Investment Institute

Working Paper 157 I **March 2024**

Causality Approach Applied to Clean-Tech Equities

Trust must be earned

Amundi ASSET MANAGEMENT

Causality Approach Applied to Clean-Tech Equities

Edmond LEZMI
Amundi Investment Institute
edmond.lezmi@amundi.com

Karl SAWAYA
Amundi Investment Institute
karl.sawaya@amundi.com

Jiali XU
Amundi Investment Institute
jiali.xu@amundi.com

Abstract

The clean-tech industry has experienced remarkable growth, bringing forth groundbreaking technologies and sustainable solutions. This research article delves into the examination of factors that shape the evaluation of net-zero assets in various sectors and themes. Through observational analysis utilizing key financial indicators, it becomes apparent that companies exclusively involved in the clean-tech industry, known as pure players, generally outperform those that have less focus in this area, referred to as non-pure players in terms of financial performance [50]. The transition towards a sustainable energy system is greatly facilitated by comprehensive policies and regulations. For instance, in the United States, the Inflation Reduction Act (IRA) and in Europe, the Net-Zero Industry Act (NZIA) play significant roles in shaping the dynamics of asset valuation. These regulatory frameworks contribute to the valuation dynamics and help drive the growth of clean-tech investments [26]. Additionally, the physical and transitional climate risk exert a substantial influence on the valuation of net-zero assets. To gain a deeper understanding of the drivers behind clean technologies and their causal relationships, our study employs a specific branch of Bayesian probabilistic approach introduced by Judea Pearl, the Ladder of Causation, explained in The Book of Why. This approach enables us to model the dependency structure among these influential factors and evaluate their direct and indirect impacts on cleantech stock returns by manipulating the explanatory variables. By creating coherent scenarios through interventions on these variables, we can address essential what-if questions, aiding investors and policymakers in making more informed

Acknowledgement

The authors are very grateful to Takaya Sekine and Thierry Roncalli for their helpful comments.

decisions in this ever-evolving and dynamic industry. Within the framework of Bayesian analysis, the do-calculus and the counterfactual concept play a pivotal role and make it possible to calculate the probability distribution of a random variable under a hypothetical scenario on the explanatory variables different from the observed data. We not only explore the direct effects of interventions on explanatory variables but also reveal sensitivity groups among clean-tech companies. These sensitivity groups consist of companies that exhibit a similar sensitivity to a specific causal factor. This insight is valuable for pinpointing which clean-tech subsectors or companies are particularly affected by certain changes or interventions, offering a more detailed understanding of the industry's dynamics.

Keywords: Bayesian network, causal effect, clean-tech, do-calculus, instrumental variable, structural causal model.

JEL classification: C11, G12, Q5.

About the authors

Edmond LEZMI

Edmond Lezmi is the Head of Multi-Asset Quant Portfolio Strategy at Amundi Institute. He is involved in the application of machine learning techniques, notably optimization, Bayesian optimization, graph neural networks and generative models.

He joined Amundi in 2002. Prior to his current role, he was Head of Quantitative Research at Amundi Alternative Investments (2008-2012), a Derivatives and Fund Structurer at Amundi IS (2005-2008), and Head of Market Risk (2002-2005). Before joining Amundi, he was Head of Market Risk at Natixis, and an Exotic FX Derivatives Quantitative Developer at Société Générale. He started his working career with Thales in 1987 as a Research Engineer in signal processing.

Edmond holds an MSc in Stochastic processes from the University of Orsay.

Karl SAWAYA

Karl Sawaya joined Amundi in June 2023 as an intern in the Quantitative Portfolio Strategy team of the Amundi Investment Institute. His work focuses on causality using do-calculus, graph theory, Bayesian analysis, and structural equations. He studied the effects on Clean-Tech stocks after perturbing macroeconomic variables. Karl was an engineering student at the CentraleSupélec engineering school, where he pursued a Master's degree in Applied Mathematics, specializing in Mathematics and Physics. He is currently pursuing a Master's degree in Probability and Random Models, jointly offered by ENS Ulm and Sorbonne University.

Jiali XU

Jiali Xu is a quantitative research analyst in Quant Portfolio Strategy at Amundi Investment Institute. He joined Amundi in 2018 as a quant in the multi-asset quantitative research team, where he is responsible for developing quantitative investment strategies. His research focuses on factor investing, portfolio construction, and the application of advanced statistical methods in portfolio management.

Previously, between 2014 and 2018, he was a quantitative analyst in the risk analytics and solutions team at Société Générale. A graduate of Ecole des Ponts ParisTech, he also holds a Master's degree in Financial Mathematics from the University of Paris-Est Marne-la-Vallée.

1 Introduction

In recent years, the global community has witnessed an increasing sense of urgency to address the environmental challenges posed by climate change and resource depletion. This heightened awareness has led to a growing emphasis on the development and adoption of clean technologies, commonly referred to as clean-techs. The clean-tech industry encompasses a diverse range of technologies, processes, and services aimed at mitigating environmental impact while promoting sustainable economic growth.

Clean-tech represents a paradigm shift in the way industries and societies operate by incorporating innovative solutions that reduce carbon emissions, optimize resource utilization, and minimize waste generation. By leveraging renewable energy sources, implementing energy-efficient systems, and adopting sustainable practices, the clean-tech industry seeks to transition societies towards a low-carbon and sustainable future. The urgency to address climate change has been underpinned by mounting scientific evidence highlighting the detrimental consequences of greenhouse gas emissions on global temperatures, ecosystems, and human health. Governments, corporations, and individuals are recognizing the need to reduce their carbon footprint and embrace cleaner alternatives. This collective push for sustainability has fueled the rapid growth of the clean-tech industry, which encompasses sectors such as as solar power, green hydrogen, electric vehicles, water treatment, and sustainable agriculture.

As the demand for clean and renewable energy solutions intensifies, understanding the intricate dynamics of the clean-tech equities market becomes imperative for both investors and policymakers. This paper aims to unravel the causal relationships that underlie the valuation of net-zero assets in the clean-tech sector, employing a rigorous Bayesian framework rooted by Judea Pearl [32].

In the context of clean-tech equities, causality plays a pivotal role in decoding the intricate web of relationships among various macro-economic factors influencing stock returns. Traditional financial models often fall short in capturing the nuanced cause-and-effect relationships that define the clean-tech sector's performance. Judea Pearl's research provides a theoretical foundation, allowing us to move beyond correlation and explore the underlying causal mechanisms. Central to our investigation is the application of Bayesian analysis, which not only accommodates uncertainty inherent in financial markets but also enables the modeling of causal relationships through directed acyclic graphs (DAGs). Bayesian networks provide a powerful tool to represent and analyze the dependencies among variables. One of the key challenges in evaluating the clean-tech sector lies in understanding the impact of interventions – policy changes, technological advancements, or market shifts – on stock returns. The do-calculus in Bayesian analysis facilitates this exploration by allowing us to estimate the effects of interventions on our observed variables. This is particularly crucial in an industry where external factors, such as regulatory changes and technological breakthroughs, can have profound effects on asset valuations. Through the counterfactual concept, we extend our analysis beyond observed data, answering essential what-if questions. What if a specific policy had not been implemented? What if a technological innovation had not occurred ? By manipulating explanatory variables, we create hypothetical scenarios, enabling us to assess the potential impacts of different trajectories on clean-tech stock returns ([36]).

In section 2, we commence with a comprehensive review of the clean-tech industry, exploring macroeconomic drivers and emphasizing the concept of purity in clean-tech companies. Following this, in section 3, we introduce the theoretical framework, leveraging Bayesian networks, DAGs, and Pearl's causality concepts such as do-calculus. This groundwork sets the stage for our empirical investigation into sector-based clean-tech equities. Section 4 outlines

our empirical approach and details the case study on sector-based clean-tech equities. We discuss the causal structure revealed through Bayesian analysis, the impact of interventions on key explanatory variables, and measures such as the average treatment effect. Finally, in section 5, we draw conclusions from our findings.

Through this research, we contribute to the growing body of knowledge surrounding clean-tech equities, providing a robust analytical framework that transcends traditional financial models. This research article enables investors and readers to benefit from both a thorough analysis of the clean-tech sector and access to a comprehensive methodology involving innovative causal tools. Designed to be applied to specific or general case studies, this methodology offers a practical approach to constructing evolution scenarios for various asset types, extending beyond the scope of clean-techs.

2 Clean-tech industry review

2.1 Valuation challenges in the clean-Tech industry

The clean-Tech industry serves as a dynamic force driving innovation and sustainability across diverse sectors. In the global effort to combat climate change, clean-techs play a pivotal role in promoting environmental responsibility and fostering economic growth. Operating within a matrix of sectors, each contributing uniquely to the overarching goal of sustainability, clean-techs span various domains. The hydrogen/chemicals sector explores innovative approaches to clean energy production, electric vehicles (EV) and batteries revolutionize transportation, Agribusiness focuses on sustainable farming practices, and environmental initiatives address pollution and conservation. Additionally, the energy sector explores renewable sources, and financial services dedicated to supporting clean-tech initiatives complete the industry's multifaceted nature.

In the United States, the IRA stands as a legislative milestone influencing the trajectory of clean-tech companies. The IRA incentivizes financial investment in sustainable technologies and introduces regulatory frameworks to ensure compliance with environmental standards. Simultaneously, Europe embraces the NZIA, reinforcing the continent's commitment to achieving carbon neutrality. These legislative measures provide a guiding framework for clean-tech companies, shaping their strategies and influencing their valuation.

The passage of the IRA in late 2022 has garnered significant attention, signaling wide-ranging implications across various sectors. The IRA's impact is anticipated to be substantial, with potential tailwinds driving over 50% incremental earnings upside. This positive effect could materialize as early as 2024, prominently reflecting in Profit and Loss statements across multiple companies. Figure 1 displays some clean-tech companies with their sectors and investments they plan to make to meet IRA criteria. This demonstrates the influence of government policies on the development of clean-techs, encouraging increased investment in necessary technologies. Specifically, the IRA presents a significant boon to the US solar and energy storage sector. The IRA extends the solar investment tax credit from 26% to 30% and ensures the continuation of these credits for at least a decade. This development is poised to fuel a decade-long runway for stable installation growth in residential, commercial, and utility-scale markets. Projections indicate robust growth of +18% compound annual growth rate (CAGR) in US solar installations from 2022 to 2026 and +16% CAGR through 2040 in US energy storage installations. Moreover, the IRA introduces generous solar manufacturing credits, offering a significant advantage to suppliers engaged in domestic manufacturing. Beyond demand tailwinds, these manufacturing credits could incentivize an increase in domestic manufacturing capacity, providing meaningful P&L benefits to manufacturers. It is estimated that manufacturing credits may account for anywhere between 10%-40% of the average selling price (ASP) of solar components. Illustrative analyses suggest that buy-rated solar stocks could potentially see upside of 60%-280% from current levels (Figure 2).

2.2 Pure-player thematics

The term pure-players designates companies exclusively dedicated to the development, implementation, and advancement of clean technologies, operating with an unwavering commitment to environmental sustainability. Purity in this context signifies a comprehensive dedication to clean-tech endeavors, encompassing a spectrum of attributes that distinguish these entities within the clean-tech ecosystem.

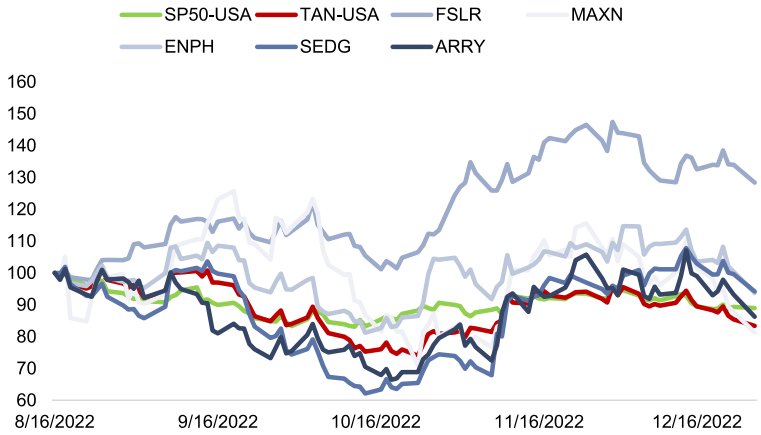
In our paper, our focus is on conducting a causal analysis exclusively on pure-player clean-techs with sufficient market capitalization to meet the IRA criteria. Here, we delve into the

Figure 1: Commentary of clean-techs companies on their leverage to IRA incentives

Sector	Company	Company commentary
Clean Tech	FSLR	Announced new 3.5GW manufacturing footprint in Alabama, US, scheduled for completion in 2025 Announced new 3.3GW module capacity expansion in Ohio, with completion expected in 2H23 Announced new R&D testing facility in Ohio, with completion expected in 2024
	MAXN	Planning to build 3GW US cell and module capacity in Southeast US with production in 2025
	ENPH	Opening 4 to 6 manufacturing lines in the US during 2H23, with estimated capacity of ~4.8GW-7.2GW (GSe)
	SEDG	Planning to establish US manufacturing capabilities in the US during 2023
	ARRY	Exploring options to split manufacturing credits with its torque tube & fastener suppliers
Hydrogen/Chemicals	NEL	Currently ongoing a site selection process in the US for its next greenfield electrolyser manufacturing capacity addition in both PEM and alkaline technologies.
	DNR	Stated the opportunity to increase presence in the US for their energy transition division through their JV
	APD	Announced a $4bn investment in a green hydrogen JV with AES in Texas capable of producing 200 mtd of green hydrogen with a targeted start-up in 2027
	LIN	Calls out $33bn of US investment opportunity driven by the IRA that LIN believes it will win
EV/Autos/Batteries	TSLA	Expects to fully meet the IRA's requirements to qualify for electric vehicle tax credits, which could be a significant boost towards accelerating Tesla's mission and scaling up the battery supply chain in the US
	FREY	Announced a new 34GWh Gigafactory in Georgia, US, that is set to start production in 2026
Industrial/Agribusiness	GE	Over 7GW of offshore wind in the pipeline could potentially benefit from the PTC extension
	DAR	Realizing ~$600mn of annual income from the BTC extension in 2024, with an additional $0.60-$0.75/gal ($360mn-$450mn GSe) in 2025-27 under the 45z credit
	ADM	Potentially realize $120mn-$180mn of annual earnings from its eight connected facilities in 2025-27, with scope for an additional $60mn-$90mn from three other facilities
	GPRE	Potentially realize $200mn-$300mn of incremental earnings from low-CI ethanol production, with separate medium-term tailwinds from industry investment in SAF; biofuel supportive provisions will continue to benefit soybean crush and edible oil refining margins
Environmental Services	WM	Securing long-term RNG pricing contracts close to 10 years in duration; citing a strong RNG demand outlook from public utilities and large institutions looking to decarbonize; potentially developing 100 landfill gas projects over time
	RSG	RSG and Archaea announced plans for an RNG facility at Middle Point landfill in Tennessee as part of its 40 project joint venture on 10/4; BP announced a $4.1 bn acquisition of Archaea on 10/20
Energy Services	MTZ	Increasing IRA exposure with close of IEA acquisition; expecting to benefit from the $370bn IRA incentives that impact its end markets
	BKR	Seeing opportunities with hydrogen and carbon capture technology following more favorable subsidies in the IRA and increased customer discussions on projects
Financials	BAM	Specialized flagship funding seeking to make investments in emission reduction/avoidance; seeing an $150 trillion investment opportunity in clean energy and decarbonization solutions ; expects projects to build faster at more attractive economics and on a derisk basis

Source : Data compiled by Goldman Sachs Global Investment Research.

