concepts. To this end we are going to adopt the Neyman-Rubin causal model, also known as the potential outcomes framework (Section 2.1). Section 2.2 introduces various quantities of interest we often need to estimate in an observational study. Section 2.3 describes some basic properties of the propensity score, the stepping stone for most recent weighting methods. In Section 2.4 we describe some common methods for average treatment effect estimation, focusing primarily on recently proposed weighting methods which will be used later in the thesis. Finally, in Section 2.5

we briefly discuss the problem of conditional average treatment effect estimation, i.e. the effect of the treatment for a specific observation in the data.

2.1 The Neyman-Rubin Causal Model

The concept of potential outcomes dates back to the analysis of randomised studies by Fisher (1937) and Neyman (1923). In randomised studies the treatment assignment is either independent of the covariates or comes from a known procedure. Alternatively, the evaluation of an intervention may be performed using historical/observational data where the treatment assignment is in general unknown. The abundance of such data as well as the inability to perform a randomised study in many cases, e.g. for ethical reasons or because it can be impractical, has attracted much attention to the analysis of observational studies. The modern set up was introduced by Rubin (Rubin, 1974, 1977, 1978) and subsequent works have made additional contributions making the potential outcomes framework a popular approach for formalising causal inference problems.

Each observation is described by pre-treatment covariates $\boldsymbol{X} = \boldsymbol{x} \in \mathbb{R}^d$ and a treatment $t \in \mathcal{T} = \{0, 1\}$. For simplicity, we will often refer to $T = 0$ as the control group and $T = 1$ as the treated group. The control group often describes the case of no-treatment, placebo or baseline treatment. The treated group refers to those who received the novel treatment, that is they have been subject to the intervention which we wish to investigate on whether it had some effect with respect to the control group. We define for each observation $\boldsymbol{x}$ and for each value of the treatment the potential outcome $Y(t)$, while the observed outcome is Y. The potential outcome $Y(t)$ denotes what would have happened if an observation had been exposed to treatment $T = t$.

For each subject we only observe the outcome under the actual treatment received, $Y = Y(1)T + Y(0)(1 - T)$, hence the potential outcomes are partially observed. This is known as the "*fundamental problem of causal inference*" (Holland, 1986). If for an observation the received treatment is z then $Y(z)$ is also referred to as the *factual* outcome (always observed), while $Y(\overline{z})$ is known as the *counterfactual* (never observed). The potential outcomes framework provides a mathematical formulation for quantifying the causal effect of a treatment. This can be defined as a comparison between the potential outcomes. A commonly

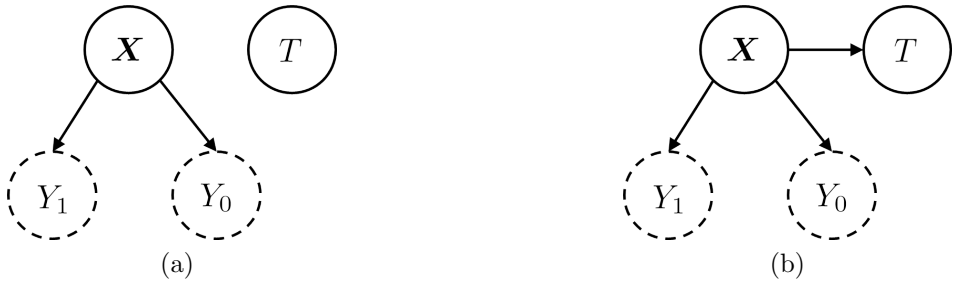


Figure 2.1: (a) **Marginally Randomised Study**: The treatment assignment is independent of the covariates $\boldsymbol{X}$. (b) **Conditionally Randomised Study**: The treatment assignment depends on the pre-treatment covariates $\boldsymbol{X}$ or a subset of those. The latter graph also describes an observational study under the assumption of no hidden confounders.

used measure of the Conditional Average Treatment Effect (CATE) for an observation $\boldsymbol{x}$ is the causal risk difference,

$$\text{CATE}(\boldsymbol{x}) = \mathbb{E}[Y(1) \mid \boldsymbol{x}] - \mathbb{E}[Y(0) \mid \boldsymbol{x}] \tag{2.1}$$

This quantity expresses the effect of intervening on T for a unit $\boldsymbol{x}$. The Average Treatment Effect (ATE) that describes the overall effect in the population can be expressed similarly without conditioning on $\boldsymbol{x}$:

$$\text{ATE} = \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)]$$

In order for treatment effects to be identifiable from the observed data certain assumptions must hold which we will briefly describe. Firstly, a subject's potential outcomes must not depend on the treatment received by other subjects (there is no interference between the subjects) and there is a single version of the treatment (e.g. patients are not administered different doses of the same drug or if they do they are considered as different treatments). This is often referred to as the Stable Unit Treatment Value Assumption (SUTVA) (Rubin, 1990; Imbens and Rubin, 2015) and in the case of a binary treatment it allows us to focus on only two potential outcomes $Y(1)$ and $Y(0)$ so that any causal quantity can be derived by comparisons of these. Under SUTVA, there is consistency of the observed potential outcomes, that is we observe $Y = Y(1) \mid T = 1$ and $Y = Y(0) \mid T = 0$, i.e. we observe $Y(1)$ for the treated and $Y(0)$ for the control.

The following assumptions ensure that treatment is *unconfounded* and there

is sufficient evidence to derive expected values of the potential outcomes.

Assumption 1. *The potential outcomes are independent of the treatment conditioned on the observed covariates:* $(Y(1), Y(0)) \perp\!\!\!\perp T \mid \boldsymbol{x}$

Assumption 2. *The probability of receiving some treatment is bounded away from zero:* $p(T = t \mid \boldsymbol{x}) > 0, \forall t \in \mathcal{T}$

Assumption 1 is also referred to as *unconfoundedness* or *no hidden confounders* while Assumption 2 is also known as *overlap* (Imbens and Wooldridge, 2009; Imbens and Rubin, 2015). Together they form *strong ignorability* (Rosenbaum and Rubin, 1983; Shalit et al., 2017) and result in identifiable estimators of the treatment effect (Imbens and Wooldridge, 2009; Shalit et al., 2017).

In marginally randomised studies (figure 2.1(a)) the treatment is assigned independently of any covariates (Hernán and Robins, 2020). In this case, the above assumptions are implied. In conditionally randomised studies (Hernán and Robins, 2020) the above assumptions hold by design. Such a scenario would occur if for example the treatment is assigned with some fixed probability within strata of the sample defined by some covariate(s) (Hernán and Robins, 2020). In observational studies the above assumptions, even though common in the literature, they are not guaranteed to hold. In the rest of the thesis we will assume that all potential confounders are included in the pre-treatment covariates $\boldsymbol{X}$, i.e. there are no unobserved covariates that are causes of both the treatment and the outcome (Assumption 1). We highlight this assumption as it can be considered a strong one in real-world studies while confirmation of its validity is itself a challenging problem. We will discuss the implications of its violation in Chapter 7. Regarding overlap we will also explore some scenarios where the probability of receiving the treatment can be close to zero or one for some subset of the data.

If follows directly from unconfoundedness and consistency that the conditional average treatment effects are identifiable from the observed data since it holds: $\mathbb{E}[Y(t) \mid \boldsymbol{x}] = \mathbb{E}[Y(t) \mid T = t, \boldsymbol{x}] = \mathbb{E}[Y \mid T = t, \boldsymbol{x}]$. For example, CATE in eq. (2.1) will become $\text{CATE}(\boldsymbol{x}) = \mathbb{E}[Y \mid \boldsymbol{x}, T = 1] - \mathbb{E}[Y \mid \boldsymbol{x}, T = 0]$. This shows that CATE can be estimated if additionally there is sufficient overlap between the treatment groups (see Assumption 2) so that we can estimate the expected values of the outcomes. In observational studies the problem is not trivial and often requires additional steps such that there is sufficient balance between the distributions of the covariates in the two treatment groups (Johansson et al., 2016;

Shalit et al., 2017; Alaa and Schaar, 2018). Even though much of our discussion will focus on average treatment effects, we will return to this subject later in this chapter. We can notice for now that if CATE is identifiable from the observed data, then this will also hold for estimators of population-level effects. The next section defines some commonly used quantities.

2.2 Average Treatment Effect in a Target Population

Besides ATE which was described earlier, practitioners might be interested in estimating average effects for target populations. Two special cases are the Average Treatment effect on the Treated (ATT) and the Average Treatment effect on the Control (ATC) defined using causal risk differences as:

$$\text{ATT} = \mathbb{E}[Y \mid T = 1] - \mathbb{E}[Y(0) \mid T = 1]$$

$$\text{ATC} = \mathbb{E}[Y(1) \mid T = 0] - \mathbb{E}[Y \mid T = 0]$$

Regarding the assumptions we made previously, we require Assumption 1 to hold for $Y(0)$ for ATT estimation and for $Y(1)$ for ATC estimation. We notice that both $\mathbb{E}[Y(1) \mid T = 1]$ and $\mathbb{E}[Y(0) \mid T = 0]$ can be estimated from the data and therefore we require uncondoundedness to hold only for the counterfactual.

The effects described so far answer inherently different questions in an observational study. We remind here the reader that we focus on intention-to-treat effects. Then, given some treatment of interest (e.g. medical treatment, a new policy etc.), ATE is the effect in the overall population, ATT is the effect amongst those who were intended to receive it (the exact mechanism that determined who received it might not be known), while ATC is the effect of switching to the new treatment amongst those who were not intended to receive it. If the treatment is medical then ATC can express the gain from moving subjects from baseline treatment, $T = 0$ to the new treatment $T = 1$ (Tao and Fu, 2019). On the other hand, as described in (Tao and Fu, 2019), ATT might be better suited when studying safety issues related to the new treatment by estimating what would happen if those who got the treatment were switched to the baseline. Another example where ATT might be the quantity of interest is the case of evaluating some new program or policy, in which case the researcher might be interested in

the population that actually participated in the program (Frölich, 2004). Hence, this is often described as the quantity of interest for policy makers, particularly when they are interested in comparing the benefits of the program with its costs (Heckman et al., 1997). On the other hand, ATC can describe the effect of a new program on a population that this has not been implemented in order explore its potential benefits (Wang et al., 2017).

In general the target population may be defined by other criteria such as a sub-population defined by demographic characteristics. For example, in Chapter 5 we discuss in more detail the estimation of the conditional average treatment effect in subgroups of the population. Other examples are the Average Treatment effect among the evenly Matchable (ATM) (Li and Greene, 2013; Samuels, 2017) and the Average Treatment effect on the Overlap population (ATO) (Li et al., 2018). These quantities (a detailed description of which can also be found in (McGowan, 2018)) were originally defined such that they satisfy certain desirable properties, such as smaller variance of the treatment effect and/or increased balance between the treatment groups. The latter will come up very frequently in our discussion and refers to the key property we wish our data to have, that is the distributions of the two treatment groups are matched with respect to specified moments. While this is expected in marginally randomised studies, it will not be the case in observational studies due to the presence of confounders. Under the assumptions described previously the sample analogues of the aforementioned treatment effects can be estimated from the observed data by weighting the observed outcomes with functions of the probability of receiving the treatment, also known as the propensity score. Hence, these are also referred to as Weighted Average Treatment Effects (WATE) (Hirano et al., 2003; Tao and Fu, 2019). Since the propensity score plays a crucial role in causal inference, we will next discuss some key properties.

2.3 The Propensity Score

The propensity score is generally defined as $p(T = 1 \mid \boldsymbol{x}, y(1), y(0))$. Strong ignorability implies that this is independent of the potential outcomes, i.e. the treatment assignment does not depend on what would have happened to a subject had she received a particular treatment and can be expressed as $p(T = 1 \mid \boldsymbol{x}) = e(\boldsymbol{x})$. The propensity score is commonly described as a balancing score. A

balancing score $b(\boldsymbol{x})$ has the following important property: If unconfoundedness holds then the potential outcomes are independent of treatment given $b(\boldsymbol{x})$, i.e. $(Y(1), Y(0)) \perp\!\!\!\perp T \mid b(\boldsymbol{x})$ (Imbens and Rubin, 2015, Lemma 12.2).

Based on this property, propensity score methods are widely applicable in causal effect estimation problems via stratification (Imbens and Rubin, 2015), but also via re-weighting (Robins et al., 2000), or matching. Weighting has recently attracted much attention due to the application of Machine Learning approaches either for more accurate estimation of the propensity score (McCaffrey et al., 2004; Gharibzadeh et al., 2018; McCaffrey et al., 2013; Xie et al., 2019) or for directly balancing the distributions of the treatment groups (Imai and Ratkovic, 2014; Fong et al., 2018; Ning et al., 2020).

For an observation $\boldsymbol{x}$ we can estimate the expected values of the potential outcomes from the observed data as follows (Imbens and Rubin, 2015):

$$\frac{\mathbb{E}[YT \mid \boldsymbol{x}]}{p(T=1 \mid \boldsymbol{x})} = \frac{\mathbb{E}[Y(1)T \mid \boldsymbol{x}]}{p(T=1 \mid \boldsymbol{x})} = \frac{\mathbb{E}[Y(1) \mid T=1, \boldsymbol{x}]p(T=1 \mid \boldsymbol{x})}{p(T=1 \mid \boldsymbol{x})} =$$
$$= \mathbb{E}[Y(1) \mid T=1, \boldsymbol{x}] = \mathbb{E}[Y(1) \mid \boldsymbol{x}]$$

and similarly for $T = 0$. Additionally, for a function of the covariates $u(\boldsymbol{X})$, which defines the population of interest it holds (Tao and Fu, 2019):

$$\mathbb{E}\Big[\frac{(1-T)u(\boldsymbol{X})}{p(T=0 \mid \boldsymbol{x})}\Big] = \mathbb{E}\Big[\frac{Tu(\boldsymbol{X})}{p(T=1 \mid \boldsymbol{x})}\Big] = \mathbb{E}\big[u(\boldsymbol{X})\big]$$

In other words propensity score re-weighting allows us to match the expected values of the covariates (or functions of those) in the population of interest.

In practice the weights are often normalised (e.g. in order to sum to one) (Robins et al., 2000; Tao and Fu, 2019). Depending on the treatment effect of interest (e.g. ATT) different sets of weights will need to be defined. Then the resulting estimators can be shown to be consistent as long as the propensity model is correctly specified. In the next section we delve deeper into the problem of treatment effect estimation with propensity score weights.

2.4 Average Treatment Effect Estimation Methods

This section introduces some common methods for average treatment effect estimation. In particular we focus on the popular Inverse Propensity Weighting, Doubly Robust as well as more recent non-parametric weighting methods.

2.4.1 IPW Methods

In the previous section we described how the propensity score can be used to estimate the potential outcomes and its balancing property. In this section we describe Inverse Propensity Weighting (IPW) methods for treatment effect estimation. There are various ways in which we can derive consistent estimators of the treatment effect of interest (Hirano et al., 2003; Tao and Fu, 2019; Hirano and Imbens, 2001; Robins et al., 2000). For example the following estimators of the sample analogues of ATE, ATT and ATC will be consistent if the estimated propensity score, denoted here as $\hat{e}(\boldsymbol{x})$, is correctly specified.

$$\widehat{\text{SATE}}_{IPW} = \frac{1}{n}\Big(\sum_i \frac{\mathbb{I}(t_i=1)y_i}{\hat{e}(\boldsymbol{x}_i)} - \sum_i \frac{\mathbb{I}(t_i=0)y_i}{1-\hat{e}(\boldsymbol{x}_i)}\Big)$$

$$\widehat{\text{SATT}}_{IPW} = \frac{1}{\sum_i \mathbb{I}(t_i=1)}\Big(\sum_i \mathbb{I}(t_i=1)y_i - \sum_i \mathbb{I}(t_i=0)y_i\frac{\hat{e}(\boldsymbol{x}_i)}{1-\hat{e}(\boldsymbol{x}_i)}\Big)$$

$$\widehat{\text{SATC}}_{IPW} = \frac{1}{\sum_i \mathbb{I}(t_i=0)}\Big(\sum_i \mathbb{I}(t_i=1)y_i\frac{1-\hat{e}(\boldsymbol{x}_i)}{\hat{e}(\boldsymbol{x}_i)} - \sum_i \mathbb{I}(t_i=0)y_i\Big)$$

The following estimators are also consistent and they use normalised weights. In small sample settings this may result in more stable results (Tao and Fu, 2019).

$$\widehat{\text{SATE}}_{IPW} = \frac{\sum_i \mathbb{I}(t_i=1)y_i/\hat{e}(\boldsymbol{x}_i)}{\sum_i \mathbb{I}(t_i=1)/\hat{e}(\boldsymbol{x}_i)} - \frac{\sum_i \mathbb{I}(t_i=0)y_i/(1-\hat{e}(\boldsymbol{x}_i))}{\sum_i \mathbb{I}(t_i=0)/(1-\hat{e}(\boldsymbol{x}_i))}$$

$$\widehat{\text{SATT}}_{IPW} = \frac{\sum_i \mathbb{I}(t_i=1)y_i}{\sum_i \mathbb{I}(t_i=1)} - \frac{\sum_i \mathbb{I}(t_i=0)y_i\hat{e}(\boldsymbol{x}_i)/(1-\hat{e}(\boldsymbol{x}_i))}{\sum_i \mathbb{I}(t_i=0)\hat{e}(\boldsymbol{x}_i)/(1-\hat{e}(\boldsymbol{x}_i))}$$

$$\widehat{\text{SATC}}_{IPW} = \frac{\sum_i \mathbb{I}(t_i=1)y_i(1-\hat{e}(\boldsymbol{x}_i))/\hat{e}(\boldsymbol{x}_i)}{\sum_i \mathbb{I}(t_i=1)(1-\hat{e}(\boldsymbol{x}_i))/\hat{e}(\boldsymbol{x}_i)} - \frac{\sum_i \mathbb{I}(t_i=0)y_i}{\sum_i \mathbb{I}(t_i=0)}$$

For illustrative purposes we will consider the estimator for $\widehat{\text{SATE}}$, then in the limit of data and assuming a correctly specified model for the propensity score we have:

$$\begin{aligned}
&\mathbb{E}\left[\frac{TY}{\hat{e}(\boldsymbol{x})}\right] - \mathbb{E}\left[\frac{(1-T)Y}{1-\hat{e}(\boldsymbol{x})}\right] = \mathbb{E}\left[\frac{1}{\hat{e}(\boldsymbol{x})}\mathbb{E}[TY \mid \boldsymbol{x}]\right] - \mathbb{E}\left[\frac{1}{1-\hat{e}(\boldsymbol{x})}\mathbb{E}[(1-T)Y \mid \boldsymbol{x}]\right] \\
&= \mathbb{E}\left[\frac{1}{\hat{e}(\boldsymbol{x})}\mathbb{E}[TY_1 \mid \boldsymbol{x}]\right] - \mathbb{E}\left[\frac{1}{1-\hat{e}(\boldsymbol{x})}\mathbb{E}[(1-T)Y_0 \mid \boldsymbol{x}]\right] \\
&= \mathbb{E}\left[\frac{1}{\hat{e}(\boldsymbol{x})}\mathbb{E}[T \mid \boldsymbol{x}]\mathbb{E}[Y_1 \mid \boldsymbol{x}]\right] - \mathbb{E}\left[\frac{1}{1-\hat{e}(\boldsymbol{x})}\mathbb{E}[(1-T) \mid \boldsymbol{x}]\mathbb{E}[Y_0 \mid \boldsymbol{x}]\right] \\
&= \mathbb{E}\left[\frac{1}{p(T=1 \mid \boldsymbol{x})}p(T=1 \mid \boldsymbol{x})\mathbb{E}[Y_1 \mid \boldsymbol{x}]\right] \\
&- \mathbb{E}\left[\frac{1}{1-p(T=1 \mid \boldsymbol{x})}(1-p(T=1 \mid \boldsymbol{x}))\mathbb{E}[Y_0 \mid \boldsymbol{x}]\right] \\
&= \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)]
\end{aligned}$$

where the second equality follows from consistency, the third from unconfoundedness and the fourth from assuming a correctly specified model for the propensity score. Note that in practice since the propensity score is an estimated probability, values close to 0/1 may result in arbitrary large or undefined estimated potential outcomes. Therefore it is common practice to perform some post-processing, such as fixing a minimum or maximum value for the weights, based on percentiles of their distribution (Lee et al., 2011).

A key challenge with IPW estimators is the assumption of correct estimation of the propensity score. Additionally, in high-dimensional settings we may need to identify the covariates to be included in the model. Intuitively, we may consider trying to identify those covariates that are predictors of the treatment if the purpose is to build the correct model. In practice the investigator might be interested additionally in other quantities rather only getting unbiased estimates, in which case the inclusion of covariates that are strong predictors of the outcome has been suggested as both empirical evidence and theoretical results show that this can result in variance reduction (Brookhart et al., 2006; Westreich et al., 2011; Williamson et al., 2014). It is worth mentioning that the use of IPW estimators for variance reduction have also been studied in the context of randomised studies (Williamson et al., 2014). The problem of identifying a suitable set of covariates to include in the propensity model can also be partially ameliorated by using Machine Learning approaches (McCaffrey et al., 2004; Gharibzadeh et al.,

2018; McCaffrey et al., 2013; Xie et al., 2019). In any case when performing our simulations, where we have knowledge of the true propensity score, we will explore both correct and incorrect specifications. In the next section we describe doubly robust estimators that relax the requirement of a correct propensity score.

2.4.2 Doubly Robust Methods

A treatment effect estimator is doubly robust if it is consistent when either the outcome model or the propensity model are correctly specified (Bang and Robins, 2005). A doubly robust estimator can correct the potential miss-estimation of the propensity score by additionally estimating the potential outcomes. This is normally performed by fitting some model on the the factual outcome and then inferring the missing counterfactuals. The literature of causal inference and missing data analysis are awash with approaches that satisfy double robustness (e.g. (Bang and Robins, 2005; Funk et al., 2011; Koch et al., 2018; Schuler and Rose, 2017; Tao and Fu, 2019)). Here we are going to describe the most commonly used Doubly Robust (DR) estimators for three common quantities of interest: ATE, ATT and ATC.

Let $\hat{m}_1(\boldsymbol{x})$, $\hat{m}_0(\boldsymbol{x})$ be some models trained using as target the observed outcomes under treatment $T = 1$ and $T = 0$ respectively. Then the following estimators are doubly robust (Tao and Fu, 2019):

$$\widehat{\text{SATE}}_{DR} = \frac{1}{n}\sum_i \Big[\hat{m}_1(\boldsymbol{x}_i) - \hat{m}_0(\boldsymbol{x}_i) + \frac{\mathbb{I}(t_i = 1)}{\hat{e}(\boldsymbol{x}_i)}(y_i - \hat{m}_1(\boldsymbol{x}_i)) - \frac{\mathbb{I}(t_i = 0)}{1 - \hat{e}(\boldsymbol{x}_i)}(y_i - \hat{m}_0(\boldsymbol{x}_i))\Big]$$

$$\widehat{\text{SATT}}_{DR} = \frac{1}{\sum_i \mathbb{I}(t_i = 1)}\sum_i \Big[\mathbb{I}(t_i = 1)y_i - \Big(\frac{\hat{e}(\boldsymbol{x}_i)\mathbb{I}(t_i = 0)}{1 - \hat{e}(\boldsymbol{x}_i)}y_i + \frac{\mathbb{I}(t_i = 1) - \hat{e}(\boldsymbol{x}_i)}{1 - \hat{e}(\boldsymbol{x}_i)}\hat{m}_0(\boldsymbol{x}_i)\Big)\Big]$$

$$\widehat{\text{SATC}}_{DR} = \frac{1}{\sum_i \mathbb{I}(t_i = 0)}\sum_i \Big[\Big(\frac{1 - \hat{e}(\boldsymbol{x}_i)}{\hat{e}(\boldsymbol{x}_i)}\mathbb{I}(t_i = 1)y_i - \frac{\mathbb{I}(t_i = 1) - \hat{e}(\boldsymbol{x}_i)}{\hat{e}(\boldsymbol{x}_i)}\hat{m}_1(\boldsymbol{x}_i)\Big) - \mathbb{I}(t_i = 0)y_i\Big]$$

For completeness of the presentation we will now describe the double robustness

property using $\widehat{\text{SATE}}_{DR}$. In the limit of data we have:

$$\widehat{\text{SATE}}_{DR} \simeq \widehat{\text{ATE}}_{DR} = \mathbb{E}\big[\hat{m}_1(\boldsymbol{x})\big] - \mathbb{E}\big[\hat{m}_0(\boldsymbol{x})\big] + \overset{\text{(a)}}{\mathbb{E}\Big[\frac{TY}{\hat{e}(\boldsymbol{x})}\Big]} - \overset{\text{(b)}}{\mathbb{E}\Big[\frac{T\hat{m}_1(\boldsymbol{x})}{\hat{e}(\boldsymbol{x})}\Big]} - \overset{\text{(c)}}{\mathbb{E}\Big[\frac{(1-T)Y}{1-\hat{e}(\boldsymbol{x})}\Big]} + \overset{\text{(d)}}{\mathbb{E}\Big[\frac{(1-T)\hat{m}_0(\boldsymbol{x})}{1-\hat{e}(\boldsymbol{x})}\Big]}$$

Suppose the propensity model is correctly specified, then in the limit of data we have $\hat{e}(\boldsymbol{x}) \simeq e(\boldsymbol{x}) = p(T = 1 \mid \boldsymbol{x})$. Each term can be written as follows:

$$\begin{aligned}
\text{(a)} &= \mathbb{E}\Big[\mathbb{E}\Big[\frac{TY}{\hat{e}(\boldsymbol{x})} \mid \boldsymbol{x}\Big]\Big] = \mathbb{E}\Big[\frac{1}{\hat{e}(\boldsymbol{x})}\mathbb{E}\Big[TY(1) \mid \boldsymbol{x}\Big]\Big] \\
&= \mathbb{E}\Big[\frac{1}{\hat{e}(\boldsymbol{x})}\mathbb{E}[T \mid \boldsymbol{x}]\mathbb{E}[Y(1) \mid \boldsymbol{x}]\Big] = \mathbb{E}\Big[\frac{1}{p(T=1 \mid \boldsymbol{x})}p(T = 1 \mid \boldsymbol{x})\mathbb{E}[Y(1) \mid \boldsymbol{x}]\Big] \\
&= \mathbb{E}[Y(1)] \\
\text{(b)} &= \mathbb{E}\Big[\mathbb{E}\Big[\frac{T\hat{m}_1(\boldsymbol{x})}{\hat{e}(\boldsymbol{x})} \mid \boldsymbol{x}\Big]\Big] = \mathbb{E}\Big[\frac{\hat{m}_1(\boldsymbol{x})}{\hat{e}(\boldsymbol{x})}\mathbb{E}\Big[T \mid \boldsymbol{x}\Big]\Big] \\
&= \mathbb{E}\Big[\frac{\hat{m}_1(\boldsymbol{x})}{p(T=1 \mid \boldsymbol{x})}p(T = 1 \mid \boldsymbol{x})\Big] = \mathbb{E}[\hat{m}_1(\boldsymbol{x})] \\
\text{(c)} &= \mathbb{E}\Big[\mathbb{E}\Big[\frac{(1-T)Y}{1-\hat{e}(\boldsymbol{x})} \mid \boldsymbol{x}\Big]\Big] = \mathbb{E}\Big[\frac{1}{1-\hat{e}(\boldsymbol{x})}\mathbb{E}\Big[(1-T)Y(0) \mid \boldsymbol{x}\Big]\Big] \\
&= \mathbb{E}\Big[\frac{1}{1-\hat{e}(\boldsymbol{x})}(1 - \mathbb{E}[T \mid \boldsymbol{x}])\mathbb{E}[Y(0) \mid \boldsymbol{x}]\Big] \\
&= \mathbb{E}\Big[\frac{1}{p(T=0 \mid \boldsymbol{x})}p(T = 0 \mid \boldsymbol{x})\mathbb{E}[Y(0) \mid \boldsymbol{x}]\Big] = \mathbb{E}[Y(0)] \\
\text{(d)} &= \mathbb{E}\Big[\mathbb{E}\Big[\frac{(1-T)\hat{m}_0(\boldsymbol{x})}{1-\hat{e}(\boldsymbol{x})} \mid \boldsymbol{x}\Big]\Big] = \mathbb{E}\Big[\frac{\hat{m}_0(\boldsymbol{x})}{1-p(T=1 \mid \boldsymbol{x})}(1 - p(T = 1 \mid \boldsymbol{x}))\Big] = \\
&= \mathbb{E}[\hat{m}_0(\boldsymbol{x})]
\end{aligned}$$

So putting everything together we have $\widehat{\text{ATE}}_{DR} = \text{ATE} = \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)]$. Now suppose that the outcome models are correctly specified, i.e. $\hat{m}_t(\boldsymbol{x}) \simeq m_t(\boldsymbol{x}) = \mathbb{E}[Y(t) \mid \boldsymbol{x}]$. Then we can easily see that (a) and (b) cancel out and the same applies for (c) and (d), hence $\widehat{\text{ATE}}_{DR} = \mathbb{E}\big[\hat{m}_1(\boldsymbol{x})\big] - \mathbb{E}\big[\hat{m}_0(\boldsymbol{x})\big] = \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)]$. We can similarly show the above property for the other estimators.

2.4.3 Weighting Estimators

A challenge with IPW estimators is that balance between the treatment groups is satisfied asymptotically and under a correct propensity model and therefore they may result to poor small-sample performance. The problem becomes more challenging under strong confounding since we also have to handle potentially extreme values. Imai and Ratkovic (2014) note that the propensity score is appropriate if it results in balanced groups, while some authors highlight the time-consuming nature of propensity weighting methods (Hainmueller, 2012; Imai and Ratkovic, 2014). As noted by Imai and Ratkovic (2014), researchers may need to continuously defined a propensity model and check for balance until the latter is satisfied. To overcome these issues most recent works focus on estimating weights that satisfy some pre-specified conditions such as equality of certain moments of the distributions of covariates in the treatment groups. We will now discuss some of these methods.

Based on the above, one choice is to seek for the parameters of the propensity model that simultaneously optimise the likelihood of treatment but also achieve balance between the treatment groups. This is the objective optimised by the Covariate Balancing Propensity Score (CBPS) (Imai and Ratkovic, 2014). Suppose again we have a dataset $\mathcal{D} = \{\boldsymbol{x}_i, t_i, y_i\}_{i=1}^n$, where $\boldsymbol{x}_i$ are d-dimensional vectors. Assuming the propensity score follows the model,

$$\hat{e}(\boldsymbol{x}) = \frac{1}{1 + \exp(-\boldsymbol{\beta}^T \boldsymbol{x})}$$

the parameters $\boldsymbol{\beta} \in \mathbb{R}^d$ are optimised so that they satisfy the following conditions (Imai and Ratkovic, 2014):

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n} \frac{\mathbb{I}(t_i = 1)\hat{e}'(\boldsymbol{x}_i)}{\hat{e}(\boldsymbol{x}_i)} - \frac{\mathbb{I}(t_i = 0)\hat{e}'(\boldsymbol{x}_i)}{1 - \hat{e}(\boldsymbol{x}_i)} = 0 \\ \frac{1}{n}\sum_{i=1}^{n} \frac{\mathbb{I}(t_i = 1)\boldsymbol{x}_i}{\hat{e}(\boldsymbol{x}_i)} - \frac{\mathbb{I}(t_i = 0)\boldsymbol{x}_i}{1 - \hat{e}(\boldsymbol{x}_i)} = 0 \end{aligned} \tag{2.2}$$

The first equation is the first-order condition satisfied by maximising the log-likelihood function for the propensity model (i.e. taking the first derivative and equating it to zero). The second equation is the balancing condition that states the re-weighted mean in the treatment group should be equal to the re-weighted

mean in the control group (Imai and Ratkovic, 2014). The authors propose learning the parameters using the generalised method of moments with the above moment conditions. This can be adapted according to the quantity of interest (i.e. ATT or ATC) and has also been extended to continuous treatments (Fong et al., 2018) and settings where $d \gg n$ (Ning et al., 2020).

We can argue that since balance is the primary objective we should try to learn weights that optimise this – an approach commonly adopted in the literature of treatment effect estimation. A prominent example is the entropy balancing estimator (Hainmueller, 2012) which minimises the Kullback Leibler divergence between the distribution of the weights and a target distribution (usually the uniform) such that the re-weighted samples match with respect to user-specified moments. The user can specify both which covariates will be matched and also their moments. For simplicity let us focus on ATT, where we only need to assign weights in the control observations and suppose we wish to match the means of all the covariates. In this case the method proposed by Hainmueller (2012) can be expressed as:

$$\begin{aligned} \min \ & \sum_{i:t_i=0} w_i \log \frac{w_i}{b_i} \\ \text{s.t.} \ & \sum_{i:t_i=0} w_i x_i^k = \sum_{i:t_i=1} x_i^k, \ k \in \{1, ..., d\} \\ & \sum_{i:t_i=0} w_i = 1, \ \boldsymbol{w} \succeq 0 \end{aligned}$$

where x_i^k is the value of the k-th covariate for the i-th example (these could be replaced by higher-order terms in order to capture moments other than the means of the covariates), and b_i are base weights selected by the user. These weights can be set to $1/n_c$, where n_c is the number of observations in the control group (Hainmueller, 2012). Intuitively, if the data were from a marginally randomised experiment then all weights should be equal to $1/n_c$. Then the above optimisation problem identifies the weights closer to that uniform distribution so that pre-specified moments of the covariates match. The other constraints ensure that the weights are positive and sum to one.

Zubizarreta (2015) solve a similar problem setting the objective function to $\|\boldsymbol{w} - \overline{\boldsymbol{w}}\|_2^2$, where $\overline{\boldsymbol{w}}$ the mean value of the weights. Their motivation is to search for the weights with minimal variance that achieve the specified balancing conditions. Athey et al. (2018) propose an approach better suited for high-dimensional data. The authors assume linearity of the outcomes and estimate the weights that

minimise the residual after applying a sparse linear model for the control units. While some of the above approaches are defined heuristically in (Kallus, 2020b; Kallus et al., 2021) the authors derive approaches by targeting directly the conditional bias and mean squared error of the sample ATT and ATE respectively. This is the approach we will adopt in Chapter 5, where we will discuss it in more detail. We will focus on this method primarily due to its motivation and empirical performance but also because it requires less effort by the user, since we do not need to specify beforehand the matching conditions that need to be satisfied.

On Balance and Bias Reduction: A Motivating Example

Let us, for the sake of exposition, focus on SATT and let us assume that the relationship between the potential outcome $Y(0)$ and the covariates follows the model:

$$m_0(\boldsymbol{X}) = \boldsymbol{X}\boldsymbol{\alpha} + \epsilon$$

where here we assume linearity in the original set of covariates and ϵ is a zero mean error term. The counterfactual $Y(0) \mid T = 1$ is estimated using a weighting method by re-weighting the observed outcome $Y(0) \mid T = 0$. In other words the estimated SATT is:

$$\widehat{\text{SATT}}^w = \frac{1}{n_t} \sum_{i:t_i=1} y_i - \sum_{i:t_i=0} w_i y_i$$

where the weights are normalised to sum to 1. Then using the linearity assumption for the counterfactual it can be shown that the absolute conditional bias is (Kuang et al., 2019):

$$|\mathbb{E}[\widehat{\text{SATT}}^w - \text{SATT} \mid \{\boldsymbol{x}_i, t_i\}_{i=1}^n]| = |((\overline{\boldsymbol{X}^1})^T - \boldsymbol{w}^T\boldsymbol{X}^0)\boldsymbol{\alpha}| \tag{2.3}$$

where we omitted the error term which will be on expectation equal to zero. Here $\boldsymbol{X}^1$ are the covariates in the treated group and $\boldsymbol{X}^0$ in the control group. Then applying Hölder's inequality results in the following:

$$|((\overline{\boldsymbol{X}^1})^T - \boldsymbol{w}^T\boldsymbol{X}^0)\boldsymbol{\alpha}| \leq \|\boldsymbol{\alpha}\|_p \|(\overline{\boldsymbol{X}^1})^T - \boldsymbol{w}^T\boldsymbol{X}^0\|_q$$

with $1/p + 1/q = 1$. The estimated counterfactual can be expressed as:

$$((\overline{\boldsymbol{X}^1})^T - \boldsymbol{w}^T\boldsymbol{X}^0)\hat{\boldsymbol{\alpha}} + \sum_{i:t_i=0} w_i y_i = \hat{\boldsymbol{\alpha}}^T\overline{\boldsymbol{X}^1} + \sum_{i:t_i=0} w_i(y_i - \hat{\boldsymbol{\alpha}}^T\boldsymbol{x}_i) \tag{2.4}$$

The first equation motivates the bias minimisation procedure under a parametric model (Kuang et al., 2019), i.e. minimise the bias and estimate the counterfactual by weighting. The second is the estimator used by Athey et al. (2018). Their motivation is to use the last term to capture the residuals of the model after re-weighting (Athey et al., 2018). The authors focus on optimising the balancing condition $\|(\overline{\boldsymbol{X}^1})^T - \boldsymbol{w}^T\boldsymbol{X}^0\|_\infty$ and estimation of the parameters $\boldsymbol{\alpha}$ separately. In (Kuang et al., 2019) the authors focus on direct minimisation of the bias term under a linear model by jointly learning the weights and parameters of the model. Both approaches target the problem of treatment effect estimation in a high dimensional setting.

