UNIVERSITY OF OKLAHOMA
GRADUATE COLLEGE

IMPROVING OUTCOMES IN MACHINE LEARNING AND DATA-DRIVEN LEARNING SYSTEMS USING STRUCTURAL CAUSAL MODELS

A DISSERTATION
SUBMITTED TO THE GRADUATE FACULTY
in partial fulfillment of the requirements for the
Degree of
DOCTOR OF PHILOSOPHY

By

HENRY MADUKA MBOGU
Norman, Oklahoma
2023

IMPROVING OUTCOMES IN MACHINE LEARNING AND DATA-DRIVEN LEARNING SYSTEMS USING STRUCTURAL CAUSAL MODELS

A DISSERTATION APPROVED FOR THE
GALLOGLY COLLEGE OF ENGINEERING

BY THE COMMITTEE CONSISTING OF

Dr. Charles D. Nicholson, Chair

Dr. Randa Shehab

Dr. Naveen Kumar

Dr. Heather Bedle

Acknowledgments

The successful completion of this work would not have been possible without the invaluable contributions and support of the members of my doctoral committee. Firstly, I extend my heartfelt gratitude to my doctoral advisor and committee chair, Dr. Charles Nicholson, for not only accepting me under his guidance but also for steadfastly leading me through the completion of my dissertation. Our collaboration has been immensely fruitful, and I grew a lot as a researcher thanks to his guidance. Secondly, I am profoundly grateful to Dr. Randa Shehab for her unwavering support throughout this journey. She has been a great mentor and sponsor, and her belief in my abilities has been instrumental in my progress. Thirdly, I extend my sincere appreciation to Dr. Naveen Kumar for his pivotal role during the early stages of my dissertation and for his continued support all the way through to the completion of my doctoral program. Lastly, I would like to thank Dr. Heather Bedle for her watchful support and genuine interest in my work.

I want to express my gratitude to my wife, Liz, for her love and support as I navigated this important stage in my life. I am thankful to my mother, Florence, for her prayers and encouragement, and for being my ultimate cheerleader throughout this journey. To my siblings and the entire extended family, I express my heartfelt thanks for always having my back. Many of you have played important roles, directly and indirectly at various stages, contributing significantly to bring me to this very moment. Your support means the world to me.

Table of Contents

List Of Tables

List Of Figures

Abstract

The field of causal inference has experienced rapid growth and development in recent years. Its significance in addressing a diverse array of problems and its relevance across various research and application domains are increasingly being acknowledged. However, the current state-of-the-art approaches to causal inference have not yet gained widespread adoption in mainstream data science practices.

This research endeavor begins by seeking to motivate enthusiasm for contemporary approaches to causal investigation utilizing observational data. It explores the existing applications and potential future prospects for employing causal inference methods to enhance desired outcomes in data-driven learning applications across various domains, with a particular focus on their relevance in artificial intelligence (AI). Following this motivation, this dissertation proceeds to offer a broad review of fundamental concepts, theoretical frameworks, methodological advancements, and existing techniques pertaining to causal inference.

The research advances by investigating the problem of data-driven root cause analysis through the lens of causal structure modeling. Data-driven approaches to root cause analysis (RCA) have received attention recently due to their ability to exploit increasing data availability for more effective root cause identification in complex processes. Advancements in the field of causal inference enable unbiased causal investigations using observational data. This study proposes a data-driven RCA method and a time-to-event (TTE) data simulation procedure built on the structural causal model (SCM) framework. A novel causality-based method is introduced for learning a representation of root cause mechanisms, termed in this work as *root cause graphs (RCGs)*, from observational TTE data. Three case scenarios are used to generate TTE datasets for evaluating the proposed method. The utility of the proposed RCG recovery method is demonstrated by using recovered RCGs to guide the estimation of root cause treatment effects. In the presence of mediation, RCG-guided models produce superior estimates of root cause total effects compared to models that adjust for all covariates.

The author delves into the subject of integrating causal inference and machine learning. Incorporating causal inference into machine learning offers many benefits including

enhancing model interpretability and robustness to changes in data distributions. This work considers the task of feature selection for prediction model development in the context of potentially changing environments. First, a filter feature selection approach that improves on the *select k-best* method and prioritizes causal features is introduced and compared to the standard select k-best algorithm. Secondly, a causal feature selection algorithm which adapts to covariate shifts in the target domain is proposed for domain adaptation. Causal approaches to feature selection are demonstrated to be capable of yielding optimal prediction performance when modeling assumptions are met. Additionally, they can mitigate the degrading effects of some forms of dataset shifts on prediction performance.

Chapter 1

Introduction

1.1 Motivation

Many scientific questions and subsequent investigations are fundamentally causal in nature. Yet, often times such questions are not answered in explicit causal terms because of the difficulty in reaching objective causal conclusions in research. Doing causal inference is a notoriously difficult and often controversial scientific proposition (Maathuis and Nandy 2016; Pearl and Mackenzie 2018). Even the very idea and definition of causality constitutes a philosophical debate (Holland 1986; Cartwright 2004; Rothman and Greenland 2005). As a result of a lack of consensus about how to conclusively establish causality in science, for a long time causal inference has been treated with extreme caution in many scientific fields. Not only did many scientific investigators avoid making explicit conclusions about causality, there was also a conservative attitude and even discouragement of causal inference in many quarters of science, with causal reasoning being thought to be 'unscientific' by some theorists (Spirtes et al. 2000; Rothman and Greenland 2005; Frosch and Johnson-Laird 2011; Gelman 2011; Hernán and Course 2018; Hernan et al. 2019; Grosz et al. 2020).

Establishing causality poses a formidable scientific challenge due to the stringent conditions required to definitively affirm causal relationships among variables within any phenomenon of reasonable scale. In addition, there has been a lack of consensus on the precise definition of causality and the necessary conditions for inferring causal

relationships. One of the most widely accepted sets of conditions requires three types of evidence: (1) time order or precedence: the cause must come before the effect, (2) association: the cause and effect must be related, and (3) non-spuriousness: the possibility that a third possibly unobserved variable induces the relationship between the presumed cause and effect must be ruled out (Zheng and Pavlou 2010; Chambliss and Schutt 2018). This third condition is often the most difficult to satisfy. Also, the first condition may not be possible to demonstrate when working with certain types of observational datasets. Given such conditions,[1] the randomized controlled experiment or randomized controlled trial (RCT) is widely regarded as the gold standard for causal inference (Cartwright 2010).

A well-designed experiment is effective for identifying causes and estimating their effects, but they are not always feasible to implement. Indeed, RCTs are often too expensive, too time consuming, unethical or impractical for many applications (Spreeuwenberg 2010). Also, when a large number of variables are involved, the number of experiments required to sufficiently identify causes makes the option of experimentation unrealistic (Eberhardt et al. 2005). Hence, there has been a growing enthusiasm about developing alternative methodologies for causal inference, which do not depend solely on experimental data. This, along with a realization that the study and resolution of problems in causality could help to solve some important challenges in automated learning systems and artificial intelligence (AI), has given rise to the development of "a new science of cause and effect" as Pearl (2018) describes it. This new science of causality emphasizes well-defined causal models and permits plausible assumptions in the modeling of a system under study. It features a rich representation for causal mechanisms, a theoretical foundation for manipulating causal models, and a growing set of methods and algorithms for extracting causal information from observational

[1] see also Hill's criteria (Thygesen et al. 2005).

data. The structural causal model (SCM) framework articulates the basic tenets of this modern approach to causal inference.

Observational data as opposed to experimental data, is data collected through studies where the researcher has little or no control over the data generating process. This type of data is often more readily available than experimental data. It is used extensively in problems of prediction where the goal is to estimate the value of a variable of interest based on its relationship to other observed variables in a system. In such problems, the relationships between variables only need to be based on association, not necessarily causation. Yet, it is often desirable to be able to explain how the estimates are obtained, or to gain insights about the nature of such systems through the techniques used for modeling them. Machine learning (ML) been tremendously successful in prediction and pattern recognition problems. However, its limitations as a technique for enabling critical decision support and the development of advanced intelligent systems are increasingly being acknowledged. ML on its own is unable to provide reliable answers to fundamental questions about causality, and struggles with generalization (Pearl and Mackenzie 2018; Pearl 2019a; Scholkopf et al. 2021).[2]

Causal inference offers tools which have the potential for addressing many of the concerns about current and future AI systems. With regard to generalization, causal approaches can be used to improve the stability of ML models in the presence of changes in the data distribution. This benefit can be exploited in transfer learning and domain adaptation applications (Zhang et al. 2020; Scholkopf et al. 2021; Yang et al. 2021; Spirtes and Zhang 2016). Such causal methods can be used to identify the sources of changes in the data distribution, predict the impact of the changes, and expose how those changes are propagated throughout the system (Guyon et al. 2007; Subbaswamy et al. 2019; Makhlouf et al. 2020). In addition, causality-based modeling approaches

[2]The problem of generalizing models trained in a specific domain to other domains.

can enhance the transparency of prediction models by revealing how they arrive at their predictions. Developing such modeling approaches could have a huge impact on some of the current challenges with ML and AI. An especially thrilling prospect lies in the capacity to reason with data beyond mere statistical associations, delving into the realms of interventions and counterfactuals through the utilization of causal inference techniques (Pearl 2019b). This capability could serve as a catalyst that propels AI research closer to the realization of the aspiration to attain a form of intelligence akin to that of humans, commonly referred to as Artificial General Intelligence (AGI) (Pearl and Mackenzie 2018; Scholkopf et al. 2021).

Modeling techniques which merely map inputs to outputs by computing parameters of a joint distribution or by estimating an approximating function using data without consideration of the data generating process can barely be considered truly intelligent (Pearl and Mackenzie 2018). These systems are often not able to immediately adapt and respond to changes in the underlying data generation process. They do not provide reliable answers to the "why" or "how" questions, as per a reliable explanation for the phenomenon predicted. Without this, their usefulness as decision-making tools is limited. Decision support is often a key need for businesses seeking to adopt AI technologies. Ryall and Bramson (2013) summarize the importance of a modern causal modeling approach for business applications by stating that "In a world of resource scarcity, a decision about which business elements to control or change – a managerial intervention, must precede any decision on how to control or change them, and understanding causality is crucial to making effective interventions."

Other emerging issues and concerns about the prominent autonomous learning and AI methods have to do with bias/fairness, explainability, trust, reproducibility, security and ethical use. As AI technologies gain wider adoption, these issues come into greater focus and the need to address them becomes more urgent. This introduction delves

into several emerging issues concerning contemporary AI methodologies and explores how causal inference could be harnessed to address these challenges.

1.2 Towards Improving ML and AI Outcomes using Causal Inference

1.2.1 Issues and Limitations of Current AI Systems

Progressive developments in machine learning techniques have been the driving force behind most recent advancements in the field of artificial intelligence. The current state-of-the-art techniques, particularly in deep learning (DL), have performed spectacularly and beyond earlier expectations, even though there is still a lack of in-depth understanding, and comprehensive theoretical explanation of why they work so well (Sejnowski 2020). Statistical learning, ML and DL techniques represent the prevailing approach to AI known as Narrow AI. Narrow AIs are designed and trained to perform a specific task. This approach has been so impressive in numerous applications leading some to believe that narrow AIs such as deep learning models provide a path to achieving the so-called holy grail of AI research – achieving AGI or "Strong AI" (Tucker et al. 2008). AGI is an idealized type of AI that is equipped with complex human-like intelligence and can solve a wide variety of problems. It is often imagined as a machine that is capable of a range of abilities in the spectrum of human intelligence (Goertzel 2014). However, many experts do not believe that the models developed using current narrow AI methods can simply be scaled up to create a Strong AI, nor that deep learning in its current form offers a direct path to AGI, due to various fundamental limitations of the current methods (Pearl 2019a; Gobble 2019; Ng and Leung 2020). Despite the excitement at the present rate of AI advancement particularly with the emergence of

successful large language models, current methods are limited in a lot of ways, and we are still a long way from a transformative AI that can truly be considered to be AGI.

Schölkopf (2022a) makes an interesting comparison of current ML-powered AI systems to animal intelligence, highlighting the problem of generalization as a key limitation of current AI systems. The validity of machine learning predictions depend on certain assumptions about the data on which models are trained and deployed. Most notably, the assumption that the source and target data are sampled from the same distribution, is routinely violated in the real world (Pan and Yang 2009; Weiss et al. 2016; Zhou et al. 2022). Generally, ML models tend to become unreliable when there is a shift in the data distribution between source and target domains (Subbaswamy et al. 2019). This also results in the inability of such systems to transfer what has been learned in one environment to another environment, or to generalize learned patterns in one problem setting in order to make predictions and inference in similar problems.

Beyond predictive analytics, machine learning is increasingly being applied to advance scientific discovery efforts. In applications focused on knowledge discovery, a significant challenge when using many statistical techniques and ML models arises from their tendency to imply relationships between variables that do not reflect real-world dependencies. This is because ML methods do not necessarily model the data generating process (Li et al. 2020). It is not uncommon for ML models to achieve predictive power by learning spurious relationships among data features. This can result in poor reproducibility of ML outcomes. Furthermore, the opacity of many contemporary successful models owing to their complexity, can create difficulties in establishing confidence that the obtained results are defensible, particularly in scientific knowledge discovery tasks.

The increasing complexity of state-of-the-art ML models renders them obscure to human comprehension, impeding our ability to understand and interpret them effectively. Consequently, many ML models are often perceived as black-box systems that simply generate predictions when given an input, but they do not readily support inferential or decision-making tasks due to the challenge of providing comprehensive explanations regarding the rationale behind their predictions. This is problematic in sensitive applications (e.g., healthcare), or where there are ethical concerns (e.g., law enforcement), and where there is need to understand how a process really works in order to make the right decisions or take appropriate actions to influence future outcomes (e.g., policy evaluation). This challenge has sparked a revival in the field of explainable artificial intelligence (XAI), as evidenced by notable works such as Berrada et al. (2018); Samek and Müller (2019); Vilone and Longo (2020); Angelov et al. (2021). This resurgence is encouraged by the observation that in some domains there is resistance to embracing new AI applications because of a lack of trust or due to ethical apprehensions on the part of users (Miller 2019; Confalonieri et al. 2021). These concerns are often linked to the limited transparency and interpretability of complex AI systems.

Despite these limitations, AI systems driven by ML and DL are experiencing widespread adoption across numerous sectors. Powerful AI algorithms are now deployed in critical operations, including medicine, law enforcement, financial systems, business processes, autonomous vehicles, various industrial automation processes, and even in military operations. However, these AI systems, while being highly capable, lack the sort of adaptable complex reasoning and emotional intelligence found in humans. This deficiency raises concerns as these systems which humanity is increasingly becoming dependent on, may not be readily alignable to human values and could be vulnerable to manipulation. This leads to the critical issue of AI ethics and the potential for misuse.

AI misuse can occur intentionally or unintentionally. Unintentional misuse has garnered considerable attention recently and is addressed in a growing body of literature on algorithmic/ML bias and fairness (Hajian et al. 2016; Barocas et al. 2017; Corbett-davies and Goel 2022; Leslie 2020; Mehrabi et al. 2021). On the other hand, intentional or malicious misuse has the potential to result in even more catastrophic consequences. Adversarial AI is emerging as a field that utilizes AI for various attack and defense strategies (Huang et al. 2011; Ng and Leung 2020). The work of Brundage et al. (2018) delves into the landscape of malicious AI use, uncovering current vulnerabilities within AI systems and the security threats posed to human society. The field of Ethical AI has emerged in response to these concerns about AI misuse (Siau and Wang 2020; Eitel-Porter 2021). Trustworthy AI, a related field, focuses on enhancing trust between humans and AI technologies. Surveys on trustworthy AI literature by Liu et al. (2021) and Kaur et al. (2022) explore various dimensions of AI-related concerns, encompassing safety, reliability, privacy, discrimination and fairness, explainability, accountability, and environmental sustainability.

1.2.2 Prospects and Opportunities in Causal Inference for Improving AI Outcomes

This section confronts the issues raised in the previous section by highlighting how causal inference concepts and methods may be useful for resolving current challenges in AI/ML and statistical learning methods. Pearl (2018; 2019b), in discussing the constraints of machine learning in the context of advanced intelligent systems, characterizes AI systems founded on ML and DL as primarily proficient in "curve fitting," which represents only a basic intelligence capability. Causal models, unlike ML models, are intentionally crafted to capture the authentic influence mechanisms inherent in the

processes they represent, linking relevant factors in a network of causal relationships that elucidate how changes in one variable can affect others. Consequently, it is likely more feasible to emulate natural reasoning patterns using causal models.

Pearl (2018; 2018; 2019b) further suggests how AI models can make the progression towards higher level reasoning using the "ladder of causation". The ladder of causation is a three-level hierarchy that describes what types of questions an intelligent agent can answer based on what causal information it is able to use. Within this hierarchy, knowledge about associations (acquired through observing or "seeing"), which represents the learning capability of standard ML techniques, constitutes the most rudimentary form of causal knowledge and is situated at the lowest rung of the causal hierarchy. True causal reasoning capabilities are attained in the second and third rungs of the ladder of causation, specifically: in the ability to reason about interventions (learned by "doing") and in the capacity to reason about counterfactuals (learned by "imagining"). The integration of these advanced causal reasoning capabilities into AI systems has the potential to elevate their overall intelligence capabilities.

Schölkopf (2022b) supports this idea by suggesting that causality's focus on modeling and reasoning about interventions can help with the understanding and resolution of the issues that are currently limiting ML and AI. With the impressive predictive performance of current machine learning methods, an enticing prospect is the incorporation of structural causal models into ML to enable it to be useful for answering questions of an interventional or counterfactual nature. Pawlowski et al. (2020) take a step in this direction by suggesting a framework for building structural causal models using deep learning networks.

A key motivation behind the growing engagement of ML&AI researchers with causal inference lies in the aspiration to enhance the generalization capabilities of ML models. This is substantiated by the recent exploration of causal inference principles for

tackling challenges in domain adaptation and transfer learning, as indicated by the works of (Spirtes and Zhang 2016; Zhang et al. 2020; Yang et al. 2021; Scholkopf et al. 2021). Causal inference can enable stable prediction in the presence of dataset shifts, enhancing the dependability of ML predictions when changes to the data distribution are likely. Furthermore, it offers the prospect of mitigating the costly need for frequent retraining of ML models to align them with the evolving state of the processes they model. Subbaswamy et al. (2019) show how generalization from source to target distributions can be improved by incorporating knowledge about the data generating process in the form of a causal graph which reveals features that may experience a distribution shift due to an intervention on one or more variables.

In data-driven knowledge discovery, achieving high prediction accuracy in ML is not an indication that the results will be highly reproducible in the real world (Li et al. 2020). The challenge of poor reproducibility in ML models can be addressed by aligning a prediction mechanism with the true causal mechanisms governing the process it models. Li et al. (2020) advocates for a shift from accuracy-based to robust, causality-based model development. An ML paradigm known as informed machine learning (also, physics-guided machine learning) revolves around the incorporation of prior knowledge about the physical world into machine learning models (Vonrueden et al. 2021). This approach involves leveraging logic rules and algebraic expressions to introduce prior knowledge, often in the form of constraints, into neural networks (Xu et al. 2018; Daw et al. 2017; Shakya et al. 2021). As knowledge representations themselves, causal models are well-suited for such integration which could further enhance the alignment of AI models with the true data generating process of the systems they model. An additional benefit of aligning a prediction mechanism with the true causal mechanism is the mitigation of spurious correlations that lead to biased predictions. This can result in enhanced prediction performance on both test and target data, with reduced model

overfitting. Demonstrations by Bahadori et al. (2017) and Kyono et al. (2020) illustrate that causality-based regularization methods can improve the prediction performance of neural network models.

In decision support applications, causal inference methods offer a direct avenue for assessing the potential impacts of different decisions and courses of action, all without necessitating direct intervention or experimentation on an existing process. Narendra et al. (2019) demonstrate this using a methodology that transforms business process modeling notation (BPMN) into DAGs, and subsequently employs non-parametric models to estimate various counterfactual effects. This can be vital for business process improvement by identifying appropriate interventions to be triggered during the execution of a process in order to optimize its performance (Shoush and Dumas 2021). Moreover, the capacity to predict the effects on the outcome, of interventions on features in prediction models can be invaluable (Blöbaum and Shimizu 2017; Kiritoshi et al. 2021).

Developing causality-based prediction models can enhance the explainability of ML models by simplifying the understanding of how the model integrates features to generate reliable predictions. The incorporation of causal models simplifies the task of understanding and explaining to users and stakeholders why the model functions as it does, ultimately contributing to the establishment of more trustworthy AI systems through improved transparency. Despite considerable recent efforts to enhance the transparency and interpretability of ML and DL, many of these approaches still rely on correlations rather than causation which limits their effectiveness. Feder et al. (2021) demonstrate why such methods are insufficient due to their inability to distinguish between strong correlations and real causes. They introduce a causal framework for explaining predictive models using counterfactual language representation models, and also show how this approach helps to mitigate bias in predictive models.

The notion that causality-based solutions are needed to properly address the problem of fairness in ML is increasingly being acknowledged (Loftus et al. 2018; Makhlouf et al. 2020). Kilbertus et al. (2017) and Makhlouf et al. (2020) illustrate the inadequacies of non-causal statistical fairness notions and provide a review of various causal characterizations of the fairness problem that could address those limitations. An emerging body of research is dedicated to characterizing, identifying, and mitigating various forms of discrimination in both data and algorithms, leveraging principles from causal inference theory and methods (Zhang et al. 2017; Bonchi et al. 2017; Zhang et al. 2016; Wu et al. 2019; Zhang and Bareinboim 2018). Kusner et al. (2017) introduce a framework for modeling fairness using causal concepts. They use the notion of counterfactual fairness to evaluate the fairness of decisions made by algorithms (e.g., label classification) based on how different the decision would be in a counterfactual world where the individual belonged to a different demographic group. A better understanding of the underlying structure of the relationships between variables also provides further insight for constructing models that are not biased or discriminatory.

One noteworthy area of opportunity for the application of modern causal inference concepts, which has received comparatively less attention is in combating the malicious use of AI. Causal methods capable of reverse-engineering the underlying process mechanism from observational data can serve as a powerful tool for assessing the effects of unusual interventions or adversarial counterfactual scenarios before they materialize. They can be designed to identify sources of the changes in data and may expose how those changes can be propagated throughout the system (see Guyon et al. 2007; Subbaswamy et al. 2019; Makhlouf et al. 2020). Consequently, an AI system equipped with such capabilities can detect and potentially prevent adversarial manipulations of its data or algorithms. Additionally, it can make informed determinations about the suitability of specific datasets for training fair prediction models. This is important

for AI applications where bias detection and mitigation are paramount, such as those that bear significant consequences on human lives and society.

1.3 Objectives

The overall goal of this research is to develop causality-based solutions capable of harnessing observational data to enhance various outcomes derived from machine learning and other data-driven learning systems. To achieve this goal, the following objectives are specified:

- Survey and discuss approaches for causal inference using observational data.
- Explore the theoretical underpinnings of structural causal models and the landscape of causal learning tasks, discussing the fundamental concepts, prominent methodologies and emerging techniques that form the bedrock of modern causal inference.
- Identify research opportunities for integrating causal methodologies into data-driven systems, aiming to bolster knowledge discovery, and address specific current limitations of machine learning.
- Develop a causality-based method that addresses an important limitation of machine learning.
- Develop a data-driven causality-based solution that can be applied to an important problem in industry.
- Assess the methods developed through this work by conducting simulation studies that employ datasets with well-understood properties and well-defined ground-truth data generation mechanisms.

- Discuss in detail the results obtained from the studies, highlighting the contributions of each study.

1.4 Organization of Dissertation

The rest of this dissertation is organized as follows. Chapter 2 features a discussion on theoretical frameworks which enable causal inference using observational data. Fundamental causal concepts are introduced through the potential outcome model framework before a more in-depth exploration of the principles of the broader structural causal model (SCM) framework.

Chapter 3 engages in a broad exploration of current methods and approaches for causal inference using observational data. The author delves into the diverse types of learning tasks utilized for addressing causal inquiries in a variety of scenarios. The causal learning tasks are categorized into two primary groups: tasks aimed at estimating treatment or causal effects, and tasks centered around modeling the causal structure of the variables within the system under study. The author contends that in a general context, the latter task is a prerequisite for the former to be accomplished. This sets up a broad discussion of algorithmic techniques for learning causal structures from observational data, commonly referred to as causal discovery methods. Finally, two strategies for integrating causal inference and machine learning are briefly introduced.

Chapter 4 confronts the problem of data-driven root cause analysis (RCA). A graphical causal model representation termed as root cause graphs (RCG) is suggested for depicting the structural mechanism that leads to an observed event of interest. In industrial settings, the event of interest is often some type of failure. A novel causal learning method based on SCMs is proposed for RCA using observational time-to-event

(TTE) data. This method is demonstrated to improve the estimation of the impacts of changes to variables within the system on the outcome of interest.

In Chapter 5, the integration of causal inference and machine learning through feature selection is explored. With the goal of improving the outcome of ML generalization through domain adaptation, as well as enhancing the interpretability of prediction models, two novel causal feature selection algorithms are proposed. The first algorithm is shown to enhance prediction performance in comparison to a related non-causal filter feature selection method. The second algorithm demonstrates the capability to adapt the feature selection process to a target domain, mitigating the impact of certain distribution shifts between source and target datasets. The findings of this study shed new light on the specific problem settings and conditions where causal feature selection approaches can excel.

Finally, the dissertation concludes in Chapter 6 with a recap of the contributions of this work. The synergies between causal inference and statistical/machine learning are further emphasized.

Chapter 2

Frameworks for Causal Learning

Cartwright (2004); Rubin (2005); Pearl (2009a, 2013) emphasize the need for the formalization of causal inference as a scientific theory with linguistic, symbolic and methodological developments, in order to complement the tools provided by statistics and probability for the effective elicitation of causal knowledge from data. Two of the most popular theoretical frameworks for causal inference are the Potential Outcome Model (POM) framework (Rubin 1974, 2005; Imbens and Rubin 2015) and the Structural Causal Model (SCM) framework (Pearl 2000, 2009a). These frameworks generalize and extend principles for causal inference in the ideal experimental setting to allow for inference in the observational setting. Many of the prominent methods for causal inference today are based on these frameworks. The theories and concepts reviewed in this section serve as a foundation for subsequent discussions in the rest of this dissertation.

Besides the POM and SCM frameworks, it is worth mentioning that other frameworks exist for causal inference. Granger Causality (Granger 1969; Stern 2011), a special notion of causation which is prediction-based is used for causal inference with time series data and is popular in the social sciences. Structural econometrics provides several tools for parametric modeling and estimation of causal effects. Elements of the econometric approach are reflected in SCMs through structural equations models (SEMs). The Sufficient Component Cause (SCC) model (Rothman and Greenland

2005; Flanders 2006) which can be found in various biology and epidemiology literature, tries to describe a complete causal mechanism including events and states that are necessary and sufficient for the realization of an outcome of interest, while recognizing that the conditions may differ between different individuals or groups. The Quantitative Comparative Analysis (QCA) method (Marx et al. 2014; Berg-Schlosser et al. 2009) is a way of using boolean logic and set theory to model situations where multiple combinations of several variables may be sufficient to produce an outcome. The SCC and QCA frameworks may be able to capture very specific and complex behaviors in certain causal mechanisms by taking a deterministic view on the nature of causality.

Assumptions are necessary for practical causal inference and these causal frameworks are useful for articulating the relevant assumptions for causal inference in different scenarios. This dissertation refrains from delving into the age-old debate on whether causality is deterministic or probabilistic - see Rosen and Press (1978); Frosch and Johnson-Laird (2011). From the author's view of existing literature, most of the recent advancements in the science of causality, including its algorithmization and progressive incorporation into ML & AI techniques, have been achieved by leaning towards a probabilistic view of causality. However, while the SCM approach which has been central to many recent advancements in causality is primarily probabilistic, it does invoke deterministic characterizations – notably through the Markov condition (Pearl 1996).

2.1 Potential Outcome Model Framework

The POM framework characterizes the fundamental challenge of causal inference and lays down principles that allow causal effects to be estimated not only from RCTs, but

also from imperfect experiments and observational settings. Causal inference practically involves assessing how the manipulation of a treatment affects the outcome, while holding other relevant variables constant (Heckman 2008).

Consider a set $\mathbf{V}$, of random variables under study with subsets X, Y, and $\mathbf{Z}$ where X represents a treatment variable for which its causal effect on the outcome variable Y is of primary interest. $\mathbf{Z}$ represents a set of covariates measured alongside X and Y which may confound the relationship between X and Y. Assuming X and Y are a binary treatment and outcome respectively, both variables can take on the value of 0 or 1. For an experimental unit or observed instance i, a **potential outcome** Y_i^x is an outcome which will be observed given a particular treatment $X = x$. For example, when $X_i = 1$ is observed, the potential outcome is expressed as Y_i^1. Note that Y_i^1 may manifest two possible outcome values: $Y_i = 0$ or $Y_i = 1$.

The potential outcome model framework hypothetically formulates the problem of causal inference as the difference in potential outcomes.[1] Practically, this implies that one needs to measure all possible outcomes for each potential treatment and compare them in order to estimate causal or treatment effects. For individual i in our example, this causal effect is the difference between the two potential outcomes as expressed in Equation 2.1.

$$Causal\ Effect = Y_i^1 - Y_i^0 \tag{2.1}$$

However, in reality only one potential outcome can be observed for an individual at any given time. This is the *fundamental problem of causal inference* – for any particular subject, we can only observe one of the potential outcomes (Imbens and Rubin 2015). Regardless of whether $X = 1$ or $X = 0$, only one of the two possible outcome values, $Y_i = 1$ or $Y_i = 0$, can be observed for any individual. A *counterfactual outcome* is an outcome that would be observed had the treatment been different. So for individual i,

[1] Other comparison operators besides difference could also be used, e.g. ratio.

if $X = 1$ is observed, the counterfactual outcome for unit i, Y_i^0, would be the outcome that would have been obtained had $X = 0$ been observed.

Because of the fundamental problem of causal inference, it is impossible to directly compute causal effects for an individual unit using Equation 2.1. Since in reality only one realization of all potential outcomes for a sample unit can be observed, the common strategy is to compare multiple units in order to estimate the average effect of a treatment on an outcome. This means that treatment effects are estimated at the population level. Assuming treatment X is binary, the average causal effect (ACE) or average treatment effect (ATE) is the difference between the expected values of Y given treatment $X = 1$ on the whole population, and treatment $X = 0$ on the whole population.[2] Hence ATE can be expressed as

$$ATE = E(Y^1) - E(Y^0) \tag{2.2}$$

where Y^1 is the potential outcome when treatment is $X = 1$, and Y^0 is the potential outcome when treatment is $X = 0$

Equation 2.2 is still a theoretical expression that cannot be directly computed since both potential outcomes cannot be observed for all samples in the same population, however it is easier to estimate compared to 2.1. Note that the ATE is not the same as a **naive causal estimator** (NCE) which is simply calculated by directly comparing the average values of the outcomes observed from two different samples given different treatments, without accounting for the differences between the individuals in the sample groups. Without ensuring that the individuals being compared are reasonably similar, the analysis becomes prone to confounding and selection bias.

$$NCE = E(Y|X = 1) - E(Y|X = 0) \tag{2.3}$$

[2]Causal effect and treatment effect are used interchangeably in this work.

NCE is the difference between those who got the treatment and those who did not. This measure by itself does not consider sample selection mechanism (e.g., treatment selection could have been by individual choice) and thus is not a reliable causal estimate when using observational data. With well-designed RCTs however, the naive estimator is expected to approximate the ATE since randomization nullifies selection bias.

At the population level, it is easier to estimate causal effects by comparing the difference between two sample groups randomly drawn from the same population that have been assigned different treatments. So, in practice different subsets of the population are actually compared since the same group can not be assigned different treatments at the same time. The goal is to have two sample groups that can be assumed to be similar to each other except that they are subjected to a different treatment. This is the basis for most methods for treatment effect estimation. The similarity between sample groups to be compared can be achieved by randomization as in RCTs or by other techniques such as propensity score methods. The average difference in outcomes between the two similar sub-populations which have been exposed to different treatments is the estimated ATE.

Several other treatment effect measures are also defined theoretically in terms of potential outcomes. For example, the average treatment effect on the treated (ATT) considers only the treated sub-population.

$$ATT = E(Y^1|X = 1) - E(Y^0|X = 1) \tag{2.4}$$

The average treatment effect on the non-treated (ATN) considers only the untreated sub-population.

$$ATN = E(Y^1|X = 0) - E(Y^0|X = 0) \tag{2.5}$$

The conditional average treatment effect considers a subset of the population with similar values of covariates or confounders **Z**.

$$CATE = E(Y^1|Z = z) - E(Y^0|Z = z) \tag{2.6}$$

The CATE is an important measure because in many applications sub-populations with different characteristics could be affected in different ways by a particular treatment. When this happens, the ATE can be a misleading indicator of causal effects when considering sections of the population since the population is not homogeneous. CATE isolates the treatment effects for different sub-populations, partitioned based on the values of relevant covariates. This accounts for the heterogeneity in the population being studied with respect to the effect of the treatment. Hence, CATE is also referred to as the heterogeneous treatment effect. The isolation of effects in sub-populations can be done up to the individual level in which case CATE would reflect the individual treatment effect (ITE). ATE is sometimes calculated from CATE by taking a weighted average of the CATE over all sub-populations.

2.1.1 Assumptions for Causal Inference

To compute reliable estimates of causal effects, certain conditions are assumed about the subjects and the variables in the data. The POM framework is grounded in the following assumptions (Imbens and Rubin 2015).

Ignorability: This assumption states that treatment assignment is independent of potential outcomes conditional on a set of covariates. This means that there can be no unmeasured or unaccounted variables which confound the causal relationship between the treatment and the outcome. Hence treatment assignment is said to be ignorable,

given the relevant set of covariates Z. This is often the most consequential assumption in most causal inference tasks. It is also sometimes referred to as the assignment assumption, or unconfoundedness assumption, or exchangeability assumption in various literature. While unconfoundedness can be achieved through randomization in experimental settings, it is typically more difficult to satisfy this assumption in observational studies. However, there are several methods for enhancing the plausibility of this assumption in observational settings.

Stable Unit Treatment Value (SUTVA): This assumption has two aspects. The first entails that there is no interference between the subjects being studied. This precludes interactions and influence between the units under study. The second aspect assumes that there are no hidden variations of the treatment values. So all treatment levels are assumed to be known.

Positivity: This states that the probability of receiving any value of treatment is non-zero for all subjects given covariates Z. So there is some chance that any subject could receive any treatment, and treatment is not deterministic as a function of Z.

2.2 Structural Causal Model (SCM) Framework

The Structural Causal Model (SCM) framework is a broad theoretical framework that combines features of the potential outcome framework, structural equations modeling, and probabilistic graphical modeling (Pearl 2009a, 2018). In this framework, causal mechanisms can be represented using graphical models or by a corresponding set of equations known as a functional causal model (FCM). These representations describe the structural dependencies that exist between a set of variables without having to specify the precise functional forms and parameters characterizing the relationships

(Pearl 2009b). The functional parameters can subsequently be fully specified as in linear structural equations models (SEMs). Thus, FCMs consist of a set of non-parametric SEMs.

In probabilistic graphical models, the nodes in the graph represent the variables in the system under study, while the edges represent dependencies between pairs of variables. The basic type of graph used to represent causal structures is the Directed Acyclic Graph (DAG) in which directed paths that start and end at the same node (cycles) are not permitted. DAGs can be used to encode information about joint distributions and conditional independencies via a criterion known as *d-separation* (see Definition 2.2.2). Graphical models are visual, easy to interpret, and make explicit the beliefs and assumptions of the modeler about the process under study.

The Bayesian Network (BN) is a type of probabilistic graphical model that is based on DAGs (Koller and Friedman 2009). It represents a set of variables, with associated marginal and conditional probabilities, and their joint probability distribution (Korb and Nicholson 2008).[3] A BN encodes the set of conditional independencies in the joint probability distribution over the variables in the system. With BNs, it is important to pay attention to the missing edges in the network because they reflect the main assumptions about the data. As Korb and Nicholson (2008) explain "the lack of an edge between two variables must be reflected in a probabilistic independence in the system being modeled."

BNs may be causal or non-causal depending on the assumptions supporting a particular representation. In this work, we are mainly interested in causal BNs where a directed edge is assumed to indicate the direction of causation between two adjacent nodes and DAGs are assumed to represent the underlying causal structure for a set of

[3]The terms BN and DAG are commonly used interchangeably in causal inference literature. The main difference between the two is that a Bayesian network maps a probability distribution, while a DAG is a representation that does not have to be linked to a probability distribution.

variables.[4] Hence, for a causal graph, G, of a set of variables, $\mathbf{V}$, a directed edge from V_i to V_j indicates that V_i is a direct cause of V_j relative to $\mathbf{V}$. For in-depth discussions on causal graphs, graphical models, and Bayesian networks, see Lauritzen (1996); Pearl (2000); Greenland and Pearl (2006); Koller and Friedman (2009); Korb and Nicholson (2010); Pearl et al. (2016).

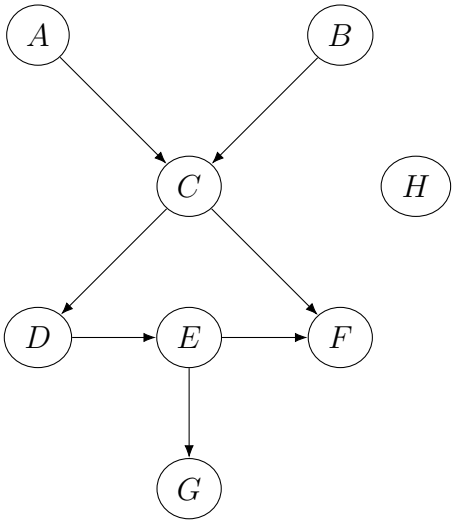


Figure 2.1: DAG representation

$$
\begin{aligned}
A &:= f_a(\epsilon_A) \\
B &:= f_b(\epsilon_B) \\
C &:= f_c(A, B, \epsilon_C) \\
D &:= f_d(C, \epsilon_D) \\
E &:= f_e(D, \epsilon_E) \\
F &:= f_f(C, E, \epsilon_F) \\
G &:= f_g(E, \epsilon_G) \\
H &:= f_h(\epsilon_H)
\end{aligned}
$$

Figure 2.2: FCM representation

Figure 2.1 is an example of a causal DAG or BN while Figure 2.2 outlines its corresponding FCM or set of SEMs[5]. This model consists of eight variables {A,B,C,D,E,F,G,H} and the dependencies between the variables are expressed in both the graph and the set of structural equations comprising the FCM. In an SEM, each variable is a function of its direct causes and an error term ϵ. Notice that the SEMs do not commit to a specific functional form at this stage. Also, the unidirectional assignment operator :=

[4](Korb and Nicholson 2008) explain that causal models are often the simplest of Bayesian networks capable of representing the probabilistic fact.