Figure 2: Market performance of solar stocks

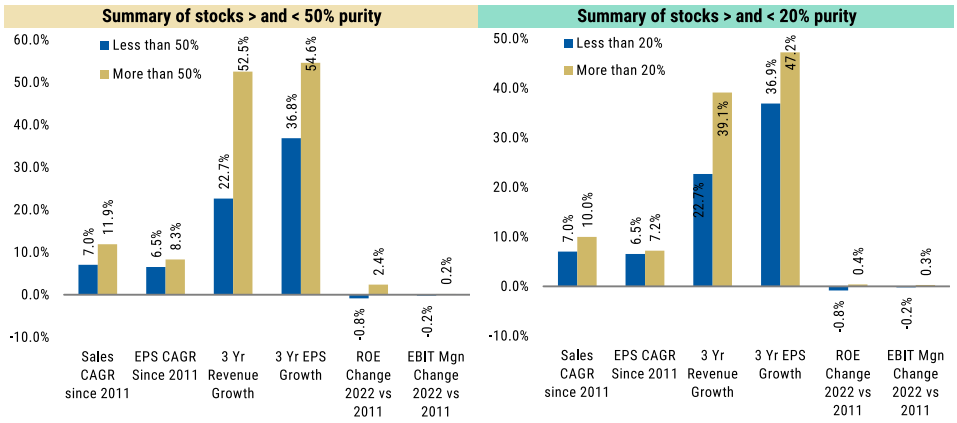


Source : Factset.

correlation between thematic purity in stocks and their financial performance. The premium for thematically pure stocks is predominantly driven by a growth phenomenon. The analysis of fundamentals (Figure 3) indicates that highly valued pure stocks outperform their non-pure counterparts in terms of sales, earnings per share (EPS) and CAGRs, suggesting higher long-term revenue growth. Ongoing debates exist about whether this growth is organic or

through acquisition. Margin analysis shows little evidence of a clear divide between pure and impure stocks, while return on equity (ROE) exhibits a fading pattern over a decade for both categories. Figure 3 illustrates that fundamentals such as growth, margin, and ROE are higher for pure-player clean-techs than for non-pure-players.

Figure 3: Fundamentals of pure vs non-pure stocks



Source : Thomson, Morgan Stanley Research.

2.3 Key drivers of clean-tech asset valuation

In our paper, we aim to construct evolution scenarios for clean-tech stocks by studying the causal effects of explanatory macroeconomic variables. We conducted sector-specific research (solar, agribusiness, EV/batteries, etc.) to identify the macroeconomic variables influencing the value of clean-techs. After thorough investigation, we identified nine key factors for causal analysis:

- **Oil price:** The cost of oil has a direct impact on clean-techs, especially in sectors like renewable energy and electric vehicles, as it influences production costs and market competitiveness.
- **US interest rate:** Changes in US interest rates affect the financing costs for clean-tech projects, influencing investment decisions and overall market dynamics.
- **EU interest rate:** Similar to the US, European interest rates play a role in shaping the financial landscape for clean-tech companies, impacting their valuation.
- **Inflation:** Inflation rates influence operational costs and pricing strategies.
- **Carbon price:** The price of carbon credits directly affects clean-tech companies involved in carbon trading and offsets, making it a crucial factor in their valuation.
- **Gas price:** Clean-techs, particularly those in the energy sector, are influenced by the cost of gas, as it can impact their competitiveness in the market.
- **Nickel price:** Given the significance of nickel in electric vehicle batteries, its price has a direct impact on the production costs and profitability of EV-focused clean-tech companies.

- **Technology companies stocks index:** As discussed in [8], the impact of both physical and transition climate risk on the short-term and long-term relationship between clean-tech stocks and technology stocks holds potential significance for investment decisions. The observed co-movement between clean energy and technology stocks, particularly the positive association with transition risk shocks, suggests a convergence of these sectors over the long run.
- **Semiconductor stocks index:** With increasing integration of semiconductors in clean-tech products, the semiconductor stocks index provides insights into technological trends and potential supply chain challenges impacting clean-tech companies.

3 Theoretical framework

Understanding and identifying cause and effect relationships between variables or events is a major goal of hard and social sciences. We describe some methodologies suitable for extracting such relationships from raw data, despite the lack of consensus within the scientific community.

The primary objective of this paper is to construct scenarios for evaluating net-zero assets, specifically clean-tech stocks. To achieve this, we started by identifying the macroeconomic factors driving the values of these stocks. We'll need then to quantify the impact of disturbances in these factors on the value of a clean-tech stock. The causal framework of Pearl will enable to generate coherent scenarios and address what-if questions. For instance, we may explore what would have happened to the value (V) of the clean-tech company A if the price of oil had increased by 10% within one year. Or, having observed a value V equal to v when the oil price P was equal to x, what would have happened to V if I had intervened to set P to the value x'?. We will delve into the concept of intervention, where altering a variable with a deterministic value changes the entire probabilistic model, making it possible to perform what-if inference. To assess these impacts effectively, we will develop various metrics, each one based on ordered moments of the new, post-intervention probability distribution. One intriguing concept is the creation of sensitivity groups for clean-tech stocks, where stocks are grouped based on the factor to which they are most sensitive. Sensitivity, in this context, relates to causality, where we identify the factor with the strongest causality – meaning a disturbance in this factor would have the most significant impact on the value of interest. Once these sensitivity groups are defined, the analysis of clean-tech stock evolution would then translate into the analysis of the evolution of the factor to which they are most sensitive. To embark on this extensive process, it is necessary to delve into the theoretical framework of graphical models. Graphical models serve as tools providing a theoretical foundation for studying causal hypotheses, non-trivial causal phenomena, and paradoxes. Judea Pearl's causal analysis introduces a set of technical rules, such as the do-calculus, to facilitate the identification of causal effects in non-parametric models. Regardless of the subject of study, as long as we explore how external factors influence a variable of interest and how disruptions occur, causal graphs should emerge as a fundamental tool in our consideration. Causal graphs are special cases of classical random graphs, as they are induced by mathematical models presented as a set of ordered equations known as structural equation models (SEM). These models are particularly powerful when the causal structure is identifiable, enabling the estimation of direct and indirect causal effects, which play a pivotal role in generating coherent scenarios. Subsequently, calculation metrics are defined to quantitatively assess these disturbances.

In the subsequent sections, we will find out that the causal graph corresponds precisely to a DAG. Consequently, the structural causal model linked with the causal graph aligns perfectly with a Bayesian network (or belief network). This alignment stems from an intuitive understanding that, when exploring causality within graphical models, we are naturally drawn to the use of DAGs (and thus Bayesian networks). This preference is supported by a straightforward mathematical argument stating that the joint distribution of the model variables equals the product of the conditional probabilities of each variable given its parents. This inherent directionality from effect to cause makes DAGs the most efficient representation for causal models.

Bayesian networks inherently adopt the structure of DAGs. Within these graphs, a topological order of variables is established, where certain variables act as parent nodes, linked by interdependence connections to a specific variable of interest, which plays the role of a child node. The set of edges converging on a specific variable embodies a function, whether linear

or nonlinear, where each variable serves as a function of its parent variables. These functions collectively constitute structural equation models, quantifying the causal relationships from parent variables to their child counterparts. In cases where these functions are linear, they give rise to structural coefficients, intricately connected to both direct and indirect causal effects. DAGs serve as tools for uncovering causal links between various variables in the model. This prompts the question of how to define the graph's structure and determine the interdependence links that best reflect reality. While experts can establish causal links manually, we aspire to automate this process using available data. Quality data is essential for this endeavor, as data quality is considered as divine in mathematical modeling, much like the Higgs boson in modern physics. To achieve this, we define various causality algorithms, grouped into four families: constraint-based, score-based, functional-based, and geometric-based algorithms. These algorithms, each following a specific approach, help establish the optimal graph structure and the most probable causal links. It's vital to emphasize that using causality algorithms in isolation, without expert guidance, may not yield sufficiently effective results and might even generate counter-intuitive or erroneous links. For instance, a causality algorithm may, depending on data quality, establish a link like *rain causes clouds*, which is not particularly relevant. Thus, expert intervention is crucial, even if it involves constraining the algorithm within an intuitive and realistic logic.

Once the graph structures are established, our next step involves estimating causal functions. These functions include coefficients, especially in the context of linear functions, which we'll predominantly employ throughout this article. These coefficients act as the connecting threads on the edges of the graph, linking each child variable (the effect) to its parent variables (the causes). To undertake this estimation, we introduce a crucial mathematical tool developed by Judea Pearl known as the do-calculus [32]. The do() operator plays a pivotal role in expressing causal effects as theoretical formulas, offering a detailed understanding of how changes in one variable translate into changes in another. The do-calculus is inherently tied to the concept of intervention. The do() operator serves as a powerful notation within the do-calculus, signifying a deliberate intervention or manipulation of a specific variable to observe its impact on the overall causal structure. The estimation of these causal effects (or structural coefficients in the case of linear functions) from data becomes feasible when the model adheres to certain identifiability criteria. These criteria are intricately linked to the structure and interdependence links within the graph. Identifiable models pave the way for estimating these coefficients through partial regressions.

Finally, we will introduce a set of metrics designed to address counterfactual queries. These queries, often framed as what-if causal questions, arise from hypothetical scenarios. The metrics we present are essentially ordered moments of counterfactual variables – that is, variables derived after intervening on the causal model. At the core of these interventions lies the do() operator. One key concept that emerges from such interventions is the notion of post-intervention probability distribution. When we apply the do() operator to a specific variable, we essentially "fix" or set that variable to a particular value, simulating a controlled experiment where that variable is deterministic. The post-intervention probability distribution then represents the distribution of the variables in the system, given this controlled intervention. It provides insights into how the system's behavior changes in response to a deliberate alteration of a particular variable, offering a valuable perspective on causal relationships and their implications. In essence, the post-intervention probability distribution allows us to quantify the impact of interventions on the overall causal structure, enabling a more comprehensive understanding of causal effects in hypothetical scenarios.

In summary, our approach follows the following roadmap:

1. We begin with a comprehensive exploration of the foundational concepts underlying

Bayesian networks and DAGs, establishing the theoretical groundwork.

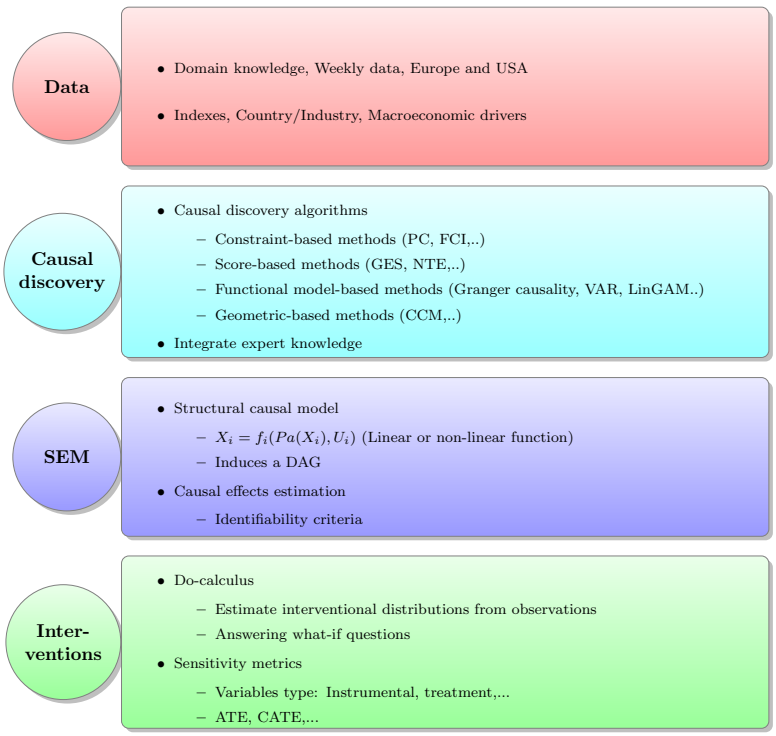
2. Next, we introduce a diverse range of causality algorithms–a crucial step to optimize the structure of our causal model, ensuring its accuracy and effectiveness.

3. Subsequently, we delve into structural equation models, as well as the theoretical framework behind the do-calculus, introducing the concept of causal effects and intervention.

4. Then, we explain the criteria for identifying coefficients and the methods for estimating them.

5. Finally, we focus on creating sensitivity metrics to quantify variable disturbances, along with addressing what-if questions to build coherent scenarios. While the theoretical part may be extensive, it is vital for a complete understanding of our approach.

Figure 4: Overview of the methodology



3.1 Bayesian networks and directed acyclic graphs

In this subsection, we embark on a rigorous exploration of causality modeling, grounded in Bayesian principles and an array of probabilistic graphical models. Our objective is to provide a systematic and analytical approach to understanding the complexities of causal inference and modeling. At the core of causality modeling lies DAGs, which serve as a graphical framework for representing causal relationships. These acyclic graphs offer a formal and

expressive means to represent the directional and causal connections between variables, facilitating a clear visualization of causal structures. Within this framework, we delve into the notion of conditional dependencies within a graph, a fundamental concept that allows us to uncover indirect causal effects and identify conditional independence relationships. Bayesian belief networks, firmly rooted in probability theory, provide a formal foundation for encoding and analyzing these dependencies, enabling the construction of structured causal representations. Structural Causal Models (SCMs) further enhance our ability to mathematically model causality. By specifying structural equations governing variables and their interplay, SCMs enable us to simulate the outcomes of interventions and rigorously estimate causal effects within a system.

Let us begin by defining the fundamental components of graphical models:

1. **Variables (X):** In a graphical model, we work with a set of random variables denoted as X. These variables can represent various observable or latent factors within a system.

2. **Nodes and Edges:** Graphical models are composed of nodes and edges. Each node in the graph corresponds to a random variable, while the edges represent probabilistic dependencies between variables.

3. **Conditional Independence:** One of the key concepts in graphical modeling is conditional independence. Two variables X_i and X_j are conditionally independent given a set of variables $\mathbf{Z}$, denoted as $X_i \perp X_j|\mathbf{Z}$, if their relationship can be explained solely by the variables in $\mathbf{Z}$. This notion plays a crucial role in modeling probabilistic relationships.