To show the relationship between balance maximisation and bias reduction we identify weights that minimise $\|(\overline{\boldsymbol{X}^1})^T - \boldsymbol{w}^T\boldsymbol{X}^0\|_2^2$. We assume the outcome is generated by the following linear model $Y = \sum_{i=1}^{8} X_i + 2T(X_7 + X_8 + 1) + \epsilon$ and the treatment assignment model is $\text{logit}(P(T = 1 \mid X)) = \sum_{i=1}^{5} X_i + \epsilon$, where $\epsilon \sim \mathcal{N}(0, 1)$. The covariates are independent and normally distributed following, $X_i \sim \mathcal{N}(0, 1),\ i = 1, ..., d$. We use $n = 1000$ observations and $d = 50$ covariates. All results are averaged over $N = 500$ realisations. We report the following (Kuang et al., 2019):

$$\text{Absolute Bias} = |\frac{1}{N}\sum_{i=1}^{N}\widehat{\text{SATT}}_i^w - \text{ATT}|$$

As we can observe in table 2.1 when both models are correctly specified the optimisation problem minises the bias. In the second case we modify the term X_1 of the outcome to $\log(1 + \exp(X_1))$ so that this is the only non-linear term. We observe that the balance here is not affected, which means that the mean of this term in the treated group is close to the mean in the control group after weighting. We note here that a method that relies on minimising the bias under an assumed parametric model would be influenced in this case. In the third scenario we add the same term in the treatment model so that now there is a confounder that affects both treatment and outcome in a non-linear fashion. Now the performance of the method starts deteriorating and we observe an increase in the bias. This is even more prominent in the fourth case where we replace the terms X_1, X_2 in both models with $\log(1 + \exp(X_1))X_2^2$. In this case balancing simply on the means of X_1 and X_2 separately is not enough to reduce the bias. This toy example highlights that simply matching the means of the covariates

Table 2.1: Results on a toy example showing how identifying weights that maximise the balance between the treatment groups affects the bias of the estimated ATT under different specifications.

specification of Y	specification of T	Absolute bias of estimated ATT
✓	✓	0.08
✗	✓	0.07
✗	✗	0.12
✗	✗	0.73

could be enough under some conditions but more complex approaches will be required when there are non-linear and non-additive terms. In Chapter 5 we will discuss such approaches in more detail.

2.5 Conditional Average Treatment Effect Estimation

In the previous sections we discussed the problem of estimating average treatment effects for a target population. In this section we will discuss some commonly used approaches for CATE estimation. We note that there is a rich literature for tackling this issue and here we will focus primarily on methods that have been used in the context of subgroup identification which will be discussed later in the thesis.

A simple approach would be to use a single model for the outcome as a function of the covariates and the treatment. In this case the target function that we wish to learn can be denoted as $f(t_i, \boldsymbol{x}_i)$ (Here we will use f to denote the functions we wish to learn instead of m in order to distinguish between the different types of learning methods). In practice, interactions between the pre-treatment covariates $\boldsymbol{X}$ and the treatment T may also be added. In the case of linear models such interactions are necessary in order to capture heterogeneous treatment effects. Given an estimate of the observed outcome $\hat{f}(t_i, \boldsymbol{x}_i)$ the causal risk difference for a subject $\boldsymbol{x}_i$ is $\widehat{\text{CATE}}(\boldsymbol{x}_i) = \hat{f}(1, \boldsymbol{x}_i) - \hat{f}(0, \boldsymbol{x}_i)$. In other words, for an observation $\boldsymbol{x}_i$ we estimate the factual outcome under the assigned treatment t_i and then modify the value t_i in order to get an estimate of the counterfactual. Using the nomenclature of Künzel et al. (2019) this is also referred to as an S-learner. This is summarised below.

Modelling Approach 1. *The Single-Model approach estimates the observed outcome y_i using the sample $\{\boldsymbol{x}_i, t_i\}_{i=1}^n$: $\hat{f}(t_i, \boldsymbol{x}_i)$. During inference the predicted causal risk difference is the difference between estimated outcomes fixing the value of t_i as either $t_i = 1$ or $t_i = 0$: $\widehat{CATE}(\boldsymbol{x}_i) = \hat{f}(1, \boldsymbol{x}_i) - \hat{f}(0, \boldsymbol{x}_i)$.*

An alternative approach is to treat the problem of estimating treatment effects as two separate problems, one for each treatment group. In this case, we perform two steps: 1. Estimate $\hat{f}_0(\boldsymbol{x}_i)$ using the data for the control group and 2. Estimate $\hat{f}_1(\boldsymbol{x}_i)$ using the treated data. Here $\hat{f}_1$ and $\hat{f}_0$ can in general be different models. For a given example the causal risk difference can be estimated as the difference between the outputs of the two models. This is also referred to as a T-learner (Künzel et al., 2019). The procedure is summarised below.

Modelling Approach 2. *The Two-Model approach estimates the outcome potential outcome for $T = 1$ from the sample $\{\boldsymbol{x}_i, y_i\}_{i:t_i=1}$: $\hat{f}_1(\boldsymbol{x}_i)$ and the potential outcome under $T = 0$ from the sample $\{\boldsymbol{x}_i, y_i\}_{i:t_i=0}$: $\hat{f}_0(\boldsymbol{x}_i)$. During inference the predicted causal risk difference is the difference between the estimated outcomes: $\widehat{CATE}(\boldsymbol{x}_i) = \hat{f}_1(\boldsymbol{x}_i) - \hat{f}_0(\boldsymbol{x}_i)$.*

Künzel et al. (2019) describe some properties of these two modelling approaches. One important property described for the S-learner is that since the treatment variable is included in the set of covariates, it might be ignored when using regularised models (will be discussed in the next chapter) or methods that rely on partitioning of the space (e.g. Tree-based methods) (Künzel et al., 2019). On the other hand the T-learner can be more flexible since it will fit a new model for the outcome under each value of the treatment. The authors find empirically this approach to perform better when the functional forms of the potential outcomes share few similarities. The two approaches will also use different sample sizes, since the first modelling approach uses all data to build the model, while the second uses only a fraction of those.

An alternative approach, commonly used in causal inference and the closely related area of uplift modelling is the change of the outcome variable (Athey and Imbens, 2016; Jaskowski and Jaroszewicz, 2012). Consider a continuous outcome $Y \in \mathbb{R}$. The transformed outcome can be expressed as:

$$Y_i^* = Y_i \cdot \frac{T_i - e(\boldsymbol{x}_i)}{e(\boldsymbol{x}_i)(1 - e(\boldsymbol{x}_i))} \tag{2.5}$$

where $e(\boldsymbol{x}_i) = p(T_i = 1 \mid \boldsymbol{x}_i)$ is the propensity score. Athey and Imbens (2015) show the following, under the assumption of unconfoundedness, which results directly from applying IPW:

$$\mathbb{E}[Y_i^* \mid \boldsymbol{x}_i] = \text{CATE}(\boldsymbol{x}_i)$$

Jaskowski and Jaroszewicz (2012) follow a similar approach for the case of binary outcomes, $Y \in \{0, 1\}$, marginal randomisation and balanced groups (1:1). They define a new variable $Y_i^* = Y_i T_i + (1 - Y_i)(1 - T_i)$. For the new variable we have $Y_i^* = 1$, if $Y_i = 1, T_i = 1$ or $Y_i = 0, T_i = 0$ and $Y_i^* = 0$ otherwise. Assuming unconfoundedness then they relate Y^* with the causal risk difference as follows.

$$p(Y_i = 1 \mid T_i = 1, \boldsymbol{x}_i) - p(Y_i = 1 \mid T_i = 0, \boldsymbol{x}_i) = 2 \cdot p(Y_i^* = 1 \mid \boldsymbol{x}_i) - 1$$

The left hand side of the above equation is the causal risk difference $\text{CATE}(\boldsymbol{x}_i)$ which can be estimated directly from the new variable Y^*.

Modelling Approach 3. *The Outcome Transformation approach first estimates the propensity score $\hat{e}(\boldsymbol{x}_i)$ and then builds a model on the covariates $\boldsymbol{X}$ using the transformed variable Y^* as the response. During inference the predicted causal risk difference is simply the output of the model: $\widehat{CATE}(\boldsymbol{x}_i) \simeq \mathbb{E}[Y_i^* \mid \boldsymbol{x}_i]$.*

A practical difference between Approaches 1,2 and Approach 3 is that the former estimate the potential outcomes while the latter estimates the treatment effect directly. In other words, with the outcome transformation method we can avoid modelling the main effect. We will refer to the first two approaches as counterfactual models, as they estimate both the factual and counterfactual outcome. The aforementioned approaches have been used extensively in the literature of causal effect estimation (Johansson et al., 2016; Shalit et al., 2017; Alaa and Schaar, 2018; Athey and Imbens, 2015; Foster et al., 2011) adopting different types of Machine Learning models.

There are however approaches that do not fall in the above categories. For example the X-learner (Künzel et al., 2019) is suited for scenarios where one of the treatment groups is significantly larger than the other. Other popular approaches are Causal Trees (CT) and Causal Forests (CF) which follow a recursive partitioning approach and estimate the treatment effects locally within the leaves (Athey and Imbens, 2016; Wager and Athey, 2018a; Athey et al., 2019;

Athey and Wager, 2019). CFs (Wager and Athey, 2018a) are built so that for each example the outcome is either used to derive the split or to estimate the treatment effect within the leaf nodes. This is referred to as "honest" splitting (Wager and Athey, 2018a). They suggest two approaches. In the first, they build a classification tree using the treatment to derive the splits. In the second, they split the data in two parts using one to split the data and one for the estimation. In the second approach, the variance of the estimated treatment effect for each example is used as the splitting criterion. In Chapter 6 we will use CF as implemented in the package *grf* (Tibshirani et al., 2020) which is more closely related to the second approach (Athey et al., 2019). As described in (Athey and Wager, 2019), firstly two forests are built in order to estimate $\mathbb{E}[Y_i \mid \boldsymbol{x}_i]$, denoted as $\hat{f}(\boldsymbol{x}_i)$ and the propensity score $\hat{e}(\boldsymbol{x}_i)$. These are then used to get the out-of-bag predictions, denoted with the superscript $(-i)$ (the model was trained without using the i-th example to derive the splits). If we set as $y_i^{res} = y_i - \hat{f}^{(-i)}(\boldsymbol{x}_i)$ and $e_i^{res} = t_i - \hat{e}^{(-i)}(\boldsymbol{x}_i)$ then the treatment effect for an example $\boldsymbol{x}$ is estimated as $\sum_i \eta_i(\boldsymbol{x}) y_i^{res} e_i^{res} / \sum_i \eta_i(\boldsymbol{x})(e_i^{res})^2$. Here the weights $\eta_i(\boldsymbol{x})$ denote how many times the i-th training example is in the same leaf node as the test example $\boldsymbol{x}$.

Additionally, in the last years there has been a significant growth of neural network based approaches for treatment effect estimation (Yao et al., 2018; Hartford et al., 2017; Künzel et al., 2018; Kallus, 2020a; Li and Fu, 2017; Shi et al., 2019; Yoon et al., 2018; Louizos et al., 2017; Alaa et al., 2017; Shalit et al., 2017; Johansson et al., 2016). In the following chapters we will discuss/apply some methods for subgroup identification that estimate the conditional average treatment effect as part of their process. They all belong in one of the aforementioned modelling approaches.

2.6 Chapter Summary

In this chapter we firstly introduced the Neyman-Rubin Causal Model (Fisher, 1937; Neyman, 1923; Rubin, 1974) which we will use in the rest of the thesis to describe mathematically the causal effect estimation problems we will try to tackle. We then introduced the problem of average treatment effect estimation and provided a brief description of some commonly used approaches. We described Inverse Propensity Weighting (IPW) and Doubly Robust (DR) methods mainly due to their popularity in the analysis of observational data. We then

focused on the most recent weighting estimators highlighting their advantages over IPW methods. These estimators will be the focus of Chapter 5 where we will use them to tackle the problem of subgroup identification in the presence of confounders. Lastly we described some simple modelling approaches for the estimation of conditional average treatment effects. In the next chapter we will see how these approaches have been used for subgroup identification. We will additionally introduce the problem of variable selection in the presence of interventions.

Chapter 3

Variable Selection and Subgroup Identification

In this chapter we discuss some methods for variable selection and subgroup identification that we will use in the rest of this thesis. With variable selection we commonly refer to the problem of identifying a few variables that are relevant for the task at hand. This results in smaller, easier to interpret and more computationally efficient predictive models. Most importantly it provides insights to the practitioner about the data generating mechanism. The latter is exemplified by the use of variable selection in clinical trial data where understanding the relationships between treatment, pre-treatment covariates (e.g. demographics, genetic factors) and outcome allow the practitioner to generate hypotheses and better understand the underlying mechanisms.

The selection of important variables may be coupled with building a predictive model and we will refer to these methods as *model-based*. We describe some of these methods in Section 3.1. Other approaches based on hypothesis testing and scoring functions, such as the mutual information, identify variables of interest without performing inference or requiring some predictive model. We discuss information theoretic variable selection in Section 3.2. The problem of identifying variables of interest becomes more complex in the presence of interventions, where in contrast to traditional supervised learning, we may not only be interested in identifying variables predictive of the outcome. In Section 3.3 we discuss the various types of variables we often need to identify when facing a causal inference problem. In this thesis we will focus primarily on identifying *predictive* pre-treatment covariates. In Section 3.4 we discuss three common frameworks

for performing this task. These approaches are all model-based, but estimate different quantities. These approaches will be compared against an information theoretic method in the next chapter.

3.1 Model-based Variable Selection

In this section we will describe approaches that perform variable selection coupled with prediction. We will focus on two common approaches: regularised linear models (Hastie et al., 2009; Tibshirani, 1996) and recursive partitioning approaches (Breiman, 2001). Throughout this section we consider a dataset $\mathcal{D} = \{\boldsymbol{x}_i, y_i\}_{i=1}^n$ where y_i is the outcome and $\boldsymbol{x}_i$ is a d-dimensional vector of the variables.

3.1.1 Regularised Models

A commonly used approach for variable selection is the Least Absolute Shrinkage and Selection Operator (LASSO) (Tibshirani, 1996). It identifies the parameters $\hat{\theta_0}, \hat{\boldsymbol{\theta}}^{lasso}$ by solving the following optimisation problem:

$$\begin{aligned} &\min_{\theta_0, \boldsymbol{\theta}} \ \frac{1}{n} \sum_{i=1}^{n} \mathcal{J}(\boldsymbol{x}_i, y_i; \theta_0, \boldsymbol{\theta}) \\ &\text{s.t. } \|\boldsymbol{\theta}\|_1 \leq k \end{aligned}$$

where $\mathcal{J}(\boldsymbol{x}_i, y_i; \theta_0, \boldsymbol{\theta})$ is the loss function for a linear model $\text{logit}(f(\boldsymbol{x}_i)) = \theta_0 + \boldsymbol{\theta}^T \boldsymbol{x}_i$ with parameters $\{\theta_0, \boldsymbol{\theta}\}$. For binary outcomes $y_i \in \{0, 1\}$ this corresponds to the negative log-likelihod:

$$\mathcal{J}(\boldsymbol{x}_i, y_i; \theta_0, \boldsymbol{\theta}) = -y_i \text{log } f(\boldsymbol{x}_i) - (1 - y_i) \text{log } (1 - f(\boldsymbol{x}_i))$$

while for continuous outcome $y_i \in \mathbb{R}$ we may use the squared error:

$$\mathcal{J}(\boldsymbol{x}_i, y_i; \theta_0, \boldsymbol{\theta}) = (y_i - f(\boldsymbol{x}_i))^2$$

The Lagrangian form of the above optimisation problem is:

$$\min_{\theta_0, \boldsymbol{\theta}} \ \frac{1}{n} \sum_{i=1}^{n} \mathcal{J}(\boldsymbol{x}_i, y_i; \theta_0, \boldsymbol{\theta}) + \lambda \|\boldsymbol{\theta}\|_1$$

In the above problem λ is a hyper-parameter that controls the amount of regularisation with larger values resulting in more sparse models. The above problem is not differentiable due to the constraint and is often solved with iterative approaches such as coordinate descent (Friedman et al., 2010). The resulting non-zero coefficients $\hat{\boldsymbol{\theta}}^{lasso}$ indicate the selected variables.

An alternative to enforcing sparsity is to add a constraint on the $L2$-norm of the weights, an approach known as Ridge Regression (Hoerl and Kennard, 1970). The optimal parameters $\hat{\theta}_0, \hat{\boldsymbol{\theta}}^{ridge}$ are chosen by solving the following optimisation problem:

$$\begin{aligned} &\min_{\theta_0, \boldsymbol{\theta}} \ \frac{1}{n} \sum_{i=1}^{n} \mathcal{J}(\boldsymbol{x}_i, y_i; \theta_0, \boldsymbol{\theta}) \\ &\text{s.t. } \|\boldsymbol{\theta}\|_2 \leq k \end{aligned}$$

Equivalently, we can solve the Lagrangian of the above optimisation problem for a given value of the hyper-parameter λ:

$$\min_{\theta_0, \boldsymbol{\theta}} \ \sum_{i=1}^{n} \frac{1}{n} \mathcal{J}(\boldsymbol{x}_i, y_i; \theta_0, \boldsymbol{\theta}) + \lambda \|\boldsymbol{\theta}\|_2$$

The $L2$-norm regulariser is traditionally used in Machine Learning models to avoid overfitting, while for linear regression models the absolute value of each coefficient can be interpreted as the importance of the corresponding variable. This allows us to rank the variables and select the top-k.

3.1.2 Recursive Partitioning Models

Decision Trees and Random Forests have long been used in the Machine Learning literature (Breiman et al., 1984; Breiman, 2001). In addition to their often good performance they are also easy to interpret and can be used to derive variable importance scores. In its simplest form, when fitting a tree, the root node partitions the initial data $\mathcal{D}$ in two non-overlapping subsets $\mathcal{D}_r, \mathcal{D}_l$ which correspond to the new nodes (a.k.a. children). This is performed by iterating over all variables and their possible values and estimating a pre-defined splitting criterion G. The variable and its value that optimise this criterion are selected and the data are split according to these. Then for each new node we repeat the procedure until certain termination criteria are met (e.g. a maximum depth has been reached).

There is a variety of methods for determining importance scores from trees

and forests. Perhaps the most common is to combine the training procedure with the variable selection task and rank the variables based on the values of the splitting criterion (Breiman et al., 1984; Hastie et al., 2009). For each variable the improvement in the splitting criterion is estimated each time this variable is used and its values are averaged across all splits. The splitting criterion used to derive the variable importance score could be the Gini index for classification or the residual sum of squares for regression. For Random Forests (Breiman, 2001) the out-of-bag predictions can also be used, in which case the variable importance can be derived from calculating the decrease in error before and after permutation of the variable. These along with other approaches for deriving variable importance scores are described in (Liaw et al., 2002; Breiman, 2001, 2002; Sandri and Zuccolotto, 2008).

3.2 Information Theoretic Variable Selection

The mutual information between two variables $\mathrm{I}(X;Y)$ (Shannon, 1948) captures the reduction in uncertainty for a variable X given that we observe the values of another variable Y. The uncertainty of a categorical random variable X is defined as its entropy $\mathrm{H}(X) = -\sum_{x\in\mathcal{X}} p(x)\log\, p(x)$ yielding the following definition of the mutual information:

$$\mathrm{I}(X;Y) = \mathrm{H}(X) - \mathrm{H}(X \mid Y) = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log\frac{p(x,y)}{p(x)p(y)}$$

where $\mathcal{X},\mathcal{Y}$ are the domains of values of the variables X and Y. It can be easily seen that the mutual information is zero if and only if the variables are statistically independent, i.e. $p(x,y) = p(x)p(y), \forall x,y$. Another core concept is the conditional mutual information $\mathrm{I}(X,Y \mid Z)$, defined as follows:

$$\mathrm{I}(X;Y \mid Z) = \mathrm{H}(X \mid Z) - \mathrm{H}(X \mid YZ) = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}}\sum_{z\in\mathcal{Z}} p(x,y,z)\log\frac{p(x,y \mid z)}{p(x \mid z)p(y \mid z)}$$

The conditional mutual information quantifies the dependence between two random variables, once the value of a third variable is known.

For categorical variables the mutual information is normally estimated from the data with a *maximum likelihood* estimator. In this case all distributions are estimated from the contingency table formed for the two variables. In scenarios

with small sample size and/or large cardinality of the variables many cells will be empty and the estimates will not be reliable. To this end, several estimators have been proposed some of which are better suited for this scenario (Paninski, 2003; Nemenman et al., 2002; Hausser and Strimmer, 2009). A review of different estimators can also be found in (Sechidis et al., 2019a). In Chapter 4 we will use a shrinkage estimator (Hausser and Strimmer, 2009) and we will describe it in more detail.

In variable selection, filter methods use a scoring criterion intended to measure how useful a set of variables is for predicting the outcome of interest (Guyon et al., 2008). Since the mutual information is a measure of dependence, it is an intuitive scoring criterion for developing a filter method. There is a large number of information theoretic methods for variable selection a summary of which can be found in (Vergara and Estévez, 2014) and (Brown et al., 2012). In particular, Brown et al. (2012) starting from a clearly defined objective, the conditional likelihood, derive the following information theoretic variable selection criterion.

$$\underset{\boldsymbol{X}_\theta \in \boldsymbol{X}}{\arg\min}\ \mathrm{I}(Y; \boldsymbol{X}_{\tilde{\theta}} \mid \boldsymbol{X}_\theta)$$

where $\boldsymbol{X}_{\tilde{\theta}} \subseteq \boldsymbol{X}$ are the unselected variables and $\boldsymbol{X}_\theta \subseteq \boldsymbol{X}$ are the selected. Alternatively using the chain rule of mutual information (Cover and Thomas, 2012) the variable selection problem can be phrased by maximising $\mathrm{I}(Y; \boldsymbol{X}_\theta)$. This states that given a joint variable $\boldsymbol{X}$, the optimal set of variables can be derived by maximising the mutual information shared with the outcome Y.

To solve the above optimisation problem Brown et al. (2012) propose two greedy procedures: forward selection and backward elimination. For example, the greedy forward selection procedure selects at each step k the variable $X_k \in \boldsymbol{X}_{\tilde{\theta}_\tau}$ that maximises the conditional mutual information (CMI): $J_{CMI}(X_k) = \mathrm{I}(X_k; Y \,|\, \boldsymbol{X}_{\theta_\tau})$, where $\boldsymbol{X}_{\theta_\tau}$ are the variables selected so far and $\boldsymbol{X}_{\tilde{\theta}_\tau}$ are those that remain unselected at the $\tau-$th step of the procedure. The backward elimination procedure starts from the full set of $\boldsymbol{X}$ and at each step removes the variable that minimises $\mathrm{I}(X_k; Y \mid \{\boldsymbol{X}_{\theta_\tau} \setminus X_k\})$ where $X_k \in \boldsymbol{X}_{\theta_\tau}$.

As the number of selected variables grows, i.e. the dimensionality of $\boldsymbol{X}_\theta$ grows, the estimates of (conditional) mutual information may be less reliable. To overcome this several low-order criteria have been proposed, each one relying on a different set of assumptions. For example, a simple ranking criterion is the mutual information between the target variable Y and a variable X_k, which takes

into account the *relevancy* of each variable (Lewis, 1992; Brown et al., 2012):

$$J_{MIM}(X_k) = \mathrm{I}(X_k; Y)$$

A popular information theoretic approach is the *Joint Mutual Information* (JMI) criterion (Yang and Moody, 2000). This criterion accounts for the three key aspects of a variable selection algorithm: the relevancy, redundancy and conditional redundancy and shows good empirical performance in terms of both predictive accuracy and stability (Kuncheva, 2007) of the selected variables (Brown et al., 2012). It is defined as follows:

$$J_{JMI}(X_k) = \sum_{X_j \in \boldsymbol{X}_\theta} \mathrm{I}(X_k X_j; Y)$$

Many other criteria have been proposed in the literature and several of those can be derived from $\mathrm{I}(X_k; Y \mid \boldsymbol{X}_\theta)$ under different assumptions (Brown et al., 2012).

3.3 Variable Categorisation in the Presence of Interventions

Variable selection in most learning environments, such as supervised and semi-supervised learning, deals with the problem of identifying variables predictive of the outcome of interest (Guyon et al., 2008; Brown et al., 2012; Sechidis and Brown, 2018). When dealing with data that include interventions, we may similarly be interested in identifying covariates that are strongly relevant with the outcome under a particular value of the treatment, i.e. the potential outcome. These covariates will be indicators of the likely outcome under some value of the treatment and in general their ranking may be different from the ranking of the covariates that are predictive of the outcome Y. In particular, since the potential outcomes are partially observed, the problem of identifying covariates predictive of the potential outcomes shares similarities with the literature of semi-supervised variable selection and requires assumptions about the missingness mechanism.

In causal effect estimation problems pre-treatment covariates can affect the outcome in several ways. In clinical trials distinguishing between different types of covariates is important and has been highlighted in numerous works (Dunn

et al., 2013; Ruberg and Shen, 2015; Lipkovich et al., 2017a). A prognostic covariate is associated with the likelihood of the event of interest. In healthcare this is sometimes described as a characteristic that is associated with the outcome in untreated patients (Lipkovich et al., 2017b; Italiano, 2011). In some cases a prognostic covariate is described as a covariate that indicates the likely outcome irrespective of the applied intervention (Ballman, 2015). On the other hand, predictive covariates are used to identify subjects who are more likely to experience a favourable or unfavourable effect from the applied treatment. In some cases predictive covariates might be defined as those that lead to enhanced treatment effect under the novel treatment (e.g. a new drug). For example, this could be the case in areas where the treatment effect is monotonic, that is the novel treatment is not expected to harm the subjects (Kallus, 2019). In our context, in order to demonstrate whether a covariate is predictive we will need to show that it exhibits an interaction with the treatment. For a (solely) prognostic covariate we will need to show that it is associated with the outcome irrespective of the treatment and it does not interact with the treatment. In practice, a covariate can be both predictive and prognostic.

We can further categorise the predictive covariates depending on whether they exhibit a quantitative or a qualitative interaction with the treatment (Lipkovich et al., 2017a). A predictive covariate is said to exhibit a quantitative interaction with the treatment if it modifies the overall treatment effect to a certain direction. For example, in a failed clinical trial, a quantitative interaction could be an indicator of enhanced treatment effect for subsets of the data (Lipkovich et al., 2017a). A predictive covariate is said to exhibit a qualitative interaction with the treatment if it defines both subgroups of enhanced and deteriorated treatment effect. In healthcare the distinction between the two is also referred to as identifying the best patient for the treatment (quantitative) or the best treatment for a patient (qualitative) (Lipkovich et al., 2017a).

A detailed discussion regarding the distinction between the different types of covariates and their use in a clinical trial setting can be found in (Lipkovich et al., 2019, 2017a). Here we will focus on four cases that are worth exploring in more detail:

- Case 1: A covariate is prognostic but not predictive
- Case 2: A covariate is predictive but not prognostic

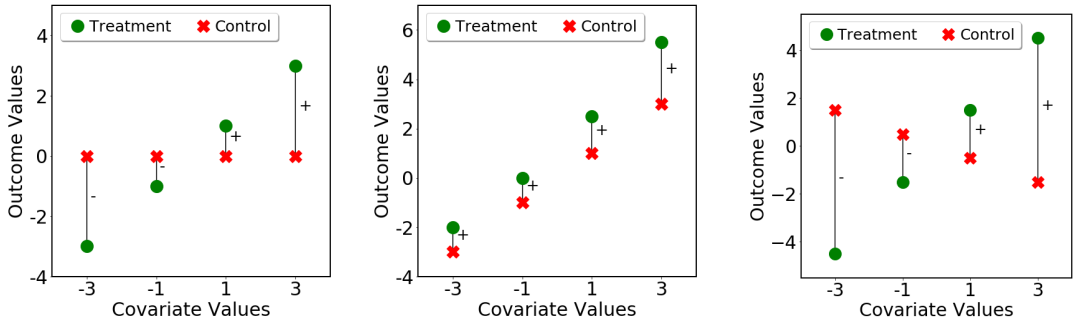


Figure 3.1: (a) X is predictive but not prognostic (b) X is predictive and exhibits a quantitative interaction with the treatment (d) X is predictive and exhibits a qualitative interaction with the treatment.

- Case 3: A covariate is both predictive and prognostic and exhibits a quantitative interaction with the treatment
- Case 4: A covariate is both predictive and prognostic and exhibits a qualitative interaction with the treatment

The first case is simulated for a continuous covariate and outcome in figure 1.2(b). The other cases are shown in figure 3.1 where each point represents an observation from a simulated randomised experiment assuming we know both potential outcomes. In particular, in figure 1.2(b) the potential outcomes are generated by $Y = X + T$, in figure 3.1(a) by $Y = TX$, in figure 3.1(b) by $Y = X + T(1 + 0.5X_+)$, where $X_+ = 0$ if $X < 0$ and $X_+ = X$ if $X > 0$ and in figure 3.1(c) as $Y = 0.5 * X + (2T - 1)X$.

To give another example, let $m_1(\boldsymbol{X}), m_0(\boldsymbol{X})$ be the functional forms of the potential outcome and suppose both are linear. Following the definition of a prognostic covariate that will be adopted in this thesis, solely prognostic covariates will be included in the functional form of both potential outcomes, $m_1(\boldsymbol{X})$ and $m_0(\boldsymbol{X})$, and in the same way –i.e. they describe the main effect. On the other hand, solely predictive covariates are those that are included in the functional form of one potential outcome but not the other. For example, if we intend to compare a new treatment, $T = 1$, with a baseline treatment, $T = 0$, then a solely predictive covariate for our task will be predictive of $Y(1)$ and therefore will be included in $m_1(\boldsymbol{X})$ but not in $m_0(\boldsymbol{X})$. In practice, a variable may be characterised by degrees of "prognosticness" and "predictiveness".

In observational studies (with no hidden confounders) we are also concerned

with the identification of confounders - a set of covariates that we need to condition, to satisfy unconfoundedness. Confounder selection is itself a variable selection problem (De Luna et al., 2011; VanderWeele and Shpitser, 2011; Pearl, 1995). Any of the covariates described previously can act as a confounder if without conditioning on it leaves an open path between the treatment T and at least one of the potential outcomes. A common criterion for identifying confounders is Pearl's Back-door criterion (Pearl, 1995). Given a causal graph over the variables $\{\boldsymbol{X}, T, Y\}$ a set of covariates $\boldsymbol{X}_c$ satisfies the back-door criterion if firstly no covariate in $\boldsymbol{X}_c$ is a descendant of T and secondly, $\boldsymbol{X}_c$ blocks all paths from T to Y that contain an arrow into T (Pearl, 1995). VanderWeele and Shpitser (2011) notice that assessing this criterion can be a difficult task as it requires knowledge of the underlying causal structure. Instead they propose a new criterion suited for practitioners that suggests including a covariate in the set of confounders if it is either a cause of the treatment or the outcome or both. The theoretical justification arrives from the observation that if there are some observed pre-treatment covariates that satisfy the back-door criterion then the set that satisfies the aforementioned causal relationships will also satisfy it (VanderWeele and Shpitser, 2011). The first criterion relies on knowledge of the causal graph, which in many cases can be infeasible. The second criterion requires at least knowledge of the causes of the outcome or the treatment. VanderWeele and Shpitser (2011) describe that in practice subject matter experts may have such knowledge which makes this easier to apply in a practical setting. From a methodological perspective if we are willing to assume unconfoundedness the main question is whether we can define a data-driven approach for confounder selection that does not make further assumptions about the causal graph. To this end some methodologies for identifying a minimal set of confounders with respect to the potential outcomes $Y(1), Y(0)$ have been proposed (De Luna et al., 2011; Häggström, 2018). They use data-driven methods to identify possible sets of covariates that may act as confounders by combining the Markov Blanket of the treatment T and the outcomes under each treatment in different ways.

We can summarize the variable selection problems we often need to tackle when dealing with data that include interventions:

1. Covariates $\boldsymbol{X}_t$ that are predictive of the potential outcome $Y(t)$ under treatment $T = t$. As opposed to covariates that are predictive of the outcome Y, they describe the importance of the covariates if we were to apply treatment

$T = t$.

2. Predictive covariates $\boldsymbol{X}_{pred}$ that exhibit an interaction with the treatment.

3. Prognostic covariates $\boldsymbol{X}_{prog}$ that provide information for predicting the outcome irrespective of the applied treatment.

4. Confounders $\boldsymbol{X}_c(t)$ that satisfy $Y(t) \perp\!\!\!\perp T \mid \boldsymbol{X}_c(t)$. Depending on the context, we may define the set of confounders $\boldsymbol{X}_c$ as those that satisfy $\big(Y(1), Y(0)\big) \perp\!\!\!\perp T \mid \boldsymbol{X}_c$.

The second category of variables is going to be the focus of the next chapter, where we will describe how information theoretic variable selection approaches can be adapted to tackle this problem. Before that let us describe some existing approaches that can be used for identifying predictive covariates.

3.4 Three Frameworks for Subgroup Identification

Subgroup identification is the task of identifying subsets of the data with desirable characteristics. For example, a common scenario in clinical trials is when a sponsor is interested in identifying subgroups that benefit from the treatment in an otherwise failed trial (a trial that did not meet its primary objective) (Lipkovich et al., 2017a). This is commonly performed in phase III and IV trials (Dmitrienko et al., 2016) where the sample size is large enough to allow us to perform such analysis. We will focus on the task of *exploratory* subgroup identification, that is using methods to generate hypotheses which can be tested in latter stages. This is in contrast to *confirmatory* subgroup analysis where pre-specified subgroups are analysed using clinical trial data (Lipkovich et al., 2017a). We note that the task of subgroup identification is a rather general task and can be of interest in scenarios other than clinical trials such as public policies (Loh et al., 2019; Alemayehu et al., 2018) and consumer analysis (Wang and Rudin, 2017). In this section we describe three approaches for subgroup identification and discuss how these can be used for identifying predictive covariates, which will be the focus of the next chapter.

Subgroup identification is closely related to the problem of conditional average treatment effect estimation. In order to connect the described algorithms with

the modelling approaches described in the previous chapter, we use the taxonomy of Lipkovich et al. (2017a). They categorise subgroup identification algorithms as global outcome modelling, global treatment effect modelling and local modelling. In our context in order to make matters simpler we distinguish algorithms that model the potential outcomes, the treatment effect or identify subgroups directly.

- **Counterfactual Modelling**: These approaches estimate the potential outcomes as part of their process. Based on the discussion of the previous chapter, this can be achieved using a Single-Model or a Two-Model method.

- **Treatment Effect Modelling**: These approaches estimate the treatment effect directly without modelling the main effect. An approach that uses the outcome transformation method as part of its process will fall in this category.

- **Subgroup Modelling**: These approaches directly search for subgroups with desirable characteristics.

The first two approaches solve the problem of treatment effect estimation before identifying subgroups of interest and they differ on the way they achieve this. The last approach in principal may not be used to estimate treatment effects but will provide subgroups directly. We shall now describe in more detail some representative examples of each category and describe how we can use them to identify predictive covariates. We will focus our discussion on methods that will be used in Chapter 4.

3.4.1 Counterfactual Modelling

In this category we have approaches that estimate the potential outcomes before searching for subgroups. Commonly used methods are (penalised) linear models and tree-based models, appropriately modified to include treatment/covariate interactions. For example Imai et al. (2013) propose a modified SVM model that imposes two $L1$ regularisation terms, one for the covariates that describe the main effect and one for the covariates that describe the treatment effect, i.e. the interactions with the treatment. A popular method of this category is Virtual Twins (Foster et al., 2011), which proceeds in two steps. In the first

step, it estimates $\mathbb{E}[Y_i \mid T_i = 1, \boldsymbol{x}_i]$ and $\mathbb{E}[Y_i \mid T_i = 0, \boldsymbol{x}_i]$ using Random Forests (Breiman, 2001). In Foster et al. (2011) the authors train a Random Forest on the variable set $\{\boldsymbol{X}, T, \boldsymbol{X}\mathbb{I}(T = 1), \boldsymbol{X}\mathbb{I}(T = 0)\}$. The factual outcome is estimated using the out-of-bag estimates while the counterfactual is predicted by switching the treatment indicator (i.e. replace $T = 1(0)$ with $T = 0(1)$). Then they define a new variable $z(\boldsymbol{x}_i)$, which is the estimated treatment effect defined as the causal risk difference. Alternatively for binary outcomes one may measure the effect as difference in logits, log-odds or some other measure of comparison of the potential outcomes. In the second step, they use either a regression tree on $z(\cdot)$ or a classification tree on the variable $Z^* = \mathbb{1}(z(\boldsymbol{x}_i) > c)$ where c is some constant, most commonly the average treatment effect in the sample. It is the second step of the approach that identifies subgroups by partitioning the space based on the values of the estimated treatment effect.

In general, Virtual Twins defines a rather general method to the problem of subgroup identification and the first step can be replaced by other modelling approaches for estimating the potential outcomes, such as using two models, one for each treatment group. A closely related problem is described in (Makar et al., 2019), where their motivation is to use the final decision tree as a surrogate of a more complex model for conditional average treatment effect estimation. In order to derive predictive covariates, we can replace the second step with a random forest, a common method to derive importance scores (Hastie et al., 2009). Using the estimated $z(\cdot)$ as the target we rank the covariates as described in 3.1.2 and use the residual sum of squares as the scoring criterion. This provides a ranking of the covariates based on how relevant they are for predicting the estimated treatment effect and hence can be interpreted as a predictive ranking.

3.4.2 Treatment Effect Modelling

Virtual Twins estimates the potential outcomes and therefore it necessarily models the main effect as well as the interactions with treatment. For subgroup identification purposes, we are interested on modelling only the interactions. This motivated the adoption of methods that use the the outcome transformation approach, which we described in the previous chapter. A representative example of this category is the Modified Covariates Regression (MCR) (Tian et al., 2014). For continuous outcomes the treatment effect can be estimated by performing regression using the modified outcome variable $Y_i^* = 2T_i^* Y_i$, where $T_i^* = 2T_i - 1$.