[5]In this context, the author uses SEM and FCM interchangeably. In some fields SEMs denote parametric sets of equations with explicitly specified functional forms and parameters, and may not necessarily imply a causal relationship. (Glymour et al. 2019) discuss the class of parametric structural equations models which describe each variable as a deterministic function of its direct causes.

is used in the SEMs instead of the equals sign (=) to make it clear that the relationship depicted is asymmetric.[6]

Error (ϵ) terms have been omitted in the graph for simplicity, as is usual in the graphical representation. The error terms represent the uncertainty or noise in the relationship. They also account for the effect of variables that may exclusively affect a particular node but are not accounted for in the model. The error terms together are assumed to be jointly independent. For example, for the variable D, ϵ_D is a source of variation that only affects D relative to the rest of the variables in the model, and ϵ_D is independent of all other ϵ's.

In BN terminology, familial relationships are used to describe connections between a set of variables or nodes in a graph. In the BN in Figure 2.1, {A,B} are parents of C, and {D,F} are children of C. Similarly, {F,G} are children of E, and this makes F and G siblings. D and F are grandchildren of both A and B. The set {C,D,E,F,G} consists of descendants of A and B, while A and B are likewise their ancestors. A is the spouse of B and vice-versa, while H is unrelated to any other node in the Bayesian Network.

2.2.1 Definitions and Assumptions

Definition 2.2.1 Conditional independence: *Given a joint probability distribution consisting of a set of variables {X,Y,Z}, Y is conditionally independent of X given Z if the conditional distribution of Y given X and Z does not depend on X. This is expressed mathematically as*

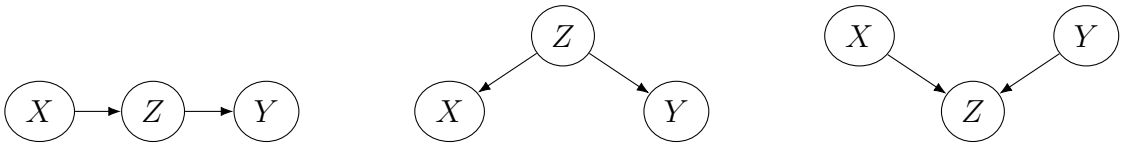
$Y \perp\!\!\!\perp X \mid Z$, *if* $P(Y \mid X, Z) = P(Y \mid Z)$.

This also implies:

$P(Y, X \mid Z) = P(Y \mid Z) \times P(X \mid Z)$.

[6] some texts use the symbol $\leftarrow$ instead.

Consider the basic graph structures graphs of the variables {X,Y,Z} with the configurations in Figure 2.3. These three types of graph junctions are useful for illustrating how conditional independencies are encoded in causal graphs.



(a) A chain (b) A fork (c) An inverted fork

Figure 2.3: Three basic graph junctions

The graph in Figure 2.3a is called a *chain*. In this graph the effect of X on Y is transmitted through Z. Hence, the variable Z is called a mediator because it explains the causal effect of X on Y. There is marginal dependence between every pair of nodes in this graph which entails that $Y \not\perp X$. However, Y is conditionally independent of X given Z ($Y \perp X \mid Z$).

Figure 2.3b is referred to as a *fork*. In this graph Z exerts causal influence on both X and Y. In this case X and Y are associated and hence marginally dependent ($Y \not\perp X$) because of their mutual dependence on Z, however, they are not causally dependent. Their marginal dependence is due to their common causal relationship with Z. Hence Z is referred to as a *confounder* because of the spurious relationship it induces between X and Y. Y becomes conditionally independent of X when Z is given ($Y \perp X \mid Z$). This confounding relationship is a great example of how correlation does not imply causation. Note that the chain and fork structures encode the same conditional independencies.

The graph in Figure 2.3c is called an *inverted fork*. Z in this case is referred to as a *collider* because the separate causal influences of X and Y meet (collide) at Z. A collider node in a graph has two or more edge arrows pointing into it. In this setting, X and Y are marginally independent ($Y \perp X$) , but given Z they become dependent

($Y \not\perp X \mid Z$). Notice that the inverted fork encodes a different set of independence constraints compared to the chain and fork structures. Such structures can be uniquely determined from a joint distribution through conditional independence testing under certain conditions.

The following graphical causal modeling concepts as formalized in Pearl (2014) are hereby introduced. First, a *path* is defined as a sequence of adjacent nodes and edges in a graph.

Definition 2.2.2 D-Separation: *A path p is said to be d-separated by a set of nodes* $\boldsymbol{Z}$ *in a graph if either*

1. $\boldsymbol{Z}$ *includes at least one arrow-emitting node in path p*

2. *p contains a collider node that is not in Z and has no descendant in Z.*

A path that satisfies the above condition is said to be *blocked* by Z, otherwise it is said to be *activated* by Z. If the collider node referred to has a descendant in $\mathbf{Z}$, then it is a partial block. Hence, we say that a set of variables Z d-separates two or more other variables, e.g., X and Y if it blocks the path between them. Z blocks the path from X to Y in both Figure 2.3a and Figure 2.3b, but activates the path in Figure 2.3c.

Definition 2.2.3 I-map DAGs: *A DAG G is an independence map (I-map) of a dependency model M if every d-separation condition in G corresponds to a conditional independence relationship in M. A DAG is a minimal I-map if none of its edges can be deleted without nullifying its I-map property*

Definition 2.2.4 Bayesian network: *Given a joint probability distribution P on a set of variables V, a DAG G is a Bayesian network if and only if it is a minimal I-map of P.*

Definition 2.2.5 Perfect maps: *A DAG G is a perfect map of a probability distribution P if P embodies all independencies present in G, and no others.*

A perfect map satisfies both the Markov and Faithfulness conditions (Markov condition is introduced below while faithfulness condition is discussed in section 3.2). Definitions 2.2.3 and 2.2.4 encapsulate two important assumptions about the type of graphical representations used in structural causal models: the causal Markov condition and the minimality condition described below. But first, causal sufficiency, a common assumption in many causal inference methods is introduced.

Causal sufficiency: The causal sufficiency assumption states that the set of measured variables $\mathbf{V}$ includes all common causes of all pairs of variables in $\mathbf{V}$. $\mathbf{V}$ is said to be causally sufficient if for every pair of variables $V_1, V_2 \in \mathbf{V}$, every common direct cause of V_1 and V_2 relative to V is also a menber of $\mathbf{V}$ (Zhang 2008). That is, there are no unmeasured confounders in the set of variables in $\mathbf{V}$.

Causal Markov Condition: For a Bayesian network G, every variable V_i in G is conditionally independent of its non-descendants (non-effects) given its parents (direct causes) $Pa(V_i)$. This assumption is known as the Markov condition. The Markov condition has an implication on the factorization of joint distributions using graphical criteria. Via the chain rule of probabilities, the probability of a joint distribution $\{V_1, V_2, ...V_n\} \in \mathbf{V}$ can be factorized as $P(\mathbf{V}) = \prod_{i=1}^{n} P(V_i|Pa(V_i))$ according to this condition. The Causal Markov Condition (CMC) interprets these dependencies expressed by a Bayesian Network as causal (Scheines and Sobel 1997; Spirtes and Zhang 2016).

Minimality Condition: The minimality condition states that any *proper sub-graph* H of graph G would violate the Markov Condition (Spirtes et al. 2000). That is, removing any edge in G would cause the graph to imply a conditional independence that does not exist in P (G is a minimal I-map of P). The CMC and minimality

conditions connect a DAG to the probability distribution it represents (Spirtes et al. 2000).

Acyclicity: Just like the DAGs that are used to represent them, the causal structures considered in this work are assumed to be acyclic. Cyclic causal mechanisms in the real world can often be modeled as acyclic by temporally unfolding relevant variables or events. It's worth noting that ongoing research is addressing cyclic causal structures within the SCM framework. The reader is referred to the paper by Bongers et al. (2021) for a detailed discussion of the desirable properties and limitations of acyclic causal graphs, and recent theoretical work on cyclic graphical models.

2.2.2 Intervention and Counterfactual Analysis

One area where association-based prediction methods struggle is when a change occurs in the modeled system that results in a modification of its joint probability distribution. Such a change can occur due to an "intervention" on one or more variables. Causal inference can enable reliable inference and prediction of outcomes of interest in the presence of interventions that affect the data distribution. Intervention-type queries are important in many applications where decision makers need to understand what actions to take in order to affect a possible outcome. Conducting controlled experiments allow the direct manipulation of variables to see how they affect the outcome. In situations where experimentation is impractical, the capability to address intervention queries using observational data takes on significant importance.

A conditional probability distribution can be useful for investigating possible confounding effects. However, the probabilistic conditioning operation, e.g., $P(Y \mid X)$, stems from mere observation and is ambiguous as to whether the observed association is causal (Eberhardt 2017). Pearl (2000; 2009a; 2016) introduced the do-operator to operationalize the intervention mechanism, enabling intervention queries using data.

Consider a situation where the value of a variable X is fixed by an intervention $X = x$ as opposed to allowing X to obtain its values organically from its natural data generation process.[7] While the probability that $Y = y$ conditional on observing $X = x$ is expressed as $P(Y = y \mid X = x)$, the probability of observing $Y = y$ after manipulating X $(X = x)$ is expressed as $\mathrm{P}((Y = y) \mid do(X = x))$. $P(Y \mid do(X))$ can be described as the post-intervention distribution of $\{X, Y\}$.

In graphical models, an intervention is performed by removing all incoming arrows into the variable such that the variable has no natural causes but takes on values set by a manipulator. Consider the task of investigating the effect of X on Y in Figure 2.3b, fixing X will result in the graph in Figure 2.4 which allows the correct estimation of $P(Y \mid do(X))$. In this graph it becomes clear that Y is no longer statistically associated with X so one can infer that X is not a cause of Y in this distribution. On the other hand, in SEMs, intervention is performed by setting the treatment $X = x$ as in the corresponding SEMs in Figure 2.4. Modeling an intervention using the

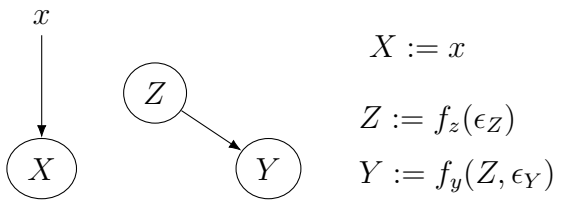


Figure 2.4: Intervention on X in Fig. 2.3b and corresponding SEMs

do-operator allows the causal effect of X on Y to be defined as the distribution of Y after performing $do(x)$, that is $P(Y \mid do(X))$. By this hypothetical definition of interventional distribution, the causal effect of an intervention x_1 relative to another x_0 is given by

$$P(Y \mid do(X = x_1)) - P(Y \mid do(X = x_0)) \tag{2.7}$$

[7] As in how a variable may be manipulated in an experiment.

This is analogous to the potential outcome definition of causal effects in Equations 2.1 and 2.2. It is also evident that the fundamental problem of causal inference applies here and so do the same constraints that make the calculation of individual treatment effects difficult as discussed in Section 2.1.

Consider the graph in Figure 2.5a. The effect of X on Y cannot be disentangled from the effect of Z on Y using observational data alone. Suppose we fix X as in a controlled experiment, we get the graph in Figure 2.5b. The causal effect of an intervention on X becomes obtainable, and can in fact be derived from the conditional probability observed from the manipulated model P_m as in Equation 2.8. Pearl (1995; 2016) further derives the *backdoor adjustment formula* (or simply *adjustment formula*) in equation 2.9 for computing causal effect of variable X on variable Y, given data on a *sufficient set* of confounders Z'.[8] Assuming the covariate Z is equivalent to Z', this operation amounts to adjusting for or controlling for Z. The derivation of the adjustment formula successfully stripped out the do-expression leaving only marginal and conditional probabilities which can be estimated from observational data. This way the adjustment formula can compute the causal effect of X on Y by evaluating the association between X and Y at each value of Z, and averaging over those values.

(a) Probability model P (b) Manipulated probability model P_m

Figure 2.5: Intervention modeling

[8] In this article, the author often refers to the set of variables sufficient for the control of confounding - Z', as simply the sufficient set. Z' is introduced to indicate that the set of measured covariates is not necessarily equivalent to the sufficient set. Z' is used instead of Z to specifically indicate that the covariates being referred to constitute the sufficient set.

$$P(Y = y|do(X = x)) = P_m(Y = y|X = x) \tag{2.8}$$

$$P(Y = y|do(X = x)) = \sum_z P(Y = y|X = x, Z = z)P(Z = z) \tag{2.9}$$

Another type of causal query is the counterfactual query. This type of query is of interest in many applications and takes the form of "If X were true, would Y have been true?"(Balke and Pearl 2013). The perspective of the query is after the fact. That is, after data about a series of events has been observed, what would have been the outcome if say one of the events had not occurred? Its essence is that of a "What if" analysis. Pearl et al. (2016) outlines this idea of counterfactuals clearly as a tool for comparing two alternate (potential) outcomes under the same exact conditions except for the antecedent, which is an alternate treatment. It differs from intervention queries in that an intervention query does not refer to a world where a different event had already occurred and its resultant outcome had been observed.

A counterfactual can be defined as follows: Given a unit i with variables X and Y related by model M (which is defined by a set of equations), the counterfactual value of Y for an individual i when $X = x$, $Y^x(i)$, is given by $Y^x(i) = Y_{M_x}(i)$. Where M_x is the modified M model with X set to $X = x$. For a binary treatment X, suppose we have observed $X = 0$ and $Y = y$, the counterfactual effect can be expressed as

$$E(Y^1 \mid X = 0, Y = Y^0 = y) \tag{2.10}$$

where Y^1 is the hypothetical outcome (potential outcome) if treatment had been $X = 1$. This hypothetical outcome is conditioned on the fact that the treatment X has already been observed as $X = 0$ resulting in the outcome Y^0.

Counterfactual analyses allow the prediction of features of a system as if the system had been different (Hernan et al. 2019). They can answer causal questions at the individual level where intervention queries may struggle due to the fundamental problem of causal inference. Hence the result of a counterfactual query is an actual value of the variable, while intervention queries are typically answered in probabilistic terms (Pearl et al. 2016). Pearl et al. (2016) note that interventional query estimates could be used to approximate counterfactuals in certain applications. For a comprehensive treatment of how SEMs encode counterfactuals and approaches to counterfactual analysis, see (Balke and Pearl 2013; Pearl et al. 2016). For practical examples and applications of counterfactual estimation, see (Bottou et al. 2013; Pan and Qiu 2021)

2.2.3 Identification, Confounding, and Covariate Selection

Causal effect identifiability involves assessing whether causal effects can be inferred from data given a structural model (Galles and Pearl 1995; Pearl 2000; Tian and Pearl 2002; Hünermund and Bareinboim 2023). Elwert (2013) aptly describes identification analysis as determining whether and when it is possible to strip an observed association of all spurious components. Causal effects are generally identifiable in the absence of unmeasured confounders (Tian and Pearl 2002). Confounding is a primary challenge in causal inference using non-experimental data.

Consider again the graph in Figure 2.5a where Z is a confounder in the relationship between X and Y. X has a direct effect on Y, but this effect is confounded by the variable Z which affects both X and Y. To estimate the causal effect of X on Y, we want to measure $P(Y \mid do(X))$ and not the conditional $P(Y \mid X)$, so the direct effect $X \rightarrow Y$ needs to be disentangled from this confounding situation. With reference to a causal graph, to estimate causal effects, we have to measure effects along causal (front-door) paths alone ($X \rightarrow Y$, not $X \leftarrow Z \rightarrow Y$).

In many types of analyses this is achieved by "adjusting" for the confounding set of covariates, Z. In practice, adjusting for Z can be achieved by controlling Z in an experiment, or by including Z in a statistical model such as a regression model, or conditioning on Z in a statistical analysis.[9] Using graphical causal models, adjustment can be done via the backdoor adjustment formula described in Equation 2.9. The challenge then becomes – how to identify the sufficient set Z', for adjusting for confounding? Pearl's (Pearl 1995, 2000) back-door criterion is a useful graphical criterion for determining the sufficient covariate set Z'.

Definition 2.2.6 Back-door criterion: *Given a set of variables $\boldsymbol{V}$ and its corresponding DAG G, a set of variables $Z' \subset V$ satisfies the back-door criterion relative to X and Y in G if:*

1. *No variable in Z' is a descendant of X.*
2. *Z' d-separates X and Y. That is Z' blocks every path between X and Y that contains an arrow into X (back-door path).*

Without the aid of a structural causal model, the process of identifying Z' becomes notably less certain. VanderWeele and Shpitser (2011) recount the debate by experts on whether all pre-treatment covariates should be adjusted for. While adjusting for all measured covariates was once considered conventional wisdom in many fields, it has been shown to be a potentially hazardous general rule that can result in biased inferences in certain cases (Greenland and Pearl 2011; Pearl and Mackenzie 2018; VanderWeele 2019; Tafti and Shmueli 2020). For example, adjusting for colliders or mediators can be counter-productive (See Figure 2.3 for reference) . Adjusting for a collider can open a causal path where none exists leading to the estimation of spurious causal effects. On the other hand adjusting for a mediator could block the very causal path

[9] Z here corresponds to the sufficient set Z'.

one is interested in measuring since the effect of interest may completely flow through a mediator. A simple example is a case such as Figure 2.3c, where adjusting collider Z opens the path between X and Y. Estimating the causal effect of X using $P(Y \mid X, Z)$ rather than $P(Y \mid X)$ in this scenario will lead to the wrong conclusion that X is a cause of Y. Notice that X is marginally independent of Y (when Z is not considered) according to the d-separation principle. It is necessary to consider principled identification criteria when trying to estimate causal effects from observational studies as over-adjustment or under-adjustment of covariates can both be problematic[10] (Greenland and Pearl 2011; Pearl and Mackenzie 2018; VanderWeele 2019; Hernán and Robins 2018; Tafti and Shmueli 2020). Elwert (2013) concisely reviews several causal effect identification criteria.

Pearl's backdoor criterion provides graphical conditions for when and how covariate adjustment can be used to sufficiently control for confounding. Thus it is a tool for covariate selection when given a graphical model. Given this criterion, it is easy to determine that variable Z is a sufficient adjustment set, given that the DAG in Figure 2.5a is true for the set of variables $\{X, Y, Z\}$. Hence we can compute the the effect of an intervention on X using $P(Y \mid do(X)) = \sum_z P(Y \mid X, Z)P(Z)$ from equation 2.9. Shpitser et al. (2012) generalize this criterion by proposing a 'complete' criterion for determining a sufficient set for covariate adjustment. Similar to the back-door criterion though, this criterion requires that all confounders are observed and measured. Pearl proposed another graphical identification criterion known as the the front-door criterion (Pearl 1995, 2000) together with an accompanying adjustment formula known as the front-door adjustment formula for a special case when a confounding variable is not observed, but a known mediator exists in the causal path between the treatment and

[10]Over-adjustment and under-adjustment are sometimes referred to using the terms included variable bias and excluded variable bias respectively.

the outcome. In such a case covariate adjusment can be performed using the front-door adjustment formula (Pearl et al. 2016).

In some cases, for example high-dimensional data, it may be unrealistic to expect a precise understanding of the underlying causal mechanism in the dataset such that a precise graphical model could be constructed. In such scenarios where there is limited knowledge of the true causal structure of a process, VanderWeele and Shpitser (2011) propose the disjunctive cause criterion which only uses knowledge of whether a variable is a cause of the treatment X, or the outcome Y to determine if it belongs in the set of variables to be adjusted.

It is worth noting how the back-door adjustment formula has only conditional probabilities and no do-operator. This means that the expression can be evaluated using observational data. This is the purpose of graphical adjustment formulas – to rid interventional expressions of the do operator so that interventional distibutions can be computed from observational data. Pearl's do-calculus (Pearl 1995, 2000, 2009b, 2012) provides a set of operations for deriving adjustment formulas given a graphical model. The equations of the do-calculus together are sound and complete for determining the identifiability of causal effects from data given a graphical causal model of the data distribution (Huang and Valtorta 2006; Shpitser and Pearl 2008; Pearl 2012). Probability distribution equations containing the do-operator can be resolved into equations without the do-operator using the do-calculus, enabling the ability to predict the effect of interventions without having to intervene in the real world.

Chapter 3

Causal Inference Tasks

Most causal investigations can be categorized under two general groups of tasks; treatment effect estimation and causal structure modeling. The first category involves estimating the magnitude of the effect of one variable (the presumed cause or treatment) on another variable (the outcome or response) in a system under study. The second category of causal learning tasks has to do with eliciting and representing the structure of the causal relationships between a set of variables. This structure delineates beliefs and assumptions about the direct connections between variables, the directions of influence, and how influence propagates through the system. This chapter provides a review of prominent methods, as well as emerging techniques used for both causal inference tasks.

3.1 Treatment Effect Estimation

With data obtained from randomized controlled experiments, average treatment effects (ATE) can be estimated using the expression in Equation 2.2. In observational studies or data obtained from "imperfect" experiments, other techniques need to be employed to mitigate confounding. Methods that adjust for confounding generally produce conditional ATE (CATE) estimates. This discussion to a degree aligns with the classification of treatment effect estimation methods in Guo et al. (2020) by grouping these methods

into the following classes: (1) traditional methods addressing ignorability, (2) traditional methods relaxing the ignorability assumption, (3) advanced estimation methods, and (4) adjustment formulas using graphical criteria. Methods that employ graphical identification criteria such as the backdoor criterion were discussed briefly in Section 2.2.2.

3.1.1 Traditional Methods Addressing Ignorability

The ignorability assumption (see Section 2.1.1) implies that there are no unmeasured covariates which affect the treatment selection and the outcome. Randomized Controlled Trials (RCTs) are widely held as the "gold standard" for causal inference by many because confounding influences can be nullified through randomization. This section focuses on methods that attempt to address ignorability by adjusting for confounders. These approaches allow causal effects to be estimated from observational data. However they can also be useful in experimental settings due to the difficulty in designing and executing perfect experiments in many scenarios. Issues of selection bias, confounding bias or covariate imbalance may yet arise even in experimental settings (Rosenberger and Lachin 2015).

Matching and Propensity Scores: The problem of covariate imbalance arises when the distribution of certain covariates is significantly uneven between the treatment group and control group of a sample under study. In a binary treatment setting, matching involves selection by pairing treatment and control units. The goal of matching methods is to eliminate copious differences between treatment and control samples on important covariates that may influence outcomes (Pattanayak et al. 2011). This mitigates confounding by ensuring that treatment and control units are directly comparable, and any differences can be attributed to the causal effect of the treatment.

There are several strategies for executing matching. For a wider discussion on the topic refer to Gu et al. (1993); Austin (2014); de los Angeles Resa and Zubizarreta (2016). A popular metric used for matching is the propensity score. The **propensity score** is the probability of a unit being selected for treatment conditional on a set of covariates. The propensity score for a sample unit i, π_i, is given by $\pi_i = P(X_i = 1 \mid Z_i)$, where X is the treatment variable and Z is ideally the set of variables that is sufficient for control of confounding in this setting. In Propensity Score Matching (PSM), matching of treated and control units is performed based on similarity of propensity scores. Propensity scores simplify the matching process by allowing matching to be done based on one derived value rather than the all values of a set of covariates. For a detailed discussion on matching techniques and propensity scores, see Steiner and Cook (2013); Rosenbaum (2020).

Re-weighting: Similar in objective to matching methods, re-weighting methods enable the estimation of average treatment effects by addressing confounding through covariate balancing. Re-weighting methods achieve balance by creating a pseudo-population where all units have equal probability of being treated given a set of covariates. Under-represented samples in treatment or control groups are up-weighted, while over-represented samples are down-weighted based on the values of relevant covariates. Propensity scores are also commonly used for re-weghting (Imai and Ratkovic 2014). **Inverse Probability of Treatment Weighting (IPTW)** is a re-weighting scheme that involves re-weighting units in the samples by the inverse of the estimated propensity score. In a binary treatment variable case, the units under treatment are assigned weights of $1/\pi_i$, and the control units are assigned weights equal to $1/(1-\pi_i)$, where π_i is the propensity score of unit i in this simple case and is obtained from the exposure/treatment model $g(Z) = P(X = 1 \mid Z)$. The goal is to estimate the average causal effect, and this is given by $E(Y^1 - Y^0) = E\{f(X = 1, Z) - f(X = 0, Z)\}$.

Hence IPTW mitigates confounding by correcting for the contribution of each unit in the sample by the assigned weight (Chatton et al. 2020). For more elaborate discussion on re-weighting and IPTW see Mansournia and Altman (2016); Thoemmes and Ong (2016); Chesnaye et al. (2022).

Sub-classification: Matching and re-weighting methods enable the direct estimation of the average treatment effect (ATE) using an expression of the form in Equation 2.2. In some applications, the conditional average treatment effect (CATE) is a more useful measure of causal effects because treatment effects may differ systematically throughout the population based on the values of certain covariates. In such cases, there may exist sub-groups of the population under study that are affected in different ways by the same treatment (heterogeneity). Sub-classification or stratification methods allow the estimation of CATE by first creating approximately homogeneous subgroups of the population using the values of relevant covariates. ATE can be estimated by taking a weighted average of the estimated CATE for all sub-group, weighting by the proportion of observations in each sub-group. Propensity scores are also often used as a basis for sub-classification. Rosenbaum and Rubin (1984); Lunceford and Davidian (2004); Brand and Thomas (2013); Imbens and Rubin (2015) provide a detailed discussion on heterogeneous causal effects and methods for sub-classification.

Regression adjustment: Regression methods are extremely popular for modeling the relationships between variables. They are often the tool of choice for economists and social scientists and are can be useful for the investigation of causal effects. Consider the simple linear regression model $Y = \alpha + \beta X + \epsilon$ that relates a continuous outcome Y to a treatment X, where ϵ is the error term, and α is the intercept. The coefficient β is by default a naive causal effect estimator. If it can be assumed that the functional form of the regression model is correctly specified, and if X is randomly assigned, then

β is an unbiased estimator of the causal effect of X on Y (Schochet 2010; Hernán and Robins 2018; Funk et al. 2011; Morgan and Winship 2015).

When treatment is not randomly assigned, it is necessary to adjust for the sufficient set of confounders in the regression model. This can be done in multiple linear regression by including the variables in the sufficient set Z' in the regression model, as in $Y = \alpha + \beta X + \delta_1 Z_1 + \delta_2 Z_2 .... + \delta_n Z_n + \epsilon$. In this model, δ_1 to δ_n isolate the effects of the covariates Z_1,Z_2,...$Z_n \in Z'$ from the primary effect of interest β. Hence β can be interpreted as the mean effect of X on Y conditional on the set Z'. For a comprehensive discussion on the conditions that are necessary for such causal interpretations of linear regression coefficients, refer to Wooldridge (2015); Rebonato (2016); Morgan and Winship (2015).

Regression-type models are particularly important when dealing with continuous treatments and outcomes where the direct estimation methods discussed previously can become problematic as demonstrated in Chapter 11 of Hernán and Robins (2018). It is important to note that without causal assumptions, regression functions are merely descriptive. Pearl et al. (2016) highlight the difference and inter-relationship between regression equations and structural equations models which have causal assumptions embedded in them.[1]

G-methods: The generalized methods (G-methods) are a family of methods introduced by Robins (1986) which require less restrictive conditions for identification of causal effects than standard regression models (Naimi et al. 2017). They enable the modeling of more complex causal effects such as effect modification (mediation analysis) and time-varying effects (Daniel et al. 2011; Coffman and Zhong 2012; Wang and Arah 2015). Models under this family include marginal structural models (MSMs) and the G-formula. MSMs model the mean of a potential/counterfactual outcome. So instead

[1]Regression models are generally considered a type of structural equations model.

of modeling $E(Y)$ directly, $E(Y^1)$ and $E(Y^0)$ are modeled separately for a binary outcome. The average causal effect is then obtained by taking the difference between the expected values of the potential outcomes. MSMs are often used in conjunction with IPTW and can be useful for modeling time-varying causal effects using longitudinal data (Thoemmes and Ong 2016).

As a maximum likelihood estimation approach for the G-formula, the G-computation approach involves the estimation of 'the full data set' comprising all potential outcomes of the observed units given various treatments (Snowden et al. 2011). The causal effect can be estimated from the full data set which is often obtained by first fitting a regression model using the observed data (the outcome model), and using this model to predict counterfactual outcomes under different treatments. The outcome model is given as $Y = f(X, \mathbf{Z}) = E(Y \mid X, \mathbf{Z})$. For detailed discussion on MSMs and other G-methods, see (Hernán and Robins 2018; Naimi et al. 2017, 2021).

Doubly robust estimators: Using IPTW, if the propensity score model is correctly specified, the causal effect estimator is unbiased. On the other hand, using the G-formula or an outcome regression model, if the outcome regression model is correctly specified, the estimator is unbiased. Doubly robust estimators combine the propensity score model $E(X \mid \mathbf{Z})$, and the outcome regression model $E(Y \mid X, \mathbf{Z})$, such that only either of both models needs to be correctly specified for the estimator to be unbiased (Funk et al. 2011; Hernán and Robins 2018). The estimator obtained is said to be robust to the misspecification of one of the two component models (Naimi et al. 2021).

3.1.2 Traditional Methods Relaxing Ignorability

The methods in the previous sub-section depend on the credibility of the ignorability assumption with respect to the specific dataset. This sub-section briefly considers research designs for causal inference in settings where there may be unobserved or

unidentified confounders in quasi-experimental settings. Such designs are often used in scenarios regarded as natural experiments in the economic, biomedical and social science literature (Meyer 1995; Kim and Steiner 2016; White and Sabarwal 2014). Propensity score methods such as matching and re-weighting are also commonly employed in these settings when the ignorability assumption is considered to be reasonable.

The methods considered here exploit unique variation patterns in certain measured covariates which may influence treatment assignment, or other special conditions which create approximate localized experiments such that all confounders do not have to be observed. These methods include *Regression Discontinuity Designs (RDD)* (Imbens and Lemieux 2008; Lee and Lemieux 2010; Stevens 2016), *Difference-in-Differences(DiD)* (Lee and Kang 2006; Dimick and Ryan 2014), and *Instrumental Variables(IVs)* (Imbens 2014; Martens et al. 2006; Angrist and Krueger 2001; Angrist et al. 1996).

In observational studies, these approaches have a restricted scope of applicability due to their reliance on particular study designs or quasi-experiments. RDD relies on a cut-off point within a narrow range of the treatment values where treatment assignment can be assumed to be randomized. DiD uses data that includes similar groups of treated and control units before and after an intervention. IV methods exploit random variation in a special variable called the instrument which directly affects treatment but indirectly affects the outcome only through its effect on treatment. These approaches to causal inference are well established in the social sciences The reader is referred to Remler and Van Ryzin (2021) for a more elaborate discussion.

3.1.3 Advanced Estimation Techniques

This section highlights advanced approaches that build on the fundamental principles for causal effect estimation, and allow the use of non-parametric models and machine

learning for the estimation of causal effects. Parametric estimators are widely used because of their simplicity, and the ease of interpretation of their model parameters. However, they are limited because of the restrictions they impose on the data distribution, e.g., linear functional form and Gaussian errors in ordinary least squares regression. Model misspecicification issues arise when parametric methods are used to model data distributions which substantially deviate from their assumptions and can lead to biased parameter estimations. To mitigate model misspecification issues, non-parametric and semi-parametric modeling approaches are gaining wider adoption for causal effect estimation (Imbens 2004; Hill 2011; Curth and van der Schaar 2021). Thoemmes and Ong (2016) highlight the potential problems with the common practice of using parametric models like the logistic regression for propensity score estimation in methods like PSM, IPTW, and weighted regression. A misspecified logistics regression model would not remove all confounding bias even if the confounders are included in the model. One way to mitigate this problem is to use non-parametric machine learning models to estimate the propensity score (as in Westreich et al. 2010; Maguire et al. 2007) or for assessing covariate balance (as in Linden and Yarnold 2016). Similarly, the use of tree ensemble methods for the estimation of counterfactuals in G-computation methods have been explored (as in Austin 2012).

Machine learning methods have demonstrated remarkable success in predicting various estimands across numerous applications and effectively handling large datasets. Thus, the growing trend of employing machine learning to support the estimation of causal effects comes as no surprise. (Alaa and Schaar 2018; Diaz 2020; Curth and van der Schaar 2021). However, procedures that use machine learning methods for the estimation of causal effects have to be carefully designed as machine learning models by themselves can be very poor estimators of causal parameters (Chernozhukov et al.

2017b; Alaa and Schaar 2018; Balzer and Petersen 2021). Naimi et al. (2021) emphasize this point through a simulation study that investigated bias in causal estimation under several scenarios. Rolling and Yang (2014); Alaa and Schaar (2018) suggest guidelines for principled choice and design of procedures and algorithms which employ non-parametric models in the estimation of causal effects.

In light of the escalating demand for modeling techniques capable of deciphering large, high-dimensional observational datasets and the wealth of big data in contemporary information systems, a pivotal challenge in causal research lies in the automated selection of a covariate set Z' that is approximately sufficient for confounding adjustment within such data. Works by Belloni et al. (2012); Belloni and Chernozhukov (2013); Belloni et al. (2014b,a, 2016); Urminsky et al. (2016); Chernozhukov et al. (2017b,a); Belloni et al. (2017); Chernozhukov et al. (2018) constitute a string of papers that seek to address "sparse" covariate selection in high-dimensional data using machine learning techniques.[2]

Some of the methods considered in this section are direct extensions of the traditional methods in Section 3.1.1 but using non-parametric estimators instead. For example, the Bayesian additive regression trees (BART) model (Hill 2011) and its extension, the Bayesian Regression Forests (Hahn et al. 2020), are non-parametric alternatives to linear parametric regression adjustment. The term **doubly robust learners** is sometimes used when non-parametric models are used flexibly in doubly robust estimation procedures.[3]

Regularized Regression and Double Selection: Regularized regression methods such as the least absolute shrinkage and selection operator (LASSO) can be used for principled variable selection for covariate adjustment (Belloni et al. 2012; Belloni

[2]sparsity here refers to a scenario where only a few covariates affect the outcome (Athey and Imbens 2016)

[3]see (Dudik et al. 2011; Jacob 2021).

and Chernozhukov 2013; Belloni et al. 2014a; Urminsky et al. 2016). The goal is to select a (sparse) set of covariates Q, which are most "relevant" to the outcome variable Y, from the set of all measured covariates Z. A further "post-LASSO" step of using a standard OLS regression of Y on Q is recommended to minimize bias in the coefficients of the covariates (Belloni and Chernozhukov 2013).

Using predictive variable selection directly for confounder selection can be problematic (Diaz 2020). This is because these methods are based on identifying strong correlations. As a result, further steps need to be taken to ensure that the most relevant variables are chosen in order to increase the likelihood that these are causally influential variables. In a related approach often described as Double Selection, regularized regression is not only used to select the set of covariates Q relevant to the outcome Y, but also the set of covariates R, relevant to the treatment X (Belloni et al. 2014b,a). These two sets of covariates are combined and assumed to form the sufficient set of variables $Z' = Q \cup R$, where $Z' \leq Z$, and Z is the original set of covariates. This approximate sufficient set Z', is then used in a final regression step for adjusting for confounding in the estimation of treatment effects.[4] This method minimizes omitted-variable bias and also provides robustness by allowing for imperfect variable selection in either selection step (Belloni et al. 2014b,a). Belloni et al. (2017) formally generalize these regularized regression approaches to allow the use of a wide variety of machine learning methods as long as they are good approximators of the data distribution and do not overfit.

Double Machine Learning: Double/debiased machine learning (Double ML or DML) takes the concept of Double Selection a step further by exploiting the predictive ability of machine learning algorithms for estimating causal effects through a set of carefully designed procedures. The procedures are set up to guard against additional

[4]Notice how the Double Selection method for covariate selection seemingly aligns with the disjunctive cause criterion mentioned in Section 2.2.3.

challenges brought about by using complex ML models for causal effect estimation including regularization and overfitting. They creatively use sampling techniques to minimize bias in the estimation of treatment effects and to produce an estimator with desirable theoretical properties such as consistency and rate of convergence (Chernozhukov et al. 2017b,a, 2018). The basic DML procedure assumes the partially linear structural form described in the following two equations (Chernozhukov et al. 2017b):

$$Y = X\theta_0 + g_0(Z) + U, \qquad E(U \mid Z, X) = 0 \tag{3.1}$$

$$X = m_0(Z) + V, \qquad E(V \mid Z) = 0 \tag{3.2}$$

where Y is the outcome, X is the treatment, θ_0 is the treatment effect parameter to be estimated, Z is the set of covariates, g_0 and m_0 are functions which relate the Z to Y and X respectively, and U and V are disturbances or error terms. The functions g_0 and m_0 can be estimated using machine learning models such as random forests, support vector machine, and neural networks. The procedure continues as follows: (1) use ML to model Y as a function of Z and predict $\hat{Y}$, (2) use ML to model X as a function of Z and predict $\hat{X}$, (3) regress the residuals from (1), $Y - \hat{Y}$, on the residuals from (2), $X - \hat{X}$, to get an estimate of θ_0. The procedure guards against confounding, reduces regularization bias, and strives to satisfy the desirable Neyman-orthogonality condition (Chernozhukov et al. 2017a; Witlox and Naghi 2018). A cross-validation technique known as cross-fitting is used to avoid overfitting and minimize bias.[5]

Targeted maximum likelihood estimation (TMLE): TMLE (Van Der Laan 2010) is a doubly robust estimation method that starts out like G-computation, but includes a "targeting" step which incorporates the selection/exposure mechanism, for

[5]see also (Jung et al. 2021) for a direct connection of SCM and DML, and discussion on the theoretical properties of DML.

minimizing bias in the target parameter (ATE) (Gruber and Van Der Laan 2009; Schuler and Rose 2017). This model of the selection mechanism is used to update the initial G-computation estimates iteratively (Van Der Laan 2010; Chatton et al. 2020). TMLE often involves the use of adaptive machine learning and ensembling algorithms such as the Super Learner (Van der Laan et al. 2007) for estimation.