4. **Adjacency (undirected graph):** In an undirected graph G, two vertices u and v are considered adjacent if there exists an edge $\{u, v\}$ in the graph, denoting that u and v are directly connected.

5. **Adjacency (directed graph):** In a directed graph G, vertex u is considered adjacent to vertex v if there exists a directed edge (u, v) in the graph, indicating that there is a one-way connection from u to v.

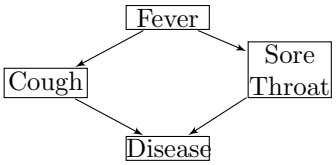
6. **Cycle in a directed graph:** A cycle in a directed graph is a sequence of distinct vertices $(v_1, v_2, \ldots, v_k)$ for $k \geq 3$, where $v_1 = v_k$, and for $1 \leq i < k$, v_i and v_{i+1} are adjacent for $1 \leq i < k$. In other words, a cycle is a closed path in the directed graph where the first and last vertices are the same, and there are directed edges between consecutive vertices, following a specific direction from v_i to v_{i+1} for all $1 \leq i < k$.
   In a directed cycle, it's important to follow the direction of the edges, meaning that you can traverse from one vertex to the next only by following the direction of the edges, not in the reverse direction. This is in contrast to undirected graphs, where cycles can be traversed in both directions.

We provide a mathematical framework of graphical models.

Graph Structure:
Let $G = (V, E)$ be a graph, where $V = \{X_1, X_2, \ldots, X_n\}$ represents the set of random variables and E represents the set of edges between the variables. Here's an example of a Bayesian network that represents the relationships between symptoms (Fever, Cough, and Sore Throat) and the presence of a disease. Fever influences the presence of Cough and Sore Throat, while both Cough and Sore Throat contribute to the likelihood of having the

Disease. This graph is oriented, meaning that hat the edges of the graph have a specific direction associated with them. The direction of the edge represents a one-way connection, indicating that there is a specific order or flow between the vertices. The graph is also acyclic, meaning there are no circular loops or circular dependencies. Each edge has a specific one-way direction, representing a clear order or flow from one vertex to another.

Medical Diagnosis Bayesian Network



Probability Distributions:
For each variable X_i in V, there is an associated probability distribution $P(X_i \mid \text{Pa}(X_i))$, where $\text{Pa}(X_i)$ denotes the set of parent variables of X_i in the graph G. The probability distribution $P(X_i \mid \text{Pa}(X_i))$ quantifies the conditional relationship between X_i and its parents given the values of the parent variables.

Markov Property:
A graphical model satisfies the Markov property if every variable X_i is conditionally independent of its non-descendants in the graph G, given its parents $\text{Pa}(X_i)$.

Factorization:
The joint probability distribution of all variables in the graphical model can be factorized as follows:

$$P(X_1, X_2, \ldots, X_n) = \prod_i \mathbb{P}(X_i \mid \text{Pa}(X_i))$$

where the product is taken over all variables X_i in V.

Inference and learning:
Graphical models allow for efficient inference and learning algorithms. Inference involves computing the probability distribution over a subset of variables given evidence, while learning aims to estimate the parameters of the probability distributions from observed data.

Now, we can define a DAG :

Definition 1. ***Directed acyclic graph (DAG)*** *A Directed acyclic graph, often abbreviated as DAG, is a finite directed graph that satisfies two essential properties:*

1. *It is a finite set of vertices or nodes.*
2. *It is a set of directed edges where each edge has a direction, and there are no cycles in the graph.*

Now, let us introduce two prominent types of graphical models:

3.1.1 Bayesian Networks

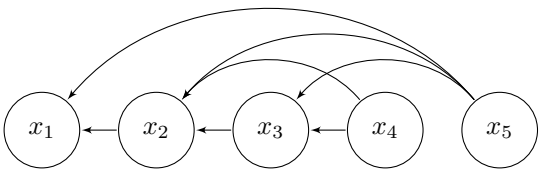
A Bayesian network is a DAG that represents probabilistic dependencies among variables. Mathematically, a Bayesian network consists of a set of nodes (random variables) and a set

of directed edges that indicate the causal relationships between them. The conditional probability distribution of each variable is defined based on its parents in the graph. The joint distribution of all variables in the network is factorized according to the graph structure using the chain rule of probability. Consider a Bayesian network with variables $X_1, X_2, \ldots, X_n$, denoted as $\mathcal{B}$, its joint probability distribution can be represented as:

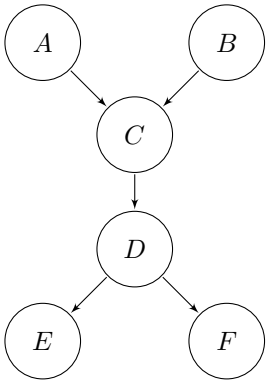
$$\mathbb{P}(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} \mathbb{P}(X_i | \text{Parents}(X_i, \mathcal{B}))$$

In a belief/Bayesian network, conditional independence can be determined by examining the graph structure and the relationships encoded by the directed edges. If two variables are not directly connected by an edge or if there is an active trail between them with all observed variables being on the trail, then these variables are conditionally independent given the observed variables. We will present some rules that help characterize conditional dependencies in Bayesian networks. To illustrate these rules, we will refer to the example depicted in Figure 5.

Figure 5: Bayesian networks



(a) A Bayesian network with 5 variables



(b) A Bayesian network with 6 variables

The formula for the joint distribution of the Bayesian network given in figure 5a is:

$$\mathbb{P}(x_1, x_2, x_3, x_4, x_5) = \mathbb{P}(x_1 \mid x_2, x_5) \cdot \mathbb{P}(x_2 \mid x_3, x_4, x_5) \cdot \mathbb{P}(x_3 \mid x_4, x_5) \cdot \mathbb{P}(x_4) \cdot \mathbb{P}(x_5)$$

Proof. The proof is based on Bayes' rule:

$$\mathbb{P}(x_1, x_2, x_3, x_4, x_5) = \mathbb{P}(x_1 \mid x_2, x_3, x_4, x_5) \cdot \mathbb{P}(x_2, x_3, x_4, x_5)$$

Given that a variable is conditionally independent of its non-descendants given its parents, we have $x_1 \perp (x_3, x_4) \mid x_2, x_5$

$$\begin{aligned} &= \mathbb{P}(x_1 \mid x_2, x_5) \cdot \mathbb{P}(x_2 \mid x_3, x_4, x_5) \cdot \mathbb{P}(x_3, x_4, x_5) \\ &= \mathbb{P}(x_1 \mid x_2, x_5) \cdot \mathbb{P}(x_2 \mid x_3, x_4, x_5) \cdot \mathbb{P}(x_3 \mid x_4, x_5) \cdot \mathbb{P}(x_4, x_5) \end{aligned}$$

Given that x_4 and x_5 are unconditionally independent since there is no edge linking these two variables, we have $\mathbb{P}(x_4, x_5) = P(x_4) \cdot \mathbb{P}(x_5)$ and thus :

$$\mathbb{P}(x_1, x_2, x_3, x_4, x_5) = \mathbb{P}(x_1 \mid x_2, x_5) \cdot \mathbb{P}(x_2 \mid x_3, x_4, x_5) \cdot \mathbb{P}(x_3 \mid x_4, x_5) \cdot \mathbb{P}(x_4) \cdot \mathbb{P}(x_5)$$

□

Following Figure 5b, we'll present some facts on conditional dependencies in a belief network. Afterward, we'll discuss the fundamental rules of these dependencies.

- Marginalizing over C makes A and B independent. We say that A and B are unconditionally independent, i.e., $\mathbb{P}(A, B) = \mathbb{P}(A) \cdot \mathbb{P}(B)$.
- A and B are conditionally dependent given C, i.e., $\mathbb{P}(A, B \mid C) \neq \mathbb{P}(A \mid C) \cdot \mathbb{P}(B \mid C)$.
- A and B are conditionally dependent given D, i.e., $\mathbb{P}(A, B \mid D) \neq \mathbb{P}(A \mid D) \cdot \mathbb{P}(B \mid D)$.
- E and F are conditionally independent given D, i.e., $\mathbb{P}(E, F \mid D) = \mathbb{P}(E \mid D) \cdot \mathbb{P}(F \mid D)$.
- Marginalizing over D makes E and F graphically dependent, i.e., $\mathbb{P}(E, F) \neq \mathbb{P}(E) \cdot \mathbb{P}(F)$.

Remark 1. ***(Markov Blanket)***
The Markov Blanket of a variable $x_i \in \chi$, denoted by $MB(x_i)$, is the set of variables that corresponds to the parents, spouse, and children of the variable x_i. For any other variable x_j that is not in the Markov Blanket of x_i, we have:

$$x_i \perp x_j \mid MB(x_i)$$

which means that x_i is conditionally independent of x_j given its Markov Blanket. For example, for figure 5a, $MB(x_3) = x_2, x_4, x_5$, so $x_3 \perp x_1 \mid MB(x_3)$.

The *d-separation* and *d-connection* criteria can be used to determine conditional independence relationships in Bayesian networks. If X and Y are d-separated given a set of variables Z, then X and Y are conditionally independent given Z. Alternatively, if X and Y are d-connected by a set of variables Z, then X and Y are conditionally dependent given Z.

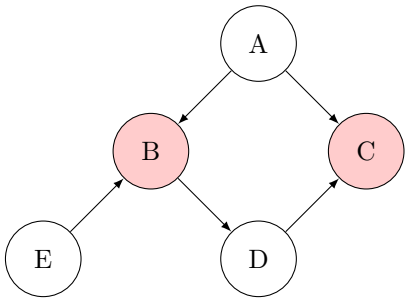
Proposition 1. ***Rules of d-Separation and d-connection***

- *X and Y are d-connected if there exists an undirected path between them that do not contain a collider. A collider is a node in the path where two arrows meet.*
- *X and Y are d-separated if there is no active path between them. An active path is one that can be traced without passing through a collider. In this case, X and Y are (unconditionally) independent.*
- *X and Y are d-connected by a set of variables Z if there is at least one active path that does not include any member of Z.*
- *X and Y are said d-separated by a set of variables Z if every undirected path between any variable in X and any variable in Y is blocked. A path is blocked if it contains a node w such that either w is a collider not included in Z and has no descendants in Z or w is not a collider but it is included in Z.*

- *If a path between X and Y includes a collider w, and w is in Z or has at least one descendant in Z, then X and Y are d-connected by Z. However, if Z contains a non-collider along this path, then X and Y are blocked given Z.*

Example 1. *According to the rules of d-separation, we may say that :*

1. *$A \perp E \mid D$: The two paths between A and E are $A \to B \leftarrow E$ and $A \to C \leftarrow D \leftarrow B \leftarrow E$. The first path is not blocked since B is a collider but it has a descendent in the conditioning set, namely D. The second path is blocked since C is a collider and neither C or its descendants is in the conditioning set.*
2. *D is graphically dependent on E given B, C : The two paths between D and E are $D \leftarrow B \leftarrow E$ which is not blocked since B is not a collider and it's not in the conditioning set and $D \to C \leftarrow A \to B \leftarrow E$ which is not blocked since C and B are colliders but they are included in the conditioning set.*
3. *The path $E \to B \leftarrow A \to C$ is unblocked given B or D, since B is a collider and D is a descendant of B. However, this path is blocked given $\{A, B\}$ or $\{A, D\}$, since the conditioning set now contains a non-collider A.*



Remark 2. ***(Markov equivalence)***
Two graphs are Markov equivalent if they imply the same set of conditional independence relationships or if they have the same underlying independence structure.
Formally, let's denote two graphs as G_1 and G_2. G_1 and G_2 are Markov equivalent if, for every pair of variables X and Y in the graph, they satisfy the following conditions:

- *X and Y are d-separated in G_1 given Z if and only if they are d-separated in G_2 given Z, for any subset of variables Z.*
- *X and Y are d-separated in G_1 given Z if and only if they are d-separated in G_2 given Z, for any separating set Z, where a separating set Z blocks all possible paths between X and Y.*

In the context of Bayesian networks, the representation of the network as a DAG is of utmost importance. This acyclic property is indispensable as it ensures the coherence of conditional dependencies within the network. In fact, in a Bayesian network, each variable should be conditionally independent of its non-descendants given its parents. The DAG's acyclic nature guarantees that variables do not depend on themselves through circular dependencies. Furthermore, the presence of directed cycles in the graph would lead to circular dependencies, making the joint probability distribution ill-defined.

3.1.2 Markov Random Fields (MRFs)

A Markov Random Field is an undirected graph that represents conditional independence relationships among variables. In an MRF, nodes are connected by undirected edges, and the absence of an edge signifies conditional independence between the connected nodes. MRFs are commonly used in image processing, spatial statistics, and other fields where pairwise interactions are significant. Let $\mathcal{M}$ represent a Markov Random Field with variables $X_1, X_2, \ldots, X_n$. The joint probability distribution of these variables is defined using a Gibbs distribution:

$$\mathbb{P}(X_1, X_2, \ldots, X_n) \propto \exp\left(-\sum_C \Psi_C(\mathbf{X}_C)\right)$$

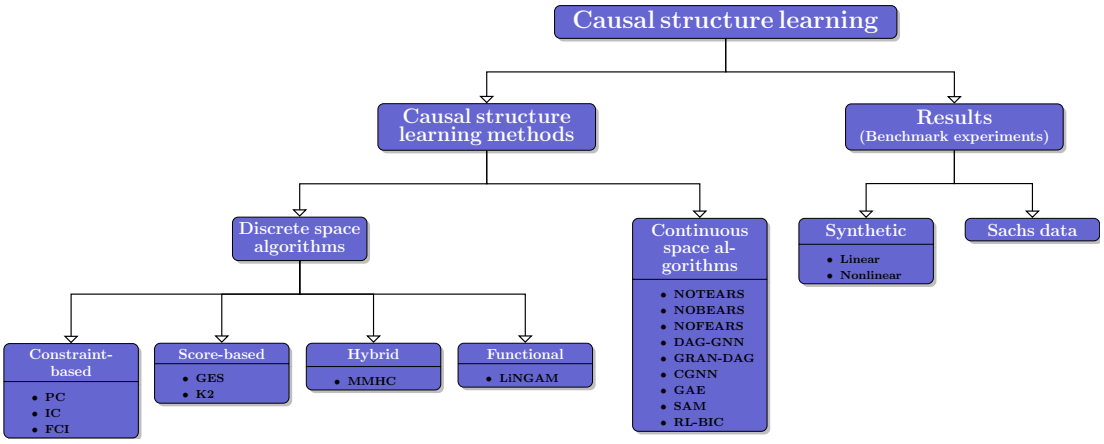
In this equation, $\Psi_C(\mathbf{X}_C)$ represents a potential function defined over cliques (fully connected subsets of nodes) in the graph.