Since $\text{CATE}(\boldsymbol{x}_i) = \mathbb{E}[Y_i^* \mid \boldsymbol{x}_i]$ this approach has a clear causal interpretation, while it also avoids modelling the main effect. The authors show that under the squared loss for continuous outcomes this is equivalent to regressing on Y and multiplying the covariates with $T^*/2$. They also extend their approach to survival and binary outcomes following a similar weighting scheme. This approach will produce an estimation of the treatment effect, which can be used to stratify the population and identify subgroups with desirable characteristics (Lipkovich et al., 2017a).

We can notice that this approach is equivalent to the one discussed in (Athey and Imbens, 2016, 2015) in 1:1 randomised experiments. In this case we have $e(\boldsymbol{x}_i) = 1/2$. Substituting in eq. (2.5) we have $Y_i^* = Y_i \cdot \frac{T_i - 1/2}{1/2 \cdot 1/2} = 2(2T_i - 1)Y_i = 2T_i^* Y_i$. The authors use $L1$ regularisation (Tibshirani, 1996) to identify a minimal set of predictive covariates, while alternatively one may use $L2$ regularisation (Hoerl and Kennard, 1970) to derive a ranking, as we discussed in 3.1.1.

Other examples of this framework are Interaction Trees (IT) (Su et al., 2009) and CF (Athey and Imbens, 2016; Athey et al., 2019) which perform recursive partitioning of the space. CF was discussed in detail in the previous chapter. IT follows the procedure of Classification and Regression Trees (CART) (Breiman et al., 1984) adapted for causal inference tasks. This approach will be discussed in more detail in Chapter 5 where we will explore extensions in observational studies with no hidden confounders. In the next chapter we will use MCR as an example of a linear model that adopts the outcome transformation approach for identifying predictive covariates.

3.4.3 Subgroup Modelling

The approaches that fall in this framework search for subgroups with desirable characteristics directly. A representative example of this category is Subgroup Identification using Differential Effect Search (SIDES) (Lipkovich et al., 2011) and its extensions SIDEScreen (Lipkovich and Dmitrienko, 2014b) and Stochastic SIDEScreen (Lipkovich et al., 2017b). SIDES is a recursive partitioning approach that splits the data into subgroups using the following splitting criterion (Lipkovich et al., 2017b; Lipkovich and Dmitrienko, 2014b):

$$G = 2\Big[1 - \Phi\Big(\frac{|Z_L - Z_R|}{\sqrt{2}}\Big)\Big]$$

where Z_L, Z_R are the normalised treatment effects of the children defined by subsets of the data split using some covariate X. Here, $\Phi(\cdot)$ is the cumulative distribution of the standard normal distribution. Larger values of the absolute difference or lower values of the estimated p-value indicate greater discrimination between the subgroups. The covariate that has the minimal p-value is added in the list of promising subgroups. The estimated p-values may be further adjusted to account for multiple comparisons (Lipkovich et al., 2011).

Various criteria are applied in order to control the number of promising subgroups. Firstly, the treatment effect p-value of a child is compared to the corresponding value of the parent and the subgroup is retained if their ratio is lower than a pre-specified threshold. The partitioning continues until a minimum sample size has been reached or the subgroup is defined by a maximum number of covariates that has been set by the user. Secondly, a child subgroup is added to the promising subgroups if the estimated p-value is lower than a pre-specified threshold. The p-values of the promising subgroups are re-estimated using a resampling-based adjustment in order to control the Type I error rate (falsely identifying a subgroup) (Lipkovich et al., 2011). In particular, the covariates of each example are permuted, so that in the permuted sample the treatment effect remains unchanged but removing any interactions between treatment and covariates. This is repeated multiple times for a set of possible thresholds calculating the proportion of times at least one promising subgroup is retained. From the threshold values for which the calculated proportion was not greater than a specified value of the Type I error rate, the largest is retained as the final one (Lipkovich et al., 2011). In contrast to methods that extend CART (Breiman et al., 1984), such as IT (Su et al., 2008), SIDES may give overlapping subgroups, since at each iteration it retains a number of promising splits rather than keeping only the most promising one.

The above splitting criterion cannot discriminate subgroups that lead to positive and negative effects. To address this, they suggest the directional splitting criterion (Lipkovich and Dmitrienko, 2014a; Lipkovich et al., 2017b) which may ignore those subgroups that lead to negative treatment effects, hence evaluating the splits based on whether they lead to enhanced effects. There have also been proposed modifications suited for handling high-dimensional data, in which case false discovery rate, i.e. forming subgroups using non-predictive covariates, is a

key concern. In particular, they introduce an additional screening step that identifies promising covariates and then proceed by identifying subgroups using only those (Lipkovich and Dmitrienko, 2014b; Lipkovich et al., 2017b). The promising covariates are selected based on the value of the splitting criterion for each subgroup that they appear in. This criterion will be used in the next chapter in order to rank the covariates based on their predictive strength. In particular, the importance of a covariate X is given by $\frac{1}{L}\sum_{i=1}^{L} v_i$, where L is the number of subgroups, $v_i = -\log G_i$ if the i-th subgroup contains X and zero otherwise (Lipkovich et al., 2017b). This procedure can be repeated for multiple bootstrap samples in order to derive a distribution of the variable importance scores as discussed in (Lipkovich et al., 2017b).

3.5 Chapter Summary

The goal of this chapter is two-fold; Firstly to introduce some key concepts and methods regarding variable selection and subgroup identification and secondly to provide background information on the problem of variable selection in the presence of interventions. In particular, in Section 3.1 we discussed variable selection using LASSO, Ridge Regression and Random Forests. In Section 3.2 we described information theoretic criteria, that belong to the family of filter variable selection - they are used to identify useful variables without building some model for inference. We then discussed the variable selection problem in the presence of interventions.

Finally, we concluded this chapter with a discussion of the problem of subgroup identification and described how such algorithms can be used to identify predictive covariates. In particular, we described Virtual Twins (VT) (Foster et al., 2011), Modified Covariates Regression (MCR) (Tian et al., 2014) and Subgroup Identification using Differential Effect Search (SIDES) (Lipkovich et al., 2011) respectively. These are representative examples of three subgroup identification frameworks. VT estimates the potential outcomes, MCR the treatment effect and SIDES performs subgroup identification directly. In the next chapter we take a closer look at the problem of predictive covariate selection and discuss a new framework, the information theoretic.

Chapter 4

Identifying Predictive Covariates: An Information Theoretic Approach

In this chapter we study a method for identifying predictive covariates and discuss its properties. This method is an information theoretic approach, which comes with the advantage that it does not require inference of treatment effects or counterfactuals. We present this method from a likelihood maximisation perspective (Section 4.1), analyse its properties (Sections 4.2, 4.3) and perform an empirical comparison with the approaches discussed in the previous chapter in synthetic scenarios (Section 4.4) and an evaluation in real data (Section 4.5). The theoretical analysis of the proposed method shows that it can be influenced by the treatment assignment mechanism. We then present extensions that can ameliorate this issue (Section 4.6).

Author contribution statement: This chapter is based in part on (Sechidis et al., 2018), where the author contributed to the theoretical analysis, experimental evaluation and writing of the manuscript. In addition, this chapter presents work not present in (Sechidis et al., 2018). In particular, we define the covariate selection criterion from a likelihood maximisation perspective (Section 4.1). We perform a theoretical analysis that shows newly identified properties of the method (Section 4.3). We add new experiments with both continuous and discrete covariates while varying both the main effect and the effect of the interaction (Section 4.4) as well as a new dataset (section 4.5). Finally we introduce two extensions that overcome some of the identified limitations (section 4.6).

4.1 An Information Theoretic Criterion for Identifying Predictive Covariates

In this section we build links between data-driven predictive covariate selection and information theoretic feature selection (Brown et al., 2012). We consider a dataset $\mathcal{D} = \{\boldsymbol{x}_i, t_i, y_i\}_{i=1}^n$, where $\boldsymbol{x}_i, t_i, y_i$ are n realisations of the following random variables: $\boldsymbol{X} \in \mathbb{R}^d$ are the covariates, $T \in \{0, 1\}$ is the treatment variable, $Y \in \{0, 1\}$ is the outcome. In a randomised study designed to evaluate the efficacy of some intervention, the covariates $\boldsymbol{X}$ can be distinguished by how much they influence Y via their main or prognostic effect and their interaction with the treatment or predictive effect. Focusing on the latter we can rank the covariates in terms of how predictive they are of the outcome Y when used in conjunction with the treatment T as opposed to without using the treatment.

Definition 1 (Predictive covariate identification objective)**.** *The objective is to identify a set of covariates $\boldsymbol{X}^* \subseteq \boldsymbol{X}$ such that they maximise,*

$$\boldsymbol{X}^* = \underset{\boldsymbol{X}_\theta \in \boldsymbol{X}}{\arg\max}\ \mathbb{E}\big[log\ p(Y \mid t, \boldsymbol{x}_\theta)\big] - \mathbb{E}\big[log\ p(Y \mid \boldsymbol{x}_\theta)\big]$$

Here the first term captures the conditional likelihood of the outcome given the interaction between treatment and covariates and the second term is the likelihood of the outcome given the covariates, irrespective of the treatment. By adding and subtracting the term $\mathbb{E}\big[\log p(Y)\big]$ we get:

$$\begin{aligned}&\mathbb{E}\big[\log p(Y \mid t, \boldsymbol{x}_\theta)\big] - \mathbb{E}\big[\log p(Y)\big] - \mathbb{E}\big[\log p(Y \mid \boldsymbol{x}_\theta)\big] + \mathbb{E}\big[\log p(Y)\big] = \\ &= \mathrm{I}(Y; \boldsymbol{X}_\theta T) - \mathrm{I}(Y; \boldsymbol{X}_\theta) = \mathrm{I}(T; Y \mid \boldsymbol{X}_\theta)\end{aligned}$$

Therefore the objective can be expressed in information theoretic terms as:

$$\boldsymbol{X}^* = \underset{\boldsymbol{X}_\theta \in \boldsymbol{X}}{\arg\max}\ \mathrm{I}(T; Y \mid \boldsymbol{X}_\theta) \tag{4.1}$$

Brown et al. (2012), in the context of feature selection, presented two heuristics for optimising problems as the above, which consider sequentially features one-by-one for adding or removal; the forward selection and the backward elimination respectively. The forward selection step starts from an empty set and sequentially

adds features, while the backward elimination step starts from the full feature set $\boldsymbol{X}$ and sequentially removes one feature at a time. To present these procedures in our setting let us denote as $\boldsymbol{X}_{\theta_\tau}$ the covariates selected up to the τ-th step and $\boldsymbol{X}_{\tilde{\theta}_\tau} = \boldsymbol{X} \setminus \boldsymbol{X}_{\theta_\tau}$ the remaining covariates.

Definition 2 (Predictive covariate forward selection step)**.** *The forward selection step adds the covariate X_k^* which maximises the conditional mutual information between T and Y given the joint variable of the currently selected set $\boldsymbol{X}_{\theta_\tau}$ and X_k^*. The operations performed are:*

$$\begin{aligned} X_k^* &= \underset{X_k \in \boldsymbol{X}_{\tilde{\theta}_\tau}}{\arg\max}\ I(T; Y \mid \boldsymbol{X}_{\theta_\tau} X_k) \\ \boldsymbol{X}_{\theta_{\tau+1}} &\leftarrow \boldsymbol{X}_{\theta_\tau} \cup X_k^* \\ \boldsymbol{X}_{\tilde{\theta}_{\tau+1}} &\leftarrow \boldsymbol{X}_{\tilde{\theta}_\tau} \setminus X_k^* \end{aligned} \tag{4.2}$$

Using the results of Brown et al. (2012), the following corollary holds.

Corollary 1. *The predictive forward selection heuristic adds the covariate that causes the largest increase in the predictive objective.*

For the backward elimination we have the following definition and corollary which follows from Brown et al. (2012).

Definition 3 (Predictive covariate backward elimination step)**.** *The backward elimination step removes the covariate X_k^* which minimises the conditional mutual information between T and Y given the joint variable of the currently selected set $\boldsymbol{X}_{\theta_\tau}$ without X_k^*. The operations performed are:*

$$\begin{aligned} X_k^* &= \underset{X_k \in \boldsymbol{X}_{\theta_\tau}}{\arg\min}\ I(T; Y \mid \{\boldsymbol{X}_{\theta_\tau} \setminus X_k\}) \\ \boldsymbol{X}_{\theta_{\tau+1}} &\leftarrow \boldsymbol{X}_{\theta_\tau} \setminus X_k^* \\ \boldsymbol{X}_{\tilde{\theta}_{\tau+1}} &\leftarrow \boldsymbol{X}_{\tilde{\theta}_\tau} \cup X_k^* \end{aligned} \tag{4.3}$$

Corollary 2. *The predictive backward elimination heuristic removes the covariate that causes the minimum possible decrease in the predictive objective.*

For simplicity from now on we will focus on the forward selection procedure. Given the set of unselected covariates $\boldsymbol{X}_{\tilde{\theta}}$ we select the covariate not ranked so

far $X_k^* \in \boldsymbol{X}_{\tilde{\theta}}$ that maximises the following score:

$$X_k^* = \underset{X_k \in \boldsymbol{X}_{\tilde{\theta}}}{\arg\max}\ J_{\text{Pred-CMI}}(X_k) = \underset{X_k \in \boldsymbol{X}_{\tilde{\theta}}}{\arg\max}\ \mathrm{I}(T; Y \mid \boldsymbol{X}_\theta X_k) \tag{4.4}$$

We will refer to this criterion as Pred-CMI, since it quantifies the predictive "strength" of each covariate X_k using the conditional mutual information between the treatment and the outcome given the joint variable between X_k and the currently selected covariates $\boldsymbol{X}_\theta$. For simplicity we will use the term $J_{\text{Pred-CMI}}$ to denote both $\mathrm{I}(T; Y \mid \boldsymbol{X}_\theta)$ and $\mathrm{I}(T; Y \mid \boldsymbol{X})$ depending on the context. In order to derive predictive rankings using Pred-CMI we need to tackle an important challenge: as the number of selected covariates grows, so does the dimension of $\boldsymbol{X}_\theta$, which makes our estimations less reliable. To overcome this problem we can use low-dimensional criteria that rely on simplifying assumptions regarding the underlying distribution of the data (Brown et al., 2012).

4.2 Low-dimensional Approximations

The simplest approximation is to measure the conditional mutual information of T and Y given each covariate *independently*. This criterion can be seen as a univariate information theoretic way to derive predictive rankings. This will be referred to as INFO. The score that INFO uses to rank the covariates is:

$$J_{\text{INFO}}(X_k) = \mathrm{I}(T; Y \mid X_k)$$

While this is a low-dimensional criterion – we simply need to estimate a joint distribution of three variables – and therefore relaxes the complexity of $J_{\text{Pred-CMI}}$, it fails to capture the dependencies between the covariates. We can illustrate these dependencies better by using the information theoretic identity $\mathrm{I}(A; B \mid CD) = \mathrm{I}(A; B \mid D) - \mathrm{I}(C; B \mid D) + \mathrm{I}(C; B \mid AD)$ to re-write the Pred-CMI criterion as

$$J_{\text{Pred-CMI}}(X_k) = \mathrm{I}(T; Y \mid X_k) - \mathrm{I}(\boldsymbol{X}_\theta; Y \mid X_k) + \mathrm{I}(\boldsymbol{X}_\theta; Y \mid TX_k)$$

Using this expression, INFO is an approximation of Pred-CMI that captures only the first term, which measures the predictive "strength" of covariate X_k but it fails to account for terms that capture the redundancy between the covariates. The

second term captures the three-way interaction between the existing covariate set $\boldsymbol{X}_\theta$, the outcome Y and the covariate X_k. Finally, the third term captures the four-way interaction between $\boldsymbol{X}_\theta$, Y, T and the covariate X_k.

We approximate these high dimensional functions of $\boldsymbol{X}_\theta$ by a sum of second-order interactions as follows.

$$J_{\text{INFO+}}(X_k) = \sum_{X_j \in \boldsymbol{X}_\theta} \text{I}(T; Y \mid X_j X_k)$$

We refer to this criterion as INFO+ and we can show that it can be derived from $J_{\text{Pred-CMI}}$ under certain simplifying assumptions. As in Brown et al. (2012) if we assume $p(\boldsymbol{x}_\theta \mid y, x_k) = \prod_{j \in \boldsymbol{x}_\theta} p(x_j \mid y, x_k)$ and $p(\boldsymbol{x}_\theta \mid x_k) = \prod_{j \in \boldsymbol{x}_\theta} p(x_j \mid x_k)$ and additionally we assume $p(\boldsymbol{x}_\theta \mid y, t, x_k) = \prod_{j \in \boldsymbol{x}_\theta} p(x_j \mid y, t, x_k)$ and $p(\boldsymbol{x}_\theta \mid t, x_k) = \prod_{j \in \boldsymbol{x}_\theta} p(x_j \mid t, x_k)$ then the criterion becomes equal to:

$$J_{\text{Pred-CMI}}(X_k) = \text{I}(T; Y \mid X_k) - \sum_{X_j \in \boldsymbol{X}_\theta} \text{I}(X_j; Y \mid X_k) + \sum_{X_j \in \boldsymbol{X}_\theta} \text{I}(X_j; Y \mid T X_k)$$

Brown et al. (2012) consider a parameterisation of a similar objective in the context of supervised learning and show that many existing feature selection criteria can be derived from this. Here, INFO+ follows by applying the weight $\frac{1}{|\boldsymbol{X}_\theta|}$ in the last two terms. Hence it shares similarities in its form with the JMI criterion in the context of supervised feature selection (Brown et al., 2012).

$$J_{\text{INFO+}}(X_k) \propto \text{I}(T; Y \mid X_k) - \frac{1}{|\boldsymbol{X}_\theta|} \sum_{X_j \in \boldsymbol{X}_\theta} \big[\text{I}(X_j; Y \mid X_k) - \text{I}(X_j; Y \mid T X_k)\big]$$

Here $|\boldsymbol{X}_\theta|$ is the number of covariates already selected. In theory this could be extended to arbitrary higher order interactions, but data constraints will always limit this. In scenarios with high dimensionality and/or small sample size the estimator of mutual information will be particularly important.

4.3 Estimation and Properties

Algorithm 1 describes the forward selection procedure for deriving predictive rankings using INFO+, where now the information theoretic terms $\text{I}(\cdot)$ are replaced with their estimates $\hat{\text{I}}(\cdot)$. We note here that since we estimate information theoretic terms for all pairs of covariates, these values can be stored every time we

estimate them and use them in subsequent iterations resulting in a much faster implementation than the one described here. The number of the selected covariates k can be selected either by the user or by adopting some stopping criterion. The former is common amongst information theoretic approaches (Vergara and Estévez, 2014). Normally the identified predictive covariates will be assessed in terms of whether they can define interesting subgroups. Since these subgroups will need to be interpretable by domain experts we could focus on only a few covariates. Regarding the stopping procedure some approaches have been introduced in the literature of supervised learning such as monitoring the changes of the objective value or performing permutation tests to assess the significance of adding or removing a covariate (François et al., 2007; Gocht et al., 2018; Beraha et al., 2019; Yu and Príncipe, 2019). Here we will not consider such stopping criteria, however including them in the algorithm could be an interesting extension.

In order to estimate the information theoretic terms for categorical variables we can use any off-the-shelf estimator suggested in the literature (Hausser and Strimmer, 2009; Sechidis et al., 2019a; Nemenman et al., 2002). Since in randomised studies and particularly in clinical trials we often encounter small-samples, we use a shrinkage estimator suitable for such scenarios (Hausser and Strimmer, 2009). The conditional mutual information is estimated as:

$$\hat{\mathrm{I}}(T;Y \mid \boldsymbol{X}_\theta) = \sum_{t\in\mathcal{T},y\in\mathcal{Y},\boldsymbol{x}_\theta\in\mathcal{X}_\theta} \hat{p}^{shrink}(t,y,\boldsymbol{x}_\theta)\log\frac{\hat{p}^{shrink}(t,y,\boldsymbol{x}_\theta)\hat{p}^{shrink}(\boldsymbol{x}_\theta)}{\hat{p}^{shrink}(t,\boldsymbol{x}_\theta)\hat{p}^{shrink}(y,\boldsymbol{x}_\theta)}$$

where $\hat{p}^{shrink}(t,y,\boldsymbol{x}_\theta)$ is the convex combination of a low-variance/high-bias estimator and a high-variance/low-bias one. Hausser and Strimmer (2009) adopt the uniform probabilities $p^{uni}(t,y,\boldsymbol{x}_\theta) = \frac{1}{|\mathcal{T}||\mathcal{Y}||\mathcal{X}_\theta|}$ as the low-variance/high-bias estimator, where $|\cdot|$ denotes the cardinality of the domain. As a high-variance/low-bias estimator, they choose the maximum likelihood estimates $\hat{p}^{ML}(t,y,\boldsymbol{x}_\theta) = \frac{n_{t,y,\boldsymbol{x}_\theta}}{n}$, where $n_{t,y,\boldsymbol{x}_\theta}$ is the number of observations with $T=t, Y=y$ and $\boldsymbol{X}_\theta = \boldsymbol{x}_\theta$. The optimal (in terms of mean squared error) parameter that controls the convex combination can be derived in a closed form expression (Hausser and Strimmer, 2009):

$$\hat{p}^{shrink}(t,y,\boldsymbol{x}_\theta) = \lambda p^{uni}(t,y,\boldsymbol{x}_\theta) + (1-\lambda)\hat{p}^{ML}(t,y,\boldsymbol{x}_\theta)$$

$$\hat{\lambda}^* = \frac{1 - \sum_{t\in\mathcal{T},y\in\mathcal{Y},\boldsymbol{x}_\theta\in\mathcal{X}_\theta}\left(\hat{p}^{ML}(t,y,\boldsymbol{x}_\theta)\right)^2}{(n-1)\sum_{t\in\mathcal{T},y\in\mathcal{Y},\boldsymbol{x}_\theta\in\mathcal{X}_\theta}\left(p^{uni}(t,y,\boldsymbol{x}_\theta) - \hat{p}^{ML}(t,y,\boldsymbol{x}_\theta)\right)^2}$$

Algorithm 1 Algorithm for identifying predictive covariates using INFO+.

Input dataset $\{\boldsymbol{x}_i, t_i, y_i\}_{i=1}^n$, Size of the returned ranking k
Output $\boldsymbol{X}_\theta$: the *INFO+* ranking of the top-k variables

```
Initialisation
X_θ ← ∅
X_θ̃ ← X
Ranking
for l := 1 to k do
    for X_i ∈ X_θ̃ do
        Ĵ(X_i) = 0
        for X_j ∈ X_θ do
            Ĵ(X_i) += Î(T; Y | X_j X_i)
        end for
    end for
    X_θ(l) ← arg max_{X_i ∈ X_θ̃} Ĵ(X_i)
    X_θ̃ ← X_θ̃ \ X_θ(l)
end for
```

For continuous variables we can still use the above estimator after discretisation, using for example methods for histogram generation. We will discuss this in the experimental section and also in Chapter 7.

Ranking predictive covariates with INFO+ has several advantages over existing methods but also has some limitations which we will discuss in this section. For subgroup identification purposes and particularly in high-dimensional settings we might be interested in only a few covariates (Lipkovich and Dmitrienko, 2014b). To this end, in contrast to the methods described in the previous chapter (i.e. SIDES, VT, MCR) which rank all covariates, the forward step-wise procedure adopted by INFO+ can return only the top-k, reducing the computational burden considerably. Compared to the frameworks described in the previous section it does not require imputation of the missing counterfactuals (such as VT) or a correctly specified model (such as MCR). Additionally due to its computational efficiency it can be adopted as a screening criterion before the application of any of the aforementioned frameworks (such as subgroup identification with SIDES).

An interesting scenario is when there is no interaction with the treatment and no overall effect, e.g. a failed study that does not exhibit treatment effect heterogeneity. In this case the following holds.

Lemma 1. *In marginally randomised experiments and in the absence of treatment effect $J_{Pred\text{-}CMI}$ becomes independent of the covariates.*

Proof. Provided in Appendix A.1

Therefore we expect that in the absence of an interaction with the treatment (and with no overall effect) $J_{\text{Pred-CMI}}$ should perform similarly to random selection. We validate this empirically in the next section for a low-dimensional criterion where interestingly we notice that this is not the case for VT.

Despite its advantages, INFO+ has also certain limitations. As we described in the previous chapter an interaction with the treatment might be quantitative or qualitative, showing an increased effect under the experimental treatment, a negative effect or both. This criterion cannot distinguish between the different types of interaction. This however can be performed by certain subgroup identification approaches which focus either on the identification of interactions that lead to enhanced effects (Lipkovich and Dmitrienko, 2014b) or qualitative interactions (Dusseldorp and Van Mechelen, 2014). Nevertheless, as we described INFO+ could be a fast filtering criterion before applying the more computationally demanding subgroup identification methods that are suited for certain types of interaction.

Using the chain rule we can notice that the predictive covariate selection objective can written as:

$$\mathrm{I}(T; Y \mid \boldsymbol{X}_\theta) = \mathrm{I}(T; Y) - \mathrm{I}(T; \boldsymbol{X}_\theta) + \mathrm{I}(T; \boldsymbol{X}_\theta \mid Y) \tag{4.5}$$

The first term is independent of the covariates and in the case of marginally randomised studies, where $T \perp\!\!\!\perp \boldsymbol{X}_\theta$ the criterion becomes equivalent to maximising the last term. In all other cases, the variables that are dependent with the treatment variable T will also affect the score and therefore the final ranking. Hence $J_{\text{Pred-CMI}}$ and consequently its low dimensional approximations will be influenced by the treatment assignment mechanism. We validate this in the last section and explore simple, yet effective extensions that use propensity-score weighting and stratification as pre-processing steps. The first approach modifies the data on which algorithm 1 is applied, while the second modifies the step that estimates the mutual information (i.e. the term $\hat{J}(X_i)+ = \hat{\mathrm{I}}(T; Y \mid X_j X_i)$). As we will see, our motivation is to perform covariate selection with INFO+ in such a way so that any influence of the treatment assignment mechanism is limited. This is useful in conditionally randomised and/or observational studies with no hidden confounders where the treatment groups are imbalanced in their covariate distributions.

4.4 Simulated Data

Following the bulk of the literature on treatment effect estimation, e.g. (Foster et al., 2011; Lipkovich et al., 2011; Nie and Wager, 2021; Wager and Athey, 2018b; Anoke et al., 2019), we will first evaluate INFO and INFO+ using simulated randomised studies, where we know the ground truth for the covariates that interact with the treatment. In this section we assume 1:1 marginally randomised studies.

Let us denote as $\boldsymbol{X}_{pred}$ the set of predictive covariates and $\widehat{\boldsymbol{X}}_{pred}$ the set of covariates identified at the top $k = |\boldsymbol{X}_{pred}|$ positions. We define the true positive rate (TPR) as the fraction of predictive covariates correctly ranked at the top k positions:

$$\text{TPR} = \frac{|\boldsymbol{X}_{pred} \cap \widehat{\boldsymbol{X}}_{pred}|}{|\boldsymbol{X}_{pred}|}$$

TPR captures how accurate is an algorithm in correctly identifying the predictive covariates. Let us also define as $\boldsymbol{X}_{prog}$ the prognostic covariates and $\boldsymbol{X}_{irr}$ the irrelevant covariates. The False Negative Rate (FNR = 1 -TPR) can be decomposed as follows:

$$\text{FNR} = \frac{|(\boldsymbol{X}_{prog} \setminus \boldsymbol{X}_{pred}) \cap \widehat{\boldsymbol{X}}_{pred}|}{|\boldsymbol{X}_{pred}|} + \frac{|\boldsymbol{X}_{irr} \cap \widehat{\boldsymbol{X}}_{pred}|}{|\boldsymbol{X}_{pred}|} = \text{FNR}_{prog} + \text{FNR}_{irr}$$

FNR_{prog} captures how often an algorithm selects as predictive covariates those that are solely prognostic.

We compare INFO+ against a counterfactual modelling method (Virtual Twins), a subgroup modelling method (SIDES) and a treatment effect modelling (MCR). The Virtual Twins (VT) approach was initially proposed for subgroup identification problems (Foster et al., 2011). Following Foster et al. (2011) we train a Random Forest using the variables $\{\boldsymbol{X}, T, \boldsymbol{X}T, \boldsymbol{X}(1-T)\}$. For an observation $\boldsymbol{x}$ with treatment t we estimate the probability of the factual outcome, $\hat{p}(y = 1 \mid \boldsymbol{x}, t)$ using the out-of-bag estimate. To estimate the counterfactual we switch the treatment to $\bar{t}$, which is 0/1 if $t = 1/0$. Following their implementation we used 1000 trees for this step. The difference between the predicted probabilities for each $\boldsymbol{x}$ can be used to express CATE. Then the authors train a decision tree on the estimated treatment effect to partition the input space into subgroups of heterogeneous treatment effects. Instead, we train a RF using 1000 trees and rank the covariates based on their importance score as described in Chapter 3.

The first step of VT was performed using the *R* package *aVirtualTwins* (Vieille, 2018). For SIDES we used the original algorithm[1] (Lipkovich et al., 2011) with the default parameters as described by Lipkovich et al. (2017a). For the Modified Covariates Method (MCR) (Tian et al., 2014) method, the user has to specify the main effects and the interactions. We assume no prior knowledge and include in the model only first order interactions with the treatment. In order to rank the covariates based on their predictive "strength" we used $L2$ regularisation and optimised the regularisation parameter via cross-validation using the package *glmnet* (Friedman et al., 2010). We used the default parameters for the number of folds, which is 10. In the simulations that use MCR this will leave 100 observations to get an estimation of the error and 900 for training with the dimensionality being 20. The regularisation parameter was selected using the one-standard-error rule, i.e. selecting the largest value such that the error is within one standard error of the minimum (Hastie et al., 2009). This takes into account that the measure that is optimised (the cross-validation error) will change over different runs, hence taking a conservative approach and picking a regularisation parameter that is likely to be optimal (Hastie et al., 2009). As we will notice VT and SIDES tend to be stronger competitors, so we will focus primarily on the results of these methods.

For all simulations we report the results averaged over 200 realisations of the outcome functions. Unless specified otherwise the covariates have a marginal distribution that is the standard normal $\mathcal{N}(0,1)$ and any two covariates X_i, X_j have a correlation equal to 0.7 if both $i \neq j$ are even or odd and 0 otherwise.

4.4.1 Correlated Covariates and Interactions

In this section we explore the low-dimensional approximations of Pred-CMI, namely INFO and INFO+. The first criterion captures only first-order interactions with the treatment, while the second can capture additionally second-order interactions. In order to validate that we compare the two criteria using the following simulated outcomes:

M1 : $\text{logit}(p(Y=1 \mid T, \boldsymbol{X})) = \sum_{j=1}^{4} X_j + \beta_{pred} T \sum_{j=5}^{8} X_j$

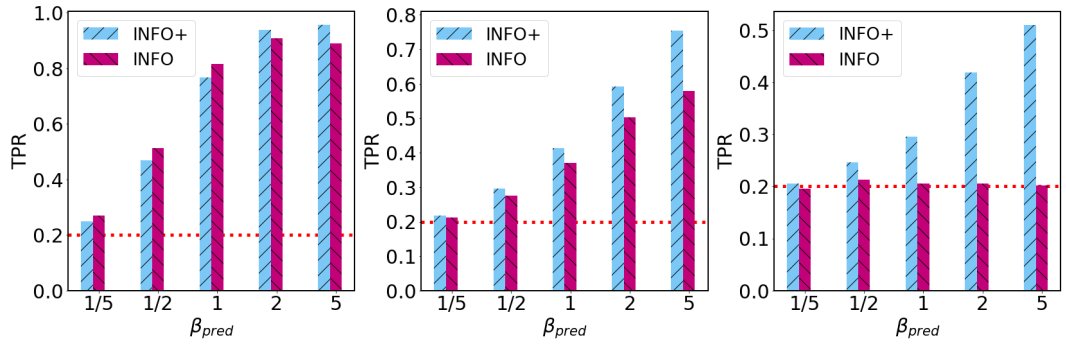
M2 : Same as M1 but with correlated covariates

M3 : $\text{logit}(p(Y=1 \mid T, \boldsymbol{X})) = X_1X_2 + X_3X_4 + \beta_{pred} T\big(X_5X_6 + X_7X_8\big)$

[1]The code can be found on the Biopharmaceutical Network web site at: `http://biopharmnet.com/subgroup-analysis/` [last accessed: 17/12/2020].

We generate samples with size $n = 1000$ and $d = 20$ covariates. The covariates are then discretised in 2-5 bins following an equal width strategy.

We vary the predictive strength β_{pred} in order to explore what happens under different degrees of difficulty. In figure 4.1(a) we observe that in the first scenario the two criteria have similar TPR. The dotted horizontal line corresponds to the expected result under random selection. The second scenario is more challenging due to the correlations between the covariates, in which case INFO+ tends to perform better (figure 4.1(b)). This is highlighted even more in third scenario where we also have interactions between the covariates. In this case the univariate criterion fails and has a TPR close to what it would be under random selection. For the rest of this chapter we will use only INFO+.



(a) M1 (b) M2 (c) M3

Figure 4.1: INFO+ that captures second-order interactions outperforms the univariate criterion in the presence of correlated covariates and interactions.

4.4.2 Varying the Predictive and Prognostic Strength

We define as β_{main} the coefficient of the main effect and β_{pred} the coefficient of the interaction. We would like to explore how different methods compare as we vary these parameters. We generate the outcomes as follows:

$$\text{M4} : \text{logit}(p(Y = 1 \mid T, \boldsymbol{X})) = \beta_{main} \sum_{j=1}^{5} X_j + \beta_{pred} T \mathbb{1}(X_1 > 0 \cap X_2 < 0)$$

$$\text{M5} : \text{logit}(p(Y = 1 \mid T, \boldsymbol{X})) = \beta_{main} \sum_{j=1}^{5} X_j + \beta_{pred} T X_1 X_2$$

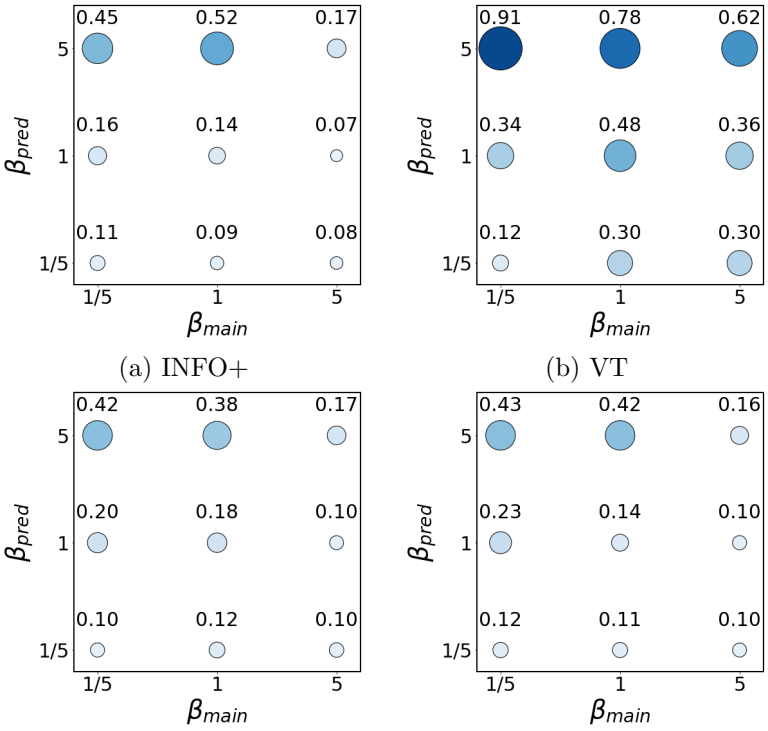
We generate $n = 1000$ observations and $d = 20$ covariates. For M4 we used continuous covariates, while for M5 we consider discrete covariates, by discretising the data in 2-5 bins following an equal-width strategy (Dougherty et al., 1995).

In order to handle the continuous case for INFO+ we used an estimator for the conditional mutual information, which is based on non-parametric density estimation procedure. The core idea is to transform the continuous covariates to categorical using a method for histogram generation, such as Scott's rule or the Freedman-Diaconis' rule (Scott, 1992) and then to use the shrinkage estimator. Here we used Scott's rule.

The results are shown in figure 4.2 for M4 and figure 4.3 for M5. Each circle in the graphs corresponds to the TPR for a pair of values for β_{main} and β_{pred}. We observe that INFO+ tends to perform better than SIDES and MCR for large predictive effects and particularly for M4, but can perform worse in some cases in the presence of small predictive effects and when using continuous data. Overall, INFO+, SIDES and MCR tend to increase their TPR for higher values of β_{pred} and/or lower values β_{main}. VT shows a different behaviour. It outperforms all competing approaches, especially in the continuous case, but its TPR tends to increase even if we keep β_{pred} to a small value and increase β_{main}. This suggests that VT tends to be affected by the main effect to a larger extend than the competing methods. However, in order to study this in more detail we need to distinguish the predictive covariates from the main effect and observe what happens in TPR and FNR_{prog} separately. In the next section we take a closer look at what happens when there is only main effect and no interaction with the treatment.