Metalearners: Metalearners are a superclass of methods for estimating the CATE function which allow flexible choice of machine learning algorithms. Künzel et al. (2019) describe Metalearners as meta-algorithms that build on base machine learning algorithms which are designed for prediction tasks (regression and classification), to estimate CATE. In general, the Metalearners take advantage of the estimation capabilities of machine learning algorithms, while remaining model-agnostic and allowing the choice of arbitrary ML models in their procedures.

Jacob (2021); Künzel et al. (2019) provide a detailed discussion on the various meta-algorithms considered as Metalearners. The T-learner is a Metalearner which involves the two-step approach of first estimating the conditional mean of the outcomes separately for treated and control units using different models, and then taking the difference to give the estimate of the treatment effect. The S-learner uses a single model to estimate the outcome using the treatment and covariates. The CATE estimate in this case is the difference in the predicted values of the outcome at different levels of the treatment, with all other covariates held constant.

The T-learner and S-learner can suffer from sample imbalance and poor choice of outcome model. Künzel et al. (2019) propose an expansion of the T learner, the X-learner, which estimates individual treatment effects (ITEs) in its first step,[6] and then estimates CATEs from the ITEs of the treated and untreated groups. Jacob (2021)

[6]Recall that ITEs cannot be directly calculated because only one potential outcome is observed (see section 2.1). X-learners estimate the unobserved outcomes using models created with observed outcomes, and then use them as if they are actually observed to estimate ITEs from both potential outcomes.

include in their classification of Metalearners, the set of Double-ML-esque estimators known as as R-learners (Nie and Wager 2021), and the doubly robust estimators referred to as as DR-learners (Kennedy 2020). Curth and van der Schaar (2021) suggest a reclassification of Metalearners, aligning them with their underlying theoretical foundations, while also offering valuable insights into the selection of algorithms, with a particular focus on the use of neural networks and meta-algorithms in the context of causal Metalearning.

Causal Trees and Forests: The sub-classification or stratification method discussed in Section 3.1.1 can be used for dividing the units under study into multiple groups so as to learn heterogeneous treatment effects (CATE) at different values of covariates. This approach is feasible when there are a reasonably low number of relevant covariates and a relatively large number of measured observations such that there are enough observations in each group after splitting on covariates to reliably calculate treatment effects within each subgroup. When the data is high-dimensional, then this approach can become problematic. Athey and Imbens (2016) note that unlike most other machine learning methods, decision trees produce a partition of the population based on covariates. They take advantage of this property to propose the Causal Tree method for partitioning a population based on relevant covariates, and the estimation of conditional average treatment effects within the partitions under the assumption of randomized treatment assignment given the covariates.

Decision tree algorithms such as Classification and Regression Trees (CART) are not built for causal inference. As a result, Causal Trees modify CART by using an "honesty" condition, and changing the splitting criterion from minimizing the prediction error, to maximizing the heterogeneity in treatment effects between leaf nodes in

the tree (Jacob 2021; Powers et al. 2018).[7] Honesty is an important quality of Causal Trees which refers to the separation (especially in terms of the information used) of the modeling step (or initial partitioning step), from the estimation step given the model structure. In the estimation step,[8] the difference between the conditional mean of the treatment and control groups within each final partition is the estimated CATE (Powers et al. 2018). An additional benefit of Causal Trees is the ability to obtain confidence intervals for estimated treatment effects, and also the interpretability of decision trees (Athey and Imbens 2016). Wager and Athey (2018) expand and improve this approach by ensembling Causal Trees into Causal Forests, similar to Random Forests. Athey and Wager (2019) demonstrate a practical application of Causal Forests.

Neural Network Approaches: The success of deep learning models has led to widespread adoption of neural network models in many applications (Sejnowski 2020). An advantage of neural networks is the flexibility in constructing modeling structures and the ability to automatically discover representations and features for prediction from unstructured raw data. Neural network models can be plugged into the previously discussed model-agnostic procedures for estimating causal effects, and used as an estimator of causal parameters at various stages of the procedure. However, some authors have recently explored deep neural networks as a special approach for causal effect estimation, investigating how to optimize them for causal effect estimation tasks.

[7] This heterogeneity of treatment effects is maximized by adjusting the mean square error (MSE) to an alternative measure of the expectation of MSE over test and estimation samples – see (Athey and Imbens 2016).

[8] This separation is achieved by dividing the dataset into two parts; one to be used for each step of the process. One part is used to construct the structure of the model through recursive splitting (like in a decision tree). The learned structure is then used to split the other part of the dataset before estimating CATE in each subgroup obtained.

Hartford et al. (2017) use deep learning for estimation tasks in the Instrumental Variable (IV) framework. Shi et al. (2019) propose two adaptations of standard multilayer perceptron (MLPs) for constructing the propensity score and outcome models before using those in a downstream causal effect model, as part of the 2-step regression approach similar to the Double Selection procedure. The first adaptation is a modified architecture (Dragonnet) employing a representation layer, while the second adaptation is a TMLE-based regularization technique (targeted regularization). This approach is similar to Shalit et al. (2017)'s CFR and TARNet, which are in turn improvements on Johansson et al. (2016)'s BNNs for estimating ITEs and counterfacuals.

Some other modifications of neural networks for causal effect estimation are the Variational Autoencoder based CEVAE (Louizos et al. 2017) for estimating latent confounders and causal effects, the Generative Adversarial Network (GAN) based GAN-ITE Yoon et al. (2018) for estimating ITE. Yao et al. (2018) propose the SITE network which uses a representation learning network that preserves local similarity information and balances data distributions in treated and control groups, for the estimation of ITE. Garrido et al. (2021) propose using a neural auto-regressive density estimator (NADE) approach for modeling causal mechanisms and estimating effects according to principles of the SCM framework.

Shalit et al. (2017); Farrell et al. (2021); Koch et al. (2021) study the theoretical properties of neural networks for non-parametric estimation of causal effects under the usual assumption of ignorability, and establish valid inference for treatment and counterfactual effects using standard deep learning architectures. Farrell et al. (2021) further provide non-asymptotic bounds for such networks and demonstrate the utility of deep learning for causal inference using an empirical study. Koch et al. (2021) provide a detailed review of deep learning methods for causal effect estimation, and classifies the main approaches into deep outcome modeling methods, balancing through

representation learning methods, methods extending inverse propensity score weighting (IPW), and methods for adversarial training of generative models, representations, and IPW.

3.2 Causal Discovery

3.2.1 Causal Structure Modeling

The term *causal structure modeling* is used in this work to describe methodologies for the identification of the structure of causal relations among a set of variables, and the representation of the identified relationships. This structure is useful for expressing one's understanding of the causal interconnections between variables, without the need to delve into the specific magnitudes of their effects. SCMs are able to encode this type of information regarding the direct and indirect connections between the variables, and the direction of the flow of causal influence. This allows SCMs to be useful for inferring which variables are expected to change as a result of the manipulation of another variable. Graphical SCM representations are invaluable for expressing such structural relationships between variables in a system explicitly.[9]

Some form of causal structure modeling always precedes any attempt to estimate treatment effects. While many studies about treatment effect estimation do not explicitly declare a causal structure before attempting to estimate causal effects, the authors of these works whether they realize it or not implicitly assume a causal structure between the treatment X and the outcome Y. Often times the implicit structural causal model seems obvious. For example, when investigating the effect of a drug on symptoms, the model $drug \rightarrow symptoms$ is assumed rather than $drug \leftarrow symptoms$.

[9]Graphical SCM representations may include various forms of graphs ranging from basic representations like DAGs or BNs, to more complex representations such as ancestral graphs and cyclic graphs.

Nevertheless, the model *drug* $\leftarrow$ *symptoms* may be appropriate in a different study where the objective is to examine how different symptoms may influence the prescription or consumption of certain medications. Such implicit assumptions about causal structure apply even in cases where randomized controlled trials are used to investigate causal relationships.

Often researchers have to contend with a set of covariates $\mathbf{Z}$, which may confound the effect of X on Y. A good deal of the discussion so far in this chapter have had to do with identifying and dealing with possible confounders when investigating causal effects. As discussed in Section 2.2.3, having a graphical causal model of the process under study enables the identification of the sufficient set of covariates for the adjustment of confounding (assuming the causal structural model is correct). Figure 3.1 illustrates a typical graphical causal model an analyst may implicitly assume when estimating the effect of a variable X on another variable Y. The dashed edges represent arbitrary relationships (may be causal or non-causal, or no association at all) between covariates $\mathbf{Z}$ and the variables of interest.

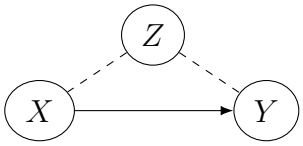


Figure 3.1: A typical form of implicit causal model for causal effect estimation tasks

Causal structure modeling, whether subconsciously or explicitly performed, is in reality one of the first steps of causal inference before any effect estimation tasks can be conducted. Depending on the research interest or application, it could be the only step needed but typically, it is a precursor to further causal analysis. Traditionally, the task of constructing some form of structural causal model for a set of variables is carried out in the absence of data using domain knowledge (see Koller and Friedman 2009; Pearl and Mackenzie 2018), a task sometimes referred to as *knowledge engineering*.

Tafti and Shmueli (2020) provide a set of guidelines for creating a causal diagram using background knowledge. In many causal inference tasks involving the estimation of causal effects, researchers are already convinced about the nature and structure of the causal relationship based on intuition, domain knowledge, experimental evidence or previous experience. However, it is often the case that this information is not clear. SCM theory makes it possible to learn aspects of causal structures from observational data under certain assumptions.

This section focuses on methods, techniques and algorithms that attempt to recover underlying causal structures from data. These techniques, referred to as causal discovery methods, enable the automation of causal structure modeling.

3.2.2 Causal Discovery Principles

Recovering causal structures from observational data usually requires additional assumptions to those introduced in Section 2.2.1. The causal Markov condition (CFC) and minimality condition together establish that every independence relation in a Bayesian network G is also present in its probability distribution P. However, this does not mean that the converse is necessarily the case. It is possible to have independence relations in P that are not present G. In such cases, G is said to be "unfaithful" to P. The faithfulness condition requires that G is faithful to P.

Causal Faithfulness Condition (CFC): A probability distribution P is faithful to its DAG G if every independence relation that exists in P is represented in G (G is a dependence map of the data). P is said to be faithful if there exists some DAG G to which it is faithful. There has been a debate about whether faithfulness is too strong a general assumption for causal modeling, but Korb and Nicholson (2008) explain that given that realistic examples of unfaithful distributions are not common, it is a methodological assumption that can be generally accepted unless there is good reason

to doubt that a faithful model exists for a particular distribution.[10] Examples of unfaithful distributions are the *chessboard problem* and the effect cancellation problem (described in Scheines and Sobel 1997; Guyon et al. 2007; Kubus 2015; Marx et al. 2021).

Note that to be faithful, P must also satisfy the Markov condition relative to G as well. Thus, for a faithful distribution there exists a DAG that encodes exactly all of its independence relations and no more. Together, the CMC and the CFC match the conditional independence relations in a probability distribution P, to the relations entailed by its causal graph or Bayesian network G. They guarantee the mutual correspondence between conditional d-separation and conditional probabilistic independence: $X \perp Y \mid Z \Leftrightarrow X \perp\!\!\!\perp Y \mid Z$ (Eberhardt 2017). This makes causal discovery possible.

Given the CMC and CFC conditions, as well as other standard assumptions such as causal sufficiency and acyclicity (see Section 2.2.1), it is possible to identify causal structures in a data distribution by testing for marginal and conditional independencies in data. Methods that learn causal structures through conditional independence constraints are known as constraint-based methods. In many cases, the direction of some edges in the causal Bayesian network cannot be identified uniquely from data. This is because more than one Bayesian network can encode the same set of conditional independence constraints. Employing constraint-based methods, a generalized DAG consisting of a mixture of edge types, including directed and undirected edges can be learned. This generalized DAG, known as a *Markov Equivalence Class* (MEC) graph, encapsulates all statistically equivalent DAGs relative to a data distribution.

[10] See also Weinberger (2018).

3.2.3 Constraint-Based Methods

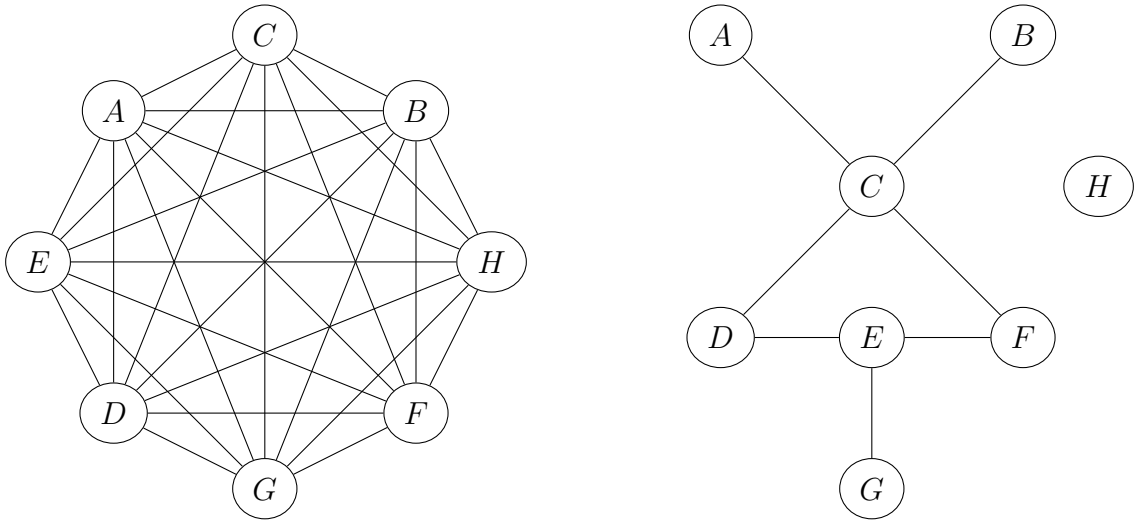
Constraint-based methods elicit causal graphs by exploiting conditional independence constraints in the data distribution.[11] Because multiple causal graphs can encode the same set of conditional independencies, constraint-based methods are often not able to identify a unique, completely directed causal graph such as a DAG or a Maximal Ancestral Graph (MAG) (Richardson and Spirtes 2002).[12] Usually, they return a *Markov Equivalence Class* (MEC) causal graph such as a Completed Partially Directed Acyclic Graph (CPDAG) or Partial Ancestral Graph (PAG) in which some edge directions may be undetermined or ambiguous. As a MEC graph, a CPDAG can represent a collection of several Markov equivalent DAGs, and may contain a mixture of edge types including directed and undirected edges. It is useful to think of MEC graphs as one compact representation of several causal models outputted by a causal discovery algorithm (Malinsky and Danks 2018). DAGs belonging to the same MEC are statistically indistinguishable based on independence relationships over the set of variables in a data distribution.

The most prominent constraint-based causal discovery algorithm is the PC algorithm (Spirtes et al. 2000). In the PC algorithm, two variables are considered to have a direct causal relationship relative to the set of observed variables if there is no subset of the remaining variables conditioning on which they are independent, under the assumptions of CMC, CFC, and causal sufficiency (Glymour et al. 2019). The PC algorithm starts with a complete (fully-connected) undirected graph (example in Figure 3.2a) from which it estimates the skeleton structure (example in Figure 3.2b) using

[11] For an overview of available conditional independence tests, refer to Kitson et al. (2021); Yu et al. (2016).

[12] Ancestral graphs, unlike DAGs allow modeling of causal relations in a distribution with latent variables, thus not requiring causal sufficiency (Zhang 2008).

conditional independence tests applied iteratively on the nodes in the graph.[13] The algorithm then proceeds to determine edge orientations by first finding unshielded colliders in *v-structures*. V-structures are variable triples, e.g. {X,Z,Y}, such that they are connected as an undirected chain-like structure as in X — Z — Y. In this structure, if Z was not part of the conditioning set that made X and Y independent, then Z is an unshielded collider (Glymour et al. 2019). Other edge orientation rules allow the algorithm to asymptotically converge to an MEC graph of the type CPDAG. For example, the edge propagation rule allows the algorithm to find directed chain structures by converting a structure like X → Z — Y to X → Z → Y.



(a) Fully connected graph (b) Skeleton structure

Figure 3.2: Fully connected and skeleton of the BN in Figure 2.1

Some other constraint-based algorithms have been developed to improve specific performance aspects of the PC algorithm or to relax some of its assumptions. For example, the FCI (Spirtes et al. 2000) and RFCI algorithms (Colombo et al. 2012) are generalizations of the PC algorithm that are able to perform causal discovery

[13]The skeleton is the undirected graph obtained after some edges have been eliminated using conditional independence testing.

in the presence of hidden or latent variables. This generalization is useful because causal sufficiency may be difficult to guarantee in practice (Yu et al. 2016; Spirtes et al. 2000). These algorithms make it possible to perform causal discovery in *causally insufficient* datasets, although the ancestral graphs they produce are more complicated (Triantafillou and Tsamardinos 2016).

3.2.4 Score-Based Methods

Score-based methods learn causal structures from data by searching for an optimal graph that maximizes the likelihood of observing the data distribution. They make use of a predefined score which measures how well a candidate graph fits the data, to search for an optimally scoring MEC graph. There are many different search strategies and score functions which could be used in this class of causal discovery algorithms, and scores could be Bayesian or information theory based (Maathuis and Nandy 2016; Huang et al. 2018; Kitson et al. 2021). The most prominent score-based method is the Greedy Equivalence Search (GES) algorithm (Chickering 2003).

The GES algorithm learns the Markov equivalence class graph by performing a two-phase search for an optimally scoring DAG while penalizing the complexity of the DAG. The forward phase starts with an empty graph (only nodes, no edges) and adds single edges sequentially with the goal of achieving the maximum improvement of the score each time. The backward phase starts with the best scoring DAG produced by the forward phase and removes single edges sequentially to achieve maximum score improvements until no more improvements can be made.

Score based methods depend on similar assumptions as constraint-based methods, but it may be harder to meaningfully relax some of the assumptions. For example, it is apparently more challenging to adequately relax the causal sufficiency assumption with score-based algorithms, but algorithms exist within the constraint-based domain

that do this and successfully recover MAGs from causally insufficient data.[14] Being a combinatorial optimization problem, these methods generally do not scale well as the number of variables becomes very large. Score-based methods may offer some computational performance advantages over constraint based methods by finding ways to limit the search space of candidate DAGs (Maathuis and Nandy 2016). Overall they are expected to converge to the same Markov equivalence causal graph as constraint-based algorithms asymptotically (Pearl et al. 2016).

3.2.5 Functional Causal Model Approach

This section explores a set of methods that exploit the functional causal model (FCM) representation for causal discovery. An FCM for {X,Y} describes the outcome Y as a function of its direct cause(s) X and an error or noise term ϵ:

$$Y := f(X, \epsilon, \theta) \tag{3.3}$$

where θ is the set of parameters of the function f, and the error term ϵ is assumed to be independent of the cause X.

Consider an example involving two continuous variables $\{X, Y\}$ where it is not known prior to investigation which is the cause and which is the effect. With only two variables, there are no conditional independence relations so constraint-based methods cannot be used to determine the direction of causal influence. One can proceed by trying to determine the asymmetry in the relationship between the two variables instead. This approach is based on the expectation that the model which correctly assigns the direction of causality will be less complex and more natural, and will be in accordance with the data generating process (Spirtes and Zhang 2016; Mooij et al. 2016; Glymour

[14] See discussions in (Ramsey et al. 2012; Triantafillou and Tsamardinos 2016; Yu et al. 2016).

et al. 2019). This will be a model that tries to recover effect from cause $Y = f(X, \epsilon)$, rather than cause from effect $X = f(Y, \epsilon)$.

Zhang et al. (2015b); Spirtes and Zhang (2016) show that the assumption of independence between the error term ϵ and the cause X, under certain structural conditions on the FCM, allows the determination of causal asymmetry between the two variables. It is also assumed that the two variables have a direct causal relationship and that there are no confounders. To exploit this for the two variable scenario, one can fit two models to characterize the relationship between X and Y in both directions of influence, and test to see in which model the error term is approximately independent of the hypothetical cause.

Shimizu et al. (2006) found that while many causal discovery algorithms require that the error terms are Gaussian,[15] assuming that they are non-Gaussian is particularly useful for finding the complete causal structure rather than the set of statistically equivalent structures typically recovered based on conditional independence. They propose the Linear Non-Gaussian Acyclic Model (LiNGAM) (Shimizu et al. 2006) which uses the statistical method known as independent component analysis for model discovery under the assumption that the causal model is linear in its functional form, and the error terms are non-Gaussian.

Hoyer et al. (2008)'s non-linear additive noise model extends this idea to non-linear models. They further show that non-linearity can be advantageous to causal identification as it helps to break the symmetry between cause and effect. Zhang and Hyvärinen (2010, 2009) generalize the approaches in LiNGAM and the non-linear additive noise model to propose the post-nonlinear (PNL) causal model which accounts for the non-linear effects of the causes and noise, specifically including measurement distortion in the model and allowing for Gaussian errors. Zhang and Hyvärinen (2009) further

[15]Independence tests often assume linear models and Gaussian error/noise (Hoyer et al. 2008). Non-linear, non-Gaussian tests however do exist - see Ramsey (2014).

discuss the model's identifiability conditions in detail and outline the five situations in which causal direction is not identifiable using this type of model. The scenarios however include the popular linear Gaussian case.[16]

Using FCMs or SEMs for causal discovery offers some advantages (Spirtes and Zhang 2016; Glymour et al. 2019). One advantage is that the faithfulness condition is not required, instead, the rather weak non-Gaussian error condition is relied upon (Eberhardt 2017). FCM methods can be used to distinguish between DAGs in the same equivalence class produced by constraint-based and score-based methods. They are particularly useful for causal identification in the two-variable case since conditional independence tests cannot be used in the two-variable case. Also, large samples may be needed to get reliable conditional independence tests, so FCM models may be preferred with small sample data. FCM methods can be used in the multivariate case by exhaustively testing every pair of variables in the set $\mathbf{V}$, although the challenge is that the complexity of such an exhaustive search increases super-exponentially (Zhang et al. 2015b).

3.2.6 Hybrid Methods and Other Bayesian Network Learning Approaches

Several other techniques have been explored for uncovering causal structures from data. Many of these techniques are hybrids of the methods discussed above. Most hybrid methods try to optimize the best of two worlds by combining two different approaches. One approach is to use a constraint-based approach to restrict the search space for a score-based method, leading to improvements in accuracy and scalability. The adaptively restricted GES (ARGES) algorithm (Nandy et al. 2018) and the Greedy Fast Causal Inference (GFCI) algorithm (Ogarrio et al. 2016) are examples of this

[16] See Eberhardt (2017).

hybrid approach. Functional methods can also be incorporated to find the orientation of undetermined edges in the Markov Equivalence class graphs produced by constraint-based methods(Zhang and Hyvärinen 2009).[17]

Another approach is to use interventional or experimental data when available in the causal discovery process so as to exploit the improved causal identifiablity of interventional data. An example of an algorithm that can incorporate interventional data is the Greedy Interventional Equivalence Search (GIES) algorithm (Hauser and Bühlmann 2012). Similarly, background knowledge can be incorporated into the search process via various means including imposing hard constraints on the search space to improve accuracy and computational performance (Tian and Pearl 2001; Perković et al. 2017; de Campos and Castellano 2007; De Campos et al. 2009). Some methods use standard Boolean satisfiability solvers to find causal structures after encoding prior knowledge as constraints in propositional logic (Triantafillou and Tsamardinos 2015; Hyttinen et al. 2013; Eberhardt 2017).[18] **Active learning** is a broad term encompassing the approach of integrating human expertise or empirical knowledge into an algorithmic causal discovery process. This approach can be used to improve edge determination after finding MEC graphs. Active learning can take various forms and the human input could come from targeted intervention experiments or expert/domain knowledge (Yu et al. 2016; He and Geng 2008; Masegosa and Moral 2013; Tong and Koller 2001; Ma et al. 2016; Hauser and Bühlmann 2014; Kitson et al. 2021).

The approaches for causal discovery outlined in the previous sub-sections employ combinatoric or search-based optimization techniques. Finding high-scoring DAGs from data is an NP-hard problem (Chickering et al. 2004). Thus these methods struggle to scale effectively with very high dimensional data. Aliferis et al. (2010a) outline

[17] See also Huang et al. (2020a) for a review of models and approaches for non-stationary data.

[18] Prior knowledge could also be obtained from conditional independence tests or graph search methods.

several approaches that have been developed to mitigate this scalability problem. These methods often employ some heuristic local neighborhood learning approach. A prominent hybrid method that combines heuristic local learning with constraint-based and score-based methods is the Max-Min Hill-Climbing (MMHC) algorithm. The algorithm first learns the skeleton of the causal graph via constraint-based method, and then orients the edges of the graph using a greedy hill-climbing score-based search.

There is a new promising approach when it comes to learning Bayesian networks from data in a manner that scales more effectively with increasing number of variables. Zheng et al. (2018) propose a reformulation of the score-based DAG learning problem from a combinatorial optimization problem to a continuous optimization problem. This allows the problem to be efficiently solved with standard numerical optimization algorithms such as gradient descent. They introduce a new constraint to the problem to enforce acyclicity in the learned graphs.[19] Their method, NOTEARS, has sparked the emergence of several methods for Bayesian network discovery via continuous optimization and especially using neural networks and deep learning. Vowels et al. (2022) provides a review of methods that take this approach. While this continuous optimization approach to learning Bayesian networks is promising and exciting, the theoretical properties and implications of the methods based on this approach are still being unpacked (Wei et al. 2020; Ng et al. 2022). It is not immediately clear what assumptions are necessary to interpret any recovered DAG as a causal BN – see discussions in (Reisach et al. 2021; Kaiser and Sipos 2022).

Several other approaches have also been explored for Bayesian network recovery from data including penalized regression (Bühlmann et al. 2014; Gu et al. 2019), information theory and entropy (Weilenmann and Colbeck 2017; Kocaoglu et al. 2020; Chaves et al. 2014), decision trees (Li et al. 2017), evolutionary algorithms (Contaldi

[19]The method actually learns linear SEMs first before translating to a causal graph.

et al. 2019; Dai et al. 2020), reinforcement learning (Huang et al. 2020b; Zhu et al. 2019), unsupervised learning (Brady 2020), dynamic programming and graph search algorithms (Koivisto and Sood 2004; Silander et al. 2006; Xiang and Kim 2013; Yuan and Malone 2013). For more detailed discussions, comparisons, and implementations of a broad range of algorithms for casual discovery and ayesian network recovery, refer to Kalisch and Bühlmann (2014); Singh et al. (2018); Neapolitan (2004); Martin (2019); Kalisch et al. (2012); Aliferis et al. (2010a); Vowels et al. (2022); Kitson et al. (2021); Ma and Statnikov (2017).

3.3 Towards Integrating Causal Inference and Machine Learning

This section briefly introduces two compelling approaches for integrating causal inference and machine learning methods.

3.3.1 Causal Feature Selection

Causal feature selection is a direct way to incorporate causal considerations into ML. Feature selection is a critical step in ML model training but most classical methods for feature selection are based on association. Guyon et al. (2007) in their seminal paper on the topic, explore causal approaches to feature selection and contrast them with classical statistical approaches. They outline the benefits that causal-based feature selection approaches can bring to a machine learning process. These benefits include robustness to violations of the assumption that source and target distributions are similar, increased parsimony of selected feature sets, improved interpretability, and enhanced data understanding.

Causal feature selection involves local causal discovery using observational data, and hence depends on the theoretical framework of structural causal models. The notion of **Markov blanket** is used to describe the set of most relevant causal features for predicting the outcome variable Y. The Markov blanket of Y, MB(Y), is a set of variables that d-separates Y from other variables which are not in the Markov blanket, hence shielding it from the influence of variables outside the Markov blanket (Guyon et al. 2007; Aliferis et al. 2010a). Using this notion, a goal of causal feature selection is to identify a minimal set of features S from the set of measured covariates in the data V, such that including any other variables in V but not in S in a model for predicting Y generally does not improve the prediction of Y.

For a causal Bayesian network that satisfies the causal Markov and causal faithfulness conditions, there is a unique Markov blanket relative to the outcome variable which includes the variable's parents (direct causes), its children (direct effects), and its spouses (direct causes of direct effects). A related set of variables to MB(Y) which is typically more parsimonious, is the set of parents & children of Y, PC(Y). This set includes only the direct causes and direct effects of the variable of interest. Guyon et al. (2007); Aliferis et al. (2010a,b); Yu et al. (2020); Pellet and Elisseeff (2008) explore these concepts in detail and present analyses and reviews of causal discovery algorithms, demonstrating their usefulness for causal feature selection.

3.3.2 Causal Representation Learning

Representation learning refers to the set of algorithmic learning methods that allow a computer to be fed with raw, unstructured data which is used to extract useful information in the form of representations that enable further learning tasks such as classification and regression (Bengio et al. 2013; Lecun et al. 2015). The ability to automatically discover useful feature representations from raw data is a major advantage

of neural networks and deep learning methods over traditional ML algorithms (Lecun et al. 2015). With the elevated interest in causal inference in the AI research community, the area of causal representation learning (CRL) is emerging as a central problem for AI and causality (Scholkopf et al. 2021). CRL involves the discovery of high-level causal features from low-level data and observations.

Causal representation learning furthers the quest to incorporate causal inference methodologies into machine learning to improve aspects of AI performance such as generalization. It can also be used to build new capabilities like counterfactual prediction into sophisticated ML methods. Specifically, new methods have been proposed to directly incorporate SCMs into neural networks, or in cases where the causal structure is unknown, to automatically learn the causal model before embedding it into the network (Zhang et al. 2020; Leeb et al. 2020; Yang et al. 2021). Yu et al. (2020) suggest a design approach for deep learning networks where the hierarchy of independencies in the input distribution is encoded in the hidden layers according to the structure of a Bayesian network automatically learned from the data. CRL makes possible the application and automation of causal reasoning in unstructured data forms such as images (Lopez-Paz et al. 2017). Scholkopf et al. (2021) provide an overview and wide-ranging discussion on causal representation learning.

Some of the methods discussed in Section 3.1.3 for the estimation of causal effects using neural networks involve some form of causal representation learning. The technique has been suggested for the determination of the direction of causality between a set of variables (causal discovery) in the context of disentangling causal representations (Bengio et al. 2019; Li et al. 2022). CRL is gaining increasing attention in the area of domain adaptation and transfer learning.

Chapter 4

Data-Driven Root Cause Analysis via Causal Discovery using Time-To-Event Data

4.1 Introduction

Root cause analysis (RCA) refers to structured investigations that are used to identify underlying causes of an observed phenomenon or event. In industrial processes RCA is used to discover the causes of process or device failures, and is critical for improving quality, reliability and safety (Rooney and Heuvel 2004; Vuković and Thalmann 2022). Traditional root cause analysis tools such as the cause and effect (Ishikawa) diagram are qualitative in nature. They rely on subjective judgments about factor relationships and often fail to address system-wide problems (Doggett 2005; Yuniarto 2012). Data-driven approaches to RCA have received growing attention recently (e.g. He et al. 2017a, 2019; Lin et al. 2020; Ma et al. 2021; Rocha et al. 2022; Thakar and Kalbande 2023). They take advantage of increasing data availability and provide quantitative tools for discovering the root causes of systemic problems.

Root cause analysis is fundamentally a causal problem. However, attempts to address data-driven root cause analysis in various domains have mostly been built on statistical methods and machine learning (ML) techniques based on association (e.g. He et al. 2017a,b; Samantha et al. 2018; Liu et al. 2018, 2020; Ma et al. 2021). Techniques that employ such methods for causal investigations should be carefully designed, and

application must be guided by robust causal theory (Alaa and Schaar 2018; Pearl 2019b; Balzer and Petersen 2021). Efforts to address RCA using a causal framework (as in Li et al. 2016; Chen et al. 2018; Alizadeh et al. 2018; Liu et al. 2020, Wang et al. 2023; Zhang et al. 2023) are mostly dependent on Granger causality (or some variant), which has been criticized for its limitations as a causal theory because of its dependence on prediction (Stern 2011; Maziarz 2015). Moreover, these methods are designed for time-series data. A goal of this study is to develop data-driven RCA methods that are based on a rigorous causal framework and can be applied to time-to-event (TTE) data. Such methods should allow flexible incorporation of statistical or ML estimation techniques in a principled manner in line with established causal theory.

Advancements in the field of causal inference have led to the development of the structural causal model (SCM) framework; a general theoretical framework for modeling and analyzing causal relationships (Pearl 2009a). Also, algorithms for learning the structure of causal relationships from data, known as causal discovery methods have received a lot of attention (Glymour et al. 2019). This work builds on such cummulative progress in the field of causal inference using observational data to develop a method for RCA that is applicable to censored TTE data.

The existing methods for data-driven RCA outlined above have been applied to regular cross-sectional and time series data. We focus on TTE data because of its popularity for studying failure trends in industrial process. Also, TTE data has received limited attention in causal discovery research. Unlike temporal datasets such as time-series data where observations are repeated measurements of the same quantity across a set of time intervals, TTE data consists of measurements of the duration from a pre-defined time of origin to the occurrence of an event of interest (often times some type of failure in industrial settings) for a set of units sampled from a population under study. This type of data has unique features that make it unsuitable to traditional

statistical and machine learning methods (Vittinghoff et al. 2006). For example, it is able to capture information about incomplete event observations such as censoring, and the outcome of interest consists of not only *whether* an event of interest occurred but also *when* the event occurred (Klein and Moeschberger 2003; Kartsonaki 2016). These features allow event data to be collected and analyzed in a way that avoids bias and loss of critical information. The field of survival analysis provides a suite of specialized techniques for modeling and analyzing TTE data (for an introduction, see Lemeshow et al. 2011; Smith 2017).

A challenge that causal learning methods must overcome to be successful for root cause analysis using censored TTE data is correctly handling the dual variable representation of the outcome of interest in censored TTE data typically consisting of the variables time of event T and status S, where T represents the duration before the event and S represents the status at time T for every sample unit in the data. Statistical associations between covariates and the observed event times in such data can be learned using techniques in survival analysis, however, the discovery of causal structures and algorithmic learning of root causes from censored TTE data has not been explored in existing literature to the best of our knowledge. This work exploits techniques in survival analysis for TTE data, but goes beyond statistical associations to reveal causal relationships in event data. The method proposed in this paper is especially suitable for investigations where events that occur up to a certain time-point is of primary interest, such as in the root cause analysis of product infant failure.

We propose a method for estimating a graphical model of the causal relationships between a set of covariates and a suitably defined outcome of interest. We refer to this graphical model as a *root cause graph* (RCG). Integral to the proposed method is the definition and estimation of an alternative single outcome variable for the dataset that becomes the focus of a subsequent causal discovery procedure. The RCG depicts

the structural mechanism that leads to the outcome of interest (often some measure of failure) relative to the set of observed variables. Graphical models like RCGs are visual, easy to interpret, and explicitly encode assumptions about the process under study.

Through RCGs learned from data, the proposed method is useful for summarizing answers to root cause investigations using a graphical depiction of how variables in the data are related. The causes of the observed event relative to the set of covariates in the dataset are made explicit. This is crucial for making decisions about how to intervene in a process to improve the outcome of interest. Since RCGs reveal the causal mechanism connecting all the variables in the data, potential intervention points suggested by this method are not limited to the root cause variables alone. As a result, other variables that mediate the effects of the root causes which might be easier to manipulate can be considered for intervention actions. Furthermore, recovering RCGs from data facilitates principled causal effect estimation. This aids decision making by enabling an evaluation of the effects on the outcome, of potential changes to covariates. This usage is demonstrated later in this paper.

The novel contributions of this study include: (i) a data-driven framework for root cause analysis using observational time-to-event data. (ii) a methodology for causal discovery using time-to-event-data. (iii) a two-part simulation framework for generating realistic time-to-event datasets from causal structures. (iv) a demonstration of how root cause treatment effect estimation can be improved using a principled approach informed by RCGs. The rest of this paper is organized as follows. Section 4.2 provides an overview of structural causal models and causal discovery. Section 5.3 presents the proposed methodology for RCG recovery and the TTE data simulation framework. Section 4.4 describes the nature of root cause analysis problems addressed by this study and summarizes the charateristics of the simulated datasets. Section 5.4

discusses the results obtained and demonstrates the use of the root cause graph for treatment effect estimation in the presence of mediation. Section 4.6 provides further discussion on the relevance of the proposed RCG recovery method and considerations for practical applications. The paper concludes in Section 5.5.

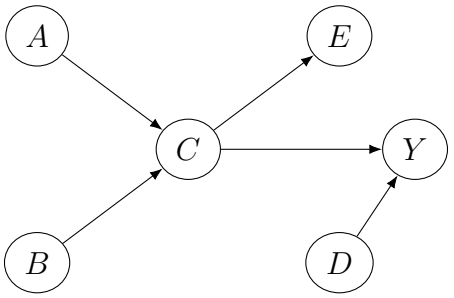
4.2 Background

4.2.1 Structural Causal Models

The Structural Causal Model (SCM) framework is a broad theoretical framework that combines features of the potential outcome framework, structural equations modeling, and probabilistic graphical modeling (Pearl 2009a, 2018). In the SCM framework, causal mechanisms between variables in a data distribution can be represented as graphical models known as Bayesian networks (BN), or by a corresponding set of non-parametric structural equations models (NPSEM) (Pearl 2009a; Maathuis and Nandy 2016). These representations describe the causal dependencies that exist between a set of variables. Both representations can be easily transformed from one to the other.

Figure 4.1 shows an example of graphical and NPSEM representations of the same causal structure of a data distribution consisting of variables A, B, C, D, E and Y. In the NPSEM (Figure 4.1b), each variable is a function of its direct causes and an error term (ϵ), and the unidirectional assignment operator := is used to depict the asymmetrical nature of causal relationships. In the BN (Figure 4.1a), error terms are commonly omitted for simplicity.

