FORECASTING NONSTATIONARY TIME SERIES WITH CAUSAL INVARIANCE

Submitted to the Department of Computer Science of Amherst College

in partial fulfillment of the requirements

for the degree of Bachelor of Arts with honors

April 10, 2024

Author: Angelica Kim

Advisor: Prof. Matteo Riondato

Abstract

The traditional machine learning paradigm assumes that the training and test data come from the same distribution. In real-world scenarios, however, the data-generating distribution may change over time, leading to nonstationarity. This work addresses the challenge of forecasting nonstationary time series data by leveraging the invariance property of causal mechanisms. We investigate a method that identifies the predictors of a target variable that remain invariant to interventions across time through hypothesis testing. This approach is grounded in the principle that the conditional distribution of a target variable given its causes is unchanged even when external interventions perturb the distribution of input features. To simulate the interventions that closely mimic real-world scenarios, we utilize the concept of covariate shift, changing the causal strengths and noise variances within the data. The method is validated on synthetic data and rigorously tested from various computational perspectives. Experimental results demonstrate that invariant causal models outperform ordinary least squares regression in predicting nonstationary time series data across various levels of sample complexity, causal strengths, and noise variances.

Acknowledgements

First and foremost, I would like to express my deepest gratitude to my advisor, Matteo, for his guidance, support, and encouragement throughout this thesis work. This thesis would not have been possible without you. I really appreciate your patience and willingness to explore the field of causality together from scratch. Whenever I was feeling lost or frustrated, your were always there to provide me with the right direction and motivation every week, reminding me that I was on the right track, just being in the process of learning.

A special thanks to Professor Scott Alfeld for his insightful comments and suggestions on my thesis. Your Machine Learning course has been a great inspiration for me to draw connections between causality and machine learning in the ways that significantly enrinched my thesis work. I would also like to thank the Computer Science Department at Amherst College for providing me with the resources and support to pursue my research interests.

I would also like to thank my family in Korea for their unwavering support and encouragement throughout my journey at Amherst. Despite the 14-hour time difference, you always made time to check in on me over Facetime, providing me with the comfort and strength that kept me going through these past 8 years.

To my friends Mia and Ahanu, I cannot even imagine how I would have made it through Amherst without you guys. We met through CS classes, but you have been so much more than just project partners. Mia, you have been the most supportive and understanding friend I could ever ask for, always there to listen to my rants and reassure me that I am more than enough whenever I was feeling I was not good enough. Ahanu, you are one of very few people who makes me feel like I can be myself without any judgment. Thank you so much for being my chillest friend, being goofy with me, and making me laugh when I needed it the most.

I cannot talk about my thesis journey without mentioning Wooseok and his unconditional love and support. Whenever I was pushing myself too hard, you were always there to remind me to take a break and take care of myself. Whenever I was drowning in doubts, you were always there to remind me of my strengths and help me see the light at the end of the tunnel. Thank you for being my best listener, my most honest critic, and my biggest cheerleader all at once.

Contents

Chapter 1

Introduction

Machine learning has demonstrated its versatility in many domains, from stock price forecasting to text generation. The appeal of machine learning models comes from their ability to employ inductive reasoning, allowing them to generalize beyond hard-coded rules and make informed predictions about previously unseen data. The underlying challenge in machine learning, however, lies in the fact that observed data are mere "snapshots" of the probability distributions that generated them. Without knowledge of the underlying distributions, models strive to minimize empirical error at best, hoping it aligns closely with the error with respect to the true distribution. The fundamental theorem of statistical learning tells us that this hope is not unfounded. When a hypothesis class has a finite Vapnik-Chervonenkis (VC) dimension, risk estimates become uniformly accurate across the hypothesis class, given enough training data [Shalev-Shwartz and Ben-David, 2014]. In essence, these conditions allow the algorithm to perform empirical risk minimization (ERM) with the confidence that it will lead to minimizing the true risk as well.

Uniform convergence, in its most generic form, implicitly assumes the stationarity of the data-generating distribution. For example, traditional covariance-based methods such as linear regression rely on the concept of covariance with the expectation that it will persist into the future. However, they encounter significant challenges in the face of non-stationarity, i.e., where the training and test sets no longer come from the same distribution. In such cases, the guarantees of uniform convergence and the effectiveness of ERM can be compromised as empirical risk (i.e. error) estimated from training data is no longer a reliable approximation of the expected risk. In regression tasks, finding invariant relationships among variables thus becomes an indispensable task for enhancing the robustness of machine learning models.

In this thesis, we answer the question of how to find invariant relationships in time series data, guided by the principles of causality. We start with the premise that data is produced by a deterministic causal mechanism that is unknown to us. The causal structure, by construction, specifies which variables will be affected by changes in the system, so it provides a natural framework for identifying invariant relationships between variables [Simon, 1979]. Invariance implies that the conditional distribution of a target variable given its causes remains identical even when external interventions alter the distribution of variables other than the target variable itself. We exploit the invariance property of causal relationships to find the predictors of a target variable that withstand unforeseen changes. Simon [1979] argues that the value of causal knowledge lies in its ability to predict the effects of changes in the system. As long as the system remains unaltered, knowledge of the causal mechanism is not required for prediction, but in the face of non-stationarity, it becomes essential, for accurate predictions, to at least partially discover such mechanisms.

Evidently, there is a clear distinction between causal and statistical concepts. Existing vocabulary in probability theory is not fully equipped to express causal assumptions and claims [Pearl, 2009]. In Chapter 3, we introduce terminologies and concepts that are central to the identification of causal relationships. Chapter 4 describes a method proposed by Pfister et al. [2019] that finds the causes of a variable of interest that are invariant to interventions across time. Through hypothesis testing, the method identifies all the models that yield invariant predictions across different environments. Pfister et al. [2019] argue that the true causal model is a member of this set of models with high probability. In Chapter 5, we demonstrate the effectiveness of the method in finding invariant relationships in time series data. We validate the method on synthetic data by comparing the estimated causal predictors with the true causes. We then compare the performance of the method with traditional regression models, and show that causal knowledge is useful for robust forecasting in non-stationary environments. Finally, we discuss the implications of our findings and suggest future directions for research in Chapter 6.

Chapter 2

Related Work

Methods for inferring causal graphs can be broadly categorized into four classes: constraint-based, score-based, structural-causal-model-based, and Granger-causality-based methods. In this section, we discuss their generalized version for multivariate time series data. Constraint-based methods like the Peter-Clark (PC) algorithm and its time series extension, PCMCI, rely on tests of conditional independence to construct a graph that outlines potential causal relationships [Runge et al., 2019; Spirtes et al., 2000]. These methods start by creating a basic structure without directed edges based on which variables are conditionally independent of each other. Then, they use rules to direct the edges, forming a graph that represents the causal relationships.

Score-based methods, such as NTS-NOTEARS, assess how well different causal models fit the observed data using probabilistic scores [Sun et al., 2021]. This approach allows for the comparison of many possible models, selecting the one that best explains the data. It contrasts with constraint-based methods, which typically suggest a single model without indicating how certain it is. Score-based methods are more flexible because they use a goodness-of-fit measure instead of relying strictly on conditional independence tests.

The two families of methods mentioned above encounter some challenges. They struggle to distinguish between the models within the same Markov equivalence class—a set of causal structures that share the same conditional independencies. They also require faithfulness, which is a strong assumption imposing that the observed conditional independencies in the data accurately reflect the true causal relationships. Another approach that goes beyond these limitations is the Structural Causal Models (SCM) framework, which combines the causal and probabilistic aspects of the data-generating process [Gong et al., 2023]. A method called Time Series Models with Independent Noise (TiMINo) leverages the Additive Noise Model framework [Peters et al., 2013]. It aims to identify causal relationships by fitting structural causal models that can capture temporal dynamics, where each variable at time t is expressed as a function of its own past values and an independent noise term. The method uses independence tests, like cross-correlations and the Hilbert-Schmidt Independence Criterion (HSIC), to detect causal relationships within the additive noise model setting.

Granger-causality-based methods evaluate the causal impact of one variable on another by testing if past values of the predictor variable can significantly improve the prediction of the current value of the target variable [Gong et al., 2023]. Temporal Causal Discovery Framework (TCDF) is one such method, leveraging attention mechanisms within deep learning models to identify and quantify causal relationships over time, focusing specifically on the temporal dynamics between variables [Nauta et al., 2019]. However, Granger causality relies on the assumption that data is stationary and the cause precedes its effect in time, which means it cannot detect instantaneous effects [Gong et al., 2023]. This limitation makes Granger causality less effective in the real world setting.

Recent work also uses deep learning models to learn causal representations from data. Ke et al. [2022] propose a method that uses transformers for causal discovery in the i.i.d. setting. They treat the inference process as a black box and design a neural network architecture that learns the mapping from both observational and interventional data to graph structures via supervised training on synthetic graphs [Ke et al., 2022]. In the early stage of this thesis work, we have attempted to extend this approach to time series data, but it was not able to capture the causal relationships from the data as effectively as we had hoped, not to mention the subtantial computational complexity of training a transformer model.

After exploring various methods, we internalized that discovering the full causal structure from scratch is a challenging task, often requiring strong, if not unrealistic, casual assumptions. What proved effective for us was to focus on finding the invariant predictors of a target variable within time series data based on the principles of causality, as introduced by Pfister et al. [2019]. We find all the models that produce invariant predictions across different environments through hypothesis testing. This process leads us to likely identify the true causal model within this set.

We find the hypothesis testing approach to be effective for causal discovery because it allows us to set up our hypothesis about the causal relationships between variables and iteratively validate them against the observed data, instead of finding the full causal structure from scratch using a black-box model. This perspective is corroborated by Pearl and Mackenzie [2018], where it is remarked that the goal of causal discovery is not to deduce the entire causal structure from bottom up, but rather to "[represent] plausible causal knowledge in some mathematical knowledge, [combine] it with empirical data, and [answer] causal queries that are of practical value." The method we use also proves practical because it does not rely on strong causal assumptions such as faithfulness. Moreover, it leverages covariate shift, which is common in the real-world data, to identify the invariant predictors of a target variable, facilitating accurate predictions in the face of non-stationarity.

Chapter 3

Preliminaries

3.1 Graph Terminology

Graphs offer a visual and intuitive means to represent causal relationships, enabling the identification of direct and indirect influences among variables. This section introduces some terminology for understanding graphical representations of causality.

A directed graph $G = (V, E)$ is denoted as a pair of vertices V (or nodes, used interchangeably) and edges $E \subseteq V \times V$ such that the pairs $(u, v) \in V \times V$ are considered ordered, that is $(u, v) \neq (v, u)$. In an edge (u, v), u is a parent of its child v. A path in a directed graph is a sequence of nodes $(u_1, u_2, ..., u_k)$ such that (u_i, u_{i+1}) or $(u_{i+1}, u_i) \in E$ for all $i = 1, ..., k - 1$ and $u_i \neq u_j$ for all $i, j = 1, ..., k$. Note that a path, by its definition, can "flow" in any direction, which may go either along or against the direction of the edges, such as $u \rightarrow v \leftarrow w$ or $u \leftarrow v \rightarrow w$. This definition is in line with causality literature, where the direction of the path is not restricted by the direction of the edges [Pearl, 2009]. A path is said to be directed if $(u_i, u_{i+1}) \in E$ for all $i = 1, ..., k - 1$. A directed graph is acyclic if and only if there are no nodes $u, v \in V, u \neq v$ such that there are directed paths from u to v and from v to u. Such a graph is called a Directed Acyclic Graph (DAG). All nodes from which there is a directed path to a node u are its ancestors, and all nodes to which there is a directed path from a node u are its descendants. We now introduce a concept that is central to the identification of causal relationships from a DAG.