4.4.3 Homogeneous Effects

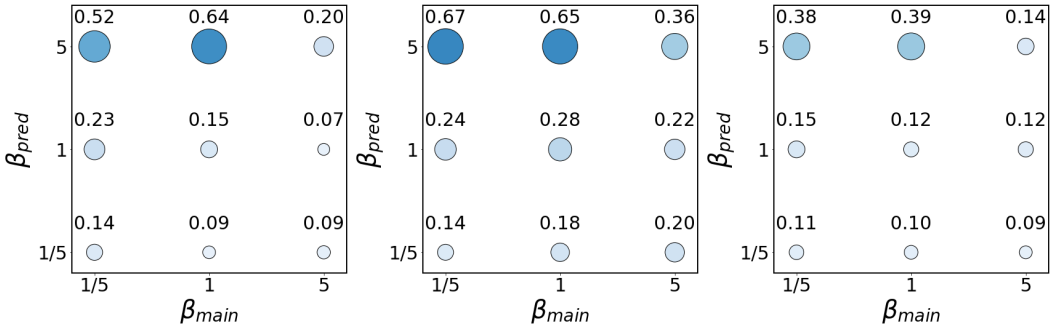
In this section we explore what happens when there is no covariate exhibiting an interaction with the treatment and no overall effect. This is the scenario of having homogeneous effects, in which case we expect that no covariate would be preferred over the others. We study a simple case where the outcome is defined as $\text{logit}(p(Y = 1 \mid T, \boldsymbol{X})) = \sum_{j=1}^{5} X_j$. We generate $n = 1000$ observations and $d = 20$ covariates and report the position of each covariate in the ranking averaged over 200 repetitions. In figure 4.4 we report the average position of each covariate. The vertical dotted line indicates the expected position of a covariate under random selection. We observe that VT tends to identify the prognostic covariates at the top positions, while the other methods do not show any preference. These results show that VT is biased towards identifying prognostic covariates as predictive. For the rest of this section we will omit MCR and focus on VT that achieves



(a) INFO+ (b) VT

(c) MCR (d) SIDES

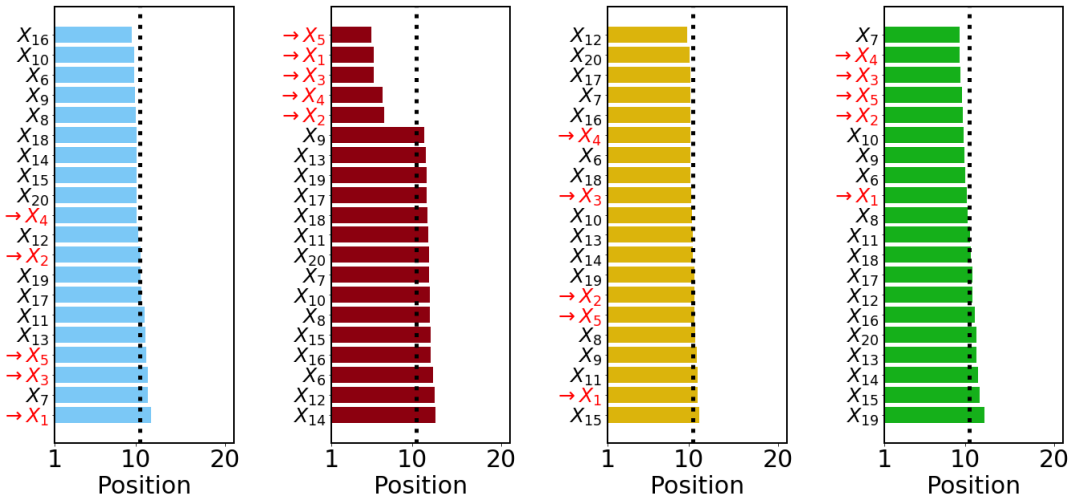
Figure 4.2: INFO+ is sensitive to the predictive "strength" and achieves a higher TPR for larger values of β_{pred} and lower values of β_{main} as we observe here for model M4. Similar behaviour is observed with MCR and SIDES. In contrast VT tends to achieve higher TPR even if we keep β_{pred} constant and increase the value of β_{main}.



(a) INFO+ (b) VT (c) SIDES

Figure 4.3: INFO+ performs similarly to VT for categorical data (model M5) and large values of the predictive "strength". For fixed β_{pred} increasing β_{main} may result in higher TPR when using VT. This is not observed to the same extend with the other methods.

high TPR but can be biased to prognostic covariates and SIDES that tends to show lower TPR but does not exhibit such biases. Additionally, these methods can handle both continuous and categorical covariates without any modification.



(a) INFO+ (b) VT (c) MCR (d) SIDES

Figure 4.4: In the absence of treatment effect, INFO+, MCR and SIDES perform similarly to random selection (the average position of each covariate is close to the vertical dotted line). In contrast VT tends to rank the prognostic covariates at the top positions.

4.4.4 Distinguishing Prognostic and Predictive Covariates

The experiment of this section focuses on two scenarios where the covariates can be either solely predictive or solely prognostic. We would like to explore how each method can distinguish between the two. We use the following outcome functions:

$$\text{M6} : \text{logit}(p(Y=1 \mid T, \boldsymbol{X})) = \sum_{j=1}^{4} X_j + 5T(X_5 + \mathbb{1}(X_6 > -0.545 \cap X_7 < 0.545)$$

$$\text{M7} : \text{logit}(p(Y=1 \mid T, \boldsymbol{X})) = \sum_{j=1}^{4} X_j + 5T(X_5 + X_6 X_7)$$

We generate $d = 20$ covariates and report TPR and FNR_{prog} for increasing sample size. For M6 we used continuous covariates, while for M7 we consider discrete covariates, by discretising the data in 2-5 bins following an equal-width strategy. In figure 4.5 we observe that INFO+ achieves both higher TPR and lower FNR_{prog} compared to VT and higher TPR compared to SIDES. On the other hand, VT often selects solely prognostic covariates as predictive, which results in a higher

FNR_{prog}. This is more clearly observed in the first scenario. In the second scenario all approaches have a slightly higher FNR_{prog} but as the sample size increases this tends to remain constant for INFO+ but it increases for SIDES and VT. The results of this section suggest that INFO+ can successfully distinguish the predictive and prognostic covariates.

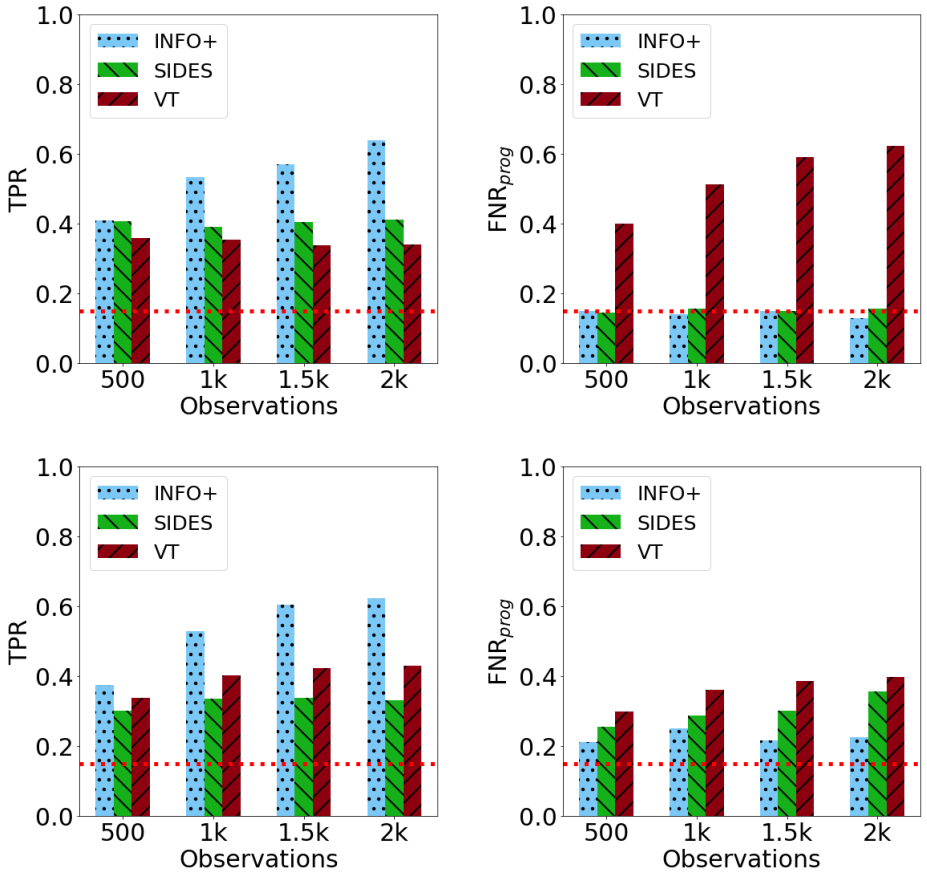


Figure 4.5: INFO+ and SIDES can distinguish between predictive and prognostic covariates. On the other hand, VT may wrongly identify solely prognostic covariates as predictive, as indicated by the large FNR_{prog}.

4.4.5 Computational Time

Information theoretic criteria allow us to identify only the top-k covariates, with $k \ll d$, as opposed to VT and SIDES which will rank all the covariates. This results in significant computational savings as shown in figure 4.6 where we report the time required (in logarithmic scale) to identify the most predictive covariates for model M4. If we were to rank all covariates then the required computational

time of the INFO+ is similar to SIDES for small sample sizes and better as the sample size increases.

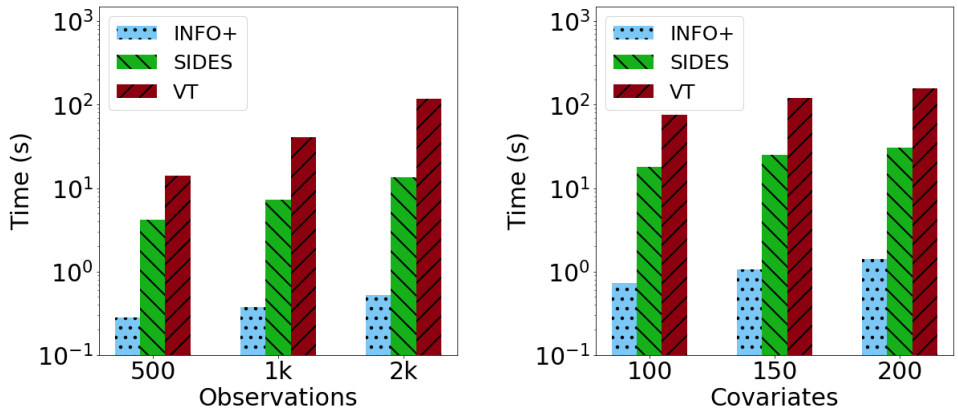


Figure 4.6: (a) Computational time required to identify the two most predictive covariates for M4. (b) Time required to identify the top 20 covariates using 1000 observations and increasing dimensionality.

4.5 Case Studies

In this section we apply INFO+ in three scenarios with diverse characteristics. In the first two there are known predictive covariates while the other is a real clinical trial dataset where we do not have prior knowledge of potentially predictive covariates. Regarding the analysis of the clinical trial data, we highlight that the purpose of our experiments is not to reproduce the results of the corresponding studies. In contrast, we treat them as binary classification tasks and explore whether the identified predictive covariates are plausible.

4.5.1 Application to Simulated Clinical Trial Data

We evaluate INFO+ on a simulated dataset, a description of which can be found in (Lipkovich et al., 2017a). The dataset consists of 470 patients with severe sepsis who were randomly assigned to either a treatment group that received a novel therapy or the control group that received standard care. The outcome is survival at 28 days. Here we use a version of the dataset that is available with the *R* package *aVirtualTwins* (Vieille, 2018). In this dataset there are 11 covariates and the true subgroup is defined by the patient's age and the pre-infusion apache-ii score.

Here we would like to explore how INFO+ would perform in a setting where all covariates are continuous that follow different distributions as indicate by their histograms, while there are also potential outliers. In order to apply INFO+ we discretise the covariates using the K-means algorithm with K=3. We repeat the algorithm for 500 bootstrap samples and observe the average score. The score of a covariate is given by $(d-r_f+1)/d$ where r_f is the position of the f-th covariate in the ranking and d is the number of covariates. We find that INFO+ ranks first the pre-infusion apache-ii score and second the patient's age. The average scores are 0.96 and 0.84 respectively (table 4.1). The results validate that INFO+ can identify known predictive covariates after discretisation.

4.5.2 Application to Real Clinical Trial Data

Firstly, INFO+ was validated in a clinical trial dataset were there is a known predictive covariate. The IPASS study (Mok et al., 2009; Fukuoka et al., 2011) was a Phase III, multi-center, randomised, open-label, parallel-group study comparing gefitinib (Iressa, AstraZeneca) with carboplatin (Paraplatin, Bristol-Myers Squibb) plus paclitaxel (Taxol, Bristol-Myers Squibb) as first-line treatment in clinically selected patients in East Asia who had advanced non small-cell lung cancer (NSCLC). The trial consisted of 1217 patients randomly assigned in the two treatment arms. The outcome of interest is progression-free survival, which was modelled as a binary endpoint, neglecting its time-to-event nature. The data were analysed at 78% maturity, when 950 subjects have had progression events. Covariates with missing data were handled by creating an additional category (Allison, 2001). It is known that gefitinib inhibits the epidermal growth factor receptor (EGFR) and is now indicated as a first-line treatment for patients with NSCLC whose tumours have specific EGFR mutations. It is therefore expected the EGFR mutation status to appear as a strongly predictive covariate. INFO+ was applied in 500 bootstrap samples and the average score was calculated. The ranking based on the average score showed that the EGFR mutation was ranked first. We will now explore a case study where the is no known predictive covariate.

The AURORA study was a randomised, double-blind, placebo-controlled, multicenter trial in which 2776 patients with end-stage renal disease were randomly assigned 1:1 to double-blind treatment with rosuvastatin at a dose of 10 mg or placebo. The primary endpoint was the time to a major cardiovascular event (MACE) defined as a nonfatal myocardial infarction, nonfatal stroke, or

death from cardiovascular cause. For full details of the trial see (Fellström et al., 2009).

Here the outcome is treated as binary indicating the presence of a MACE. Following Lipkovich et al. (2017a) for the patients with missing values in a categorical covariate an additional category is created (i.e. missing indicator method), while the missing values for continuous covariates were replaced by the mean of that covariate (i.e. mean imputation) (Allison, 2001). Furthermore, the continuous covariates were discretised in 5 bins following an equal width strategy. Similarly to the previous studies the covariate selection algorithm is repeated for 500 bootstraps. INFO+ identifies at the top positions the following: Blood Lymphocytes, Serum Apolipoprotein B and Blood Leukocyte Particle Concentration. In contrast to the previous studies, in this case there is no prior information on potentially predictive covariates. In this case INFO+ does not place high confidence on a specific covariate and the average scores are not very high (below 0.9). The detailed ranking can be found in table 4.1. The results of INFO+, i.e. not showing a strong preference towards a covariate are in agreement with the trial findings. In contrast we note that VT identified the patient's age at the top position with high confidence (average score close to 1), a covariate that has previously been identified as a risk factor for MACE in a post hoc analysis (Schneider et al., 2013).

Table 4.1: The top selected covariates for two studies based on their average score over 500 bootstrap samples.

Data	1st	2nd	3rd
Sepsis	Apache	Age	Glasgow coma scale
Aurora	Lymphocytes	Apolipoprotein B	Leukocyte conc.

4.6 Addressing a Limitation of INFO+: Extensions in the Presence of Confounders

The approaches studied in this chapter are primarily designed for randomised studies where there is sufficient balance between the covariates of the treatment groups. In particular, for INFO+ we know from our previous analysis and from eq. (4.5) that is affected by the treatment assignment mechanism. We would like to explore how standard pre-processing steps such as propensity-score weighting

and stratification can be used along with INFO+ when facing studies where the treatment is not marginally randomised. An advantage of INFO+ over some existing approaches is that it does not require a model for inference, i.e. it is a filter method (Guyon et al., 2008). Hence we can avoid the problem of model selection and/or introducing additional errors from estimation. With the aforementioned pre-processing steps we still avoid the use of a model for the outcome since we use only the covariates and the treatment. However, we now need to introduce hyperparameters relating to the estimation of the treatment assignment.

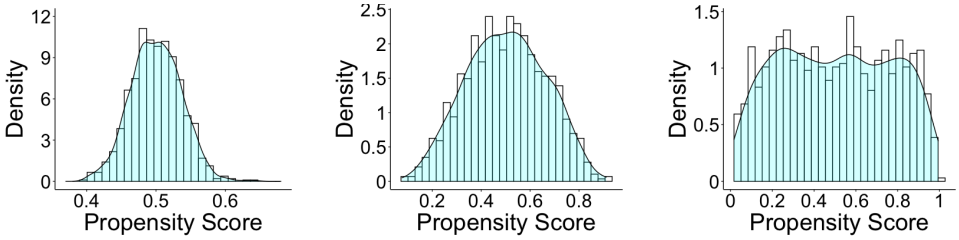
Following the discussion of Chapter 2 we can estimate the propensity score and use the inverse as a weight in order to up-weight each observation accordingly (Robins et al., 2000). Based on this we create a new sample, where each observation is repeated as many times as the corresponding weight and apply INFO+ in the new sample. We refer to this approach as INFO+W. An alternative approach is to perform stratification on the propensity score. In particular, we divide the sample into strata based on the value of their propensity score, so that observations with similar propensity score fall within the same stratum. For example, when using 5 equal-width strata, for the observations with propensity score in (0,0.2) we assume that they come from an approximately marginally randomised study and apply INFO+ as usual. We repeat this for the rest of the sub-samples, i.e. [0.2,0.4), [0.4,0.6], [0.6,0.8], [0.8,1) and average the scores of INFO+ for each covariate to get its final score. Therefore, the covariates are ranked based on the value of the criterion averaged over all strata. We refer to this as INFO+S.

In order to explore how INFO+ performs in the presence of confounders, as well as how the aforementioned approaches can be used to correct it, we revisit the model M5 but change the treatment assignment mechanism considering the following scenarios:

$$
\begin{aligned}
\text{PM1}: \ & \text{logit}(p(T=1 \mid \boldsymbol{X})) = \gamma(X_1 + X_2) \\
\text{PM2}: \ & \text{logit}(p(T=1 \mid \boldsymbol{X})) = \gamma(X_3 + X_4 + X_5) \\
\text{PM3}: \ & \text{logit}(p(T=1 \mid \boldsymbol{X})) = \gamma(X_1 + X_2 + X_1X_2 + X_1^2 - X_2^2) \\
\text{PM4}: \ & \text{logit}(p(T=1 \mid \boldsymbol{X})) = \gamma(X_3 + X_4 + X_5 + X_3X_4 + X_3X_5 + X_4X_5 - X_3^2 + \\
& \qquad + X_4^2 - X_5^2)
\end{aligned}
$$

In the first case the treatment assignment depends on predictive covariates, while

in the second case it depends on solely prognostic covariates. Additionally, we consider modifications of PM1 and PM2 so that the treatment assignment depends on the covariates, their squared terms and all pairwise interactions. We explore what happens as we increase the value of γ, i.e. increasing the confounding strength and hence creating more imbalanced samples. The result of increasing γ on the values of the propensity score is shown in figure 4.7 for propensity model PM1. As the confounder strength increases so does the number of observations with propensity score close to 0 and 1. This is particularly the case for PM3 and PM4 under $\gamma = 1$. We would like to explore what happens under such a scenario in terms of predictive covariate selection. In all cases we do not assume knowledge of the true propensity score, which needs to be estimated.

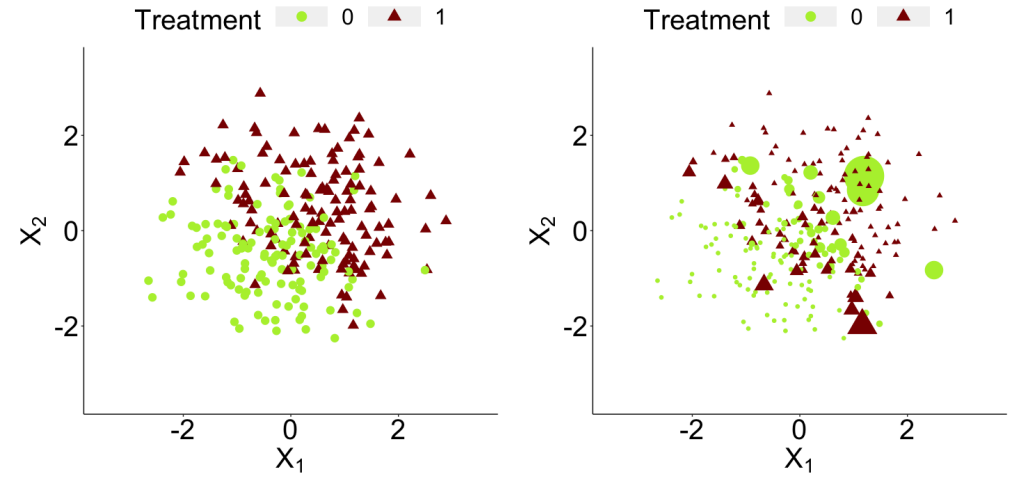


(a) $\gamma = 0.1$ (b) $\gamma = 0.5$ (c) $\gamma = 1$

Figure 4.7: Histograms of values of the propensity score for PM1 and different values of the confounding strength γ. As the confounding strength increases the values of the propensity score move away from 0.5 resulting in increased imbalance between the treatment groups.

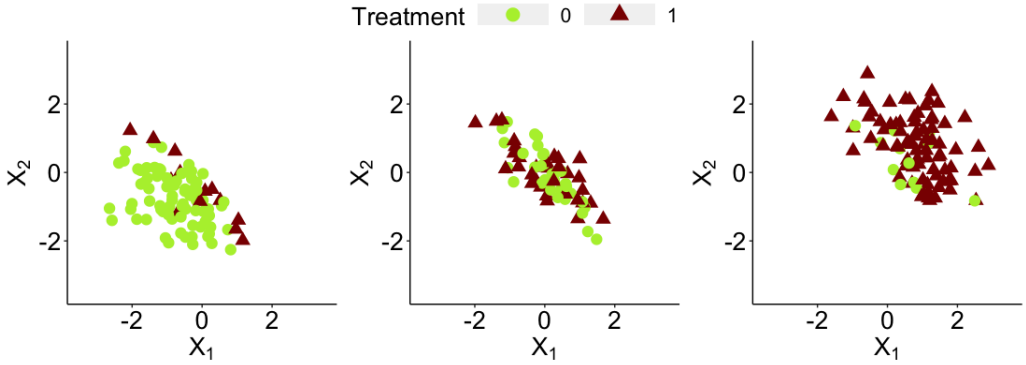
In figure 4.8 we show an example of how the described pre-processing steps perform in practice. With propensity score weighting observations that have $T = 0/1$ and lie in regions where treatment group $T = 1/0$ is over-represented are over-sampled in order to create a balanced sample. With propensity score stratification the observations are grouped so that each group can be approximately treated as a sample from a randomised study with a constant propensity score.

In all cases we fit a linear propensity score model. We study two scenarios: using the estimated propensity score and re-weighting and using propensity score stratification. The number of strata can be considered a parameter to be selected by the user. Here we choose 5 as this is a common choice in the literature (Lunceford and Davidian, 2004; Austin, 2009b). A data-driven approach could also be considered here. In particular, we could determine the minimum sample size that would be required to get a reliable estimation of the conditional mutual information. This will depend on the cardinality of the involved covariates as well



(a) Original Data

(b) After Weighting



(c) After Stratification

Figure 4.8: We generate a dataset using PM1 and $\gamma = 1$. In (a) we report the distribution of the covariates X_1 and X_2. In (b) we show how with propensity score weighting certain observations are over-sampled in areas with limited overlap. In (c) we observe how propensity score stratification with three strata creates groups with different probabilities of treatment assignment. These probabilities are from left to right, 0.15, 0.5 and 0.85.

as the used estimator. In this case some preliminary analysis would be required in which case we could explore the behaviour of the estimation error under different values of the conditional mutual information e.g. by simulating contingency tables or referring to existing empirical analyses of the used estimator. We could then define the number strata so that each stratum has the minimum sample size that is required. On the other hand, we could select the number of strata by exploring the balance of the covariates between the treatment groups and then search for an estimator that would perform well given the sizes of selected strata. In any

case larger number of strata will result in smaller sample sizes hence making the estimation more challenging.

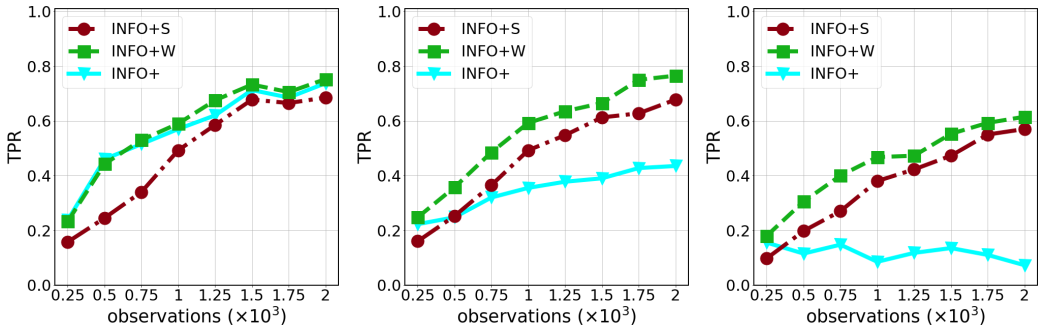
In figure 4.9 we plot the TPR averaged over 200 simulated datasets, using 20 covariates and increasing sample size. We report the results for increasing value of γ. More specifically the left-hand side plots correspond to $\gamma = 0.1$, the plots in the middle are for $\gamma = 0.5$ and the right-hand size plots correspond to $\gamma = 1$. We observe that we can partially correct INFO+ using simple pre-processing techniques, with stratification being the best performing in most settings with $\gamma > 0.1$. For example when $\gamma = 0.5$ both stratification and propensity score weighting perform in most cases similarly to the case with $\gamma = 0.1$. The case of $\gamma = 1$ becomes more challenging since more observations have propensity score close to 0 and 1. We note that even after these pre-processing steps we still observe some bias towards the treatment assignment mechanism. In particular, when the treatment assignment depends on prognostic covariates INFO+ tends to perform better compared to when the dependency is on predictive covariates (we remind the reader that according to eq. (4.5) the covariates that have larger mutual information with the treatment are penalised more heavily). This is also the case for INFO+S which shows higher TPR compared to the other methods when the treatment assignment depends on prognostic covariates (PM2,PM4). Therefore, even though we can improve standard INFO+, the resulting methods may still be influenced but to a lower degree by the treatment assignment.

From the conducted simulations we cannot conclude whether INFO+S or INFO+W would be better suited in a particular setting. We can however note some properties of the two approaches. INFO+S uses a smaller sample size as it replaces $\hat{J}(X_i)+ = \hat{\mathrm{I}}(T; Y \mid X_j X_i)$ in Algorithm 1 with:

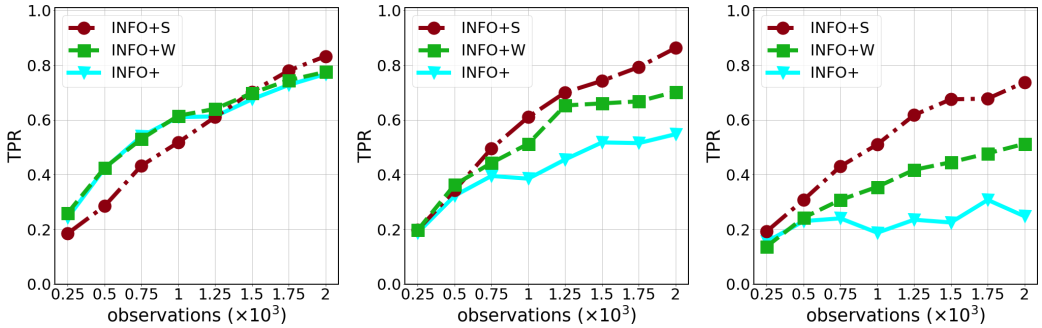
for $\mathcal{D}_m \in D_{all}$ **do**
  $\hat{J}(X_i)+ = \frac{n_m}{n}\hat{\mathrm{I}}(T^m; Y^m \mid X_j^m X_i^m)$
end for

where $\mathcal{D}_m$ is the dataset for m-th stratum with n_m examples and D_{all} denotes the set of all created datasets after stratification. We notice in the simulations that when $\gamma = 0.1$ where we expect all methods to perform similarly due to the small confounding strength, INFO+S shows lower TPR than the other methods particularly for small sample sizes which could be attributed to this. INFO+S also introduces an additional parameter, the number of strata, which needs to be chosen by the user as we discussed above. On the other hand INFO+W does

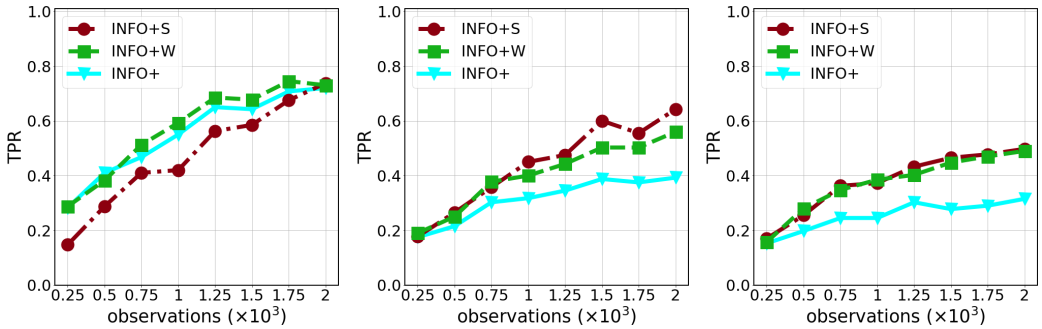
not have any additional parameters and uses the full dataset, but it acts as an approximation, since it does not use the estimated IPW weights directly in the estimation and instead it up-weights the observations based on the closest integer value of the IPW weight (this can also be adjusted by the user for better precision). To this end, we could potentially explore extensions of INFO+W that incorporate the weights directly in the estimation. This would likely require revisiting the predictive covariate identification objective given in Definition 1 (perhaps by considering a weighted likelihood), since as we discussed the information theoretic criterion that follows from this definition will be sensitive to the treatment assignment. In this chapter we chose to not change the definition and instead perform some pre-processing steps so that it would be applicable in the presence of confounders.



(a) PM1

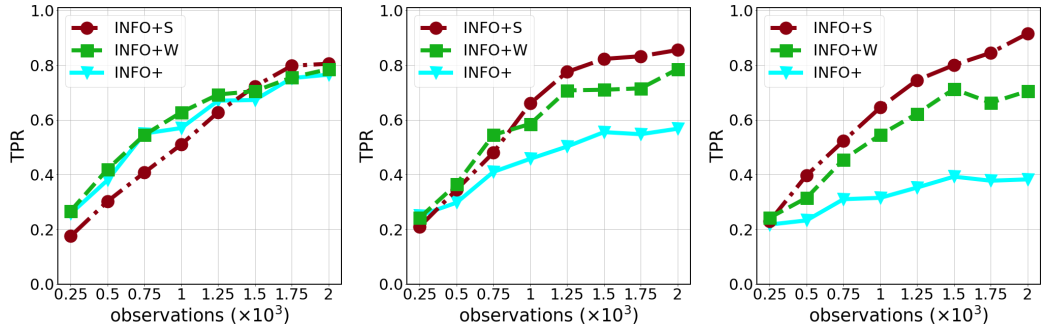


(b) PM2



(c) PM3

Figure 4.9: INFO+ is influenced by the treatment assignment mechanism. Combining INFO+ with propensity score weighting and stratification can ameliorate some of the issues that arise. Here we plot the TPR for four different scenarios and increasing sample size. From left to right we report the results for $\gamma = 0.1, 0.5$ and 1.



(d) PM4

Figure 4.9: INFO+ is influenced by the treatment assignment mechanism. Combining INFO+ with propensity score weighting and stratification can ameliorate some of the issues that arise. Here we plot the TPR for four different scenarios and increasing sample size. From left to right we report the results for $\gamma = 0.1, 0.5$ and 1. *(cont.)*

4.7 Chapter Summary

In this chapter we discussed methods for identifying predictive covariates using information theory. These methods are all derived from the objective of maximising the difference of two log-likelihood functions which in turn results in an information theoretic objective. We discussed some theoretical properties of INFO+ and compared it against three approaches, representative of the frameworks described in the previous chapter: VT (Foster et al., 2011), SIDES (Lipkovich et al., 2011) and MCR (Tian et al., 2014). Unlike VT, the information theoretic approach does not require to build a prediction model to estimate the potential outcomes. Also, in contrast to recursive partitioning methods, it uses all available data for estimating the predictive strength and for capturing possible interactions between the covariates. In addition, for INFO+ the user does not need to define a functional form of the outcome (in contrast to MCR where this is required) and perform hyper-parameter selection. In particular the studied low-dimensional criteria (INFO+, INFO+S, INFO+W) can capture second-order interactions between the covariates. We may account for higher order interactions with a simple modification of the criterion and by adopting an appropriate estimator for the mutual information (Sechidis et al., 2019a).

The experimental evaluation was performed such that we can explore what happens when both discrete and continuous covariates are used but also when

the predictive covariates are also prognostic. From the conducted simulations we notice that each method has different characteristics. In particular, VT can achieve high TPR when the predictive covariates are also prognostic but it often fails to distinguish between the two types and can be influenced by the prognostic strength. This behaviour can result in more false discoveries. In the simulations SIDES did not show a bias towards identifying prognostic as predictive, but it showed lower TPR than INFO+ when the prognostic and predictive covariates were different. INFO+ can better distinguish the two types of covariates, but it also has a limitation that may influence its applicability. Its performance will depend on the estimator for the mutual information and the discretisation of the continuous covariates (if applied). Hence applying it in a mixed data scenario will require these to be taken into account by the researcher, while other methods, such as VT, can be applied directly. It is also worth noting that information theoretic methods allow us to return the top-k covariates, which is practically relevant since domain experts are often interested in only a few covariates (the most predictive). This results in substantial computational savings compared to the other approaches.

We also studied the behaviour of INFO+ in the presence of confounders and showed that it can be influenced by the treatment assignment mechanism. We proposed two simple yet effective modifications that allow us to ameliorate these issues namely INFO+W and INFO+S. We validated these criteria using simulated data. Note that a characteristic of filter criteria (such as the information theoretic) is that they are independent of any predictor. Our extensions use a model for the treatment, however they remain independent of any predictor of the outcome.

In this chapter we studied the problem of predictive covariate selection. This procedure allows us to generate hypotheses for potential interactions with the treatment and reduce the dimensionality of the dataset. Alternatively, we may wish to directly search for subgroups, especially when the sample size and the dimensionality allows us to do it. Most existing approaches, such as SIDES (Lipkovich et al., 2011) and IT (Su et al., 2008) focus on subgroup identification in marginally randomised studies. We now turn our attention to exploring this problem in the presence of confounders using non-parametric weighting methods.

Chapter 5

Subgroup Identification using Weighting Methods

In this chapter we turn to the problem of subgroup identification and we focus on methods that perform recursive partitioning of the space. These are attractive due to their simplicity and interpretability as the resulting subgroups are described by a set of rules. Most of the methods discussed in the previous chapter cannot be applied in studies where there is imbalance between the covariate distributions of treatment groups due to the presence of confounders. For example, MCR (Tian et al., 2014), SIDES (Lipkovich et al., 2011), Qualitative Interaction Trees (QUINT) (Dusseldorp and Van Mechelen, 2014) and Interaction Trees (IT) (Su et al., 2009), to name a few, are applicable in randomised studies where there is balance between the treatment groups. Subgroup identification in scenarios of increased imbalance becomes more and more common with the increased availability of observational data (e.g. medical databases).

This chapter discusses modifications of recursive partitioning approaches that allow us to perform subgroup identification in these scenarios. We focus on IT (Su et al., 2009) and study firstly how it performs in the presence of confounders and secondly how we can modify the methodology in order to tackle the issues that arise. The choice of IT over other recursive partitioning approaches, such as SIDES, is motivated primarily due to its simplicity as it requires less hyper-parameters and follows the rules of CART (Breiman et al., 1984) a well studied approach in the literature. This has been extended recently including modifications that combine IT with regression (Steingrimsson and Yang, 2019) or propensity score weighting (Yang et al., 2021).

In this chapter we revisit the problem discussed in (Yang et al., 2021), where the authors propose Causal Interaction Trees (CIT), an extension of IT for observational studies with no hidden confounders using IPW, DR and regression estimators. We suggest a new methodology that adopts recently proposed non-parametric weighting methods (Kallus and Santacatterina, 2019b; Kallus et al., 2021). The motivation is two-fold; Firstly, IPW estimators require correct specification of a propensity score model and secondly balance between the re-weighted data is not always achieved especially under strong confounding. The latter is particularly important in the context of subgroup identification with recursive partitioning where we encounter small sample sizes as we partition the space. As we discussed in Chapter 2 these observations have motivated weighting methods that optimise the balance between the treatment groups directly. Here we follow the method proposed in (Kallus et al., 2021) to get weights that optimise clearly defined quantities, such as the conditional bias or MSE of the sample average treatment effect. This approach has shown improved empirical results compared to other weighting estimators, including IPW, and here we study its use in the context of subgroup identification. Similarly, to CIT we modify the splitting criterion, while additionally we explore what happens as we increase the confounding strength, as well as how the original IT algorithm performs in the presence of imbalance between the treatment groups.

Section 5.1 defines the problem and shows why some existing methods may be problematic in the presence of confounders. Section 5.2 describes the general principles of the recursive partitioning approach followed by IT and CIT which will also be adopted in this chapter. Section 5.3 describes the non-parametric weighting methods and the modifications performed in the original IT. Finally, Section 5.4 validates the approach using simulated data and Section 5.5 shows the results in two case studies.

5.1 Problem Definition

Consider a study conducted to evaluate the efficacy of a novel treatment. We assume an i.i.d. dataset $\{\boldsymbol{x}_i, t_i, y_i\}_{i=1}^n$, where $\boldsymbol{x}_i$ denotes the d-dimensional vector of covariates of the i-th subject, t_i is the assigned treatment and y_i is the outcome. Additionally, suppose that for the support of the treatment variable we have $\mathcal{T} = \{0, 1\}$. For ease of exposition, we use $t_i = 0$ to denote subjects in the control

arm and $t_i = 1$ for subjects in the experimental treatment arm.