3.2 Causal discovery

Causality discovery involves understanding the causal relationships between variables in a dataset. This knowledge is crucial for making informed decisions, predicting outcomes, and uncovering hidden mechanisms in complex systems. In this section, we will explore different discrete space algorithms used in causality discovery, each with its unique approach to infer causal relationships from data. By discrete space, we refer to algorithms applied to discrete-time Bayesian networks where the random variables take values in a discrete space. These algorithms, corresponding to the left part of Figure 6 can be broadly categorized into three groups: Constraint-based, score-based, and functional-based algorithms. We will specifically delve into the details of three key algorithms used in causality discovery: Chow-Liu, DirectLiNGAM, and Transfer Entropy. This exploration will naturally lead us to discuss the significance of the topological ordering of variables. Topological ordering is crucial for determining the sequence in which variables influence each other. This is particularly crucial in constraint-based methods and, more specifically, in functional methods such as PCMCI and DirectLiNGAM. PCMCI is a functional-based algorithm that assesses time-lagged causal relationships between child variables and their parents using statistical measures, while DirectLiNGAM focuses on identifying causal relationships in observational data through a linear, non-Gaussian causal structure. This order will be necessary for constructing the causal network, demonstrating that each variable is a function of its parents in the causal network.

In the context of Bayesian networks, the notion of topological ordering is pivotal. It is the arrangement of variables in such a way that if there is an edge from variable A to variable B, then variable A precedes variable B in the ordering. This elegant structure adheres to the causal relationships encoded within the network, ensuring that no variable depends on another occurring later in the order. The significance of topological ordering in Bayesian networks extends to several key domains. First and foremost, it is the linchpin of efficient inference. By providing a systematic sequence for variable processing, topological ordering enables streamlined procedures such as variable elimination and message passing, ultimately minimizing redundant computations. This efficiency is indispensable in the analysis of probabilistic models. On another hand, topological ordering empowers causal reasoning. When variables are meticulously ordered, it becomes easier to discern the causal relationships between them. Variables positioned at the beginning of the ordering are often seen as potential causes for those found further along. This structured perspective is invaluable in investigating and understanding the factors influencing a particular variable of

Figure 6: Continuous and discrete space causal learning algorithms



Source : [35]

interest. The application of topological ordering is not limited to static models. In dynamic Bayesian networks (DBNs), it plays a pivotal role in representing the temporal evolution of variables. By adhering to the temporal sequence, DBNs capture how variables change and depend on each other over time, making them essential in modeling dynamic systems. To illustrate the concept, consider a Bayesian network modeling a student's performance. If we have nodes representing *Intelligence*, *Difficulty of Course*, *SAT Score*, and *Grade*, a valid topological ordering might be as follows:

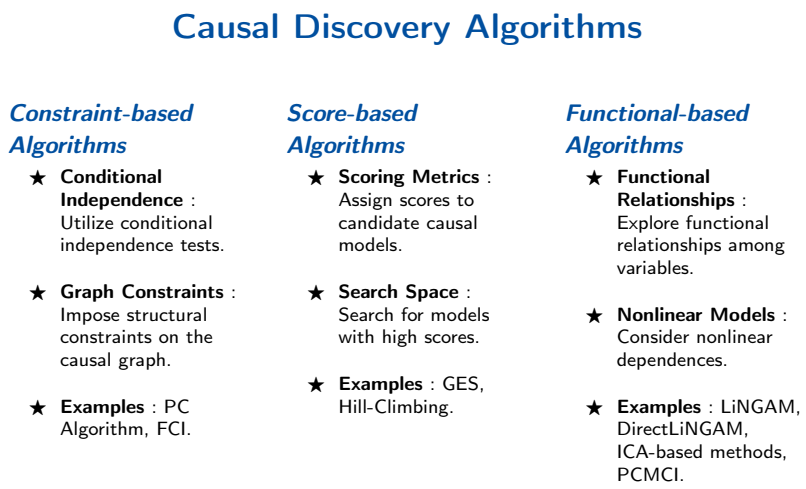
$$\text{Intelligence} \rightarrow \text{Difficulty of Course} \rightarrow \text{SAT Score} \rightarrow \text{Grade}$$

We will present a list of causal discovery algorithms. We will outline these algorithms and go into slightly more detail for the algorithms Chow-Liu, DirectLiNGAM, and Transfer Entropy.

- **Hill-Climbing:** Hill-Climbing is a score-based algorithm that starts with an empty graph and iteratively adds or removes edges to maximize a scoring metric, such as BIC (Bayesian Information Criterion) [3].
- **Chow-Liu:** Chow-Liu is a constraint-based algorithm that constructs a tree structure (tree-shaped Bayesian network) by selecting the most informative conditional independence relationships among variables [3].
- **DirectLinGAM:** DirectLinGAM is a functional-based algorithm that aims to discover causal relationships by considering the linearity and non-linearity of dependencies between variables [43].
- **Normalized Transfer Entropy:** Normalized Transfer Entropy measures information flow between variables and can reveal causal relationships when the transfer entropy is significantly different from zero [38] [39].
- **GES (Greedy Equivalence Search):** GES is a score-based algorithm that explores the space of directed acyclic graphs (DAGs) to find the one that best fits the data by adding or removing edges.

- **PC (Peter and Clark) Algorithm:** The PC Algorithm is a constraint-based approach that infers causality by determining conditional independence relationships among variables and imposing constraints on the graph structure [48].
- **PCMCI:** PCMCI is a functional-based algorithm that assesses time-lagged causal relationships between child variables and their parents using statistical measures [38] [**?**].

Figure 7: Different types of causal algorithms



3.2.1 Chow-Liu algorithm

The Chow-Liu algorithm [3] is a probabilistic graphical model-based approach used to discover causal relationships between variables in a dataset. Its primary idea is rooted in constructing a Bayesian network or a tree structure that represents the conditional independence relationships among variables. The algorithm is particularly effective in situations where we want to uncover the causal structure among variables when we have observational data but lack interventions, making it suitable for many real-world scenarios.

1. **Conditional independence:** The fundamental concept behind the Chow-Liu algorithm is conditional independence. It assumes that if two variables are conditionally independent, there is no direct causal relationship between them. In other words, they are not dependent on each other once we consider the influence of other variables.
2. **Mutual information:** The algorithm quantifies the strength of associations between variables using mutual information. Mutual information measures how much knowing the value of one variable reduces the uncertainty about the other, which is indicative of their statistical dependence. The formula for mutual information $I(X;Y)$ between

two random variables X and Y is given by:

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log \left(\frac{p(x,y)}{p(x)p(y)} \right)$$

3. **Tree structure:** The primary goal of the algorithm is to find a tree structure that represents the conditional independence relationships. This tree is called the "maximum-weight spanning tree" and is derived from the pairwise mutual information scores. It connects the variables in a way that best captures the conditional dependencies.

4. **Causal directions:** While the Chow-Liu algorithm establishes an undirected tree, it does not directly indicate causal directions. However, we can infer potential causal directions based on prior knowledge or by conducting further experiments. For instance, in a causal chain, we might assume that the earlier variables influence the later ones.

Algorithm 1 Chow-Liu Algorithm

Input: Data D with variables $X_1, X_2, \ldots, X_n$
Output: Tree-structured Bayesian network G
Initialize an empty undirected graph G.
Calculate mutual information $MI(X_i, X_j)$ for all pairs of variables.
for each variable X_i in $X_1, X_2, \ldots, X_n$ **do**
  Add X_i to G.
end for
for each pair of variables X_i and X_j in G **do**
  Calculate the mutual information $MI(X_i, X_j)$ based on data D.
  Update the edge weight between X_i and X_j in G to $MI(X_i, X_j)$.
end for
Find a maximum spanning tree T in G using a suitable algorithm (e.g., Kruskal's algorithm).
Orient the edges in T to create a tree-structured Bayesian network G.

3.2.2 DirectLiNGAM

DirectLiNGAM [43] is an algorithm used in causality discovery, that builds upon the LiNGAM (Linear Non-Gaussian Acyclic Model) framework. Its primary objective is to unveil causal relationships within a dataset by modeling linear associations with non-Gaussian additive noise components.

A linear structural causal model (SCM) is defined by the system of structural equations:

$$X = BX + U \tag{1}$$

where $B \in \mathbb{R}^{d \times d}$ is the matrix of coefficients that defines X_i as a linear combination of its parents and the disturbance U_i. Under the assumption of a non-Gaussian distribution of the noise terms, the model is identifiable. This SCM is known as the LiNGAM. DirectLiNGAM advances the LiNGAM framework by enhancing the identification of causal directions based on these relationships and non-Gaussianity. It can estimate a causal graph that illustrates the direction of influence between variables, making it a valuable tool for causal inference. More specifically, assuming that observed data is generated by a DAG, represented by an adjacency matrix $B = \{b_{ij}\}$, each b_{ij} signifying the strength of the connection from variable

x_j to x_i in the DAG. The causal order of variables $k(i)$ ensures that no later variable determines or has a directed path to any earlier variable in the DAG.

These relationships are assumed to be linear. Each observed variable x_i is assumed to have zero mean, and it can be expressed as:

$$x_i = \sum_{k(j)<k(i)} b_{ij} x_j + u_i$$

Here, e_i represents an external influence. All external influences e_i are continuous random variables with non-Gaussian distributions, zero means, and non-zero variances. They are independent of each other, eliminating the presence of latent confounding variables.

This model can be expressed in matrix and we can retrieve the linear SCM defined in equation 1:

$$X = BX + U$$

where X is a p-dimensional random vector, and B can be permuted to become strictly lower triangular through simultaneous row and column permutations due to the acyclicity assumption. Each element b_{ij} in the matrix is a structural coefficient and represents the direct causal effect of x_j on x_i, while each (i, j)-th element of the matrix $A = (I - B)^{-1}$ represents the total causal effect of x_j on x_i. We will delve into the details of causal effects in subsequent sections. It's crucial to note that x_i is equal to e_i if no other observed variable x_j $(j \neq i)$ in the model has a directed edge to x_i. In this case, an external influence e_i is observed as x_i, and this x_i is termed an exogenous observed variable. When there are directed edges from other observed variables to x_i, e_i is considered as an error.

For example, consider the model where:

$$\begin{aligned} x_2 &= e_2 \\ x_1 &= 0.5x_2 + e_1 \\ x_3 &= 0.1x_1 - 2x_2 + e_3 \end{aligned}$$

Here, x_2 is equal to e_2 since it is not influenced by x_1 or x_3. Therefore, x_2 is classified as an exogenous observed variable, while e_1 and e_3 are considered errors. Due to the acyclicity and the assumption of no latent confounders,[1] there will always be at least one exogenous observed variable $(x_i = e_i)$. An exogenous observed variable is typically defined as an observed variable determined outside of the model. In other words, an exogenous observed variable is a variable that any other observed variable inside the model does not have a directed edge to. It doesn't necessarily have to be independent of external influences, and the external influences of exogenous observed variables may be dependent. The DirectLiNGAM algorithm can be summarized as in Algorithm 2.

In Figure 8, we can observe a non-Gaussian linear dynamic model. In this context, each variable is time-indexed, and it's natural to consider that variables at time $t - 1$ precede the variables at time t in terms of a topological order. The coefficients associated with each edge represent direct causal effects and can be estimated using data, provided the model satisfies the identification criteria, which we will discuss in the following sections. These effects quantify the causality from the parent variable to the child variable. For instance, if you increase the value of x_0 at time t by one unit, the value of x_2 at time t will decrease

[1] In causal modeling, confounders are unobserved variables that can affect both the presumed cause and effect, leading to a misleading association between them. We will explore the concept of confounders in the next section, but in essence, confounders are unobserved variables that can introduce bias in causal relationships.

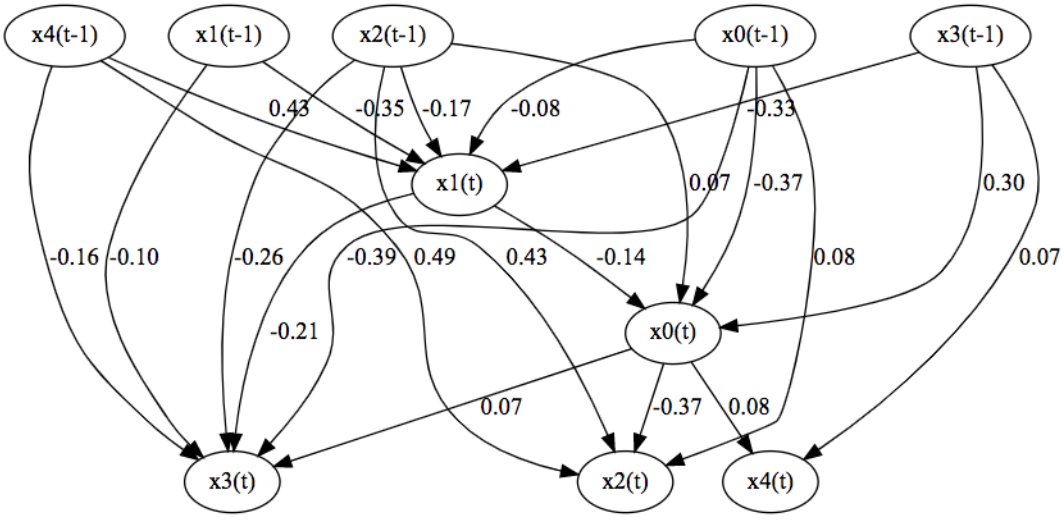
Algorithm 2 DirectLinGAM Algorithm

Require: Observational data matrix $X \in \mathbb{R}^{n \times d}$
Ensure: Causal graph G representing the direction of influence between variables
Preprocessing:
- Center the columns of X to have zero mean.
- Compute the covariance matrix of X: $C = \frac{1}{n} X^T X$.
- Perform independent component analysis (ICA) to estimate non-Gaussian components.

- Obtain the matrix $A = (I - B)^{-1}$, where B is a lower triangular matrix such that $X = BX + E$.

Causal Graph Estimation:
for i in 1 to d **do**
  for j in 1 to d **do**
    if $i \neq j$ **then**
      Compute the causal effect a_{ij} from A.
      if $a_{ij} \neq 0$ **then**
        if $a_{ji} = 0$ **then**
          Add a directed edge $x_j \rightarrow x_i$ to the causal graph G.
        else if $a_{ji} < 0$ **then**
          Add a bidirectional edge $x_i \leftrightarrow x_j$ to the causal graph G.
        end if
      end if
    end if
  end for
end for
return Causal graph G

Figure 8: LinGAM model with identified causal coefficients



by 0.37 units. It's essential to note that these structural coefficients, also known as causal

effects, exist solely within linear models. In cases where each child variable is connected to its parents through nonlinear functions, it becomes necessary to employ methods based on neural networks to estimate these functions.

3.2.3 Normalized Transfer Entropy

The Transfer Entropy (TE) [38] is a powerful tool for detecting non-linear causal relationships, complementing Granger Causality. The Normalized Transfer Entropy (NTE) normalizes the strength of causality to the range [0, 1], which is well-suited for integration into the proposed ensemble model.

