FRIEDRICH-SCHILLER-
UNIVERSITÄT
JENA

Deep Networks for Time-Series Causality Analysis

PhD Thesis in Computer Science

submitted by

Wasim Ahmad

born April 1, 1990 in Mardan, Pakistan

written at

Computer Vision Group

Department of Mathematics and Computer Science

Friedrich-Schiller-Universität Jena

in cooperation with

Max-Planck-Institute (MPI)

07745, Jena

Germany

Advisor:	Prof. Dr.-Ing. Joachim Denzler
Started:	May 15, 2020
Finished:	June 30, 2025

Erklärung

Ich versichere, dass ich die vorliegende Arbeit ohne fremde Hilfe und ohne Benutzung anderer als der angegebenen Quellen angefertigt habe. Alle Ausfüührungen, die wörtlich oder sinngemäß übernommen wurden, sind als solche gekennzeichnet. Die vorliegende Arbeit wurde in gleicher oder ähnlicher Form noch keiner anderen Prüfungsbehörde vorgelegt und von dieser als Teil einer Prüfungsleistung angenommen.
Die Richtlinien des Lehrstuhls für Examensarbeiten habe ich gelesen und anerkannt, insbesondere die Regelung des Nutzungsrechts.

Jena, den 30. Juni 2025 Wasim Ahmad

Zusammenfassung

This study explores how deep learning can uncover causal relationships in complex, nonlinear time series. Real-world systems involve intricate variable interactions, making causal understanding crucial for informed decisions in various domains, i.e., climate science, healthcare and economics. Traditional methods for causality struggle with high-dimensional, nonlinear data, often missing subtle dependencies. Deep learning, however, excels at modeling complex patterns, offering a powerful approach to improve causal discovery and enhance our understanding of dynamic systems.

In the initial phase of my research, I aimed to improve Granger causality (GC) by incorporating deep learning and intervention-based techniques. GC assesses whether one time series X_t helps predict another Y_t, but its traditional form assumes linearity and struggles with high-dimensional, nonlinear systems. To overcome these limitations, I developed a novel approach combining deep learning's ability to model nonlinearities with controlled interventions using knockoff variables.

Knockoffs are synthetic variables that retain the statistical properties of the original data but remain uncorrelated with it. Introducing these variables enables controlled perturbations, allowing observation of counterfactual outcomes. Statistical hypothesis testing evaluates whether the post-intervention residual error increases, indicating a causal relationship. This method effectively handles nonlinear dependencies and incorporates counterfactual reasoning for improved causal inference.

Expanding on this foundation, I developed an advanced approach called causal discovery with model invariance (CDMI). This method is based on the principle of model invariance, which states that the model response Y should remain stable across different interventional settings E, as long as the true causal predictors C are preserved:

$$P(Y \perp E \mid C)$$

A violation of the invariance property upon an intervention indicates the presence of causal dependencies. A key assumption underlying CDMI is stationarity, meaning that the statistical properties of the system remain constant over time. However, real-world systems frequently exhibit non-stationarity, where data-generating processes evolve, posing significant challenges for causal inference.

To address this issue, we introduced a regime identification (RegID) technique. This technique identifies distinct regimes or segments in the data based on their statistical properties, i.e., covariance, with each regime representing a period of local stability. By applying the CDMI method within each identified regime, I was able to retrieve causal graphs that capture the relationships between variables in locally stable environments. This allows for a more accurate assessment of causality in multivariate nonlinear time series data, even in non-stationary systems.

Furthermore, to address the growing complexity of real-world systems, I extended CDMI to capture subsystem and group-wise causal interactions beyond traditional pairwise causality. This extension, termed group causal discovery with model invariance (gCDMI), introduces group-level interventions to assess the causal influence of one set of variables on another. This approach is particularly valuable in domains such as climate science, ecological systems, and neuroscience, where understanding interactions among variable groups provides deeper insights into system dynamics and answers a distinct causal query from pairwise analysis. For instance, in neuroscience, it can reveal how different brain regions interact to shape cognitive functions or behavior. By incorporating group-level interventions, gCDMI enables the discovery of causal relationships that may remain hidden when analyzing individual variables in isolation.

Additionally, I utilized multi-set canonical correlation (MCC) to enable pairwise causality methods, such as CDMI, PCMCI, and VAR, to uncover group-level causal relationships while addressing high-dimensionality challenges. Traditional pairwise methods are not inherently suited for discovering causal interactions among groups of variables, but MCC overcomes this limitation by representing each subsystem as a set of canonical variables. This approach not only reveals causal dependencies between groups but also reduces the dimensionality of the data, making it particularly useful in fields like neuroscience. For example, a single brain region may consist of a vast number of time series signals, and MCC helps condense this complexity, allowing for more efficient causal analysis.

To validate the proposed methods, I conducted comprehensive experiments using both synthetic and real-world time series datasets. The synthetic data were generated using structural causal models (SCMs) to create causal graphs with known relationships, enabling controlled evaluation. Additionally, real-world datasets such as FluxNet (climate-ecosystem interactions), NetSim (brain network simulations), MOXA (tectonic-climate relationships), and river discharges in Germany were used

to assess performance in practical settings. The results demonstrated substantial improvements in causal discovery, particularly in identifying complex relationships within multivariate nonlinear time series.

The integration of Knockoff-based interventions in trained deep learning models, combined with the principle of model invariance, emerged as a robust framework for enhancing causal analysis. This approach not only strengthens our capacity to detect cause-effect relationships in high-dimensional and nonlinear data but also sheds light on the underlying dynamics of multivariate time series. The versatility of these methods makes them applicable across various domains, such as climate science, and neuroscience. By advancing our understanding of causal mechanisms in complex systems, these tools support more informed decision-making and offer valuable insights into the intricate processes driving real-world phenomena.

Abstract

This research investigates the use of deep learning to infer causal relationships in multivariate nonlinear time series, driven by the complex interactions found in real-world systems across domains such as climate science, and healthcare. Traditional causality methods, including Granger causality and LiNGAM, often assume linear relationships and struggle with non-linear data, limiting their effectiveness in complex settings. To address these challenges, I developed a deep learning-based Granger causality framework that incorporates interventions, using Knockoff variables to introduce controlled perturbations and evaluate causal effects on model predictions. This approach enables a more refined estimation of causality by capturing nonlinear dependencies and incorporating counterfactual reasoning. Building on this foundation, I introduced the causal discovery via model invariance (CDMI) method, which leverages model invariance principles to identify causal relationships while handling non-stationary data through a regime identification (RegID) technique. To extend causal analysis beyond pair-wise variable interactions, I further enhanced CDMI, which employs group-wise interventions to uncover causal interactions between subsystems or groups of variables. This extension is particularly valuable in fields like climate science and neuroscience, where understanding group-level dynamics is crucial. Additionally, I integrated multi-set canonical correlation (MCC) with pairwise causality methods to enable the discovery of multi-group causal structures while reducing data dimensionality. This approach is especially useful in high-dimensional systems, such as brain networks, where a single region may consist of numerous time series signals. Extensive experiments on both synthetic and real-world datasets-including FluxNet (climate-ecosystem interactions), NetSim (brain network simulations), MOXA (tectonic-climate relationships), and river discharges time series-demonstrated substantial improvements in causal discovery. The results highlight the effectiveness of deep learning in capturing complex causal structures that traditional methods struggle to identify. Overall, this research introduces novel methodologies that enhance causal analysis, offering deeper insights into the dynamics of complex systems and supporting informed decision-making across various applications.

Contents

Chapter 1

Introduction

1.1 Causality

The concept of causality has long been at the core of human inquiry, representing our fundamental need to understand why things happen. From ancient philosophical debates to modern scientific exploration, the quest to uncover cause and effect relationships has shaped how we interpret the world. Causality isn't merely about identifying patterns; it is about discovering the underlying mechanisms that explain and predict phenomena ultimately support decision making.