Definition 1 (Blocking) *A path p is said to be blocked by a set of nodes Z if and only if*

1. *p contains a chain $i \to m \to j$ or a fork $i \leftarrow m \to j$ such that the middle node m is in Z,*

2. *p contains a collider $i \to m \leftarrow j$ such that the middle node m is not in Z and no descendant of m is in Z.*

Definition 2 (d-separation) *A set Z is said to d-separate X from Y if and only if Z blocks every path from a node in X to a node in Y.*

Paths that are not blocked by any set of nodes are *d-connected*. The concept of d-separation is useful for identifying conditional independence relationships among variables in a DAG, which we discuss in the next section.

DAGs are a powerful tool for modeling causal relationships. Their directionality and sense of hierarchy are useful for expressing the precedence of causes over effects, which is both temporal and logical. However, graphs alone do not provide a complete picture of the causal relationships. To establish causality, we need to go beyond graphs, and consider the probabilistic relationships between the variables. In the next section, we introduce a framework that formalizes the causal generative process by relating both causal and probabilistic statements. Using this framework, we can mathematically define what it means for a variable to be a cause of another variable.

3.2 Structural Causal Model

We start from the premise that the observed data are generated by causal mechanisms that are deterministic functional relationships between variables. We assume that the mechanisms take the form of acyclic structures that can be represented by DAGs.

Definition 3 *A **causal structure** of a set of variables V is a directed acyclic graph $G = (V, E)$ in which each node corresponds to a distinct element of V, and each edge represents a direct functional relationship where the parent node is a cause of the child node.*

Causal structures are a graphical representation of the causal relationships between variables. It serves as a blueprint for forming a causal model that specifies the types of the functional relationships and the probability distribution of the noise terms that inject variability into the system [Pearl, 2009].

Definition 4 *A **Structural Causal Model (SCM)** $\mathbb{C} :=< G, \Theta_G >$ consists of a causal structure $G = (V, E)$ and a set of parameters Θ_G that assign a function*

$$X_i := f_i(PA_i, U_i) \tag{3.1}$$

*to each variable $X_i \in V$ and a probability distribution $P(U_i = u_i)$ to each noise term U_i. PA_i is a set of parents or **direct causes** of $X_i \in G$ and U_i is a noise term that is jointly independent of all other $U_{j \neq i}$.*

$$\begin{aligned} X &:= U_X \\ W &:= X + U_W \\ Y &:= X + W + U_Y \\ Z &:= Y + U_Z \end{aligned} \tag{3.2}$$

where $U_X \overset{\text{ind}}{\sim} \mathcal{N}(\mu_X, \sigma_X^2)$,
$U_W \overset{\text{ind}}{\sim} \mathcal{N}(\mu_W, \sigma_W^2), U_Y \overset{\text{ind}}{\sim} \mathcal{N}(\mu_Y, \sigma_Y^2)$,
$U_Z \overset{\text{ind}}{\sim} \mathcal{N}(\mu_Z, \sigma_Z^2)$.

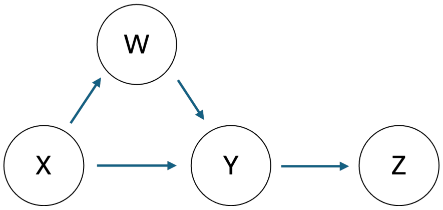

Figure 3.1: An example of an SCM (left) and its corresponding causal structure (right).

Figure 3.1 shows an example of an SCM and its corresponding causal structure. The structural assignments in the SCM specify how the value of each variable is determined by the values of its parents and noise terms. Note that they are different from algebraic equations. They are inherently directional (like a variable assignment in a programming language), whereas algebraic equations treat both sides symmetrically (See Section 3.6 for more details). While the structural assignments are also called structural equations in econometrics, we intentionally avoid using the term "equation" [Peters et al., 2017].

An SCM combines the causal and probabilistic aspects of the data-generating process. It entails a unique joint distribution P over the variables $\mathbf{X}$ = $(X_1, X_2, ..., X_d)$ in the sense that the values of $\mathbf{X}$ are uniquely determined by the distribution of the noise terms $P(U_i = u_i)$.

Definition 5 *An SCM $\mathbb{C}$ defines a unique distribution of the variables $X_1, X_2, ..., X_d$: any $X_1, X_2, ..., X_d$, $U_1, ..., U_d$ satisfying $X_i = f_i(PA_i, U_i)$ almost surely, where $(U_1, U_2, ..., U_d)$ has the desired distribution, induce the same distribution over $\boldsymbol{X}$ = $(X_1, X_2, ..., X_d)$. We denote it as the **entailed distribution** $P_{\boldsymbol{X}}^{\mathbb{C}}$.*

The joint independence of the noise terms in Definition 4 follows from two assumptions: (1) every variable that is a cause of another variable is included in the model; and (2) "no correlation without causation," stating that if any two variables are dependent, then one is a cause of the other or there is a third variable causing both [Pearl, 2009]. If an unobserved common cause had not been included in the model, the noise terms would have been correlated. Hence, these two assumptions allow the noise terms to be jointly independent. By virtue of the joint independence assumption, we can facilitate an economical representation of the entailed distribution $P_{\mathbf{X}}^{\mathbb{C}}$. The chain rule of probability states that the joint distribution of d variables can be decomposed into a product of d conditional distributions:

$$P(X_1, X_2, ..., X_d) = \prod_j P(X_j | X_1, X_2, ..., X_{j-1}).$$

Given an SCM, a variable X_j follows a distribution that is determined by its parents PA_j and a noise term U_j that introduces variability into the variable. Hence, we can use joint independence of the noise terms to decompose the joint distribution of the variables into a product of conditional distributions of each variable given its parents [Pearl, 2009]:

$$P_{\mathbf{X}}^{\mathbb{C}} = P(X_1, X_2, ..., X_d) = \prod_j P(X_j | PA_j). \tag{3.3}$$

This factorization implies that once we know the values of the parents of a variable, the value of the variable is independent of all other non-descendants. In other words, knowing the values of other variables does not provide any additional information about the variable beyond what is already provided by its parents. This property equips us with a powerful tool for identifying causal relationships from the data.

Assume we are given a joint distribution of the variables whose causal structure is *unknown*. If we find a DAG such that the distribution factorizes according to Equation 3.3, we can infer that the graph represents the causal structure of the variables.

Definition 6 (Markov Compatibility) *If a probability distribution P admits the factorization of Equation 3.3 relative to DAG G, we say that G represents P, that G and P are compatible, or that P is Markovian* *relative to G.*

Definition 6 tells us that if the value x_j of each variable X_j is chosen at random with some probability $P_j(X_j = x_j|PA_j = pa_j)$ based solely on the value pa_j chosen for PA_j as determined by a causal structure G that we discovered, then the overall distribution P will be Markovian relative to the DAG G. In other words, compatibility between a distribution and a DAG that we infer from the data allows the graph to represent the causal mechanism of the variables. A convenient way to verify the compatibility is to list all conditional independence relationships that the graph implies and check if they hold in the data. These conditional independencies can be read off from the graph using d-separation.

Theorem 1 *For any disjoint sets of nodes X, Y, Z in a DAG G and for all probability distributions P,*

$$X \perp\!\!\!\perp_G Y \,|\, Z \implies X \perp\!\!\!\perp Y \,|\, Z$$

when P is Markovian with respect to a graph $\mathcal{G}$. The symbol $\perp\!\!\!\perp_G$ denotes d-separation.

If conditional independencies implied by the graph hold in the data, we can understand which variables are direct causes of others. This understanding is critical when predicting the effect of changes in the system because it tells us which variables' relationships will remain invariant under such changes, which we will further discuss in Section 3.4.

3.3 SCM for Time Series

In our work, we are interested in finding the causes of a variable of interest within time series data. So far, we have only considered the settings where samples are i.i.d. drawn from a joint distribution P. We adapt the terminologies to a multivariate time series setting to incorporate the notion of time in the causal model.

Assume we have a d-variate time series $(\mathbf{X}_t)_{t\in\mathbb{Z}}$, where $\mathbf{X}_t = (X_t^1, ..., X_t^d) \in \mathbb{R}^{1\times d}$ for each t. X_t^i represents a measurement of the variable i at time t. We assume that the time series is generated by an

SCM in which at most the past q values (for some q) of all variables occur.

$$X_t^i := f^i(PA_{t-q}^i, ..., PA_{t-1}^i, PA_t^i, U_t^i),$$

where

$$..., U_{t-1}^1, ..., U_{t-1}^d, U_t^1, ..., U_t^d, U_{t+1}^1, ..., U_{t+1}^d, ...$$

are jointly independent noise terms. PA_{t-s}^i denotes the set of variables observed at time $t - s$ that cause X_t^i for $s = 1, ..., q$. If PA_{t-s}^i is non-empty for X_t^i, we say there is a delayed effect. If PA_t^i is non-empty for X_t^i, we say there is an instantaneous effect. The inherent temporal ordering guides the identification of causal relationships, as it imposes a natural constraint that causes must precede their effects in time or occur simultaneously.

We restrict the function class of f^i to be linear functions with additive Gaussian noise of the form:

$$X_t^i := \sum_{s=1}^{q} \mathbf{X}_{t-s}\beta_s^i + U_t^i, \tag{3.4}$$

where $\beta_t^i \in \mathbb{R}^d$ and $U_t^i \sim N(0, \sigma_{U_i}^2)$. This restriction is a popular special case of SCMs in time series analysis, known as Vector Autoregressive Models (VAR). However, the method we describe to identify invariant predictors of a variable can also be extended to non-linear functions with non-Gaussian noise [Pfister et al., 2019].

3.4 Interventions

SCMs provide a convenient framework for quantifying the impact of changing the distribution of one variable on another. By articulating the data-generating mechanisms, they enable accurate predictions about the effects of these changes. Essentially, they allow us to understand the causal influence of one variable on another by modifying the former's distribution and observing the change in the latter. In this section, we introduce the notion of interventions, which provide a structured method to describe how changing a variable's distribution affects the that of another in an SCM.