A subgroup $\mathcal{S}$ is most commonly defined by a rule that determines which observations belong to a favourable group according to the values of their covariates. For example, a subgroup may be defined as the subset of the data for which some covariate Z is positive, $\mathcal{S} = \{\boldsymbol{x}_i : z_i > 0\}$. There is an abundance of methods for subgroup identification, some of them discussed in detail in the previous chapters. Since we perform exploratory analysis, any resulting hypothesis should be easy to interpret by domain experts, particularly since this is going to define the sample on which a confirmatory analysis might be performed. To this end, subgroups are usually defined by only a few covariates (e.g. 2–3) with the strongest predictive strength (Foster et al., 2011). Considering the above we will focus on recursive partitioning approaches due to their flexibility (they can incorporate different splitting criteria) and interpretability.

The partitioning of the space is most commonly performed using either some scoring criterion or a statistical test that uses the distribution of such a criterion. In any case a key component is the correct estimation of the treatment effect in the subset of the data determined by the parent node. This can be written as,

$$\text{ATE}_S = \mathbb{E}[Y(1) \mid \boldsymbol{x} \in \mathcal{S}] - \mathbb{E}[Y(0) \mid \boldsymbol{x} \in \mathcal{S}] \tag{5.1}$$

To make the above identifiable from the observed data the following assumptions (Imbens and Wooldridge, 2009; Rosenbaum and Rubin, 1983; Imbens and Wooldridge, 2009) are made, which we discussed in detail in Chapter 2.

1. *Consistency*: We observe $Y = Y(1) \mid T = 1$ and $Y = Y(0) \mid T = 0$, i.e. we observe $Y(1)$ for the experimental treatment arm and $Y(0)$ for the control arm.

2. *Unconfoundedness*: The potential outcomes are independent of the treatment conditioned on the observed variables: $(Y(1), Y(0)) \perp\!\!\!\perp T \mid \boldsymbol{x}$

3. *Overlap*: The probability of receiving the treatment is bounded away from zero: $p(T = t \mid \boldsymbol{x}) > 0, \forall t \in \mathcal{T}$

Let $S(\boldsymbol{x}) = \mathbb{1}\{\boldsymbol{x} \in \mathcal{S}\}$ denote which observations belong in the subgroup for which we want to estimate the treatment effect. In practice we wish to estimate

the sample average treatment effect in the subgroup which can be defined as:

$$\text{SATE}_S = \frac{\sum_i S(\boldsymbol{x}_i)\mathbb{E}[y(1) \mid \boldsymbol{x}_i] - S(\boldsymbol{x}_i)\mathbb{E}[y(0) \mid \boldsymbol{x}_i]}{\sum_i S(\boldsymbol{x}_i)} \tag{5.2}$$

Subgroup identification methods such as IT (Su et al., 2009) and SIDES (Lipkovich et al., 2011) use the raw mean difference of the outcomes under each treatment value as the estimate of the treatment effect. Hence they are suitable for balanced randomised studies where this would be an unbiased estimator of the effect. We will refer to averaging of the outcomes within each group as the unadjusted estimator:

$$\widehat{\text{SATE}}_S^{unadj} = \frac{\sum_{i:t_i=1,S(\boldsymbol{x}_i)=1} y_i}{\sum_i S(\boldsymbol{x}_i)\mathbb{I}(t_i = 1)} - \frac{\sum_{i:t_i=0,S(\boldsymbol{x}_i)=1} y_i}{\sum_i S(\boldsymbol{x}_i)\mathbb{I}(t_i = 0)} \tag{5.3}$$

One approach for estimating the subgroup effects in imbalanced studies is the use of IPW methods. Let $e(\boldsymbol{X})$ denote the propensity score. The estimated treatment effect in the subgroup can be expressed as:

$$\widehat{\text{SATE}}_S^{ipw} = \frac{\sum_i S(\boldsymbol{x}_i)\frac{\mathbb{I}(t_i=1)y_i}{\hat{e}(\boldsymbol{x}_i)} - S(\boldsymbol{x}_i)\frac{\mathbb{I}(t_i=0)y_i}{1-\hat{e}(\boldsymbol{x}_i)}}{\sum_i S(\boldsymbol{x}_i)} \tag{5.4}$$

As we discussed in Chapter 2 the theoretical properties of IPW estimators hold under correct specification of the propensity score. Studies have found that miss-specifications may lead to substantial bias of the estimated effect (Kang et al., 2007). Additionally, the estimation of SATE_S with IPW estimators can become particularly challenging in the presence of small samples and under strong confounding. This is crucial in subgroup identification via recursive partitioning, where we need to estimate effects in smaller subsets of the data as we move deeper in the tree. To this end we instead focus on methods that either optimise the balance of the groups directly or some error function of the treatment effect.

Most recently, optimisation methods such as Entropy Balancing (EBAL) (Hainmueller, 2012), Stable Weighting (SW) (Zubizarreta, 2015) and Kernel Optimal Matching (KOM) (Kallus, 2020b; Kallus and Santacatterina, 2019b; Kallus et al., 2021) have shown good empirical performance over standard IPW approaches for estimation of SATT and SATE. From these methods KOM has additionally some desirable properties, since it directly optimises the worst-case conditional Mean Squared Error (MSE) of SATE (the worst-case conditional bias is a special case).

Additionally, such weighting methods can allow us to achieve balance on non-linear transformations of the covariates by postulating a representation for the potential outcomes.

In this chapter we make use of KOM to get the estimates of the treatment effect within the subgroups. In particular, we replace the biased unadjusted estimator with the weighting estimator:

$$\widehat{\text{SATE}}_S^w = \frac{\sum_i S(\boldsymbol{x}_i)\mathbb{I}(t_i = 1)w_i y_i}{\sum_i S(\boldsymbol{x}_i)\mathbb{I}(t_i = 1)w_i} - \frac{\sum_i S(\boldsymbol{x}_i)\mathbb{I}(t_i = 0)w_i y_i}{\sum_i S(\boldsymbol{x}_i)\mathbb{I}(t_i = 0)w_i} \tag{5.5}$$

When $w_i = 1/\hat{e}(\boldsymbol{x}_i)$ for $t_i = 1$ and $w_i = 1/(1 - \hat{e}(\boldsymbol{x}_i))$ for $t_i = 0$ we get $\widehat{\text{SATE}}_S^{ipw}$ with normalised weights. We compare algorithms that use the above estimator with standard IT. As we will see the methods that use weighting estimators can successfully identify subgroups in the presence of confounders, where standard IT may fail. Additionally, the method that adopts KOM does not require training a parametric model for the outcome or the treatment. Let us now describe the approach in more detail.

5.2 Subgroup Identification via Recursive Partitioning

In this section we describe the recursive partitioning procedure followed by IT (Su et al., 2009) and its extensions (Steingrimsson and Yang, 2019; Yang et al., 2021). At the beginning all observations $\mathcal{D} = \{\boldsymbol{x}_i, t_i, y_i\}_{i=1}^n$ are in a single node, the root node. Then for a continuous covariate X_j and each value of the covariate c we split the sample into two sub-samples – in one holds $X_j \leq c$ and the other $X_j > c$. For a categorical covariate the splitting point is defined for each possible combination of values. The optimal split is selected by maximising a criterion, which is a function of the sample average treatment effects in the two sub-samples $\widehat{\text{SATE}}_L$ and $\widehat{\text{SATE}}_R$. In order to maximise the treatment effect contrast, the squared standardised difference between the treatment effects in the two samples is adopted, defined as follows:

$$G = \Big(\frac{\widehat{\text{SATE}}_L - \widehat{\text{SATE}}_R}{\sqrt{\widehat{\text{var}}_L + \widehat{\text{var}}_R}}\Big)^2$$

When the treatment effects are estimated using $\widehat{\text{SATE}}_S^{unadj}$ normalised with the pooled variance then we get the criterion of Su et al. (2009). When the treatment effects are estimated using $\widehat{\text{SATE}}_S^{ipw}$ then we get the criterion of Yang et al. (2021) with a modified denominator. The authors suggest estimators of the variance as well as DR and regression-based estimators for the numerator, but in this chapter we will focus on weighting. The covariate that maximises the above splits the current node, referred to as parent node, into two nodes, which will be referred to as child nodes. This is repeated for each new node until some pre-defined termination criteria are met. Here the termination criteria are the depth of the tree and the size of the node, which also controls the minimum size of a subgroup. The final nodes which cannot be split further are referred to as terminal or leaf nodes.

Once an initial tree has been derived, the second step is the *pruning* of the tree. This step creates a sequence of smaller trees by removing the weaker nodes. It is commonly performed in order to reduce computational complexity and memory requirements of large trees as well as to control overfitting. Each sub-tree T_i created by removing a node is assigned a score:

$$G_\rho(T_i) = \sum_{j \in \mathcal{I}_{T_i}} G(T_j) - \rho |\mathcal{I}_{T_i}| \tag{5.6}$$

where $\mathcal{I}_{T_i}$ is the set of internal nodes, i.e. all nodes without the terminal ones and $|\mathcal{I}_{T_i}|$ is the cardinality of this set, i.e. number of internal nodes in the tree T_i. The first term captures the overall score of the tree, while the second term captures the complexity measured as the number of splits. The parameter ρ controls the importance of the complexity of the tree. Following the procedure described in (Su et al., 2009) at each iteration a sub-tree is created by removing the node that minimises $\sum_{j \in \mathcal{I}_{T_i}} G(T_j)/|\mathcal{I}_{T_i}|$ until the remaining tree consists of only the root node. This step creates a sequence of trees of varying size. The last step is the selection of the final tree, which can be done using various approaches. When the size of the dataset is large enough we can split the data into a training set and a validation set (Su et al., 2009; Steingrimsson and Yang, 2019). The training set is used to fit the initial tree and create the sequence that results from pruning, while the validation set is used to select the final tree that has the maximum score $G_\rho(T_i)$.

In small sample datasets the final tree can be selected using the bootstrap

method described in (Su et al., 2009, 2008; Dusseldorp and Van Mechelen, 2014). This method corrects the bias of the estimated value of the splitting criterion by re-evaluating it using multiple bootstrap samples. First the original data are used to perform both the construction of the tree and estimate the splitting criterion for each sub-tree $G^{init}(T_i)$. Then B bootstrap samples are drawn. For each bootstrap sample b a tree is grown and the value of the splitting criterion is computed $G_b^{boot}(T_i)$. The splitting criterion is re-evaluated in the original sample $G_b^{orig}(T_i)$. The final value of the criterion is $G^{init}(T_i) - \frac{1}{B}\sum_{b=1}^{B}\Big(G_b^{boot}(T_i) - G_b^{orig}(T_i)\Big)$. For the details of the procedure followed by IT we refer the reader to (Su et al., 2009, 2008; Calhoun et al., 2018). In each repetition b the original data are used as the external dataset. The motivation is to correct the initial optimistic value by subtracting the difference between evaluating the criterion in the same data used for learning the structure and using different data for learning the structure and evaluating the criterion (Su et al., 2008). The final tree can be selected using $G_\rho(\cdot)$ for some value of ρ as described previously, or using the one-standard-error rule as described in (Dusseldorp and Van Mechelen, 2014). In the latter case we may choose the smallest tree that has a bias-corrected value of the splitting criterion within 1 standard error (1-SE) of the maximal value.

5.3 Estimation of the Splitting Criterion

We now describe the modifications to IT and in particular how to estimate $\widehat{\text{SATE}}_S$ and $\widehat{\text{var}}_S$ for each subgroup. The estimation of $\widehat{\text{SATE}}_S$ is done via weighting using Kernel Optimal Matching (KOM) (Kallus et al., 2021). In particular, the authors represent the conditional expectations of the outcomes in a Reproducing Kernel Hilbert Space (RKHS) and derive an approach that optimises the worst-case conditional MSE (conditioned on the covariates and treatment). Here we will also consider the approach that targets the worst-case conditional bias. We will describe these two approaches along with the other components required for implementing the subgroup identification algorithms.

Targeting the Bias

The treatment effect within each subgroup S is estimated using eq. (5.5), where if we normalise the weights to sum to one within each treatment group we can

equivalently express it as:

$$\widehat{\text{SATE}}_S^w = \sum_{i=1}^{n_s} \big(\mathbb{I}(T_i = 1) w_i y_i(1) - \mathbb{I}(T_i = 0) w_i y_i(0)\big) \tag{5.7}$$

Let us define as $k(\cdot, \cdot) : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ a positive semi-definite symmetric kernel and $\phi(\boldsymbol{x}) : \mathcal{X} \to \mathcal{H}$ a feature map. We also consider that for such a kernel function $k(\cdot, \cdot)$ we can find a feature mapping $\phi(\cdot)$ that transforms $\boldsymbol{x}_i$ to a new space where for any $\boldsymbol{x}_j$ we have $k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \langle \phi(\boldsymbol{x}_i), \phi(\boldsymbol{x}_j) \rangle$. Let $\boldsymbol{w}(z)$ denote the weights for the observations that have $t = z$. In the first version of the subgroup identification algorithm these are chosen solving the following optimisation problem:

$$\begin{aligned} &\underset{\boldsymbol{w}(z)}{\arg\min} \;\; \boldsymbol{w}^T(z) K_z^{zz} \boldsymbol{w}(z) - \frac{2}{n_s} \mathbf{1}^T K_z^{\cdot z} \boldsymbol{w}(z) \\ &\text{s.t.} \sum_{i:t_i=z} w_i(z) = 1, \;\; \boldsymbol{w}(z) \succeq 0 \end{aligned} \tag{5.8}$$

where $\boldsymbol{w}(z) \succeq 0$ indicates that all entries in the vector of weights are non-negative, K_z is a $n_s \times n_s$ kernel matrix associated with the potential outcome $Y(z)$ and $\mathbf{1}$ denotes a vector of ones. Here K_z^{zz} corresponds to the entries of the matrix for the group $T = z$ and $K_z^{\cdot z}$ has all rows of the initial matrix and the columns for which $T = z$. Let us denote the above objective function as $\mathcal{J}_{Y(z)}$. This problem can be motivated by minimising the worst-case squared conditional bias of each potential outcome (Kallus et al., 2021) and we will refer to the recursive partitioning method that uses these weights as Bias reducing IT (B-IT).

We can alternatively approach the above by assuming the potential outcomes can be expressed as linear functions in the feature space denoted by $\phi(\cdot)$ as in (Hazlett, 2020). In particular, we can write the potential outcomes as $y_i(z) = m_z(\boldsymbol{x}_i) + \epsilon_z(\boldsymbol{x}_i) = \alpha_z^T \phi(\boldsymbol{x}_i) + \epsilon_z(\boldsymbol{x}_i)$, where $\epsilon_z(\boldsymbol{x}_i)$ is a zero-mean error term. Then we can express the optimisation problem so that the following balancing condition holds within the subgroup:

$$\sum_{i=1}^{n_s} \mathbb{I}(t_i = z) w_i \phi(\boldsymbol{x}_i) = \frac{1}{n_s} \sum_{i=1}^{n_s} \phi(\boldsymbol{x}_i) \tag{5.9}$$

for $z \in \{0, 1\}$ and all weights w_i are positive and sum to one within each treatment group. The relationship between balance maximisation and bias reduction was also discussed in Chapter 2 and described for SATT estimation using kernels in

(Hazlett, 2020). For completeness of the presentation let us describe this in more detail. We know that $\text{SATE}_{\mathcal{S}}$ in the subgroup can be expressed as:

$$\text{SATE}_{\mathcal{S}} = \frac{1}{n_s}\sum_{i=1}^{n_s}\big(\mathbb{E}[Y(1) \mid \boldsymbol{x}_i] - \mathbb{E}[Y(0) \mid \boldsymbol{x}_i]\big) = \frac{1}{n_s}\sum_{i=1}^{n_s}\big(m_1(\boldsymbol{x}_i) - m_0(\boldsymbol{x}_i)\big)$$

Taking the difference of the above and its estimate given in eq. (5.7) and replacing $y_i(z)$ with $\alpha_z^T\phi(\boldsymbol{x}_i)+\epsilon_z(\boldsymbol{x}_i)$ results in three terms: $\widehat{\text{SATE}}_{\mathcal{S}}^{w} - \text{SATE}_{\mathcal{S}} = \mathcal{B}_1 - \mathcal{B}_0 + \mathcal{E}$. The expected values of these conditioned on the covariates and treatment are the conditional biases and error.

The error term $\sum_{i=1}^{n_s}\mathbb{I}(T_i = 1)w_i\epsilon_1(\boldsymbol{x}_i) - \sum_{i=1}^{n_s}\mathbb{I}(T_i = 0)w_i\epsilon_0(\boldsymbol{x}_i)$ will be on expectation equal to zero and therefore we can focus on minimising the two terms $\mathcal{B}_1$ and $\mathcal{B}_0$, which will result in the aforementioned balancing conditions since for $\mathcal{B}_z$ we have:

$$\sum_{i=1}^{n_s}\big(\mathbb{I}(T_i = z)w_i - \frac{1}{n_s}\big)\alpha_z^T\phi(\boldsymbol{x}_i) = 0$$

Alternatively, the absolute value of the above can be upper-bounded by the product of two $L2$-norms as described in Chapter 2. By re-arranging eq. (5.9) and taking the squared $L2$-norm results in:

$$\boldsymbol{w}^T(z)K_{zz}\boldsymbol{w}(z) - \frac{2}{n_s}\mathbf{1}^T K_{\cdot z}\boldsymbol{w}(z) + \frac{1}{n_s^2}\mathbf{1}^T K\mathbf{1}$$

which corresponds to the optimisation problem of eq. (5.8), after omitting the last term that is independent of $\boldsymbol{w}(\cdot)$ and assuming the potential outcomes can be expressed as linear functions in the same feature space. Hence, the resulting optimisation problem can be solved so that balance is achieved in the new space defined by the mapping $\phi(\cdot)$.

In the simplest case we can assume a linear model on the initial space of pre-treatment covariates in which case we will be matching the means of the treatment groups. More complex functions allow for matching higher moments, while the use of a kernel allows for adopting feature spaces where the number of dimensions might not be finite. Some examples of commonly used kernels are the RBF, $k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp(-\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^2/2l^2)$ and the polynomial $k(\boldsymbol{x}_i, \boldsymbol{x}_j) = (\boldsymbol{x}_i^T\boldsymbol{x}_j + l^2)^d$. In this chapter we will use the latter with $d = 1, 2$ in order to either match the means of the covariates or additionally interactions between those. It should be noted that a potential problem with the above approach is

that no restrictions are imposed on the variability of the weights. Controlling the variance of weights was suggested by Zubizarreta (2015) who proposed minimising the squared distance of the weights from their means. Here we will instead focus on using an approach that adds a regularisation term for the weights which can be associated with controlling the variance of the estimated effect. Before describing the second method we will adopt in this chapter let us give an example of how the aforementioned method optimises the balance.

The properties of the described approach over using IPW become particularly important in the case of subgroup identification, where treatment effects need to be estimated using small sample sizes as we keep partitioning the space but also in the presence of strong confounding. In order to show this consider the following example. The outcome follows: $Y = X_1 + X_2 + 2T\mathbb{I}(X_1 > 0) + \mathcal{N}(0, 1)$ and the treatment assignment model is: $\text{logit}(p(T = 1 \mid \boldsymbol{X})) = \gamma(X_1 + X_2)$ where the covariates are independent and follow the standard normal distribution. Suppose we run the recursive partitioning algorithm to identify the subgroup in which case we will need to estimate the treatment effect within subsets of the data. We use IPW and weights that optimise eq. (5.8) with a linear kernel. For IPW the propensity score is estimated using a correctly specified model and applying weight truncation using the 1st and 99th percentile of their distribution (Lee et al., 2011; Cole and Hernán, 2008). In figure 5.1 we report the absolute standardised mean difference between the treatment groups for the confounder X_2 estimated within the subgroup averaged over 500 runs. In this figure, we plot the cases of $\gamma = 0.5$ (dashed line) and $\gamma = 1$ (solid line). We vary the sample size from 50 to 500 observations and in the horizontal axis we report the approximate subgroup size which will correspond to $\sim 50\%$ of the observations. We observe that in the initial data there is imbalance between the treatment groups and the absolute standardised mean difference is approximately 0.5 for $\gamma = 0.5$ and 0.9 for $\gamma = 1$. Using IPW results in better balance which is improved as we increase the sample size. The weights that optimise the balance within the subgroup will result in an absolute standardised mean difference close to zero even with a few observations.

Targeting the Variance and Augmented Estimators

Recent works focusing additionally on the variance of the effect showed that this can be controlled by the squared norm of the weights (Kallus, 2020b; Kuang et al., 2019; Kallus et al., 2021). In this study we will adopt a version of KOM

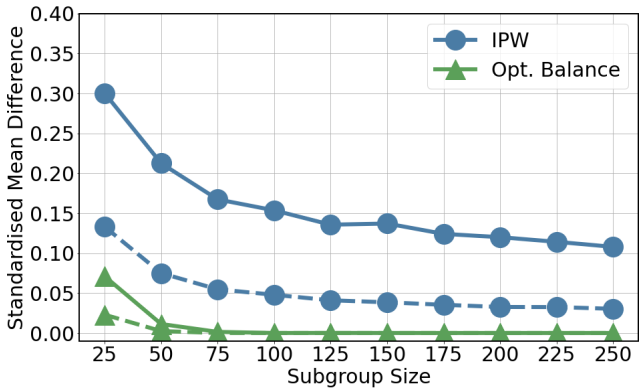


Figure 5.1: When performing subgroup identification with IT we may encounter small sample sizes as we keep partitioning the space. IPW with a correctly specified model will balance the covariates better as the sample size increases and this can be further improved by optimising it directly (Opt. Balance). Here the solid line corresponds to $\gamma = 1$ and the dashed line to $\gamma = 0.5$.

(Kallus et al., 2021) by assuming constant conditional variance of the outcomes under each value of the treatment. Their problem is more general and allows for including the conditional variance for each observation. The objective $\mathcal{J}_{Y(z)}$ is re-written as a function of $\boldsymbol{w}$: $\mathcal{J}_{Y(z)} = \boldsymbol{w}^T I_z K_z I_z \boldsymbol{w} - \frac{2}{n_s} \mathbf{1}^T K_z I_z \boldsymbol{w}$, where I_z is a $n_s \times n_s$ diagonal matrix with one where $t_i = z$ and zero otherwise. Then, under the assumption of constant variance, the optimisation problem can be written as:

$$\begin{aligned} &\arg\min_{\boldsymbol{w}} \ \mathcal{J}_{Y(0)} + \mathcal{J}_{Y(1)} + \lambda_0 \sum_{i:t_i=0} w_i^2 + \lambda_1 \sum_{i:t_i=1} w_i^2 \\ &\text{s.t.} \ \sum_{i:t_i=0} w_i = \sum_{i:t_i=1} w_i = 1, \ \boldsymbol{w} \succeq 0 \end{aligned} \tag{5.10}$$

This approach is motivated by optimising the (worst-case) conditional MSE of SATE when λ_t are the conditional variances of the outcomes under each treatment group and the conditional expectations of the outcomes are represented in a RKHS (Kallus et al., 2021). Additionally the authors discuss an approach for selecting the parameters of the used kernels and the variances by fitting a Gaussian Process (GP) (Williams and Rasmussen, 2006) for the outcome under each treatment arm. Here we will also follow this approach.

We will refer to the method that uses IT and applies the aforementioned weights as MSE-IT. The values λ_z can also be seen as parameters in which case as the value of λ_z increases more uniform weights are achieved and when they are equal to zero MSE-IT becomes B-IT with the weights optimised jointly. We note

that we will focus on the absolute error of the estimated treatment effect within the subgroup and not the MSE, hence the use of both MSE-IT and B-IT in our setting is motivated so that we can explore whether targeting the variance can provide different results in terms of identifying the correct trees under different values of the confounding strength.

Since combinations of weighting and regression adjustment have shown good empirical performance (Athey et al., 2018; Kuang et al., 2019; Kallus, 2020b), we shall explore these using the predicted outcomes of models fitted for each treatment group. These are often described as augmented estimators in the literature (Kang et al., 2007; Kallus, 2020b) and we will follow this terminology. In this case the predicted treatment effect is given by:

$$\begin{aligned}\widehat{\text{SATE}}_S^{aug} = \overline{\hat{m}_1} + \sum_{i:\boldsymbol{x}_i\in\mathcal{S},t_i=1} w_i(1)(y_i - \hat{m}_1(\boldsymbol{x}_i)) \\ - \overline{\hat{m}_0} - \sum_{i:\boldsymbol{x}_i\in\mathcal{S},t_i=0} w_i(0)(y_i - \hat{m}_0(\boldsymbol{x}_i))\end{aligned} \tag{5.11}$$

where $\hat{m}_1(\boldsymbol{x}_i), \hat{m}_0(\boldsymbol{x}_i)$ are the estimated potential outcomes and $\overline{\hat{m}_1}, \overline{\hat{m}_0}$ their means.

Such augmented estimators have been studied with the use of IPW weights in (Kang et al., 2007) where their DR properties are discussed and with the use of the aforementioned weights in (Kallus, 2020b). It has been observed empirically that when using IPW weights and both models are incorrect then the augmented estimator can perform worse than weighting alone (Kang et al., 2007). However, under correct specification they can show better efficiency (Kang et al., 2007). In this chapter we will focus mainly on weighting and we will present some results using augmented estimators in Section 5.4.1 for completeness.

Balancing Considerations

So far we have considered that the weights are going to be estimated within each candidate subgroup. However, creating new weights for each possible split may be computationally impractical especially if the covariates are continuous and/or we face a high-dimensional problem. Similarly to Yang et al. (2021) we consider approaches of increasing complexity. Firstly, we derive the weights once in the root node. In other words we identify weights that optimise the quantity

of interest in the initial dataset and then apply recursive partitioning on the re-weighted dataset. As we partition the data, this will not guarantee balance (or optimisation of the corresponding quantity) on each resulting sub-sample, but it may act as an approximation. The second approach estimates new weights in each parent node. Lastly we consider fitting new weights for each candidate split. This however can be computationally demanding in the presence of a large number of covariates with a large number of distinct values. In this case we may use only a subset of the possible splits of a continuous covariate to derive the balancing weights. More specifically, if the covariate is continuous with unique values more than a specified number we treat it as such for the purposes of estimation of the optimal split but the weights are fitted on only a number of equally distributed values of the covariate.

Notice that by fitting the weights once in the root node and using them for the rest of the analysis we implicitly assume that the functional form of the potential outcomes assumed in the root node will also hold in subsets of the data. Moreover, the derived weights that optimise the corresponding quantity (bias, MSE) in the full sample, will also optimise this quantity in subsets. As we will see in certain scenarios this simple and computationally efficient approach may provide good empirical results, but we need to stress that the used weights may not be the optimal. The other two approaches relax this assumption and may consider a different form of the outcome in each parent node or each possible child node. Overall, we would suggest finding new weights for each possible split (or at least a subset of the possible splits). If the other two approaches are considered, e.g. due to limited computational resources, the resulting balance may need to be assessed.

Subgroup-Specific Variance

In order to normalise the estimated effects we require an estimate of the variance. There is a long literature regarding the estimation of standard errors or confidence intervals for weighting estimators (Little and Rubin, 2002, Chapter 3; Kallus, 2020b; Athey et al., 2018; Abadie and Imbens, 2006; Imbens, 2004). A general approach for deriving the variance of the sample treatment effect conditioned on $\boldsymbol{X}, T$ is described in (Imbens and Rubin, 2015, Chapter 19). For a weighting

Method	Estimator	Weighting Algorithm
IT (Su et al., 2009)	eq. (5.3)	-
CIT-IPW (Yang et al., 2021)	eq. (5.4)	propensity model
B-IT	eq. (5.7)	eq. (5.8)
MSE-IT	eq. (5.7)	eq. (5.10)
B-IT (augmented)	eq. (5.11)	eq. (5.8)
MSE-IT (augmented)	eq. (5.11)	eq. (5.10)

Table 5.1: Summary of IT-based methods. The estimator refers to the equation used to estimate the treatment effect within a potential subgroup and the weighting algorithm refers to the approach used to estimate the weights where this is applicable.

estimator with normalised weights this is given by:

$$\widehat{\text{var}}_S = \sum_{i:\boldsymbol{x}_i \in \mathcal{S}, t_i=1} w_i^2 \hat{\sigma}_i^2(1) + \sum_{i:\boldsymbol{x}_i \in \mathcal{S}, t_i=0} w_i^2 \hat{\sigma}_i^2(0) \tag{5.12}$$

For the estimation of $\hat{\sigma}_i(t)$, Imbens and Rubin (2015) (see also (Abadie and Imbens, 2006)) propose a nearest neighbour estimator where $\hat{\sigma}_i^2(t)$ is approximated by the mean squared difference between the outcome of the i-th example and the outcome of the nearest neighbour in the group with $T = t$. An alternative approach that we adopt in this chapter when using the augmented estimator is to use the predicted outcomes to estimate the residuals. In this case, $\hat{\sigma}_i^2(t) = (y_i(t) - \hat{m}_t(\boldsymbol{x}_i))^2$, where $\hat{m}_t(\boldsymbol{x}_i)$ is the estimated value of the potential outcome. This is in similar spirit with the variance estimator described in (Athey et al., 2018) for linear models. Another option which is adopted in our case studies when using the weighting estimator is to perform weighted least squares with the estimated weights and use the robust sandwich estimator (Freedman, 2006; Zeileis, 2004, 2006).

5.4 Simulated Data

In this section we compare IT with unadjusted treatment effect estimation against B-IT and MSE-IT. Additionally, we will consider IT with IPW weighting. The methods that are used in this chapter are summarised in table 5.1. For their

implementation we modified the splitting criterion of IT[1] (Su et al., 2009). We will use the following evaluation metrics (Loh et al., 2019; Yang et al., 2021).

- Proportion of Correct Trees (PCT): The percentage of simulations where the resulting tree has the splits that define the true subgroup and only those.

- Absolute Error (AE) expressed as the difference between the estimated and expected treatment effect within the subgroup $|\widehat{\text{SATE}}_s - \text{ATE}_s|$.

- Mean Squared Error (MSE) of the estimated treatment effect for each observation evaluated in a separate test set, $\text{MSE} = \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} \big((\hat{y}_i(1) - \hat{y}_i(0)) - (y_i(1) - y_i(0))\big)^2$. Here $\hat{y}_i(1), \hat{y}_i(0)$ are the estimated potential outcomes if we were to use our algorithms for estimating the individual treatment effect $y_i(1) - y_i(0)$. For a given observation this is estimated as the weighted average within the leaf of the tree where this observation belongs.

- True Positive Rate (TPR): Let $\mathcal{P}$ be the covariates that define the splits and $\mathcal{V}$ be the covariates that are used by the subgroup identification algorithm to define the resulting subgroups. The TPR is the ratio $\frac{|\mathcal{P} \cap \mathcal{V}|}{|\mathcal{P}|}$ averaged over the number of simulations.

- Number of False Discoveries (NFD): The number of non-predictive covariates wrongly used to define subgroups averaged over the number of simulations.

5.4.1 Varying the Confounding Strength

Firstly we test IT with unadjusted estimator and the weighting versions in the presence of confounders. We simulate data using the following outcome model from Foster et al. (2011), but using a continuous outcome instead of binary.

$$Y = \boldsymbol{\alpha}^T(1, X_1, X_2, X_3, X_2X_3) + T\boldsymbol{\beta}^T(1, \mathbb{1}(X_1 > 0 \cap X_2 < 0) + \epsilon$$

[1] An implementation of IT can be found on the Biopharmaceutical Network web site at: http://biopharmnet.com/subgroup-analysis/ [last accessed: 17/12/2020]

where $\boldsymbol{\alpha} = (-1, 0.5, 0.5, -0.5, 0.5)^T$, $\boldsymbol{\beta} = (0.1, 0.9)^T$ and $\epsilon \sim \mathcal{N}(0, 0.01)$. The covariates are normally distributed with zero mean and unit variance, while odd-numbered covariates $\{X_1, X_3, X_5, ...\}$ have an internal correlation of 0.7, and even-numbered, $\{X_2, X_4, X_6, ...\}$, have the same internal correlation. In this dataset there is one subgroup that corresponds to approximately 25% of the sample size where we have $\text{ATE}_S = \mathbb{E}[Y(1) \mid \boldsymbol{x} \in \mathcal{S}] - \mathbb{E}[Y(0) \mid \boldsymbol{x} \in \mathcal{S}] = 1$. In order to introduce imbalance between the covariates in the two treatment groups, the treatment assignment is generated according to the following model:

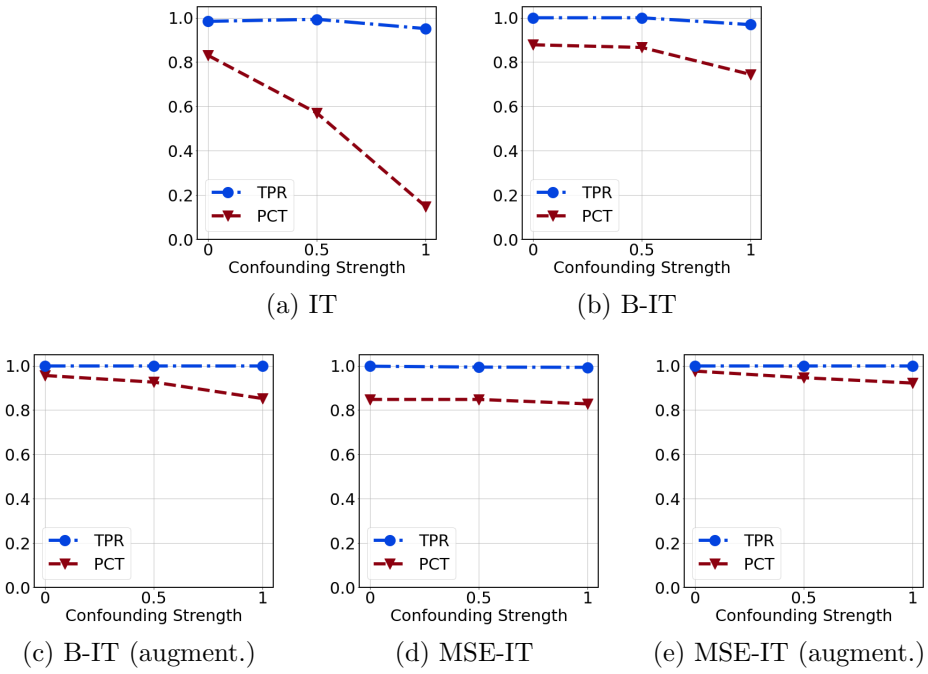
$$\text{logit}(p(T = 1 \mid \boldsymbol{X})) = \gamma(X_1 + X_2 - X_3) \tag{5.13}$$

where γ is the confounding strength. We consider three cases: no confounding, $\gamma = 0$, which corresponds to a 1:1 randomised study, and two cases of increased confounding $\gamma = 0.5$ and $\gamma = 1$.

In this section we consider the simplest implementation of the proposed methods optimising the weights once using the full data. We choose to use a linear kernel (i.e. matching the means of the covariates) for B-IT and MSE-IT. We additionally consider B-IT and MSE-IT combined with the augmented estimators. Notice here that the linear kernel corresponds to miss-specification of the outcome due the the interaction between X_2 and X_3 but also the presence of the indicator function. Nevertheless, initial results showed that assuming linearity in this case results in treatment effects close to the ground truth. We create a sample of 2000 observations and 6 covariates and use 1000 for training, 500 for tree selection and 500 for estimating the MSE. From the 6 covariates, 1 is solely prognostic, 2 are both prognostic and predictive and 3 are irrelevant. All results reported in this section have been averaged over 500 realisations of the treatment assignment and outcome functions. The initial tree was grown with maximum depth equal to 15, minimum size required for splitting a node equal to 50 observations and minimum size of a terminal node (i.e. minimum size of a subgroup) equal to 20 observations. The final tree was selected as the one that maximises $G_\rho(\cdot)$ estimated in a separate validation set (Su et al., 2009). The value of ρ was chosen so that it corresponds to the 0.05 significance level on the χ^2 distribution with 1 degree of freedom as suggested in previous works (Su et al., 2009; Yang et al., 2021).

In figure 5.2 we firstly observe that all approaches have high TPR, hence using the true predictive covariates to split the data. Regarding PCT, when the

treatment assignment is marginally randomised all approaches perform similarly. In the scenario studied in this section the PCT is defined as the number of times the depth of the final tree is 2 and the data are split using only X_1 and X_2. As we increase the confounding strength we observe that IT has lower PCT. B-IT is also influenced by the confounding strength but to a lower degree compared to IT. On the contrary, MSE-IT that controls the variance of the estimated treatment effect is not influenced by the increased confounding strength resulting in almost the same results as if the treatment assignment was marginally randomised. Additionally, we observe that combining weighting with regression adjustment can further improve the results, particularly in the challenging case of $\gamma = 1$.