In the 20th century, this philosophical inquiry was formalized mathematically by Wiener (1956) and further developed by Granger (1969) to study cause and effect between variables, particularly in econometric applications. Granger causality (GC) quantifies interactions between variables, identifying cause-effect relationships through modeling and prediction. It assesses the impact of incorporating past information from one variable (the cause) on the prediction of another variable (the effect), providing a more rigorous approach to understanding interactions beyond simple correlation.

In our work, we focus on uncovering the causal structure from observed multivariate time series data. This process, known as *causal discovery*, aims to identify the underlying causal relationships between variables in a system, often without prior knowledge of the mechanisms governing that system. Unlike traditional statistical methods that primarily reveal associations, causal discovery seeks to determine the direction of influence and the true cause-effect dynamics driving the observed data. Understanding causality is crucial across numerous fields ranging from natural sciences and economics to medicine and machine learning where grasping the underlying mechanisms enables more informed decision-making, policy formulation, and

predictive modeling. As systems grow more complex, the challenge of identifying causal relationships intensifies, making causal discovery an essential tool in modern science and technology.

In many fields, relying solely on statistical correlation between variables can be misleading. While correlation measures the strength of the association between two variables, it does not reveal the *direction* of influence or establish whether one variable *causes* changes in another. This limitation is captured by the well-known statistical principle: 'correlation does not imply causation.' For instance, if two time-series variables, X_t and Y_t, exhibit correlation, this relationship could be driven by an underlying confounding factor, Z_t, that simultaneously affects both X_t and Y_t. In such cases, any inferred connection between X_t and Y_t would be misleading, obscuring the actual causal structure governed by Z_t. Figure 1.1 illustrates this idea, demonstrating how a confounding variable can create an apparent but deceptive relationship between two observed variables.

Mathematically, the relationship between two variables X_t and Y_t can be quantified using the Pearson correlation coefficient Lee Rodgers and Nicewander (1988).

$$\rho_{XY} = \frac{\mathrm{Cov}(X_t, Y_t)}{\sigma_X \sigma_Y}$$

where $\mathrm{Cov}(X_t, Y_t)$ is the covariance between the two variables and σ_X and σ_Y are their respective standard deviations. While this equation quantifies the linear relationship between X_t and Y_t, it says nothing about which variable causes the other to change or whether the relationship is driven by a third factor. In contrast, causality seeks to uncover the functional form of the cause-effect relationship, such as:

$$Y_{t-\tau} := f(X_t, \eta_t)$$

where X_t influences $Y_{t-\tau}$ with functional relationship f and time delay τ, and η_t represents noise or unobserved factors. In this framework, establishing causality means identifying the correct direction of influence and ruling out spurious correlations caused by confounders or latent variables.

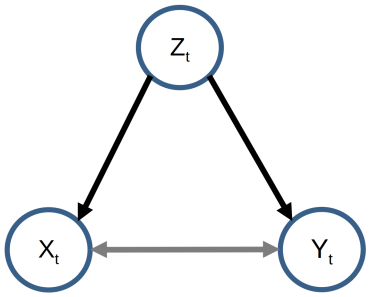


Figure 1.1: Common cause scenario in 3-variable directed acyclic graph (DAG).

1.2 Causal discovery and deep learning

Causal discovery seeks to identify cause-and-effect relationships from data, offering insights into underlying mechanisms beyond mere statistical associations. It employs various methods, including structural causal models (SCMs), which use directed acyclic graphs (DAGs) to represent causal relationships Pearl (2009a), constraint-based approaches that rely on conditional independence tests to infer causal structures Spirtes et al. (2000), and score-based techniques that optimize a predefined score to determine the most likely causal graph Chickering (2002). Meanwhile, deep learning excels at capturing complex patterns in data but primarily focuses on correlations rather than true causation, making models susceptible to biases and poor generalization in out-of-distribution settings Schölkopf et al. (2021).

Recent efforts have sought to integrate deep learning with causal discovery, aiming to make neural networks learn causal structures rather than just correlations. Approaches such as neural causal models Goudet et al. (2018), attention mechanisms Zheng et al. (2020), and reinforcement learning-based graph search Bechavod et al. (2021) have been explored to infer causal relationships from data. Additionally, causal regularization Arjovsky et al. (2019) and invariant learning Peters et al. (2016) have been proposed to encourage deep networks to learn stable causal relationships across different environments. However, enforcing causal constraints during training remains a significant challenge, as deep networks naturally tend to exploit spurious correlations rather than uncover true causal mechanisms.

In this work, we take an alternative approach by leveraging deep networks' structure-learning capabilities while bypassing the challenge of enforcing causal constraints during training. Instead of explicitly regularizing for causality, we conduct post hoc causal inference analysis by applying controlled interventions to the trained deep models in a sophisticated way and observing its responses. This allows us to probe learned representations and evaluate causal behavior without restricting the model's expressivity. Our framework aims to train deep networks for learning high-dimensional representations without explicit causal constraints and perform structured interventions to assess whether its internal representations exhibit causal properties. By perturbing inputs, and applying counterfactual reasoning Pearl et al. (2000), we aim to reveal meaningful causal dependencies.

This post hoc evaluation has broad applications. In healthcare, it helps assess treatment effects in deep models for personalized medicine Bica et al. (2021). In

economics, it aids policy evaluation and forecasting through causal interventions Athey and Imbens (2017). By shifting from enforcing causality during training to analyzing it post hoc, our approach bridges deep learning and causal inference. This framework harnesses deep networks' representational power while enabling causal interpretability through structured interventions, making it a promising path toward robust and generalizable AI systems.

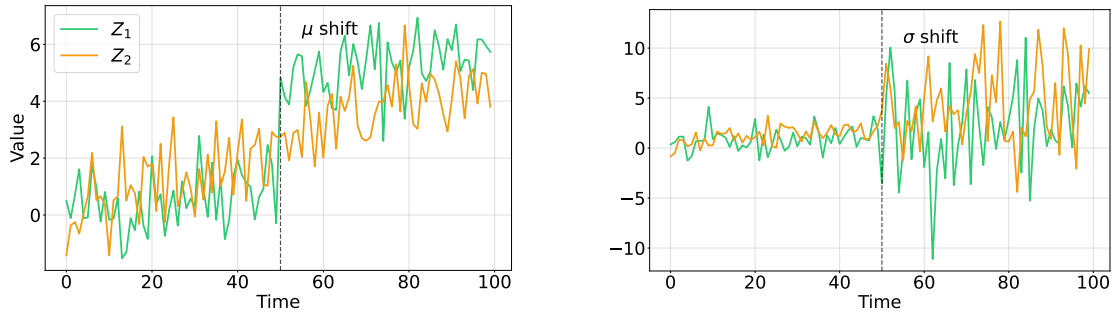
1.3 Research Challenges

Extracting causal relationships from observational data presents numerous difficulties due to both data-related and methodological complexities. These challenges include non-stationarity, high dimensionality, and confounding effects, as well as the intricate and dynamic interactions between variables. To deal with these issues, causal discovery often depends on key assumptions about the underlying data-generating process. Two fundamental assumptions are stationarity, which presumes that statistical properties remain stable over time, and causal sufficiency, which assumes that all relevant common causes are observed and that there are no unmeasured confounders. When these assumptions are violated, causal inference can become biased or unreliable. Overcoming these challenges requires the development of more sophisticated techniques that can extract meaningful causal insights from observational data while accounting for its inherent complexities.

In the following, we discuss some of the key challenges in causal discovery.