Definition 7 (Intervention) *Consider an SCM $\mathbb{C}$ and its entailed distribution $P_{\boldsymbol{X}}^{\mathbb{C}}$. We replace one (or several) of the structural assignments to obtain a new SCM $\tilde{\mathbb{C}}$. Assume we replace the assignment for X_i by*

$$X_i := \tilde{f}_i(\widetilde{PA}_i, \tilde{U}_i).$$

*We call the entailed distribution of the new SCM an **intervention distribution**, and say that the variables whose structural assignments we have replaced have been **intervened** on. We denote the intervention distribution as*

$$P_{\boldsymbol{X}}^{\tilde{\mathbb{C}}} := P_{\boldsymbol{X}}^{\mathbb{C};do(X_i:=\tilde{f}(\widetilde{PA}_i,\tilde{U}_i))}.$$

The set of noise variables in $\tilde{\mathbb{C}}$ are still required to be jointly independent.

Judea Pearl observed that the language of probability lacked a means to differentiate between setting a variable's value and simply observing it. He pointed out that this limitation hinders the modeling of causal relationships [Pearl, 2009]. He introduced the notation "$do(X := x)$" for setting $X = x$. The do-notation deletes the original structural assignment of a variable $X_i := f_i(PA_i, U_i)$ in the SCM and adds the assignment $X_i := x$.

Replacing the structural assignment of a variable changes the data-generating distribution of the variable. In case of hard interventions where value of a variable is set to a constant value, the distribution of the intervened variable is a point mass at the constant value. Graphically, one may think of hard interventions as a surgical operation that cuts off the edges between the intervened variable and its parents in the graph. This operation renders the distribution of the intervened variable no longer dependent on its parents, allowing us to isolate its causal effect on another variable.

Joint independence of the noise terms in Definition 7 is a crucial assumption in using interventions to identify causal relationships. Recall that in Equation (3.3), it enabled the decomposition of the joint distribution into a product of conditional distributions of each variable given its direct causes. Each conditional distribution in the product is a causal mechanism that specifies how the value of a variable is determined by the values of its direct causes and noise terms; variables that are not direct causes are not involved in the determination. Equation (3.3) thus suggests the independence of the causal mechanisms where knowledge of one mechanism does not inform or influence another. Hence, even if we intervene on a variable, then the other mechanisms would remain unchanged. Given an SCM $\mathbb{C}$, we have the following invariance property:

$$P^{\tilde{\mathbb{C}}}_{X_j|PA_j=pa_j} \overset{d}{=} P^{\mathbb{C}}_{X_j|PA_j=pa_j}$$

for any SCM $\tilde{\mathbb{C}}$ that is constructed from $\mathbb{C}$ by intervening on some variable other than X_j (Proof in Section 3.5). The equation shows that causal relationships are *autonomous* or *invariant* under interventions. Peters et al. [2017] state that the observed data is an instantiation of the following general principle of independent mechanisms.

Principle 1 *The causal generative process of a system's variables is composed of autonomous modules that do not inform or influence each other.*

This autonomy allows us to apply localized interventions where we can change the distribution of a variable without affecting the distribution of other variables [Pearl, 2009]. The atomic intervention $do(X_i := x'_i)$ results in the truncated factorization of the joint distribution, where the distribution of the intervened variable is no longer dependent on its parents:

$$P^{\tilde{\mathbb{C}}}_{\mathbf{X}}(x_1, ..., x_d) = \begin{cases} \prod_{j \neq i} P(x_j|pa_j) & \text{if } X_i = x'_i, \\ 0 & \text{otherwise} \end{cases}$$

We can now define the causal effect of one variable on another using interventions.

Example Consider the following SCM we introduced in Figure 3.1:

$$\begin{aligned} X &:= U_X \\ W &:= X + U_W \\ Y &:= X + W + U_Y \\ Z &:= Y + U_Z \end{aligned} \tag{3.5}$$

where $U_X \sim \mathcal{N}(\mu_X, \sigma_X^2), U_W \sim \mathcal{N}(\mu_W, \sigma_W^2), U_Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2), U_Z \sim \mathcal{N}(\mu_Z, \sigma_Z^2)$ and U_X, U_W, U_Y, U_Z are jointly independent.

Hence,

$$P_Y^{\mathbb{C}} := \mathcal{N}(E[U_X] + E[U_W] + E[U_Y], \sigma^2_{U_X} + \sigma^2_{U_W} + \sigma^2_{U_Y}).$$

Assume we are interested in predicting Y from X and Z. If we intervene to set Z to a value of 100, the distribution of Y is unaffected:

$$P_Y^{\mathbb{C};do(Z:=100)} = P_Y^{\mathbb{C}}.$$

However, intervening on X by setting it to 100 changes the distribution of Y:

$$P_Y^{\mathbb{C};do(X:=100)} := \mathcal{N}(200 + E[U_W] + E[U_Y], \sigma^2_{U_W} + \sigma^2_{U_Y}) \neq P_Y^{\mathbb{C}}.$$

Given that intervening on X alters Y's distribution, while an intervention on Z leaves it unchanged, it follows that X causes Y, whereas Z does not. Pearl calls the mapping from the value x of X to $P_Y^{\mathbb{C};do(X:=x)}(y)$ for all x the *causal effect* of X on Y [Pearl, 2009]. Note that this effect includes both direct and indirect effects of X on Y through W. Direct effects of X on Y can be identified by intervening on *all* direct causes of Y. Cutting off all the incoming edges to Y essentially blocks all the directed paths that X can take to reach the target variable. Pearl calls the mapping from the values $(x, pa_{Y\setminus X})$ to $P_Y^{\mathbb{C};do(X:=x, PA_{Y\setminus X}:=pa_{Y\setminus X})}$ for all $(x, pa_{Y\setminus X})$, where $pa_{Y\setminus X}$ is a realization of the parents of Y other than X, the *direct effect* of X on Y [Pearl, 2009].

It is also possible to summarize the direct effect of X on Y by a single number. Pearl [2009] defines the direct effect of X on Y as the partial derivative with respect to x of the expectation of Y when X is set to x:

$$\frac{\partial}{\partial x}\mathbb{E}^{\mathbb{C};do(X:=x)}[Y].$$

Indeed, assuming that the noise terms have zero expectation, we can obtain the structural coefficient $\alpha = 1$ of X on Y in the SCM 3.5 as follows:

$$\frac{\partial}{\partial x}\mathbb{E}[Y|do(X := x), W = w] = \frac{\partial}{\partial x}(x + w) = 1.$$

Pearl [2009] states that when we are not given the joint distribution but a sample from it, the direct effect of X on Y can be estimated using the regression coefficient of Y on X and a set of variables that meet certain conditions known as the *single-door criterion* for direct effect, which we discuss below.

Assume we have a set of variables U that lie on d-connected paths between X and Y. Consider the partial regression coefficient $r_{YX \cdot U} = \rho_{YX \cdot U} \sigma_{Y \cdot U} / \sigma_{X \cdot U}$, which measures the direct relationship between X and Y after controlling for U. If U contains no descendants of Y, we can write

$$r_{YX \cdot U} = \alpha + I_{YX \cdot U},$$

where α is the direct effect of X on Y and $I_{YX \cdot U}$ is the partial correlation between X and Y in a model whose graph G_α is obtained by removing the edge between X and Y. If U d-separates X and Y in G_α, $I_{YX \cdot U} = 0$ and $r_{YX \cdot U} = \alpha$. We can then estimate the direct effect of X on Y by regressing Y on X and U:

$$Y = \alpha X + \beta U + \epsilon.$$

The following theorem provides a graphical criterion where a regression coefficient provides a consistent estimate of the direct causal effect.

Theorem 2 (Single-Door Criterion for Direct Effect) *Let G be a causal structure in which α is the direct effect of X on Y associated with edge $X \rightarrow Y$. Let G_α denote the graph obtained by deleting the edge $X \rightarrow Y$. The coefficient α is identifiable if there is a set U such that (1) U contains no descendant of Y and (2) U d-separates X and Y in G_α. If these conditions are satisfied, then α is equal to the regression coefficient $r_{YX \cdot U}$. Conversely, if U does not satisfy these conditions, then $r_{YX \cdot U}$ is not a consistent estimate of α (except in rare instances).*

3.5 Invariance to Interventions

Each structural assignment in an SCM is identified using interventions and thus specifies, by design, which variables will be affected by interventions. In other words, causal mechanisms are inherently invariant under interventions [Simon, 1979]. By invariance, we mean that the conditional distribution $P(x_j | pa_j)$ remains unchanged across different environments, including both observational and interventional settings:

$$P^{\tilde{\mathbb{C}}}_{X_j|PA_j=pa_j} \overset{d}{=} P^{\mathbb{C}}_{X_j|PA_j=pa_j}. \tag{3.6}$$

The intervention should not involve the variable X_j itself because it would change its own structural assignment and hence the causal mechanism. Proof of this property is immediate from the definition of SCMs.

Proof of Equation 3.6 Assume we are given an SCM where $x_j := f_j(pa_j, \epsilon_j)$ where $\epsilon_j \perp\!\!\!\perp PA_j$. After intervening on some variable X_i but not on X_j, we obtain a new environment e where $x^e_j := f_j(pa^e_j, \epsilon^e_j)$. Because we do not intervene on X_j, the structural assignment for X_j and the distribution of ϵ^e_j remain the same. Hence, for a given pa_j, the function f does not depend on the environment e. Thus, $P^{\tilde{\mathbb{C}}}_{X_j|PA_j=pa_j} \overset{d}{=} P^{\mathbb{C}}_{X_j|PA_j=pa_j}$.

Additionally, if we were to compare two different interventional settings (e.g. $do(X_k := x)$ and $do(X_k := x'))$ where the intervened variable $X_k \in PA_j$, the conditional distribution would remain the same in terms of the structural parametrization. In other words, changing one variable due to an intervention does not alter the functional form describing the causal mechanisms. This invariance property is a key principle we leverage to identify causal predictors in our method.

3.6 Structural Causal Models vs Algebraic Equations

In Section 3.4, we discussed how the direct effect of X on Y can be estimated using the regression coefficient of Y on X and a set of variables that meet the single-door criterion for direct effect. However, it is important to note that SCMs and regression equations are inherently different.

Consider two SCMs:

$$\begin{aligned} X &:= \epsilon + U_X \\ Z &:= \alpha X + U_Z \\ Y &:= \beta Z + \gamma\epsilon + U_Y \end{aligned} \tag{3.7}$$

and

$$\begin{aligned} X &:= U_X \\ Z &:= \alpha' X + U_z \\ Y &:= \beta' Z + \delta X + U_Y \end{aligned} \tag{3.8}$$

where ϵ is an unobserved noise term that follows $\mathcal{N}(0, \sigma_\epsilon^2), U_X \sim \mathcal{N}(0, \sigma_X^2), U_Y \sim \mathcal{N}(0, \sigma_Y^2), U_Z \sim \mathcal{N}(0, \sigma_Z^2)$ and they are jointly independent.

Upon setting $\alpha = \alpha', \beta = \beta'$ and $\delta = \gamma$, the two models will yield the same probabilistic predictions for Y. However, each depicts a different story about data-generating processes, and their predictions will differ under interventions.