In probabilistic graphical models, the nodes in the graph represent the variables in a data distribution while the edges represent dependencies between pairs of variables. The Bayesian Network (BN) is a type of probabilistic graphical model that is based on



(a) Graphical (BN) representation

$$
\begin{aligned}
A &:= f_a(\epsilon_A) \\
B &:= f_b(\epsilon_B) \\
C &:= f_c(A, B, \epsilon_C) \\
E &:= f_e(C, \epsilon_E) \\
D &:= f_d(\epsilon_D) \\
Y &:= f_y(C, D, \epsilon_Y)
\end{aligned}
$$

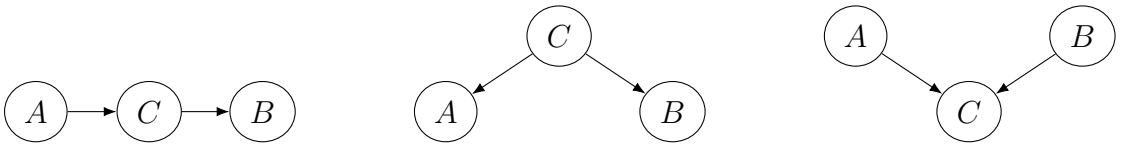
(b) NPSEM representation

Figure 4.1: Graphical and NPSEM representations of the same causal structure.

directed acyclic graphs (DAGs). It represents a set of variables and their joint probability distribution (Korb and Nicholson 2008; Koller and Friedman 2009). Familial terms are used to describe connections between variables or nodes in Bayesian networks. For example, in the BN in Figure 4.1, A and B are parents of C, and ancestors of Y. In causal BNs, a directed edge in the graph is assumed to represent the direction of causation between two adjacent nodes.

A Bayesian Network encodes the set of conditional independencies in the joint probability distribution of the variables in some observed data. Given a joint probability distribution consisting of a set of variables $\{A, B, C\}$, B is said to be conditionally independent of A given C if the conditional distribution of B given A and C does not depend on A. This is expressed mathematically as $B \perp\!\!\!\perp A \mid C$, if $P(B \mid A, C) = P(B \mid C)$. Consider the variables A, B, C with the relationships depicted in Figure 4.2. These three types of graph junctions are useful for illustrating how conditional independencies are encoded in causal graphs.

The graph in Figure 4.2a is called a *chain*. In this graph the effect of A on B is transmitted through C, hence the variable C is known as a mediator because it explains the causal effect of A on B. There is marginal dependence between every pair of nodes in this graph. Notably, B is not independent of A (mathematically, $B \not\perp\!\!\!\perp A$). However,



(a) A chain, C acts as a mediator here.

(b) A fork, C acts as a confounder here.

(c) An inverted fork, C acts as a collider here.

Figure 4.2: Three basic graph structures with node C playing different roles in each case.

B is conditionally independent of A given C ($B \perp\!\!\!\perp A \mid C$). Figure 4.2b is known as a *fork*, with C exerting influence on both A and B. In this case, A and B are associated and hence marginally dependent ($B \not\perp\!\!\!\perp A$) because of their mutual dependence on C. However, they are not causally dependent. C is referred to as a *confounder* as it is said to confound the relationship between A and B. B becomes conditionally independent of A given C ($B \perp\!\!\!\perp A \mid C$). Notice that the chain and fork structures encode the same marginal and conditional independencies and so cannot be uniquely distinguished by conditional independence in observational data. They are said to belong to the same *equivalence class*.

Figure 4.2c depicts an *inverted fork*, and C in this case is known as a *collider*. A collider node in a graph has two or more edge arrows pointing into it. This particular type of collider in Figure 4.2c is known as an *unshielded collider* because its parents are not adjacent in the causal graph. In this setting, A and B are marginally independent ($B \perp\!\!\!\perp A$), but given C they become dependent, ($B \not\perp\!\!\!\perp A \mid C$). The inverted fork encodes a different set of independence constraints compared to the chain and fork structures and can be uniquely identified from a joint probability distribution.

4.2.2 Causal Discovery

Causal discovery methods are techniques for recovering a data distribution's underlying structural causal model. Constraint-based algorithms for causal discovery identify

causal structures in a data distribution by testing for marginal and conditional independencies in the data distribution. They typically assume certain conditions about the data distribution and its Bayesian network including the causal Markov condition and causal faithfulness condition (Spirtes and Zhang 2016; Glymour et al. 2019). Another common assumption is *causal sufficiency*. Causal sufficiency means that the set of measured variables includes all common causes of all pairs of variables in the set. That is, there are no latent confounders relative to the set of observed variables.

In many cases, the direction of some edges in the causal Bayesian network can not be identified uniquely from data though conditional independence testing. This is because more than one Bayesian network can encode the same set of conditional independence constraints. However, some edges like the edges incident to unshielded colliders, can be identified uniquely using constraint-based methods (Eberhardt 2017; Maathuis and Nandy 2016). Hence, using constraint-based methods, a generic type of causal graph consisting of a mixture of edge types (including directed and undirected edges) can be recovered. This graph is known as a *Markov equivalence class* (MEC) graph and it represents all statistically indistinguishable DAGs that can be elicited from a particular data distribution.

The most prominent constraint-based causal discovery algorithm is the PC algorithm (Spirtes et al. 2000). It returns an MEC graph known as a Completed Partially Directed Acyclic Graph (CPDAG) which may include undirected edges or bi-directed edges in addition to directed edges. An ambiguous undirected or bi-directed edge between a pair of variables signifies when the algorithm is not able to uniquely determine which variable is the cause and which is the effect. Incorporating background knowledge into algorithmic causal discovery can help with correctly orienting ambiguous edges and improve the correctness of recovered graphs. This is a significant opportunity for causal discovery applications in industrial processes since aspects of causal

dependencies in the process are likely to be well known by process subject matter experts.

4.3 Methodology

4.3.1 Root Cause Graphs (RCGs)

We define the *root cause graph* (RCG) as a graphical model that describes the structural mechanism between features in data and their relationship to an outcome of interest. The true RCG for any process is a causal Bayesian network. Using a constraint-based method, it is possible to learn a Markov equivalence class of the RCG. In this work, the ground truth RCG is referred to as the *true RCG* or *true causal BN*, while the RCG learned from data is referred to as the *recovered RCG* or *estimated RCG*. As a representation of the data generating process that leads to the values of the outcome event observed, the root cause graph is a data-driven alternative to traditional representations of cause and effect mechanisms like the Ishikawa diagram which are derived from subjective approaches to root cause analysis.

4.3.2 Procedural Framework

To precisely evaluate methods for recovering the root cause graph from observational TTE data, the true causal structure and data generating process must be known. To achieve this, a simulation framework is developed to simulate TTE data for scenarios where root cause analysis techniques may be employed. A novel root cause graph recovery method is then applied to the simulated datasets. Figure 4.3 is a high-level depiction of this procedure.

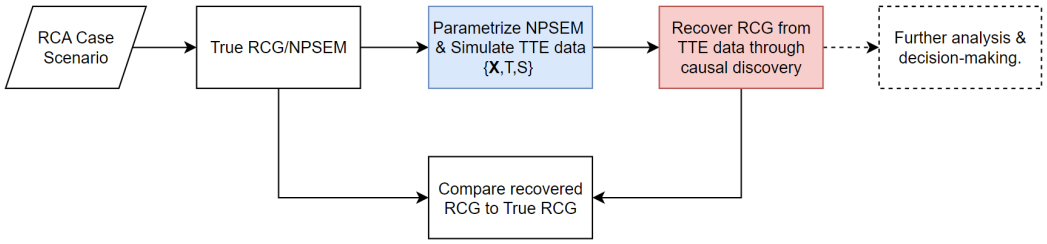


Figure 4.3: Overall procedure for simulating TTE data and recovering its root cause graph.

Consider a hypothetical system where it is desirable to learn the root cause graph of a failure trend using time-to-event data. For evaluation purposes, assume the SCM of the system is known, all data generating parameters are known, and all relevant covariates are represented in a causal BN. This BN may be transformed to an NPSEM, which can then be parameterized for data simulation. Censored time-to-event data along with covariate data is generated according to the specified SCM using the simulation framework described in Section 4.3.3. The method described in Section 4.3.4 is used to recover the root cause graph from the simulated data. The recovered RCG is then evaluated against the ground truth.

4.3.3 Data Simulation

Let $\{\mathbf{X}, T, S\}$ be the data distribution of a right censored time-to-event data to be simulated, where $\mathbf{X}$ is the set of covariates $\mathbf{X} = X_1, X_2, ..., X_m$ measured alongside event times T for every observed unit in the study sample. Let S denote the status variable indicating whether a particular observation is censored ($S = 0$) or not ($S = 1$). Our goal is to generate realistic datasets of the form $\{\mathbf{X}, T, S\}$, sampled from standard distributions. Existing simulation methods are designed to simulate either causal covariate data $\mathbf{X}$ (as in Sofrygin et al. 2017; Al Hajj et al. 2023) or the event times $\{T, S\}$ (as in Crowther and Lambert 2012; Harden and Kropko 2019) alone. Our simulation

approach builds on existing methods and provides an integrated framework for simulating both covariate data and event times in censored TTE data. This framework enables the flexible simulation of covariate data from complex Bayesian networks whose nodes may exert direct or indirect influence on the generation of censored event times through variously specified direct and mediated covariate effects. The ability to easily specify mediated covariate effects on event times is a key feature of this framework.

We implement a two-part simulation procedure. The first generates the covariate data $\mathbf{X}$ based on the specified SCM while the second part generates event times and status, $\{T, S\}$ from a combination of the covariates $\mathbf{X}$ and a parametric function that describes the baseline hazard of the event time distribution $h_0(t)$. This two-part simulation procedure is depicted in Figure 4.4.

In the first part of the simulation process, a pre-specified NPSEM is used to generate covariate data in a stochastic fashion. The variables in the NPSEM are assumed to follow a particular parametric distribution which needs to be specified prior to data generation. Hence, to generate data for a specific variable in the NPSEM, we sample from its standard parametric distribution. The parameters of exogenous variables are precisely specified while the parameters of endogenous variables are defined as a function of their parents in the SCM.

In the second part of the simulation procedure, an event time for each sample unit is generated under a proportional hazards assumption based on the covariates $\mathbf{X}$ and a baseline hazard function. One method that can be used to simulate realistic event times is the *cummulative hazard inversion* method (Bender et al. 2005; Brilleman et al. 2021). Through this method, a function is derived for computing survival times by inversion of the cummulative hazard function.

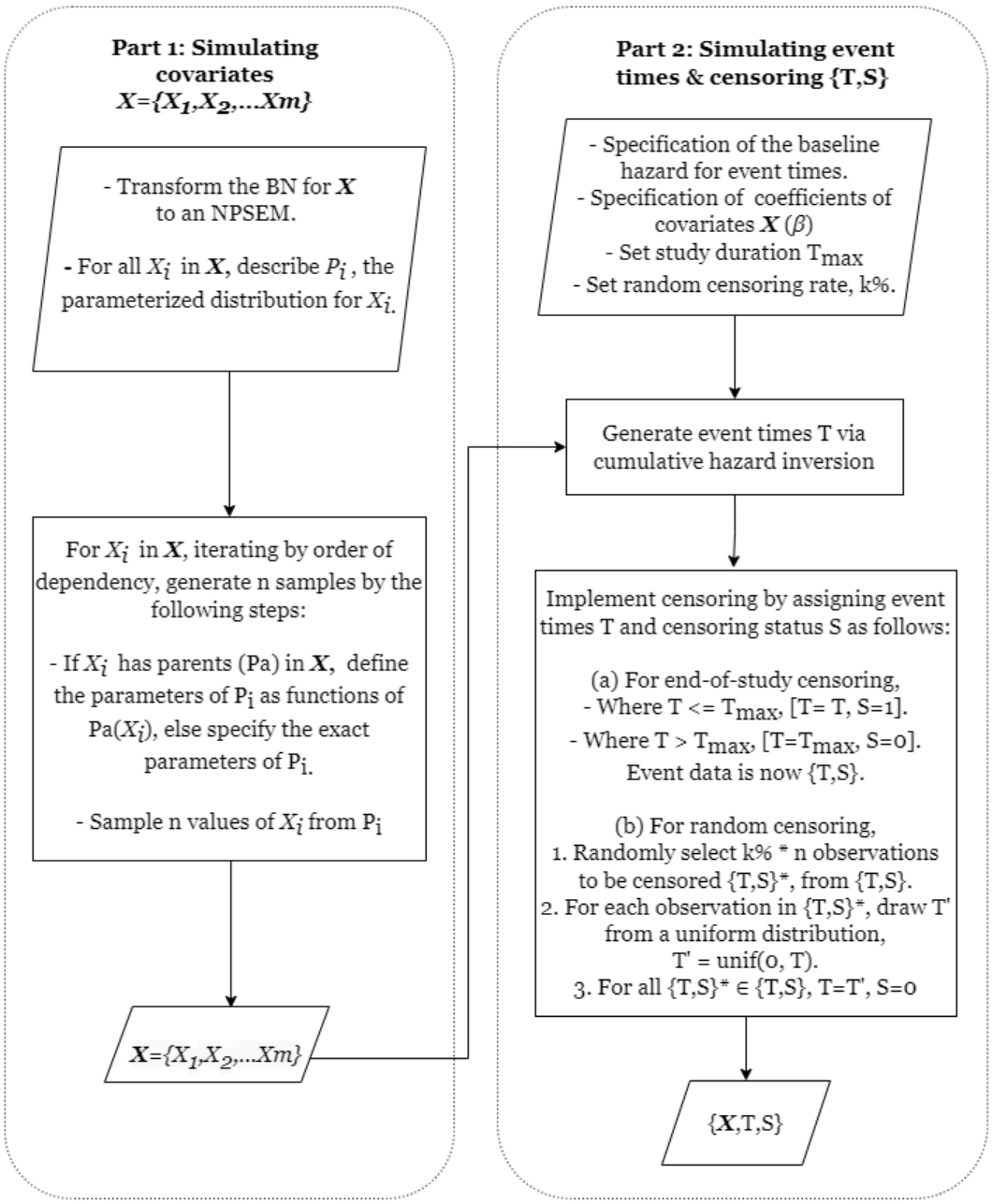


Figure 4.4: Two part simulation procedure for generating time-to-event data distribution $\mathbf{X}, T, S$ according to specified SCM. Part 1 generates $\mathbf{X}$ while part 2 generates $\{T, S\}$

.

The survival function $S(t) = P(T > t)$ of a proportional hazard model can be expressed as

$$S(t \mid \mathbf{x}) = exp[-H_0(t)exp(\beta \mathbf{x})] \tag{4.1}$$

Where $\mathbf{x}$ is the realized values of covariates $\mathbf{X}$, β is the vector of the coefficients of $\mathbf{x}$, and $H_0(t)$ is the cumulative baseline hazard function.

Now, the probability of survival at a particular event time T_i for a sample unit i in the study, is given by the cumulative distribution function (CDF) of its survival function. The CDF of a continuous random variable is expected to follow the uniform distribution $U(0,1)$, therefore we can write,

$$S_i(T_i) = U_i \sim U(0,1) \tag{4.2}$$

From Equation 4.1,

$$S_i(T_i) = exp[-H_0(T_i)exp(\beta_i \mathbf{x}_i)] = U_i \sim U(0,1) \tag{4.3}$$

The event times can then be obtained by inverting the baseline cumulative hazard function and rearranging Equation 4.3 to give

$$T_i = H_0^{-1}[-log(U_i)exp(-\beta_i \mathbf{x}_i)] \tag{4.4}$$

where H_0^{-1} is the inverted baseline hazard.

Assuming a parametric distribution for the event times, this expression allows event times to be computed after sampling individual survival probabilities from a uniform distribution (Brilleman et al. 2021). For a discussion including methods for inverting the hazard function, see Crowther and Lambert (2013).

Two right-censoring mechanisms are implemented on the generated event times. The first is end-of-study censoring which occurs when for a sample unit, the event is not observed before the study comes to an end. By this mechanism, simulated event times that are larger than a pre-specified study duration T_{max} are considered censored. The

event time T is accordingly assigned the value T_{max} and censoring status $S = 0$. The second mechanism of censoring implemented is random censoring which is a common censoring mechanism in real survival data that occurs for a variety of reasons but in a non-systematic manner. To implement random censoring, a proportion of the study sample (including samples which could have been subject to end-of-study censoring) are randomly selected, and for each observation with simulated event time T, we draw its random censoring time T' from a uniform distribution $U(0, T)$ and reassign this value to T. For all censored observations, the status variable S is set to 0, while S remains 1 when event has been observed at recorded time T.

4.3.4 Root Cause Graph Recovery

To recover the root cause graph for an event of interest from observational TTE data, we propose using causal discovery techniques to reverse-engineer the structure of the causal relations between variables associated with the event. We employ in this work the constraint based causal discovery algorithm, the PC algorithm (Spirtes et al. 2000). The PC algorithm starts with a fully-connected undirected graph of the variables, from which it estimates the graph's skeleton structure. The skeleton is the undirected graph obtained after edges between variables in the fully connected graph that are not directly related have been eliminated using conditional independence tests. The algorithm then proceeds to determine edge orientations by first finding unshielded colliders in *v-structures*. V-structures are variable triples, e.g. $\{A, B, C\}$, such that they are connected as an undirected chain-like structure as in A — C — B. In this structure, the algorithm determines that if C was not part of the conditioning set that made A and B independent, then we have an unshielded collider in C. Other edge orientation rules allow the algorithm to asymptotically converge to a Markov equivalence class DAG (Glymour et al. 2019). For example, the edge propagation rule

allows the algorithm to find directed chain structures by converting $A \rightarrow C - B$ to $A \rightarrow C \rightarrow B$.

One challenge with causal discovery using censored time-to-event data is that the outcome of interest is described by two variables, the event time T and the event status S. Choosing to use either of these variables alone as the outcome variable would result in loss of critical information. Our solution is to estimate an alternative outcome variable Y that captures relevant information from T and S in a single variable. This single outcome variable becomes the focus of a causal discovery procedure that reveals how the outcome is related to other variables in the data. Ideally, it will be a variable that can be used as a causal estimand. One such measure that can be estimated from event data is the survival probability for individuals in the study sample at a specific time of interest. As a causal estimand, it can be used for quantifying the effect of a change in any of the covariates on the outcome in terms of a risk scale that measures the difference in the marginal survival functions given different treatments. For example, the average causal effect (ACE) at time t of a change in a variable x from a value of 0 to 1 can be evaluated as

$$ACE(t) = P(T > t \mid x = 1) - P(T > t \mid x = 0)$$

where T is the survival time, and all other relevant covariates are kept constant.

Considering a specific time-point of interest t_{int} for which an analyst seeks to understand the process during $0 \leq t \leq t_{int}$. We define the alternative outcome variable Y as the survival probability at time t_{int} given $\mathbf{x}$ $S(t = t_{int} \mid \mathbf{x})$. Given an observed TTE data distribution $\{\mathbf{X}, T, S\}$, and a stipulated time of interest t_{int}, the estimator

for the new outcome variable for each sample unit $\hat{Y}_i$ is the probability that the sample unit survives beyond t_{int} given its covariate values $\mathbf{x}_i$.

$$\hat{Y}_i = P(T_i > t_{int} \mid \mathbf{x}_i) \tag{4.5}$$

where T_i is the actual survival time for the sample unit. This definition leaves the analyst with the flexibility of choosing what time-point is most relevant to the research question. This single outcome replaces the event time and event status in the original data distribution.

$$\{\mathbf{X}, T, S\} \Rightarrow \{\mathbf{X}, Y\}$$

The root cause graph recovery procedure is described in Figure 4.5. Survival regression models like proportional hazards and accelerated failure time models can be used to estimate the probability of survival for individual units at a specific time. Other specialized machine learning survival models such as the multi-task logistic regression (Yu et al. 2011) and random survival forests (Ishwaran et al. 2008) models can also be used for predicting individual survival probabilities at a specific time.

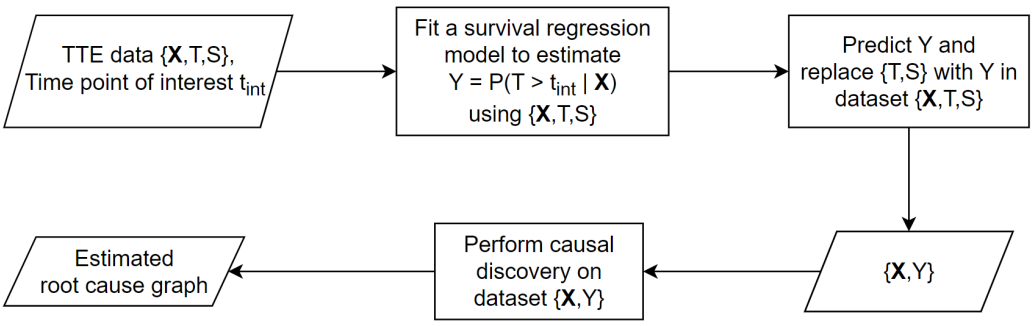


Figure 4.5: Method for recovering the root cause graph via causal discovery.

4.3.5 Implementation

Three datasets are generated from different case scenarios using the simulation framework described in Section 4.3.3 for the purpose of evaluating the proposed method for root cause analysis using TTE data. The case scenarios have different data generation processes representing different applications where the methods proposed in this paper can be utilized. Not only have different Bayesian network structures been explored, different sample sizes, data types, and parameters of the hazard function have been experimented with during data simulation. Baseline hazards are specified using the Weibull family of distributions along with reasonable fixed effects for the covariates in each case.

The proposed root cause recovery method is applied to each dataset before evaluating the degree to which the true casual DAG was recovered. The semi-parametric Cox model is used to fit the hazard function of the event as

$$h(t \mid \mathbf{x}) = h_0(t) \times exp(\beta \mathbf{x}) \tag{4.6}$$

Using Equation 4.6, the survival function can be obtained from the expression the $S(t) = e^{-H(t)}$ where $H(t)$ is the cummulative hazard. Sample unit covariate values can then be applied to the survival function to obtain the desired single outcome estimate of survival probabilities (Equation 4.5).

Finally, the stable PC algorithm is used to recover the structural causal model of the dataset which gives the estimated root cause graph. Stable PC is an order-independent version of the PC algorithm (Colombo and Maathuis 2014). The likelihood-ratio test for mixed data types proposed by Andrews et al. (2018) is used as the conditional independence test in this implementation.

4.4 Problem Description

The proposed methodology for root cause analysis can be applied to a variety of situations where principled root cause analysis can lead to improvements on an outcome of interest. This includes problems under the domains of operations management, process improvement, device management, product life cycle management and failure diagnostics.

For our current implementation of the proposed RCG recovery method, the following conditions about the problem and its associated data are assumed:

1. Causal sufficiency: The set of measured variables includes all common causes of all pairs of variables in the data. There are no latent confounders in the data.

2. The set of relevant covariates $\mathbf{X}$ are random variables which form a multivariate Gaussian distribution. The relationships between the covariates are approximately linear and the hazard function of the event can be modeled using a Weibull distribution.

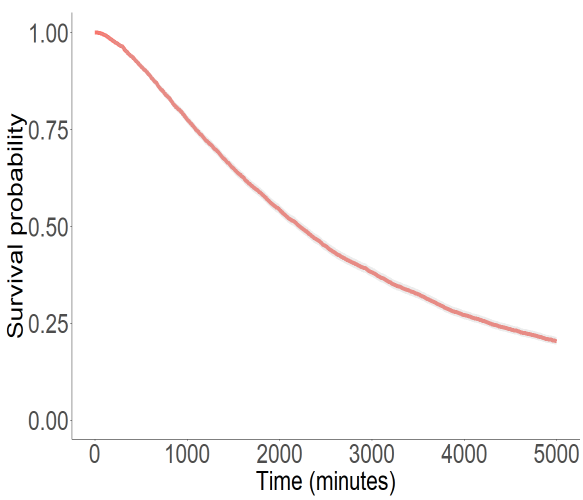
3. Proportional hazards: The covariates have a constant multiplicative effect on the hazard function of the survival outcome, and the hazard functions for any two subjects at any point in time are proportional.

4. The TTE data is right-censored, and censoring is uninformative.

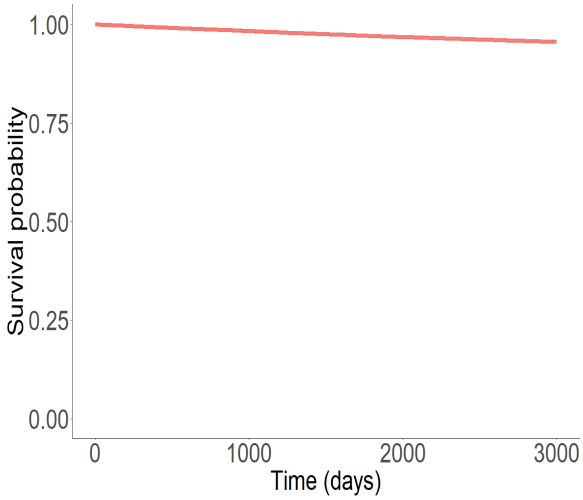
4.4.1 Case Scenarios

The case scenarios in this work are hypothetical problems which mimic realistic scenarios in industry where the proposed method for root cause analysis could be applied. These scenarios are described in subsections 4.4.1.1 to 4.4.1.3. In each case, the

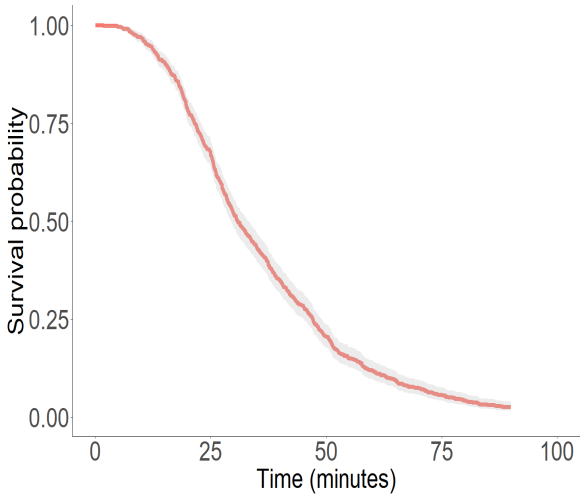
outcome variable Y is the censored event time. The case datasets are simulated according to the structural causal models assumed to govern the data generating process of these scenarios. Several problem characteristics and parameters for data generation are varied in the different scenarios. The properties of the three different cases are summarized in Table 4.1 and the survival plots of the datasets are shown in Figure 4.6.



(a) Case 1



(b) Case 2



(c) Case 3

Figure 4.6: Survival plots for the case scenarios.

Table 4.1: Characteristics of the case scenarios.

	Case 1	Case 2	Case 3
Objective for outcome (Y)	maximize	maximize	minimize
No. of observations	6000	50000	730
No. of nodes	6	8	11
No. of edges	5	10	9
No. of {discrete, continuous} covariates in $\mathbf{X}$	{1, 4}	{3, 4}	{4, 6}
Baseline hazard trend	Increasing	Decreasing	Increasing
Weibull distribution parameters {Shape(γ), Scale(λ)}	{2, 0.5}	{0.99, 50}	{3.5, 0.0001}
End-of-study censoring rate	20%	95.5%	2.5%
Random censoring rate[a]	10%	0.1%	0%
Overall censoring rate	28%	95.5%	2.5%
Duration of study	5000 minutes	3000 days	90 minutes
Time-point of interest	100th minute	365th day	30th minute

[a] End-of study censoring candidates could also be random censored.

4.4.1.1 Case 1

The SCM depicted in Figure 4.1 is used for case 1. This case relates to the observed failures of different models of some rotational equipment (such as a drill) when operated continuously for a long duration. 6000 manufactured units of the equipment are studied for a maximum of 5000 minutes of continuous operation and their failure (event) times and status recorded as part of a large-scale testing program. Equipment which do not fail within 5000 minutes are considered censored via end of study. About 10% of the study samples are also randomly censored due to measurement errors. This specific study uses the data to investigate the root causes of systematic early failures at or before the 100th minute of continuous operation. The set of covariates $\{A, B, C, D, E\}$ are summary properties of the equipment and their operating conditions and can be described as follows: A - surface hardness; B - average operating torque; C - relative aggregated stress; D - equipment model; and E - average wear rate.

4.4.1.2 Case 2

The true causal BN for case 2 is depicted in Figure 4.7. This case portrays the challenge of understanding the factors driving observed failure trends of computer hardware components in data center infrastructure. In this case, 50000 units of a specific hardware component deployed in data center servers across geographic locations are monitored for 3000 days. The goal is to use the TTE data from this study to unravel the root causes of early failures of the components under study. The set of covariates $\{A, B, C, D, E, F, G\}$ are summary characteristics of the hardware units and the conditions under which they operate. They can be described as follows: A - overall

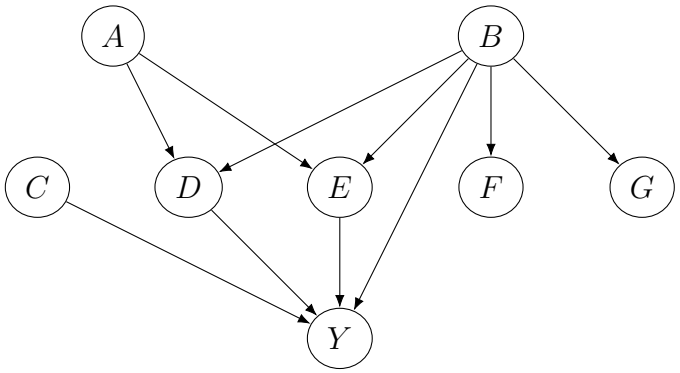


Figure 4.7: Case 2 Bayesian network

monitoring at rack level; B - location of data center; C - machine functional designation; D - rack temperature regulation score; E - voltage regulation score; F - outside temp; and G - data center energy consumption.

4.4.1.3 Case 3

The causal BN for case 3 is depicted in Figure 4.8. This scenario mimics a use case for data-driven root cause analysis in logistics operations improvement. In this scenario, the goal is the improvement of a warehouse operations key performance index; the truck-turn-around (TTA) time. The task is to investigate the contributing factors to patterns of delays being observed in the operational work flow for unloading and reloading delivery trucks. The duration of interest in this case is the average TTA time per shift, and the TTE data is at shift level; with different work crews rotating shift duties. Unlike the previous case scenarios, in this case high event times are undesirable, and the goal of the logistics department is to minimize average TTA times. The covariates in this study can be described as follows: A - forklift crew on duty; B - storekeeper on duty; C - operational losses; D - check-out staff experience level; E - check-in staff experience level; F - average forklift availability; G - warehouse

layout adherence score; H - housekeeping score; Q - luminance; and R - fresh order level.

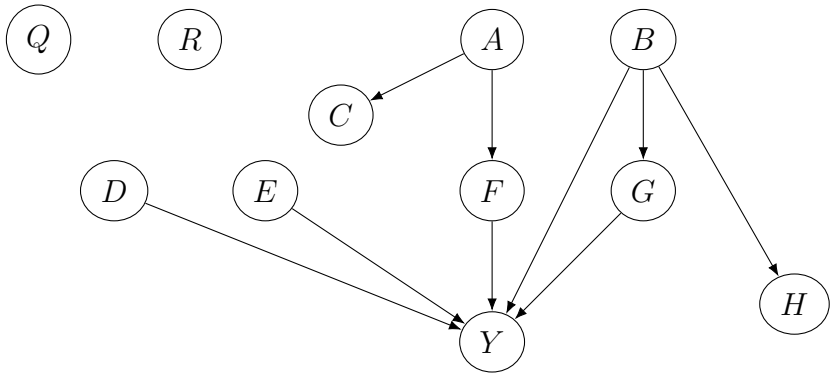
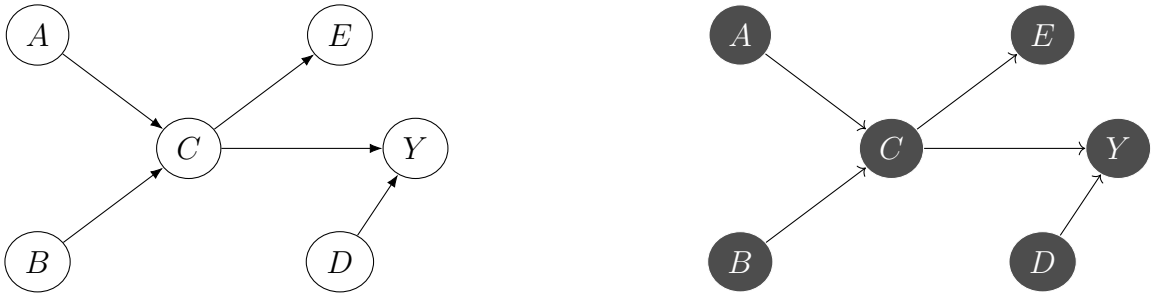


Figure 4.8: Case 3 Bayesian network

4.5 Results and Analysis

The recovered root cause graph is a Markov equivalence class graph known as a Completed Partially Directed Acyclic Graph (CPDAG). The directionality of some recovered edges may remain unresolved and ambiguous. We compare the true causal BNs of the case scenarios to the RCGs recovered using the proposed method. Figure 4.9 shows the true causal BN and the recovered RCG for case 1. In this ideal scenario, the proposed method is able to recover the true causal BN exactly.

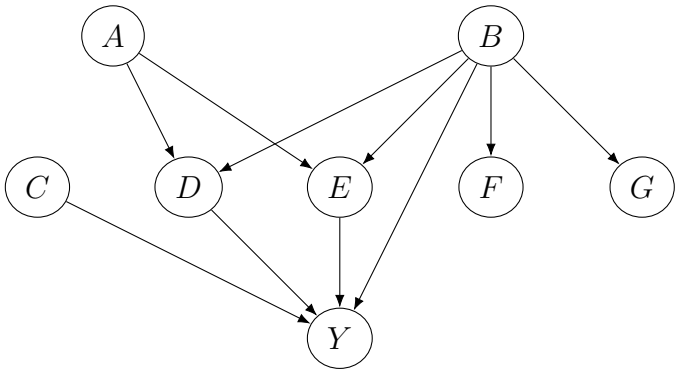


(a) True causal BN for case 1.

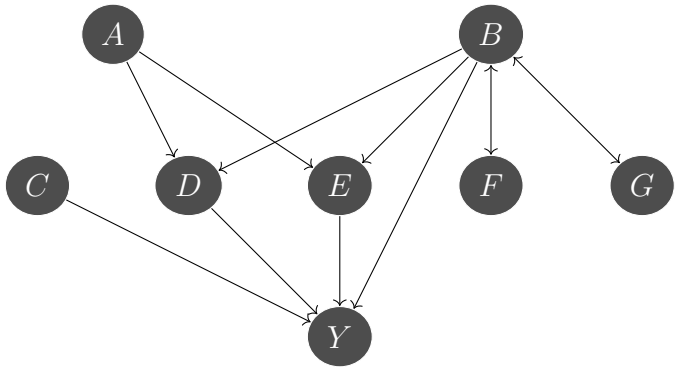
(b) Recovered RCG for case 1.

Figure 4.9: Comparing the true RCG and recovered RCG in case 1

The recovered RCG in case 2 includes two ambiguous edges as shown in Figure 4.10. The remaining eight edge directions are correctly oriented. All node and edge adjacencies are also correct. The ambiguous bi-directional edges, $B \longleftrightarrow F$ and $B \longleftrightarrow G$, present an example where background knowledge if available, could be used to reorient edges to improve the result of an algorithmic causal discovery process.



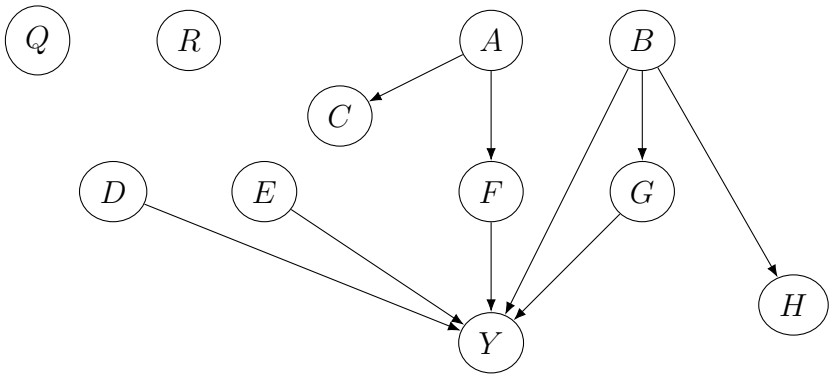
(a) True causal DAG for case 2.



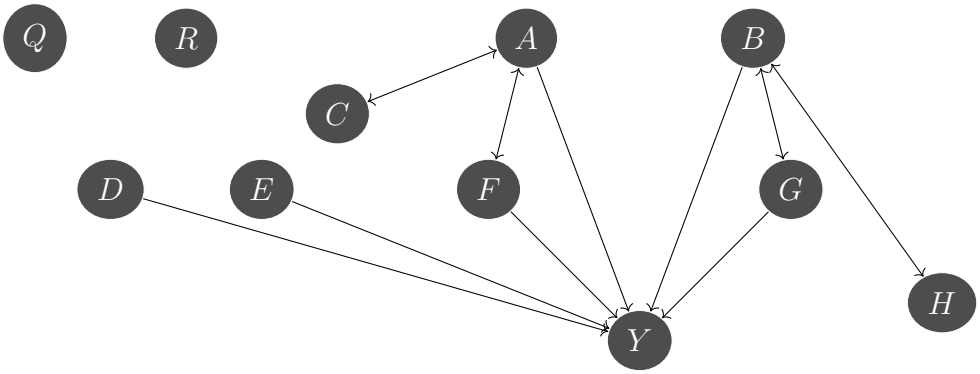
(b) Recovered root cause graph for case 2.

Figure 4.10: Comparing the true RCG and recovered RCG in case 2.

The recovered RCG in case 3 (Figure 4.11) includes four ambiguous edges; $A \longleftrightarrow C$, $A \longleftrightarrow F$, $B \longleftrightarrow G$ and $B \longleftrightarrow H$, but also one extra edge $A \longrightarrow Y$. Similar to case 2, background knowledge could potentially be useful for correctly reorienting the ambiguous edges.



(a) True causal DAG for case 3.



(b) Recovered root cause graph for case 3.

Figure 4.11: Comparing the true RCG and recovered RCG in case 3.

We use the pattern metrics in Table 4.2 to evaluate the recovered root cause graphs. Other metrics for comparing the performance of different causal discovery methods including adjacency/arrowhead precision and recall, and structural hamming distance, can be computed from these pattern metrics (Raghu et al. 2018; Nogueira et al. 2022).

Table 4.2: Pattern metrics for recovered RCGs relative to their true RCGs in the three case scenarios.