Non-Stationarity Non-stationarity arises when the statistical characteristics of data evolve over time, making causal inference more challenging. In fields such as economics and climate science, fundamental metrics like the mean, variance, and autocorrelation can fluctuate due to long-term trends, seasonal variations, or abrupt structural changes, as depicted in Figure 1.2. Such changes can distort causal relationships, potentially resulting in incorrect conclusions if not appropriately addressed. To address this issue, researchers have developed methods such as state-space models Durbin and Koopman (2012), time series decomposition Dagum (2010), and transfer learning approaches Pan (2010). These techniques aim to adapt to temporal variations in data, but their effectiveness depends on factors such as the quality of available data and computational efficiency. Developing more robust

approaches for handling non-stationarity remains an essential step toward improving causal inference in dynamic environments.



(a) (b)

Figure 1.2: Two types of non-stationarity in time series: (a) mean shift (μ) and (b) standard deviation shift (σ), illustrating how these changes affect the underlying data dynamics.

Hidden Confounders Confounders influence both treatment and outcome variables, leading to biased causal estimates. Hidden confounders, which are unmeasured variables affecting the system, introduce spurious correlations that obscure true causal relationships, as we shown in Figure 1.1. Mathematically, if Z is a hidden confounder influencing both X and Y, the observed correlation $\text{Corr}(X, Y)$ may not reflect the true causal effect $P(Y|do(X))$. Methods such as instrumental variables Angrist and Krueger (2001), and proxy-based approaches Trifunov et al. (2022) help address this issue, but they often require strong, unverifiable assumptions.

Nonlinear and Complex Relationships Many real-world causal interactions are nonlinear, making traditional linear models insufficient. Systems in biology, economics, and neuroscience exhibit intricate dependencies, requiring flexible functional forms to model causal effects. Machine learning methods, particularly deep learning Schölkopf et al. (2021) and kernel-based approaches Sejdinovic (2024), improve causal discovery in such settings but introduce challenges related to interpretability and generalization. Ensuring that these models capture causal mechanisms rather than spurious patterns remains an open research problem.

High Dimensionality Modern datasets, such as those in genomics, social networks, and sensor data, often contain thousands of variables, leading to the 'curse of dimensionality'. When the number of variables p far exceeds the number of observations n

($p \gg n$), causal discovery becomes computationally expensive and statistically unreliable. Dimensionality reduction techniques e.g., PCA Jolliffe and Cadima (2016), CCA Härdle et al. (2015) , sparsity-based regularization (e.g., LASSO Tibshirani (1996)), and constraint-based methods Spirtes et al. (2000, 2001) leverage assumptions like sparsity in causal graphs to improve tractability. However, selecting appropriate priors or constraints is critical to maintaining causal validity.

Temporal and Spatial Dynamics Causal relationships and their effects can change over time due to external shocks or evolving environments and may also vary across different spatial regions. Time-varying structural equation models (SEMs) Bollen (1989), dynamic Bayesian networks Murphy (2002), and RNNs Hochreiter and Schmidhuber (1997) capture these dependencies, but they require extensive data and complex inference, making them challenging with limited or noisy data.

Causal Complexity Many domains, including healthcare, ecosystems, and social sciences, involve complex causal structures with multivariate interactions, feedback loops, and hierarchical dependencies. For example, bidirectional influences between economic indicators or biological pathways create cyclical dependencies that traditional DAG-based methods struggle to represent. Advanced techniques such as structural causal models Pearl et al. (2016), variational inference Rezende and Mohamed (2015), and graph neural networks Kipf and Welling (2016) enable modeling of these complexities, but their scalability and robustness remain active areas of research.

Data Limitations Observational data is often the only available source for causal inference due to ethical, financial, or logistical constraints. However, such data is prone to biases from confounding, measurement errors, and missing values. Techniques like *do-calculus* Pearl (2022) provide a theoretical framework for identifying causal effects from observational data, while Bayesian methods Edition (2013) and imputation strategies Little and Rubin (1987) address uncertainty in incomplete datasets. Nonetheless, ensuring that causal conclusions remain valid in real-world settings requires careful study design, sensitivity analysis, and robustness checks.

1.4 Research objectives

The main objectives of this thesis is to advance causal discovery by addressing key challenges in complex systems, particularly those involving nonlinear dynamics, non-stationarity, and group-level interactions.

Non-linearity First, we aim to apply causal discovery to nonlinear systems, as traditional methods often assume linear relationships, limiting their applicability to real-world systems where variables interact in more complex ways. By focusing on systems with nonlinear interactions, this research seeks to uncover deeper and more intricate causal mechanisms that are otherwise obscured by linear approaches.

Non-Stationarity A key challenge in time-series data is non-stationarity, where statistical properties such as the mean (μ) and variance (σ^2) change over time, complicating causal analysis. Many existing approaches rely on the assumption of stationarity, which does not hold in many real-world scenarios, potentially leading to incorrect or incomplete causal conclusions. This thesis seeks to develop methods that can identify and adapt to regime shifts, enabling more accurate detection of evolving causal relationships over time.

Group-level interactions Additionally, this work addresses causal discovery at the group level, as many real-world systems exhibit interactions that operate between groups of variables rather than individual ones. For instance, in neuroscience, brain regions often work in networks where groups of neurons or brain areas collectively impact cognitive functions or behaviors, rather than individual neurons acting in isolation. By focusing on the collective behavior of groups, this thesis aims to reveal higher-order causal structures and interactions, which are crucial in complex domains such as climate science, and neuroscience. These objectives will provide a more nuanced understanding of causality in dynamic, high-dimensional environments.

1.5 Overview of the proposed approaches

My research on multivariate time series causal discovery introduces advanced methods that tackle key challenges in causality analysis. By leveraging deep learning, invariance testing, interventions, and counterfactual reasoning, these methods enhance the reliability of causal analysis and have been successfully applied to both real-world and synthetic datasets.

Granger causality using deep networks with Knockoff counterfactuals Granger causality Granger (1969) determines whether one time-series variable influences another by evaluating its contribution to predictive accuracy. If incorporating X_t improves the prediction of Y_t, X_t is deemed Granger causal for Y_t. Although commonly implemented using VAR models Sims (1980), these models perform well in linear settings but struggle with nonlinear dependencies. To address this limitation, we propose leveraging deep learning models, which excel at capturing intricate, nonlinear relationships in multivariate time series, offering a more robust approach to causal inference in complex systems.

In traditional Granger causality testing, the approach involves comparing the outputs of two models–one that includes the variable of interest and one that excludes it. However, this approach introduces complications, as comparing two independently trained deep networks can lead to inconsistencies. The networks internal weights adjust dynamically based on their inputs and training conditions, making it difficult to conduct a fair comparison. To overcome this challenge, we rather adopt an approach based on the notion of *do-calculus*, which involves intervening on a variable directly within a trained model.

Intervention, denoted as $\text{do}(X = x)$, allows us to actively change or *intervene* on a variable to see how it affects other variables in the system. It forces the variable X to take a specific value x, independent of its natural causes, allowing us to observe its effect on the rest of the system. Mathematically, the key idea is to compute the counterfactual distribution $P(Y \mid \text{do}(X = x))$, which represents the probability distribution of Y after intervening on X. This differs from the observational distribution $P(Y \mid X = x)$, which might be confounded by other factors.

For example, if we have a simple causal relationship like: $X \to Y$. The observational conditional probability $P(Y \mid X = x)$ might reflect correlations influenced by confounders. However, the interventional distribution $P(Y \mid \text{do}(X = x))$ removes those confounding influences, giving us the true causal effect of X on Y. Using do-calculus, we can express the causal effect of X on Y as:

$$P(Y \mid \text{do}(X = x)) = \sum_{Z} P(Y \mid X = x, Z)P(Z)$$

where Z represents all the other variables that might mediate or confound the relationship between X and Y. This approach allows us to simulate interventions

and observe their causal effect without the need to compare multiple networks, thus providing a more controlled and consistent framework for causality.