For example, suppose we are interested in predicting the expectation of Y after intervening on X in the first SCM (3.7).

$$\mathbb{E}^{do(X:=x)}[Y] = \mathbb{E}[\beta(\alpha x + U_Z) + \gamma\epsilon + U_Y] = \beta\alpha x$$

The second SCM (3.8) will yield a different prediction:

$$\mathbb{E}^{do(X:=x)}[Y] = \mathbb{E}[\beta'(\alpha' x + U_Z) + \delta x + U_Y] = (\beta'\alpha' + \delta)x$$

Even upon setting $\alpha = \alpha', \beta = \beta'$ and $\delta = \gamma$, we can see that the two models assign different causal effects to X on Y: a unit change in X will result in a change of $\beta\alpha$ in Y in the first model and $\beta\alpha + \gamma$ in the second model.

One might wonder if we can subsitute $X - U_X$ for the ϵ term in Y in the first model prior to taking the expectation to make the two quantities equal. This argument highlights the difference between SCMs and algebraic equations. In algebraic equations, we can substitute one variable for another as long as they are equal. However, SCMs are not to be treated as "immutable mathematical equalities" [Pearl, 2009]. Since they describe data-generating processes and thus "a state of equilibrium" reached through causal mechanisms, altering these processes by substitution would change the meaning and interpretation of the model [Pearl, 2009].

Regression equations are primarily concerned with prediction and capture covariances between variables. Substitution is often used as a mathematical convenience to simplify expressions or solve for variables, because the equations primarily capture associative relationships rather than causal mechanisms. This substitution can lead to equivalent forms of the model that still accurately predict outcomes, as long as the statistical relationships between variables remain consistent.

SCMs, on the other hand, encode how interventions on one variable causally affect others, maintaining a focus on the directionality and mechanism of causation. Simon [1979] suggests that the value of structural models lies in their ability to predict the effects of changes in the system. Structural coefficients can answer the counterfactual question such as ”What would be the change in the expected value of Y if we change the value of X from x to x+1?,” which is evidently different from the question answered by regression coefficients, ”What would be the difference in the expected value of Y if we were to *find* X at level of x+1 instead of x?”. These questions lead to the difference between intervention and conditioning, which we will discuss in the next section.

Simon [1979] discusses the operational significance of structural models. He posits that distinguishing feature of SCMs from non-structural sets of equations describing identical sets of observations lies in the application of interventions. As long as structures remain unchanged, identifiability of causal mechanisms is not required to estimate the parameters that are needed for prediction. When a change in structure occurs, however, the causal mechanisms must be identifiable if correct predictions are to be made in the new structure. He suggests that “these epistemological considerations” reveal the conditions under which structural causal models can be distinguished from nonstructural equations, which are also the conditions that lend operational meaning to causal structures.

3.7 Intervention vs Conditioning

Consider the first SCM in Equation 3.7:

$$
\begin{aligned}
X &:= \epsilon + U_X \\
Z &:= \alpha X + U_Z \\
Y &:= \beta Z + \gamma \epsilon + U_Y
\end{aligned}
$$

where ϵ is an unobserved noise term that follows $\mathcal{N}(0, \sigma_\epsilon^2), U_X \sim \mathcal{N}(0, \sigma_X^2), U_Y \sim \mathcal{N}(0, \sigma_Y^2), U_Z \sim \mathcal{N}(0, \sigma_Z^2)$ and they are jointly independent.

In Section 3.6, we saw that $\mathbb{E}^{do(X:=x)}[Y] = \alpha\beta x$. Below we show that this quantity is different from $E[Y|X = x]$. If we treated the SCM as a set of regression equations, we could substitute $X - U_X$ for ϵ in Y to obtain

$$
\begin{aligned}
Y &= \beta\alpha X + \gamma(X - U_X) + U_Y \\
&= (\beta\alpha + \gamma)X - \gamma U_X + U_Y.
\end{aligned}
$$

Hence,

$$E[Y|X = x] = r_{YX} x = (\alpha\beta + \gamma)x$$

where r_{YX} is the regression coefficient of Y on X.

Conditioning does not change the distribution of the variable being conditioned on. In other words, variance of X remains the same as σ_X^2 under conditioning. By conditioning on $X = x$, we are simply finding X at a certain level x from its distribution. In contrast, intervening on X by fixing it to x sets its distribution to a point mass at x, hence $P(X = x) = 1$ and variance of X is now 0.

3.8 Covariate Shift

Although fixing a variable at a constant value allows us to isolate the causal effect of a specific variable on another variable, it is not always feasible to intervene on a variable in such a deterministic manner in practice. Instead, we can assume that the distribution of the observational data changes over time and identify the causal effects that remain invariant to these changes. This phenomenon is referred to as covariate shift, a common problem in machine learning characterized by the training and test sets originating from distinct distributions, while the conditional distribution of the target variable given the input variables remains the same [Sugiyama and Kawanabe, 2012]. This approach offers a more feasible strategy for identifying causal relationships from observational data, considering that the data distributions are often subject to change in the real-world settings.

We leverage covariate shift as a form of interventions that naturally occur in the data-generating process to detect an invariant causal mechanism from observational data. We divide the data into multiple blocks or environments and assume that the distributions of the input variables change across these environments. We simulate the covariate shift by (1) changing the variance of the distribution of the noise terms in (3.4) by a multiplicative factor, or (2) changing the strength of causal effects by a multiplicative factor. Not only does this appproach model the effect of external interventions on the distribution of the input variables, but also reflects more realistic settings where the distribution of the data changes over time.

Chapter 4

Methods

We aim at finding the causes of a variable of interest in time series data, rather than to recover the entire causal structure of the data. In this chapter, we describe a method that finds the causal predictors that are invariant to interventions across time. We adopt and elaborate on the approach introduced in [Pfister et al., 2019].

4.1 Invariant Causal Prediction

Assume we are given an input data from a time series $(\mathbf{Y}, \mathbf{X}) = (Y_t, X_t)_{t \in \{1,2,...,n\}} \in \mathbb{R}^{n \times (d+1)}$, where $X_t \in \mathbb{R}^{1 \times d}$ is a set of predictor variables at time t and $Y_t \in \mathbb{R}$ is a target variable of interest. We are interested in a setting where distributions of the variables are subject to change. Assumption 1 underpins the method: there is a set of variables whose relationship with the target variable remains invariant under distributional shifts such as interventional settings.

Assumption 1 (Invariant Prediction) *There exists a subset $S \subseteq \{1, ..., d\}$ that satisfies the following:*

(a) $\forall t \in \{1, 2, ..., n\}$*:* $Y_t = \mu + X_t^S \beta + \epsilon_t$ *and* $\epsilon_t \perp\!\!\!\perp X_t^S$,

(b) $\epsilon_1, ..., \epsilon_n \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$

where $\mu \in \mathbb{R}, \beta \in (\mathbb{R} \setminus \{0\}^{|S| \times 1})$ *and* $\sigma \in \mathbb{R}_{>0}$*. Without loss of generality, we often ignore the intercept term* μ*. We call the set* S *an* ***invariant set*** *with respect to (**Y**, **X**).*

Assumption 1 implies that, for an invariant set S, it holds $Y_t|X_t^S \sim \mathcal{N}(\mu + X_t^S\beta, \sigma^2)$ across $t = 1, 2, ..., n$. For a given X_t^S, the noise terms ϵ_t dictate the variability of the conditional distribution as all the other terms are constant.

Note that Assumption 1 makes no claims about causality between Y and S. In fact, S may not necessarily be unique. Given the invariance of SCMs under interventions as shown in Section 3.5, however, the set of parents of Y form a valid invariant set S. Since Assumption 1 does not restrict the distribution of X_t^S, any type of intervention on the predictor variables in an SCM can be performed, with the goal of finding the ones that are invariant under interventions. However, since we are interested in discovering an invariant model for Y, the method does not allow for interventions on the target variable. This limitation explains why this method cannot be used to recover the entire causal structure, but only to find the causes of a variable of interest.

Under Assumption 1, our goal is to estimate an invariant set S from the observed data (**Y**, **X**). Pfister et al. [2019] define this esimate as the intersection of all sets $S \subseteq \{1, 2, ..., d\}$ detected as invariant with respect to (**Y**, **X**) through hypothesis testing. Given a set S, it tests the null hypothesis

$H_{0,S}$: S is an invariant set with respect to $(\mathbf{Y}, \mathbf{X})$.

The set of plausible causal predictors is then obtained as the intersection of all sets S for which $H_{0,S}$ is not rejected at a specified significance level α:

$$\hat{S} := \bigcap_{S \subseteq \{1,2,...,d\}} \{S : H_{0,S} \text{ is not rejected at } \alpha\}. \tag{4.1}$$

Peters et al. [2016] demonstrate that this definition of $\hat{S}$ guarantees it is a subset of the true invariant set S^* with high probability (Refer to [Peters et al., 2016] for the proof).

Theorem 3 *Assume that the estimator $\hat{S}$ is constructed according to Definition 4.1 with a valid test for $H_{0,S}$ for all sets $S \subseteq \{1, 2, ..., d\}$ at level α in the sense that for all S, $sup_{P:H_{0,S} \text{ not rejected}} P[H_{0,S} rejected] \leq \alpha$. Consider now a distribution P over (**Y**, **X**) and consider any S^* such that Assumption 1 holds. Then, $\hat{S}$ satisfies $P[\hat{S} \subseteq S^*] \geq 1 - \alpha$.*

Additionally, Peters et al. [2016] proves identifiability of the causal predictors (i.e. $\hat{S} = PA_Y$ where PA_Y is the true causes of Y) under certain conditions.

We believe Definition 4.1 is intuitively reasonable because variables consistently present across all invariant sets are more likely to be the true invariant causal predictors. Furthermore, the definition adheres to the principle of causal minimality, advocating for the simplest causal structure by including only the essential variables needed to accurately describe the causal relationships among the variables [Pearl, 2009]. By focusing on a minimal set of variables that remain invariant over time, this approach aligns with Occam's Razor, which favors simpler models as long as they adequately capture the underlying phenomenon [Pearl, 2009].

We acknowledge that such a conservative approach may incur false negatives (i.e. some parents of Y may be missed) in practice. However, our goal puts more emphasis on finding a set of variables useful for predicting Y that are robust to distributional shifts, rather than recovering the exact set of parents of Y, as long as they are invariant across time.

Pfister et al. [2019] also discuss the method's robustness to the presence of hidden variables. In the settings with an arbitrary set of hidden variables that does not include direct causes of Y, the plausible causal set estimator $\hat{S}$ still satisfies the coverage property $P[\hat{S} \subseteq AN_Y] \geq 1 - \alpha$, where AN_Y is the set of all ancestors of Y in the causal graph.

4.2 Hypothesis Testing for Invariance

We now introduce the method for testing the null hypothesis $H_{0,S}$. Pfister et al. [2019] propose a class of tests under the assumption that a linear Gaussian model, including a linear Gaussian SCM, exists for an invariant set (See Assumption 1). This assumption allows using a goodness of fit test of the linear Gaussian model as a test for $H_{0,S}$, treating the causal coefficients in Equation 3.4 as identical to the regression coefficients in the linear model for Y given X_t^S. Recall that Theorem 2 provides a graphical criterion for estimating the direct causal effects of a variable on another variable using regression coefficients. Pfister et al. [2019] formulate the null hypothesis $H_{0,S}$ as a test of the following linear regression model in matrix notation:

Definition 8

$$H_{0,S} : \exists \beta \in (\mathbb{R} \setminus \{0\}), \sigma \in (0, \infty) : \boldsymbol{Y} = \boldsymbol{X}^S \beta + \epsilon,$$

with $\epsilon \perp\!\!\!\perp \boldsymbol{X}^S$ *and* $\epsilon \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{Id})$.