Information theory is a prominent research domain for analyzing information flow between two processes in a temporal order. Transfer Entropy is a non-parametric statistical measurement and a fundamental method for inferring non-linear causality connections. It is based on conditional mutual information (CMI) given the past values of the influenced variable. When measuring information using Shannon's entropy, TE from a time series X to another time series Y can be defined as:

$$\mathrm{TE}_{X\to Y} = I(Y_t; X_{t-L:t-1} \mid Y_{t-L:t-1}) = H(Y_t \mid Y_{t-L:t-1}) - H(Y_t \mid Y_{t-L:t-1}, X_{t-L:t-1}), \tag{2}$$

where $\mathrm{TE}_{X\to Y}$ is the TE from X to Y, $I(x)$ represents CMI, and $H(x \mid y)$ represents the conditional Shannon entropy, as given in Equation (3):

$$H(X \mid Y) = -\sum_{x,y} \mathbb{P}(x,y) \log \mathbb{P}(x \mid y). \tag{3}$$

Here, $\mathbb{P}(x,y)$ is the joint probability density function, and $\mathbb{P}(x \mid y)$ denotes the conditional probability density. When $\mathrm{TE}_{X\to Y} > 0$, it indicates that X is the cause of Y, and the causal strength becomes stronger with an increase in transfer entropy.

Algorithm 3 Normalized Transfer Entropy for Causality

Require:
Two time series: X and Y.
Maximum time lag L.
Number of bins N for discretization.
A significance level α.
Ensure:
Normalized Transfer Entropy $\mathrm{NTE}_{X\to Y}$.

Procedure NormalizeTransferEntropy
Input: X, Y, L, N, α
Output: $\mathrm{NTE}_{X\to Y}$

Discretize X and Y into N bins.
Compute conditional and joint probability distributions.
for $l = 1$ to L **do**
    Compute transfer entropy $\mathrm{TE}_{X\to Y}(l)$ as in Equation 2.
    Compute transfer entropy $\mathrm{TE}_{Y\to X}(l)$ by swapping X and Y.
end for
Determine if NTE is statistically significant at α.
Return $\mathrm{NTE}_{X\to Y}$.

3.3 Structural causal model

Bayesian networks offer an elegant framework for capturing probabilistic dependencies among variables, enabling inferences and predictions in uncertain environments. However, when it comes to understanding the underlying causal relationships within these systems and answering counterfactual questions–what would happen if we intervened on a particular variable–Bayesian networks have limitations. To address these limitations and introduce causal effects, it becomes imperative to define structural models. As an extension of Bayesian networks, structural models consist of structured functions that explicitly relate child variables to their parents, capturing the causal relationships between them. These functions, representing causal effects, play a crucial role in providing a foundation for exploring and answering counterfactual queries. A counterfactual distribution is defined as the probability distribution of a random variable in a hypothetical scenario different from the observed data. The modification only operates on the parents of this variable on the graph. This is where SCMs [32], also known as graphical causal models (GCMs), come into play.

We start first by pointing out the difference between probabilistic graphical models (PGMs) and GCMs. A PGM [21] is a model that combines graph theory with probability theory to develop new algorithms and present models in an intuitive framework. A probabilistic graphical model is a DAG over variables, representing how the joint distribution over these variables can be factorized. Notably, any missing edge in the graph implies a conditional independence relation in the joint distribution. Multiple valid probabilistic graphical model representations exist for a given joint distribution. For instance, any joint distribution over two variables (X, Y) can be represented as both $X \to Y$ and $X \leftarrow Y$. A CGM is a probabilistic graphical model with the additional assumption that a link $X \to Y$ indicates that X causes Y. This additional assumption implies naturally a topological ordering of variables in the Bayesian Network. CGMs are a natural extension of Bayesian networks, designed explicitly to address causal questions and provide insights into the mechanisms driving observed data. They consist of a collection of structural equations, each defining how a variable is causally influenced by its direct predecessors within the model. SCMs encompass various types of models, including structural functional models, where the functions denoted by f can be nonlinear, capturing complex causal relationships, and linear models, where f is linear, simplifying the representation while retaining interpretability. Here is a formal definition of a structural causal model :

Definition 2. ***Causal Model***
A causal model over a set of variables V is a tuple $M = \langle V, U, F, P_a, P(U) \rangle$, where:

- *$V = \{V_1, V_2, \ldots, V_n\}$ is a set of n variables that are determined by the model (endogenous or observed variables).*
- *U is a set of random variables that are determined outside the model (exogenous or unobserved variables) but that can influence the rest of the model.*
- *F is a set of n functions such that $V_i = f_i(P_a(V_i), U_i)$.*
- *$P_a(V_i)$ is a subset of $V \setminus \{V_i\}$ (observed parents of V_i).*
- *U_i is a subset of U (unobserved parents of V_i).*
- *$P(U)$ is a joint probability distribution over the variables in U.*

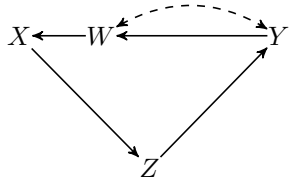
A causal model has an associated graph in which each observed variable V_i corresponds to a vertex. There is one edge pointing to V_i from each of its observed parents $P_a(V_i)$, and there is

a doubly-pointed edge between pairs of vertices influenced by a common unobserved parent in U. In other words, in a causal model, the probability distribution of each variable V_i is assigned by a function f_i determined by a subset of $V \setminus \{V_i\}$ called the observed parents of V_i ($P_a(V_i)$) and a subset of U (U_i) called the unobserved parents of V_i. The joint probability distribution of the observed variables in a causal model M is given by [7]:

$$\mathbb{P}(V) = \mathbb{P}(V_1, V_2, \ldots, V_n) = \prod_U \prod_i \mathbb{P}(V_i \mid P_a(V_i), U_i) \prod_i \mathbb{P}(U_i) \tag{4}$$

The graphical representation of a causal model is also called the induced graph of the causal model or causal graph. It contains vertices V_i, edges from $P_a(V_i)$ to V_i, and bidirected edges between pairs of vertices influenced by a common unobserved variable, that is, between V_i and V_j if $U_i \cap U_j \neq \emptyset$. We refer to the unobserved variables U as hidden confounders.

Figure 9: Causal graph with vertices representing variables X, Y, W, Z, edges representing functions $X = f_1(W)$, $Z = f_2(X)$, $W = f_3(U)$, $Y = f_3(W, Z, U)$. The hidden confounder that has an effect on W and Y is represented by the double dashed edge.



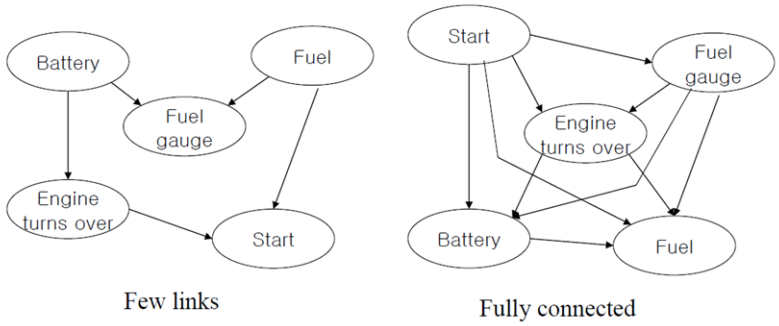
A SCM is therefore associated with a directed graph G where there is an edge $W \to V$ if and only if W is in the set of observed and unobserved parents of V, denoted as $\mathrm{Pa}^*(V)$. We will focus on recursive SCMs, where the set F defines a topological order for the variables $V \cup U$. This order ensures that $\mathrm{Pa}^*(Z) < Z$ for all $Z \in V \cup U$. The graph associated with a recursive SCM is a DAG. Notation $\tau(V)$ refers to the variables that precede V in the topological order, indicating variables $W \in V \cup U$ such that $W < V$. Therefore, each variable V_i can be expressed as :

$$V_i = f_i(\tau(V_i)).$$

We assume that variables in set U precede those in set V in the topological order. Notation $\mathrm{Pa}(V)$ is a concise representation for $\mathrm{Pa}^*(V) \cap V$, denoting the observed parents of V. This topological ordering is crucial for understanding causal relationships. In fact, as shown in Figure 10, both causal graphs represent the same joint probability distribution. However, altering the order of the variables leads to different interpretations of causal relationships.

Causal graphs encode causal relations between variables in a model. The primary purpose of causal graphs is to help estimate the joint probability of some of the variables in the model upon controlling some other variables by forcing them to specific values; this is called an action, experiment, or intervention. For every probabilistic model M, every set of variables $X \subset V$ and every set of values $X = x$ we define the model $M_{\mathrm{do}}(X = x)$ to be the same as M except that every function f_i for variable $X_i \in X$ assigns a probability distribution of 1 to the value x_i and 0 to the rest of values. Graphically, this is represented by removing all the incoming edges (which represent the causes) of the variables in the graph that we control in the intervention. In this scenario, these variables are considered exogenously determined, no longer influenced by their previous causes. In this way, a CGM encodes more than just the factorization or conditional independence structure of the joint distribution among its variables; it also defines how the system responds to atomic interventions. Mathematically, the do() operator represents this intervention on the variables, by transforming M into

Figure 10: Topological ordering of variables changed



$M_{\text{do}}(X = x)$. For a causal graph with sets of variables X and Y, the expression $\mathbb{P}(Y \mid \text{do}(X = x))$ represents the joint probability of Y under an intervention on the controlled set X. This rigorously corresponds to applying equation 4 to $M_{\text{do}}(X = x)$ instead of M. As elucidated by Pearl, the do() operator acts like conditioning on a *mutilated* graph and performs calculus on this truncated graph. A causal relation represented by the expression $\mathbb{P}(Y \mid \text{do}(X = x))$ is said to be identifiable if it can be uniquely determined from the graph G induced by causal model M, and from the joint distribution P of its observed variables. We will provide more details about the do-calculus in the subsequent section.

We can extend the defintion of a structural causal model by adding noise components :

Definition 3. ***Causal model with driving noise***
More common than the above approach is the assumption that the randomness enters inside the structural equations. Formally, a stochastic structural causal model over n random variables $V_1, \ldots, V_n$ is a set of n functions such that

$$V_i = f_i(P_a(V_i), U_i, \varepsilon_i), \quad i = 1, \ldots, n, \tag{5}$$

together with a distribution over the noise variables $\varepsilon_1, \ldots, \varepsilon_n$.

We obtain a corresponding graphical representation of the causal structure over the vertices $(1, \ldots, n)$ by drawing directed edges from $P_a(V_i)$ to V_i for all $i \in \{1, \ldots, n\}$. We further assume that the joint noise distribution is absolutely continuous with respect to a product measure and that it factorizes, i.e., the noise components are assumed to be jointly independent. As before, we require the system (5) to be uniquely solvable, which is always satisfied if the graph is acyclic. LinGAM is a particular case of a causal model with driving noise where the function f_i are linear (leading to structural coefficients) and the noises have non-Gaussian distributions.

We can also extend the causal models defined above by considering time as a factor. This leads us to dynamic causal models, consisting of a set of stochastic differential equations:

Definition 4. ***Dynamic causal model***
Formally, a dynamic stochastic structural causal model over n random variables $V_1, \ldots, V_n$ is a set of n functions such that [33]:

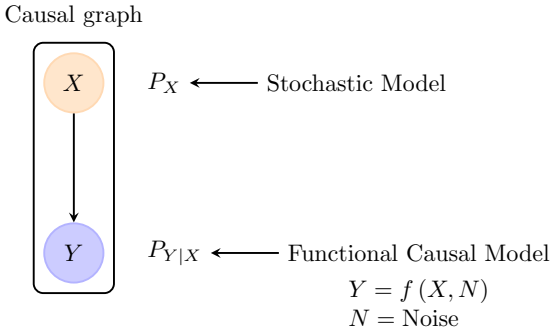
$$dV_t^i = f^i(P_a(V_t^i), V_t^i, U_t^i)dt + \sigma(P_a(V_t^i), V_t^i, U_t^i)d_{\epsilon_t^i}, \quad i = 1, \ldots, n$$

where ϵ_i is a Wiener (white noise) process. In the case where time evolution is assumed to be deterministic, we get structural equations of the form:

$$\frac{dV_t^t}{dt} = f_i(P_a(V_t^i), V_t^i, U_t^i), \quad i = 1, \ldots, n$$

The functions $\{f_i\}$ can be referred to collectively as the evolution function of the system.

Figure 11: Structural causal model with driving noise

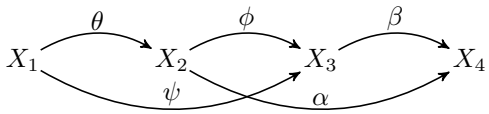


For the subsequent part of this article, we will explore situations where these functions take on a linear form. The rationale behind this focus lies in the fact that the presence of non-linearity in the structural equations can introduce complexities into their estimation. When faced with such scenarios, it becomes essential to employ neural networks for the estimation of causal effects. To keep matters simple, we will confine ourselves to linear functions. These linear functions yield structural coefficients that provide a quantification of the causal impact one variable has on another. We will introduce structural coefficients with an example. Let's consider a scenario where we aim to understand the factors contributing to a student's final grade. It is evident that analyzing study hours alone won't provide a comprehensive understanding of the determinants of success. This is because other factors, such as prior knowledge, also play a significant role. To represent this relationship, we can utilize structural equations. These equations capture how the variables directly influence one another:

$$\begin{aligned} X_1 &= \epsilon_1 \\ X_2 &= \theta \cdot X_1 + \epsilon_2 \\ X_3 &= \phi \cdot X_2 + \psi \cdot X_1 + \epsilon_3 \\ X_4 &= \alpha \cdot X_2 + \beta \cdot X_3 + \epsilon_4, \end{aligned}$$

In this model, X_1 represents prior knowledge, X_2 represents study hours, X_3 represents study habits, and X_4 represents the student's final grade. The terms $\epsilon_1, \epsilon_2, \epsilon_3$, and ϵ_4 represent the error terms of the variables X_1, X_2, X_3, and X_4 respectively. The coefficients $\theta, \psi, \alpha, \phi$, and β are referred to as structural coefficients. These coefficients are derived from the linear functions f_i in a way that $X_i = f_i((Pa(X_i), \epsilon_i)$. These structural equations offer insights into how changes in one variable can propagate through the system, influencing other variables in a direct and interconnected manner. For example, if we increase the value of X_2 by one unit, the value of X_4 will change by α units. Assuming $\alpha = 2$, this implies that if a student increases their study hours by 10%, their grade will also increase by 20%. The structural model defined above has a graphical representation. It takes the form of a causal graph, as depicted in Figure 12. Each variable in the model has a corresponding node or vertex in the graph. Additionally, for each equation, arrows are drawn from the independent variables to the dependent variables, indicating the causal relationships.

Figure 12: Causal graph representing the SEM specification Model 1.