Metric	Case 1	Case 2	Case 3
Correct directed edges[a]	5	8	6
Incorrect directed edges[b]	0	2	4
Correct adjacencies[c]	5	10	9
Incorrect adjacencies[d]	0	0	1
Missing edges[e]	0	0	0
Extra edges[f]	0	0	1

[a] Number of correctly directed edges in the recovered RCG.
[b] Number of incorrectly directed edges in the recovered RCG.
[c] Number of undirected edges (disregarding arrowheads) that are present in both true and recovered RCGs.
[d] Number of undirected edges (disregarding arrowheads) present in one RCG but absent in the other.
[e] Number of edges that are present in the true RCG but absent in the recovered RCG.
[f] Number of edges that are present in the recovered RCG but absent in the true RCG.

4.5.1 Root Cause Effect Estimation

Root cause graphs are useful towards effecting desirable changes on an outcome of interest such as increasing survival times of a device fleet. Root cause graph recovery from data allows further analyses, decisions and corrective actions to be approached in a principled, informed manner, taking advantage of the ability of a causal Bayesian network to encode knowledge about the process under study. In this section, we consider how the effects of root causes (treatment effects) can be estimated in a principled manner using a recovered RCG. This is useful for predicting the effect that changes to a root cause would have on the outcome.

Regression models can be used for estimating the causal effect of a treatment variable on an outcome. This requires the absence of selection bias in the data, and adjusting for any confounders in the relationship between the treatment and the effect. Assuming that the regression model is correctly specified, and confounders have been

properly adjusted, the coefficient of the treatment variable in such a regression model can be interpreted as an unbiased estimate of the causal effect of the treatment variable on the outcome (Schochet 2010; Hernán and Robins 2018; Funk et al. 2011).

Covariate selection for confounder adjustment is critical for causal inference using observational data but remains a major challenge (VanderWeele 2019). Traditional approaches to confounder selection can be prone to over-adjustment which can amplify biases in the estimation of causal effects (Greenland and Pearl 2011; Shrier and Platt 2008). It is not uncommon for researchers to try to adjust for most/all measured covariates or all covariates that are correlated with the outcome variable (VanderWeele and Shpitser 2011; Pearl and Mackenzie 2018). Causal DAGs provide a basis for principled confounder selection by explicitly revealing which variables confound the relationships being studied. Several recent criteria that have emerged for reliable covariate selection depend on knowledge of the causal DAG (Greenland and Pearl 2011; VanderWeele 2019; Tafti and Shmueli 2020).

Consider a root cause variable for which some treatment effect measure on the outcome Y is to be estimated. A naive approach to covariate adjustment in a regression model for estimating such treatment effect, is adjusting for the full set of measured covariates. The independent variables in such model is the full set of $\mathbf{X}$ variables including the treatment variable whose coefficient will be the estimated treatment effect. However, this approach often leads to over-adjustment which can lead to biased causal estimates. Specifically, adjusting for variables which act as mediators or colliders would likely diminish or exaggerate the estimated total effect of the treatment on the outcome (Greenland and Pearl 2011; Pearl and Mackenzie 2018). Mediators are of particular interest in root cause analysis applications because it is often the case that some or all of the effects of certain root causes of an outcome are mediated by another intermediary variable.

For root cause treatment effect estimation in the three case scenarios, we apply principled covariate selection to Cox regression models using their recovered RCGs. We call such models *RCG-informed models* or *RCG-guided models*. We also fit another model that adjusts for all measured covariates which we call the *full model*, and obtain the coefficients of the root cause treatment variables from this model. The estimated total effects of root causes from the RCG-informed model and the full model are compared to the ground truth effect sizes which are obtained from the data simulation functions.

Table 4.3 shows the the coefficients of mediated root cause variables from the two estimating models: the full model and the RCG-informed model. These coefficients of the treatment are considered to be estimates of the total effects of the root causes on the outcome which in this case is the hazard of survival times. The true effects on the hazard are shown in the last column of the table. Also included in the table are measures of prediction performance and model fit for the two estimating models; the concordance index (CI) and Akaike Information Criteria (AIC).

Notice that the estimates from the RCG-informed model are much closer to the true values of total root cause effects while the estimates from the full model are either diminished or exaggerated. In several cases, the +/- signs of the full model coefficients are reversed. The poor estimation of mediated root cause treatment effects in the full model occurs notwithstanding the fact that the full model achieves superior predictive performance than the RCG-informed model in all cases, as indicated by the concordance-index and AIC. Adjusting for mediators when estimating treatment effects leads to poor estimates because the mediators tend to 'explain away' the true effects being investigated. This clearly demonstrates the value of learning the root cause graph before attempting to estimate treatment effects of potential root causes.

Table 4.3: Comparison of effect estimates for root causes with mediated effects on the outcome. {CI,AIC} stand for the concordance index and Akaike Information Criteria.

Case Study	Variable	Full model	RCG-informed model	True effect
Case 1	A	0.0016	-0.289	-0.3
	B	0.0118	-0.164	-0.18
	$\{CI,AIC\}$	$\{0.802,64094\}$	$\{0.794,64474\}$	
Case 2	A[a]	-0.0644	-1.908	-1.95
	$B2$[b]	-0.031	0.2	0.2
	$B3$[b]	0.0194	-1.441	-1.485
	$\{CI,AIC\}$	$\{0.891,42079\}$	$\{0.866,43056\}$	
Case 3	$B2$[c]	-0.507	0.07	0
	$B3$[c]	0.0066	0.812	1
	$A2$[c]	0.0275	1.89	2.4
	$\{CI,AIC\}$	$\{0.798,7336\}$	$\{0.757,7610\}$	

[a] A is an ordinal categorical variable that is treated like a continuous variable under the assumption that its effect is proportional to the ordinal progression of its factor levels.
[b] B2 and B3 are different levels of the categorical variable B in case 2. Their values are with respect to the reference level B1.
[c] Similarly, the values of B2 and B3 reference B1 in case 3, while A2 references A1.

Tables 4.4, 4.5, 4.6 show all model coefficients for the full model and RCG-informed model, as well as the true values of the effects for all three cases. Notice that the estimated covariate effects in the full model are close to the true values for variables where there is no mediation but considerably far off for variables where there is mediation.

Table 4.4: Comparison of all covariate coefficients in case 1 outcome models.

Variable	Full Cox model	RCG-informed model	True effect
A^*	0.0016	-0.2886	-0.3
B^*	0.0118	-0.1644	-0.18
C	0.6057		0.6
E	0.0055		0
$D2$	-1.1085	-1.08	-1.1
$D3$	-0.5339	-0.5175	-0.5

* Mediated variables.

Table 4.5: Comparison of all covariate coefficients in case 2 outcome models.

Variable	Full Cox model	RCG-informed model	True effect
A^*	-0.0644	-1.9084	-1.95
$C2$	-0.0318	-0.021	-0.07
$C3$	-0.2552	-0.265	-0.3
$B2^*$	0.031	0.2	0.2
$B3^*$	-0.0194	-1.4411	-1.485
D	-0.5615		-0.6
E	-0.7728		-0.75
F	0.0067		0
G	-0.0064		0

* Mediated variables.

For non-mediated root causes, including mediators from other root causes can lead to better estimates of their effects.

Table 4.6: Comparison of all covariate coefficients in case 3 outcome models.

Variable	Full Cox model	RCG-informed model	True effect
E	-0.6097	-0.4764	-0.6
D	-0.3147	-0.2448	-0.4
$B2^*$	-0.5065	0.0738	0
$B3^*$	0.0066	0.8116	1
G	0.4371		0.4
$A2^*$	0.0276	1.8934	2.4
F	-0.6413		-0.6
C	0.0081		0
H	-0.0108		0
Q	-0.0056		0
R	-0.0138		0

* Mediated variables.

4.6 Discussion

The general framework for RCG recovery proposed in this work as depicted in Figure 4.5 is algorithm/model-agnostic. This allows for flexibility in the choice of causal discovery algorithms and models for estimating the survival probability outcomes during implementation. These choices should depend on what is known about the data and what can be assumed about its generating process. For example, in a process where the observed set of variables in the data cannot reasonably be assumed to be causally sufficient, a causal discovery algorithm which does not assume causal sufficiency such as the RFCI (Colombo et al. 2012) or GFCI (Ogarrio et al. 2016) algorithms should be

used instead of the PC algorithm used in this implementation. Likewise, for datasets where the proportional hazards assumption is not tenable, or where more complex estimators might be required for predicting survival probabilities, the Cox model used in this work would need to be replaced by a suitable alternative model such as the accelarated failure time model (Saikia and Barman 2017) or random survival forests model (Ishwaran et al. 2008). Other tradeoffs such as the loss of explainability when using more complex models like random survival forests would need to be considered. Therefore, for any application of the proposed RCG recovery method, it is important to check the assumptions of the candidate estimating models and causal discovery algorithms against the particular dataset RCG recovery is to be performed on.

In research involving causal structure learning, simulations are commonly used and are often necessary for the validation of new methods for learning causal models like RCGs. This is because with simulated data the ground truth causal structure is known. With real world data there is no way to guarantee accurate knowledge of the precise structure and parameters of data generation. The RCG recovery framework is applied to three datasets generated from known Bayesian networks from real world inspired industrial case scenarios, as described in Section 4.4.1.

The datasets simulated from these case scenarios are used to evaluate the effectiveness of the proposed root cause graph learning method. The method is able to recover useful approximations of the ground truth causal BNs with high rates of correct edge adjacencies. The recovered RCGs depict causal dependencies between measured covariates and how their effects are propagated through the system to affect the outcome of interest. The proposed method can be applied to various operations and processes under the domains of manufacturing, process improvement, reliability and maintenance engineering, and device/product life cycle management. It can also be extended to other domains like healthcare where time-to-event data are commonly used.

In real world applications, the recovered RCGs can be used by process experts as a tool for checking their understanding and assumptions about the processes they manage while also evaluating the plausibility of the recovered RCGs. Learning about root causes often involves expensive experimentation in industry. The proposed RCG recovery method can help to eliminate the need for costly experiments or simplify these experiments by revealing potential confounders that need to be controlled for. Also, RCG recovery unveils aspects of the process where causal direction is identifiable from data so that experimentation and other knowledge discovery efforts can focus on aspects where the direction of causation is difficult to determine.

In all three cases studied in this paper, the recovered RCGs are shown to be useful for improving downstream analyses and decision-making tasks. This is demonstrated through principled estimation of the total effects of the identified root causes using the RCGs as a guide for covariate adjustment. In particular, using RCG-guided models for estimating the treatment effects of mediated root cause variables produces effect estimates that are much closer to their true values compared to models that do not consider mediation mechanisms. Adjusting for a mediator variable while estimating treatment effects of a mediated root cause variable diminishes or exaggerates the estimated effect of the root cause relative to the true effects. RCGs are useful for clearly identifying mediator variables.

Future research should address assumptions used in the current implementation of the RCG recovery method which may not be reasonable in different datasets. For example, the assumption of causal sufficiency is difficult to guarantee in many datasets collected from real world processes. When that is the case, causal discovery algorithms which relax this assumption may be used for RCG recovery. Additionally, in collaboration with industry process owners, domain-specific research that incorporate RCG recovery can be used to evaluate how much the recovered root cause graphs validate or

challenge the contemporary understanding of stakeholders about the causes of systematic failures in their respective processes leading to improvements in component failure times.

Chapter 5

Causal Feature Selection for Machine Learning Interpretability and Domain Adaptation

5.1 Introduction

Causal machine learning is an emerging field that incorporates ideas and techniques from causality research into machine learning (ML). This approach to ML offers several benefits including enhancing ML generalization and domain adaptation, facilitating model transparency and interpretability, improving fairness in artificial intelligence (AI), and amplifying opportunities for knowledge discovery (Pearl 2019b; Schölkopf 2022b; Kaddour et al. 2022).

With the current emphasis on interpretable/explainable ML&AI, developing models with causal intuitions is highly desirable for many reasons (Miller 2019; Molnar et al. 2020; Cheng et al. 2021; Saeed and Omlin 2023). Such models are useful for decision-making, knowledge discovery, and building safety, ethics and trust into AI systems. Traditional methods for explaining ML models are based on superficial correlation and may lead to misleading interpretations of the real world mechanism generating the observed data (Xu et al. 2020; Feder et al. 2021). The field of causal inference offers tools that can be leveraged for deeper insights into how ML models arrive at their predictions and the real-world implications of the models.

The assumption in machine learning that source data and target data are similar, and are drawn from the same distribution is routinely violated in practice (Pan and Yang 2009; Weiss et al. 2016; Zhou et al. 2022). This same source/target distribution (SSTD) assumption forms an important limitation for most ML algorithms which often leads to significant performance decline in changing environments. As a result, research fields which tackle this problem such as domain adaptation and transfer learning are growing in importance. This factor drives the increasing interest among machine learning researchers in causal inference, given its recognized potential for bolstering the generalization capabilities of machine learning models (Zhang et al. 2015a; Subbaswamy et al. 2019; Zhang et al. 2020; Yang et al. 2021; Scholkopf et al. 2021).

This work considers the subject of feature selection for prediction modeling. Causal approaches to feature selection offer a direct and accessible way to integrate causal considerations in machine learning. We explore the suggestion that causal feature selection techniques can improve prediction performance in machine learning over comparable methods that are not based on causality. A feature selection algorithm that prioritizes relevant causal relationships is proposed and its prediction performance is compared to a related classical feature selection approach based on statistical dependence alone.

Furthermore, we focus on the problem of domain adaptation and the performance of prediction models when the SSTD assumption does not hold. Introducing a new causal feature selection algorithm, a domain adaptation approach that dynamically adjusts selected causal features to the target environment is proposed. This is achieved through the identification of univariate covariate shifts and the subsequent removal of predictors that are likely to degrade performance in the target environment.

The feature selection algorithms introduced in this paper are experimentally evaluated on regression tasks, with the results strongly corroborating the theory that the

set of causal features known as the Markov blanket delivers optimal prediction performance, while upholding the interpretability of prediction models. The experiments on simulated datasets with well-understood properties demonstrate the impact of dataset shift on prediction models and how causal strategies for feature selection can improve in-distribution and out-of-distribution prediction performance.

5.1.1 Related Work

Domain adaptation and transfer learning have garnered substantial attention in the realm of machine learning research (Zhou et al. 2022). Supervised methods for domain adaptation require labeled data from the target domain (Daumé 2007; Pan et al. 2011). A problem setting that is increasingly being studied is when some data samples from the target domain with no labels (or few labels) are available during model development rather than fully labeled target data which is often unavailable. This allows for an unsupervised/semi-supervised strategy for domain adaptation. Feature-based unsupervised methods focus on the data's features rather than individual sample instances.

Most feature-based domain adaptation methods involve some form of transformation, remapping, or acquisition of new representations of the original feature space before training a prediction model (Pan et al. 2011; Sun et al. 2016; Shen et al. 2018; Farahani et al. 2021; Dhaini et al. 2023). Such manipulations add a layer of complexity and abstraction which impacts the interpretability of prediction models. Also, it is possible that the new representations may be less informative to the prediction model thereby impacting the discriminative ability of the model (Sun et al. 2019). Furthermore, as highlighted in Dhaini et al. (2023), methods based on this approach may exhibit performance limitations when applied to regression problems.

One way to avoid such feature transformations and reduce prediction model performance degradation in the target domain is through a suitable feature selection strategy. The goal is to select a set of invariant features across source and target distributions in order to minimize the effects of distribution shifts when predicting in the target domain. Uguroglu and Carbonell (2011) propose a feature-based domain adaptation method that identifies invariant features through an unsupervised approach using the maximum mean discrepancy statistic. Their method requires that at least one feature is variant across the two domains and involves solving an optimization problem which can be computationally limiting for practical problems. Sun et al. (2019); Yan et al. (2022) follow a similar feature selection strategy but their methods lose the model-agnostic property of the feature selection step due to its integration with a specific prediction model class. Deng et al. (2019) highlight the limitations of the approach in Uguroglu and Carbonell (2011) including two issues addressed in this work: the lack of consideration of the conditional distribution and the local structure of the data, and the use of computationally expensive optimization procedures. However, their proposed method also involves feature transformation.

Evolutionary optimization algorithms such as particle swarm optimization have been explored for domain adaptation via feature selection (e.g., Nguyen et al. 2018; Sanodiya et al. 2020; Dhrif et al. 2020), but in these methods the feature selection step is not independent of the prediction step as the prediction error is used for optimizing the selected feature set in a computationally costly procedure. Castillo-García et al. (2023) attempt to separate the feature selection step from the prediction step by using complexity measures instead, but under the assumption that good feature sets would have lower data complexity.

Some published works have investigated the impact of causal feature selection on predictive performance through experimental studies. The Causation and Prediction

Challenge at the 2008 IEEE World Congress on Computational Intelligence (WCCI 2008) inspired a series of articles compiled in Guyon et al. (2010) which explore the use of causal inference techniques for machine learning. In the analysis of the results of the challenge, although the organizers confirmed the link between causation and prediction, they found that the prediction performance of causal feature selection and causal discovery methods generally did not meet expectations (Guyon et al. 2008). Overall, they did not consistently perform better than methods which do not consider causality (non-causal methods) both in tasks where the target sets were drawn from the same distribution as the training set, and those where some variables in the test set had undergone some manipulation or intervention.

However, there is supporting evidence from other studies indicating that causal feature selection can indeed improve prediction performance. Aliferis et al. (2010a,b) compare the prediction performance of various feature selection algorithms and found that the causal methods generally did achieve theoretically expected performance in terms of optimal prediction performance and feature set parsimony. In Yu et al. (2020) the causal feature selection algorithms mostly achieve better prediction performance compared to the non-causal methods evaluated. These studies however did not examine the impact of distribution shifts.

Some recent works have explored causality-based approaches for dealing with distribution shift. Causal methods allow the relaxation of the covariate shift assumption and have the advantage of improved robustness. Rojas-Carulla et al. (2018); Subbaswamy et al. (2019); Kügelgen et al. (2019) suggest causality-based approaches for stable modeling and prediction in potentially changing environments using only invariant causal features. Their methods however require prior knowledge of the causal structure of the data generating process and which features may change in new environments. Magliacane et al. (2018) relax this need for prior background knowledge but makes additional

assumptions about the causal structure based on the joint causal inference framework of Mooij et al. (2020). The techniques introduced in this present work offer the benefits of a causal approach to feature selection and domain adaptation without requiring prior knowledge of the data generating process. This is achieved without imposing further restrictive assumptions beyond those inherent in the SCM framework and specific causal discovery algorithms used.

5.1.2 Contribution

The novel contributions of this study include: (i) a new filter feature selection algorithm that prioritizes relevant causal relationships. (ii) a new domain adaptation strategy via a causal feature selection algorithm which adapts to univariate distributional changes between source and target data. (iii) a demonstration of the ability of causal approaches to feature selection to improve prediction performance for regression tasks in both in-distribution and out-of-distribution target datasets. (iv) new insights into the conditions that may lead to the Markov blanket delivering sub-optimal prediction performance as observed in some previous studies.

5.2 Background

5.2.1 Covariate Shift Adaptation

Changes in the probability distribution between source and target data distributions are an important challenge and key source of failure in machine learning (Moreno-Torres et al. 2012; Subbaswamy et al. 2022; Polo et al. 2023; Rahmani et al. 2023). This problem, often referred to as *dataset shift* has mostly been studied under three main categories: covariate shift, label shift, and concept shift. Let X be the set of

predictors and y the response or outcome variable in a data distribution, with the probability distribution of the source and target data represented by P_s and P_t respectively. Covariate shift is when $P_s(X) \neq P_t(X)$, prior probability or label shift is when $P_s(y) \neq P_t(y)$, and concept shift is when $P_s(y \mid X) \neq P_t(y \mid X)$. The covariate shift is the most commonly observed and studied form of dataset shift (Dharani et al. 2019; Xu et al. 2021).

Methods for covariate shift adaptation often rely on the covariate shift assumption which requires that the conditional distribution of the outcome remains invariant (i.e., no concept shift) (Sugiyama and Kawanabe 2012; Kügelgen et al. 2019). In this work, in addition to the standard covariate shift, we relax the covariate shift assumption so as to also consider a more extreme case of covariate shift which is caused by manipulations on covariates forcing them to take on a specific value. This type of shift which may or may not lead to some form of concept shift is referred to as *intervention shift* in this work.

Domain adaptation methods can be effective for mitigating the effects of covariate shifts on machine learning model performance (Kouw and Loog 2021; Xu et al. 2021). Most unsupervised domain adaptation methods can be classified into instance-based approaches and feature-based approaches. Instance or sample-based approaches are used for sample bias correction and often involve importance sampling and sample reweighting. Feature-based methods usually entail some transformation of the feature space for mapping the source distribution to the target distribution. Feature transformations often have a detrimental effect on model transparency and interpretability (Molnar 2019; Gosiewska et al. 2021; Fuchs et al. 2022).

The domain adaptation approach introduced in this paper is feature-based but does not entail any transformations of the feature space. Instead, we employ a univariate covariate shift detection strategy and a causality-based feature selection approach,

taking advantage of continued advancements in structural causal models (Pearl 2009b) and causal discovery (Glymour et al. 2019).

5.2.2 Causal Feature Selection

One aspect of machine learning that can benefit considerably from advancements in causal discovery is feature selection. Causal feature selection involves the learning of local causal structure around the outcome variable y for predictive modeling (Guyon et al. 2007; Aliferis et al. 2010a). Algorithms for learning such local causal structure from observational data can be grouped into two categories: (i) parents and children $PC(y)$ algorithms infer the direct causes (parents) and direct effects (children) of y, and (ii) Markov blanket $MB(y)$ algorithms infer the parents, children and spouses (direct causes of effects) of y.

The Markov blanket of y, $MB(y)$, by definition refers to any set of variables such that y is conditionally independent of all other variables in the data when conditioned on the variables in $MB(y)$ (Guyon et al. 2007). However, this term is often used loosely (as we do in this paper) to refer to the unique and minimal set of variables including the graphical neighborhood of y that consists of its parents, children and spouses, and which render all other variables conditionally independent of y (e.g., Tsamardinos et al. 2003; Pellet and Elisseeff 2008; Aliferis et al. 2010a). Some texts do distinguish the latter (minimal set) using the term Markov boundary (e.g., Yu et al. 2020).

Causal feature selection methods can be considered as part of the class of feature selection methods known as filters which function independently of a prediction model. The most common type of filter feature selection algorithm is the *select-k best* family of algorithms (Aliferis et al. 2010a). The select k-best approach is used for executing feature selection before prediction model training by first ranking the available features using some statistical criteria before selecting the k highest ranked features where k

is determined by the user (Miao and Niu 2016). Causal feature selection methods on the other hand generally attempt to identify an 'optimal' set of features for prediction. Optimal in this sense is in terms of the feature set with the minimum size that achieves the maximum predictive performance (Guyon and Elisseeff 2003; Tsamardinos et al. 2003; Yu et al. 2021).

Guyon et al. (2007) outline some benefits that a causal approach to feature selection may bring to machine learning over classical feature selection methods. These benefits include robustness to violations of the SSTD assumption, improved parsimony of selected feature sets, and enhanced data understanding and model interpretability. It has also been suggested that causality-based feature selection approaches can improve prediction performance in machine learning (Guyon et al. 2007; Kulynych 2022). This is to be expected from a theoretical point of view considering that causal feature selection methods are designed to find theoretically optimal feature sets like the $MB(y)$. This should help to avoid using features which contribute spurious information for model development, thereby improving the stability of machine learning models and reducing prediction errors.

In practice though, there is a lack in consistency of evidence that supports the optimality of the Markov blanket for prediction and the superiority of causal feature selection methods over non-causal (classical) filter methods in terms of prediction performance. The reported prediction performance in comparative studies have been somewhat mixed or heterogeneous (e.g. Cawley 2009; Guyon et al. 2008; Yu et al. 2021; Lemmon et al. 2023), with only a few studies demonstrating consistent maximal classification performance for causal feature selection (e.g. Aliferis et al. 2010a; Yu et al. 2020). It has also been suggested that the more parsimonious $PC(y)$ feature set generally does not produce inferior predictive performance compared to the theoretically optimal $MB(y)$ set in prediction tasks (e.g. Aliferis et al. 2010a; Yu et al. 2020, 2021).

The apparent contrasts between theoretical expectations and empirical observations with regard to the prediction performance of causal feature selection methods raises an interesting research question.

5.3 Methods

Let $X = \{x_1, x_2, ..., x_p\}$ be a set of variables that may be used for fitting or training a model for estimating an outcome variable y, $\{X_s, y_s\}$ forms the source data distribution to be used for model training. The model be used for obtaining predictions of the outcome in the target domain y_t using the covariates X_t. However, without employing special techniques to maintain model stability within the target domain, model performance is likely to deteriorate significantly in cases where substantial distributional shifts occur between the source and target domains.

The following two feature selection problem settings are considered in this work in the context of potentially different source and target data distributions:

1. The problem of selecting the k-best features from all available p features in X_s, for predicting y_t where k and p are integers.

2. The problem of selecting an optimal feature set for predicting y_t using labelled data from the source domain $\{X_s, y_s\}$ and unlabelled data samples $\{X_{t0}\}$ from the target domain, where $\{X_{t0}\} \subseteq \{X_t\}$.

5.3.1 Causal Feature Prioritization

The *select k-best* feature selection method works by calculating some score of each feature based on the strength of its relationship to the outcome variable before selecting

the k features with the best ranked scores. The scores are usually a measure of statistical dependence or mutual information. A commonly used ranking score for regression tasks is the linear correlation coefficient ρ given by Equation 5.1 for a pair of variables $\{x, y\}$, where $\bar{x}$ and $\bar{y}$ are the means of x and y respectively.

$$\rho_i = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2}\sqrt{\sum_i (y_i - \bar{y})^2}} \tag{5.1}$$

A causal approach to the select k-best method would need to consider the data generation mechanism in order to give priority to features that hold the highest causal relevance to the outcome. We suggest that prioritizing causal features in the selection of the k-best features could improve overall prediction performance across the range of k values. In this context, a feature selection algorithm that improves on the classical select k-best method by prioritizing causal features is hereby introduced.

As described in Algorithm 1, the proposed algorithm first learns two causal feature subsets consisting of the set of parents and children of the outcome $PC(y)$ and the Markov Blanket of the outcome $MB(y)$. It then proceeds to rank all features using a traditional ranking score such as ρ in Equation 5.1. Finally, k features are selected by considering the contents of the inferred causal sets $PC(y)$ and $MB(y)$, as well as the score rankings. Effectively the algorithm prioritizes the selection of features in the folowing order: 1) parents & children set, 2) spouses 3) statistical dependence ranking.

In the present implementation of Algorithm 1, $PC(y)$ set is learned using the Semi-Interleaved Hiton-PC (SI-HITON-PC) algorithm (Aliferis et al. 2010a) while $MB(y)$ set is learned using the Interleaved Incremental Association MB (Inter-IAMB) algorithm (Tsamardinos et al. 2003). The $score(x, y)$ function in Algorithm 1 calculates ρ and uses it to rank each predictor in the data.

Algorithm 1 Causal Feature Prioritizing (CFP) select k-best

Input: k, X_s, y_s
$PC_y \leftarrow$ find $PC(y)$ of y_s
$MB_y \leftarrow$ find $MB(y)$ of y_s
$NC_y \leftarrow$ score all features in X_s using $score(x, y)$ and sort
$FS_y \leftarrow$ initialize empty list of selected features
$c \leftarrow 1$; initialize counter
if $k \leq length(PC_y)$ **then**
  while $length(FS_y) \leq k$ **do**
    if $NC_y[c] \in PC_y$ **then**
      $FS_y.insert(NC_y[c])$
    end if
    $c = c + 1$
  end while
  return FS_y
else
  if $k \leq length(MB_y)$ **then**
    while $length(FS_y) \leq k$ **do**
      if $NC_y[c] \in MB_y$ **then**
        $FS_y.insert(NC_y[c])$
      end if
      $c = c + 1$
    end while
  end if
  return FS_y
else
  $FS_y \leftarrow MB_y$
  while $length(FS_y) \leq k$ **do**
    if $NC_y[c]$ **not in** FS_y **then**
      $FS_y.insert(NC_y[c])$
    end if
    $c = c + 1$
  end while
  return FS_y
end if

5.3.2 Shifted Child Feature Elimination for Domain Adaptation

Consider a setting where the objective is to train a prediction model using the source data $\{X_s, y_s\}$, for prediction of y_t given X_t in a target domain. Additionally, an unlabelled sample of the target data X_{t0} is available during model development. This setting which has received significant attention in the field of domain adaptation and transfer learning enables unsupervised adaptation to the target domain (Sun et al. 2016; Kouw and Loog 2021; Dhaini et al. 2023). We develop a feature-based domain adaptation strategy in the form of a feature selection method based on structural causal models (SCMs) that is capable of adapting to the target domain.

One of the benefits of representing a data generation process using an SCM is that SCMs clearly depict how influence flows through the network of variables in the data distribution and how changes are propagated through the network. For a quick introduction to SCMs and how they can be represented using both Bayesian networks (BNs) and structural equations models, see Mbogu and Nicholson (2023). In a causal BN representation, changes in the system follow a directional flow from parent nodes to children nodes, and not in the opposite direction. This knowledge can be exploited to design a feature selection strategy that is robust to distributional changes from source to target data by helping to decide which shifted covariates may affect the prediction of y_t.

A special case of covariate shift caused by 'perfect' interventions which fix the values of affected variables in a system is hereby considered. This type of fixed or *surgical* intervention as modeled by Pearl's (2009b) do-operator is useful for illustrating how a severe type of distribution shift in one node may affect the other variables in a data distribution (Pearl et al. 2016). Consider Figure 5.1 where an intervention on a variable x_1 brought about by a manipulation by an external agent forces x_1 to take on a specific value $x_1 = 0$ in the target domain. The effect of this type of manipulation is to

disconnect x_1 from all of its natural causes by externally setting its value as depicted in Figure 5.1b. Supposing x_3 is the outcome variable in this case, as a child of x_3, x_1 will likely prove to be a useful feature for predicting x_3 in the source domain (pre-intervention distribution). However, in the post-intervention target distribution using the model trained in the source domain, the presence of x_1 as a predictor for x_3 will indeed hamper prediction performance in the target domain because the changes to x_1 are not propagated to its parent. Hence, this work proposes to identify when such significant univariate distribution changes have occurred in specific features between the source and target distributions, and exclude children of the outcome variable affected by such a change from the training feature set in the source domain.

(a) Source: pre-intervention model (b) Target: post-intervention model

Figure 5.1: A BN illustration of a fixed/surgical intervention within the target domain.

We introduce a causal feature selection method that learns the Markov blanket of the outcome variable in the source domain, tests for distribution changes in individual features in the data, and adapts the Markov blanket to the target domain by excluding children of the outcome variable that have been remarkably impacted by a distribution shift. The adjusted Markov blanket forms the set of selected features for training a model for predicting y_t. The proposed method is described in Algorithm 2.

In the current implementation of Algorithm 2, the $MB(y)$ set is selected using the Inter-IAMB algorithm. The children of the outcome are identified after learning the full causal structure using the stable PC algorithm (Colombo and Maathuis 2014). The

Algorithm 2 MB Shifted Child Feature Elimination (SCFE)

Input: X_s, y_s, X_{t0}, α
$MB_y \leftarrow$ find $MB(y)$ of y_s
$CH_y \leftarrow$ find set of children of y_s
$SH_x \leftarrow$ find shifted covariates using $isShifted(X_s, X_{t0}, \alpha)$
$SH_{ch} \leftarrow$ obtain set of shifted children of y $SH_x \cap CH_y$
$FS_y \leftarrow$ initialize list of selected features
for each variable v in MB_y **do**
  if $v \notin SH_{ch}$ **then**
    $FS_y.insert(v)$
  end if
end for
return (FS_y)

function $isShifted(X_s, X_{t0}, \alpha)$ in Algorithm 2 performs a univariate distribution test on all features in the data, with α representing the significance level for the test. Its purpose is to identify which features have undergone a substantial distribution shift between the source and target domains. A small α is recommended to ensure that only features with remarkable distribution changes are singled out by the function.

The distributional test employed in the current implementation is the non-parametric Kolmogorov-Smirnov (KS) two-sample test or Smirnov test (Simard and L'Ecuyer 2011; Berger and Zhou 2014). The KS test is used for deciding whether two data samples come from the same distribution. For a given variable x_i the test statistic is the maximum value of the difference between the cummulative distribution functions (CDF) of the two samples. Suppose that x_i in the source data has sample size m with CDF $F_s(x_i)$ and in the target data has sample size n with CDF $F_t(x_i)$, the test statistic $D_{m,n}$ is given by Equation 5.2. The null hypothesis which states that the two samples are from equal distribution functions is rejected if $D_{m,n} > c(\alpha)\sqrt{\frac{m+n}{m.n}}$ where $c(\alpha)$ is the critical value at α.

$$D_{m,n} = max_x \mid F_s(x_i) - F_t(x_i) \mid \tag{5.2}$$

5.3.3 Experimental Setup

Five sets of data are simulated for evaluating different filter feature selection methods under three types of covariate shift in the target dataset: (1) no shift (baseline/unshifted target data), (2) simple covariate shift (covariate-shifted target data), and (3) fixed intervention shift (intervention-shifted target data). All simulated datasets are generated from linear structural causal models with Gaussian distributions. For each dataset, one out of its p features is selected as the outcome variable to be predicted while the rest of the $q = p - 1$ features make up the full set of potential predictors.

Two experiments are conducted to evaluate the utility of causal approaches to feature selection. Algorithms 1 and 2 are respectively assessed in Experiments 1 (Section 5.3.3.3) and 2 (Section 5.3.3.4). Two regression models, ordinary least squares (OLS) regression and support vector regression (SVR) with polynomial kernel are used for prediction of the outcome in the experiments. Given that the simulated datasets represent linear systems with Gaussian data distributions, OLS is a simple and appropriately specified estimating model. The support vector regression model is used as a more complex, more adaptive, but possibly slightly mispecified and overfitting alternative.

5.3.3.1 Data

Datasets used in the experiments are simulated from randomly generated structural causal models with different data generation hyperparameters. The procedure for data simulation is similar to the first part of the two-part simulation framework described in Mbogu and Nicholson (2023) where the data is generated from parametrized structural equations models. In this case, the SCMs and their model parameters are randomly generated.

First, a Bayesian network (BN) or directed acyclic graph (DAG) is generated at random given the number of nodes in the network ($numNodes$) and a probability value ($probConnect$) which specifies the probability that each node would be connected to other nodes in the network with higher topological ordering. The direction of graph edges in this DAG can only go from lower to higher topological ordering. This means we can have $x1 \rightarrow x2$ or $x10 \rightarrow x20$ but never $x1 \leftarrow x2$ or $x10 \leftarrow x20$ in the true DAG. Thus, a higher value of $probConnect$ produces a more densely connected DAG.

The generated DAG is converted to an equivalent non-parametric structural equations model (SEM) which is then parametrized and used to simulate data. In the SEM, each variable is a function of its parents in the DAG, and the coefficients of the variables in these functions are drawn from Gaussian distributions whose parameters, the mean and standard deviation, are in turn drawn from uniform distributions. The training datasets are then sampled from these parametrized Gaussian distributions.

Five sets of data **D1**, **D2**, **D3**, **D4** and **D5** are generated with each set including a source dataset and three target datasets. The first target dataset $BaseTarget$ is drawn from the same SEM and standard distributions as the training data. The second and third target datasets are generated by simulating a covariate shift and intervention shift respectively, on some features of the training data. This is achieved by introducing some disruption or perturbation in the parameters of the SEM. A binary vector of size p is used to indicate which variables will be perturbed and the elements of this vector are drawn from a Bernoulli distribution (the outcome variable is never perturbed). A hyperparameter, $disruptVecProb$, is used to specify the parameter of the Bernoulli distribution which sets the probability that each variable in the data will undergo some perturbation.

The second target dataset $CovTarget$, is covariate shifted. The covariate shift is achieved by adding an integer drawn from a uniform distribution to the mean of

the Gaussian distribution from which each perturbed variable is sampled from. For the third target dataset $IntTarget$, the perturbed variables are set to a specific value (zero), mimicking the effect of a fixed intervention. The user specified data generation parameters used for simulating the five datasets are summarized in Table 5.1.

Table 5.1: Data generation hyperparameters for the 5 simulated datasets: **D1**, **D2**, **D3**, **D4**, **D5**.

Hyperparameter	**D1**	**D2**	**D3**	**D4**	**D5**
$numNodes$[a]	31	31	61	81	101
$probConnect$[b]	0.1	0.07	0.07	0.05	0.03
$disruptVecProb$[c]	0.6	0.3	0.5	0.7	0.7
$nSource$[d]	20000	5000	20000	5000	40000
$nTarget$[e]	10000	2000	10000	2000	15000

[a] Number of nodes in the generated graph. Also translates to number of variables p in the simulated data.
[b] Probability that a node in the DAG is a parent of any other node of higher topological ordering. A measure of density/sparsity of the DAG.
[c] Probability that a node will experience some form of externally caused distribution change for target datasets $CovTarget$ and $IntTarget$.
[d] Source data sample size.
[e] Target data sample size.

5.3.3.2 Assumptions

Overall, the datasets on which Algorithms 1 and 2 have been evaluated have the following characteristics which are generally favorable to the causal discovery methods employed: relatively large samples, fairly sparse graphs, moderately sized feature set sizes, and causal sufficiency. A significant proportion of the features in the out-of-distribution target datasets are manipulated or shifted. The algorithms rely on the assumption that the data generating process can be modeled by a structural causal model and that the relevant parts of the SCM can be learned reliably from the data

using an appropriate causal discovery method. Algorithms 1 and 2 allow for freedom in the choice of causal discovery algorithms – this means that the standard assumptions of any causal discovery method employed applies. Additionally for Algorithm 2, an unlabelled sample of the target data X_{t0}, large enough for reliably detecting distribution changes in individual features is assumed to be available during model development.

5.3.3.3 Experiment 1

Experiment 1 is designed to investigate how the strategy of prioritizing causal features during feature selection may improve machine learning outcomes, particularly prediction performance. Using the seleck k-best approach, the goal is to compare prediction errors at various values of k for the Causal Feature Prioritizing (CFP) algorithm to those obtained using the correlation-based select k-best method (CB). This allows for a fair comparison in terms of similar selected feature set sizes. In addition, the performance of the theoretical optimal feature set, $MB(y)$, and its more parsimonious alternative, the $PC(y)$ set are to be noted. This experiment enables an evaluation of the Markov blanket induction theory which predicts maximal feature compactness/parsimony and optimal prediction performance when using the Makov blanket set (Aliferis et al. 2010a).