The method is based on the principle that the residuals of the linear regression model should be i.i.d. Gaussian distributed under the null hypothesis. In other words, we can check for invariance of the conditional distribution of Y_t given X_t^S across $t = 1, 2, ..., n$ by testing whether the residuals of the linear regression model are i.i.d. Gaussian distributed. Pfister et al. [2019] construct a test statistic based on the scaled residuals.

Recall that the Ordinary Least Squares (OLS) estimator for β is given by

$$\hat{\beta} = ((\mathbf{X}^S)^\top \mathbf{X}^S)^{-1} (\mathbf{X}^S)^\top \mathbf{Y},$$

where $\mathbf{P}_\mathbf{X}^S := ((\mathbf{X}^S)^\top \mathbf{X}^S)^{-1} (\mathbf{X}^S)^\top$ is the projection matrix that projects Y onto the column space of X, resulting in the predicted values of $\hat{Y}$:

$$\hat{Y} = \mathbf{X}^S \hat{\beta} = \mathbf{X}^S \mathbf{P}_\mathbf{X}^S \mathbf{Y}.$$

Hence, the residuals are given by $\mathbf{R}^S := (\mathbf{Id} - \mathbf{P}_\mathbf{X}^S)\mathbf{Y}$. Under $H_{0,S}$, the scaled residuals $\tilde{\mathbf{R}}^S := \mathbf{R}/\|\mathbf{R}^S\|_2$ are then given by

$$\tilde{\mathbf{R}}^S := \frac{(\mathbf{Id} - \mathbf{P}_\mathbf{X}^S)\mathbf{Y}}{\|(\mathbf{Id} - \mathbf{P}_\mathbf{X}^S)\mathbf{Y}\|_2} = \frac{(\mathbf{Id} - \mathbf{P}_\mathbf{X}^S)\epsilon}{\|(\mathbf{Id} - \mathbf{P}_\mathbf{X}^S)\epsilon\|_2} = \frac{(\mathbf{Id} - \mathbf{P}_\mathbf{X}^S)\tilde{\epsilon}}{\|(\mathbf{Id} - \mathbf{P}_\mathbf{X}^S)\tilde{\epsilon}\|_2},$$

where $\tilde{\epsilon} := \epsilon/\|\epsilon\|_2$ is the scaled noise.

Assume there is a function $T : \mathbb{R}^n \to \mathbb{R}$ that maps the scaled residuals to a test statistic. In Section 4.2.1, we describe possible choices for T that allow us to detect the shifts in mean or variance of the residual distribution. The null distribution of $T(\tilde{\mathbf{R}}^S)$ is then approximated by either permuting the scaled residuals or bootstrapping the residuals under the null hypothesis, taking into account that the scaled noise $\tilde{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{Id})$. The p-value of the test is then obtained by comparing the observed test statistic to the null distribution. If the p-value is greater than the specified significance level α, we fail to reject $H_{0,S}$ and conclude that S may be a potential invariant set. It is important to note that while null hypotheses usually argue for the absence of an effect, $H_{0,S}$ argues for the existence of an invariant set (See Definition 8).

The method evaluates the causal coefficients, β, providing both p-values and confidence intervals.

When assessing the coefficient of the variable X_t^j, it takes into account the p-values from all fitted models that include this variable. Should the maximal p-value among these be less than the predetermined significance level α (i.e. none of the models containing the variable are invariant), the coefficient corresponding to this variable is set to 0, and neither p-values nor confidence intervals are computed for it.

Conversely, if the maximal p-value exceeds α, suggesting the inclusion of X_t^j in at least one invariant set, the p-value assigned to its coefficient is determined as the highest among the models that exclude it. This rationale is grounded in the definition of an invariant set (See Definition 4.1): if a variable X_t^j is part of the estimated invariant set, then it logically follows that all models excluding it must be rejected. In such cases, the regression coefficient of the variable is assigned a significant p-value.

The method does not provide p-values for lagged terms of S that represent non-instantaneous effects. Recall that definition 8 of $H_{0,S}$ focuses on the instantaneous effects of S on $\mathbf{Y}$ since $\mathbf{X}^S = (X_1^S, ..., X_n^S) \in \mathbb{R}^{n \times |S|}$ and $\mathbf{Y} \in \mathbb{R}^{n \times 1}$. However, the method still includes all lagged terms in fitting the OLS model, thereby returning coefficients and confidence intervals for the past terms as well. Let $\mathbf{K} \in \mathbb{R}^{(t-k)\times(k+1)|S|}$ be a design matrix of both instantaneous and lagged terms $X_{t-q}^S \in \mathbb{R}^{(t-q)\times|S|}$ for $q = 0, 1, ..., k$ (For $\mathbf{K}$, we drop the first k missing rows across all variables due to lagging). $(1-\alpha)$-confidence intervals for their coefficients β are given by $\hat{\beta}_{\mathbf{K}} \pm t_{1-\alpha/(2|\mathbf{K}|),n-|\mathbf{K}|-1}\hat{\sigma} diag(((\mathbf{K})^\top \mathbf{K})^{-1})$ as in OLS, where $t_{1-\alpha/(2|\mathbf{K}|),n-|\mathbf{K}|-1}$ is the $(1-\alpha/(2|\mathbf{K}|))$-quantile of the t-distribution with $n - |\mathbf{K}| - 1$ degrees of freedom and $\hat{\sigma}^2$ is the estimated residual variance [Peters et al., 2016]. Since we also consider the lagged terms as potential causes of Y (See Equation 3.4), we use confidence intervals, in place of p-values, as our proxy for the evidence for an invariant relationship between Y and the lagged terms of S. If the confidence intervals for the lagged terms contain 0, we conclude that the lagged terms are not invariant with respect to Y.

4.2.1 Constructing Test Statistics for Invariance to Interventions

We now introduce test statistics that are capable of detecting (non-)invariances. We consider two types of violation of invariance that can occur under $H_{0,S}$: a shift in the mean and a shift in the variance of the residual distribution. We simulate these violations by performing interventions on the predictor variables.

Pfister et al. [2019] introduce the concept of an underlying change point model, wherein violations in the invariance (i.e. distributional shifts) occur at specific time points. At each change point, the time series is divided into segments, or "environments," each of which is characterized by a distinct interventional distribution. Then, test statistics are constructed such that they simultaneously test for invariances over all potential environments. Multiple testing is accounted for by adjusting the p-values using the Bonferroni correction.

In practice, however, true change points are unknown. Pfister et al. [2019] argue that one can simply split the time series into equally-spaced blocks, although it is also possible to incorporate some prior knowledge about the change points. Pfister et al. [2019] suggest that consistency in identifying the invariant set is achievable when the as long as the size of the set of change points is of the order of $log(n)$ for a time series of length n and the size of each environment tends to infinity as n gets large.

Given a set of change points $CP = (g_1, ..., g_m)$ where $0 < g_1 < ... < g_m < n$, we define the i-th environment e_i as the time points between g_{i-1} and g_i for $i = 1, ..., m$ (for $i = 1, e_i = \{0, ..., g_i\}$; for $i = m, e_i = \{g_i, ..., g_n\}$). The collection of environment is denoted as $\mathcal{E} = \bigcup_{i \in \{1,...,m\}} e_i$. Pfister et al. [2019] proposes that a computationally efficient way of comparing the test statistics across the environments is to compare each environment against its complement.

$$\mathcal{F} := \{(e, f) \in \mathcal{P}(\{1, ..., n\})^2 : e \in \mathcal{E} \text{ and } f = \{1, ..., n\} \setminus e\}.$$

For all $e, f \in \mathcal{F}$, we construct test statistics $T^i_{e,f}$ that detect differences in the residual distribution between environments e and f. We then combine the all pairwise comparisons to obtain a single test statistic by

$$T_i^{max,\mathcal{F}}(\tilde{\mathbf{R}}^S) := \max_{(e,f)\in\mathcal{F}} |T^i_{e,f}(\tilde{\mathbf{R}}^S)| \text{ or } T_i^{sum,\mathcal{F}}(\tilde{\mathbf{R}}^S) := \sum_{(e,f \in \mathcal{F})} |T^i_{e,f}(\tilde{\mathbf{R}}^S)|.$$

Pfister et al. [2019] introduce various types of test statistics that can detect either block-wise or gradual shifts in the residuals. Here we only explain the one detecting block-wise shifts, but the others can be found in their paper. There are two types of violations of invariance that can occur under $H_{0,S}$ in definition 8:

(a) A difference in the regression coefficients: $\beta_{e,S} \neq \beta_{f,S}$

(b) A difference in the noise variance: $\sigma^2_{e,S} \neq \sigma^2_{f,S}$

where $\beta_{e,S}$, $\sigma^2_{e,S}$, $\beta_{f,S}$, $\sigma^2_{f,S}$ are the population regression coefficients and the noise variance in environments e, f when regressing $\mathbf{Y}$ on $\mathbf{X}^S$. Pfister et al. [2019] argue that both types of violations can be detected by regressing the scaled residuals $\tilde{\mathbf{R}}^S$ on $\mathbf{X}^S$ for each environment e and f. For an environment $h \subseteq \{1, ..., n\}$, the regression coefficient and biased sample variance of the scaled residuals regressed on $\mathbf{X}^S_h$ are

$$\hat{\beta}_{h,S} = ((\mathbf{X}^S_h)^\top \mathbf{X}^S_h)^{-1} (\mathbf{X}^S_h)^\top \tilde{\mathbf{R}}^S_h,$$

$$\hat{\sigma}^2_{h,S} = \frac{(\mathbf{R}^S_h - \mathbf{X}^S_h \hat{\beta}_{h,S})^\top (\mathbf{R}^S_h - \mathbf{X}^S_h \hat{\beta}_{h,S})}{|h|}.$$

It is then possible to test for either of the two violations respectively using the test statistics

$$T^1_{e,f}(\tilde{\mathbf{R}}^S) = \|\hat{\beta}_{e,S} - \hat{\beta}_{f,S}\|_2 \text{ and } \quad T^2_{e,f}(\tilde{\mathbf{R}}^S) = \frac{\hat{\sigma}^2_{e,S}}{\hat{\sigma}^2_{f,S}} - 1$$

for differences in the regression coefficients and for differences in the variance of the noise. The two tests can then be combined with a Bonferroni correction.