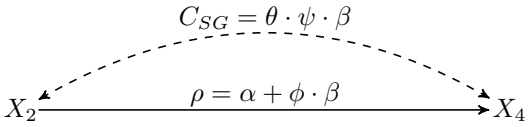
On each arrow, we notice the presence of the structural coefficients: θ represents the direct causal effect of X_1 on X_2, ϕ represents the direct causal effect of X_2 on X_3, β represents the direct causal effect of X_3 on X_4. Same thing applies to α and ψ.

Now, let's explore the concepts of direct and total causal effects using this example:

- **Direct causal effect**: α represents the direct impact of study hours X_2 on the student's final grade X_4.
- **Total causal effect**: The coefficient ρ represents the total impact of study hours X_2 on the student's final grade X_4. This includes both the direct effect of X_2 on X_4 and any indirect effects through study habits X_3. To formalize this concept, the total causal effect of a variable i on a variable j can be expressed as the product of the structural coefficients along the active paths.

If we want to visualize the equivalent graph representing the causal effect between the observed variables X_2 and X_4, assuming that the variables X_1 and X_3 are not observed, we obtain Figure 13. The effect of study hours on grade is now summarized by the coefficient ρ. Similarly, the bi-directed arc between X_2 and X_4 (representing the correlation of the error terms ϵ_2 and ϵ_4) summarizes the correlation between X_2 and X_4 due to the path $X_2 \leftarrow X_1 \rightarrow X_3 \rightarrow X_4$ and therefore depends on the parameters θ, ψ, and β .

Figure 13: Graph representing the causal effect of S on G, assuming K and H to be latent variables.



In order to estimate ρ, the total causal effect of the number of study hours on a student's final grade, the coefficients must have a unique solution in terms of the covariance matrix or probability distribution over the observed variables, X_2 and X_4. The task of finding this solution is known as identification and is discussed later. In some cases, one or more coefficients may not be identifiable, meaning that no matter the size of the dataset, it is impossible to obtain point estimates for their values.

3.4 Do-calculus

The do-calculus [32] is a fundamental concept in the domain of causal inference and probabilistic modeling. It enables us to explore the consequences of interventions on variables within a probabilistic model. These interventions are the key to answering what-if questions, allowing us to understand hypothetical/counterfactual scenarios. As we will explore in the following section, the do-calculus also provides a set of rules and techniques for estimating causal effects from observational and interventional data, helping us quantify the causal

relationships between variables in a system. The following notes are inspired from [25], [51], [32] and [28].

Consider a scenario where a patient's health is in question, and a medical practitioner is tasked with prescribing a dosage of a particular medication. What if the physician had chosen to prescribe a 75 mg dosage of Zocor rather than the usual 40 mg? What impact would this intervention have for example on the patient's cholesterol levels? This is precisely where the power of interventions comes into play. When we intervene on a variable in a causal model, such as the medication dosage, we make it deterministic and the chosen dosage becomes a fixed value. As a result, all arrows or causal pathways that point to this variable are severed in the causal graph. This abrupt change in the variable's value triggers a ripple effect throughout the probabilistic model. The joint probability distribution of the model undergoes a transformation, reflecting the consequences of this intervention, and we obtain a post-intervention distribution derived from the original joint distribution. In essence, the intervention acts as a perturbation within the causal network, creating a shift in the causal relationships that define the system. This concept is akin to modifying the course of events in a complex web of causation. It allows us to discern how changing one variable can lead to a cascade of effects throughout the model, providing answers to "what-if" questions. It is this power to explore and quantify causal effects that makes do-calculus an invaluable tool in fields such as healthcare, economics, and social sciences, where interventions can lead to meaningful insights and informed decision-making. In this section, we will delve into the mathematical foundations of do-calculus, elucidating how it allows us to rigorously study causal relationships and quantify the impact of interventions on complex probabilistic models. First, we need to introduce the do-operator. The do-operator, denoted as $do(X = x)$, is used to represent interventions or changes applied to a variable in a causal graph. It signifies that we are setting the value of variable X to x, regardless of its actual value in the observed data. Consider a causal graph with variables X, Y, and Z, where arrows indicate causal relationships. Let's express the effect of an intervention using the do-operator:

$$\begin{aligned} &\text{Causal graph:} \quad X \to Y \to Z \\ &\text{Effect of intervention:} \quad do(X = x) : Y(x), Z(x) \end{aligned}$$

In this graph, the arrow from X to Y implies a direct causal effect of X on Y. Similarly, the arrow from Y to Z implies a direct causal effect of Y on Z. To express the effect of intervening on X and setting it to a specific value x, we use the do-operator. Consider a study investigating the effect of a new drug X on blood pressure Y and heart rate Z. The causal graph is as follows:

$$\text{Causal graph:} \quad X \to Y \to Z$$

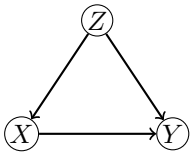
We want to know the effect of administering the new drug ($X = 1$) on blood pressure and heart rate. We express this using the do-operator:

$$\text{Effect of intervention:} \quad do(X = 1) : Y(1), Z(1)$$

By applying the do-operator, we are simulating an intervention where the drug is administered, regardless of the individual's actual drug exposure in the observed data. The do-operator helps us distinguish between natural associations (based on observed data) and causal effects (based on interventions) in causal graphs.

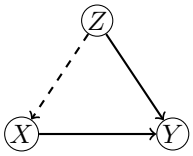
Intervening on a variable X corresponds to removing the incoming edge to X in the causal graph, effectively disconnecting X from its parent. Let's illustrate this concept with the help of causal graphs:

Figure 14: Original causal graph



In the original causal graph, we see that X is causally connected to Z and, in turn, influences Y. However, if we were to intervene and set X to a specific value, say $\text{do}(X = x)$, the causal graph changes significantly:

Figure 15: Causal graph after intervention $\text{do}(X = x)$



The intervention, represented by $do(X = x)$, severs the causal link between Z and X, rendering X as an exogenous variable with a fixed value. Consequently, the influence of X on Y becomes direct and deterministic, untethered from its previous causal pathway through Z. This exemplifies the transformative power of the do-operator in reshaping causal relationships, allowing us to explore the consequences of specific interventions in a causal system. When intervening on a variable X, it is assumed that everything else in the system remains unchanged, in particular the functions or conditional distributions that determine the value of a variable given its parents in the graph. Answering causal queries such as : what would the distribution of the patient's cholesterol level look like if we were able to prescribe 75 mg dosage of Zocor ? requires inference about the distributions of variables in the post-interventional system. The do-notation is a short-hand for describing the distribution of variables post-intervention, and the do-calculus is a set of three rules for identifying which (conditional) distributions are equivalent pre and post-intervention. If it is possible to derive an expression for the desired post-interventional distribution purely in terms of the joint distribution over the original system via the do-calculus, then the causal query is said to be identifiable, meaning assuming positive density and infinite data we obtain a point estimate for it. The do-calculus is complete : a query is identifiable if and only if it can be solved via the do-calculus. The reduction rules identifying equivalence between the conditional distributions pre and post-intervention are as follows :

Let G be a CGM. Let $do(x)$ represent intervening to set a single variable X to x.

- Rule 1: $\mathbb{P}(Y \mid do(x), z, w) = \mathbb{P}(Y \mid do(x), w)$ if $Y \perp Z \mid (X, W)$ in $G_{\overline{X}}$.
- Rule 2: $\mathbb{P}(Y \mid do(x), do(z), w) = \mathbb{P}(Y \mid do(x), z, w)$ if $Y \perp Z \mid (X, W)$ in $G_{\overline{X},\underline{Z}}$.
- Rule 3: $\mathbb{P}(Y \mid do(x), do(z), w) = \mathbb{P}(Y \mid do(x), w)$ if $Y \perp Z \mid (X, W)$ in $G_{\overline{X},\overline{Z(W)}}$.

The reduction rules are based on the structure of the DAG and d-separation. Given a set of variables X, Y, Z, W, and a graph G, the rules allow the values of some of the parent variables to be ignored for some configurations of G. $G_{\overline{X}}$. represents the graph G with the incoming links of Z removed. Rule 1 states that if, under such a graph (corresponding to the do intervention on X), Y and Z are independent given evidence X and W, it means that Z has no impact on Y and can be ignored. $G_{\overline{X},\underline{Z}}$ represents the graph G with the

incoming links of X and the outgoing links of Z removed. The do(Z) operation forces the value of Z to ignore the confounders that can affect Z. If, after removing the direct path from Z to Y, Y and Z are independent, we can deduce that there are no such confounders. Therefore, rule 2 states that an intervention on Z has no effect and can be considered an observation. $G_{\overline{X},\overline{Z(W)}}$ is the graph G where the incoming links of X are removed and the incoming links of Z are removed if Z is not an ancestor of W. If under such a graph, Y and Z are independent, all paths between Z and Y pass through W or X. Therefore, rule 3 states that the value of Z has no effect on Y and can be ignored. These rules allow us to estimate the values of some variables under interventional settings using observations only. Whether or not the query can be answered depends on the local identifiability of the graph. A graph is identifiable if and only if it can be reduced to observations using the three rules of do-calculus. Various identifiability criteria will be explored later, with the most well-known being the single-door and back-door criteria. In broad terms, identification allows us to determine whether we can nonparametrically discern the causal effect on the outcome variable, given a set of variables X and specific conditions.

As discussed earlier, the concept of intervention induces changes in the properties of DAGs, as intervening on a variable severs all incoming edges to that variable. Consequently, it becomes necessary to redefine causal Bayesian networks (CBNs), as interventions alter the conditional independence structure: In the context of Bayesian Networks, the conditional independence structure encapsulates the relationships between variables in the absence of any interventions. However, when interventions are introduced, the conditional independence relations may be modified, necessitating a reevaluation of the network. We provide the following notations :

- V: Set of nodes (variables) in the DAG
- X: Set of nodes in V (i.e., $X \subset V$), with the condition that $X \neq V$.
- $\mathbb{P}(V)$: Joint probability distribution over the variables of V
- $\mathbb{P}(V_i) = \mathbb{P}(V_i|do(X = x))$: Set of all interventional distributions, including no intervention $\mathbb{P}(V_i)$.

The technical definition of a CBN for a DAG G compatible with $\mathbb{P}$ requires the following three conditions to hold for every $\mathbb{P}(V_i \mid do(X = x)) \in \mathbb{P}$:

- $\mathbb{P}(V_i \mid do(X = x))$ is compatible with G. This implies that even after intervention, G can represent $\mathbb{P}(V \mid do(X = x))$. Removing incoming edges into X creates new independencies in the graph, but these do not affect Markov compatibility.
- $\mathbb{P}(V_i \mid do(X = x)) = 1, \forall V_i \in X$ whenever $V_i = v_i$ is consistent with $X = x$. Intervening on the same variable as on the left of the conditioning bar collapses it to a point mass : 0 or 1.
- $\mathbb{P}(V_i \mid do(X = x, P_a(V_i)) = \mathbb{P}(V_i \mid P_a(V_i), \forall V_i \notin X$ whenever $PA(V_i)$ is consistent with $X = x$. Interventions have no effect on the conditional probability distribution when V_i is conditioned on its parents $P_a(V_i)$.

The consequences of the CBN definition include a simplified factorization of $\mathbb{P}(V \mid do(X = x))$, which drops factors related to nodes intervened on ($V_i \in X$) and we obtain the following expression :

$$\mathbb{P}(V \mid do(X = x)) = \prod_{i:V_i \notin X} \mathbb{P}(V_i \mid P_a(V_i)). \tag{6}$$

When we perform an intervention $\text{do}(X = x)$ on a variable X within a probabilistic model, as illustrated in Figure 15, we alter the probability distribution of the system. The new distribution, denoted as $\mathbb{P}(Y \mid do(X = x))$, reflects the outcomes for other variables, such as Y, under the specified intervention. If the causal model is identifiable, which is the case for the model depicted in Figure 14, it becomes possible to express the post-intervention distribution mathematically:

$$\mathbb{P}(Y \mid do(X = x)) = \sum_z \mathbb{P}(Y \mid X = x, Z = z)\mathbb{P}(Z = z). \tag{7}$$

Here, Z represents the set of variables that satisfy the back-door criterion (we will see this criterion in the subsequent section in Theorem 2). That is, no member of Z is a descendant of X, and Z d-separates X from Y in the sub-graph formed by deleting all arrows emanating from X. This set of variables allows the model to be identifiable, and thus, it is possible to derive the post-intervention distribution in terms of the pre-intervention conditional distributions. In Figure 14, the variable Z satisfies the back-door criterion, enabling us to derive an expression for the conditional distribution of Y in the model $M_{do}(X = x)$. Equation 7 arises from the fact that variable X becomes deterministic. Therefore, the joint distribution $\mathbb{P}(Y, Z, X = x)$ factorizes into the conditional distribution of each variable given its parents, which is $\mathbb{P}(Y \mid Z, X = x) \cdot \mathbb{P}(Z)$ since the edge from Z to X is removed. To derive the distribution of Y, we need to sum the joint distribution over Z only since X has a fixed value, resulting in the corresponding equation. The expected value of the outcome variable Y under the intervention $\text{do}(X = x)$ is then:

$$\begin{aligned}\mathbb{E}[Y \mid \text{do}(X = x)] &= \sum_y y\, \mathbb{P}(Y \mid \text{do}(X = x)) \\ &= \sum_y y \sum_z \mathbb{P}(Y \mid X = x, Z = z)\mathbb{P}(Z = z)\end{aligned}$$

More generally, the classical way to compute the conditional distribution $\mathbb{P}(X_k \mid \text{do}(x_i))$, even if the model is not identifiable, is as follows :

- **Step 1:** Calculate the joint probability $\mathbb{P}(X_1, X_2, \ldots, X_n)$ under the intervention $\text{do}(x_i)$ using the formula obtained in equation 7:

$$\mathbb{P}(X_1, X_2, \ldots, X_n) = \prod_{j \neq i} \mathbb{P}(X_j \,|\, \text{Pa}_j(x_i))\delta_{x_i}$$

  Where:

  - $\text{Pa}_j(x_i)$ represents the parents of X_j when X_i is intervened with x_i.

- **Step 2:** Marginalize $\mathbb{P}(X_1, X_2, \ldots, X_n)$ to obtain $\mathbb{P}(X_k \mid do(x_i))$ by summing over all other variables except X_k:

$$\mathbb{P}(X_k \mid do(x_i)) = \sum_{X_1} \sum_{X_2} \cdots \sum_{X_{k-1}} \sum_{X_{k+1}} \cdots \sum_{X_n} \mathbb{P}(X_1, X_2, \ldots, X_n)$$

Of course, we can adapt these definitions in the context of continuous variables : We simply replace sums by integrals.