For the CFP and CB methods, the value of k starts from $k = q/15$ for $q = 30$, or $k = q/20$ for $q = 60$, 80 or 100, where q is the number of predictors in the dataset. The value of k is then progressively increased in each iteration by the same starting value while $k \leq q$. This way the maximum value of k corresponds to minimal or no feature selection. In each iteration the two select k-best algorithms, the classical CB and the CFP algorithms, are applied before ML models are trained using their selected feature sets. The models are then used to make predictions on the various target datasets.

5.3.3.4 Experiment 2

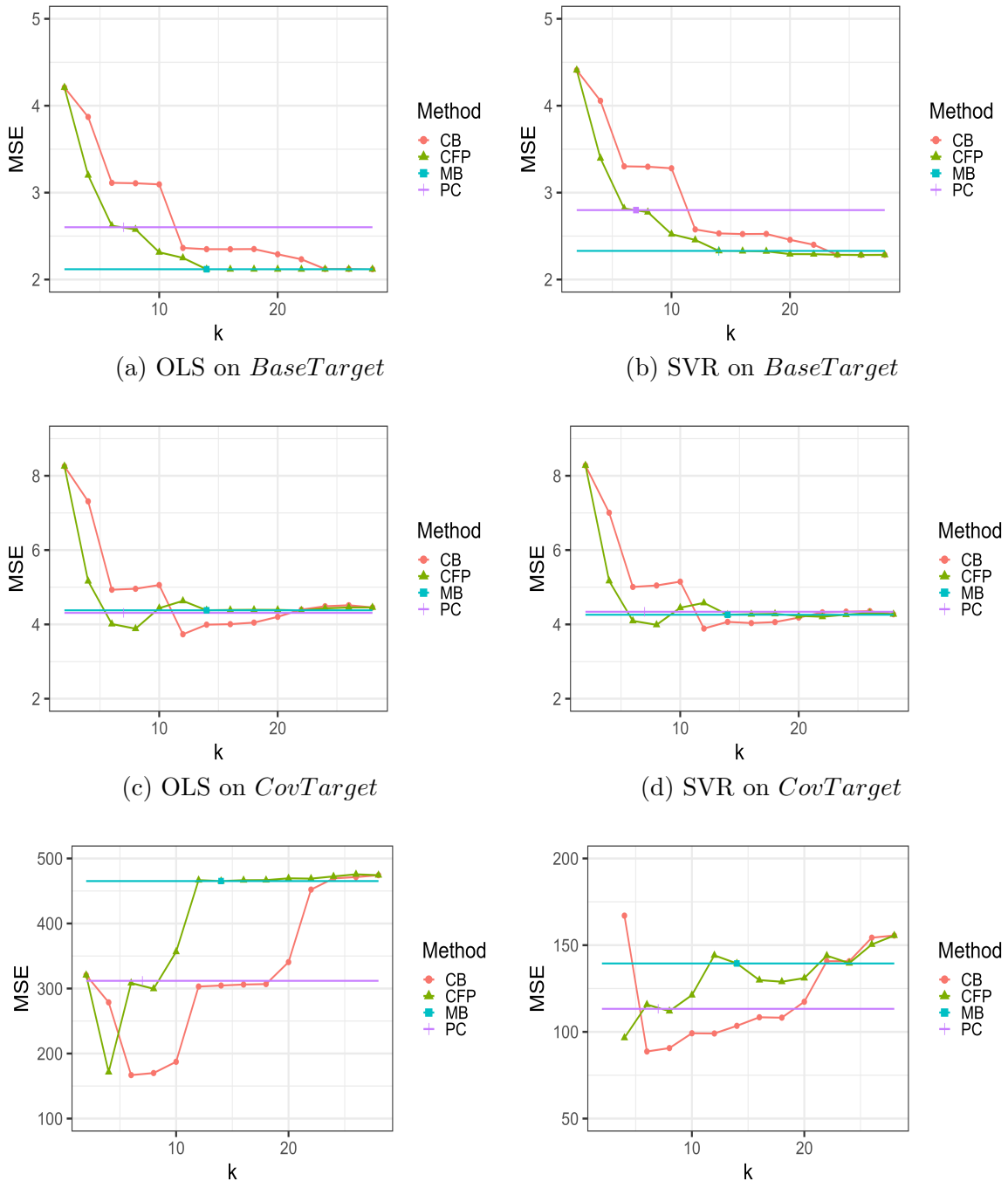
Experiment 2 is a follow up to Experiment 1 designed to test the hypothesis that excluding shifted children of the outcome from the training features could mitigate the effects of covariate shift on machine learning prediction performance in the target domain. The experiment compares prediction errors on the target datasets for models trained with the following feature sets: the set selected by Algorithm 2 (SCFEPred), the Markov blanket set (MBPred), the parents and children set (PCPred), and the full set of predictors with no feature selection applied (FullPred).

5.4 Results

5.4.1 Experiment 1 Results

Figure 5.2 depicts results obtained from Experiment 1 using data **D1**. It shows plots of mean squared error (MSE) values obtained from the OLS and SVR models using the three target sets $BaseTarget$, $CovTarget$ and $IntTarget$. The results for data **D2** to **D5** generally follow similar trends as Figure 5.2, and are shown in Appendix A. Notice the trends of the prediction performance of the feature sets selected using CFP and CB methods as k increases. The MSEs for the $MB(y)$ and $PC(y)$ sets are included for reference. Note the single k values associated with the $MB(y)$ and $PC(y)$ methods. As these methods do not take k as input, k in those instances is the size of the selected feature sets after the algorithms converge.

On the $BaseTarget$ dataset which follows a data distribution similar to the source, the CFP algorithm consistently outperforms the CB select k-best algorithm across the range of k values (e.g. Figures 5.2a and 5.2b). This indicates that causal feature prioritization can improve prediction performance when the SSTD assumption holds.



(a) OLS on *BaseTarget* (b) SVR on *BaseTarget*

(c) OLS on *CovTarget* (d) SVR on *CovTarget*

(e) OLS on *IntTarget* (f) SVR on *IntTarget*

Figure 5.2: Plots of prediction errors on data **D1** using OLS & SVR models.

Also, in this setting the $MB(y)$ feature sets consistently achieve the lowest MSEs using the fewest number of features particularly in the correctly specified OLS model. This is in line with theoretical expectations regarding the optimality of the Markov blanket for prediction modeling.

In the covariate-shifted (e.g., Figures 5.2c and 5.2d) and intervention-shifted target datasets (e.g., Figures 5.2e and 5.2f), it becomes evident that the CFP approach no longer assures enhanced performance compared to the CB approach. Similarly, the $MB(y)$ set seems to lose its optimality under these circumstances. Indeed the $PC(y)$ set seems likely to perform better under severe distribution shifts. These observations are more pronounced in the intervention-shifted target datasets where smaller values of k tend to perform better than larger k values. The performance of the $MB(y)$ set declines more dramatically compared to the $PC(y)$ set on $IntTarget$ as it appears that some features which are helpful for prediction in the $BaseTarget$ dataset become detrimental in the shifted target sets. Given that this trend arises when the SSTD assumption is violated, these results may explain why some previous experimental studies which did not consider the validity of the SSTD assumption in their datasets did not observe consistently superior performance of the $MB(y)$ set over the $PC(y)$ set.

5.4.2 Experiment 2 Results

Tables 5.2 and 5.3 show the mean squared error (MSE) values respectively achieved by OLS and SVR models using the three target datasets for data **D1** to **D5**. The number of features in the selected feature sets used in the models are provided in parenthesis. Again, $BaseTarget$, $CovTarget$, and $IntTarget$ are the unshifted baseline target set, covariate-shifted target set and intervention-shifted target set respectively. The selected feature sets used for training the models include: the full feature

set without any feature selection ($FullPred$), the feature set selected by the SCFE algorithm ($SCFEPred$), the $MB(y)$ feature set ($MBPred$), and the $PC(y)$ feature set ($PCPred$).

Again, it can be observed that the $MB(y)$ set ($MBPred$) achieves optimal prediction performance in the $BaseTarget$ set where the SSTD assumption holds (compared to $FullPred$ and $PCPred$), particularly in Table 5.2 where the model is correctly specified. As in Experiment 1, this optimal performance is no longer sustained in the presence of distribution shifts as observed in the $CovTarget$ and $IntTarget$ datasets, with the more compact $PC(y)$ set performing better than the $MB(y)$ set in several instances. In addition, the deterioration of prediction performance is much more severe in the intervention-shifted target sets.

In Table 5.3, again it is observed that the more adaptive SVR model seems to benefit from the availability of more features. For example, using the $FullPred$ set it is able to outperform $MBPred$ in some instances. Nonetheless, it is clear that the OLS model is a more precisely specified model as it always outperforms the SVR model in the in-distribution target set $BaseTarget$. In the shifted target sets however, the SVR model exhibits greater robustness to severe distribution changes in the target set.

Interestingly, the MSEs from the $SCFEPred$ set show that the SCFE algorithm usually succeeds in minimizing the detrimental effects of covariate and intervention shifts in the target data on prediction performance. When the prediction model is correctly specified as in the OLS models (Table 5.2), the SCFE algorithm always improves on or at least maintains the $MB(y)$ performance when predicting on the shifted target sets. The improvements are more remarkable in the intervention-shifted sets.

The SCFE algorithm appears to exhibit some sensitivity to model mispecification as its performance improvements on the $MBPred$ and $FullPred$ feature sets are diminished in the SVR model (Table 5.3). This apparent sensitivity may also have to

do with the superior adaptability of the SVR model which allows it to better tolerate distribution shifts when there are more variables in the prediction set to learn from, as is normally the case with *FullPred*. This adaptability would be derailed in the more parsimonious causal feature sets.

Notice from the feature set sizes in parenthesis that unlike the other methods which select the same number and set of features without any consideration of the target distribution, the SCFE feature selection method may produce different sets of features given different target/test sets. This is because the SCFE algorithm adapts the Markov blanket to the intended target domain in an attempt to retain the optimality of the Markov blanket in that domain.

Table 5.2: Comparison of MSE scores of **OLS** regression models trained using different feature sets (*FullPred*, *SCFEPred*, *MBPred*, *PCPred*) and predicting on different target sets (*BaseTarget*, *CovTarget*, *IntTarget*) from the five sets of data used in Experiment 2. In parenthesis beside every MSE score is an integer denoting the number of features in the selected feature set.

Data	Target Set	*FullPred*	*SCFEPred*	*MBPred*	*PCPred*
D1	$BaseTarget$	2.12 (30)	2.12 (14)	2.12 (14)	2.60 (7)
	$CovTarget$	4.39 (30)	3.69 (12)	4.38 (14)	4.31 (7)
	$IntTarget$	473.91 (30)	79.31 (12)	465.23 (14)	311.83 (7)
D2	$BaseTarget$	5.37 (30)	5.34 (8)	5.34 (8)	5.95 (6)
	$CovTarget$	7.34 (30)	7.00 (8)	7.00 (8)	6.56 (6)
	$IntTarget$	161.93 (30)	6.94 (7)	156.34 (8)	26.89 (6)
D3	$BaseTarget$	4.06 (60)	4.05 (17)	4.05 (17)	6.61 (8)
	$CovTarget$	8.61 (60)	6.65 (15)	8.54 (17)	6.73 (8)
	$IntTarget$	1112.34 (60)	6.68 (15)	1156.46 (17)	58.83 (8)
D4	$BaseTarget$	7.22 (81)	7.09 (8)	7.09 (8)	7.59 (6)
	$CovTarget$	8.06 (81)	7.82 (8)	7.82 (8)	8.60 (6)
	$IntTarget$	9.18 (81)	7.95 (8)	7.95 (8)	16.05 (6)
D5	$BaseTarget$	0.75 (100)	0.75 (22)	0.75 (22)	0.90 (8)
	$CovTarget$	1.73 (100)	1.41 (19)	1.70 (22)	1.99 (8)
	$IntTarget$	119.16 (100)	14.36 (19)	119.06 (22)	121.08 (8)

Table 5.3: Comparison of MSE scores of **SVR** regression models trained using different feature sets (*FullPred*, *SCFEPred*, *MBPred*, *PCPred*) and predicting on different target sets (*BaseTarget*, *CovTarget*, *IntTarget*) from the five sets of data used in Experiment 2. In parenthesis beside every MSE score is an integer denoting the number of features in the selected feature set.

Data	Target Set	*FullPred*	*SCFEPred*	*MBPred*	*PCPred*
D1	$BaseTarget$	2.28 (30)	2.33 (14)	2.33 (14)	2.80 (7)
	$CovTarget$	4.24 (30)	3.90 (12)	4.26 (14)	4.34 (7)
	$IntTarget$	155.75 (30)	55.60 (12)	139.46 (14)	113.29 (7)
D2	$BaseTarget$	5.69 (30)	5.66 (8)	5.66 (8)	6.26 (6)
	$CovTarget$	10.76 (30)	11.49 (8)	11.49 (8)	9.24 (6)
	$IntTarget$	40.39 (30)	10.72 (7)	35.06 (8)	18.48 (6)
D3	$BaseTarget$	4.32 (60)	4.26 (17)	4.26 (17)	6.76 (8)
	$CovTarget$	15.21 (60)	9.06 (15)	16.35 (17)	9.25 (8)
	$IntTarget$	23.82 (60)	103.90 (15)	21.14 (17)	55.22 (8)
D4	$BaseTarget$	7.80 (81)	7.26 (8)	7.26 (8)	7.79 (6)
	$CovTarget$	37.34 (81)	62.13 (8)	62.13 (8)	58.75 (6)
	$IntTarget$	432.25 (81)	393.76 (8)	393.76 (8)	390.76 (6)
D5	$BaseTarget$	0.79 (100)	0.81 (22)	0.81 (22)	0.99 (8)
	$CovTarget$	1.56 (100)	1.49 (19)	1.52 (22)	1.74 (8)
	$IntTarget$	43.07 (100)	17.32 (19)	46.36 (22)	44.45 (8)

5.5 Discussion

The value of causal inference is widely recognized in its capacity to provide tools for improving specific outcomes and addressing emerging concerns within ML/AI, particularly in the realms of model interpretability/explainability and decision-making. However, questions remain about whether employing causal methods to tackle these issues inadvertently leads to a reduction in prediction performance. In this work, we focus on investigating how causal machine learning can also lead to improvements in prediction performance, and consider prediction settings with both in-distribution and out-of-distribution target data. This inquiry leads us to consider the class of feature selection methods known as filters, which attempt to select the features most relevant to the outcome variable independent of any subsequent prediction model. The prominent causal feature selection methods belong to this class. Furthermore, we conduct experiments using simulated datasets in which the data properties are well understood.

5.5.1 CFP Algorithm

This paper suggests an improvement to the classical select k-best filter feature selection method that prioritizes causal features. This algorithm is referred to as the causal feature prioritizing (CFP) select k-best algorithm in this work. An experiment is conducted to investigate if when given a feature set size k, whether this feature selection approach can yield improved prediction performance over a comparable non-causal method. The results indicate that this is the case with in-distribution target datasets when the model is correctly specified. However, when dealing with out-of-distribution target datasets with covariate or intervention shifts, the consistent performance improvement observed in the in-distribution target set is not sustained. Furthermore, it's worth noting that although the theoretically optimal Markov blanket feature set

demonstrates the expected optimal performance within the in-distribution target set, it falls short in out-of-distribution target sets.

5.5.2 SCFE Algorithm

Following the first experiment, we explore how causal theory could be used to restore the optimality of the Markov blanket in out-of-distribution target datasets. A causal feature selection algorithm is proposed as a domain adaptation strategy for improving prediction performance within a target environment in the presence of substantial distribution shifts between the source and target datasets. This algorithm referred to as the MB Shifted Child Feature Elimination (SCFE) algorithm is evaluated in a second experiment using the same sets of source and target datasets as in the first experiment.

The results from the second experiment indicate that the proposed algorithm is able to mitigate the detrimental effect of covariate-type shifts on the performance of the Markov blanket set. In the correctly specified OLS model, the algorithm retains the optimality of the Markov blanket by delivering the best performance among the feature sets considered across all target datasets. When there is no detectable distribution shift in relevant features, it returns the same prediction performance as the Markov blanket feature set. Therefore, the algorithm exhibits the capability to adapt the Markov blanket to shifted target datasets.

Two limitations of previous works addressed by the domain adaptation strategy introduced in this paper are mentioned in Section 5.1.1. First, the lack of consideration of the conditional distribution and the local structure of the data is addressed by relaxing the covariate shift assumption in order to tolerate changes in the conditional distribution of the outcome from source to target domain, particularly those caused by interventions. Learning the Markov blanket ensures that the local structure around

the outcome variable is considered. Secondly, computationally expensive optimization procedures are avoided through a simple technique that employs covariate shift detection using a univariate distribution test which allows only features that are unlikely to degrade performance in the target domain to be retained in the Markov blanket. Furthermore, the proposed approach preserves model interpretability by avoiding the common feature-based domain adaptation strategy of learning new feature representations for adapting to a target domain.

5.5.3 Markov Blanket Induction Theory

The Markov blanket induction theory posits that the Markov blanket delivers optimal prediction performance using a minimal set of features (Aliferis et al. 2010a). In essence, one can achieve efficient and effective predictive modeling by focusing on the minimal set of features that are closely related to the outcome variable and which render all other variables independent of the outcome. This has the advantage of reducing the dimensionality of the problem, avoiding predicting based on spurious relationships and potentially improving model interpretability. Results from some previous experimental studies have raised questions about both the maximal prediction performance and minimal feature set properties expected of the Markov blanket. For example, Guyon et al. (2008); Cawley (2009) found that the Markov blanket did not outperform non-causal feature selection or no feature selection in their studies, while (Aliferis et al. 2010b; Yu et al. 2020) suggest that the $PC(y)$ feature set which usually has fewer features than $MB(y)$ does not exhibit inferior prediction performance compared to the $MB(y)$. Indeed, these results represent an apparent deviation from the theoretical expectations established by the Markov blanket theory. However, the findings from this study provide a new perspective and offer empirical support for the theory.

The results from this study indicate that the optimality of the Markov blanket remains valid in close to ideal conditions where the SSTD assumption is met, the prediction model is correctly specified, and the assumptions of the Markov blanket induction algorithms are reasonable with respect to the data. Causal discovery algorithms depend on specific assumptions which need to be taken into account in order to derive sound conclusions (Vonk et al. 2023). Therefore, the methods in this study are assessed using datasets with known properties and reasonable adherence to the assumptions of the causal methods employed.

Not adequately accounting for factors related to distribution shifts, bias in prediction models, and assumptions of Markov blanket learning algorithms may be responsible for the observed underwhelming prediction performance of the Markov blanket in some previous studies, especially those part of the causation and prediction challenge. As acknowledged by Guyon et al. (2008, 2010), the datasets used in the challenge involve complicated dependencies, and may likely violate some commonly made causal modeling assumptions. There were also a limited number of sound causal discovery and Markov blanket learning techniques available at the time of the challenge and most participants depended on a set of similar techniques based on the "work of Aliferis and Tsamardinos and their collaborators" (Aliferis et al. 2010a). Fortunately, since the challenge several causal discovery algorithms which relax various standard assumptions have emerged – see Zanga et al. (2022) for a recent review of the landscape.

These findings offer experimental evidence that helps clarify the specific prediction scenarios in which causal feature selection strategies excel. We observe that while using the $MB(y)$ for prediction is optimal when the prediction model is correctly specified, more adaptable non-parametric models may perform better in less-than-ideal settings given larger sets of statistically relevant features than the $MB(y)$. Also, given that most prior experimental studies have focused on classification problems, our results

in the context of regression problems demonstrate the significance of the concept of causal feature selection to regression.

5.5.4 Limitations and Future Work

From a feature selection point of view (not considering domain adaptation), the incorporation of causal discovery by the algorithms introduced in this paper introduces additional computational costs relative to comparable methods that rely on statistical dependence alone. In the current implementation of the CFP algorithm, the children of y are identified after learning the full structure of the data. An improvement in terms of computational cost could be achieved by exploring causal discovery methods that can distinguish parents from children without learning the full causal structure of the data. As Yu et al. (2020) indicate, score-based causal discovery algorithms may be useful for this specific purpose. It would also allow for causal features to be prioritized in the order of $parents > children > spouses$ instead of $parents \ \& \ children > spouses$, which may lead to further performance improvements in target datasets.

The causal feature selection methods proposed in this work exhibit sensitivity to model misspecification. Exploring the underlying causes of this sensitivity and developing approaches to mitigate it represents an interesting avenue for future research. Furthermore, there is need for research to explore methods for adapting the Markov blanket in other dynamic environments beyond the covariate or intervention shifts in the target distribution studied here.

Finally, the next logical step for this study is to validate the findings using real-world datasets. This will also enable a more in-depth exploration of important considerations for the selection of appropriate causal discovery methods and prediction models to be employed within the model-agnostic methods proposed in this work.

Chapter 6

Conclusions

The research endeavor culminating in this dissertation set out to investigate how causal inference methods can be used to improve outcomes in data-driven learning systems. This inquiry is motivated by an increasing appreciation among researchers of the potentials of causal approaches, with respect to offering solutions for resolving problems in AI. The major novel contributions of this work are elucidated in Chapters 4 and 5.

In Chapter 4, a data-driven root cause analysis method is developed for learning a graphical representation of root cause mechanisms, termed in this work as root cause graphs (RCGs), from time-to-event data. RCGs can be useful for identifying and analyzing the underlying causes of systemic problems in a wide range of settings. A simulation framework is suggested for generating realistic time-to-event datasets based on known causal structures and data generation parameters. Datasets generated from this framework are used to evaluate the root cause graph recovery method. In all scenarios tested, useful approximations of the ground truth causal structures are recovered with high rates of correct edge adjacencies. The utility of learning these root cause graphs from data is demonstrated in their use for improving the estimation of causal effects of mediated root cause variables.

In Chapter 5, two causal feature selection methods are proposed for improving machine learning outcomes in two different prediction settings. Outcomes of particular interest include domain adaptability and model interpretability. This study demonstrates that a strategy centered on prioritizing causal features in prediction models can

lead to improvements in both in-distribution and out-of-distribution prediction performance. Also, a causal feature selection approach that leverages structural causal model (SCM) theory for adjusting the Markov blanket feature set to a target dataset can be a successful approach to domain adaptation in scenarios where severe and widespread distribution changes are anticipated between source and target distributions.

Future research should work on relaxing common assumptions for causal inference so that they can be reasonably applied to a wider range of problem settings with more complex datasets. Collaboration with industry process owners and domain experts on research that incorporates causal discovery can be used to learn how to adapt existing methods to domain-specific problems. They can also be useful for validating or rethinking the contemporary understanding of stakeholders about the workings of their processes. Furthermore, the causal feature selection methods proposed in this work exhibit sensitivity to model misspecification. Exploring the underlying causes of this sensitivity and developing approaches to mitigate it represents an interesting avenue for future research. Additionally, there is need for research to explore methods for adapting the Markov blanket feature set in other dynamic environments beyond the covariate or intervention shifts in the target distribution studied here.

The remaining objectives of this research outlined in Section 1.3 are addressed by initially delving into the theoretical frameworks that support causal inference using observational data. The SCM framework incorporates aspects of the potential outcome framework, probabilistic graphical modeling, and structural equations modeling, for a broad theory of causality that includes tools for representing, manipulating and identifying causal relationships in various scenarios. This comprehensive framework enables advanced causal investigations using observational data. Additionally, several other frameworks and causal methodologies are continually being developed, building upon the foundation laid by the SCM framework.

Expanding upon the concepts introduced within the frameworks for causal learning, this work conducts a review of techniques for causal inference, with a focus on utilizing non-experimental data. The majority of causal learning methods are categorized into two primary tasks: treatment effect estimation, and causal structure modeling. Causal discovery methods fall within the realm of causal structure modeling tasks and have garnered significant attention recently for their capacity to uncover elements of causal structure from observational data, contingent on specific assumptions. This work explores a diverse range of techniques for achieving both causal learning tasks, as well as strategies for integrating causal inference and machine learning.

A central focus of this dissertation has been the intersection between causal inference and statistical/machine learning. The objectives of prediction and inference can both be improved significantly by combining these modeling approaches. The reciprocal benefits that each learning paradigm affords the other are summarized as follows:

Causal inference aids ML by:

- improving ML generalization capabilities.
- enhancing model transparency, interpretability and trustworthiness.
- improving reproducibility and external validity of ML models.
- facilitating inference and knowledge discovery.
- providing tools for mitigating bias and advancing fair ML.

ML aids causal inference by:

- providing flexible tools for modeling and estimating treatment effects.
- providing efficient estimators for counterfactual effects.
- providing techniques for efficiently handling high dimensional data.
- providing new approaches for Bayesian network and causal structure learning.

The new methods presented in this work successfully integrate causal inference techniques and statistical/machine learning methodologies to offer solutions for resolving challenges in data-driven learning applications. Specifically, the proposed RCA approach synergizes causal discovery techniques with effect estimation models to enhance the identification of the underlying causes of events of interest and the estimation of their effects, using time-to-event data. Furthermore, the causal feature selection methods introduced here incorporate causal learning strategies into machine learning and prediction modeling techniques. This approach not only improves model generalization for domain adaptation problems where the target data may be out-of distribution, it is also shown to improve prediction performance on in-distribution target data when modeling assumptions are met.

Although specific models have been implemented to demonstrate the utility of the proposed methods, it is important to note that these methods are intentionally designed to be model-agnostic. This allows for flexibility in the selection of models and algorithms for specific tasks within the procedures such as outcome prediction, causal discovery, and causal effect estimation. This flexibility is important as it empowers data scientists to make informed choices regarding models and algorithms to be employed within the proposed methods. These choices can be tailored to the unique requirements of a given problem and properties of the data when implementing the suggested solutions. This ensures that the assumptions underlying the chosen approaches are reasonable with respect to the problem at hand.

In summary, this work represents a significant step toward climbing the ladder of causation. It contributes to the ongoing efforts aimed at developing more capable, reliable, and trustworthy AI technologies.

Reference List

Al Hajj, G. S., J. Pensar, and G. K. Sandve, 2023: DagSim: Combining DAG-based model structure with unconstrained data types and relations for flexible, transparent, and modularized data simulation. *Plos one*, **18 (4)**, e0284 443.

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

Aliferis, C. F., A. Statnikov, I. Tsamardinos, S. Mani, and X. D. Koutsoukos, 2010a: Local causal and markov blanket induction for causal discovery and feature selection for classification. Part I: Algorithms and empirical evaluation. *Journal of Machine Learning Research*, **11**, 171–234.

Aliferis, C. F., A. Statnikov, I. Tsamardinos, S. Mani, X. D. Koutsoukos, and M. Meila, 2010b: Local causal and markov blanket induction for causal discovery and feature selection for classification. Part II: Analysis and extensions. *Journal of Machine Learning Research*, **11**, 235–284.

Alizadeh, E., M. E. Koujok, A. Ragab, and M. Amazouz, 2018: A Data-Driven Causality Analysis Tool for Fault Diagnosis in Industrial Processes. *IFAC-PapersOnLine*, **51 (24)**, 147–152, https://doi.org/10.1016/j.ifacol.2018.09.548.

Andrews, B., J. Ramsey, and G. F. Cooper, 2018: Scoring Bayesian Networks of Mixed Variables. *International Journal of Data Science and Analytics*, **6 (1)**, 3–18, https://doi.org/10.1007/s41060-017-0085-7.

Angelov, P. P., N. I. Arnold, E. A. Soares, P. M. Atkinson, R. Jiang, N. I. Arnold, and P. M. Atkinson, 2021: Explainable artificial intelligence: an analytical review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, **11 (5)**, e1424, https://doi.org/10.1002/widm.1424.

Angrist, J. D., G. W. Imbens, and D. B. Rubin, 1996: Identification of causal effects using instrumental variables. *Journal of the American statistical Association*, **91 (434)**, 444–455.

Angrist, J. D., and A. B. Krueger, 2001: Instrumental variables and the search for identification: From supply and demand to natural experiments. *Journal of Economic Perspectives*, **15 (4)**, 69–85, https://doi.org/10.1257/jep.15.4.69.

Athey, S., and G. Imbens, 2016: Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences of the United States of America*, **113 (27)**, 7353–7360, https://doi.org/10.1073/pnas.1510489113, 1504.01132.

Athey, S., and S. Wager, 2019: Estimating Treatment Effects with Causal Forests: An Application. *Observational Studies*, **5 (2)**, 37–51, https://doi.org/10.1353/obs.2019.0001, 1902.07409.

Austin, P. C., 2012: Using ensemble-based methods for directly estimating causal effects: an investigation of tree-based G-computation. *Multivariate behavioral research*, **47 (1)**, 115–135.

Austin, P. C., 2014: A comparison of 12 algorithms for matching on the propensity score. *Statistics in Medicine*, **33 (6)**, 1057–1069, https://doi.org/10.1002/sim.6004.

Bahadori, M. T., K. Chalupka, E. Choi, R. Chen, W. F. Stewart, and J. Sun, 2017: Causal Regularization. *arXiv preprint*, 1–18, 1702.02604.

Balke, A., and J. Pearl, 2013: Counterfactuals and policy analysis in structural models. *arXiv preprint arXiv:1302.4929*.

Balzer, L. B., and M. L. Petersen, 2021: Invited Commentary: Machine Learning in Causal Inference-How Do I Love Thee? Let Me Count the Ways. *American Journal of Epidemiology*, **190 (8)**, 1483–1487, https://doi.org/10.1093/aje/kwab048.

Barocas, S., M. Hardt, and A. Narayanan, 2017: Fairness in machine learning. *Nips tutorial*, **1**, 2.

Belloni, A., D. Chen, V. Chernozhukov, and C. Hansen, 2012: Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, **80 (6)**, 2369–2429.

Belloni, A., and V. Chernozhukov, 2013: Least squares after model selection in high-dimensional sparse models. *Bernoulli*, **19 (2)**, 521–547, https://doi.org/10.3150/11-BEJ410, 1001.0188.

Belloni, A., V. Chernozhukov, I. Fernandez-Val, and C. Hansen, 2017: Program Evaluation and Causal Inference With High-Dimensional Data. *Econometrica*, **85 (1)**, 233–298, https://doi.org/10.3982/ecta12723, 1311.2645.

Belloni, A., V. Chernozhukov, and C. Hansen, 2014a: High-dimensional methods and inference on structural and treatment effects. *Journal of Economic Perspectives*, **28 (2)**, 29–50.

Belloni, A., V. Chernozhukov, and C. Hansen, 2014b: Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, **81 (2)**, 608–650.

Belloni, A., V. Chernozhukov, C. Hansen, and D. Kozbur, 2016: Inference in high-dimensional panel models with an application to gun control. *Journal of Business & Economic Statistics*, **34 (4)**, 590–605.

Bender, R., T. Augustin, and M. Blettner, 2005: Generating survival times to simulate Cox proportional hazards models. *Statistics in Medicine*, **24 (11)**, 1713–1723, https://doi.org/10.1002/sim.2059.

Bengio, Y., A. Courville, and P. Vincent, 2013: Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **35 (8)**, 1798–1828, https://doi.org/10.1109/TPAMI.2013.50, 1206.5538.

Bengio, Y., T. Deleu, N. Rahaman, R. Ke, S. Lachapelle, O. Bilaniuk, A. Goyal, and C. Pal, 2019: A Meta-Transfer Objective for Learning to Disentangle Causal Mechanisms. *In Eighth International Conference on Learning Representations*, 1–26, 1901.10912.

Berg-Schlosser, D., G. De Meur, B. Rihoux, and C. C. Ragin, 2009: Qualitative comparative analysis (QCA) as an approach. *Configurational comparative methods: Qualitative comparative analysis (QCA) and related techniques*, **1**, 18.

Berger, V. W., and Y. Zhou, 2014: Kolmogorov–Smirnov Test: Overview. *Wiley statsref: Statistics reference online*.

Berrada, M., A. Adadi, and M. Berrada, 2018: Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE access*, **6**, 52 138–52 160, https://doi.org/10.1109/ACCESS.2018.2870052.

Blöbaum, P., and S. Shimizu, 2017: Estimation of interventional effects of features on prediction. *2017 IEEE 27th International Workshop on Machine Learning for Signal Processing (MLSP)*, IEEE, 1–6.

Bonchi, F., S. Hajian, B. Mishra, and D. Ramazzotti, 2017: Exposing the probabilistic causal structure of discrimination. *International Journal of Data Science and Analytics*, **3 (1)**, 1–21, https://doi.org/10.1007/s41060-016-0040-z.

Bongers, S., P. Forré, J. Peters, and J. M. Mooij, 2021: Foundations of structural causal models with cycles and latent variables. *Annals of Statistics*, **49 (5)**, 2885–2915, https://doi.org/10.1214/21-AOS2064.

Bottou, L., and Coauthors, 2013: Counterfactual reasoning and learning systems: The example of computational advertising. *Journal of Machine Learning Research*, **14**, 3207–3260.

Brady, S., 2020: Using Unsupervised Learning to Help Discover the Causal Graph. *arXiv preprint*, **(2016)**, 2009.10790.

Brand, J. E., and J. S. Thomas, 2013: Causal effect heterogeneity. *Handbook of causal analysis for social research*, Springer, 189–213.

Brilleman, S. L., R. Wolfe, M. Moreno-Betancur, and M. J. Crowther, 2021: Simulating survival data using the simsurv R package. *Journal of Statistical Software*, **97 (3)**, 1–27, https://doi.org/10.18637/jss.v097.i03.

Brundage, M., and Coauthors, 2018: The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. *arXiv preprint arXiv:1802.07228*, **(February)**, https://doi.org/10.17863/CAM.22520, 1802.07228.

Bühlmann, P., J. Peters, and J. Ernest, 2014: CAM: Causal additive models, high-dimensional order search and penalized regression. *The Annals of Statistics*, **42 (6)**, 2526–2556.

Cartwright, N., 2004: Causation: One word, many things. *Philosophy of Science*, **71 (5)**, 805–819, https://doi.org/10.1086/426771.

Cartwright, N., 2010: What are randomised controlled trials good for? *Philosophical studies*, **147 (1)**, 59–70.

Castillo-García, G., L. Morán-Fernández, and V. Bolón-Canedo, 2023: Feature selection for domain adaptation using complexity measures and swarm intelligence. *Neurocomputing*, **548**, https://doi.org/10.1016/j.neucom.2023.126422.

Cawley, G. C., 2009: Causal & Non-Causal Feature Selection for Ridge Regression. *Causation and Prediction Challenge: Challenges in Machine Learning*, PMLR, Vol. 3, 107–128.

Chambliss, D. F., and R. K. Schutt, 2018: *Making sense of the social world: Methods of investigation*. Sage Publications.

Chatton, A., and Coauthors, 2020: G-computation, propensity score-based methods, and targeted maximum likelihood estimator for causal inference with different covariates sets: a comparative simulation study. *Scientific reports*, **10 (1)**, 1–13.

Chaves, R., L. Luft, T. O. Maciel, D. Gross, D. Janzing, and B. Schölkopf, 2014: Inferring latent structures via information inequalities. *30th Conference on Uncertainty in Artificial Intelligence*, 1–14.

Chen, H. S., Z. Yan, Y. Yao, T. B. Huang, and Y. S. Wong, 2018: Systematic Procedure for Granger-Causality-Based Root Cause Diagnosis of Chemical Process Faults. *Industrial and Engineering Chemistry Research*, **57 (29)**, 9500–9512, https://doi.org/10.1021/acs.iecr.8b00697.

Cheng, L., A. Mosallanezhad, P. Sheth, and H. Liu, 2021: Causal Learning for Socially Responsible AI. *IJCAI International Joint Conference on Artificial Intelligence*, 4374–4381., https://doi.org/10.24963/ijcai.2021/598.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, and W. Newey, 2017a: Double/debiased/Neyman machine learning of treatment effects. *American Economic Review*, **107 (5)**, 261–265, https://doi.org/10.1257/aer.p20171038, 1701.08687.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins, 2017b: Double/Debiased Machine Learning for Treatment and Causal Parameters. *Econometrics Journal*, 1608.00060.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins, 2018: Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, **21 (1)**.

Chesnaye, N. C., V. S. Stel, G. Tripepi, F. W. Dekker, E. L. Fu, C. Zoccali, and K. J. Jager, 2022: An introduction to inverse probability of treatment weighting in observational research. *Clinical Kidney Journal*, **15 (1)**, 14–20, https://doi.org/10.1093/ckj/sfab158.

Chickering, D. M., 2003: Optimal structure identification with greedy search. *Journal of Machine Learning Research*, **3 (3)**, 507–554, https://doi.org/10.1162/153244303321897717.

Chickering, M., D. Heckerman, and C. Meek, 2004: Large-sample learning of Bayesian networks is NP-hard. *Journal of Machine Learning Research*, **5**, 1287–1330.

Coffman, D. L., and W. Zhong, 2012: Assessing mediation using marginal structural models in the presence of confounding and moderation. *Psychological methods*, **17 (4)**, 642.

Colombo, D., and M. H. Maathuis, 2014: Order-independent constraint-based causal structure learning. *Journal of Machine Learning Research*, **15 (1)**, 3741–3782.

Colombo, D., M. H. Maathuis, M. Kalisch, and T. S. Richardson, 2012: Learning high-dimensional directed acyclic graphs with latent and selection variables. *Annals of Statistics*, **40 (1)**, 294–321, https://doi.org/10.1214/11-AOS940, 1104.5617.

Confalonieri, R., L. Coba, B. Wagner, and T. R. Besold, 2021: A historical perspective of explainable Artificial Intelligence. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, **11 (1)**, e1391.

Contaldi, C., F. Vafaee, and P. C. Nelson, 2019: Bayesian network hybrid learning using an elite-guided genetic algorithm. *Artificial Intelligence Review*, **52 (1)**, 245–272, https://doi.org/10.1007/s10462-018-9615-5.

Corbett-davies, S., and S. Goel, 2022: The measure and mismeasure of fairness: A critical review of fair machine learning. *ACM Computing Surveys (CSUR)*, **3 (55)**, 1–44, arXiv:1808.00023v2.

Crowther, M. J., and P. C. Lambert, 2012: Simulating complex survival data. *Stata Journal*, **12 (4)**, 674–687, https://doi.org/10.1177/1536867x1201200407.

Crowther, M. J., and P. C. Lambert, 2013: Simulating biologically plausible complex survival data. *Statistics in Medicine*, **32 (23)**, 4118–4134, https://doi.org/10.1002/sim.5823.

Curth, A., and M. van der Schaar, 2021: Nonparametric estimation of heterogeneous treatment effects: From theory to learning algorithms. *International Conference on Artificial Intelligence and Statistics*, PMLR, Vol. 130, 1810–1818.