Find Invariant Set

Input: Time series data (Y, X) where Y is $n \times 1$ target variable and X is $n \times d$ predictor variables

Choose: set of pairwise environments $\mathcal{E}$, significance level α

```
for $S \subseteq \{1, 2, ..., d\}$
    Set $H_{0,S}$: $S$ is an invariant set
    for $(e, f) \in \mathcal{E}$
        Fit Ordinary Least Squares regression to $Y_e$, $Y_f$ using $X^S_e$, $X^S_f$, respectively
        Compute the scaled residuals $\tilde{\mathbf{R}}^S_e$, $\tilde{\mathbf{R}}^S_f$
        Compute test statistic $T_{e,f}(\tilde{\mathbf{R}}^S)$ that measures the difference between the residuals
    Combine test statistics across all environments and compute p-value
    Reject $H_{0,S}$ if p-value $< \alpha$
Output: Estimated invariant set $\hat{S} = \bigcap_{S: H_{0,S} \text{ not rejected}} S$
```

Chapter 5

Experiments

5.1 Data

We evaluate the performance of our method on synthetic data generated from a linear Gaussian SCM. A synthetic DAG is randomly generated given a specified number of nodes and an edge probability that determines the density of the graph. The nodes represent different variables in our system. The time series data for each node is generated through ancestral sampling, following the causal structure. We allow both instantaneous and lagged effects in the data, up to a specified maximum lag. Instantaneous effects are ignored for the root nodes, and lagged effects include autoregressive effects. Hence, the data generation process entails moving through the time series step by step, where the value of each node is determined by both current and previous values of its parent nodes in the DAG, with a noise term added. The noise term for each variable at every time step is sampled from a Gaussian distribution. Interventions are applied to the data at random time steps to simulate the effect of external factors or deliberate manipulations on the system.

Our evaluation of the method's performance is twofold. First, we assess its ability to accurately identify the invariant causal predictors of a target variable by computationally stressing the method's capacity to detect the invariant mechanism under various conditions. Secondly, our assessment focuses on the method's predictive accuracy, especially in forecasting the target variable when faced with distributional shifts. We benchmark the performance of our method's forecasts against those produced by Ordinary Least

Squares (OLS) regression.

5.2 Causal Discovery

In this section, we evaluate the method's ability to correctly identify the invariant causal predictors of a target variable under various conditions. We ran 50 simulations for each of five graphs with different parameter settings. Except for Section 5.2.1, the number of nodes is fixed at 5. The majority of the experiments yielded perfect precision, recall, and F1 scores, indicating that the method was able to accurately identify the invariant set. Instead of reporting the average precision, recall, and F1 score, we report the proportion of perfect precision, recall, and F1 score across the simulations.

5.2.1 Number of nodes

Table 5.1: Proportion of Perfect Precision, Recall, and F1 Score by Number of Nodes

Number of Nodes	F1_prop	precision_prop	recall_prop
5	91.11	100.00	91.11
7	90.00	93.33	96.67
9	85.33	90.67	91.33
11	74.67	93.33	78.00

As the number of nodes increases, the fraction of simulations with perfect F1 score decreases. While the proportion of perfect precision remains relatively high across all number of nodes, recall is more sensitive to the number of nodes, as the number of possible combinations of variables increases exponentially with the number of nodes, making it more difficult to detect the true invariant set.

5.2.2 Sample Size

In this experiment, we vary the sample size of the data. The sensitivity of the method to the sample size is more pronounced in recall than in precision. The consistency of the fraction of perfect precision across different sample sizes in Table 5.2 confirms that the method behaves in line with theoretical guarantees for the proportion of false positives, maintaining it at the predetermined significance level. A significance

Table 5.2: Proportion of Perfect Precision, Recall, and F1 Score by Sample Size

Sample Size	F1_prop	precision_prop	recall_prop
50	7.78	95.56	7.78
100	55.56	95.56	58.89
150	81.11	98.89	82.22
200	90.00	100.00	90.00
500	90.00	98.89	91.11
1000	91.11	100.00	91.11

level dictates how stringent the criteria are for deciding whether an apparent causal relationship between variables may be due to chance. In essence, it sets the probability of making a Type I error, or having a false positive. Given the significance level, the expectation is that the method should limit the number of false positives to a rate consistent with this level. Thus, if the method is correctly calibrated to its significance level (in this case, 0.05), it should not produce more than 5% false positives over a long series of experiments under a null hypothesis. Indeed, this theoretical guarantee is observed in the precision column in Table 5.2.

5.2.3 Noise variance

Table 5.3: Proportion of Perfect Precision, Recall, and F1 Score by Noise Variance

Noise Variance	F1_prop	precision_prop	recall_prop
0.01	94.44	97.78	96.67
0.05	97.78	100.00	97.78
0.1	94.44	100.00	94.44
0.5	68.89	97.78	71.11
1	10.00	98.89	11.11
1.5	0.00	98.89	0.00

Table 5.3 shows that the method maintains a high level of precision across all levels of noise variance, illustrating its conservative approach by prioritizing precision over recall. However, as the noise variance increases, the method's ability to identify the true invariant set diminishes, leading to a higher incidence of false negatives. This trend is visually depicted in Figure 5.1, where the F1 score distribution's center

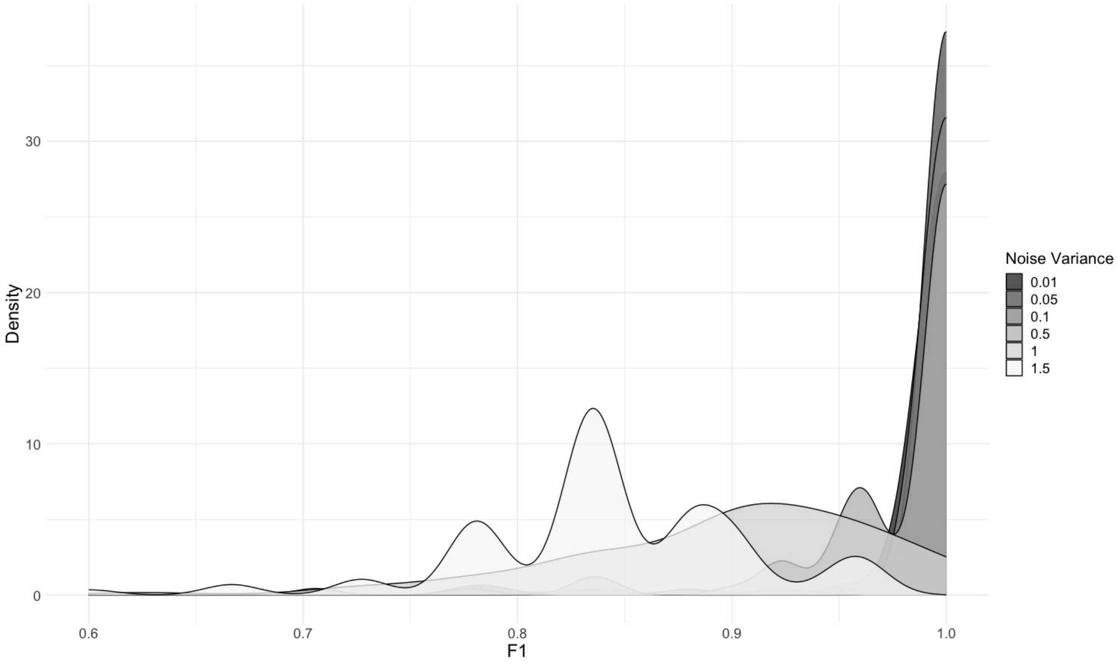

Figure 5.1: Distribution of F1 Score by Noise Variance

gradually approaches 1 with decreasing noise variance. Specifically, at a noise variance of $\sigma^2 = 1.5$, the F1 score distribution centers around 0.85 and is multimodal, highlighting the method's inconsistency in detecting the true invariant set under this condition. Conversely, at lower noise variances, the distribution becomes left-skewed towards 1, indicating a closer alignment with perfect precision and recall.

5.2.4 Graph Density

Table 5.4: Proportion of Perfect Precision, Recall, and F1 Score by Graph Density

Edge Probability	F1_prop	precision_prop	recall_prop
0.3	83.33	85.56	96.67
0.8	91.11	100.00	91.11

This experiment compares the method's performance under two graph densities. The proportion of perfect F1 score is higher for denser graphs. While the proportion of perfect precision is also higher for denser graphs, the proportion of perfect recall is higher for sparser graphs. This result suggests that dense graphs, by their nature, may inherently support higher precision, possibly because the likelihood of correctly identifying a parent increases with the number of available parents. Conversely, sparser graphs demonstrate a higher proportion of perfect recall, which could be attributed to their inherent sparsity facilitating the identification of true relationships due to fewer potential parents.

5.2.5 Causal effect

In this experiment, we manipulated the strength of the causal effects to assess its impact on our method's performance. The causal effects were varied as follows:

- weak: $\beta \sim \mathrm{U}(0.05, 0.2)$
- moderate: $\beta \sim \mathrm{U}(0.2, 0.6)$
- strong: $\beta \sim \mathrm{U}(0.6, 1)$
- mixed: $\beta \sim \mathrm{U}(0.05, 1)$

To accommodate these changes, we make the SCM sparser by removing some lagged terms of the target variable, aiming to reduce multicollinearity. This adjustment, however, generally decreases the method's performance, as indicated by an increased frequency of false positives—a trend we observe in the previous section (See Table 5.4).

Table 5.5: Proportion of Perfect Precision, Recall, and F1 Score by Causal Effect Strength

Causal Effect Strength	F1_prop	precision_prop	recall_prop
weak	66.00	76.00	80.67
moderate	72.67	75.33	94.00
strong	78.00	78.67	95.33
mixed	67.33	68.67	87.33

Our results demonstrate that both the proportions of perfect precision and perfect recall improve with the strength of the causal effects. This observation aligns with the expectation that identifying the true invariant set becomes more straightforward as the causal effects become more pronounced. Notably, while weak causal effects exhibit slightly higher precision than moderate effects, the difference is minimal. Furthermore, the method's performance drops significantly under mixed causal effects, underscoring the challenge of dealing with varied intensities of causal relationships.

5.2.6 Covariate shift

In this experiment, we leveraged the concept of covariate shift to more realistically simulate interventions on the input variables' distribution. Initially, causal effects are determined by sampling from a uniform distribution, $\beta \sim \mathrm{U}(0.3, 0.5)$. We then modify the strength of these causal effects using a multiplicative factor, aiming to evaluate the method's ability to accurately identify the true invariant set in these varied settings.