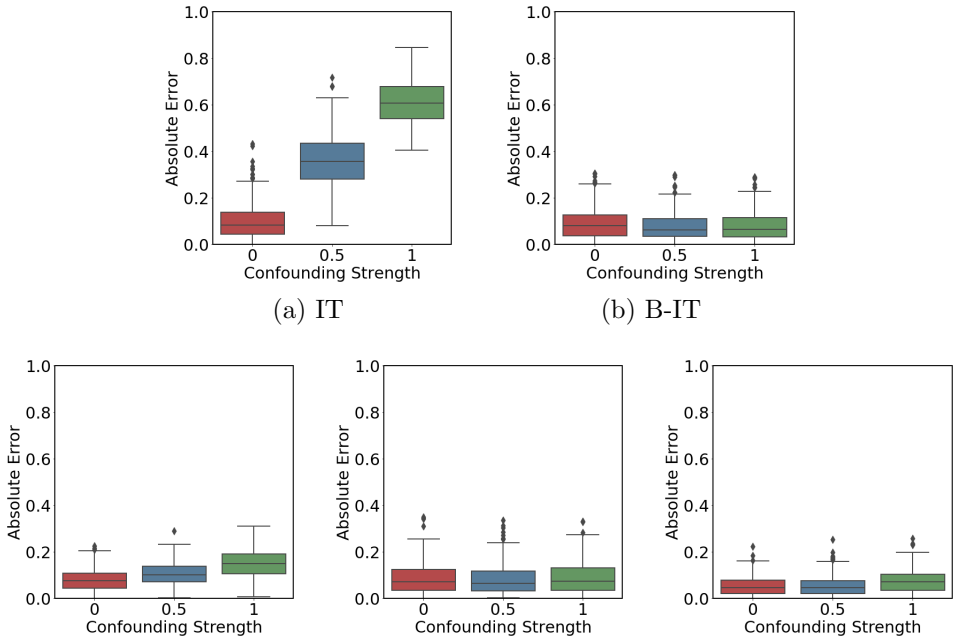
(a) IT (b) B-IT

(c) B-IT (augment.) (d) MSE-IT (e) MSE-IT (augment.)

Figure 5.2: Proportion of correct trees and True Positive Rate for various values of the confounding strength.

In figure 5.3 we show the absolute error of the estimated treatment effect within the subgroup. The box-plots show the distribution of the error over the number of simulations where the correct tree was identified. Here, we observe that B/MSE-IT and the variants that use regression adjustment tend to estimate treatment effects that are more concentrated towards the ground truth. Interestingly, even in the case of $\gamma = 0$ B/MSE-IT with regression adjustment tend to achieve errors close to zero for more simulations compared to IT with unadjusted estimator. As expected IT with unadjusted estimator does not estimate

the treatment effect correctly when $\gamma > 0$.



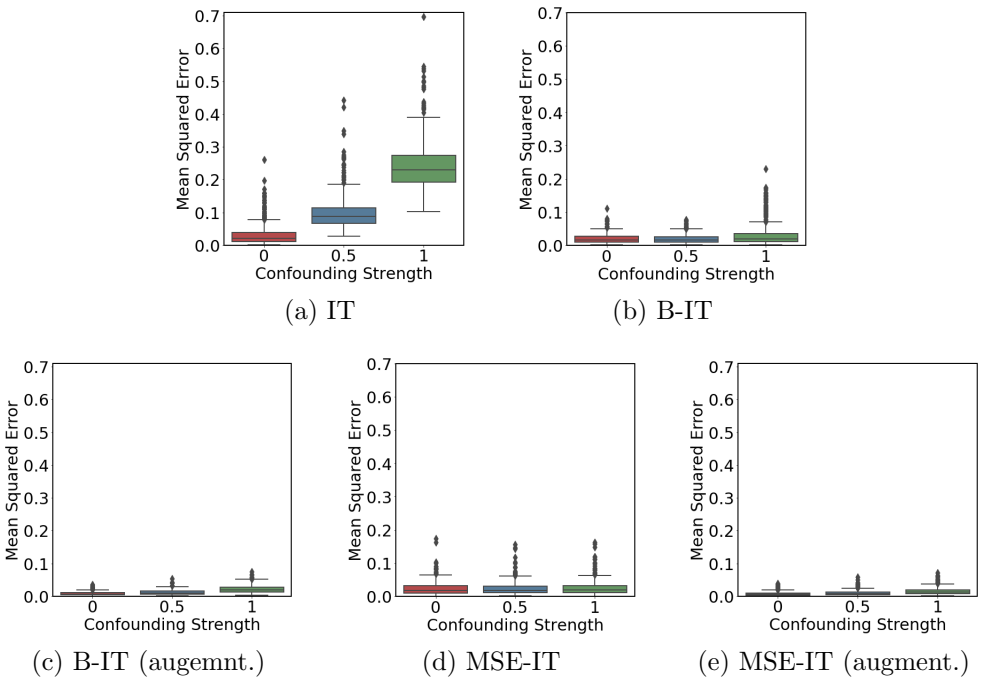
(a) IT (b) B-IT

(c) B-IT (augment.) (d) MSE-IT (e) MSE-IT (augment.)

Figure 5.3: The absolute error of the estimated treatment effect within the subgroup for various values of the confounding strength.

Even though our primary task is to identify subgroups an interesting question is how these methods would perform if they were to be used to perform inference – i.e. estimate the individual treatment effect for a given observation. In figure 5.4 we report the MSE in a separate test set where we observe that B/MSE-IT and the augmented versions tend to achieve better results, particularly in the case with $\gamma = 1$ where the problem becomes more challenging.

We repeat this experiment using larger subgroup size as described in Foster et al. (2011). The results can be found in Appendix A.2, where we find that the augmented estimators tend to provide worse results in terms of PCT than using weighting alone. Besides miss-specification, the definition of PCT may also be a reason for the observed difference. Based on this section we can validate that IT with unadjusted estimator is not suited in the presence of confounders while the use of weighting estimators may improve the increased bias resulting in an overall performance similar to as if the data were from a marginally randomised study. In the next section we will explore how these approaches compare to using IT with IPW estimators.



(a) IT (b) B-IT

(c) B-IT (augemnt.) (d) MSE-IT (e) MSE-IT (augment.)

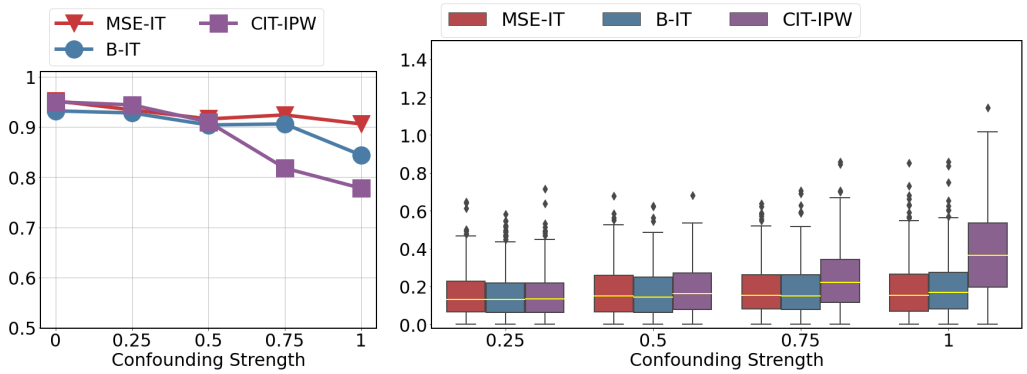
Figure 5.4: The mean squared error of the estimated treatment effect in a separate test set for various values of the confounding strength.

5.4.2 Comparison with Recursive Partitioning using IPW

In order to show the properties of B/MSE-IT compared to IT with an IPW estimator we will consider a simple example. We assume the outcome follows $Y = X_1 + X_2 + X_3 + 2T\mathbb{I}(X_1 > 0) + \epsilon$ and the propensity score is $\text{logit}(p(T = 1 \mid \boldsymbol{X})) = \gamma(X_1 + X_2 + X_3)$. The covariates take values in $\{-1.5, -1.25, ..., 1.25, 1.5\}$ and $\epsilon \sim \mathcal{N}(0, 1)$. We simulate a dataset with 500 observations and 6 covariates and consider the proportion of times (out of 500 repetitions) each method splits the data correctly. For this simulation we consider only trees of depth 1. We compare B-IT and MSE-IT, with CIT using IPW (CIT-IPW). We fit the weights in the node that is considered for splitting instead of every possible split, which showed better results in (Yang et al., 2021).

In figure 5.5 we report the PCT, defined here as the proportion of trees of depth 1 that split the data correctly and the absolute error of the treatment effect in the subgroup. We observe that the subgroup identification algorithms with non-parametric weighting methods tend to identify the correct split more often and provide more accurate estimates of the treatment effect within the subgroup for larger values of the confounding strength. For values $\gamma \leq 0.5$ all

methods perform similarly. We need to highlight that by fitting the weights in the root node and not in each possible split this is going to guarantee optimisation of the corresponding quantity only in the initial sample. We additionally find that for the largest value of the confounding strength MSE-IT tends to achieve better PCT compared to B-IT, which was also observed in the previous section. For the rest of this chapter we will focus on MSE-IT.



(a) PCT (b) Absolute Error

Figure 5.5: For small values of the confounding strength all approaches perform similarly. Both B-IT and MSE-IT find the correct split more often than using IPW estimators for larger values of the confounding strength. They also provide more accurate estimates of the treatment effect within the subgroup in this case.

5.4.3 Varying the Outcome Specification

In this section we will explore what happens under miss-specification of the outcome model. For this purpose we generate data using the following model:

$$Y = \boldsymbol{\alpha}^T(1, X_1X_2, X_3^2) + T\boldsymbol{\beta}^T(1, \mathbb{1}(X_3 > 0)) + \epsilon$$

where $\boldsymbol{\alpha} = (1, 2, 1)^T$, $\boldsymbol{\beta} = (1, 2)^T$ and $\epsilon \sim \mathcal{N}(0, 1)$. The treatment assignment depends on the covariates according to:

$$\text{logit}(p(T = 1 \mid \boldsymbol{X})) = \gamma(X_1X_2 + X_3)$$

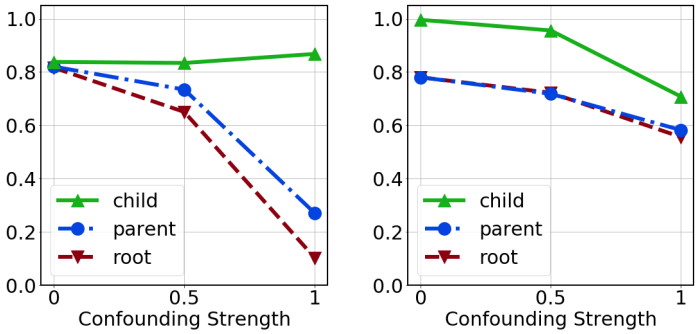
The pre-treatment covariates are independent and their values are randomly drawn from the domain $\{-1.5, -1, -0.5, 0, 0.5, 1, 1.5\}$. We create a sample of 1000 observations and 6 covariates and use 500 for training and 500 for tree

selection. The results reported in this section have been averaged over 500 realisations of the treatment assignment and outcome functions. This scenario is more challenging compared to the one we studied in Section 5.4.1. Here both the outcome and the treatment assignment depend on interactions between the covariates while we also have larger main effects, noise with larger variance and we use a smaller sample size.

We first apply IT with unadjusted estimator and we find that the PCT is 0.80 for $\gamma = 0$, 0.54 for $\gamma = 0.5$ and 0.07 for $\gamma = 1$. In this section we will focus on MSE-IT and explore the results under two model specifications. In the first setting we assume the outcome is a linear function of the covariates, therefore we simply match the means of the covariates. In the second setting we use a second-degree polynomial kernel. For the recursive partitioning method we use the same parameters as in the previous section.

The PCT is estimated as the proportion of resulting trees that have a single split on X_3 with cut-off 0. This is shown in figure 5.6. Under linear specification fitting the weights once in the root node or in each parent node is heavily influenced by the confounding strength. Interestingly, when fitting the weights for each possible split we find that this results in high PCT. However, this should be interpreted with caution since it also results in large errors on treatment effect estimation as shown in figure 5.7. In this case the algorithm becomes more conservative, finding trees with only X_3 as the splitting covariate more often, but still introducing a significant error when estimating the treatment effect. This can be witnessed also by the average number of false discoveries reported in figure 5.8. This result could occur if, for example, the method used to estimate the treatment effect tends to consistently over-estimate or under-estimate it. Interestingly, even under incorrect specification and without fitting the weights for each possible split we can still achieve higher PCT compared to using unadjusted estimators.

We repeat the experiment but this time we fit a second-degree polynomial kernel. The results are shown in figure 5.6(b) for PCT, figure 5.7(b) for the absolute error and figure 5.8(b) for the number of false discoveries. We observe that this modification improves the results, particularly when fitting the weights in the full data or each parent node. We also observe that the PCT decreases when $\gamma = 1$. In figure 5.8 we notice that in this case NFD is small. This suggests that MSE-IT may split the data using X_3 either with an incorrect cut-off point



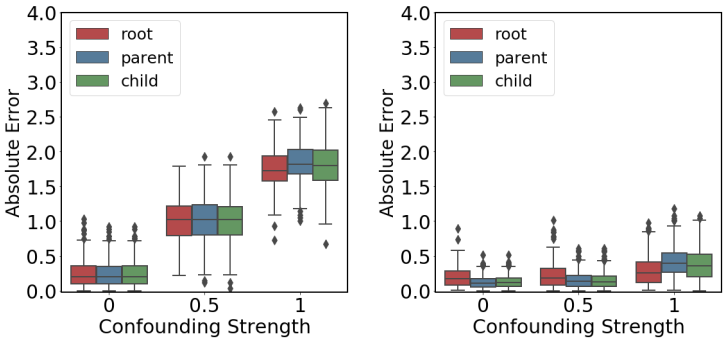
(a) $d = 1$ (b) $d = 2$

Figure 5.6: Proportion of Correct Trees (PCT) resulting by estimating the weights in the root node, parent node or each possible split of a parent node. We use MSE-IT and we either assume a linear kernel $d = 1$ or a second degree polynomial kernel $d = 2$. We notice that even when the model is not correctly specified all methods perform similarly or better than if we used the unadjusted estimator.

or by performing multiple splits on the same covariate.

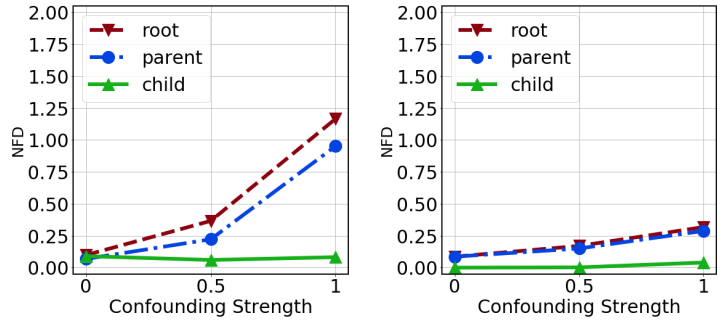
In Section 5.3 we described that when the covariates are continuous and/or take a large number of possible values, we may reduce the computational complexity by fitting new weights in a number of pre-specified cut-off values rather than each possible split. We repeat the previous experiment, but this time the covariates are normally distributed and are correlated with correlation 0.3. Instead of fitting new weights for each possible split, we consider three and five equally spaced cut-off values for each covariate and derive weights for each one of the splits that result from these cut-off values. Then while performing recursive partitioning for each possible split we use the weights of the nearest cut-off. The resulting PCT is shown in figure 5.9. A correct tree here is defined as one that has depth equal to one and splits the data using only X_3.

When using IT with unadjusted estimator we find that the PCT is 0.6 when $\gamma = 0$, 0.26 for $\gamma = 0.5$ and 0.07 for $\gamma = 1$. From the figure we observe that the aforementioned approximation results in higher PCT than IT in most cases. Similarly to our previous observations, even under incorrect specification the method may result in higher PCT compared to using the unadjusted estimator except when the weights are fitted in the root node or in each parent node and $\gamma = 1$. When the outcome accounts for higher-order terms (in this case pairwise interactions) simple approximations, such as fitting the weights once in the root node can result in high PCT. In this scenario, when the weights are fitted for three/five splits we observe a reduced PCT when $\gamma = 1$. This could be attributed to the



(a) $d = 1$ (b) $d = 2$

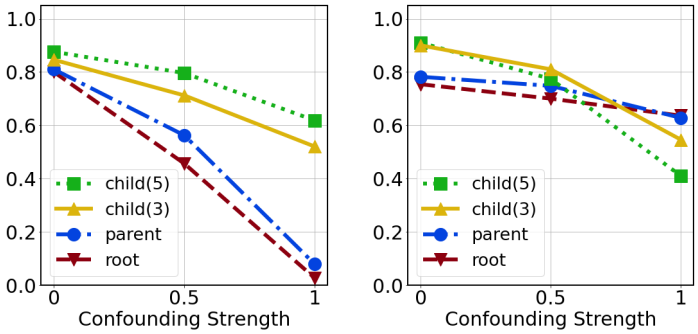
Figure 5.7: Absolute error resulting by estimating the weights in the root node, parent node or each possible split of a parent node. As the confounding strength increases so does the error, however this increase is lower under a correctly specified model in the subgroup.



(a) $d = 1$ (b) $d = 2$

Figure 5.8: Number of non-predictive covariates identified by MSE-IT and averaged over 500 realisations of the outcome and treatment assignment. This tends to be lower for smaller values of the confounding strength and and/or when using the second degree kernel.

used cut-off values for estimating the weights, but also to the definition of the metric. In particular, we find that in these cases the algorithm identifies the correct subgroup of enhanced effect as often as the other considered approaches, but it also tends to retain more splits in the final tree, resulting in a lower PCT. Note that a difference compared to the previous experiment is that here we consider a tree as correct if it uses the correct predictive covariate without taking into account the actual cut-off value. The full results including the absolute error and the number of false discoveries are reported in the Appendix A.3.



(a) $d = 1$ (b) $d = 2$

Figure 5.9: Proportion of Correct Trees (PCT) using normally distributed covariates.

5.4.4 Homogeneous Effects

In the previous sections we studied whether the proposed algorithms can identify the correct subgroup in the presence of confounders. An interesting question is what happens when there are no known subgroups. In this case we would like an algorithm to return a root-only tree. We repeat the experiment of Section 5.4.1, but now without the presence of the subgroup, i.e. there is no treatment effect heterogeneity. We perform 500 simulations and report the PCT as the number of times a root-only tree is returned.

In figure 5.10 we observe that the approach that uses weighting estimators has a PCT close to one which remains almost unaffected by the presence of confounders. In contrast IT with unadjusted estimator tends to split the data using the confounders, $X_1 - X_3$. The proportion of root-only trees decreases to close to zero when $\gamma = 1$. Therefore, in the absence of subgroups, methods that use unadjusted estimators of the treatment effect can be biased to the treatment assignment mechanism and identify subgroups defined by the confounders. We can ameliorate these issues with weighting estimators.

5.5 Case Studies

In this section we apply MSE-IT in two data sets. In the first experiment we consider a simulated trial and introduce artificially imbalance in the data. We would like to explore how imbalance can affect IT as well as the benefits from estimating the splitting criterion with weighting methods. For the second study we consider a dataset where previous works have not identified any subgroups

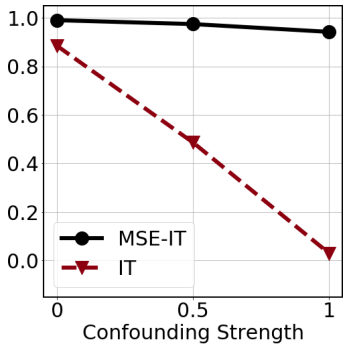


Figure 5.10: In the absence of heterogeneous effects MSE-IT correctly identifies an only root tree and exhibits a PCT close to 1. IT with unadjusted estimator tends to identify trees defined by the confounders.

and we explore whether we can reproduce these results using MSE-IT.

5.5.1 Application to Simulated Study

In this section we use the simulated trial data described in Chapter 4. In order to simulate an observational study we artificially introduce imbalance by removing a non-random proportion of the data. In particular we remove all treated patients with pre-infusion apache-ii score lower or equal to 23 and age lower or equal to 70. We note that these covariates were also found to have the strongest prognostic effect as indicated by the importance scores derived using a random forest model with 500 trees. In the resulting dataset we expect a higher imbalance between the treatment groups compared to the initial data.

In figure 5.11 we report the absolute standardised mean difference (Austin, 2009a) in the two treatment groups before and after removing a non-random proportion of the data. The vertical dotted line indicates the difference that is normally accepted as being sufficient, so that the covariates that have a lower value can be considered as balanced (Austin, 2009a). We observe that for most covariates there is an increased imbalance and this holds particularly for the patient's age and the pre-infusion apache-ii score. After we create the new sample the outcome is generated as follows: $Y = \sum_i X_i + 2T\mathbb{I}(\text{PRAPACHE} > 25)$, where X_i are the standardised covariates and PRAPACHE is the apache-ii score. This covariate acts both as a predictive covariate and a confounder. In this way we will know the true subgroup so that we can evaluate the results but in contrast to the previous section we do not know the treatment assignment mechanism.

We apply IT and MSE-IT with first-degree polynomial kernel and the weights estimated for each possible split. For both methods we use a minimum depth of 5, minimum node size for performing a split equal to 40 and minimum size of a terminal node equal to 20. The final tree is selected following the bootstrap-based approach (Su et al., 2008, 2009; Dusseldorp and Van Mechelen, 2014) with 25 bootstrap samples and using the 1-SE rule as described in (Dusseldorp and Van Mechelen, 2014). In this experiment we prefer this method over keeping a validation dataset due to the small sample size.

The fully grown trees using IT and MSE-IT are shown in figure 5.12. IT first splits the data using the patient's age, which acts as a confounder. The final tree after pruning does not retain any subgroup and is a root-only tree. By using weighting estimators for the treatment effect with MSE-IT we can identify the correct tree. The data are first split using the predictive covariate, while also the estimated treatment effects are those expected according the outcome function. After pruning, MSE-IT retains only the split on the pre-infusion apache-ii score, hence identifying the correct subgroup. Therefore, we can validate that by using weighting methods we can recover the true subgroup, while with unadjusted estimators the resulting subgroups can be biased towards splitting the data using

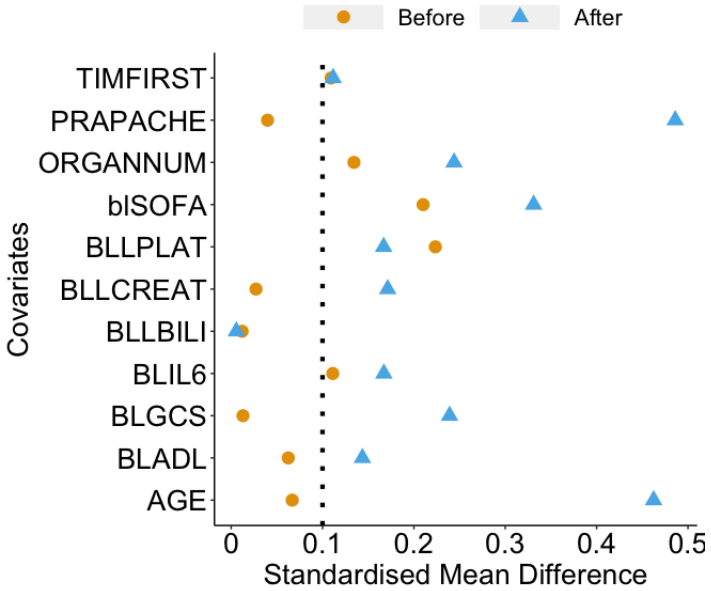
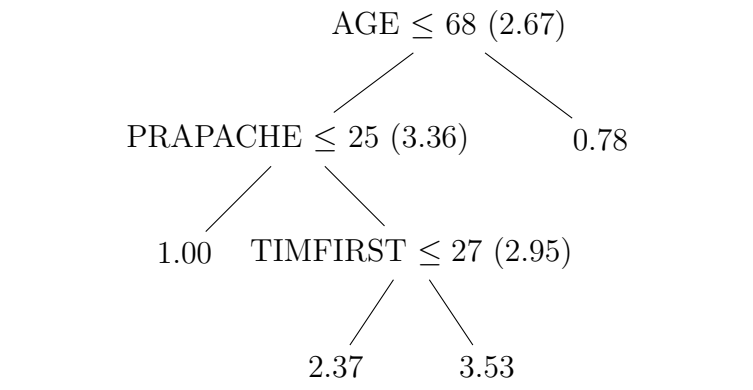
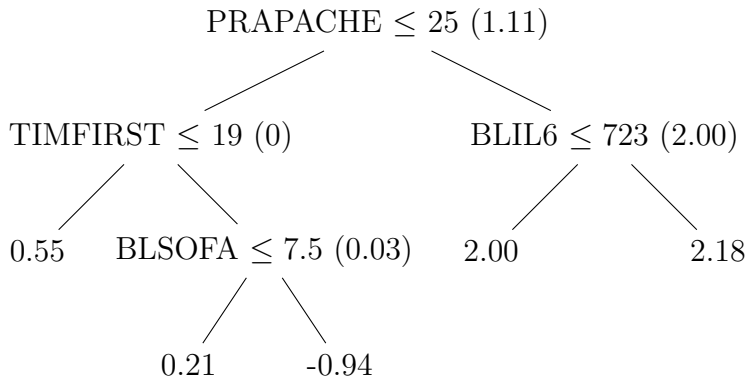


Figure 5.11: Covariate balance (Absolute Standardised Mean Difference) between the two treatment groups in the initial data and after removing a non-random proportion.



(a) IT



(b) MSE-IT

Figure 5.12: Fully grown trees using IT and MSE-IT in the simulated data where the patient's age (AGE) and the pre-infusion apache-ii score (PRAPACHE) are the most imbalanced covariates, while the latter is the only predictive covariate. The final trees after pruning are a root-only tree using IT and a tree that splits the data on PRAPACHE for MSE-IT.

confounders. This was also observed in Section 5.4.4 where in the absence of predictive covariates, using unadjusted estimators tends to split the data using the confounders. In the next section we validate our method in a case study where there are no known predictive covariates or subgroups.

5.5.2 Application to Right Heart Catheterization Study

We evaluate MSE-IT on an observational study that evaluated the effectiveness of right heart catheterization (RHC) in the initial care of critically ill patients[2] (Connors et al., 1996). The dataset contains 2184 participants who received

[2]Data obtained from `http://hbiostat.org/data` [last accessed: January 2021] courtesy of the Vanderbilt University Department of Biostatistics

RHC during the first day of their hospitalisation and 3551 who did not receive it. The patients are described with 51 covariates. A subgroup analysis was performed in (Yang et al., 2021), where the authors used the 30-day survival as the outcome and applied IPW, DR and regression methods to estimate the effects. After splitting the data in 80% for creating the tree and 20% for pruning they found that any subgroups where not retained in the final tree. We repeat the aforementioned analysis using MSE-IT with first and second degree polynomial kernels and fitting the weights either once in the root node or in each parent node. We choose these approaches due to their lower computational complexity. Missing data were handled by including a new category. With this approach we can use the full data, we note however that a pitfall is that if the covariate with missing values is a confounder, then residual confounding can be an issue (Bennett, 2001; Pedersen et al., 2017). Hence, while this approach may not be problematic when the covariates are independent of the treatment (where missing data are likely to be balanced between the treatment groups), it should be considered when evaluating treatment effects in observational studies. In this section we compare the final tree (after pruning) with the results of previous studies, but do not evaluate the resulting treatment effects.

We find that in all cases the subgroups of the initial tree where not retained after pruning. The first covariate used to split the data is the probability of surviving 2 months at study entry, a covariate that has also been identified as potentially promising in previous works (Yang et al., 2021; Connors et al., 1996).

5.6 Chapter Summary

Subgroup identification has been studied extensively in the context of (marginally) randomised studies, where the treatment groups are balanced. In this chapter we studied this problem when there are observed (but possibly unknown) confounders that cause imbalance between the covariates in the two treatment groups. This can be the case in conditionally randomised studies or observational studies under the assumption of unconfoundedness. This problem has been studied in (Yang et al., 2021), where the authors propose approaches that use IPW, DR and regression estimators. In this chapter we used non-parametric weighting methods for treatment effect estimation, which have shown good empirical results while they avoid modelling the treatment assignment (Kallus and Santacatterina,

2019b; Kallus et al., 2021).

The simulations show that using weighting methods to derive unbiased estimations of the subgroup-specific treatment effects can improve the results of standard methods, particularly in the presence of strong confounding. In the absence of treatment effect heterogeneity, the proposed methods can provide better results regarding the false discovery of subgroups, while IT may split the data using the confounders. We note that a key difference between MSE-IT and B-IT is the regularisation of the weights which controls the variance of the effect. The two approaches use weights that are motivated from different perspectives and in our context, where we are interested in identifying the correct tree and calculating the absolute error of the treatment effect, we find them to perform similarly under moderate values of the confounding strength. We do note however, that under the largest confounding value, MSE-IT tends to identify the correct tree more often (Sections 5.4.1,5.4.2) showing an advantage over B-IT. In some scenarios under correct specification of the outcome within the subgroup and under various confounding mechanisms the resulting methods show similar performance to the case of $\gamma = 0$ (marginal randomisation). However, even under incorrect specification we can balance selected moments of the distributions of the covariates in the two treatment groups, which improves the results under moderate confounding compared to using the unadjusted estimator of the treatment effect.

We note here that the subgroups identified using the described procedure can only be considered as promising candidates and further analysis will be required to assess them (e.g. estimating p-values, controlling the Type I error rate). Even though we do not discuss these issues here and we focus primarily on the treatment effect estimation problem, we need to highlight their importance. For example, if test data are available, the resulting subgroups could be re-evaluated in order to retain the final ones (Su et al., 2009; Lipkovich et al., 2011). Alternatively, in (Lipkovich et al., 2011) the authors discuss an approach for controlling the overall Type I error rate, when test data are not available (we briefly described this in Chapter 3). When performing multiple tests (as in the case of subgroup identification), adjustment and interpretation of the p-values is also an important concern (Dmitrienko et al., 2017; Alosh et al., 2014).

Chapter 6

A Multi-objective Evaluation Framework for Subgroup Identification Algorithms

In the previous chapters we studied the problem of treatment effect heterogeneity from different perspectives. A common theme was the application of subgroup identification algorithms and we explored some representative examples. However, there is a plethora of methods suited for this problem in marginally randomised studies. For example, Lipkovich et al. (2017a) review and categorise 16 subgroup identification algorithms, while more recently Loh et al. (2019) perform an empirical comparison of 13 methods. Given such a large number of algorithms to choose from, an important question is how to select the best for the task at hand. This is a challenging problem since we never observe the true treatment effect or the predictive covariates that interact with the treatment.

In this chapter we phrase the algorithm selection problem in a multi-objective framework[1]. One objective is to evaluate an algorithm in terms of the quality of the subgroups it identifies. Following the literature of subgroup identification, we define the *subgroup quality* as the excess treatment effect in the identified subgroup compared to the average treatment effect (Foster et al., 2011). For algorithms that estimate the individual treatment effect as part of their process we can also estimate some approximation of the error on the estimated effect (Steingrimsson et al., 2017; Schuler et al., 2018). Besides evaluating the efficacy

[1]An initial version of this chapter was presented in the ECML/PKDD 2020 workshop "*Machine Learning for Pharma and Healthcare Applications*".

of the treatment we would like the resulting subgroups to be reproducible. To quantify this, we introduce the *subgroup stability*, a measure that captures how small changes in the data affect the selected subgroups. We demonstrate the use of the proposed framework in a number of cases, assuming marginal randomisation throughout the chapter.

6.1 Measuring the Quality of Subgoups

Suppose we are given a dataset $\mathcal{D} = \{\boldsymbol{x}_i, t_i, y_i\}_{i=1}^n$, with n realisations of the variables $\boldsymbol{X}, T, Y$, where $\boldsymbol{X} \in \mathbb{R}^d$ are the covariates, $T \in \{0, 1\}$ is the treatment and Y the outcome. In the literature of subgroup identification, an algorithm is commonly evaluated first in simulated scenarios where we have knowledge of the ground truth. Measures that capture whether it selects the correct predictive covariates, identifies the right splits and estimates the true effects can be used for the evaluation. In non-simulated scenarios such measures are not applicable and the results are commonly evaluated with respect to the characteristics of the identified subgroups. For example by comparing the effects in the subgroups with the average effect in the data, or by validating their plausibility based on the results of previous studies or the knowledge of domain experts. Suppose our objective is to identify subgroups of enhanced effect. Then a measure that quantifies the efficacy of the treatment within the identified subgroup could be used to assess whether the algorithm has achieved our primary objective. Foster et al. (2011) introduce the subgroup quality, a measure that captures the treatment benefit.

Let $\hat{\mathcal{S}}$ denote a region in the covariate space identified by some subgroup identification algorithm. Foster et al. (2011) define the quality of this region as the excess treatment effect over the average treatment effect in the population:

$$Q(\hat{\mathcal{S}}) = \mathbb{E}[Y \mid T = 1, \boldsymbol{X} \in \hat{\mathcal{S}}] - \mathbb{E}[Y \mid T = 0, \boldsymbol{X} \in \hat{\mathcal{S}}] - \text{ATE}$$

In practice we get an estimate of the above $\hat{Q}(\hat{\mathcal{S}})$ from the observed data. Following Foster et al. (2011) and Huling and Yu (2018), we evaluate the quality of a subgroup using the bootstrap bias correction approach (Harrell Jr et al., 1996). Let $\hat{\mathcal{S}}$ denote the subgroup derived using the full data. In practice some algorithms may identify multiple subgroups, some of which may lead to enhanced and some to deteriorated effects. Here with $\hat{\mathcal{S}}$ we denote the sample that has

an estimated enhanced treatment effect, i.e. it includes all subgroups with an estimated treatment effect greater than some predefined threshold. The quality is estimated as follows:

1. Estimate $\hat{Q}_{\mathcal{D}}(\hat{\mathcal{S}}_{\mathcal{D}}) = \widehat{\text{SATE}}_{\mathcal{S}} - \widehat{\text{SATE}}$, where $\widehat{\text{SATE}}_{\mathcal{S}}$ is an estimation of the treatment effect in the subgroup and $\widehat{\text{SATE}}$ in the sample $\mathcal{D}$. This quantifies the excess treatment effect in the subgroup over the whole sample assuming the data come from a marginally randomised study.

2. Construct B bootstraps by sampling with replacement. For the b-th sample run the subgroup identification algorithm and calculate $\hat{Q}_{\mathcal{D}}(\hat{\mathcal{S}}_b)$ and $\hat{Q}_b(\hat{\mathcal{S}}_b)$. The first term is the quality of the subgroup estimated using the b-th sample and evaluated in the full sample. The second term is the quality of the subgroup derived in the b-th sample and estimated using the same sample. The bias of the quality is:

$$\text{bias}(\hat{Q}_{\mathcal{D}}(\hat{\mathcal{S}}_{\mathcal{D}})) = \frac{1}{B}\sum_{b=1}^{B} \hat{Q}_b(\hat{\mathcal{S}}_b) - \hat{Q}_{\mathcal{D}}(\hat{\mathcal{S}}_b)$$

   Therefore, here $\hat{Q}_b(\hat{\mathcal{S}}_b)$ is an estimate of the quality using the same data that were used to derive the subgroup and $\hat{Q}_{\mathcal{D}}(\hat{\mathcal{S}}_b)$ is the quality evaluated in the full data, which acts as the external dataset.

3. The bias corrected estimate of the subgroup quality is:

$$\hat{Q}_{\mathcal{D}}(\hat{\mathcal{S}}_{\mathcal{D}}) - \text{bias}(\hat{Q}_{\mathcal{D}}(\hat{\mathcal{S}}_{\mathcal{D}}))$$

Foster et al. (2011) propose various approaches for estimating the quality and find that the bias corrected approach we described above to be the most promising. Therefore, this is the measure we will use for the rest of this chapter. For completeness, we describe some alternative choices. One option is to estimate the above using cross-validation or repeated splitting of the data (Foster et al., 2011; Huling and Yu, 2018). For example, in the latter case we perform B random splits of the data intro training/validation sets and estimate the subgroup quality as $(1/B)\sum_{b=1}^{B} \hat{Q}_b^{validation}(\hat{\mathcal{S}}_b^{train})$, i.e. the quality of the subgroup identified using the training data, evaluated in the validation set. Another choice would be to use the out-of-sample data of each bootstrap as the validation set. This

approach however could also introduce some bias in the estimation (Lipkovich et al., 2017a). To this end, one may adopt the approach of Efron (1983) proposed in the context of prediction errors and described for quality estimation in (Lipkovich et al., 2017a). The estimation in this case would be a weighted average of $\hat{Q}_{\mathcal{D}}(\hat{\mathcal{S}}_{\mathcal{D}})$ and the quality estimated in the out-of-bag data $\frac{1}{B}\sum_{b=1}^{B}\hat{Q}_{-b}(\hat{\mathcal{S}}_b)$.

6.2 Measuring the Stability of Subgroups

Clearly the identified subgroups should ideally capture the primary objective of the analysis, whether this is salvaging a failed study, identifying super responders or some other objective (Lipkovich et al., 2017a). Any formed hypothesis will likely need to be tested further e.g. by domain experts and/or by performing a confirmatory analysis. Since this might be a time-consuming procedure, we would like to select the most promising hypotheses. These should not only achieve our primary objective, captured for example by the subgroup quality, but should also be robust to small changes in the data enhancing their reproducibility.

A subgroup identification algorithm should not vary its preferences with small changes in the data. The instability of an algorithm can be due to various factors such as noise, data size, data dimensionality, class imbalance, irrelevant and redundant covariates etc. To quantify the *stability* of subgroup identification algorithms due to such changes we borrow concepts from the area of feature/variable selection stability (Kalousis et al., 2007; Kuncheva, 2007; Nogueira et al., 2017; Sechidis et al., 2019b).

Suppose we run a subgroup identification algorithm and identify a region $\hat{\mathcal{S}}$. We would like to quantify how robust this result is to small changes in the data. To this end we repeat the procedure B times, each time performing some change in the data, e.g. adding some small amount of noise to the outcome. An algorithm would be stable if such a change does not affect the definition of the identified subgroup $\hat{\mathcal{S}}$. We form a subgroup membership matrix $\mathcal{M}$ with B rows and n columns, where the entries denote whether an example has been selected in a subgroup in the b-th sample. To fix ideas, consider the following examples of

membership matrices:

$$\mathcal{M}_1 = \begin{array}{c} \begin{matrix} M_1 & M_2 & M_3 & M_4 & M_5 \end{matrix} \\ \begin{pmatrix} 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 \end{pmatrix} \end{array} \quad \mathcal{M}_2 = \begin{array}{c} \begin{matrix} M_1 & M_2 & M_3 & M_4 & M_5 \end{matrix} \\ \begin{pmatrix} 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 1 \\ 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 \end{pmatrix} \end{array}$$

In $\mathcal{M}_1$ the observations are consistently either included or not included in the subgroup, while $\mathcal{M}_2$ shows an example of an unstable procedure. Based solely on the stability of subgroup membership we would trust the first algorithm.