Our approach involves applying various types of interventions to the variable of interest. These interventions include substituting the variable X with its mean value $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, drawing from a uniform distribution $f_X(x) = \frac{1}{b-a}$ for $a \leq x \leq b$, introducing out-of-distribution (OOD) data X' where $P(X) \neq P(X')$, and using *Knockoff* variables $\tilde{X}$ Barber and Candès (2019). Knockoffs $\tilde{X}$ are synthetic variables designed to maintain the covariance structure Σ_X of the original variables X and generated independent of the system's output. Through the application of these different intervention methods, we systematically explore the influence of each variable on the model's predictions. Our findings reveal that while all types of interventions provide useful insights into the causal structure of the system, Knockoff variables consistently outperform the others in uncovering the true underlying causal relationships. This methodology has been tested on both synthetic and real-world datasets. Our approach demonstrates improved results compare to other causality methods.

Causal discovery using model invariance This method is based on model invariance, the principle that a model's response remains unchanged across different interventional environments, as long as the causal predictors of the target variable remain intact Peters et al. (2016). The mathematical formulation of model invariance is:

$$P(Y \perp E \mid C)$$

where Y is the model response, E represents the interventional environment, and C is the set of causal predictors. If an intervention disrupts a causal driver, the model's invariance property no longer holds, revealing the causal structure within multivariate time-series data.

We leverage this by learning the structure of the multivariate system using deep networks and exposing them to various interventional settings. Following previous work, we use the Knockoff framework to generate interventional variables. By assessing model invariance property through its residual distribution in observational and interventional settings, we evaluate causal relationships. A significant shift in these distributions, indicating a violation of the model's invariance property, suggests that the intervened variable has a causal effect on the target. This method, called causal discovery using model invariance (CDMI), effectively uncovers causal links and has

shown strong performance across both synthetic and real-world datasets, including those from environmental and tectonic domains.

Regime-wise causal discovery in non-stationary data Causal analysis in non-stationary time series presents a formidable challenge due to the constantly evolving nature of the underlying processes that govern the data. Unlike stationary systems, where statistical properties, i.e., mean μ and variance σ^2 remain constant over time t, non-stationary systems exhibit time-varying relationships, trends, and structural breaks, as illustrated in Figure 1.2. These changes can occur due to external factors, internal dynamics, or abrupt shifts in the system itself, leading to fluctuations in data distribution.

Identifying dynamic causal links is crucial for accurate inference, but traditional stationary methods often fail to capture shifting causal structures and transient effects. Non-stationary systems are further complicated by seasonal patterns, external interventions, and temporal variability, making it difficult to distinguish genuine causal relationships from transient effects, especially with limited data. Instead of static cause-effect discovery, the challenge lies in tracking how these relationships evolve over time. This necessitates advanced approaches like dynamic causal modeling or machine learning techniques that adapt to temporal changes.

To address non-stationary behavior in our work, we use a regime identification (RegID) method to segment multivariate time series into stable periods. This allows for more accurate causal analysis by applying our CDMI method within each regime to uncover regime-specific causal graphs. By analyzing evolving causal structures, we gain insights into temporal dynamics. Beyond CDMI, we also apply other causality methods in tandem with RegID, improving overall causal analysis in complex, evolving systems.

Causal discovery in groups of time series Traditional causality methods primarily focus on pairwise or node-level causal discovery, often neglecting the collective impact of groups or subsystems of variables on a target variable or another group of variables. In domains like neuroscience, understanding interactions between brain regions is essential for grasping overall system dynamics. Similarly, in climate science, subsystems interact in intricate ways, making group-level causal interactions crucial for capturing broader system behavior. Identifying causal direction within

variable groups requires uncovering the influence and directional dependencies between different components of complex systems.

To address this, we expanded our previous pairwise causal discovery approach to include group-wise causality through model invariance (gCDMI), where interventions are performed at the group level to evaluate their effects on other groups. Furthermore, we applied canonical correlation analysis (CCA) to reduce the dimensionality of the groups while maximizing the correlation between them. The resulting canonical variables were then used as inputs for pairwise causality methods such as CDMI, VAR-GC, and PCMCI Runge et al. (2019). CCA helps by condensing each group into a smaller set of canonical variables, which retain the most significant relationships at a lower dimension, ultimately enabling pairwise causality methods to capture group interactions more effectively.

Leveraging canonical variables in pairwise causality methods for group-level causal analysis has shown promising results on both synthetic and real-world datasets, such as NetSim, Fluxnet and Moxa. In practical applications, uncovering accurate causal structures provides deeper insights into brain region interactions during cognitive tasks, the influence of climate variables on ecosystems across different regimes, and the evolving relationship between tectonic activity and environmental factors over time.

1.6 Contribution of this research

- ✧ Granger causality with knockoff counterfactuals
- ✧ Causal discovery via model invariance
- ✧ Causal relationship in subsystems or groups of variables
- ✧ Causality analysis of non-stationary time series

1.7 Structure of the thesis

This thesis is organized into five chapters, each addressing key aspects of the research, methodologies, and applications of causal discovery in the context of deep learning and time series analysis. A brief outline of each chapter is provided below.

Chapter 1 (Introduction) The introductory chapter provides a foundation for the thesis by outlining the key concepts of causal discovery and deep learning. It highlights the research challenges and objectives that guide this work, followed by an overview of the proposed approaches. This chapter concludes with a summary of the contributions of the research and an outline of the thesis structure.

Chapter 2 (Theoretical Background) This chapter presents the fundamental concepts and theories that underpin the research. Key topics include the distinction between correlation and causation, the ladder of causation, Granger causality, structural causal models (SCMs), and other essential causal inference methodologies. In addition, it covers deep forecasting models, K-Means clustering, Riemannian geometry, and Knockoff generation for feature importance, and statistical hypothesis testing which form the basis of the proposed approaches.

Chapter 3 (Methodology) This chapter details the methodologies developed and employed in this thesis for causal discovery. It explores the application of Granger causality using DeepAR networks and Knockoff counterfactuals, followed by the introduction of CDMI. Methods for causal discovery in groups of variables, gCDMI, regime identification (RegID) in non-stationary systems, and related interventions using Knockoffs are also discussed in detail.

Chapter 4 (Applications and Experimental Results) This chapter demonstrates the practical application of the proposed methods through experiments on synthetic and real-world datasets. The chapter covers causal discovery in nonlinear systems, group-level causal interactions, and handling non-stationarity in synthetically generated time series data. Further, various experimental case studies are presented, including river discharges time series, brain region connectivity, and climate-ecosystem interactions, illustrating the effectiveness of the approaches.

Chapter 5 (Discussion and Conclusions) The final chapter summarizes the contributions of the thesis, highlighting its significance in advancing causal discovery methods. It discusses limitations such as hidden confounding and suggests avenues for future research. The chapter concludes by summarizing the overall findings and implications of this research.

Chapter 2

Theoretical background (13-55)

This chapter provides a formal introduction to the essential concepts, along with more advanced ideas, necessary to fully understand the contributions of this thesis. It is assumed that the reader has a background in Calculus, Linear Algebra, Statistics, Probability, Deep learning and Graph Theory throughout the manuscript.

2.1 Foundational concepts and definitions

In this section, we introduce the fundamental concepts required for understanding the core ideas explored in this work. These concepts are organized in a hierarchical structure, where more general concepts are defined first, followed by specific instances.

Definition 2.1.1 (Set)**.** A **set** is a group of distinct elements viewed collectively. A set can either be finite or infinite. For instance, the set of natural numbers, denoted as $\mathbb{N}$, is represented as $\{0, 1, 2, 3, \dots\}$.

Definition 2.1.2 (Element)**.** An **element** of a set is a single object that is a member of that set. For instance, in the set $A = \{1, 2, 3\}$, the number 2 is an element of A.