Table 5.6: Proportion of Perfect Precision, Recall, and F1 Score by Causal Effect Change Factor

Causal Effect Change Factor	Min	Q1	Median	Q3	Max
0.1	0.25	0.55	0.62	0.71	1.00
0.5	0.22	0.51	0.60	0.69	0.86
1	0.29	0.53	0.59	0.67	0.80
1.5	0.27	0.53	0.61	0.67	0.86
2	0.17	0.53	0.62	0.71	0.92
4	0.36	0.71	0.80	0.86	1.00
6	0.44	0.78	0.87	0.91	1.00
8	0.15	0.78	0.89	0.95	1.00

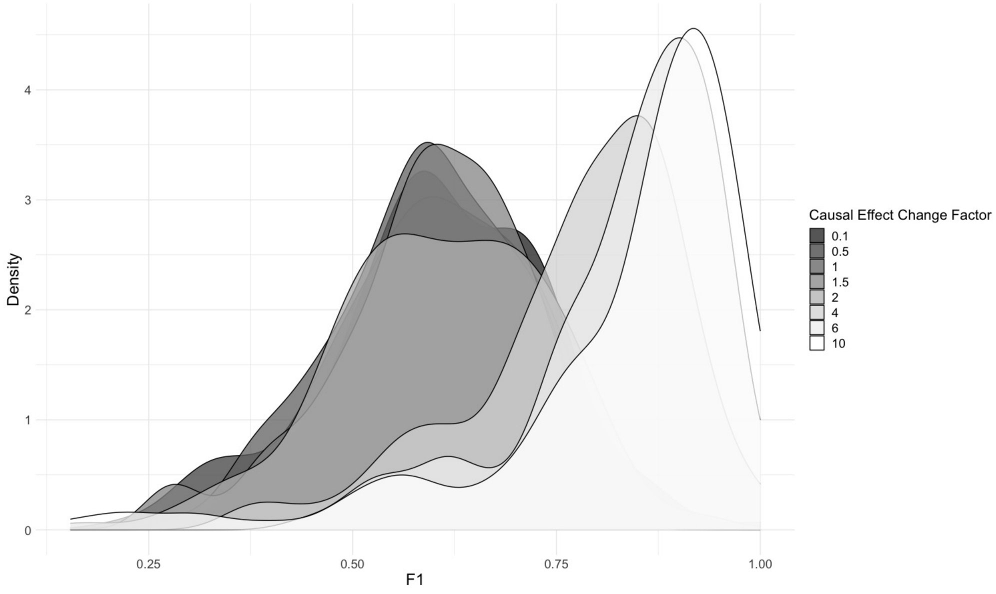

Figure 5.2: Distribution of F1 Score by Causal Effect Change Factor

In Table 5.6, change factor greater than 1 signifies stronger causal effects, which in turn produce more unusual values in the data, thereby making the interventions more conspicuous. The distribution of F1 is divided into two groups: those with a change factor ≤ 2 and those with a change factor > 2 as illustrated in Figure 5.2. The former is roughly symmetric and centered around 0.6. The latter group has higher median F1 score and is more right-skewed, indicating that the method is more consistent in detecting the true invariant set when the causal effects increase by a larger factor.

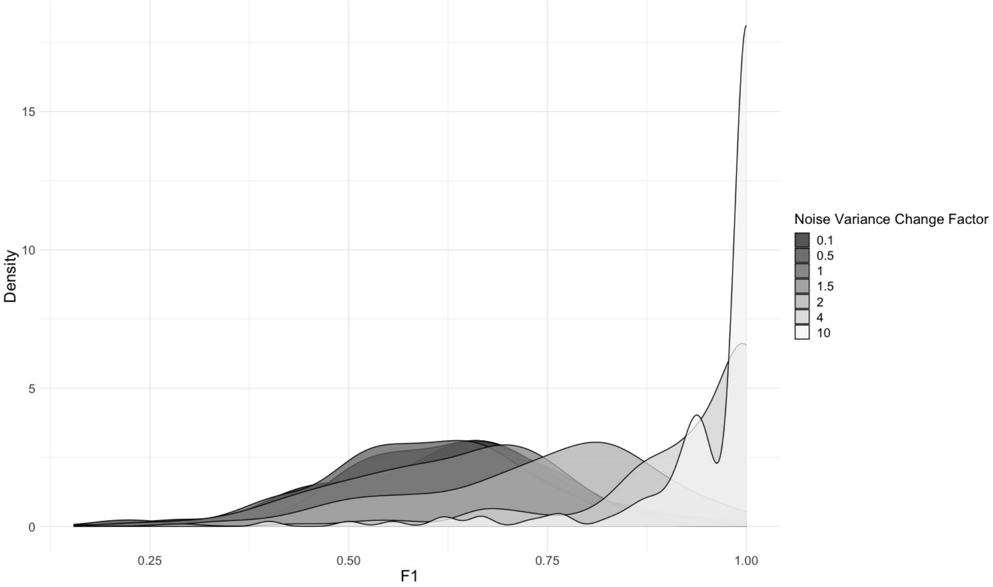

Figure 5.3: Distribution of F1 Score by Noise Variance Change Factor

Similar to the previous experiment, we changed the noise variances by a multiplicative factor to simulate the effect of external interventions on the distribution of the noise terms. Prior to interventions, the data was generated with noise variance $\sigma^2 = 0.01$. As the noise variance change factor increases, the distribution of F1 score shifts toward 1. This result indicates that the method is more consistent in detecting the true invariant set when the noise variance increases by a larger factor as a result of the interventions.

5.2.7 Hidden Variables

This section evaluates the method's robustness to the presence of hidden variables. If direct causes of a target variable are hidden from the data, the true invariant causal set would not be identifiable because the true causes compose the minimal set that forms the causal mechanism of the target variable under the principle of causal minimality [Pearl, 2009]. Recall from Definition 8 that the null hypothesis $H_{0,S}$ states that the set of variables S is an invariant set. We thus expect the method to reject the null hypothesis for any set of observed predictor variables.

Table 5.7: Proportion of Simulations with Model Rejection by Number of Hidden Causes

Number of Hidden Direct Causes	prop_rejected
1	1
2	1
3	1

In this experiment, we vary the number of hidden direct causes from 1 to 3 for a target variable with 4 direct causes. Table 5.7 shows that the method consistently rejects the null hypothesis for any set of observed variables in all simulations. In other words, despite the presence of hidden variables, the method does not falsely identify an invariant set when the true invariant set is not present in the observed variables. This result confirms the method's robustness to hidden variables, but it comes at the cost of a loss of detection power [Pfister et al., 2019].

5.3 Nonstationary Time Series Forecasting Performance

In this section, we evaluate the method's performance in forecasting the target variable in the presence of distributional shifts in the test data. We compare the Mean Squared Error (MSE) of the forecasts made by the method to the that of the forecasts made by OLS regression and OLS with Least Absolute Shrinkage and Selection Operator (LASSO) regularization. For each experiment, we report the median and the InterQuartile Range (IQR) of MSE across 200 simulations for a fixed graph structure with 5 nodes and edge probability 0.8.

5.3.1 Least Absolute Shrinkage and Selection Operator (LASSO) Regression

LASSO regression is a type of linear regression that adds a regularization term to the loss function [Tibshirani, 1996]. This regularization term is the sum of absolute values of the regression coefficients (i.e. the L1 norm of the coefficients). LASSO minimizes the following objective function:

$$\sum_{i}^{n}(y_i - \beta_0 - \sum_{j}\beta_j x_{ij})^2 + \lambda\|\beta\|_1. \tag{5.1}$$

The effect of this regularization is to reduce some regression coefficients to zero, selecting a more parsimonious model that does not include the corresponding predictors. In the shrinkage process, predictors that are less associated with the target are more likely to be shrunk to zero. This property makes LASSO regression particularly useful for models that contain a large number of predictors or suffer from multicollinearity. The choice of the regularization parameter λ dictates how many predictors are selected. A higher λ means more regularization, leading to more coefficients being set to zero. It is important to choose the right λ that balances the bias-variance trade-off.

In the following experiments, we compare the performance of our method to that of OLS and LASSO regression in forecasting the target variable in the presence of distributional shifts. Kitchen sink OLS regression is used as a baseline, and LASSO regression can be used to compare the performance of covariance-based variable selection to that of invariant causal predictors.

5.3.2 Sample complexity

Figure 5.4 shows the MSE of the forecasts made by OLS, LASSO, and the causal model across different sample sizes. LASSO clearly performs worst among the three methods, with a highest median MSE and widest IQR across all sample sizes. This result suggests that covariance-based variable selection may not be as effective as causal-based methods in the presence of nonstationarity.

We zoom in on the performance of OLS and SCM in Figure 5.5 to better compare their performance. The median MSE is consistent around 1 across different sample sizes when invariant causal predictors are used, indicating that the method's performance is not sensitive to the sample size. In other words, the method is able to make accurate forecasts even with a small sample size. For OLS, not only is the median MSE higher than that of the causal model, but it also suffers from a higher variance in MSE, as indicated by the wider IQR. This result suggests that our method is more robust to distributional shifts in the data than OLS, showing more stable forecasting performance across different sample sizes. It is important to note, however, that our method might perform poorly when it fails to find the true invariant set.