Causality, or the idea of cause and effect, has a complicated history and has sparked debates in the field of statistics and mathematics. One fundamental question is whether we

can effectively use probability theory to address causal problems or if we need an entirely new mathematical framework, such as the do-calculus, to handle these issues. This debate has been ongoing for some time. As long as we are clear about the assumptions we make regarding the impact of intervening in a system, we can estimate causal effects using the standard Bayesian approach. In other words, it may not be necessary to introduce entirely new mathematical tools like the do-calculus to tackle causal problems [37] [24]. As said in [25], *while it is critical to explicitly model our assumptions on the impact of intervening in a system, provided we do so, estimating causal effects can be done entirely within the standard Bayesian paradigm.* The assumptions that underlie causal graphical models can be effectively represented using a type of mathematical model known as PGMs. By using PGMs, Bayesian practitioners, who are familiar with this framework, can better represent and reason about the assumptions required for causal inference. However, one potential downside to explicitly modeling causal questions as a single PGM is that it can be more complex and computationally intensive compared to using the do-calculus for appropriate re-parameterizations. The do-calculus is a mathematical tool specifically designed for causal inference in the post-intervention world, making it more efficient in some cases.

3.5 Identification

In this section, we will define mathematically the causal effects using the do-operator and we'll provide some criteria for identification. Causal effects quantify the change in one variable that results from a change in another variable. In the context of causal graphs and SEMs, there are different types of causal effects, including direct causal effects and total causal effects. It's worth noting that this section is comprehensive and challenging, and it may be opportune to avoid delving too deeply into it. Therefore, readers who are looking for a more concise overview or are seeking to focus on other aspects of the material may consider skipping this section. The content of this section is inspired from [15], [51], [32], [5] [27] and [34].

3.5.1 Causal effects

Definition 5. ***Total causal effect*** *[32]*

Let $\Pi = \{\pi_1, \pi_2, \ldots, \pi_k\}$ be the set of directed paths from X to Y, and p_i be the product of the structural coefficients along path π_i. The total effect or average causal effect (ACE) of X on Y is defined as $\sum_i p_i$.
The reason for this additive formula and its extension to non-linear systems can best be seen if we define the total causal effect of X on Y as the expected change in Y when X is assigned to different values by intervention, as in a randomized experiment. The act of assigning a variable X to the value x is represented by removing the structural equation for X and replacing it with the equality $X = x$. This replacement dislodges X from its prior causes and ensures that causality between X and Y reflects causal paths from X to Y only. The expected value of a variable, Y, after X is assigned the value x by intervention is denoted $\mathbb{E}[Y \mid do(X = x)]$, and the ACE of X on Y is defined as

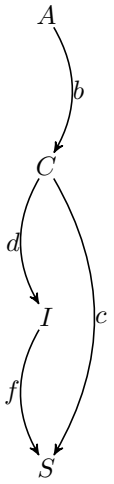
$$ACE = \mathbb{E}[Y \mid do(X = x + 1)] - \mathbb{E}[Y \mid do(X = x)], \quad \text{where } x \text{ is some reference point.}$$

In nonlinear systems, the effect will depend on the reference point, but in the linear case, x will play no role, and we can replace the equation above with the derivative,

$$ACE = \frac{\partial}{\partial x}\mathbb{E}[Y \mid do(X = x)].$$

Let's consider the causal graph presented in Figure 16. A represents the strength of a student's application, C is the variable indicating whether the student attends an elite college or not, I represents the variable indicating the quality of an internship, and S represents the salary of the student when he graduates from college. The total effect of C on S is $c + d \cdot f$, and that of A on S is $b \cdot (c + d \cdot f)$.

Figure 16: Causal relationships for the student case



The structural equation associated with this graph is given by:

$$\begin{aligned} A &= \epsilon_A \\ C &= b \cdot A + \epsilon_C \\ I &= d \cdot C + \epsilon_I \\ S &= c \cdot C + f \cdot I + \epsilon_S \end{aligned} \tag{8}$$

Let's compute the total effect of C on S by means of the expectation and see if we get back to the corresponding result : Suppose C is a binary variable taking value 1 for elite colleges and 0 for non-elite colleges. To estimate the total effect of attending an elite college on salary, we would hypothetically assign each member of the population to an elite college and observe the average salary, $\mathbb{E}[S \mid do(C = 1)]$. Then we would rewind time and assign each member to a non-elite college, observing the new average salary, $\mathbb{E}[S \mid do(C = 0)]$. Intuitively, the causal effect of attending an elite college is the difference in average salary,

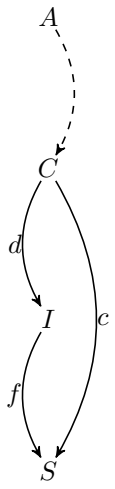
$$\mathbb{E}[S \mid do(C = 1)] - \mathbb{E}[S \mid do(C = 0)].$$

The above operation provides a mathematical procedure that mimics this hypothetical (and impossible) experiment using a structural equation model. The intervention $do(C = c_0)$ modifies the equations in the following way:

$$\begin{aligned} A &= \epsilon_A \\ C &= c_0 \\ I &= d \cdot C + \epsilon_I \\ S &= c \cdot C + f \cdot I + \epsilon_S \end{aligned}$$

The corresponding causal graph is displayed in Figure 17. Notice that back-door paths, due to common causes, between C and S have been cut, and as a result, all unblocked paths between C and S now reflect the causal effect of C on S only.

Figure 17: Causal relationships for student case after intervention $C = c_0$



We assume model variables have been standardized to mean 0 and variance 1, implying that $\mathbb{E}[\epsilon_i] = 0$ for all i. We see that setting C to c_0 gives the following expectation for S:

$$\begin{aligned}\mathbb{E}[S \mid \text{do}(C = c_0)] &= \mathbb{E}[c \cdot C + f \cdot I + \epsilon_S] \\ &= c \cdot \mathbb{E}[C] + f \cdot \mathbb{E}[I] + \mathbb{E}[\epsilon_S] \\ &= c \cdot c_0 + f \cdot \mathbb{E}[d \cdot C + \epsilon_I] \\ &= c \cdot c_0 + f \cdot d \cdot c_0 + f \cdot \mathbb{E}[\epsilon_I] \\ &= c \cdot c_0 + f \cdot d \cdot c_0.\end{aligned}$$

As a result,

$$\mathbb{E}[S \mid \text{do}(C = c_0 + 1)] - \mathbb{E}[S \mid \text{do}(C = c_0)] = c + f \cdot d$$

which coincides with the initial result.

Definition 6. ***Direct causal effect*** *[32]*

We saw that the total causal effect of one variable on another encompasses all causal pathways, including both direct and indirect links, that lead to changes in the dependent variable. It considers the effect caused only by altering the independent variable. Conversely, the direct causal effect of one variable on another is the change that occurs in the dependent variable when the independent variable is altered, while keeping all other variables constant. This means that we're isolating the impact of the independent variable to observe how it directly affects the dependent variable.

We thus define the direct causal effect of X on Y is as

$$DCE = \mathbb{E}[Y \mid do(X = x + 1, Z = z)] - \mathbb{E}[Y \mid do(X = x, Z = z)],$$

where Z is a set containing all variables other than X and Y, x is some reference point, and z is a set of reference values. In nonlinear systems, the effect will depend on the reference

point, but in the linear case, x will play no role, and we can replace the equation above with the derivative,

$$DCE = \frac{\partial}{\partial x}\mathbb{E}[Y \mid do(X = x, Z = z)].$$

Let's consider again the causal graph given by figure 16. We want to compute, by the means of the expectation, the direct causal effect of C on S, which is obviously equal to c. "keeping all other variables constant" can be simulated by intervening on all variables other than C and S and assigning them an arbitrary set of reference values. Doing so removes all causal links in the model other than those leading into S. As a result, all links from C to S other than the direct link will be removed. Figure 18 shows the path diagram after intervention on all variables other than C and S.

Figure 18: Causal relationships for student case after intervention $C = c_0, A = a, I = i$

Now, the direct effect of C on S can be defined as $\mathbb{E}[S \mid \text{do}(C = c_0 + 1, Z = z)] - \mathbb{E}[S \mid \text{do}(C = c_0, Z = z)]$, where Z is a set containing all model variables other than C and S, and $\{c_0 \cup z\}$ a set of reference values. We have then :

$$\begin{aligned}\mathbb{E}[S \mid \text{do}(C = c_0, I = i, A = a)] &= \mathbb{E}[c \cdot C + f \cdot I + \epsilon_S] \\ &= c \cdot \mathbb{E}[C] + f \cdot \mathbb{E}[I] + \mathbb{E}[\epsilon_S] \\ &= c \cdot c_0 + f \cdot i\end{aligned}$$

As a result,

$$\mathbb{E}[S \mid \text{do}(C = c_0 + 1, I = i, A = a)] - \mathbb{E}[S \mid \text{do}(C = c_0, I = i, A = a)] = c$$

In earlier sections, we highlighted the necessity for the existence of a singular solution concerning the causal coefficients. This solution should be expressible in relation to either the covariance matrix or the probability distribution encompassing the observed variables. To attain the crucial aspect of identification – that is, the process of estimating structural coefficients through observed data points – it becomes imperative to introduce the concepts of partial covariance and partial regression coefficients.

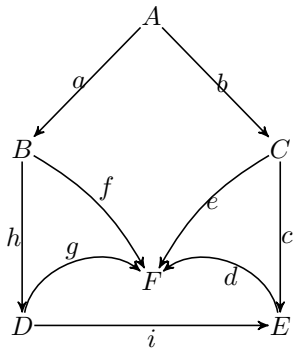
Definition 7. ***Wright's rule***

Let $\Pi = \{\pi_1, \pi_2, ..., \pi_k\}$ *denote the paths between* X *and* Y *that do not trace a collider. Recall that a collider is a node that has incoming arrows from two or more other nodes. Let* p_i *be the product of structural coefficients along path* π_i*. Then the covariance between variables* X *and* Y*, denoted as* σ_{YX}*, is equal to* $\sum_i p_i$*.*

Consider the causal graph given in Figure 16. We can compute the covariance between A and S as follows: First, we note that there are two paths between A and S and neither trace a collider, $\pi_1 = A \to C \to S$ and $\pi_2 = C \leftarrow A \to C \to I \to S$. The product of the coefficients along these paths are $p_1 = b \cdot c$ and $p_2 = b \cdot d \cdot f$. Summing these products together we obtain the covariance between A and S, $\sigma_{AS} = b \cdot (df + c)$.

Consider now another causal graph given by figure 19.

Figure 19: Causal graph illustrating Wright's rule



The paths between A and E that do not trace a collider are : $A \to C \to E$ and $A \to B \to D \to E$, since F is a collider. Summing the products of coefficients along these paths gives $\sigma_{AE} = a \cdot h \cdot i + b \cdot c$.

Now that we have explored how to calculate the covariance between two variables within a causal graph using Wright's rule, let's proceed to define three key concepts: partial covariance, partial correlation, and partial regression. These definitions are crucial as they will contribute to our understanding of the identification criteria.
Let X and Y be two random variables in a causal graph, Z a set of random variables. We denote σ_{XY} as the covariance between X and Y, ρ_{XY} as the correlation between X and Y and β_{YX} as the regression coefficient of Y on X. To express the partial covariance, $\sigma_{YX|Z}$, partial correlation, $\rho_{YX|Z}$, or regression coefficient, $\beta_{YX|Z}$, of Y on X given Z in terms of structural coefficients, we can first apply the following reductions before using Wright's rule. When Z is a singleton, these reductions are:

$$\rho_{YX|Z} = \frac{\rho_{YX} - \rho_{YZ}\rho_{XZ}}{\sqrt{(1-\rho_{YZ}^2)(1-\rho_{XZ}^2)}}$$

$$\sigma_{YX|Z} = \sigma_{YX} - \frac{\sigma_{YZ}\sigma_{ZX}}{\sigma_Z^2}$$

$$\beta_{YX|Z} = \frac{\sigma_Y}{\sigma_X}\frac{\rho_{YX} - \rho_{YZ}\rho_{ZX}}{1-\rho_{XZ}^2}$$

When Z is a singleton and S a set, we can reduce $\rho_{YX|ZS}$, $\sigma_{YX|ZS}$, or $\beta_{YX|ZS}$ as follows:

$$\rho_{YX|ZS} = \frac{\rho_{YX|S} - \rho_{YZ|S}\rho_{XZ|S}}{\sqrt{(1-\rho^2_{YZ|S})(1-\rho^2_{XZ|S})}}$$

$$\sigma_{YX|ZS} = \sigma_{YX|S} - \frac{\sigma_{YZ|S}\sigma_{ZX|S}}{\sigma^2_{Z|S}}$$

$$\beta_{YX|ZS} = \frac{\sigma_{Y|S}}{\sigma_{X|S}} \frac{\rho_{YX|S} - \rho_{YZ|S}\rho_{ZX|S}}{1-\rho^2_{XZ|S}}$$

Let's consider the causal graph given in Figure 16. Then , we can compute the regression coefficient of S on C given A as follows :

$$\begin{aligned}
\beta_{SC|A} &= \frac{\sigma_S}{\sigma_C} \frac{\rho_{SC} - \rho_{SA}\rho_{CA}}{1-\rho^2_{CA}} \\
&= \frac{1}{1} \frac{(c + d \cdot f) - b^2 \cdot (df + c)}{1-b^2} \\
&= \frac{\left(1-b^2\right)(c+df)}{1-b^2} \\
&= c + df \\
&= \sigma_{CS} \\
&= \text{ACE}\,(C \to S)
\end{aligned}$$

Now, let's consider the causal graph given in Figure 19 and let's compute the regression coefficient of E on C given D :

$$\begin{aligned}
\beta_{EC|D} &= \frac{\sigma_E}{\sigma_C} \frac{\rho_{EC} - \rho_{ED}\rho_{CD}}{1-\rho^2_{CD}} \\
&= \frac{1}{1} \frac{c + bahi - hab \cdot (i + cbah)}{1-(hab)^2} \\
&= \frac{c \cdot \left[1-(hab)^2\right]}{1-(hab)^2} \\
&= c \\
&= \text{DCE}\,(C \to E)
\end{aligned}$$

We note that if X and Y are d-separated given a set Z, then $\sigma_{XY|Z} = \rho_{XY|Z} = \beta_{XY|Z} = \beta_{YX|Z} = 0$.

3.5.2 Identification with admissible sets : Backdoor and Single-door criteria

Identification refers to the capacity of a model to uniquely estimate the causal effects or parameters of interest from observed data. A model parameter is considered identified if it can be uniquely determined based on the probability distribution of the variables within the model. When a parameter is identified, it can be estimated accurately from data. On the other hand, if a parameter is not identified, there are multiple potential values for that parameter that could match the given dataset, making it impossible to estimate it reliably. For example, consider the model given in Figure 16. We showed, using Wright's rule, that

the causal effect of C on S, ACE $(C \to S)$, is identified and equal to $\beta_{SC|A} = \frac{\sigma_S}{\sigma_C} \frac{\rho_{SC} - \rho_{SA}\rho_{CA}}{1-\rho_{CA}^2}$. Consider now the model given by Figure 13. The parameter ρ is not identified since, using Wright's rule, we have $\sigma_{SG} = \rho \cdot C_{SG}$, which will provide infinite solutions for ρ. When every parameter in a model can be clearly determined, the whole model is considered identified. But if just one parameter can't be clearly figured out, then the whole model is unidentified.