Definition 2.1.3 (Function)**.** A **function** is a relation that assigns each element from one set, known as the domain, to exactly one element in another set, referred to as the range. Formally, a function f from set A to set B is denoted as $f : A \to B$, where for every $a \in A$, there is a unique $b \in B$ such that $f(a) = b$.

Definition 2.1.4 (Distance)**.** A **distance** is a way of quantifying the separation between two elements within a specific space. Formally, a distance function $d : X \times X \to \mathbb{R}$ must satisfy several key properties, also known as axioms:

1. Non-negativity: For all $x, y \in X$, $d(x, y) \geq 0$, and $d(x, y) = 0$ if and only if $x = y$.
2. Symmetry: For all $x, y \in X$, $d(x, y) = d(y, x)$.
3. Triangle inequality: For all $x, y, z \in X$, $d(x, z) \leq d(x, y) + d(y, z)$.

Distances can be defined in various ways depending on the context, such as the Euclidean distance, Manhattan distance, or more abstract measures like Riemannian distances.

Definition 2.1.5 (Metric)**.** A **metric** is a specific type of distance function that defines a measure of distance between points in a space. A function d is a metric if it satisfies the following conditions for all points x, y, z in a set X:

1. $d(x, y) \geq 0$ with equality if and only if $x = y$ (identity of indiscernibles),
2. $d(x, y) = d(y, x)$ (symmetry),
3. $d(x, z) \leq d(x, y) + d(y, z)$ (triangle inequality).

Definition 2.1.6 (Vector)**.** A **vector** in an n-dimensional space is an ordered tuple of numbers, represented as:

$$\vec{v} = (v_1, v_2, \ldots, v_n)$$

where $v_1, v_2, \ldots, v_n$ are the components of the vector. The magnitude (or length) of the vector $\vec{v}$ is given by:

$$\|\vec{v}\| = \sqrt{v_1^2 + v_2^2 + \cdots + v_n^2}$$

The vector $\vec{v}$ has both a **magnitude** and a **direction**, where the magnitude represents its length and the direction is determined by the relative values of its components.

Definition 2.1.7 (Matrix)**.** A **matrix** is a rectangular array of numbers arranged in rows and columns. It is typically represented as:

$$A = \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \ldots & a_{mn} \end{pmatrix}$$

where a_{ij} represents the element in the i-th row and j-th column of the matrix, and m and n are the number of rows and columns, respectively. The matrix A is said to be of size $m \times n$, where m is the number of rows and n is the number of columns.

Definition 2.1.8 (Norm)**.** A **norm** on a vector space X over a field F is a function $\|\cdot\| : X \to \mathbb{R}$ that assigns a non-negative real number to each vector in the space and satisfies the following properties for all vectors $x, y \in X$ and scalar $\alpha \in F$:

1. $\|x\| \geq 0$, with equality if and only if $x = 0$ (non-negativity),
2. $\|\alpha x\| = |\alpha|\|x\|$ (absolute homogeneity),
3. $\|x + y\| \leq \|x\| + \|y\|$ (triangle inequality).

If $\|\cdot\|$ satisfies these properties, we say that it is a valid norm on X.

Definition 2.1.9 (Euclidean Norm)**.** The **Euclidean norm** (also known as the L2 norm) of a vector $x = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$ is defined as:

$$\|x\|_2 = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2}.$$

This is the most common norm used in many applications like machine learning and optimization.

Definition 2.1.10 (Manhattan Norm)**.** The **Manhattan norm** (also known as the L1 norm) of a vector $x = (x_1, x_2, \ldots, x_n)$ is defined as:

$$\|x\|_1 = |x_1| + |x_2| + \cdots + |x_n|.$$

This norm is used in sparse representations and in machine learning algorithms such as Lasso regression.

Definition 2.1.11 (Infinity Norm)**.** The **infinity norm** (also known as the max norm or inf norm) of a vector $x = (x_1, x_2, \ldots, x_n)$ is defined as:

$$\|x\|_\infty = \max_i |x_i|.$$

This norm is used in optimization problems where the goal is to minimize the maximum deviation of the vector components.

Definition 2.1.12 (Frobenius Norm)**.** The **Frobenius norm** is used for matrices and is analogous to the Euclidean norm for vectors. For a matrix $A \in \mathbb{R}^{m \times n}$, the Frobenius norm is defined as:

$$\|A\|_F = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} |a_{ij}|^2},$$

where a_{ij} are the entries of the matrix A.

Definition 2.1.13 (Euclidean Distance)**.** The **Euclidean distance** between two points $x = (x_1, x_2, \ldots, x_n)$ and $y = (y_1, y_2, \ldots, y_n)$ in $\mathbb{R}^n$ is given by:

$$d(x, y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2}.$$

Definition 2.1.14 (Manhattan Distance)**.** The **Manhattan distance** between two points $x = (x_1, x_2, \ldots, x_n)$ and $y = (y_1, y_2, \ldots, y_n)$ is defined as:

$$d(x, y) = \sum_{i=1}^{n} |x_i - y_i|.$$

Definition 2.1.15 (Cosine Similarity)**.** The **cosine similarity** between two vectors x and y is given by:

$$\text{cosine similarity}(x, y) = \frac{x \cdot y}{\|x\| \|y\|},$$

where $\cdot$ denotes the dot product and $\|x\|$ and $\|y\|$ are the magnitudes of x and y, respectively.

Definition 2.1.16 (Riemannian Distance)**.** The **Riemannian distance** is a generalized distance function that uses the geometry of curved spaces to define distances between points. It is given by:

$$d(x, y) = \inf_{\gamma} \int_a^b \|\dot{\gamma}(t)\| dt,$$

where $\gamma(t)$ is a curve connecting points x and y, and $\|\dot{\gamma}(t)\|$ is the norm of the tangent vector along γ.

Definition 2.1.17 (Random Variable)**.** A **random variable** X is a function that maps each outcome ω in the sample space Ω of a random experiment to a real

number $X(\omega) \in \mathbb{R}$. The random variable X assigns a numerical value to each possible outcome of the experiment.

- If X is discrete, it takes values from a countable set, and the probability distribution is described by a PMF.
- If X is continuous, it takes values from a continuous interval, and the probability distribution is described by a PDF.

Discrete case: $X : \Omega \to \mathbb{R}$, where Ω is the sample space, and $X(\omega)$ is the numerical value assigned to each outcome $\omega \in \Omega$.

Continuous case: $X : \Omega \to \mathbb{R}$, where Ω is typically a real interval, and X takes real values from this interval. The probability that X falls within a range $[a, b]$ is given by the integral of the PDF $f_X(x)$ over that range:

$$P(a \leq X \leq b) = \int_a^b f_X(x)\, dx$$

Definition 2.1.18 (Probability Distribution)**.** A **probability distribution** describes how the probabilities of a random variable are distributed over different possible values. It provides a mathematical rule that assigns probabilities to outcomes of a random experiment.

A function $P(X)$ is a valid probability distribution if it satisfies the following properties:

1. Non-negativity:
$$P(X) \geq 0, \quad \forall X$$
2. Total Probability:
   - For discrete random variables: $\sum_{x \in S} P(X = x) = 1$
   - For continuous random variables: $\int_{-\infty}^{\infty} f(x)\, dx = 1$
3. Probability Calculation:
   - Discrete case: The probability of a specific value is given by: $P(X = x) = p(x)$
   - Continuous case: The probability of an interval $[a, b]$ is given by: $P(a \leq X \leq b) = \int_a^b f(x)\, dx$

Probability distributions are classified into two major types:

1. Discrete Probability Distributions: A discrete probability distribution applies to a discrete random variable, where outcomes are countable and described by probability mass function (PMF). Examples are:

- Binomial Distribution: Models the number of successes in n independent Bernoulli trials, each with success probability p.