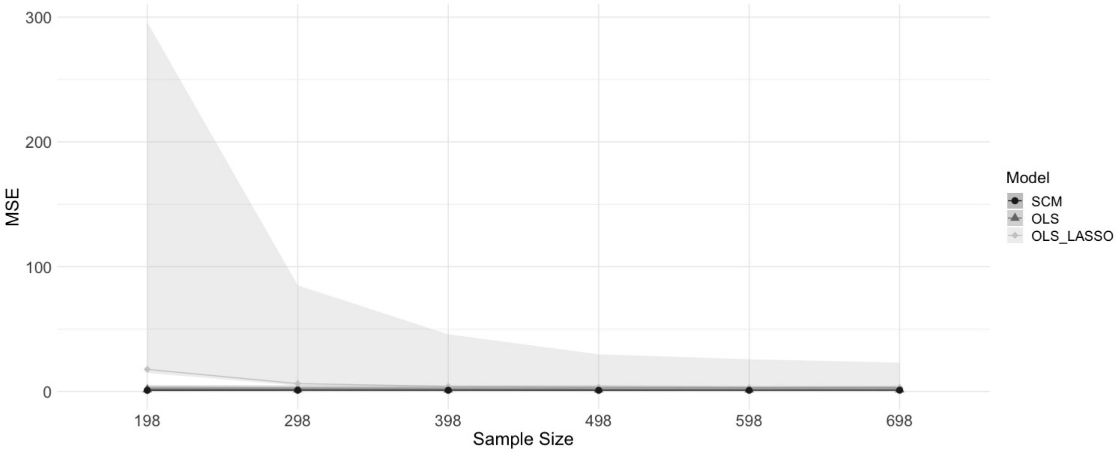

Figure 5.4: Median MSE by Sample Size with Q1 and Q3 Shaded

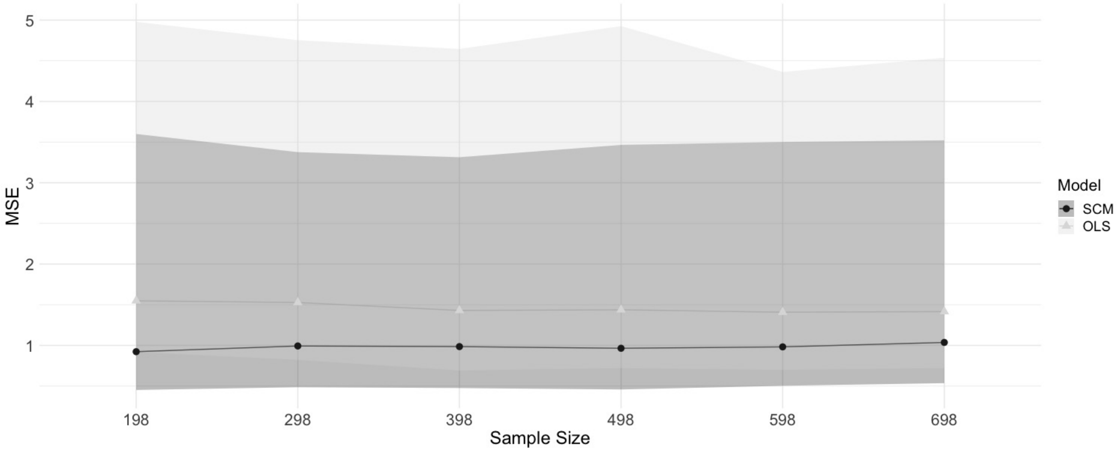

Figure 5.5: Median MSE by Sample Size with Q1 and Q3 Shaded, OLS and SCM Only

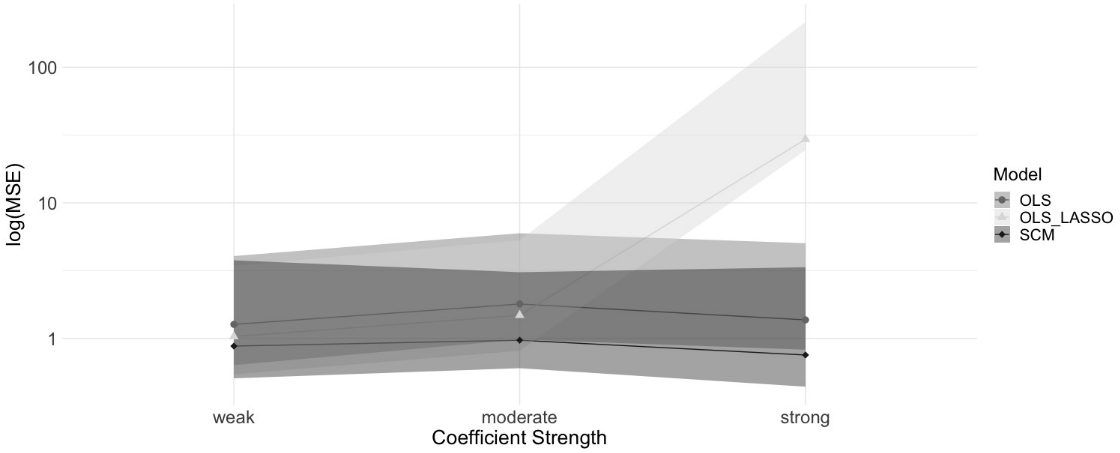

Figure 5.6: Median MSE by Causal Effect Strength with Q1 and Q3 Shaded (Log Scale)

5.3.3 Causal effect

This experiment tests how the strength of the causal effects affects the forecasting performance of three methods. In Figure 5.6, the MSE for OLS, LASSO, and the causal model is plotted in log scale across different causal effect strengths because the MSE for LASSO is too high to be displayed in the original scale. It clearly shows that LASSO performs worst among the three methods for strong causal effects, but slightly better than OLS for weak and moderate causal effects.

Figure 5.7 plots the MSE for OLS and the causal model in the original scale to compare the stability of the performance of the two methods. Both median MSE and IQR for OLS are higher than those of the causal model across all causal effect strengths. In other words, not only is the causal model more accurate in predicting the target variable, but it also shows more stable performance. For weak causal effects, however, our method suffers from a higher variance in MSE, as indicated by the slightly wider IQR, presumably because it is more prone to making false positives and false negatives as seen in Table 5.5.

5.3.4 Covariate Shift

This experiment tests how the change in noise variances affects the forecasting performance of three methods. Prior to covariate shifts, data was generated with noise variance $\sigma^2 = 0.01$. As in the previous experiments, LASSO performs worst among the three methods across all noise variance change factors, with significantly wider IQR than the other two methods (See Figure 5.8). The causal model has the lowest

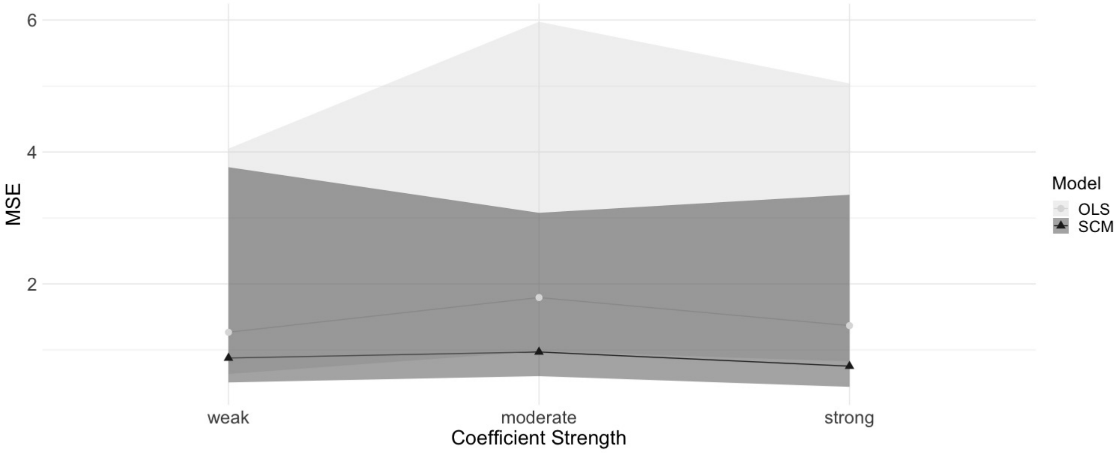

Figure 5.7: Median MSE by Causal Effect Strength with Q1 and Q3 Shaded, OLS and SCM Only

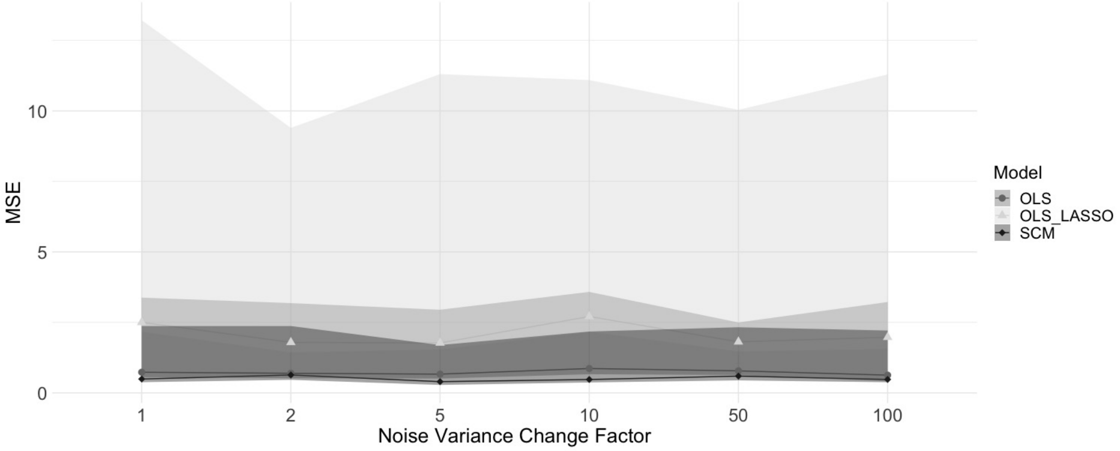

Figure 5.8: Median MSE by Noise Variance Change Factor with Q1 and Q3 Shaded

median MSE and IQR, indicating that it is most accurate and stable in forecasting the target variable. Thus, we conclude that our method is more robust to distributional shifts in the noise terms than OLS and LASSO regression.

Chapter 6

Conclusion

In this thesis, we discuss a method the finds the invariant causal predictors of a target variable of interest in a nonstationary time series using the invariance property of causal mechanisms. Our contribution lies in the strategic use of covariate shift to emulate realistic interventions on the input variables' distribution. This method marks a departure from the traditional reliance on hard interventions prevalent in causality research. Our findings reveal that using invariant predictors yields nonstationary forecasting outcomes that are notably more accurate and consistent than the results obtained via OLS and LASSO across sample complexity, strengths of causal effects, and degrees of covariate shift. Therefore, we argue that identifying the causal mechanisms described by invariant sets holds significant practical value. It ensures the models' stable performance in out-of-distribution generalization, a key factor in enhancing the reliability of machine learning models.

We argue that the concept of causal invariance paves the way for an alternative machine learning paradigm, overcoming the limitations imposed by the No Free Lunch theorem. The theorem states that for every learning algorithm, there exists a distribution over the data on which it performs poorly [Shalev-Shwartz and Ben-David, 2014]. From a causal perspective, this theorem translates to the existence of a causal structure of the data on which the learner falters. Yet, this statement does not hold up against our empirical evidence, which shows that leveraging invariant causal predictors significantly enhances the accuracy of forecasting outcomes. Essentially, causal structures act as strong and effective prior knowledge, eliminating the distributions that are problematic for all learners and thus improving generalization capabilities. This insight highlights causality’s contribution to enhancing the robustness and reliability of machine learning models. In future work, we aim to extend our method to handle non-Gaussian noise and nonlinear relationships between variables and to explore the method’s performance in real-world datasets with ground truth causal structures.

Bibliography

Chang Gong, Di Yao, Chuzhe Zhang, Wenbin Li, and Jingping Bi. Causal discovery from temporal data: An overview and new perspectives. *arXiv preprint arXiv:2303.10112*, 2023.

Nan Rosemary Ke, Silvia Chiappa, Jane Wang, Anirudh Goyal, Jorg Bornschein, Melanie Rey, Theophane Weber, Matthew Botvinic, Michael Mozer, and Danilo Jimenez Rezende. Learning to induce causal structure. *arXiv preprint arXiv:2204.04875*, 2022.

Meike Nauta, Doina Bucur, and Christin Seifert. Causal discovery with attention-based convolutional neural networks. *Machine Learning and Knowledge Extraction*, 1(1):19, 2019.

Judea Pearl. *Causality*. Cambridge university press, 2009.

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

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Causal inference on time series using restricted structural equation models. *Advances in neural information processing systems*, 26, 2013.

Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 78(5):947–1012, 2016.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

Niklas Pfister, Peter Bühlmann, and Jonas Peters. Invariant causal prediction for sequential data. *Journal of the American Statistical Association*, 114(527):1264–1276, 2019.

Jakob Runge, Peer Nowack, Marlene Kretschmer, Seth Flaxman, and Dino Sejdinovic. Detecting and quantifying causal associations in large nonlinear time series datasets. *Science advances*, 5(11):eaau4996, 2019.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

H.A. Simon. *Models of Discovery*. 3Island Press, 1979. ISBN 9789401095228. URL `https://books.google.com/books?id=RcwBswEACAAJ`.

Peter Spirtes, Clark N Glymour, and Richard Scheines. *Causation, prediction, and search*. MIT press, 2000.

Masashi Sugiyama and Motoaki Kawanabe. *Machine learning in non-stationary environments: Introduction to covariate shift adaptation*. MIT press, 2012.

Xiangyu Sun, Oliver Schulte, Guiliang Liu, and Pascal Poupart. Nts-notears: Learning nonparametric dbns with prior knowledge. *arXiv preprint arXiv:2109.04286*, 2021.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288, 1996.