There are many stability measures suggested in the literature. Nogueira et al. (2017) proposes an axiomatic framework and suggests a novel measure that satisfies a set of desirable properties. According to this measure each column in the membership matrix can be treated as a Bernoulli variable with B realisations. Hence, a key advantage of this measure is its probabilistic definition (as opposed to the most commonly used set-theoretic definition) which, allows its extension to other settings, such as correlated variables (Sechidis et al., 2019b). The measure is defined as:

$$\hat{\Phi}(\mathcal{M}) = 1 - \frac{\frac{1}{n}\sum_{f=1}^{n} s_f^2}{\mathbb{E}\left[\frac{1}{n}\sum_{f=1}^{n} s_f^2 \mid H_0\right]}$$

where s_f^2 is the (unbiased) sample variance of the f-th column and H_0 denotes the null model. The null model states that for all rows b in the membership matrix $\mathcal{M}$, all subsets of size k_b have an equal probability of being drawn (Nogueira et al., 2017).

In particular, the variance of a column under the null model can be expressed as (Nogueira et al., 2017; Sechidis et al., 2019b):

$$\text{var}\left(M_f \middle| H_0\right) = p_f^0(1 - p_f^0), \ \forall f \in \{1, ..., n\}$$

where p_f^0 captures the probability an observation being selected in a matrix $\mathcal{M}$ under the null model. Under this model there is a random procedure that generates a matrix $\mathcal{M}$. We assume that in each row b all permutations are equally likely. In each row we have k_b values set to 1 hence the probability of observing

the value 1 for a column is equal to (Nogueira et al., 2017, B.2):

$$p\left(M_f = 1 \middle| b, H_0\right) = \frac{\#\{\text{possible rows with } k_b \text{ 1s and } M_f = 1\}}{\#\{\text{possible rows with } k_b \text{ 1s}\}}.$$

The numerator is $\binom{n-1}{k_b-1}$ and the denominator is equal to $\binom{n}{k_b}$. Hence for a given row this is equal to $\frac{k_b}{n}$. Since we have B independent rows we can write $p_f^0 = \frac{1}{B}\sum_b \frac{k_b}{n} = \bar{k}/n$. Putting everything together we can estimate the stability as:

$$\hat{\Phi}(\mathcal{M}) = 1 - \frac{\frac{1}{n}\sum_{f=1}^{n} s_f^2}{\frac{\bar{k}}{n}\left(1 - \frac{\bar{k}}{n}\right)} \tag{6.1}$$

This measure is upper-bounded by 1 when a procedure is perfectly stable (e.g. matrix $\mathcal{M}_1$) and its lower bound is 0 (asymptotically if we use the unbiased sample variance in the numerator (Nogueira et al., 2017)).

In this chapter we adopt this measure and consider a mechanism that introduces noise to the outcome. For binary outcomes this can be achieved by flipping the labels of a random sample of the data (in the case of imbalanced classes this is also discussed in (Wald et al., 2012; Altidor et al., 2011)). From a more practical perspective, such a mechanism can represent scenarios where the outcome is measured based on methods that may introduce noise such as questionnaires, manual labelling or based on devices that may introduce measurement noise. Additionally, this could represent changes in the number of responders when a binary outcome is formed by dichotomisation of a continuous response. For example, in Rheumatoid Arthritis a commonly used measure is ACRx which indicates an improvement of $x\%$ in various disease activity measures and assessments (Felson et al., 1993). In such a scenario we could consider how a change in a small number of responders (which could be attributed to small changes in the continuous variables) could affect the resulting subgroups.

Another scenario most commonly found when measuring the stability of feature selection algorithms, is varying the sample size by performing bootstrapping or removing a constant number of examples (Nogueira et al., 2017; Sechidis et al., 2019b). In the case of subgroup identification such a mechanism will introduce missing values in the membership matrix making the above stability measure not directly applicable. Suppose we have a constant number of missing values equal to l, e.g. dropping a percentage of the data. Since the values are missing completely at random the probability of a Bernoulli variable can be estimated by

ignoring the missing values:

$$\hat{p}_f = \frac{\sum_{b=1}^{B} \mathbb{I}[M_f^b = 1]}{\sum_{b=1}^{B} \left(\mathbb{I}[M_f^b = 1] + \mathbb{I}[M_f^b = 0]\right)}$$

Here M_f^b denotes the value of the b-th row and f-th column in $\mathcal{M}$. For the denominator we can follow a similar procedure as the one described above (Nogueira et al., 2017, B.2). In each row b of a matrix we have k_b values set to 1 and $n - l$ observed values, hence the probability for the probability $p\left(M_f = 1 \middle| b, H_0\right)$ the numerator is $\binom{n-l-1}{k_b-1}$ and the denominator is equal to $\binom{n-l}{k_b}$. In a matrix $\mathcal{M}$ we have a total of B independent rows, hence we can write $p_f^0 = \frac{1}{B}\sum_b \frac{k_b}{n-l} = \frac{\bar{k}}{n-l}$. Here we assume that under the null model all permutations of 1s, 0s and missing values per row are equally likely and missing values are ignored in the estimation. For the rest of this chapter we will adopt a mechanism that introduces noise and use the measure of eq. (6.1).

6.3 Experiments

We study five subgroup identification algorithms with diverse characteristics. VT (Foster et al., 2011) follows the counterfactual modelling approach by first learning the potential outcomes using random forests and then using a decision tree with the estimated Individual Treatment Effect as the target in order to derive the final subgroups. Estimation of the potential outcomes is performed either with separate models for each treatment group or with a single model using the treatment as an additional covariate as well as first-order interactions between the treatment and the covariates. Following the nomenclature of Künzel et al. (2019) these are referred to as VT-T and VT-S respectively. MCR (Tian et al., 2014) follows the outcome transformation approach that avoids modelling the main effect (see also Chapter 3). PRIM (Huang et al., 2017) follows the bump-hunting procedure (Friedman and Fisher, 1999) to identify the optimal partition that maximises the treatment effect, while ensuring its statistical significance. SIDES (Lipkovich et al., 2011) follows a recursive partitioning approach to identify subsets of the data with desirable characteristics. Finally IT (Su et al., 2009) follows the procedure of CART (Breiman et al., 1984) and builds a decision tree by recursively partitioning the space and maximising the treatment effect heterogeneity. We discussed this methodology in detail in the previous chapter.

We will examine synthetic data and a simulated trial in order to show how the described measures can be used in practice. For the simulated scenarios we consider the models B1, B2 and B5 based on (Loh et al., 2019). The treatment is binary and the outcomes are generated as follows:

$$\text{B1}: \ \text{logit}(p(Y=1 \mid \boldsymbol{X})) = 0.5(X_1 + X_2 - X_5) + 2T\mathbb{I}(X_6 = \text{odd})$$
$$\text{B2}: \ \text{logit}(p(Y=1 \mid \boldsymbol{X})) = 0.5X_2 + 2T\mathbb{I}(X_1 > 0)$$
$$\text{B5}: \ \text{logit}(p(Y=1 \mid \boldsymbol{X})) = 0.2(X_1 + X_2 - 2) + 2T\mathbb{I}(X_6 = \text{odd} \cap X_1 < 1)$$

In all datasets we consider 10 covariates with the following marginal distributions as described in (Loh et al., 2019): $X_j \sim \mathcal{N}(0,1)$ for $j = 1,2,3,7,8,9,10$, $X_4 \sim Exp(1)$, $X_5 \sim Ber(0.5)$ and $X_6 \sim Multi(10)$. Here $\mathcal{N}(0,1)$ denotes the standard normal distribution, $Exp(1)$ the exponential distribution with mean 1, $Ber(0.5)$ Bernoulli with success probability 0.5 and $Multi(10)$ is the multinomial distribution with values $\{1, ..., 10\}$ all having equal probability. All covariates are independent except the pairs X_2, X_3 and X_j, X_k for $j, k = 7,8,9,10,\ j \neq k$, which have a correlation equal to 0.5. We choose these outcomes due to their diversity, as they use predictive covariates with different distributions.

6.3.1 Predictive Covariate vs Subgroup Stability

Predictive covariate selection and subgroup identification are two closely related areas. As we discussed there are methods that perform the first but not the second, unless additional steps are included in the algorithm (e.g. MCR). On the other hand, algorithms that perform subgroup identification will also identify potentially predictive covariates, since the subgroups are defined by these covariates. For algorithms that perform both, an interesting question is to quantify the stability of the two tasks and explore whether there is a relationship between the two. Finding one to be much more stable than the other can allow us to form hypotheses about the algorithm and the data. For example, if the covariate selection task is stable, but the subgroup definition is unstable, then we could explore the criteria (e.g. thresholds) used to define the final subgroup. Another example, is the presence of strongly correlated covariates, in which case the algorithm may be unstable on the covariate identification task (e.g. see (Sechidis et al., 2019b) for examples where this can occur and how it can affect the stability) but stable on the subgroup definition.

To show an example of how these measures could be used in practice we generate 100 datasets for each outcome model and plot the predictive covariate stability versus the subgroup stability estimated using 50 samples (number of rows in the membership matrix). A sample is formed by changing the label in a randomly selected 10% of the data. Then for each sample we use VT-S with the default parameters and retain as final subgroups those that have an effect larger than $\widehat{\text{SATE}}+0.1$ (Foster et al., 2011), where $\widehat{\text{SATE}}$ is the estimated average effect in the sample. In figure 6.1 we observe that in the second scenario the algorithm tends to give more stable results. Additionally, we can observe a statistically significant correlation (Pearson correlation coefficient) between predictive covariate stability and the subgroup stability measure in all scenarios. Therefore in these simulations we find that the more often the algorithm identifies the same predictive covariates it also identifies the same subgroups.

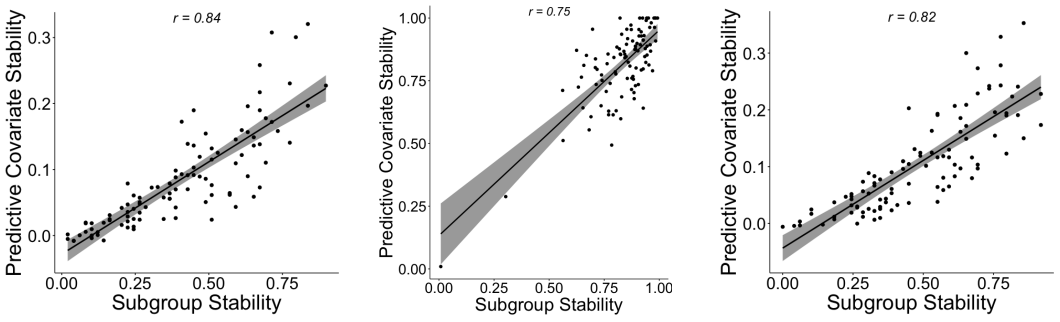


Figure 6.1: Predictive covariate versus subgroup stability in three simulated outcomes. The stability is estimated by flipping the labels for 10% of the data. Each point corresponds to a realisation of the outcome function and we report the stability for a total of 100 realisations.

We repeat the experiment for model B2 but we consider three modifications. In the first modification we change the threshold, so that we retain the subgroups that have a positive estimated effect. By reducing the threshold we expect the algorithm to be less stable with respect to the subgroup membership, particularly if a large number of observations exhibit a treatment effect close to that threshold. The results are reported in figure 6.2(a) where we observe that the algorithm has a lower subgroup stability compared to our initial setting. In particular, when using $\widehat{\text{SATE}}+0.1$ as the threshold (figure 6.1(b)) the average subgroup stability is 0.8. By reducing the threshold now this becomes equal to 0.7. We additionally observe that there is no longer a linear correlation between the two measures and

we can observe large predictive covariate stability but low subgroup stability for some realisations of the data. In the second modification of B2 we reduce the coefficient of $\mathbb{I}(X_1 > 0)$ from 2 to 1, hence making the problem more challenging. In figure 6.2(b) we observe that both measures are reduced when compared to figure 6.1(b). Lastly, in the third modification we replace the term $\mathbb{I}(X_1 > 0)$ with X_1. The results (figure 6.2(c)) indicate that when the subgroup is not defined by a clear cut-off value in the covariate space the correlation between the measures we observed in the initial dataset is no longer present. Therefore, the observed linear correlation could be a combination of factors such as retaining subgroups that have a large effect and defining the subgroups by a partition of the space, which can be identified by a tree-based algorithm like VT.

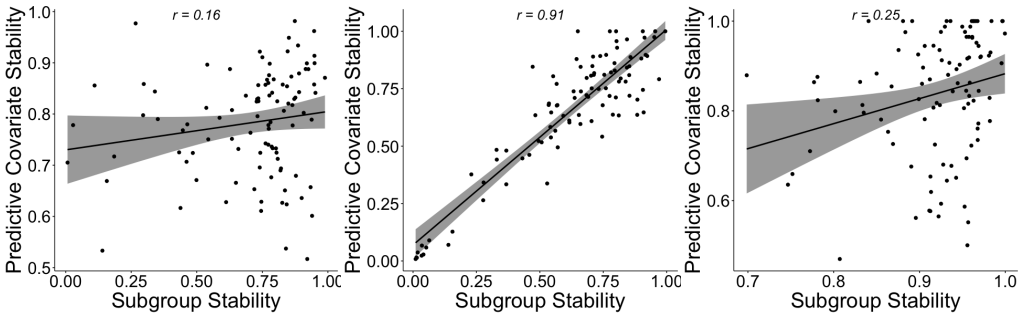


Figure 6.2: Predictive covariate versus subgroup stability for three modifications of outcome model B2. In all scenarios the stability is estimated by flipping the label for 10% of the data. In (a) we reduce the threshold of VT that controls the final subgroup selection, in (b) we reduce the effect in the subgroup and in (c) the subgroup is not defined by a clear cut-off.

6.3.2 Subgroup Quality vs Stability

In this section we explore how stability can be used in conjunction with existing measures to perform hyper-parameter selection. We consider VT with different number of trees (100, 500, 1000) as well as different types of modelling, i.e using a Two-model approach (T) or Single-model (S). We use 50 bootstraps and estimate the subgroup quality as described in this chapter. The stability is estimated by flipping the labels for 10% of the data and using again 50 samples. In figure 6.3 we report the quality versus subgroup stability for two realisations of outcome models B1 and B2. The annotated points (triangles) correspond to the pareto

front, i.e. no other point achieves both a higher quality and higher stability.

Focusing on the first plot we notice that using a Two-model approach with 1000 trees achieves the highest quality. However, we observe that also a T-learner with 100 trees and a S-learner with 500 or 1000 trees are practically equivalent in terms of subgroup quality. If we were to use this measure to identify the optimal method then any of the above could be selected. However, once we consider the subgroup stability then a T-learner with 100 trees becomes the clear choice, since it achieves a higher stability without sacrificing quality. In figure 6.3 we also show an example using B2 where we observed that the algorithm tends to achieve large values of subgroup stability. In this case a Single-model approach with 500 trees is clearly the optimal choice since it achieves the highest stability while its quality is close the largest value. This is an example of how plotting the quality and stability can allow us to identify an algorithm with the desirable operational characteristics. We will now describe how these measures can be used to select an algorithm in a simulated clinical trial.

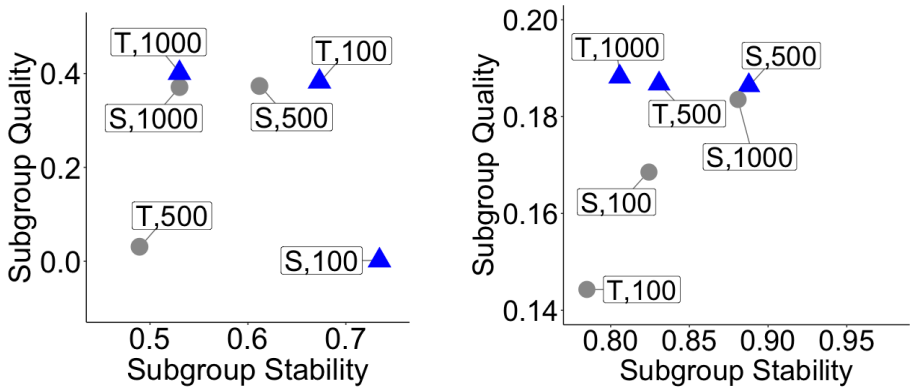


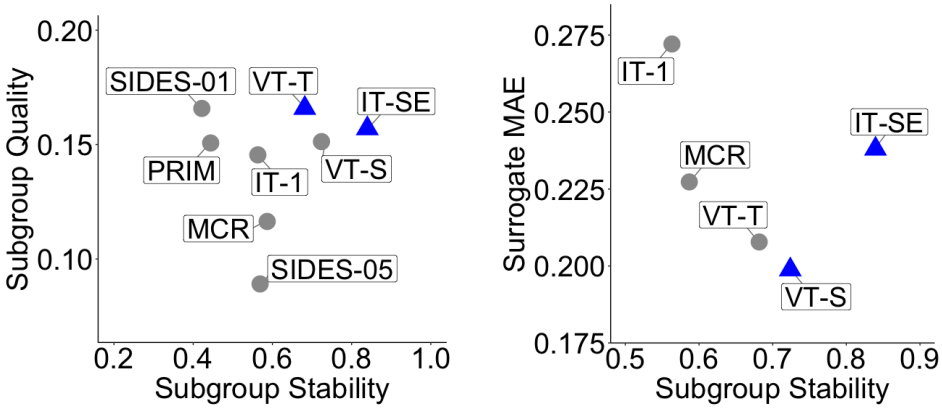
Figure 6.3: Examples of how subgroup quality and stability can be used to perform hyper-parameter selection for VT. We notice that we may choose an algorithm that achieves slightly lower quality compared to the optimal but comes with a much higher stability.

6.3.3 Algorithm Selection

We evaluate five subgroup identification algorithms with different characteristics on the simulated study the details of which we discussed in Chapter 4. Here the outcome of interest will indicate survival. For MCR and VT we keep the subgroups that have an estimated treatment effect larger than the estimated SATE

in the overall sample. For SIDES we keep the subgroups that have both larger effect than the estimated SATE *and* exhibit a statistically significant effect with $p-\text{value} \leq 0.01$ or $p-\text{value} \leq 0.05$. We report results for both cases, which will be denoted by SIDES-01 and SIDES-05 respectively. We expect the first approach to lead to higher quality since it will retain subgroups with higher statistical significance. For IT we use the bootstrap-based approach to identify the final tree (Su et al., 2009, 2008; Calhoun et al., 2018). We consider different values for the parameter ρ in eq. (5.6). Larger values penalise heavier the complexity of the final tree, resulting in smaller trees. We report results using $\rho = 1$ as well as using the 1-SE rule as described in the previous chapter. The former tends to lead to trees with many subgroups, while the latter is more conservative. These two approaches are denoted as IT-1 and IT-SE respectively.

In figure 6.4(a) we report the subgroup quality versus subgroup stability. The quality is estimated as described in the previous section and using 100 bootstraps. The stability of the algorithms is estimated by changing the labels for 10% of the data. This is repeated 100 times and the estimated stability is reported. We notice that VT-T and IT-SE are in the pareto front. The results indicate that we could choose IT-SE since we can sacrifice a small amount of quality in order to get a much more stable algorithm. We can also validate that SIDES-01 and IT-SE lead to higher values of quality compared to their less conservative counterparts SIDES-05 and IT-1.



(a) (b)

Figure 6.4: (a) Comparison of subgroup identification algorithms with respect to subgroup quality and stability on a simulated trial and (b) Comparison of subgroup identification algorithms with respect to MAE and subgroup stability.

What happens if we consider the error of the estimated treatment effect?

Some of the algorithms described in this chapter provide estimations of the individual treatment effect as part of their process. These include VT (Foster et al., 2011) and MCR (Tian et al., 2014) the details of which can be found in Chapter 3. Also, in the previous chapter we discussed how IT (Su et al., 2009) can be used for this task. Therefore, a relevant question is how these algorithms would perform if they were to be used for individual treatment effect estimation. Using the Mean Absolute Error (MAE), the error of the estimated ITE can be expressed as:

$$\text{MAE} = \frac{1}{n_{eval}} \sum_{i=1}^{n_{eval}} |\widehat{\text{ITE}}(\boldsymbol{x}_i) - \text{ITE}(\boldsymbol{x}_i)|$$

where n_{eval} is the size of the subset of the data on which the algorithm is evaluated, $\widehat{\text{ITE}}(\boldsymbol{x}_i)$ is the estimated treatment effect for $\boldsymbol{x}_i$ and $\text{ITE}(\boldsymbol{x}_i)$ is the true treatment effect.

The main challenge comes from the fact that $\text{ITE}(\boldsymbol{x}_i)$ is never observed. To this end, some recent works focus on replacing it with an estimate derived from the data. For example, Shalit et al. (2017) use a nearest-neighbour matching estimator to approximate the ground truth in order to perform hyper-parameter selection. In these cases the counterfactual of an observation is approximated by the factual outcome of its nearest neighbour that belongs to the opposite treatment group. In (Schuler et al., 2018) the authors focus on the MSE of ITE and discuss several approaches for approximating $\text{ITE}(\boldsymbol{x}_i)$ using a model fitted on the evaluation dataset. Some choices include using matching and IPW estimators. Here we will follow a similar approach and use a plug-in estimate of $\text{ITE}(\boldsymbol{x}_i)$. We note however that this comes with certain limitations. Clearly this requires the plug-in estimates to be close to the ground truth. To this end, Alaa and Van Der Schaar (2019) propose a method for correcting the potential bias that may arise from the above procedure, considering that the true loss will be close (but not equal) to the one estimated using the plug-in model. In particular, they express the loss (with parameters the conditional potential outcomes, propensity score and distribution of the covariates) as the von Mises expansion around the loss that uses plug-in estimates for its parameters and show how the relevant influence functions can be calculated.

In order to evaluate the algorithms with respect to their surrogate MAE, we

use 90% of the data for training the subgroup identification algorithm and 10% of the data for evaluation and report the average value over 100 random splits. We choose CF as the plug-in model as implemented in the *grf* package (Tibshirani et al., 2020). In particular the surrogate MAE is estimated as:

$$\frac{1}{n_{eval}} \sum_{i=1}^{n_{eval}} |\widehat{\text{ITE}}(\boldsymbol{x}_i) - \widehat{\text{ITE}}_{CF}(\boldsymbol{x}_i)|$$

where $\widehat{\text{ITE}}_{CF}(\cdot)$ is an estimate using the CF model trained on the evaluation dataset. Due to the small sample size we do not use the "honest" splitting procedure suggested in (Wager and Athey, 2018a). For VT we use the estimated ITE derived from the final tree and not the one derived in the first step. This is in order to evaluate the whole subgroup identification algorithm in terms of treatment effect estimation. The interested reader can also find information about a similar subject in (Makar et al., 2019) where the final tree is referred to as the distilled model and the procedure of training a simpler and easier to interpret model is called distillation. In figure 6.4(b) we plot the surrogate MAE and subgroup stability. Based on the results we could choose VT-S as the algorithm that minimises the estimated error while achieving the second largest stability. We note that the quality as a measure could be misleading if the treatment effect is not estimated correctly, which is particularly important if we were to perform subgroup identification in observational data. We highlight, however, that the results should be interpreted with caution, since the MAE acts here solely as a surrogate based on the estimations of CF.

6.4 Summary

We introduced a multi-objective evaluation framework for subgroup identification algorithms based on two desirable characteristics: subgroup quality and subgroup stability. We showed how predictive covariate and subgroup stability can be used to get insights about the task at hand. Furthermore, we showed how we can use our framework to tune the hyper-parameters of a popular subgroup identification algorithm. Lastly, we studied how quality and subgroup stability can guide the selection of an algorithm in a simulated trial and explored an alternative evaluation framework suited for algorithms that give estimates of the individual treatment

effect. We emphasise that in this chapter we used outcome noise as our mechanism for illustrative purposes but also due its wide applicability in real-world settings. As we discussed other mechanisms may require revisiting the definition of the stability measure and this would dependent on the specific application.

In this chapter we focused on marginally randomised studies and the subgroup identification algorithms that were discussed have been introduce in this setting. If we were to apply this framework in the case of observational studies we would first need to ensure that for the measure of quality we get unbiased estimations of the treatment effect, while we also use subgroup identification algorithms that give unbiased estimates within the subgroups. We discussed some approaches for ATE estimation in observational studies under the assumption of no hidden confounders in Chapter 2 and some approaches for subgroup identification in the previous chapter.

Chapter 7

Conclusions and Future Directions

7.1 Conclusions

In the beginning of this thesis we posed a number of questions regarding predictive covariate selection, subgroup identification in the presence of confounders and evaluation of subgroup identification algorithms. In order to tackle these questions we proposed new methods and evaluated them in a number of scenarios. Here we summarise the answers to these questions and our findings.

Q1 : "*How should we adapt information theoretic criteria to identify predictive covariates?*"

We phrased the problem of predictive covariate selection as an optimisation problem involving two log-likelihood functions – the log-likelihood of the outcome given the interaction between covariates and treatment and the log-likelihood of the outcome given only the covariates (Definition 1). Identifying the covariates that maximise the difference between the two functions corresponds to optimising an information theoretic quantity (eq. (4.1)). Borrowing concepts from the information theoretic feature selection literature we focused on low-dimensional criteria that are better suited for the small-sample scenarios we often encounter in randomised studies (Section 4.2). We then identified a limitation of the approach: it can be biased in non-marginally randomised studies, i.e. when the treatment assignment

depends on the covariates (eq. (4.5)). To this end, we proposed simple preprocessing steps that can ameliorate this issue and validated them empirically (Section 4.6). In particular we propose INFO+S, which applies INFO+ in strata with different values of the propensity score and INFO+W, which applies INFO+ in a new sample derived from over-sampling observations based on IPW.

Q2 : "*How do information theoretic criteria compare to subgroup identification approaches on the task of predictive covariate selection in marginally randomised studies and how do they perform when the treatment assignment depends on the covariates?*"

As we discussed most existing approaches do not tackle the problem of predictive covariate selection directly, but this is commonly part of their objective, which can be individual treatment effect estimation and/or subgroup identification. We compared a low dimensional information theoretic criterion (INFO+) with various methods from the literature designed for subgroup identification, namely VT (Foster et al., 2011) and SIDES (Lipkovich et al., 2011), as well as a method designed for identifying treatment-covariate interactions and modelling the individual treatment effect (Tian et al., 2014). The results show that INFO+ is a strong competitor when the covariates are categorical and is only influenced by the predictive strength (Section 4.4.2). When the covariates carry both prognostic and predictive information VT achieves the highest TPR, however it is also influenced by the prognostic strength (Section 4.4.2) and may be biased towards identifying prognostic covariates as predictive (Section 4.4.3, 4.4.4). INFO+ successfully distinguishes between prognostic and predictive covariates (Section 4.4.4) and is more computationally efficient than VT and SIDES (Section 4.4.5). The simulations that consider the presence of confounders show that INFO+ can be influenced by the treatment assignment mechanism. The proposed extensions, INFO+S and INFO+W can ameliorate these issues and in some studied scenarios they perform as if the data were from a marginally randomised study.

Q3 : "*How can we modify existing recursive partitioning approaches for subgroup identification in order to account for the presence of confounders in the data?*"

In order to answer this question we studied a methodology based on IT (Su et al., 2009) that uses non-parametric weighting estimators of the treatment effect in order to optimise clearly defined quantities (Kallus et al., 2021). In particular, this methodology uses estimators that optimise either the worst-case conditional biases of the potential outcomes (B-IT) or the worst-case conditional MSE of the sample average treatment effect (MSE-IT). This method overcomes some limitations of existing methodologies, such as those that use IPW estimators, which become particularly important in the context of subgroup identification.

Q4 : "*What are the benefits from using weighting estimators in the context of subgroup identification?*"

We validate the proposed method in simulated scenarios and a real-world study. By varying the confounding strength we show that using weighting methods can reduce the error of the estimated effects and achieve higher proportion of correct trees (Section 5.4.1). In Section 5.4.2 we explored using a simple example the properties of our modifications to IT compared to performing subgroup identification with IPW estimators. In Section 5.4.3 we studied what happens under incorrect specification of the outcome model. In the absence of subgroups, a method that uses unadjusted estimators tends to identify subgroups defined by the confounders, while weighting estimators can overcome these issues (Section 5.4.4). In Section 5.5 we firstly consider a simulated study and introduce artificially imbalance in the data. We observe that MSE-IT can identify the correct tree, while this is not the case when using unadjusted estimators. We then revisited a real-world case study and validated that similarly to previous findings, the suggested approach does not identify a subgroup.

Q5 : "*How should we evaluate subgroup identification algorithms in order to account their robustness to small changes?*"

The stability of subgroup identification algorithms refers to their ability to reproduce similar subgroups under small changes in the data. This is closely connected with the well-studied problem of feature selection stability (Nogueira et al., 2017). We show how such measures can be used for various tasks, such as hyper-parameter selection and algorithm selection. Since, we would additionally like the subgroups to show some desirable properties

such as enhanced effects, we propose a multi-objective framework and show how an algorithm can be selected using the pareto optimal solution rather than a single criterion.

Overall the thesis describes tools and methodologies for evaluating the existence of heterogeneous effects given randomised as well as observational studies with no hidden confounders. Table 7.1 summarises the results of this thesis.

	Task	What is new in this thesis?
Chapter 4	Identifying predictive covariates in potentially high-dimensional studies	Information theoretic approaches for randomised studies along with modifications that use the propensity score when the treatment assignment depends on the covariates
Chapter 5	Identifying subgroups of heterogeneous effects in the presence of confounders in the data	Modifications to a recursive partitioning approach using weighting methods for treatment effect estimation in the subgroups
Chapter 6	Evaluation of subgroup identification algorithms	A multi-objective evaluation framework that uses the stability of the selected subgroups as a measure that captures their robustness to small changes in the data

Table 7.1: Summary of topics studied in this thesis and the methodologies suggested in each chapter.

7.2 Future Work

The methods described in this thesis can be extended and further improved to handle different types of data and tasks. Also, the results of this thesis suggest a number of interesting research avenues for future work. Here we discuss some methodological improvements and potential new directions.

An interesting problem is the extension of the predictive covariate selection criterion presented in Chapter 4 in order to handle mixed data. In Chapter 4 we

discretised continuous covariates using unsupervised methods such as K-means and histogram-based approaches in order to estimate the mutual information. Information theoretic approaches, such as INFO+, can be influenced by how continuous covariates are handled, while other methods can be applied directly using all the information provided in the data. Therefore, it would be interesting to explore the role of the used estimator. Recent works have proposed estimators suited for continuous and mixed data that can outperform common baselines, such as those used in this thesis. Some examples include estimators that build on the nearest neighbour principle and extend the commonly used Kraskov-Stögbauer-Grassberger estimator (Kraskov et al., 2004; Gao et al., 2017), while others make use of kernel density estimators (Beknazaryan et al., 2019). It could be interesting to explore how these estimators could be adopted to extend INFO+ in the context of mixed covariates, continuous outcomes and treatments.

The methodology described in Chapter 5 results in multiple non-overlapping subgroups that share common covariates. In practice we may have potentially overlapping subgroups that are defined by different sets of covariates. A potential solution to handle such scenarios is to retain in each node the best split identified for a number of covariates, which can be a parameter specified by the user (Lipkovich et al., 2011). This procedure would result in multiple trees with different root nodes. Another concern comes from the multiple comparisons that may need to be performed. This also concerns the final subgroups that may be retained as discussed in the last section of Chapter 5. Additionally, as the sample size decreases (e.g. by performing multiple splits), the overlap between the treatment groups may be poor. To this end an interesting approach is explored in (Kallus and Santacatterina, 2019b), where the target population for which SATE is going to be estimated is optimised.

Another interesting extension is to modify the splitting criterion regarding the subgroup identification algorithm in order to identify subgroups with pre-defined characteristics. Dusseldorp and Van Mechelen (2014) propose a splitting criterion that identifies subgroups defined by qualitative interactions. For the two subsets derived by each possible split, the evaluation criterion is analogous to maximising the effect in one subset and minimising it in the other. The primary objective is to split in the data so that in one subgroup there is enhanced effect while in the other the opposite holds. In the initial algorithm unadjusted estimators of the treatment effect are used. To this end, we can explore the use of weighting

methods as described in Chapter 5 in order to handle the presence of confounders.

We would like to highlight that this thesis has focused on problems where there is a binary treatment. Even though this is a very common setting, extensions of the described methodology, particularly regarding the subgroup identification algorithm of Chapter 5 are not straightforward and require further research. If the treatment is categorical then the procedure might need to be repeated multiple times in order to account for the different comparisons or could require defining the treatment effect differently (Feng et al., 2012; Lopez et al., 2017). In any case, both the methodology and the weighting methods will need to be revisited. Extensions to the case of a continuous treatment is more challenging since both the recursive partitioning procedure and the estimation of the effects will need to be adjusted (Kallus and Santacatterina, 2019a; Fong et al., 2018).

Lastly, throughout this thesis we considered either marginally randomised or observational studies, where the latter were effectively treated as conditionally randomised. As we described the latter comes from making assumptions that are not guaranteed to hold in real-world scenarios. In particular, the assumption of no hidden confounders can be considered particularly strong in a real-world setting. In our context violations of this assumption can affect the results in both the predictive covariate selection problem, since we model the treatment assignment, and the subgroup identification problem, since we estimate treatment effects for the splitting criterion. There is a long literature on this problem in the context of treatment effect estimation, with different methods for assessing how unmeasured confounders can affect the conclusions such as performing sensitivity analysis (e.g. (Stürmer et al., 2005; VanderWeele and Arah, 2011; Hong et al., 2021; Kilbertus et al., 2020)). An interesting avenue for future work would be to explore how such methodologies can be adopted to validate the procedures described in this thesis.

Bibliography

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

Alaa, A. and Schaar, M. (2018). Limits of estimating heterogeneous treatment effects: Guidelines for practical algorithm design. In *International Conference on Machine Learning*, pages 129–138.

Alaa, A. and Van Der Schaar, M. (2019). Validating causal inference models via influence functions. In *International Conference on Machine Learning*, pages 191–201.

Alaa, A. M., Weisz, M., and van der Schaar, M. (2017). Deep counterfactual networks with propensity-dropout. In *ICML Workshop on Principled Approaches to Deep learning*.

Alemayehu, D., Chen, Y., and Markatou, M. (2018). A comparative study of subgroup identification methods for differential treatment effect: Performance metrics and recommendations. *Statistical methods in medical research*, 27(12):3658–3678.

Allison, P. D. (2001). *Missing Data*. Sage University Papers Series on Quantitative Applications in the Social Sciences, 07-136.

Alosh, M., Bretz, F., and Huque, M. (2014). Advanced multiplicity adjustment methods in clinical trials. *Statistics in medicine*, 33(4):693–713.

Altidor, W., Khoshgoftaar, T. M., and Napolitano, A. (2011). A noise-based stability evaluation of threshold-based feature selection techniques. In *2011 IEEE International Conference on Information Reuse & Integration*, pages 240–245. IEEE.

Anoke, S. C., Normand, S.-L., and Zigler, C. M. (2019). Approaches to treatment effect heterogeneity in the presence of confounding. *Statistics in medicine*, 38(15):2797–2815.

Athey, S. and Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360.

Athey, S. and Imbens, G. W. (2015). Machine learning methods for estimating heterogeneous causal effects. *stat*, 1050(5).

Athey, S., Imbens, G. W., and Wager, S. (2018). Approximate residual balancing: debiased inference of average treatment effects in high dimensions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(4):597–623.

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

Athey, S. and Wager, S. (2019). Estimating treatment effects with causal forests: An application. *Observational Studies*, 5(2):37–51.

Austin, P. C. (2009a). Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Statistics in medicine*, 28(25):3083–3107.

Austin, P. C. (2009b). The relative ability of different propensity score methods to balance measured covariates between treated and untreated subjects in observational studies. *Medical Decision Making*, 29(6):661–677.

Austin, P. C. and Stuart, E. A. (2015). Moving towards best practice when using inverse probability of treatment weighting (iptw) using the propensity score to estimate causal treatment effects in observational studies. *Statistics in medicine*, 34(28):3661–3679.

Ballman, K. V. (2015). Biomarker: predictive or prognostic? *Journal of clinical oncology: official journal of the American Society of Clinical Oncology*, 33(33):3968–3971.

Bang, H. and Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973.

Beknazaryan, A., Dang, X., and Sang, H. (2019). On mutual information estimation for mixed-pair random variables. *Statistics & Probability Letters*, 148:9–16.

Bennett, D. A. (2001). How can i deal with missing data in my study? *Australian and New Zealand journal of public health*, 25(5):464–469.

Beraha, M., Metelli, A. M., Papini, M., Tirinzoni, A., and Restelli, M. (2019). Feature selection via mutual information: new theoretical insights. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–9. IEEE.

Bloniarz, A., Liu, H., Zhang, C.-H., Sekhon, J. S., and Yu, B. (2016). Lasso adjustments of treatment effect estimates in randomized experiments. *Proceedings of the National Academy of Sciences*, 113(27):7383–7390.

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

Breiman, L. (2002). Manual on setting up, using, and understanding random forests v3. 1. *Statistics Department University of California Berkeley, CA, USA*, 1:58.

Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. (1984). *Classification and regression trees*. CRC press.

Brookhart, M. A., Schneeweiss, S., Rothman, K. J., Glynn, R. J., Avorn, J., and Stürmer, T. (2006). Variable selection for propensity score models. *American journal of epidemiology*, 163(12):1149–1156.