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, 2, \ldots, n$$

- Poisson Distribution: Represents the number of events occurring in a fixed interval of time or space with an average rate λ.

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}, \quad k = 0, 1, 2, \ldots$$

- Geometric Distribution: Describes the number of trials until the first success in repeated independent Bernoulli trials with success probability p.

$$P(X = k) = (1-p)^{k-1} p, \quad k = 1, 2, 3, \ldots$$

2. Continuous Probability Distributions: A continuous probability distribution applies to a continuous random variable, where outcomes are uncountable and described by a probability density function (PDF).

- Normal Distribution: Also called the Gaussian distribution, it describes many natural phenomena, characterized by mean μ and standard deviation σ.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad x \in \mathbb{R}$$

- Exponential Distribution: Models the time between events in a Poisson process with rate λ.

$$f(x) = \lambda e^{-\lambda x}, \quad x \geq 0$$

- Uniform Distribution: Describes a situation where all values in a range $[a, b]$ are equally likely.

$$f(x) = \begin{cases} \frac{1}{b-a}, & a \leq x \leq b \\ 0, & \text{otherwise} \end{cases}$$

Definition 2.1.19 (Likelihood)**. Likelihood** is a function that measures the probability of observing given data under a specific statistical model with unknown parameters. Unlike a probability distribution, which describes the probability of outcomes, likelihood quantifies how well a set of parameters explains the observed data. Mathematically, for a parameter θ and observed data $X = \{x_1, x_2, \ldots, x_n\}$, the likelihood function is defined as:

$$L(\theta \mid X) = P(X \mid \theta)$$

where $P(X \mid \theta)$ represents the probability of the observed data given the parameter θ. The maximum likelihood estimation (MLE) seeks the value of θ that maximizes $L(\theta \mid X)$.

Definition 2.1.20 (Mean)**.** The **mean** (or expected value) of a random variable X is defined as:

$$\mu_X = E[X] = \sum_{x \in \mathcal{X}} xP(X = x) \quad \text{(for discrete variables)}.$$

Definition 2.1.21 (Variance)**.** The **variance** of a random variable X measures the spread of its values around the mean, defined as:

$$\text{Var}(X) = E[(X - \mu_X)^2].$$

Definition 2.1.22 (Covariance)**.** The **covariance** between two random variables X and Y is defined as:

$$\text{Cov}(X, Y) = E[(X - \mu_X)(Y - \mu_Y)].$$

Definition 2.1.23 (Correlation)**.** The **correlation** between two random variables X and Y is the normalized covariance, given by:

$$\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y},$$

where $\text{Cov}(X, Y)$ is the covariance and σ_X, σ_Y are the standard deviations of X and Y, respectively.

The correlation value ranges from -1 to 1, which indicate the strength and direction of the linear relationship.:

- (1): Perfect positive linear relationship.
- (-1): Perfect negative linear relationship.
- (0): No linear relationship.

Definition 2.1.24 (Noise)**. Noise** refers to random fluctuations in data or signals that interfere with the intended information. It is often modeled as a random process with a specific probability distribution, such as Gaussian noise.

Mathematically, an observed signal Y with true signal X and additive noise N is given by:

$$Y = X + N$$

where N is a random variable representing noise. Different types of noise have distinct statistical properties and play key roles in signal processing and modeling.

- Gaussian Noise (Additive White Gaussian Noise): The noise follows a normal (Gaussian) distribution:

  $$N \sim \mathcal{N}(\mu, \sigma^2)$$

  where μ is the mean (often 0 for white noise) and σ^2 is the variance. It is common in communication systems, image processing, and electronics.

- White Noise: A random signal with equal intensity across all frequencies. Its autocorrelation function is a delta function:

  $$R_N(\tau) = \sigma^2 \delta(\tau)$$

  It is used in modeling random fluctuations in signals and financial data.

- Quantization Noise: Arises from rounding errors in analog-to-digital conversion (ADC). It is modeled as uniform noise within a range:

  $$N \sim U\left(-\frac{\Delta}{2}, \frac{\Delta}{2}\right)$$

  where Δ is the quantization step size. In machine learning (ML) and deep learning (DL), quantization noise occurs when reducing the precision of model weights (e.g., from FP32 to INT8), potentially affecting accuracy but improving efficiency.

- Impulse Noise (Salt-and-Pepper Noise): This appears as sudden spikes or drops in signal values. Commonly occurs in digital image processing and communication errors.

Definition 2.1.25 (Time Series)**.** A **time series** is a sequence of observations indexed by time, represented as $\{X_t\}_{t\in\mathbb{T}}$, where X_t is the value at time t. It can be deterministic (e.g., $X_t = at+b$) or stochastic (e.g., autoregressive: $X_t = \phi X_{t-1}+\epsilon_t$).

Definition 2.1.26 (Signal)**.** A **signal** is a function $S : \mathbb{T} \to \mathbb{R}$ that conveys information, where $\mathbb{T}$ represents time.

It is continuous if $\mathbb{T} = \mathbb{R}$ (real-valued time, e.g., audio signals) and **discrete** if $\mathbb{T} = \mathbb{Z}$ (integer-valued time, e.g., digital signals).

Definition 2.1.27 (Frequency of a Signal)**.** The **frequency** of a signal is the number of oscillations per unit time. For a continuous sinusoidal signal:

$$S(t) = A\cos(2\pi ft + \phi)$$

where f is the frequency in Hertz (Hz). The **angular frequency** ω is $\omega = 2\pi f$. For discrete signals, the normalized frequency is $f_d = \frac{f}{f_s}$, where f_s is the sampling frequency.

Definition 2.1.28 (Stationarity)**.** A time series is said to be **stationary** if its statistical properties, such as mean, variance, and autocovariance, remain constant over time.

Formally, a time series $\{X_t\}$ is **strictly stationary** if the joint distribution of $(X_{t_1}, X_{t_2}, ..., X_{t_k})$ is the same as $(X_{t_1+h}, X_{t_2+h}, ..., X_{t_k+h})$ for all $t_1, t_2, ..., t_k$ and any shift h.

Mathematically, a weaker form, *weak stationarity* (or second-order stationarity), requires that:

- $E[X_t] = \mu$, a constant mean,
- $\text{Var}(X_t) = \sigma^2$, a constant variance,
- $\text{Cov}(X_t, X_{t+h})$ depends only on h and not on t.

Definition 2.1.29 (Eigenvalues and Eigenvectors)**.** Let A be a square matrix of size $n \times n$. A scalar λ is called an **eigenvalue** and a non-zero vector $v \in \mathbb{R}^n$ is called a corresponding **eigenvector** if they satisfy the equation:

$$Av = \lambda v$$

This equation can also be written as:

$$(A - \lambda I)v = 0$$

where I is the identity matrix of the same size as A. The eigenvalue λ is found by solving the characteristic equation:

$$\det(A - \lambda I) = 0$$

The solutions to this equation give the eigenvalues, and the corresponding eigenvectors can be found by solving the system $(A - \lambda I)v = 0$.

Definition 2.1.30 (Kullback-Leibler Divergence)**.** The **Kullback-Leibler (KL) divergence** between two probability distributions P and Q is defined as:

$$D_{KL}(P \parallel Q) = \int_{-\infty}^{\infty} p(x) \log \left(\frac{p(x)}{q(x)} \right) dx,$$

where $p(x)$ and $q(x)$ are the probability density functions of P and Q, respectively. In discrete settings, this is approximated by a sum.