Dai, J., J. Ren, W. Du, V. Shikhin, and J. Ma, 2020: An improved evolutionary approach-based hybrid algorithm for Bayesian network structure learning in dynamic constrained search space. *Neural Computing and Applications*, **32 (5)**, 1413–1434.

Daniel, R. M., B. L. De Stavola, and S. N. Cousens, 2011: gformula: Estimating causal effects in the presence of time-varying confounding or mediation using the g-computation formula. *The Stata Journal*, **11 (4)**, 479–517.

Daumé, H., 2007: Frustratingly easy domain adaptation. *ACL 2007 - Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, 256–263.

Daw, A., and Coauthors, 2017: Physics-guided neural networks (pgnn): An application in lake temperature modeling. *arXiv preprint arXiv:1710.11431*, arXiv:1710.11431v3.

De Campos, C. P., Z. Zeng, Q. Ji, and C. P. D. Campos, 2009: Structure learning of Bayesian networks using constraints. *Proceedings of the 26th Annual International Conference on Machine Learning*, 113–120.

de Campos, L. M., and J. G. Castellano, 2007: Bayesian network learning algorithms using structural restrictions. *International Journal of Approximate Reasoning*, **45 (2)**, 233–254.

de los Angeles Resa, M., and J. R. Zubizarreta, 2016: Evaluation of subset matching methods and forms of covariate balance. *Statistics in medicine*, **35 (27)**, 4961–4979.

Deng, W. Y., A. Lendasse, Y. S. Ong, I. W. H. Tsang, L. Chen, and Q. H. Zheng, 2019: Domain Adaption via Feature Selection on Explicit Feature Map. *IEEE Transactions on Neural Networks and Learning Systems*, **30 (4)**, 1180–1190, https://doi.org/10.1109/TNNLS.2018.2863240.

Dhaini, M., M. Berar, P. Honeine, and A. Van Exem, 2023: Unsupervised domain adaptation for regression using dictionary learning. *Knowledge-Based Systems*, **267**, 110 439.

Dharani, G., N. G. Nair, P. Satpathy, and J. Christopher, 2019: Covariate Shift: A Review and Analysis on Classifiers. *2019 Global Conference for Advancement in Technology, GCAT 2019*, 20–25., https://doi.org/10.1109/GCAT47503.2019.8978471.

Dhrif, H., V. Bolon-Canedo, and S. Wuchty, 2020: Gene Subset Selection for Transfer Learning using Bilevel Particle Swarm Optimization. *Proceedings - 19th IEEE International Conference on Machine Learning and Applications, ICMLA 2020*, 1317–1323., https://doi.org/10.1109/ICMLA51294.2020.00206.

Diaz, I., 2020: Machine learning in the estimation of causal effects: targeted minimum loss-based estimation and double/debiased machine learning. *Biostatistics*, **21 (2)**, 353–358.

Dimick, J. B., and A. M. Ryan, 2014: Methods for Evaluating Changes in Health Care Policy: The Difference-in-Differences Approach. *JAMA - Journal of the American Medical Association*, **312 (22)**, 2401–2402, https://doi.org/10.1001/jama.2014.16153.

Doggett, A. M., 2005: Root Cause Analysis: A Framework for Tool Selection. *Quality Management Journal*, **12 (4)**, 34–45, https://doi.org/10.1080/10686967.2005.11919269.

Dudik, M., J. Langford, and H. Li, 2011: Doubly robust policy evaluation and learning. *Proceedings of the 28th International Conference on Machine Learning, ICML 2011*, 1097–1104, 1103.4601.

Eberhardt, F., 2017: Introduction to the foundations of causal discovery. *International Journal of Data Science and Analytics*, **3 (2)**, 81–91, https://doi.org/10.1007/s41060-016-0038-6.

Eberhardt, F., C. Glymour, and R. Scheines, 2005: On the number of experiments suficient and in the worst case necessary to identify all causal relations among N variables. *Proceedings of the 21st Conference on Uncertainty in Artificial Intelligence, UAI 2005*, 178–184, 1207.1389.

Eitel-Porter, R., 2021: Beyond the promise: implementing ethical AI. *AI and Ethics*, **1 (1)**, 73–80.

Elwert, F., 2013: Graphical causal models. *Handbook of causal analysis for social research*, Springer, 245–273.

Farahani, A., S. Voghoei, K. Rasheed, and H. R. Arabnia, 2021: A Brief Review of Domain Adaptation. *Advances in data science and information engineering*, 877–894, October, https://doi.org/10.1007/978-3-030-71704-9_65.

Farrell, M. H., T. Liang, and S. Misra, 2021: Deep neural networks for estimation and inference. *Econometrica*, **89 (1)**, 181–213.

Feder, A., N. Oved, U. Shalit, and R. Reichart, 2021: Causalm: Causal model explanation through counterfactual language models. *Computational Linguistics*, **47 (2)**, 333–386.

Flanders, W. D., 2006: On the relationship of sufficient component cause models with potential outcome (counterfactual) models. *European Journal of Epidemiology*, **21 (12)**, 847–853, https://doi.org/10.1007/s10654-006-9048-3.

Frosch, C. A., and P. N. Johnson-Laird, 2011: Is everyday causation deterministic or probabilistic? *Acta psychologica*, **137 (3)**, 280–291.

Fuchs, C., S. Spolaor, U. Kaymak, and M. S. Nobile, 2022: The Impact of Variable Selection and Transformation on the Interpretability and Accuracy of Fuzzy Models. *2022 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology, CIBCB 2022*, https://doi.org/10.1109/CIBCB55180.2022.9863019.

Funk, M. J., D. Westreich, C. Wiesen, T. Stürmer, M. A. Brookhart, and M. Davidian, 2011: Doubly robust estimation of causal effects. *American Journal of Epidemiology*, **173 (7)**, 761–767, https://doi.org/10.1093/aje/kwq439.

Galles, D., and J. Pearl, 1995: Testing Identifiability of Causal Effects. *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, 185–195.

Garrido, S., S. Borysov, J. Rich, and F. Pereira, 2021: Estimating causal effects with the neural autoregressive density estimator. *Journal of Causal Inference*, **9 (1)**, 211–218.

Gelman, A., 2011: Causality and Statistical Learning. *American Journal of Sociology*, **117 (3)**, 955–966.

Glymour, C., K. Zhang, and P. Spirtes, 2019: Review of causal discovery methods based on graphical models. *Frontiers in Genetics*, **10 (JUN)**, 1–15, https://doi.org/10.3389/fgene.2019.00524.

Gobble, M. M., 2019: The road to artificial general intelligence. *Research-Technology Management*, **62 (3)**, 55–59, https://doi.org/10.1080/08956308.2019.1587336.

Goertzel, B., 2014: Artificial general intelligence: concept, state of the art, and future prospects. *Journal of Artificial General Intelligence*, **5 (1)**, 1.

Gosiewska, A., A. Kozak, and P. Biecek, 2021: Simpler is better: Lifting interpretability-performance trade-off via automated feature engineering. *Decision Support Systems*, **150 (July 2020)**, https://doi.org/10.1016/j.dss.2021.113556.

Granger, C. W. J., 1969: Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: journal of the Econometric Society*, 424–438.

Greenland, S., and J. Pearl, 2006: Causal diagrams. *Encyclopedia of Epidemiology*, **(June)**, 1–12.

Greenland, S., and J. Pearl, 2011: Adjustments and their consequences—collapsibility analysis using graphical models. *International Statistical Review*, **79 (3)**, 401–426.

Grosz, M. P., J. M. Rohrer, and F. Thoemmes, 2020: The taboo against explicit causal inference in nonexperimental psychology. *Perspectives on Psychological Science*, **15 (5)**, 1243–1255.

Gruber, S., and M. J. Van Der Laan, 2009: Targeted maximum likelihood estimation: A gentle introduction. *U.C. Berkeley Division of Biostatistics Working Paper Series*, **(252)**.

Gu, J., F. Fu, and Q. Zhou, 2019: Penalized estimation of directed acyclic graphs from discrete data. *Statistics and Computing*, **29 (1)**, 161–176.

Gu, X. S., P. R. Rosenbaum, X. Sam, P. R. Rosenbaum, S. Journal, and G. Statistics, 1993: Comparison of multivariate matching methods: Structures, distances, and algorithms. *Journal of Computational and Graphical Statistics*, **2 (4)**, 405–420.

Guo, R., L. Cheng, J. Li, P. R. Hahn, and H. Liu, 2020: A Survey of Learning Causality with Data: Problems and Methods. *ACM Computing Surveys*, **53 (4)**, 1–36, https://doi.org/10.1145/3397269, 1809.09337.

Guyon, I., C. Aliferis, W. F. Carvalho, and L. Zarate, 2007: Causal feature selection. *Computational methods of feature selection*, Chapman and Hall/CRC, 79–102, https://doi.org/10.4018/978-1-7998-5781-5.ch007.

Guyon, I., C. Aliferis, and G. Cooper, 2010: *Causation and Prediction Challenge: Challenges in Machine Learning, Volume 2*. Challenges in machine learning, Microtome Publishing.

Guyon, I., C. Aliferis, G. Cooper, A. A. Elisseeff, J.-P. Pellet, P. Spirtes, and A. Statnikov, 2008: Design and analysis of the causation and prediction challenge. *Causation and Prediction Challenge*, PMLR, Vol. 3, 1–33.

Guyon, I., and A. Elisseeff, 2003: An introduction to variable and feature selection. *Journal of Machine Learning Research*, **3 (Mar)**, 1157–1182.

Hahn, P. R., J. S. Murray, and C. M. Carvalho, 2020: Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects (with discussion). *Bayesian Analysis*, **15 (3)**, 965–1056, https://doi.org/10.1214/19-BA1195.

Hajian, S., F. Bonchi, and C. Castillo, 2016: Algorithmic bias: From discrimination discovery to fairness-aware data mining. *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2125–2126.

Harden, J. J., and J. Kropko, 2019: Simulating duration data for the cox model. *Political Science Research and Methods*, **7 (4)**, 921–928, https://doi.org/10.1017/psrm.2018.19.

Hartford, J., G. Lewis, K. Leyton-Brown, and M. Taddy, 2017: Deep IV: A flexible approach for counterfactual prediction. *International Conference on Machine Learning*, PMLR, 1414–1423.

Hauser, A., and P. Bühlmann, 2012: Characterization and greedy learning of interventional markov equivalence classes of directed acyclic graphs. *Journal of Machine Learning Research*, **13**, 2409–2464, 1104.2808.

Hauser, A., and P. Bühlmann, 2014: Two optimal strategies for active learning of causal models from interventional data. *International Journal of Approximate Reasoning*, **55 (4)**, 926–939, https://doi.org/10.1016/j.ijar.2013.11.007, 1205.4174.

He, Y., C. Zhu, Z. He, C. Gu, and J. Cui, 2017a: Big data oriented root cause identification approach based on Axiomatic domain mapping and weighted association rule mining for product infant failure. *Computers and Industrial Engineering*, **109**, 253–265, https://doi.org/10.1016/j.cie.2017.05.012.

He, Y. B., and Z. Geng, 2008: Active learning of causal networks with intervention experiments and optimal designs. *Journal of Machine Learning Research*, **9**, 2523–2547.

He, Z., Y. He, F. Liu, and Y. Zhao, 2019: Big Data-Oriented Product Infant Failure Intelligent Root Cause Identification Using Associated Tree and Fuzzy DEA. *IEEE Access*, **7**, 34 687–34 698, https://doi.org/10.1109/ACCESS.2019.2904759.

He, Z., Y. He, and Y. Wei, 2017b: Big data oriented root cause identification approach based on PCA and SVM for product infant failure. *Proceedings of 2016 Prognostics and System Health Management Conference, PHM-Chengdu 2016*, 1–5, https://doi.org/10.1109/PHM.2016.7819776.

Heckman, J. J., 2008: Econometric causality. *International Statistical Review*, **76 (1)**, 1–27, https://doi.org/10.1111/j.1751-5823.2007.00024.x.

Hernán, M. A., and O. F. Course, 2018: The C-word: scientific euphemisms do not improve causal inference from observational data. *American journal of public health*, **108 (5)**, 616–619, https://doi.org/10.2105/AJPH.2018.304337.

Hernan, M. A., J. Hsu, and B. Healy, 2019: A Second Chance to Get Causal Inference Right: A Classification of Data Science Tasks. *Chance*, **32 (1)**, 42–49, https://doi.org/10.1080/09332480.2019.1579578.

Hernán, M. A., and J. M. Robins, 2018: *Causal Inference*. Boca Raton, FL: Chapman & Hall/CRC.

Hill, J. L., 2011: Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, **20 (1)**, 217–240, https://doi.org/10.1198/jcgs.2010.08162.

Holland, P. W., 1986: Statistics and causal inference. *Journal of the American Statistical Association*, **81 (396)**, 945–960, https://doi.org/10.1080/01621459.1986.10478354.

Hoyer, P. O., D. Janzing, J. Peters, B. Sch, J. M. Mooij, J. Peters, and B. Schölkopf, 2008: Nonlinear causal discovery with additive noise models. *Advances in neural information processing systems*, **21**.

Huang, B., K. Zhang, Y. Lin, B. Schölkopf, and C. Glymour, 2018: Generalized score functions for causal discovery. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1551–1560, https://doi.org/10.1145/3219819.3220104.

Huang, B., K. Zhang, J. Zhang, J. Ramsey, R. Sanchez-Romero, C. Glymour, and B. Schölkopf, 2020a: Causal discovery from heterogeneous/nonstationary data. *Journal of Machine Learning Research*, **21**, 1–53.

Huang, L., A. D. Joseph, B. Nelson, B. I. P. Rubinstein, and J. D. Tygar, 2011: Adversarial machine learning. *Proceedings of the 4th ACM workshop on Security and artificial intelligence*, 43–58.

Huang, X., F. Zhu, L. Holloway, and A. Haidar, 2020b: Causal Discovery from Incomplete Data using An Encoder and Reinforcement Learning. *arXiv preprint*, **(1)**, 2006.05554.

Huang, Y., and M. Valtorta, 2006: Identifiability in causal bayesian networks: A sound and complete algorithm. *Proceedings of the national conference on artificial intelligence*, AAAI, 1149–1154.

Hünermund, P., and E. Bareinboim, 2023: Causal Inference and Data Fusion in Econometrics. *The Econometrics Journal*, 1–62.

Hyttinen, A., P. O. Hoyer, F. Eberhardt, M. Jarvisalo, and J. Matti, 2013: Discovering cyclic causal models with latent variables: A general SAT-based procedure. *Uncertainty in Artificial Intelligence*, 301.

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

Imbens, G., 2014: Instrumental variables: an econometrician's perspective. *Statistical science*, **(29)**, 323–358.

Imbens, G. W., 2004: Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics*, **86 (1)**, 4–29, https://doi.org/10.1162/003465304323023651.

Imbens, G. W., and T. Lemieux, 2008: Regression discontinuity designs: A guide to practice. *Journal of Econometrics*, **142 (2)**, 615–635, https://doi.org/10.1016/j.jeconom.2007.05.001.

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

Ishwaran, H., U. B. Kogalur, E. H. Blackstone, and M. S. Lauer, 2008: Random survival forests. *Annals of Applied Statistics*, **2 (3)**, 841–860, https://doi.org/10.1214/08-AOAS169.

Jacob, D., 2021: CATE meets ML. *Digital Finance*, **3 (2)**, 99–148, https://doi.org/10.1007/s42521-021-00033-7.

Johansson, F. D., U. Shalit, and D. Sontag, 2016: Learning representations for counterfactual inference. *33rd International Conference on Machine Learning, ICML 2016*, **6**, 4407–4418, 1605.03661.

Jung, Y., J. Tian, and E. Bareinboim, 2021: Estimating Identifiable Causal Effects through Double Machine Learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, **35 (13)**, 12 113–12 122, URL https://ojs.aaai.org/index.php/AAAI/article/view/17438.

Kaddour, J., A. Lynch, Q. Liu, M. J. Kusner, and R. Silva, 2022: Causal Machine Learning: A Survey and Open Problems. *arXiv preprint*, **(1)**, 2206.15475.

Kaiser, M., and M. Sipos, 2022: Unsuitability of NOTEARS for Causal Graph Discovery when Dealing with Dimensional Quantities. *Neural Processing Letters*, 1–9.

Kalisch, M., and P. Bühlmann, 2014: Causal structure learning and inference: A selective review. *Quality Technology and Quantitative Management*, **11 (1)**, 3–21, https://doi.org/10.1080/16843703.2014.11673322.

Kalisch, M., M. Mächler, D. Colombo, M. H. Maathuis, and P. Bühlmann, 2012: Causal inference using graphical models with the R package pcalg. *Journal of Statistical Software*, **47 (11)**, https://doi.org/10.18637/jss.v047.i11.

Kartsonaki, C., 2016: Survival analysis. *Diagnostic Histopathology*, **22 (7)**, 263–270, https://doi.org/10.1016/j.mpdhp.2016.06.005.

Kaur, D., S. Uslu, K. J. Rittichier, and A. Durresi, 2022: Trustworthy Artificial Intelligence: A Review. *ACM Computing Surveys (CSUR)*, **55 (2)**, 1–38.

Kennedy, E. H., 2020: Optimal doubly robust estimation of heterogeneous causal effects. *arXiv preprint arXiv:2004.14497*.

Kilbertus, N., M. Rojas Carulla, G. Parascandolo, M. Hardt, D. Janzing, and B. Schölkopf, 2017: Avoiding discrimination through causal reasoning. *Advances in neural information processing systems*, **30**.

Kim, Y., and P. Steiner, 2016: Quasi-Experimental Designs for Causal Inference. *Educational Psychologist*, **51 (3-4)**, 395–405, https://doi.org/10.1080/00461520.2016.1207177.

Kiritoshi, K., T. Izumitani, K. Koyama, T. Okawachi, K. Asahara, and S. Shimizu, 2021: Estimating individual-level optimal causal interventions combining causal models and machine learning models. *The KDD'21 Workshop on Causal Discovery*, PMLR, 55–77.

Kitson, N. K., A. C. Constantinou, Z. Guo, Y. Liu, and K. Chobtham, 2021: A survey of Bayesian Network structure learning. *arXiv preprint arXiv:2109.11415*.

Klein, J. P., and M. L. Moeschberger, 2003: *Survival analysis: Techniques for censored and truncated data*, Vol. 1230. Springer.

Kocaoglu, M., S. Shakkottai, A. G. Dimakis, C. Caramanis, and S. Vishwanath, 2020: Applications of common entropy for causal inference. *Advances in Neural Information Processing Systems*, **December**.

Koch, B., T. Sainburg, P. Geraldo, S. Jiang, Y. Sun, and J. G. Foster, 2021: Deep Learning of Potential Outcomes. *arXiv preprint arXiv:2110.04442*.

Koivisto, M., and K. Sood, 2004: Exact Bayesian structure discovery in Bayesian networks. *The Journal of Machine Learning Research*, **5**, 549–573.

Koller, D., and N. Friedman, 2009: *Probabilistic graphical models: Principles and techniques*. MIT press.

Korb, K. B., and A. E. Nicholson, 2008: The causal interpretation of Bayesian networks. *Innovations in Bayesian Networks*, Springer, 83–116.

Korb, K. B., and A. E. Nicholson, 2010: *Bayesian artificial intelligence*. CRC press.

Kouw, W. M., and M. Loog, 2021: A Review of Domain Adaptation without Target Labels. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **43 (3)**, 766–785, https://doi.org/10.1109/TPAMI.2019.2945942, 1901.05335.

Kubus, M., 2015: Feature Selection and the Chessboard Problem. *Acta Universitatis Lodziensis. Folia Oeconomica*, **1 (311)**, 17–26, https://doi.org/10.18778/0208-6018.311.03.

Kügelgen, J., A. Mey, and M. Loog, 2019: Semi-generative modelling: Covariate-shift adaptation with cause and effect features. *The 22nd International Conference on Artificial Intelligence and Statistics*, PMLR, 1361–1369.

Kulynych, B., 2022: Causal prediction can induce performative stability. *ICML 2022: Workshop on Spurious Correlations, Invariance and Stability*.

Künzel, S. R., J. S. Sekhon, P. J. Bickel, and B. Yu, 2019: Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences of the United States of America*, **116 (10)**, 4156–4165, https://doi.org/10.1073/pnas.1804597116, 1706.03461.

Kusner, M. J., J. Loftus, C. Russell, and R. Silva, 2017: Counterfactual fairness. *Advances in neural information processing systems*, **30**.

Kyono, T., Y. Zhang, and M. van der Schaar, 2020: Castle: Regularization via auxiliary causal graph discovery. *Advances in Neural Information Processing Systems*, **33**, 1501–1512.

Lauritzen, S. L., 1996: *Graphical models*. Clarendon Press.

Lecun, Y., Y. Bengio, and G. Hinton, 2015: Deep learning. *Nature*, **521 (7553)**, 436–444, https://doi.org/10.1038/nature14539.

Lee, D. S., and T. Lemieux, 2010: Regression Discontinuity designs in economics. *Journal of Economic Literature*, **48 (2)**, 281–355, https://doi.org/10.1257/jel.48.2.281.

Lee, M.-j., and C. Kang, 2006: Identification for difference in differences with cross-section and panel data. *Economics Letters*, **92 (2)**, 270–276, https://doi.org/https://doi.org/10.1016/j.econlet.2006.03.007.

Leeb, F., Y. Annadani, S. Bauer, and B. Schölkopf, 2020: Structural Autoencoders Improve Representations for Generation and Transfer.

Lemeshow, S., S. May, and D. W. Hosmer Jr, 2011: *Applied survival analysis: Regression modeling of time-to-event data*. John Wiley & Sons.

Lemmon, J., and Coauthors, 2023: Evaluation of Feature Selection Methods for Preserving Machine Learning Performance in the Presence of Temporal Dataset Shift in Clinical Medicine. *Methods of Information in Medicine*, **62 (01/02)**, 60–70, https://doi.org/10.1055/s-0043-1762904.

Leslie, D., 2020: Understanding Bias in Facial Recognition Technologies. *SSRN Electronic Journal*, https://doi.org/10.2139/ssrn.3705658.

Li, G., S. J. Qin, and T. Yuan, 2016: Data-driven root cause diagnosis of faults in process industries. *Chemometrics and Intelligent Laboratory Systems*, **159**, 1–11.

Li, J., L. Liu, T. D. Le, and J. Liu, 2020: Accurate data-driven prediction does not mean high reproducibility. *Nature Machine Intelligence*, **2 (1)**, 13–15, https://doi.org/10.1038/s42256-019-0140-2.

Li, J., S. Ma, T. Le, L. Liu, and J. Liu, 2017: Causal Decision Trees. *IEEE Transactions on Knowledge and Data Engineering*, **29 (2)**, 257–271, https://doi.org/10.1109/TKDE.2016.2619350, 1508.03812.

Li, Y., J. Hestness, M. Elhoseiny, L. Zhao, and K. Church, 2022: Efficiently Disentangle Causal Representations. *arXiv preprint arXiv:2201.01942*, 1–17, 2201.01942.

Lin, F., K. Muzumdar, N. P. Laptev, M.-V. Curelea, S. Lee, and S. Sankar, 2020: Fast Dimensional Analysis for Root Cause Investigation in a Large-Scale Service Environment. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, **4 (2)**, 1–23, https://doi.org/10.1145/3392149.

Linden, A., and P. R. Yarnold, 2016: Using machine learning to assess covariate balance in matching studies. *Journal of Evaluation in Clinical Practice*, **22 (6)**, 848–854.

Liu, C., K. G. Lore, and S. Sarkar, 2018: Data-driven root-cause analysis for distributed system anomalies. *2017 IEEE 56th Annual Conference on Decision and Control, CDC 2017*, **January (Cdc)**, 5745–5750, https://doi.org/10.1109/CDC.2017.8264527.

Liu, H., and Coauthors, 2021: Trustworthy AI: A computational perspective. *arXiv preprint arXiv:2107.06641*, **1 (1)**, 1–55, arXiv:2107.06641v3.

Liu, Y., H.-S. Chen, H. Wu, Y. Dai, Y. Yao, and Z. Yan, 2020: Simplified Granger causality map for data-driven root cause diagnosis of process disturbances. *Journal of Process Control*, **95**, 45–54.

Loftus, J. R., C. Russell, M. J. Kusner, and R. Silva, 2018: Causal reasoning for algorithmic fairness. *arXiv preprint arXiv:1805.05859*, 1–21.

Lopez-Paz, D., R. Nishihara, S. Chintala, B. Schölkopf, and L. Bottou, 2017: Discovering causal signals in images. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, **January**, 58–66, https://doi.org/10.1109/CVPR.2017.14.

Louizos, C., U. Shalit, J. M. Mooij, D. Sontag, R. Zemel, and M. Welling, 2017: Causal effect inference with deep latent-variable models. *Advances in neural information processing systems*, **30**.

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

Ma, Q., H. Li, and A. Thorstenson, 2021: A big data-driven root cause analysis system: Application of Machine Learning in quality problem solving. *Computers and Industrial Engineering*, **160 (November)**, 107 580, https://doi.org/10.1016/j.cie.2021.107580.

Ma, S., P. Kemmeren, C. F. Aliferis, and A. Statnikov, 2016: An Evaluation of Active Learning Causal Discovery Methods for Reverse-Engineering Local Causal Pathways of Gene Regulation. *Scientific Reports*, **6 (February)**, 1–14, https://doi.org/10.1038/srep22558.

Ma, S., and A. Statnikov, 2017: Methods for computational causal discovery in biomedicine. *Behaviormetrika*, **44 (1)**, 165–191, https://doi.org/10.1007/s41237-016-0013-5.

Maathuis, M. H., and P. Nandy, 2016: A review of some recent advances in causal inference. *Handbook of Big Data*, 387–408., https://doi.org/10.1201/b19567-32.

Magliacane, S., T. Van Ommen, T. Claassen, S. Bongers, J. M. Mooij, and P. Versteeg, 2018: Domain adaptation by using causal inference to predict invariant conditional distributions. *Advances in Neural Information Processing Systems*, **December (NeurIPS)**, 10 846–10 856.

Maguire, A., I. Douglas, L. Smeeth, and M. Thompson, 2007: Determinants of cholesterol and triglycerides recording in patients treated with lipid lowering therapy in UK primary care. *Pharmacoepidemiology and drug safety*, **16 (March)**, 228–228, https://doi.org/10.1002/pds.

Makhlouf, K., S. Zhioua, and C. Palamidessi, 2020: Survey on causal-based machine learning fairness notions. *arXiv preprint arXiv:2010.09553*.

Malinsky, D., and D. Danks, 2018: Causal discovery algorithms: A practical guide. *Philosophy Compass*, **13 (1)**, 1–11, https://doi.org/10.1111/phc3.12470.

Mansournia, M. A., and D. G. Altman, 2016: Inverse probability weighting. *BMJ*, **352 (January)**, 1–2, https://doi.org/10.1136/bmj.i189.

Martens, E. P., W. R. Pestman, A. de Boer, S. V. Belitser, and O. H. Klungel, 2006: Instrumental variables: application and limitations. *Epidemiology*, 260–267.

Martin, M., 2019: An Overview of the pcalg Package for R. 1–47.

Marx, A., A. Gretton, and J. M. Mooij, 2021: A weaker faithfulness assumption based on triple interactions. *Uncertainty in Artificial Intelligence*, PMLR, 451–460.

Marx, A., B. Rihoux, C. Ragin, and T. C. Method, 2014: The origins, development, and application of Qualitative Comparative Analysis: the first 25 years. *European Political Science Review*, **6 (1)**, 115–142, https://doi.org/10.1017/S1755773912000318.

Masegosa, A. R., and S. Moral, 2013: An interactive approach for Bayesian network learning using domain/expert knowledge. *International Journal of Approximate Reasoning*, **54 (8)**, 1168–1181.

Maziarz, M., 2015: A review of the Granger-causality fallacy. *The Journal of Philosophical Economics: Reflections on economic and social issues*, **8 (2)**, 86–105.

Mbogu, H., and C. Nicholson, 2023: Data-Driven Root Cause Analysis Via Causal Discovery Using Time-To-Event Data. URL https://ssrn.com/abstract=4414871.

Mehrabi, N., F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, 2021: A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, **54 (6)**, 1–35.

Meyer, B. D., 1995: Natural and quasi-experiments in economics. *Journal of Business and Economic Statistics*, **13 (2)**, 151–161, https://doi.org/10.1080/07350015.1995.10524589.

Miao, J., and L. Niu, 2016: A Survey on Feature Selection. *Procedia Computer Science*, **91 (Itqm)**, 919–926, https://doi.org/10.1016/j.procs.2016.07.111.

Miller, T., 2019: Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, **267**, 1–38, https://doi.org/10.1016/j.artint.2018.07.007.

Molnar, C., 2019: Interpretable machine learning: a guide for making black box models explainable. *[online] Available:*, URL https://christophm.github.io/interpretable-ml-book/.

Molnar, C., G. Casalicchio, and B. Bischl, 2020: Interpretable machine learning–a brief history, state-of-the-art and challenges. *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, Springer, 417–431.

Mooij, J. M., S. Magliacane, and T. Claassen, 2020: Joint causal inference from multiple contexts. *Journal of Machine Learning Research*, **21 (1)**, 3919–4026.

Mooij, J. M., J. Peters, D. Janzing, J. Zscheischler, and B. Schölkopf, 2016: Distinguishing cause from effect using observational data: Methods and benchmarks. *Journal of Machine Learning Research*, **17**, 1–102, 1412.3773.

Moreno-Torres, J. G., T. Raeder, R. Alaiz-Rodríguez, N. V. Chawla, and F. Herrera, 2012: A unifying view on dataset shift in classification. *Pattern Recognition*, **45 (1)**, 521–530, https://doi.org/10.1016/j.patcog.2011.06.019.

Morgan, S. L., and C. Winship, 2015: *Counterfactuals and causal inference*. Cambridge University Press.

Naimi, A. I., S. R. Cole, and E. H. Kennedy, 2017: An introduction to g methods. *International journal of epidemiology*, **46 (2)**, 756–762.

Naimi, A. I., A. E. Mishler, and E. H. Kennedy, 2021: Challenges in Obtaining Valid Causal Effect Estimates with Machine Learning Algorithms. *American Journal of Epidemiology*, **(Ml)**, https://doi.org/10.1093/aje/kwab201, 1711.07137.

Nandy, P., A. Hauser, M. H. Maathuis, and S. N. S. F. Grant, 2018: High-dimensional consistency in score-based and hybrid structure learning. *The Annals of Statistics*, **46 (6A)**, 3151–3183.

Narendra, T., P. Agarwal, M. Gupta, and S. Dechu, 2019: Counterfactual reasoning for process optimization using structural causal models. *International Conference on Business Process Management*, Springer, 91–106.

Neapolitan, R. E., 2004: *Learning bayesian networks*, Vol. 38. Pearson Prentice Hall, Upper Saddle River.

Ng, G. W., and W. C. Leung, 2020: Strong artificial intelligence and consciousness. *Journal of Artificial Intelligence and Consciousness*, **7 (01)**, 63–72.

Ng, I., S. Lachapelle, N. R. Ke, and S. Lacoste-Julien, 2022: On the convergence of continuous constrained optimization for structure learning. *International Conference on Artificial Intelligence and Statistics*, PMLR.

Nguyen, B. H., B. Xue, and P. Andreae, 2018: A particle swarm optimization based feature selection approach to transfer learning in classification. *GECCO 2018 - Proceedings of the 2018 Genetic and Evolutionary Computation Conference*, 37–44., https://doi.org/10.1145/3205455.3205540.

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

Nogueira, A. R., A. Pugnana, S. Ruggieri, D. Pedreschi, and J. Gama, 2022: Methods and tools for causal discovery and causal inference. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, **12 (2)**, 1–39, https://doi.org/10.1002/widm.1449.

Ogarrio, J. M., P. Spirtes, J. R. B. T. P. o. t. E. I. C. o. P. G. Models, and Diering, 2016: A Hybrid Causal Search Algorithm for Latent Variable Models. *Physiology & behavior*, **176 (1)**, 368–379, https://doi.org/10.1016/j.physbeh.2017.03.040.

Pan, S. J., I. W. Tsang, J. T. Kwok, and Q. Yang, 2011: Domain adaptation via transfer component analysis. *IEEE Transactions on Neural Networks*, **22 (2)**, 199–210, https://doi.org/10.1109/TNN.2010.2091281.

Pan, S. J., and Q. Yang, 2009: A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, **22 (10)**, 1345–1359.

Pan, Y., and L. Qiu, 2021: How Ride-Sharing Is Shaping Public Transit System: A Counterfactual Estimator Approach. *Production and Operations Management*, **0 (0)**, 1–22, https://doi.org/10.1111/poms.13582.

Pattanayak, C. W., D. B. Rubin, and E. R. Zell, 2011: Propensity score methods for creating covariate balance in observational studies. *Revista Española de Cardiología (English Edition)*, **64 (10)**, 897–903.

Pawlowski, N., D. Coelho de Castro, and B. Glocker, 2020: Deep structural causal models for tractable counterfactual inference. *Advances in Neural Information Processing Systems*, **33**, 857–869.

Pearl, J., 1995: Causal diagrams for empirical research. *Biometrika*, **82 (4)**, 669–688, https://doi.org/10.1093/biomet/82.4.669.

Pearl, J., 1996: Structural and probabilistic causality. *Psychology of learning and motivation*, **34**, 393–435.

Pearl, J., 2000: Models, reasoning and inference. *Cambridge, UK: CambridgeUniversityPress*, **19**.

Pearl, J., 2009a: Causal inference in statistics: An overview. *Statistics Surveys*, **3 (September)**, 96–146, https://doi.org/10.1214/09-SS057.

Pearl, J., 2009b: *Causality: Models, reasoning and inference*. Cambridge University Press, Cambridge, https://doi.org/10.1017/CBO9780511803161.

Pearl, J., 2012: The do-calculus revisited. *Uncertainty in Artificial Intelligence - Proceedings of the 28th Conference, UAI 2012*, 4–11, 1210.4852.

Pearl, J., 2013: The mathematics of causal inference. *Joint Statistical Meetings Proceedings*, American Statistical Association.

Pearl, J., 2014: *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Elsevier.

Pearl, J., 2018: Theoretical Impediments to Machine Learning With Seven Sparks from the Causal Revolution. *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, 3, https://doi.org/10.1145/3159652.3176182.

Pearl, J., 2019a: The limitations of opaque learning machines. *Possible minds: twenty-five ways of looking at AI*, 13–19.

Pearl, J., 2019b: The seven tools of causal inference, with reflections on machine learning. *Communications of the ACM*, **62 (3)**, 54–60, https://doi.org/10.1145/3241036.

Pearl, J., M. Glymour, and N. P. Jewell, 2016: *Causal inference in statistics: A primer*. John Wiley & Sons.

Pearl, J., and D. Mackenzie, 2018: *The book of why: the new science of cause and effect*. Basic books.

Pellet, J. P., and A. Elisseeff, 2008: Using Markov blankets for causal structure learning. *Journal of Machine Learning Research*, **9**, 1295–1342.

Perković, E., M. Kalisch, and M. H. Maathuis, 2017: Interpreting and using CPDAGs with background knowledge. *Proceedings of the 2017 Conference on Uncertainty in Artificial Intelligence (UAI2017)*, AUAI Press.

Polo, F. M., R. Izbicki, E. G. Lacerda Jr, J. P. Ibieta-Jimenez, and R. Vicente, 2023: A unified framework for dataset shift diagnostics. *Information Sciences*, 119612.

Powers, S., J. Qian, K. Jung, A. Schuler, N. H. Shah, T. Hastie, and R. Tibshirani, 2018: Some methods for heterogeneous treatment effect estimation in high dimensions. *Statistics in Medicine*, **37 (11)**, 1767–1787, https://doi.org/10.1002/sim.7623, 1707.00102.

Raghu, V. K., A. Poon, and P. V. Benos, 2018: Evaluation of Causal Structure Learning Methods on Mixed Data Types. *Proceedings of machine learning research*, **92**, 48–65.

Rahmani, K., R. Thapa, P. Tsou, S. Casie Chetty, G. Barnes, C. Lam, and C. Foon Tso, 2023: Assessing the effects of data drift on the performance of machine learning models used in clinical sepsis prediction. *International Journal of Medical Informatics*, **173**, 1–27, https://doi.org/10.1016/j.ijmedinf.2022.104930.

Ramsey, J., P. L. Spirtes, J. Zhang, and P. L. Spirtes, 2012: Adjacency-faithfulness and conservative causal inference. *Proc. Conf. on Uncertainty in Artificial Intelligence*, 401–408, 1206.6843.

Ramsey, J. D., 2014: A Scalable Conditional Independence Test for Nonlinear, Non-Gaussian Data. 1–16, URL http://arxiv.org/abs/1401.5031, 1401.5031.

Rebonato, R., 2016: Mostly Harmless Econometrics: An Empiricist's Companion; Mastering 'Metrics: The Path from Cause to Effect. *Quantitative Finance*, **16 (7)**, 1009–1013, https://doi.org/10.1080/14697688.2015.1080490.

Reisach, A. G., C. Seiler, and S. Weichwald, 2021: Beware of the simulated dag! causal discovery benchmarks may be easy to game. *Advances in Neural Information Processing Systems*, **34**, 27 772–27 784.

Remler, D. K., and G. G. Van Ryzin, 2021: *Research methods in practice: Strategies for description and causation*. Sage Publications.

Richardson, T., and P. Spirtes, 2002: Ancestral graph Markov models. *Annals of Statistics*, **30 (4)**, 962–1030, https://doi.org/10.1214/aos/1031689015.

Robins, J., 1986: A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical modelling*, **7 (9-12)**, 1393–1512.

Rocha, E. M., A. F. Brochado, B. Rato, and J. Meneses, 2022: Benchmarking and Prediction of Entities Performance on Manufacturing Processes through MEA, Robust XGBoost and SHAP Analysis. *IEEE International Conference on Emerging Technologies and Factory Automation, ETFA*, **2022-Septe**, https://doi.org/10.1109/ETFA52439.2022.9921593.

Rojas-Carulla, M., B. Schölkopf, R. Turner, and J. Peters, 2018: Invariant models for causal transfer learning. *Journal of Machine Learning Research*, **19**, 1–34, 1507.05333.

Rolling, C. A., and Y. Yang, 2014: Model selection for estimating treatment effects. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **76 (4)**, 749–769.