Brown, G., Pocock, A., Zhao, M.-J., and Luján, M. (2012). Conditional likelihood maximisation: a unifying framework for information theoretic feature selection. *Journal of machine learning research*, 13(Jan):27–66.

Calhoun, P., Su, X., Nunn, M., and Fan, J. (2018). Constructing multivariate survival trees: the mst package for r. *Journal of Statistical Software*, 83(12).

Chen, S., Tian, L., Cai, T., and Yu, M. (2017). A general statistical framework for subgroup identification and comparative treatment scoring. *Biometrics*, 73(4):1199–1209.

Cole, S. R. and Hernán, M. A. (2008). Constructing inverse probability weights for marginal structural models. *American journal of epidemiology*, 168(6):656–664.

Connors, A. F., Speroff, T., Dawson, N. V., Thomas, C., Harrell, F. E., Wagner, D., Desbiens, N., Goldman, L., Wu, A. W., Califf, R. M., et al. (1996). The effectiveness of right heart catheterization in the initial care of critically iii patients. *Jama*, 276(11):889–897.

Cover, T. M. and Thomas, J. A. (2012). *Elements of information theory*. John Wiley & Sons.

De Luna, X., Waernbaum, I., and Richardson, T. S. (2011). Covariate selection for the nonparametric estimation of an average treatment effect. *Biometrika*, 98(4):861–875.

Dmitrienko, A., Millen, B., and Lipkovich, I. (2017). Multiplicity considerations in subgroup analysis. *Statistics in Medicine*, 36(28):4446–4454.

Dmitrienko, A., Muysers, C., Fritsch, A., and Lipkovich, I. (2016). General guidance on exploratory and confirmatory subgroup analysis in late-stage clinical trials. *Journal of biopharmaceutical statistics*, 26(1):71–98.

Dougherty, J., Kohavi, R., and Sahami, M. (1995). Supervised and unsupervised discretization of continuous features. In *Machine learning proceedings 1995*, pages 194–202. Elsevier.

Dunn, G., Emsley, R., Liu, H., and Landau, S. (2013). Integrating biomarker information within trials to evaluate treatment mechanisms and efficacy for personalised medicine. *Clinical Trials*, 10(5):709–719.

Dusseldorp, E. and Van Mechelen, I. (2014). Qualitative interaction trees: a tool to identify qualitative treatment–subgroup interactions. *Statistics in medicine*, 33(2):219–237.

Efron, B. (1983). Estimating the error rate of a prediction rule: improvement on cross-validation. *Journal of the American statistical association*, 78(382):316–331.

Fellström, B. C., Jardine, A. G., Schmieder, R. E., Holdaas, H., Bannister, K., Beutler, J., Chae, D.-W., Chevaile, A., Cobbe, S. M., Grönhagen-Riska, C., et al. (2009). Rosuvastatin and cardiovascular events in patients undergoing hemodialysis. *New England Journal of Medicine*, 360(14):1395–1407.

Felson, D. T., Anderson, J. J., Boers, M., Bombardier, C., Chernoff, M., Fried, B., Furst, D., Goldsmith, C., Kieszak, S., Lightfoot, R., et al. (1993). The american college of rheumatology preliminary core set of disease activity measures for rheumatoid arthritis clinical trials. *Arthritis & Rheumatism: Official Journal of the American College of Rheumatology*, 36(6):729–740.

Feng, P., Zhou, X.-H., Zou, Q.-M., Fan, M.-Y., and Li, X.-S. (2012). Generalized propensity score for estimating the average treatment effect of multiple treatments. *Statistics in medicine*, 31(7):681–697.

Fisher, R. A. (1937). *The design of experiments*. Oliver And Boyd; Edinburgh; London.

Fong, C., Hazlett, C., Imai, K., et al. (2018). Covariate balancing propensity score for a continuous treatment: Application to the efficacy of political advertisements. *The Annals of Applied Statistics*, 12(1):156–177.

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

François, D., Rossi, F., Wertz, V., and Verleysen, M. (2007). Resampling methods for parameter-free and robust feature selection with mutual information. *Neurocomputing*, 70(7-9):1276–1288.

Freedman, D. A. (2006). On the so-called huber sandwich estimator and robust standard errors. *The American Statistician*, 60(4):299–302.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1.

Friedman, J. H. and Fisher, N. I. (1999). Bump hunting in high-dimensional data. *Statistics and Computing*, 9(2):123–143.

Frölich, M. (2004). Programme evaluation with multiple treatments. *Journal of Economic Surveys*, 18(2):181–224.

Fukuoka, M., Wu, Y.-L., Thongprasert, S., Sunpaweravong, P., Leong, S.-S., Sriuranpong, V., Chao, T.-Y., Nakagawa, K., Chu, D.-T., Saijo, N., Duffield, E. L.,

Rukazenkov, Y., Speake, G., Jiang, H., Armour, A. A., To, K.-F., Yang, J. C.-H., and Mok, T. S. (2011). Biomarker analyses and final overall survival results from a phase III, randomized, open-label, first-line study of gefitinib versus carboplatin/paclitaxel in clinically selected patients with advanced non–small-cell lung cancer in Asia (IPASS). *Journal of Clinical Oncology*, 29(21):2866–2874.

Funk, M. J., Westreich, D., Wiesen, C., Stürmer, T., Brookhart, M. A., and Davidian, M. (2011). Doubly robust estimation of causal effects. *American journal of epidemiology*, 173(7):761–767.

Gao, W., Kannan, S., Oh, S., and Viswanath, P. (2017). Estimating mutual information for discrete-continuous mixtures. In *Advances in neural information processing systems*, pages 5986–5997.

Gharibzadeh, S., Mansournia, M. A., Rahimiforoushani, A., Alizadeh, A., Amouzegar, A., Mehrabani-Zeinabad, K., and Mohammad, K. (2018). Comparing different propensity score estimation methods for estimating the marginal causal effect through standardization to propensity scores. *Communications in Statistics-Simulation and Computation*, 47(4):964–976.

Gocht, A., Lehmann, C., and Schöne, R. (2018). A new approach for automated feature selection. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 4915–4920. IEEE.

Guyon, I., Gunn, S., Nikravesh, M., and Zadeh, L. A. (2008). *Feature extraction: foundations and applications*, volume 207. Springer.

Häggström, J. (2018). Data-driven confounder selection via markov and bayesian networks. *Biometrics*, 74(2):389–398.

Hainmueller, J. (2012). Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis*, 20(1):25–46.

Harrell Jr, F. E., Lee, K. L., and Mark, D. B. (1996). Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in medicine*, 15(4):361–387.

Hartford, J., Lewis, G., Leyton-Brown, K., and Taddy, M. (2017). Deep iv: A flexible approach for counterfactual prediction. In *Proceedings of the 34th*

International Conference on Machine Learning-Volume 70, pages 1414–1423. JMLR. org.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media.

Hausser, J. and Strimmer, K. (2009). Entropy inference and the james-stein estimator, with application to nonlinear gene association networks. *Journal of Machine Learning Research*, 10(Jul):1469–1484.

Hazlett, C. (2020). Kernel balancing: A flexible non-parametric weighting procedure for estimating causal effects. *Statistica Sinica*, pages 1155–1189.

Heckman, J. J., Ichimura, H., and Todd, P. E. (1997). Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *The review of economic studies*, 64(4):605–654.

Hernán, M. and Robins, J. (2020). *Causal Inference: What If*. Boca Raton: Chapman & Hall/CRC.

Hirano, K. and Imbens, G. W. (2001). Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization. *Health Services and Outcomes research methodology*, 2(3-4):259–278.

Hirano, K., Imbens, G. W., and Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4):1161–1189.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

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

Hong, G., Yang, F., and Qin, X. (2021). Did you conduct a sensitivity analysis? a new weighting-based approach for evaluations of the average treatment effect for the treated. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 184(1):227–254.

Huang, X., Sun, Y., Trow, P., Chatterjee, S., Chakravartty, A., Tian, L., and Devanarayan, V. (2017). Patient subgroup identification for clinical drug development. *Statistics in medicine*, 36(9):1414–1428.

Huling, J. D. and Yu, M. (2018). Subgroup identification using the personalized package. *arXiv preprint arXiv:1809.07905*.

Imai, K. and Ratkovic, M. (2014). Covariate balancing propensity score. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):243–263.

Imai, K., Ratkovic, M., et al. (2013). Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics*, 7(1):443–470.

Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and statistics*, 86(1):4–29.

Imbens, G. W. and Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Imbens, G. W. and Wooldridge, J. M. (2009). Recent developments in the econometrics of program evaluation. *Journal of economic literature*, 47(1):5–86.

Italiano, A. (2011). Prognostic or predictive? its time to get back to definitions. *J Clin Oncol*, 29(35):4718.

Jaskowski, M. and Jaroszewicz, S. (2012). Uplift modeling for clinical trial data. In *ICML Workshop on Clinical Data Analysis*.

Johansson, F., Shalit, U., and Sontag, D. (2016). Learning representations for counterfactual inference. In *International Conference on Machine Learning*, pages 3020–3029.

Kallus, N. (2019). Classifying treatment responders under causal effect monotonicity. In *International Conference on Machine Learning*, pages 3201–3210.

Kallus, N. (2020a). Deepmatch: Balancing deep covariate representations for causal inference using adversarial training. In *International Conference on Machine Learning*, pages 5067–5077. PMLR.

Kallus, N. (2020b). Generalized optimal matching methods for causal inference. *Journal of Machine Learning Research*, 21(62):1–54.

Kallus, N., Pennicooke, B., and Santacatterina, M. (2021). More robust estimation of average treatment effects using kernel optimal matching in an observational study of spine surgical interventions. *Statistics in Medicine*, 40(10):2305–2320.

Kallus, N. and Santacatterina, M. (2019a). Kernel optimal orthogonality weighting: A balancing approach to estimating effects of continuous treatments. *arXiv preprint arXiv:1910.11972*.

Kallus, N. and Santacatterina, M. (2019b). Optimal estimation of generalized average treatment effects using kernel optimal matching. *arXiv preprint arXiv:1908.04748*.

Kalousis, A., Prados, J., and Hilario, M. (2007). Stability of feature selection algorithms: a study on high-dimensional spaces. *Knowledge and information systems*, 12(1):95–116.

Kang, J. D., Schafer, J. L., et al. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical science*, 22(4):523–539.

Kilbertus, N., Ball, P. J., Kusner, M. J., Weller, A., and Silva, R. (2020). The sensitivity of counterfactual fairness to unmeasured confounding. In *Uncertainty in Artificial Intelligence*, pages 616–626. PMLR.

Koch, B., Vock, D. M., and Wolfson, J. (2018). Covariate selection with group lasso and doubly robust estimation of causal effects. *Biometrics*, 74(1):8–17.

Kraskov, A., Stögbauer, H., and Grassberger, P. (2004). Estimating mutual information. *Physical review E*, 69(6):066138.

Kuang, K., Cui, P., Li, B., Jiang, M., Wang, Y., Wu, F., and Yang, S. (2019). Treatment effect estimation via differentiated confounder balancing and regression. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 14(1):1–25.

Kuncheva, L. I. (2007). A stability index for feature selection. In *Artificial intelligence and applications*, pages 421–427.

Künzel, S. R., Sekhon, J. S., Bickel, P. J., and Yu, B. (2019). Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, 116(10):4156–4165.

Künzel, S. R., Stadie, B. C., Vemuri, N., Ramakrishnan, V., Sekhon, J. S., and Abbeel, P. (2018). Transfer learning for estimating causal effects using neural networks. *arXiv preprint arXiv:1808.07804*.

Lee, B. K., Lessler, J., and Stuart, E. A. (2011). Weight trimming and propensity score weighting. *PloS one*, 6(3):e18174.

Lewis, D. D. (1992). Feature selection and feature extract ion for text categorization. In *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*.

Li, F., Morgan, K. L., and Zaslavsky, A. M. (2018). Balancing covariates via propensity score weighting. *Journal of the American Statistical Association*, 113(521):390–400.

Li, L. and Greene, T. (2013). A weighting analogue to pair matching in propensity score analysis. *The international journal of biostatistics*, 9(2):215–234.

Li, S. and Fu, Y. (2017). Matching on balanced nonlinear representations for treatment effects estimation. In *Advances in Neural Information Processing Systems*, pages 929–939.

Liaw, A., Wiener, M., et al. (2002). Classification and regression by randomforest. *R news*, 2(3):18–22.

Lipkovich, I. and Dmitrienko, A. (2014a). Biomarker identification in clinical trials. *Clinical and Statistical Considerations in Personalized Medicine*, pages 211–264.

Lipkovich, I. and Dmitrienko, A. (2014b). Strategies for identifying predictive biomarkers and subgroups with enhanced treatment effect in clinical trials using sides. *Journal of biopharmaceutical statistics*, 24(1):130–153.

Lipkovich, I., Dmitrienko, A., Denne, J., and Enas, G. (2011). Subgroup identification based on differential effect search - a recursive partitioning method for establishing response to treatment in patient subpopulations. *Statistics in medicine*, 30(21):2601–2621.

Lipkovich, I., Dmitrienko, A., et al. (2017a). Tutorial in biostatistics: data-driven subgroup identification and analysis in clinical trials. *Statistics in medicine*, 36(1):136–196.

Lipkovich, I., Dmitrienko, A., Patra, K., Ratitch, B., and Pulkstenis, E. (2017b). Subgroup identification in clinical trials by stochastic sidescreen methods. *Statistics in Biopharmaceutical Research*, 9(4):368–378.

Lipkovich, I., Dmitrienko, A., and Ratitch, B. (2019). Statistical methods for biomarker and subgroup evaluation in oncology trials. In Halabi, S. and Michiels, S., editors, *Textbook of Clinical Trials in Oncology: A Statistical Perspective (1st ed.)*, chapter 16, pages 317–346. Chapman and Hall/CRC.

Little, R. J. and Rubin, D. B. (2002). Complete-case and available-case analysis, including weighting methods. *Statistical Analysis with Missing Data*, pages 41–58.

Loh, W.-Y., Cao, L., and Zhou, P. (2019). Subgroup identification for precision medicine: A comparative review of 13 methods. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9(5):e1326.

Lopez, M. J., Gutman, R., et al. (2017). Estimation of causal effects with multiple treatments: a review and new ideas. *Statistical Science*, 32(3):432–454.

Louizos, C., Shalit, U., Mooij, J. M., Sontag, D., Zemel, R., and Welling, M. (2017). Causal effect inference with deep latent-variable models. In *Advances in Neural Information Processing Systems*, pages 6446–6456.

Lunceford, J. K. and Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in medicine*, 23(19):2937–2960.

Makar, M., Swaminathan, A., and Kıcıman, E. (2019). A distillation approach to data efficient individual treatment effect estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 4544–4551.

McCaffrey, D. F., Griffin, B. A., Almirall, D., Slaughter, M. E., Ramchand, R., and Burgette, L. F. (2013). A tutorial on propensity score estimation for multiple treatments using generalized boosted models. *Statistics in medicine*, 32(19):3388–3414.

McCaffrey, D. F., Ridgeway, G., and Morral, A. R. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological methods*, 9(4):403.

McGowan, L. E. D. (2018). *Improving Modern Techniques of Causal Inference: Finite Sample Performance of ATM and ATO Doubly Robust Estimators, Variance Estimation for ATO Estimators, and Contextualized Tipping Point Sensitivity Analyses for Unmeasured Confounding*. Vanderbilt University.

Mok, T. S., Wu, Y.-L., Thongprasert, S., Yang, C.-H., Chu, D.-T., Saijo, N., Sunpaweravong, P., Han, B., Margono, B., Ichinose, Y., Nishiwaki, Y., Ohe, Y., Yang, J.-J., Chewaskulyong, B., Jiang, H., Duffield, E. L., Watkins, C. L., Armour, A. A., and Fukuoka, M. (2009). Gefitinib or Carboplatin/Paclitaxel in Pulmonary Adenocarcinoma. *New England Journal of Medicine*, 361(10):947–957.

Nemenman, I., Shafee, F., and Bialek, W. (2002). Entropy and inference, revisited. In *Advances in neural information processing systems*, pages 471–478.

Neyman, J. S. (1923). On the application of probability theory to agricultural experiments. essay on principles. section 9.(tlanslated and edited by dm dabrowska and tp speed, statistical science (1990), 5, 465-480). *Annals of Agricultural Sciences*, 10:1–51.

Nie, X. and Wager, S. (2021). Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, 108(2):299–319.

Ning, Y., Sida, P., and Imai, K. (2020). Robust estimation of causal effects via a high-dimensional covariate balancing propensity score. *Biometrika*, 107(3):533–554.

Nogueira, S., Sechidis, K., and Brown, G. (2017). On the stability of feature selection algorithms. *The Journal of Machine Learning Research*, 18(1):6345–6398.

Paninski, L. (2003). Estimation of entropy and mutual information. *Neural computation*, 15(6):1191–1253.

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, 82(4):669–688.

Pedersen, A. B., Mikkelsen, E. M., Cronin-Fenton, D., Kristensen, N. R., Pham, T. M., Pedersen, L., and Petersen, I. (2017). Missing data and multiple imputation in clinical epidemiological research. *Clinical epidemiology*, 9:157.

Robins, J. M., Hernán, M. A., and Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5):550–560.

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

Ruberg, S. J. and Shen, L. (2015). Personalized medicine: four perspectives of tailored medicine. *Statistics in Biopharmaceutical Research*, 7(3):214–229.

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

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

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *The Annals of statistics*, pages 34–58.

Rubin, D. B. (1990). Formal mode of statistical inference for causal effects. *Journal of statistical planning and inference*, 25(3):279–292.

Samuels, L. R. (2017). *Aspects of Causal Inference within the Evenly Matchable Population: The Average Treatment Effect on the Evenly Matchable Units, Visually Guided Cohort Selection, and Bagged One-to-One Matching*. Vanderbilt University.

Sandri, M. and Zuccolotto, P. (2008). A bias correction algorithm for the gini variable importance measure in classification trees. *Journal of Computational and Graphical Statistics*, 17(3):611–628.

Schneider, A., Jardine, A. G., Schneider, M. P., Holdaas, H., Holme, I., Fellstroem, B. C., Zannad, F., Schmieder, R. E., Group, A. S., et al. (2013). Determinants of cardiovascular risk in haemodialysis patients: post hoc analyses of the aurora study. *American journal of nephrology*, 37(2):144–151.

Schuler, A., Baiocchi, M., Tibshirani, R., and Shah, N. (2018). A comparison of methods for model selection when estimating individual treatment effects. *arXiv preprint arXiv:1804.05146*.

Schuler, M. S. and Rose, S. (2017). Targeted maximum likelihood estimation for causal inference in observational studies. *American journal of epidemiology*, 185(1):65–73.

Scott, D. W. (1992). *Multivariate density estimation: theory, practice, and visualization*. John Wiley & Sons.

Sechidis, K., Azzimonti, L., Pocock, A., Corani, G., Weatherall, J., and Brown, G. (2019a). Efficient feature selection using shrinkage estimators. *Machine Learning*, 108(8-9):1261–1286.

Sechidis, K. and Brown, G. (2018). Simple strategies for semi-supervised feature selection. *Machine Learning*, 107(2):357–395.

Sechidis, K., Papangelou, K., Metcalfe, P. D., Svensson, D., Weatherall, J., and Brown, G. (2018). Distinguishing prognostic and predictive biomarkers: an information theoretic approach. *Bioinformatics*, 34(19):3365–3376. doi: `10.1093/bioinformatics/bty357`.

Sechidis, K., Papangelou, K., Nogueira, S., Weatherall, J., and Brown, G. (2019b). On the stability of feature selection in the presence of feature correlations. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 327–342. Springer.

Shalit, U., Johansson, F. D., and Sontag, D. (2017). Estimating individual treatment effect: generalization bounds and algorithms. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3076–3085. JMLR. org.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell system technical journal*, 27(3):379–423.

Shi, C., Blei, D., and Veitch, V. (2019). Adapting neural networks for the estimation of treatment effects. In *Advances in neural information processing systems*, pages 2507–2517.

Simon, R. (2010). Clinical trials for predictive medicine: new challenges and paradigms. *Clinical trials*, 7(5):516–524.

Steingrimsson, J. A., Hanley, D. F., and Rosenblum, M. (2017). Improving precision by adjusting for prognostic baseline variables in randomized trials with binary outcomes, without regression model assumptions. *Contemporary clinical trials*, 54:18–24.

Steingrimsson, J. A. and Yang, J. (2019). Subgroup identification using covariate-adjusted interaction trees. *Statistics in medicine*, 38(21):3974–3984.

Stürmer, T., Schneeweiss, S., Avorn, J., and Glynn, R. J. (2005). Adjusting effect estimates for unmeasured confounding with validation data using propensity score calibration. *American journal of epidemiology*, 162(3):279–289.

Su, X., Tsai, C.-L., Wang, H., Nickerson, D. M., and Li, B. (2009). Subgroup analysis via recursive partitioning. *Journal of Machine Learning Research*, 10(Feb):141–158.

Su, X., Zhou, T., Yan, X., Fan, J., and Yang, S. (2008). Interaction trees with censored survival data. *The international journal of biostatistics*, 4(1).

Tao, Y. and Fu, H. (2019). Doubly robust estimation of the weighted average treatment effect for a target population. *Statistics in medicine*, 38(3):315–325.

Tian, L., Alizadeh, A. A., Gentles, A. J., and Tibshirani, R. (2014). A simple method for estimating interactions between a treatment and a large number of covariates. *Journal of the American Statistical Association*, 109(508):1517–1532.

Tibshirani, J., Athey, S., Friedberg, R., Hadad, V., Hirshberg, D., Miner, L., Sverdrup, E., Wager, S., and Wright, M. (2020). *grf: Generalized Random Forests*. R package version 1.2.0.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288.

VanderWeele, T. J. and Arah, O. A. (2011). Bias formulas for sensitivity analysis of unmeasured confounding for general outcomes, treatments, and confounders. *Epidemiology*, pages 42–52.

VanderWeele, T. J. and Shpitser, I. (2011). A new criterion for confounder selection. *Biometrics*, 67(4):1406–1413.

Vergara, J. R. and Estévez, P. A. (2014). A review of feature selection methods based on mutual information. *Neural computing and applications*, 24(1):175–186.

Vieille, F. (2018). *aVirtualTwins: Adaptation of Virtual Twins Method from Jared Foster*. R package version 1.0.1.

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

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

Wald, R., Khoshgoftaar, T. M., and Shanab, A. A. (2012). The effect of measurement approach and noise level on gene selection stability. In *2012 IEEE International Conference on Bioinformatics and Biomedicine*, pages 1–5. IEEE.

Wang, A., Nianogo, R. A., and Arah, O. A. (2017). G-computation of average treatment effects on the treated and the untreated. *BMC medical research methodology*, 17(1):1–5.

Wang, T. and Rudin, C. (2017). Causal rule sets for identifying subgroups with enhanced treatment effect. *arXiv preprint arXiv:1710.05426*.

Westreich, D., Cole, S. R., Funk, M. J., Brookhart, M. A., and Stürmer, T. (2011). The role of the c-statistic in variable selection for propensity score models. *Pharmacoepidemiology and drug safety*, 20(3):317–320.

Williams, C. K. and Rasmussen, C. E. (2006). *Gaussian processes for machine learning*. MIT press Cambridge, MA.

Williamson, E. J., Forbes, A., and White, I. R. (2014). Variance reduction in randomised trials by inverse probability weighting using the propensity score. *Statistics in medicine*, 33(5):721–737.

Xie, Y., Zhu, Y., Cotton, C. A., and Wu, P. (2019). A model averaging approach for estimating propensity scores by optimizing balance. *Statistical methods in medical research*, 28(1):84–101.

Yang, H. H. and Moody, J. (2000). Data visualization and feature selection: New algorithms for nongaussian data. In *Advances in neural information processing systems*, pages 687–693.

Yang, J., Dahabreh, I. J., and Steingrimsson, J. A. (2021). Causal interaction trees: Finding subgroups with heterogeneous treatment effects in observational data. *Biometrics*.

Yao, L., Li, S., Li, Y., Huai, M., Gao, J., and Zhang, A. (2018). Representation learning for treatment effect estimation from observational data. In *Advances in Neural Information Processing Systems*, pages 2633–2643.

Yoon, J., Jordon, J., and van der Schaar, M. (2018). Ganite: Estimation of individualized treatment effects using generative adversarial nets. In *International Conference on Learning Representations*.

Yu, S. and Príncipe, J. C. (2019). Simple stopping criteria for information theoretic feature selection. *Entropy*, 21(1):99.

Zeileis, A. (2004). Econometric computing with hc and hac covariance matrix estimators. *Journal of Statistical Software, Articles*, 11(10):1–17.

Zeileis, A. (2006). Object-oriented computation of sandwich estimators. *Journal of Statistical Software, Articles*, 16(9):1–16.

Zhang, M., Tsiatis, A. A., and Davidian, M. (2008). Improving efficiency of inferences in randomized clinical trials using auxiliary covariates. *Biometrics*, 64(3):707–715.

Zubizarreta, J. R. (2015). Stable weights that balance covariates for estimation with incomplete outcome data. *Journal of the American Statistical Association*, 110(511):910–922.

Appendix A

Supplementary Material

A.1 Proof of Lemma 1

Lemma 1. *In marginally randomised experiments and in the absence of treatment effect $J_{Pred\text{-}CMI}$ becomes independent of the covariates.*

Proof. Under the assumption of no treatment effect we have $p(y \mid \boldsymbol{x}, t = 1) = p(y \mid \boldsymbol{x}, t = 0) = p(y \mid \boldsymbol{x}),\ \forall \boldsymbol{x}$. The following holds.

$$
\begin{aligned}
&\mathrm{I}(T; Y \mid \boldsymbol{X}) = \\
&= \sum_{\mathcal{X}} \sum_{\mathcal{Y}} \sum_{\mathcal{T}} p(y, \boldsymbol{x}, t) \log \frac{p(y, t \mid \boldsymbol{x})}{p(t \mid \boldsymbol{x}) p(y \mid \boldsymbol{x})} = \\
&= \sum_{\mathcal{X}} \sum_{\mathcal{Y}} \sum_{\mathcal{T}} p(y, \boldsymbol{x}, t) \log \frac{p(y \mid t, \boldsymbol{x})}{p(y \mid \boldsymbol{x})} = \\
&= \sum_{\mathcal{X}} \sum_{\mathcal{Y}} \sum_{\mathcal{T}} p(y, \boldsymbol{x}, t) \log\, p(y \mid t, \boldsymbol{x}) - \sum_{\mathcal{X}} \sum_{\mathcal{Y}} \sum_{\mathcal{T}} p(y, \boldsymbol{x}, t) \log\, p(y \mid \boldsymbol{x}) = \\
&= \sum_{\mathcal{X}} \sum_{\mathcal{Y}} p(y \mid \boldsymbol{x}, t = 1) p(t = 1 \mid \boldsymbol{x}) p(\boldsymbol{x}) \log\, p(y \mid t = 1, \boldsymbol{x}) + \\
&+ \sum_{\mathcal{X}} \sum_{\mathcal{Y}} p(y \mid \boldsymbol{x}, t = 0) p(t = 0 \mid \boldsymbol{x}) p(\boldsymbol{x}) \log\, p(y \mid t = 0, \boldsymbol{x}) - \\
&- \sum_{\mathcal{X}} \sum_{\mathcal{Y}} p(y, \boldsymbol{x}) \log\, p(y \mid \boldsymbol{x}) = \\
&= p(t = 1) \sum_{\mathcal{X}} \sum_{\mathcal{Y}} p(y \mid \boldsymbol{x}, t = 1) p(\boldsymbol{x}) \log\, p(y \mid t = 1, \boldsymbol{x}) + \\
&+ \big(1 - p(t = 1)\big) \sum_{\mathcal{X}} \sum_{\mathcal{Y}} p(y \mid \boldsymbol{x}, t = 0) p(\boldsymbol{x}) \log\, p(y \mid t = 0, \boldsymbol{x}) -
\end{aligned}
$$

$$
\begin{aligned}
&-\sum_{\mathcal{X}}\sum_{\mathcal{Y}} p(y \mid \boldsymbol{x})p(\boldsymbol{x})\log\, p(y \mid \boldsymbol{x}) = \\
&= p(t=1)\sum_{\mathcal{X}}\sum_{\mathcal{Y}} p(y \mid \boldsymbol{x}, t=1)p(\boldsymbol{x})\log\, p(y \mid t=1, \boldsymbol{x}) + \\
&+ \big(1 - p(t=1)\big)\sum_{\mathcal{X}}\sum_{\mathcal{Y}} p(y \mid \boldsymbol{x}, t=1)p(\boldsymbol{x})\log\, p(y \mid t=1, \boldsymbol{x}) - \\
&- \sum_{\mathcal{X}}\sum_{\mathcal{Y}} p(y \mid t=1, \boldsymbol{x})p(\boldsymbol{x})\log\, p(y \mid t=1, \boldsymbol{x}) = 0
\end{aligned}
$$

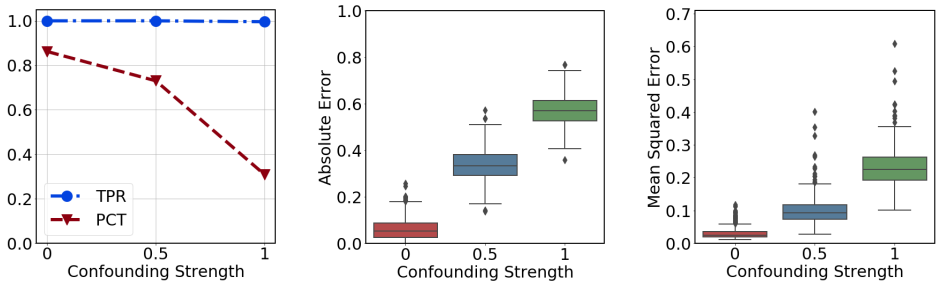
where the fifth equality follows from randomisation, $T \perp\!\!\!\perp \boldsymbol{X}$ and the last equality from the assumption of no treatment effect. □

A.2 Varying the Confounding Strength: Larger Subgroups

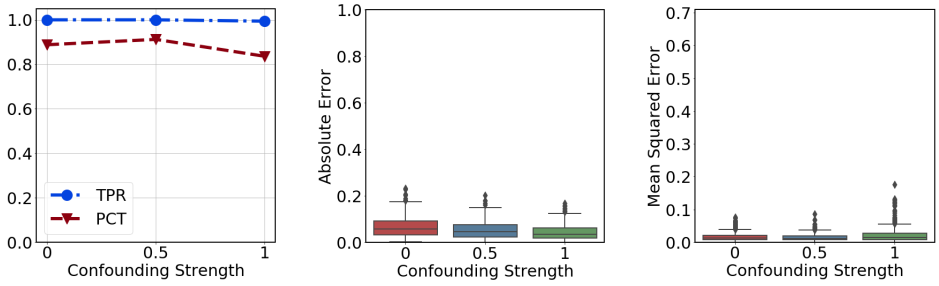
We repeat the experiment of section 5.4.1 generating data according to the outcome function (Foster et al., 2011):

$$Y = -1 + 0.5(X_1 + X_2 - X_3 + X_2X_3) + T(0.1 + 0.9\mathbb{1}(X_1 > -0.545 \cap X_2 < 0.545)$$

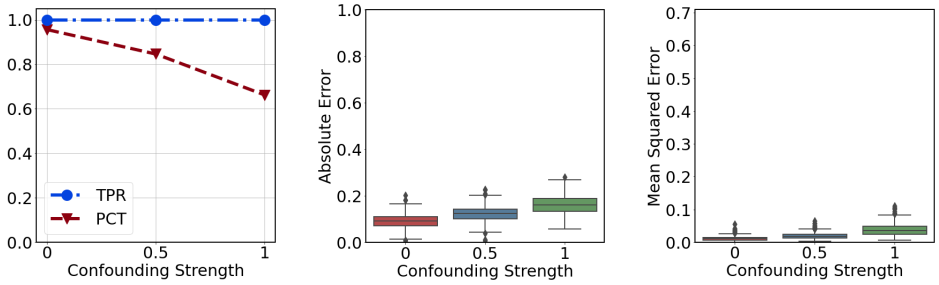
This defines a subgroup with effect equal to 1 and larger size compared to the problem studied in Chapter 5. Here the subgroup corresponds to $\sim 50\%$ of the data. The treatment is generated according to the same model. In figure A.1 we observe in general similar findings with the main document. However, we also notice that in this case using the augmented estimator tends to provide worse results for PCT and increased error for $\gamma = 1$ (particularly when using B-IT). We note again that both the weights and the outcome were generated under an incorrectly specified model since it does not include the interaction between X_2 and X_3 as well as functions of the covariates (in our case indicators). The weights may balance the means of the covariates, but as has been described this will not necessarily hold for subsets of the data. From the results it appears that miss-specification may not be the only reason, but also the definition of PCT. In particular, we found that often these methods tend to identify meaningful subgroups (close to the true one) but these are defined using more than two splits, hence they are not considered by the definition of the metric as correct.



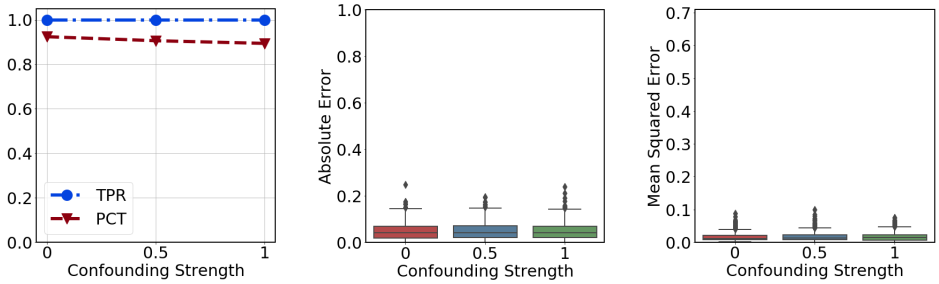
(a) IT



(b) B-IT

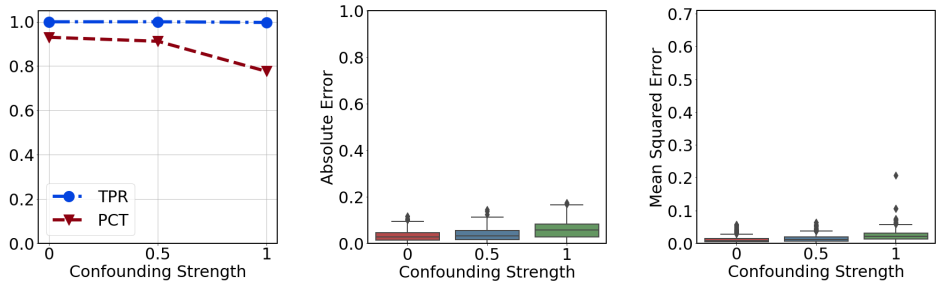


(c) B-IT (augemnted)



(d) MSE-IT

Figure A.1: PCT and TPR (first column), absolute error in the subgroup (second column) and MSE of the estimated effect in a separate test set (third column) using IT and the proposed alternatives

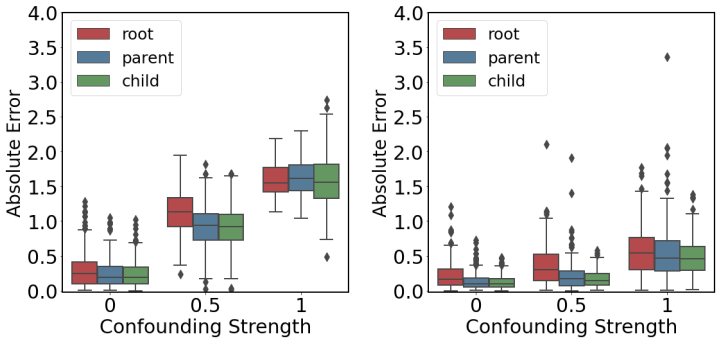


(e) MSE-IT (augmented)

Figure A.1: PCT and TPR (first column), absolute error in the subgroup (second column) and MSE of the estimated effect in a separate test set (third column) using IT and the proposed alternatives *(cont.)*

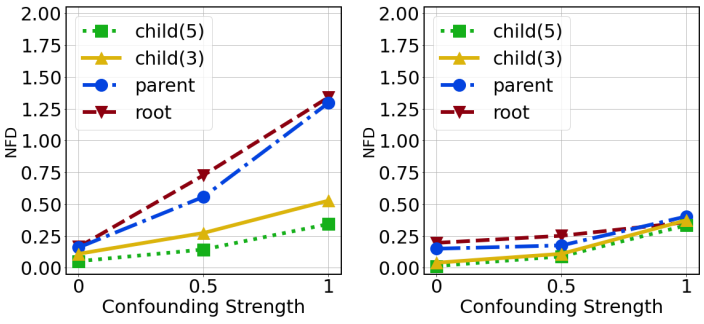
A.3 Varying the Outcome Specification: Normally Distributed Covariates

We repeat the experiment of Section 5.4.3 but instead we use normally distributed covariates. In this case the number of possible splits for each covariate is equal to the sample size of each node, hence optimising the weights for each possible split would be computationally expensive. Instead we optimise the weights using 3 and 5 equally spaced cut-off values for each covariate. Then for each possible split of a covariate we use the estimated weights derived for the nearest cut-off. In this way the covariates may not be matched in their means, but at least approximately we can achieve balance between samples with similar values of the covariates. The PCT is reported in Chapter 5. For completeness here we report the absolute error in the subgroup in figure A.2 (for the case of using 5 splits) and the average number of false discoveries in figure A.3.



(a) MSE-IT ($d = 1$) (b) MSE-IT ($d = 2$)

Figure A.2: Absolute error resulting by estimating the weights in the root node, parent node or for each split.



(a) MSE-IT ($d = 1$) (b) MSE-IT ($d = 2$)

Figure A.3: Number of false discoveries averaged over the number of simulations.