Definition 2.1.31 (Entropy)**.** In the context of information theory, **Entropy** is a measure of the uncertainty or randomness of a random variable. For a discrete random variable X with a probability distribution $P(X) = \{p_1, p_2, \ldots, p_n\}$, the entropy $H(X)$ is defined as:

$$H(X) = -\sum_{i=1}^{n} p_i \log_2 p_i$$

where p_i is the probability of the i-th outcome, and the logarithm is typically base 2. The entropy $H(X)$ represents the average amount of information (in bits) produced by the random variable X, or the uncertainty associated with predicting the outcome of X.

2.1.1 Correlation and causation

Understanding the difference between *correlation* and *causation* is essential for interpreting data correctly. While both terms describe relationships between variables,

they imply very different concepts and are often confused, leading to incorrect conclusions.

As stated in Definition 2.1.23, *correlation* refers to a statistical relationship or association between two variables X and Y, meaning that as one variable changes, the other tends to change in a specific way. Figure 2.1 demonstrates two highly correlated variables, however, their correlation does not imply that one variable causes the other to change.

Two variables may be correlated due to a *common cause* Z or other underlying factors that drive both variables, rather than one variable causing the other to change. This leads to the famous principle: *Correlation does not imply causation.* A classic example is the relationship between ice cream sales and drowning incidents, both increase in the summer, but buying ice cream does not cause drowning. Instead, both are related to the warmer weather, a confounding variable. Mere correlation does not answer any of the causal queries presented in Figure 2.1.

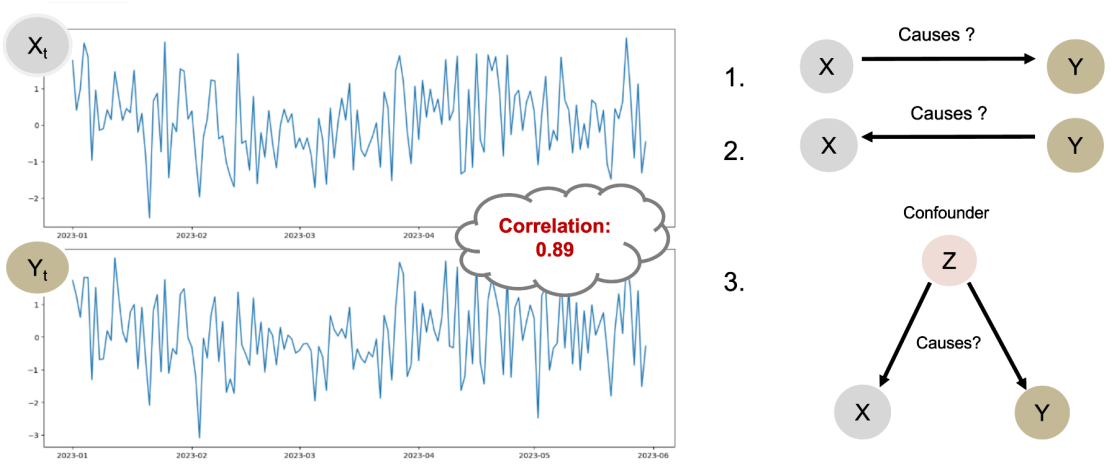


Figure 2.1: Show correlation between two variables X_t and Y_t (left side) and the corresponding causal queries [1, 2, 3] (right side).

On the other hand, *causation* implies a *cause-and-effect* relationship, where a change in one variable directly causes a change in another. To establish causation, researchers often use controlled experiments, interventions, and counterfactual reasoning. In causal studies, variables are manipulated in a controlled setting, and the resulting effects on the outcome variable are observed to determine if a causal relationship between variables exists.

Mathematically, a causal relationship between variables X and Y can be represented as:

$$Y := f(X, \text{other factors})$$

where f describes the function through which X influences Y. Interventions and experiments allow researchers to isolate this relationship and eliminate confounding factors. For hidden confounding cases, techniques like instrumental variables (IV) Angrist and Krueger (2001) are used.

The key difference is that *correlation* indicates a statistical relationship between two variables but does not imply one causes the other. While, *causation* implies that one variable directly influences the other, and the effect is not explained by other factors. Hence causation provides answers to the causal queries given in Figure 2.1.

For example, in randomized controlled trials (RCTs) Friedman et al. (2015), subjects are randomly assigned to a treatment group, which receives the intervention, and a control group, which does not, often receiving a placebo or standard care. The causal effect is measured by comparing the outcomes between these groups. The average treatment effect (ATE) Imbens and Rubin (2015) is used to quantify the causal effect:

$$\text{ATE} = \mathbb{E}[Y_{\text{treatment}}] - \mathbb{E}[Y_{\text{control}}]$$

where $Y_{\text{treatment}}$ and Y_{control} represent the observed outcomes for the treatment and control groups, respectively. where $\mathbb{E}[Y_{\text{treatment}}]$ is the expected outcome for the treatment group, $\mathbb{E}[Y_{\text{control}}]$ is the expected outcome for the control group.

By comparing these expectations, researchers can infer the causal impact of the treatment, isolating the effect of X on Y while controlling for other factors.

It is crucial to distinguish between correlation and causation to avoid misleading conclusions. Misinterpreting correlation as causation can lead to faulty decisions, which is why it is vital to employ rigorous methods, such as experiments and causal inference techniques, to establish true cause-and-effect relationships.

2.1.2 Ladder of Causation

Causality, a key concept in understanding relationships between events, has been significantly refined by Judea Pearl, particularly in his book *The Book of Why* Pearl and Mackenzie (2018, 2019). One of his major contributions is the *Ladder of Causation*, which outlines a hierarchy of causal relationships from association

to intervention to counterfactuals. This framework helps researchers move from observing correlations to understanding the underlying mechanisms of causation.

At the base of the ladder is *association*, where events co-occur without implying causality. *Interventions*, or deliberate manipulations of variables, allow researchers to observe direct effects and distinguish causality from correlation. *Counterfactuals* consider hypothetical scenarios, helping assess the causal impact of variables.

Association: At the lowest rung of the Ladder of Causation is *association*, which involves identifying statistical relationships or correlations between variables based on observational data. Mathematically, association is often expressed in terms of conditional probability:

$$P(Y \mid X)$$

which represents the probability of observing an outcome Y given that X has been observed. This allows researchers to quantify relationships between variables without establishing causality. Another common measure of association is the correlation coefficient:

$$\rho(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

where $\text{Cov}(X, Y)$ is the covariance of X and Y, and σ_X, σ_Y are their standard deviations.

However, associations may arise due to confounding variables Z, where Z influences both X and Y, leading to a spurious correlation. This confounding effect is often analyzed using conditional independence:

$$X \perp Y \mid Z$$

indicating that X and Y are conditionally independent given Z, suggesting that any observed association between X and Y could be explained by Z.

Intervention: The second rung of the ladder is *intervention*. This level involves conducting controlled experiments or interventions to manipulate variables and observe their effects on outcomes. In causal inference, interventions are often expressed using Pearl's do-operator:

$$P(Y \mid do(X))$$

where $do(X)$ represents an intentional manipulation of X, rather than just observing it. Mathematically, an intervention can be represented as changing a variable X to a specific value X' and observing the effect on an outcome variable Y:

$$Y = f(X')$$

where f describes how Y is influenced by X'. This level of causation allows researchers to establish causal relationships by actively manipulating variables, as is commonly done in randomized controlled trials (RCTs).

Counterfactual: The highest rung of Pearl's Ladder of Causation is the *counterfactual* level, which involves reasoning about what would have happened under different conditions or interventions. Unlike interventions, counterfactuals require imagining alternative realities.