Rooney, J. J., and L. N. V. Heuvel, 2004: Root cause analysis for beginners. *Quality progress*, **37 (7)**, 45–56.

Rosen, D. A., and C. Press, 1978: In defense of a probabilistic theory of causality. *Philosophy of Science*, **45 (4)**, 604–613.

Rosenbaum, P. R., 2020: Modern algorithms for matching in observational studies. *Annual Review of Statistics and Its Application*, **7**, 143–176.

Rosenbaum, P. R., and D. B. Rubin, 1984: Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American statistical Association*, **79 (387)**, 516–524.

Rosenberger, W. F., and J. M. Lachin, 2015: *Randomization in clinical trials: theory and practice*. John Wiley & Sons.

Rothman, K. J., and S. Greenland, 2005: Causation and causal inference in epidemiology. *American Journal of Public Health*, **95 (SUPPL. 1)**, https://doi.org/10.2105/AJPH.2004.059204.

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

Rubin, D. B., 2005: Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, **100 (469)**, 322–331, https://doi.org/10.1198/016214504000001880.

Ryall, M. D., and A. Bramson, 2013: *Inference and intervention: Causal models for business analysis*. Routledge.

Saeed, W., and C. Omlin, 2023: Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities. *Knowledge-Based Systems*, **263**, 110 273, https://doi.org/10.1016/j.knosys.2023.110273.

Saikia, R., and M. P. Barman, 2017: A review on accelerated failure time models. *International Journal of Statistics and Systems*, **12 (2)**, 311–322.

Samantha, R., D. Almalik, T. Mueller, J. Greipel, T. Weber, and R. H. Schmitt, 2018: Automated root cause analysis of non-conformities with machine learning algorithms. *Journal of Machine Engineering*, **18 (2)**, 58–66, URL http://www.tjyybjb.ac.cn/CN/article/downloadArticleFile.do?attachType=PDF{\&}id=9987.

Samek, W., and K.-R. Müller, 2019: Towards explainable artificial intelligence. *Explainable AI: interpreting, explaining and visualizing deep learning*, Springer, 5–22.

Sanodiya, R. K., M. Tiwari, J. Mathew, S. Saha, and S. Saha, 2020: A particle swarm optimization-based feature selection for unsupervised transfer learning. *Soft Computing*, **24 (24)**, 18 713–18 731, https://doi.org/10.1007/s00500-020-05105-1.

Scheines, R., and M. E. Sobel, 1997: An introduction to causal inference. *Sociological Methods & Research*, **24 (3)**, 353–379, https://doi.org/10.1177/0049124196024003004.

Schochet, P. Z., 2010: Is regression adjustment supported by the Neyman model for causal inference? *Journal of Statistical Planning and Inference*, **140 (1)**, 246–259.

Schölkopf, B., 2022a: Causality for Machine Learning. *Probabilistic and Causal Inference: The Works of Judea Pearl*, 765–804., 1911.10500.

Schölkopf, B., 2022b: Causality for Machine Learning. *Probabilistic and Causal Inference*, 765–804., https://doi.org/10.1145/3501714.3501755.

Scholkopf, B., F. Locatello, S. Bauer, N. R. Ke, N. Kalchbrenner, A. Goyal, and Y. Bengio, 2021: Toward Causal Representation Learning. *Proceedings of the IEEE*, **109 (5)**, 612–634, https://doi.org/10.1109/JPROC.2021.3058954.

Schuler, M. S., and S. Rose, 2017: Practice of Epidemiology Targeted Maximum Likelihood Estimation for Causal Inference in Observational Studies. **185 (1)**, 65–73, https://doi.org/10.1093/aje/kww165.

Sejnowski, T. J., 2020: The unreasonable effectiveness of deep learning in artificial intelligence. *Proceedings of the National Academy of Sciences*, **117 (48)**, 30 033–30 038, https://doi.org/10.1073/pnas.1907373117.

Shakya, A., V. Rus, and D. Venugopal, 2021: Student Strategy Prediction Using a Neuro-Symbolic Approach. *International Educational Data Mining Society*, **(Edm)**, 118–129.

Shalit, U., F. D. Johansson, and D. Sontag, 2017: Estimating individual treatment effect: generalization bounds and algorithms. *International Conference on Machine Learning*, PMLR, 3076–3085.

Shen, J., Y. Qu, W. Zhang, and Y. Yu, 2018: Wasserstein distance guided representation learning for domain adaptation. *32nd AAAI Conference on Artificial Intelligence, AAAI 2018*, 4058–4065, https://doi.org/10.1609/aaai.v32i1.11784.

Shi, C., D. Blei, and V. Veitch, 2019: Adapting neural networks for the estimation of treatment effects. *Advances in neural information processing systems*, **32**.

Shimizu, S., P. O. Hoyer, A. Hyvärinen, and A. Kerminen, 2006: A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, **7**, 2003–2030.

Shoush, M., and M. Dumas, 2021: Prescriptive Process Monitoring Under Resource Constraints: A Causal Inference Approach. *International Conference on Process Mining*, Cham: Springer International Publishing.

Shpitser, I., and J. Pearl, 2008: Complete identification methods for the causal hierarchy. *Journal of Machine Learning Research*, **9**, 1941–1979.

Shpitser, I., T. VanderWeele, and J. M. Robins, 2012: On the validity of covariate adjustment for estimating causal effects. *Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence*, AUAI Press.

Shrier, I., and R. W. Platt, 2008: Reducing bias through directed acyclic graphs. *BMC Medical Research Methodology*, **8**, 1–15, https://doi.org/10.1186/1471-2288-8-70.

Siau, K., and W. Wang, 2020: Artificial intelligence (AI) ethics: ethics of AI and ethical AI. *Journal of Database Management (JDM)*, **31 (2)**, 74–87.

Silander, T., P. Myllymaki, P. Myllym, and P. Myllymäki, 2006: A simple approach for finding the globally optimal Bayesian network structure. *Conference on Uncertainty in Artificial Intelligence*, 445–452.

Simard, R., and P. L'Ecuyer, 2011: Computing the two-sided Kolmogorov-Smirnov distribution. *Journal of Statistical Software*, **39**, 1–18.

Singh, K., G. Gupta, V. Tewari, and G. Shroff, 2018: Comparative benchmarking of causal discovery algorithms. *ACM International Conference Proceeding Series*, 46–56, https://doi.org/10.1145/3152494.3152499.

Smith, P. J., 2017: *Analysis of failure and survival data*. Chapman and Hall/CRC.

Snowden, J. M., S. Rose, and K. M. Mortimer, 2011: Implementation of G-computation on a simulated data set: demonstration of a causal inference technique. *American journal of epidemiology*, **173 (7)**, 731–738.

Sofrygin, O., R. Neugebauer, and M. J. van der Laan, 2017: Conducting Simulations in Causal Inference with Networks-Based Structural Equation Models. *arXiv preprint*, 1–20, 1705.10376.

Spirtes, P., C. Glymour, R. Scheines, S. N., and Richard, 2000: *Causation, Prediction, and Search*, Vol. 39. Mit Press: Cambridge, 137–140 pp.

Spirtes, P., and K. Zhang, 2016: Causal discovery and inference: Concepts and recent methodological advances. *Applied Informatics*, **3 (1)**, https://doi.org/10.1186/s40535-016-0018-x.

Spreeuwenberg, M. D., 2010: *Selection Bias in (Quasi-) Experimental Research*. Ridderprint.

Steiner, P. M., and D. Cook, 2013: Matching and Propensity Scores 13. *The Oxford Handbook of Quantitative Methods in Psychology, Vol. 1*, **1**, 237.

Stern, D. I., 2011: From correlation to Granger causality. *Crawford School Research Paper.*, **(13)**.

Stevens, K., 2016: Regression discontinuity designs: an introduction. *Australian Economic Review*, **49 (2)**, 224–233.

Subbaswamy, A., B. Chen, and S. Saria, 2022: A unifying causal framework for analyzing dataset shift-stable learning algorithms. *Journal of Causal Inference*, **10 (1)**, 64–89, https://doi.org/10.1515/jci-2021-0042.

Subbaswamy, A., P. Schulam, and S. Saria, 2019: Preventing failures due to dataset shift: Learning predictive models that transport. *The 22nd International Conference on Artificial Intelligence and Statistics*, PMLR, 3118–3127.

Sugiyama, M., and M. Kawanabe, 2012: *Machine learning in non-stationary environments: Introduction to covariate shift adaptation*. MIT press, https://doi.org/10.7551/mitpress/9780262017091.001.0001.

Sun, B., J. Feng, and K. Saenko, 2016: Return of frustratingly easy domain adaptation. *30th AAAI Conference on Artificial Intelligence, AAAI 2016*, **(Figure 1)**, 2058–2065, https://doi.org/10.1609/aaai.v30i1.10306, 1511.05547.

Sun, F., H. Wu, Z. Luo, W. Gu, Y. Yan, and Q. Du, 2019: Informative Feature Selection for Domain Adaptation. *IEEE Access*, **7**, 142 551–142 563, https://doi.org/10.1109/ACCESS.2019.2944226.

Tafti, A., and G. Shmueli, 2020: Beyond overall treatment effects: Leveraging covariates in randomized experiments guided by causal structure. *Information Systems Research*, **31 (4)**, 1183–1199, https://doi.org/10.1287/isre.2020.0938.

Thakar, S., and D. Kalbande, 2023: A Pipeline for Business Intelligence and Data-Driven Root Cause Analysis on Categorical Data. *Proceedings of Third International Conference on Sustainable Expert Systems: ICSES 2022*, Springer, 389–398.

Thoemmes, F., and A. D. Ong, 2016: A primer on inverse probability of treatment weighting and marginal structural models. *Emerging Adulthood*, **4 (1)**, 40–59.

Thygesen, L. C., G. S. Andersen, and H. Andersen, 2005: A philosophical analysis of the Hill criteria. *Journal of Epidemiology and Community Health*, **59 (6)**, 512–516, https://doi.org/10.1136/jech.2004.027524.

Tian, J., and J. Pearl, 2001: Causal Discovery from Changes. *Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence*, 512–521, 1301.2312.

Tian, J., and J. Pearl, 2002: A general identification condition for causal effects. *Proceedings of the National Conference on Artificial Intelligence*, 567–573.

Tong, S., and D. Koller, 2001: Active learning for structure in Bayesian networks. *IJCAI International Joint Conference on Artificial Intelligence*, 863–869.

Triantafillou, S., and I. Tsamardinos, 2015: Constraint-based causal discovery from multiple interventions over overlapping variable sets. *The Journal of Machine Learning Research*, **16 (1)**, 2147–2205.

Triantafillou, S., and I. Tsamardinos, 2016: Score based vs constraint based causal learning in the presence of confounders. *CEUR Workshop Proceedings*, **1792**, 59–67.

Tsamardinos, I., C. Aliferis, A. Statnikov, and E. Statnikov, 2003: Algorithms for Large Scale Markov Blanket Discovery. *The 16th international FLAIRS conference*, **(i)**, 376–381, 1855670437.

Tucker, B. P., C. Babbage, and P. Tucker, 2008: The AI Chasers. *The Futurist*, **42 (2)**, 14.

Uguroglu, S., and J. Carbonell, 2011: Feature selection for transfer learning. *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, Springer, Vol. 6913 LNAI, 430–442, https://doi.org/10.1007/978-3-642-23808-628.

Urminsky, O., C. Hansen, and V. Chernozhukov, 2016: Using double-lasso regression for principled variable selection. *Available at SSRN 2733374*.

Van Der Laan, M. J., 2010: Targeted maximum likelihood based causal inference: Part I. *International Journal of Biostatistics*, **6 (2)**, https://doi.org/10.2202/1557-4679.1211.

Van der Laan, M. J., E. C. Polley, and A. E. Hubbard, 2007: Super learner. *Statistical applications in genetics and molecular biology*, **6 (1)**.

VanderWeele, T. J., 2019: Principles of confounder selection. *European Journal of Epidemiology*, **34 (3)**, 211–219, https://doi.org/10.1007/s10654-019-00494-6.

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

Vilone, G., and L. Longo, 2020: Explainable artificial intelligence: a systematic review. *arXiv preprint arXiv:2006.00093*, **(Dl)**, arXiv:2006.00093v4.

Vittinghoff, E., D. V. Glidden, S. C. Shiboski, and C. E. McCulloch, 2006: *Regression methods in biostatistics: Linear, logistic, survival, and repeated measures models*. Springer.

Vonk, M. C., N. Malekovic, T. Bäck, and A. V. Kononova, 2023: *Disentangling causality: Assumptions in causal discovery and inference*. Springer Netherlands, 10613–10649 pp., https://doi.org/10.1007/s10462-023-10411-9.

Vonrueden, L., and Coauthors, 2021: Informed Machine Learning - A Taxonomy and Survey of Integrating Prior Knowledge into Learning Systems. *IEEE Transactions on Knowledge and Data Engineering*, 1–20, https://doi.org/10.1109/TKDE.2021.3079836, 1903.12394.

Vowels, M. J., N. C. Camgoz, and R. Bowden, 2022: D'ya like DAGs? A Survey on Structure Learning and Causal Discovery. *ACM Computing Surveys*, **55 (4)**, 1–36, 2103.02582.

Vuković, M., and S. Thalmann, 2022: Causal Discovery in Manufacturing: A Structured Literature Review. *Journal of Manufacturing and Materials Processing*, **6 (1)**, https://doi.org/10.3390/JMMP6010010.

Wager, S., and S. Athey, 2018: Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. *Journal of the American Statistical Association*, **113 (523)**, 1228–1242, https://doi.org/10.1080/01621459.2017.1319839, 1510.04342.

Wang, A., and O. A. Arah, 2015: G-computation demonstration in causal mediation analysis. *European journal of epidemiology*, **30 (10)**, 1119–1127.

Wang, S., Q. Zhao, Y. Han, and J. Wang, 2023: Root cause diagnosis for complex industrial process faults via spatiotemporal coalescent based time series prediction and optimized Granger causality. *Chemometrics and Intelligent Laboratory Systems*, **233**, 104 728.

Wei, D., T. Gao, and Y. Yu, 2020: DAGs with No Fears: A closer look at continuous optimization for learning Bayesian networks. *Advances in Neural Information Processing Systems*, **33 (NeurIPS)**, 3895–3906.

Weilenmann, M., and R. Colbeck, 2017: Analysing causal structures with entropy. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, **473 (2207)**, 20170 483.

Weinberger, N., 2018: Faithfulness, Coordination and Causal Coincidences. *Erkenntnis*, **83 (2)**, 113–133, https://doi.org/10.1007/s10670-017-9882-6.

Weiss, K., T. M. Khoshgoftaar, and D. D. Wang, 2016: *A survey of transfer learning*, Vol. 3. Springer International Publishing, https://doi.org/10.1186/s40537-016-0043-6.

Westreich, D., J. Lessler, and M. J. Funk, 2010: Propensity score estimation: neural networks, support vector machines, decision trees (CART), and meta-classifiers as alternatives to logistic regression. *Journal of Clinical Epidemiology*, **63 (8)**, 826–833, https://doi.org/10.1016/j.jclinepi.2009.11.020.

White, H., and S. Sabarwal, 2014: Quasi-experimental design and methods. *Methodological briefs: impact evaluation*, **8 (2014)**, 1–16.

Witlox, C., and A. Naghi, 2018: The Empirical Validation of Double/Debiased Machine Learning. *Thesis[Erasmus University Rotterdam]*.

Wooldridge, J. M., 2015: *Introductory econometrics: A modern approach*. Cengage learning.

Wu, Y., L. Zhang, X. Wu, and H. Tong, 2019: Pc-fairness: A unified framework for measuring causality-based fairness. *Advances in Neural Information Processing Systems*, **32**.

Xiang, J., and S. Kim, 2013: A* Lasso for learning a sparse Bayesian network structure for continuous variables. *Advances in neural information processing systems*, **26**, 1–9.

Xu, G., T. D. Duong, Q. Li, S. Liu, and X. Wang, 2020: Causality Learning: A New Perspective for Interpretable Machine Learning. *IEEE Intelligent Informatics Bulletin*, 2006.16789.

Xu, J., Z. Zhang, T. Friedman, Y. Liang, and G. V. D. Broeck, 2018: A semantic loss function for deep learning with symbolic knowledge. *International conference on machine learning*, PMLR, 5502–5511.

Xu, R., P. Cui, Z. Shen, X. Zhang, and T. Zhang, 2021: Why Stable Learning Works? A Theory of Covariate Shift Generalization. *arXiv preprint*, 1–25, 2111.02355v1.

Yan, Y., H. Wu, Y. Ye, C. Bi, M. Lu, D. Liu, Q. Wu, and M. K. Ng, 2022: Transferable Feature Selection for Unsupervised Domain Adaptation. *IEEE Transactions on Knowledge and Data Engineering*, **34 (11)**, 5536–5551, https://doi.org/10.1109/TKDE.2021.3060037.

Yang, S., K. Yu, F. Cao, L. Liu, H. Wang, and J. Li, 2021: Learning Causal Representations for Robust Domain Adaptation. *IEEE Transactions on Knowledge and Data Engineering*, **4347 (c)**, https://doi.org/10.1109/TKDE.2021.3119185.

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

Yoon, J., J. Jordon, and M. Van Der Schaar, 2018: GANITE: Estimation of individualized treatment effects using generative adversarial nets. *International Conference on Learning Representations*.

Yu, C. N., R. Greiner, H. C. Lin, and V. Baracos, 2011: Learning patient-specific cancer survival distributions as a sequence of dependent regressors. *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011*, 1–9.

Yu, K., X. Guo, L. Liu, J. Li, H. Wang, Z. Ling, and X. Wu, 2020: Causality-based feature selection: Methods and evaluations. *ACM Computing Surveys (CSUR)*, **53 (5)**, 1–36.

Yu, K., J. Li, and L. Liu, 2016: A review on algorithms for constraint-based causal discovery. *arXiv preprint arXiv:1611.03977*, 1–17, 1611.03977.

Yu, K., L. Liu, and J. Li, 2021: A Unified View of Causal and Non-causal Feature Selection. *ACM Transactions on Knowledge Discovery from Data*, **15 (4)**, https://doi.org/10.1145/3436891.

Yuan, C., and B. Malone, 2013: Learning optimal Bayesian networks: A shortest path perspective. *Journal of Artificial Intelligence Research*, **48**, 23–65.

Yuniarto, H. A., 2012: The shortcomings of existing root cause analysis tools. *Lecture Notes in Engineering and Computer Science*, **3**, 1549–1552.

Zanga, A., E. Ozkirimli, and F. Stella, 2022: A Survey on Causal Discovery: Theory and Practice. *International Journal of Approximate Reasoning*, **151**, 101–129, https://doi.org/10.1016/j.ijar.2022.09.004.

Zhang, H., K. Peng, and L. Ma, 2023: A systematic nonstationary causality analysis framework for root cause diagnosis of faults in manufacturing processes. *Control Engineering Practice*, **131**, 105 404.

Zhang, J., 2008: Causal reasoning with ancestral graphs. *Journal of Machine Learning Research*, **9**, 1437–1474.

Zhang, J., and E. Bareinboim, 2018: Equality of opportunity in classification: A causal approach. *Advances in Neural Information Processing Systems*, **31 (NeurIPS)**, 1–11.

Zhang, K., M. Gong, and B. Scholkopf, 2015a: Multi-source domain adaptation: A causal view. *Proceedings of the National Conference on Artificial Intelligence*, **4**, 3150–3157, https://doi.org/10.1609/aaai.v29i1.9542.

Zhang, K., M. Gong, P. Stojanov, B. Huang, Q. Liu, and C. Glymour, 2020: Domain adaptation as a problem of inference on graphical models. *Advances in Neural Information Processing Systems*, **(NeurIPS)**, 1–12.

Zhang, K., and A. Hyvärinen, 2009: On the identifiability of the post-nonlinear causal model. *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence, UAI 2009*, 647–655.

Zhang, K., and A. Hyvärinen, 2010: Distinguishing causes from effects using nonlinear acyclic causal models. *Causality: Objectives and Assessment*, PMLR, 157–164.

Zhang, K., Z. Wang, J. Zhang, and B. Schölkopf, 2015b: On estimation of functional causal models: General results and application to the post-nonlinear causal model. *ACM Transactions on Intelligent Systems and Technology*, **7 (2)**, 1–22, https://doi.org/10.1145/2700476.

Zhang, L., Y. Wu, and X. Wu, 2016: On discrimination discovery using causal networks. *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation*, Springer, 83–93.

Zhang, L., Y. Wu, and X. Wu, 2017: A causal framework for discovering and removing direct and indirect discrimination. *IJCAI International Joint Conference on Artificial Intelligence*, **0**, 3929–3935, https://doi.org/10.24963/ijcai.2017/549.

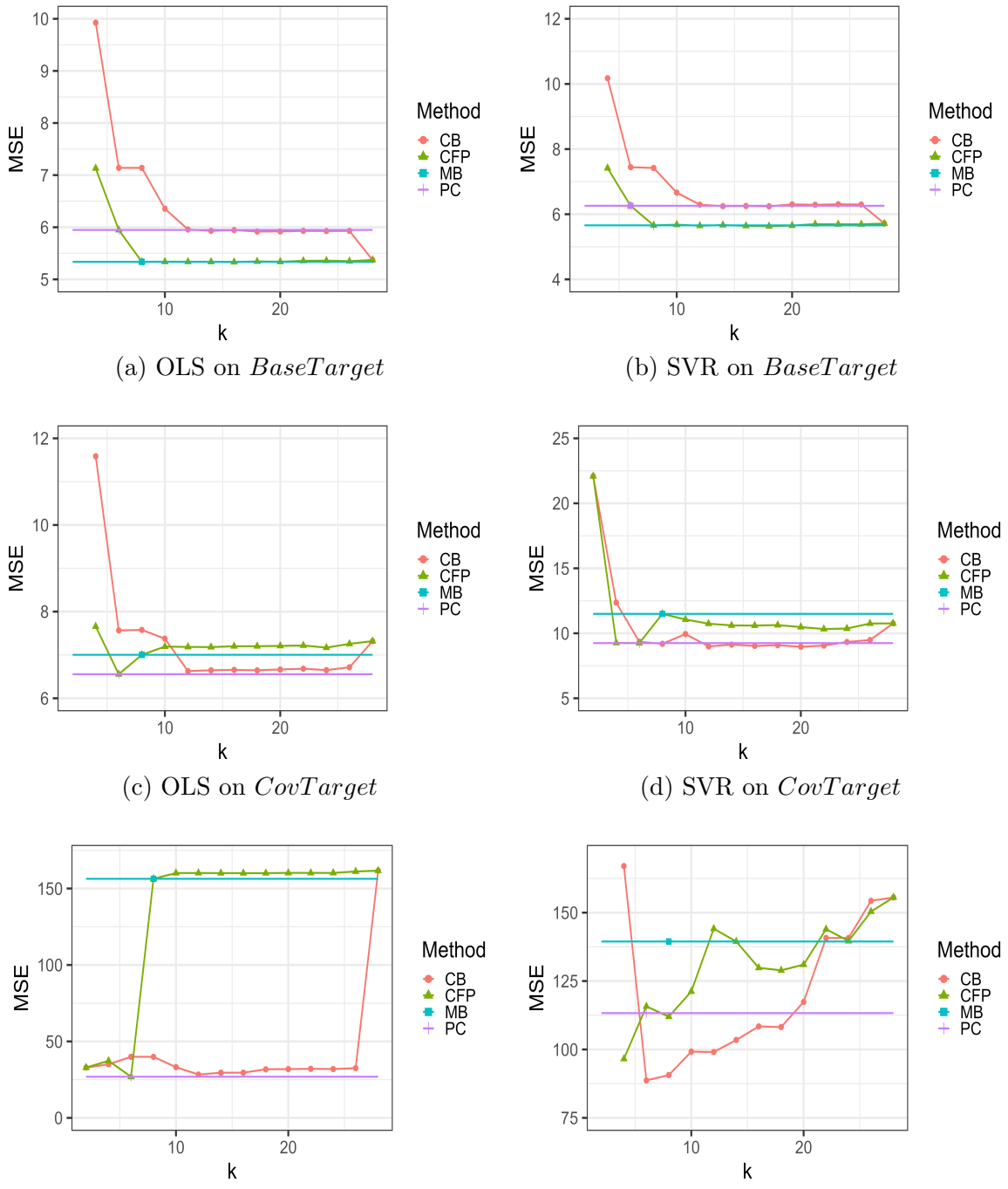
Zheng, X., B. Aragam, P. Ravikumar, and E. P. Xing, 2018: Dags with no tears: Continuous optimization for structure learning. *Advances in Neural Information Processing Systems*, **December (1)**, 9472–9483, 1803.01422.

Zheng, Z. E., and P. A. Pavlou, 2010: Toward a causal interpretation from observational data: A new bayesian networks method for structural models with latent variables. *Information Systems Research*, **21 (2)**, 365–391, https://doi.org/10.1287/isre.1080.0224.

Zhou, K., Z. Liu, Y. Qiao, T. Xiang, and C. C. Loy, 2022: Domain Generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **45 (4)**, 1–20, https://doi.org/10.1109/TPAMI.2022.3195549.

Zhu, S., I. Ng, and Z. Chen, 2019: Causal discovery with reinforcement learning. *arXiv preprint arXiv:1906.04477*.

1 Appendix A

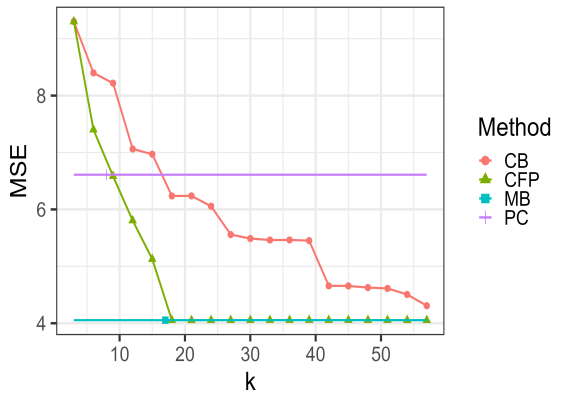


(a) OLS on *BaseTarget* (b) SVR on *BaseTarget*

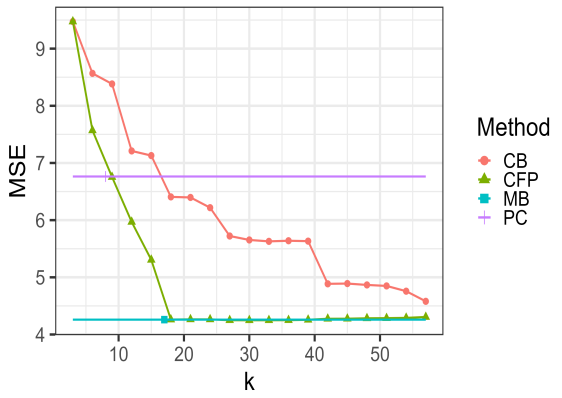
(c) OLS on *CovTarget* (d) SVR on *CovTarget*

(e) OLS on *IntTarget* (f) SVR on *IntTarget*

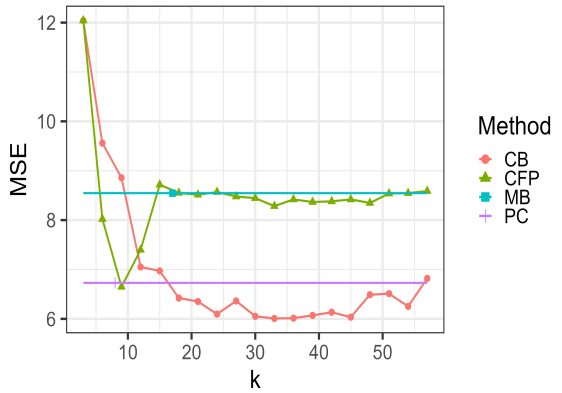
Figure A.1: Plots of prediction errors on data **D2** target datasets for OLS & SVR models.



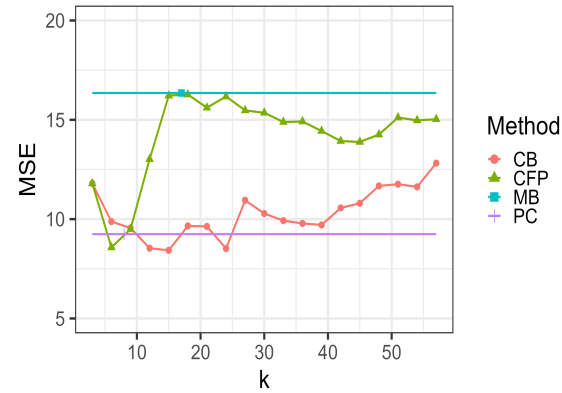
(a) OLS on *BaseTarget*



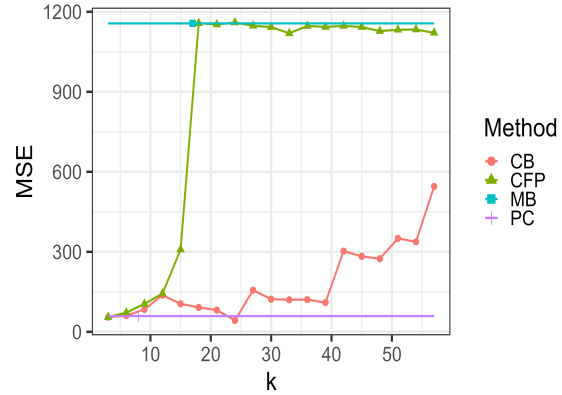
(b) SVR on *BaseTarget*



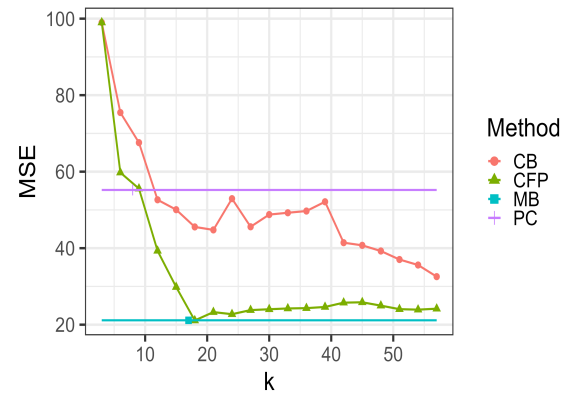
(c) OLS on *CovTarget*



(d) SVR on *CovTarget*

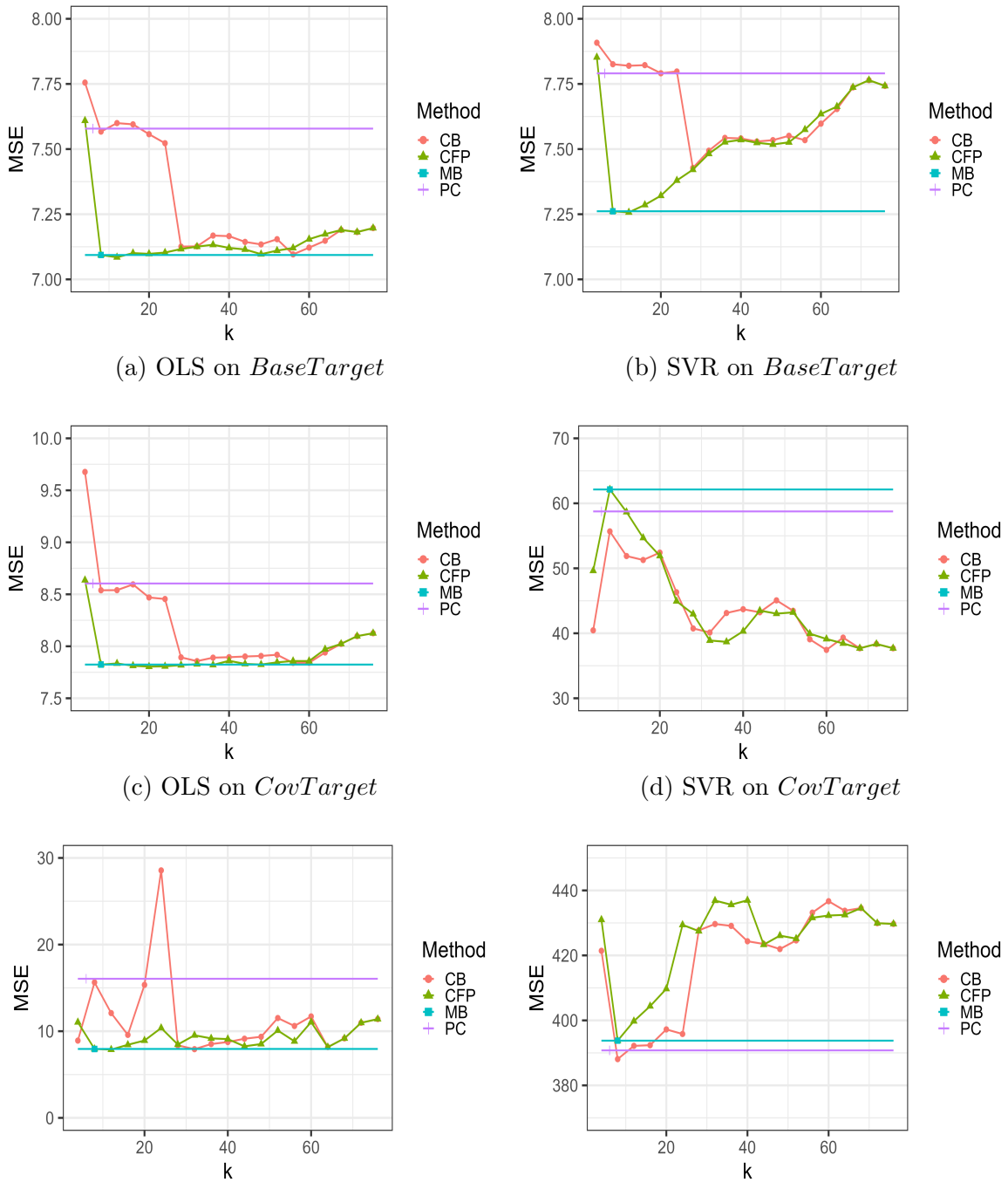


(e) OLS on *IntTarget*



(f) SVR on *IntTarget*

Figure A.2: Plots of prediction errors on data **D3** target datasets for OLS & SVR models..

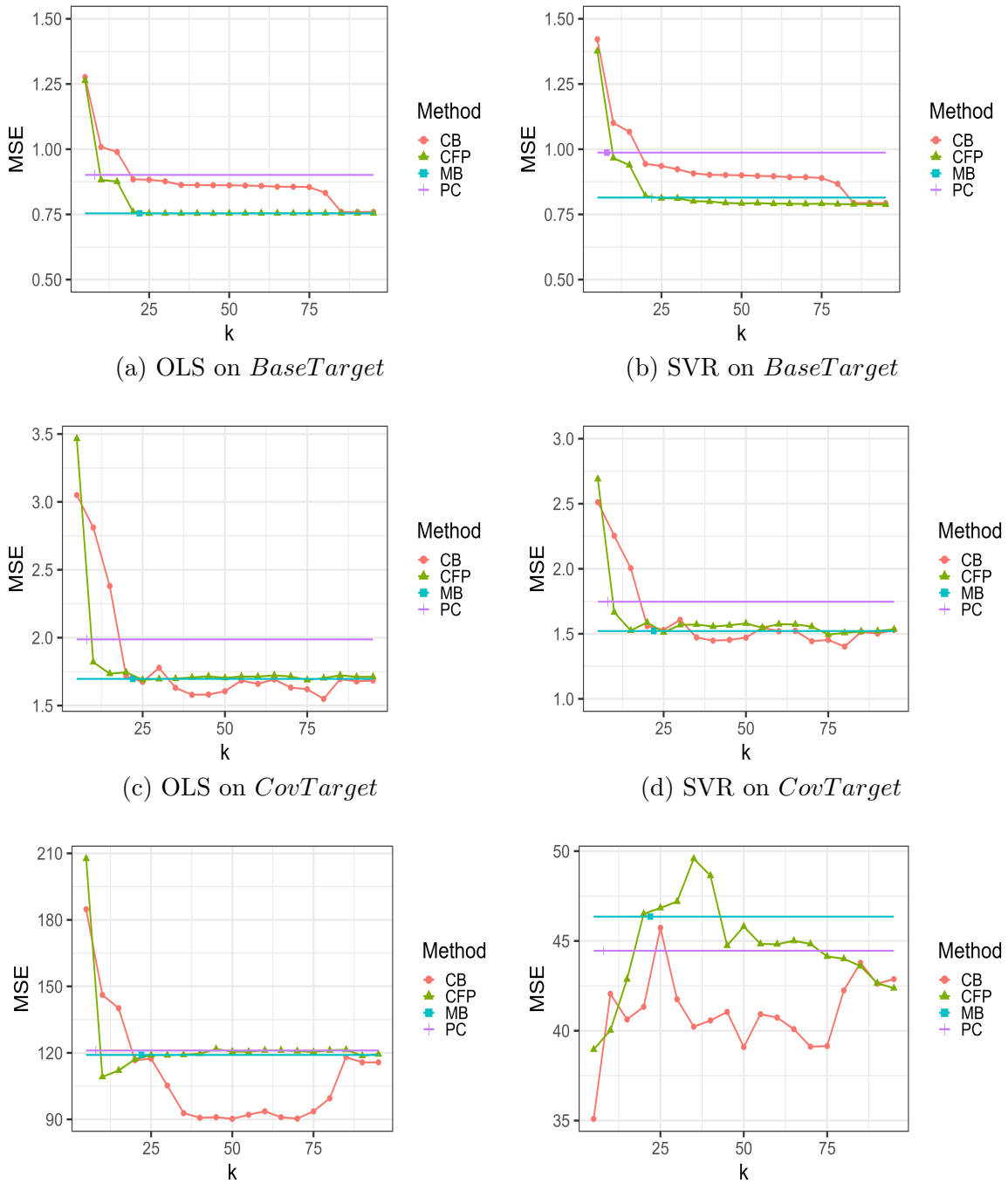


(a) OLS on *BaseTarget* (b) SVR on *BaseTarget*

(c) OLS on *CovTarget* (d) SVR on *CovTarget*

(e) OLS on *IntTarget* (f) SVR on *IntTarget*

Figure A.3: Plots of prediction errors on data **D4** target datasets for OLS & SVR models.



(a) OLS on *BaseTarget* (b) SVR on *BaseTarget*

(c) OLS on *CovTarget* (d) SVR on *CovTarget*

(e) OLS on *IntTarget* (f) SVR on *IntTarget*

Figure A.4: Plots of prediction errors on data **D5** target datasets for OLS & SVR models.