Mathematically, counterfactuals are expressed as hypothetical outcomes under alternative scenarios. Given an observed outcome Y, an actual intervention X, and a counterfactual intervention X', the counterfactual outcome can be denoted as:

$$P(Y \mid do(X'))$$

which represents the probability of Y occurring if X had been set to X', even though X was actually different. The causal effect of an intervention X' is typically expressed as the difference between the observed and counterfactual outcomes:

$$\text{Causal Effect} = Y(X') - Y(X)$$

Since counterfactuals involve unobserved realities, the *potential outcomes framework* is often applied. If Y_1 represents the potential outcome if the individual receives the treatment, and Y_0 represents the potential outcome if they do not, the average treatment effect (ATE) is defined as:

$$\text{ATE} = \mathbb{E}[Y_1 - Y_0]$$

This compares the average outcomes across a population, enabling causal inference even in non-experimental settings. Counterfactual reasoning provides a robust framework for understanding cause-and-effect relationships beyond direct observations.

2.1.3 Granger causality

The concept of Granger causality (GC) was introduced by British economist Clive Granger in his seminal 1969 work Granger (1969). Granger causality is a statistical method used to determine whether one time series provides predictive information about another. Specifically, if a variable X Granger-causes a variable Y, then past values of X contribute to forecasting Y beyond what is possible using only the past values of Y itself.

It is important to note that Granger causality does not imply true causation in the philosophical or mechanistic sense. Instead, it measures whether the historical values of one variable contain statistically significant information for predicting another. If X Granger-causes Y, this suggests that including the past values of X improves the prediction of Y, even when accounting for the historical behavior of Y alone.

For two time series X_t and Y_t, we assess whether past values of X_t contribute to predicting Y_t by comparing two models. The unrestricted model predicts Y_t using only its own past values:

$$Y_t = \alpha_0 + \alpha_1 Y_{t-1} + \alpha_2 Y_{t-2} + \cdots + \alpha_p Y_{t-p} + \epsilon_t$$

where α_0 is a constant, $\alpha_1, \ldots, \alpha_p$ are autoregressive coefficients, and ϵ_t is the error term. This model captures the time dependence of Y_t without considering X_t.

The restricted model extends this by incorporating past values of X_t:

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \cdots + \beta_p Y_{t-p} + \gamma_1 X_{t-1} + \gamma_2 X_{t-2} + \cdots + \gamma_p X_{t-p} + \nu_t$$

where $\gamma_1, \ldots, \gamma_p$ are the coefficients associated with X_t, and ν_t is the error term. If the coefficients $\gamma_1, \ldots, \gamma_p$ are statistically significant, it suggests that past values of X_t improve the prediction of Y_t, indicating Granger causality from X_t to Y_t.

Hypothesis Test The test for Granger causality involves comparing the fit of the restricted model with the unrestricted model.

- Null Hypothesis H_0: X does not Granger cause Y. This implies that the coefficients $\gamma_1, \gamma_2, \ldots, \gamma_p$ are all zero.
- Alternative Hypothesis H_1: X Granger causes Y. This implies that at least one of the coefficients $\gamma_1, \gamma_2, \ldots, \gamma_p$ is non-zero.

To test for Granger causality, we perform an F-test or likelihood ratio test to compare the unrestricted model with the restricted model. The steps are as follows:

- Estimate the unrestricted and restricted models.
- Compute the residual sum of squares (rss) for both models: $\text{rss}_{\text{unres}}$ for the unrestricted model. And rss_{res} for the restricted model.
- Compute the test statistic:

$$F = \frac{(\text{rss}_{\text{res}} - \text{rss}_{\text{unres}})/p}{\text{rss}_{\text{unres}}/(T - 2p)}$$

  where p is the number of lag terms and T is the number of time periods.
- Compare the test statistic to the critical value from the F-distribution or use the p-value to determine whether to reject the null hypothesis.

Interpretation If the test statistic is significant (i.e., the p-value is smaller than the chosen significance level α, commonly 0.05), we reject the null hypothesis and conclude that X Granger-causes Y. If the test statistic is not significant, we fail to reject the null hypothesis and conclude that there is no evidence that X Granger-causes Y.

GC makes strong assumptions about the data-generating process and has several limitations:

- Stationarity: GC assumes the time series is stationary, meaning its statistical properties remain constant over time (Definition 2.1.28).
- Lag selection: Choosing the right lag length (p) is crucial, too short may omit key information, while too long may cause overfitting. Criteria like *Akaike Information Criterion* (AIC) help optimize this.
- Bidirectional causality: If X and Y Granger-cause each other, the relationship is bidirectional, common in economic and financial data.
- Nonlinearity: Standard GC assumes linearity; if the true relationship is non-linear, the test may fail to detect causality.
- Confounding: GC identifies relationships based on prediction rather than true cause-and-effect, meaning that unseen factors or hidden influences can skew the results and lead to misleading conclusions.

Variations and Extensions

- Vector Autoregressive Model (VAR): In practice, the concept Granger causality is often implemented in the context of a VAR model Lütkepohl (2005), where multiple time series are modeled together, and the test is conducted to assess the relationships among them.

- Bivariate vs Multivariate Granger Causality: The Granger causality test can be extended to more than two variables (multivariate). In such cases, the analysis would test whether one or more variables Granger cause the others, considering their joint effects.

Granger causality provides a powerful tool for understanding the temporal relationships between time series data, though it should be used with caution. It provides causality based on predictive relationships that can guide further investigation.

2.1.4 Model invariance

Model invariance is a foundational concept in causality, referring to the property that certain aspects of a causal model remain stable or consistent across different environments, interventions, or distributions Pearl (2009a); Peters et al. (2016). This principle is crucial for identifying causal relationships that are robust and generalizable, as opposed to merely capturing associations that may be specific to particular datasets or contexts. Invariance allows researchers to distinguish between true causal mechanisms and spurious correlations by focusing on relationships that persist despite changes in external conditions or interventions.

In a structural causal models (SCMs), causal relationships are represented using equations such as $Y = f(X, \epsilon_Y)$, where X represents the causal predictors, Y is the outcome, and ϵ_Y is an independent noise term Pearl (2009b). A model exhibits invariance if the functional form of the causal mechanism f remains unchanged across environments, and the noise terms (ϵ) stay independent of the causes and retain their distribution. This implies that the relationship between the causal predictors and the outcome is stable, even when external factors like interventions or environmental shifts alter other aspects of the system.

Techniques such as *Invariant Causal Prediction* (ICP) leverage this principle by identifying subsets of predictors whose relationships with the outcome remain stable

across diverse settings, ensuring that the causal model generalizes to new, unseen environments Peters et al. (2016).

Testing for invariance often involves collecting data from multiple environments or performing interventions to perturb the system. For example, consider a scenario where we aim to understand the causal relationship between exercise (X) and health (Y), while accounting for diet (Z). If the relationship $Y = f(X, Z, \epsilon_Y)$ remains consistent across populations with varying access to healthy food, it suggests that the causal mechanism is invariant. However, if the relationship changes due to differences in how diet interacts with exercise, this indicates a lack of invariance, pointing to possible confounding or non-causal associations.

Despite its power, applying the principle of invariance comes with challenges. Hidden confounders, non-stationary mechanisms, or limited availability of interventional data can complicate the identification of invariant relationships Pearl (2009b); Schölkopf (2022). Nevertheless, the principle of invariance provides a rigorous foundation for distinguishing causal relationships from spurious correlations, enabling robust causal discovery and generalization.

In this work, we leverage the invariance principle to ensure that the model's response remains stable across different interventional settings when conditioned on its causal predictors. This property allows us to differentiate between causal and non-causal variables.

Formally, let Y represent the response variable, C denote the set of causal predictors, and E indicate the environment or intervention setting that may influence non-causal predictors. The central idea is that the distribution of Y remains unchanged when conditioned on C, expressed as:

$$P(Y \mid C, E) = P(Y \mid C)$$

This directly implies that Y is independent of E given C:

$$Y \perp E \mid C$$

In other words, changes in the environment E, which may influence non-causal predictors, do not affect the conditional distribution of Y given C. However, variations in C can influence Y, revealing the underlying causal relationships.