Their are several algorithms to check if structural models are identified. These algorithms try to find the best values for the parameters based on the data by minimizing a cost function. However, if a model is unidentified, the program can't give good estimates and warns about it. Relying only on those algorithms has some problems. For example, if the initial values for parameters are not good, the program might wrongly say the model is unidentified. Also, it doesn't tell us exactly which parameters are causing the issue. Instead of relying solely on software, we can provide criterion to see if parameters are identified. This method will help us express identified parameters in terms of partial regression coefficients and therefore in terms of the population covariance matrix. As a result, we will be able to estimate their values from the sample covariance matrix with a small set of data, and these estimates stay consistent as long as the model reflects the data generation process. This approach helps avoid problems related to bad initial parameter values, lets us identify parameters even if the whole model is not identified, and helps us determine parameter identifiability before collecting data.

Proposition 2. *The coefficients of a structural equation, $Y = \alpha_1 X_1 + \alpha_2 X_2 + \ldots + \alpha_k X_k + U_Y$, are identified and can be estimated using regression if the error term, U_Y, is independent of $X = \{X_1, X_2, \ldots, X_k\}$.*

One needs to distinguish between structural equations, in which the parameters $\alpha_1, \ldots, \alpha_k$ represent causal effects, and regression equations, in which the coefficients $\beta_1, \ldots, \beta_k$ represent regression slopes. The equation $Y = \beta_1 X_1 + \beta_2 X_2 + \epsilon_Y$ is a regression equation, where $\beta_1 = \frac{\partial}{\partial X_1}\mathbb{E}[Y \mid X_1, X_2]$, $\beta_2 = \frac{\partial}{\partial X_2}\mathbb{E}[Y \mid X_1, X_2]$, and $\epsilon_Y = Y - \beta_1 X_1 - \beta_2 X_2$ is the residual term. The equation is not necessarily structural since β_2 is not necessarily equal to the direct effect of X_2 on Y, $\frac{\partial}{\partial X_2}\mathbb{E}[Y \mid \text{do}(X_1, X_2)]$. Recall that the total effect of X_2 on Y is $\frac{\partial}{\partial X_2}\mathbb{E}[Y \mid \text{do}(X_2)]$. In Figure 16, we can have $S = \beta_1 C + \beta_2 A$ with $\beta_2 \neq 0$ but the direct causal effect of A on S is equal to 0.

In the context of regression analysis, the incorporation of a set of variables, denoted as Z, is commonly referred to as adjustment for Z. This practice raises a fundamental point: we can systematically determine whether a specific set of variables is suitable for this adjustment. This consideration becomes particularly relevant when our aim is to identify a structural coefficient (or direct causal effect). In other words, it is crucial to establish whether introducing a variable set Z would lead to the regression coefficient of Y on X being identical to the desired structural coefficient DE $(X \to Y)$. This matter can be addressed using a specific criterion outlined below, which facilitates a visual evaluation through causal graphs.

Theorem 1. ***Single-door criterion : Admissible set for direct causal effect***
Let G be DAG in which α is the structural coefficient associated with arrow $X \to Y$, and let G_α denote the graph that results when $X \to Y$ is deleted from G. The coefficient α is identifiable if there exists a set of variables Z such that :

- *Z contains no descendant of Y, and*
- *Z d-separates X from Y in G_α.*

If Z satisfies these two conditions, then α is equal to the regression coefficient $\beta_{YX|Z}$. Conversely, if Z does not satisfy these conditions, then $\beta_{YX|Z}$ is not a consistent estimator of α.

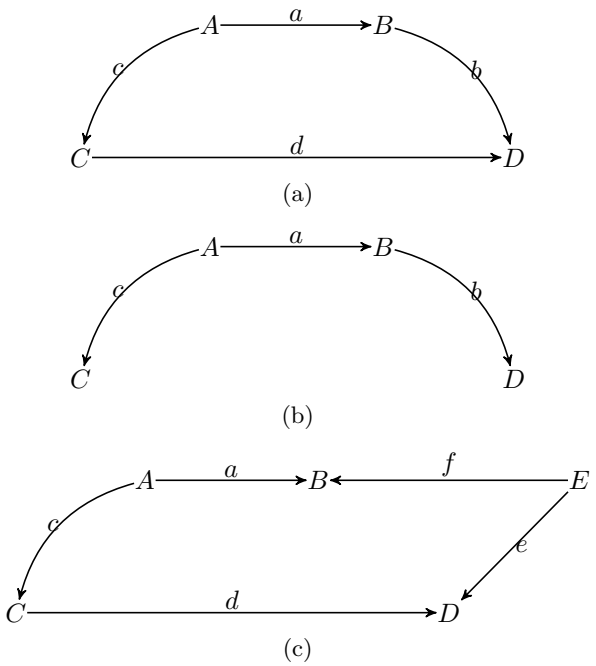
Moving forward, we shall introduce a similar criterion for identifying the total causal effect $\text{ACE}(X \to Y)$:

Theorem 2. ***Back-door criterion : Admissible set for total causal effect***
For any two variables X and Y in a DAG G, the total effect of X on Y is identifiable if there exists a set of variables Z such that :

- *No member of Z is a descendant of X; and*
- *Z d-separates X from Y in the subgraph G_X formed by deleting from G all arrows emanating from X.*

Moreover, if these two conditions are satisfied, then the total effect of X on Y is given by $\beta_{YX|Z}$.

Figure 20: Causal graphs illustrating the identifiability criteria



Let's consider the causal graphs given by Figure 20. We want to check if the causal effect of C on D, that is d is identifiable. In Figure 20a, we see that B blocks the path $C \leftarrow A \to B \to D$ and C is d-separated from D by B in Figure 20b. Therefore, d is identified and equal to $\beta_{DC|B}$. This is to be expected since C is independent of ϵ_D in the structural equation, $D = d \cdot C + b \cdot B + \epsilon_D$. The theorem above tells us, however, that A can also be used for adjustment since A also d-separates C from D in Figure 20b, and we obtain $d = \beta_{DC|A}$. We will see in a subsequent section, however, that the choice of B is superior to that of A in terms of estimation power. Consider, however, Figure 20c. A satisfies the

criterion but B does not. Being a collider, B unblocks the path, $C \leftarrow A \rightarrow B \leftarrow E \rightarrow D$, in violation of the theorem, leading to bias if adjusted for. In conclusion, d is equal to $\beta_{DC|A}$ in Figures 20a and 20c. However, d is equal to $\beta_{DC|B}$ in Figure 20a only. Returning to Figure 19, let's verify whether the direct causal effect $\text{ACE}(C \rightarrow E)$ is identifiable. We observe that E is d-separated from C given D when we eliminate the edge from C to E. In fact, all the paths between C and E are blocked, either because they pass through the collider F or they pass through D, which is included in the conditioning set. This confirms that $\beta_{EC|D} = c = \text{DCE}(C \rightarrow E)$.

It is not uncommon to encounter causal models where there are no admissible sets satisfying the single-door criterion or back-door criterion. In such cases, it becomes challenging to identify direct and total causal effects. However, instrumental variables, a specific type of variable, can be employed to help us identify the direct causal effects, as we will explore in the following subsection.

3.5.3 Identification with instrumental variables

Instrumental variables (IV) play a crucial role in identifying causal effects when dealing with endogeneity and omitted variable bias. An instrumental variable is a variable that is correlated with the treatment variable of interest, but it is not directly related to the outcome variable except through its influence on the treatment. IVs are particularly useful in situations where randomization is not feasible, and the presence of unobserved confounders makes causal inference challenging. It addresses potential issues of endogeneity and confounding in observational studies when estimating causal relationships between variables. Endogeneity refers to situations where the relationship between a treatment variable and an outcome variable is confounded by unobserved factors or reverse causation. A treatment variable represents a variable that is manipulated or controlled in an experiment or study. It is the independent variable that researchers or analysts are interested in studying to understand its effect on other variables. In a causal graph, a directed arrow points from the treatment variable to the outcome variable to indicate the causal relationship being investigated. The treatment variable is the cause or the input that influences the outcome variable. An outcome variable represents a variable that is affected by the treatment variable. It is the dependent variable whose changes are being observed or measured based on variations in the treatment variable. In a causal graph, the outcome variable is depicted with a directed arrow pointing towards it, indicating that it is influenced by the treatment variable.

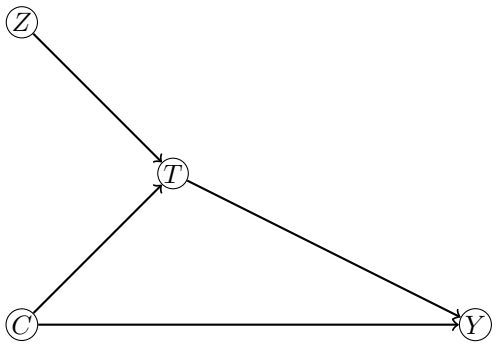
Figure 21 illustrates the various types of variables used when modeling a situation. This visual representation provides a clear overview of the different types of variables involved in the model.
The instrumental variable is used to indirectly estimate the causal effect of a treatment variable on an outcome variable when a direct causal relationship cannot be easily established due to confounding.

Example 2. ***Scenario: Education and Income***
Treatment Variable (Endogenous Variable): *Education level (years of schooling completed by an individual) is the treatment variable in this scenario. We want to understand how changes in education level impact an individual's income.*

Outcome Variable (Endogenous Variable): *Income is the outcome variable. We want to determine how education influences an individual's income.*

Instrumental Variable (Exogenous Variable): *Let's say we use "Proximity to a College" as the instrumental variable. The idea is that proximity to a college affects education (people living closer to a college are more likely to attend), but it does not have a direct effect on income except through its influence on education.*

Figure 21: Z is an instrumental variable, C is a confounder, T is the treatment and Y is the outcome variable



1. ***Causal path (direct effect):***

   $$Education \rightarrow Income$$

   A higher level of education is expected to lead to higher income due to increased skills and qualifications.

2. ***Confounding path (indirect effect):***

   $$Proximity\ to\ college \rightarrow Education \rightarrow Income$$

   Proximity to a college might influence education levels, but it might also indirectly influence income through factors like access to resources, job opportunities, and other confounding variables.

The direct causal relationship between education and income can be confounded by factors like natural ability, family background, and individual motivation. These confounders can lead to biased estimates of the education-income relationship.

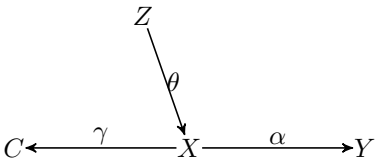
Using the intrumental variable:

1. *We use Proximity to college as the instrumental variable because it is correlated with education but is not directly related to income, except through its impact on education.*
2. *In the first stage of the analysis, we regress education on proximity to college. This estimates the effect of proximity to college on education.*
3. *In the second stage, we regress income on proximity to college. This gives us an estimate of the causal effect of education on income while controlling for the potential confounding effects.*

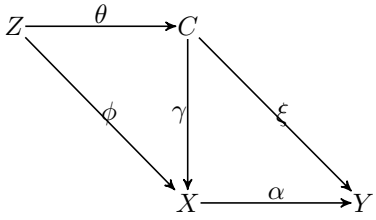
By using an instrumental variable like Proximity to College, we aim to mitigate the endogeneity and confounding issues that could affect the estimation of the causal relationship between education and income. This approach helps us separate the true causal effect of education on income from the potential biases introduced by unobserved confounders.

In Figure 22a , no admissible set exists for α and it cannot be estimated using regression. However, using Wright's equations we see that $\sigma_{YZ} = \theta\alpha$ and $\sigma_{XZ} = \theta$. As a result,

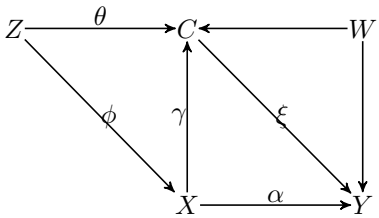
Figure 22: Instrumental variable examples



(a) Graph 1: Using IV to identify causal effect



(b) Graph 2: Conditional IV satisfying identifiability criterion



(c) Graph 3: Conditional IV not satisfying identifiability criterion

$\alpha = \frac{\sigma_{YZ}}{\sigma_{XZ}}$. In this case, we were able to identify α using the instrumental variable Z. We will provide a graphical method that allows us to quickly determine whether a given variable is an instrumental variable by inspecting the path diagram. Additionally, we will introduce conditional instrumental variables and instrumental sets, which will significantly increase the identification power of the instrumental variable method. The following is a formal definition of an instrumental variable :

Definition 8. ***Instrumental variable*** *For a structural equation, $Y = \alpha_1 X_1 + \ldots + \alpha_k X_k + U_Y$, Z is an instrumental variable if :*

- ***Relevance :*** *Z is correlated with $X = \{X_1, \ldots, X_k\}$*
- ***Exogeneity:*** *Z is uncorrelated with U_Y.*

The following graphical characterization rectifies such ambiguities and allows us to determine through quick inspection of the path diagram whether a given variable is an instrument for a given parameter. Moreover, it provides a necessary and sufficient condition for when α_i in the equation $Y = \alpha_1 X_1 + \ldots + \alpha_k X_k + UY$ is identified by $\frac{\beta_{YZ}}{\beta_{X_i Z}}$.

Proposition 3. *A variable Z qualifies as an instrumental variable for coefficient α from X to Y if:*

1. *Z is d-separated from Y in the subgraph G_α obtained by removing edge $X \to Y$ from G, and*
2. *Z is not d-separated from X in G_α.*

Moreover, if these two conditions are satisfied, then $\alpha = \frac{\beta_{YZ}}{\beta_{XZ}}$.

In Figure 22a, Z is d-separated from Y when we remove the edge associated with α, but it is d-connected with X. As a result, Z is an instrumental variable for α, and we have $\alpha = \frac{\beta_{YZ}}{\beta_{XZ}}$.

Consider Figure 22b. In this graph, Z is not an instrument for α because it is d-connected to Y through the path $Z \to C \to Y$, even when we remove the edge associated with α. However, if we condition on C, this path is blocked, i.e. C d-separates Z from Y but it does not d-separate Z from X. Thus, we see that some variables may become instruments by conditioning on covariates.

Definition 9. *A variable Z is a conditional instrumental variable given a set W for coefficient α (from $X \to Y$) if:*

1. *W contains only non-descendants of Y,*
2. *W d-separates Z from Y in the subgraph G_α obtained by removing edge $X \to Y$ from G, and*
3. *W does not d-separate Z from X in G_α.*

Moreover, if these conditions are satisfied, then $\alpha = \frac{\beta_{YZ|W}}{\beta_{XZ|W}}$.

In Figure 22b, we saw that Z is a conditional instrument for α given C. This means that $\alpha = \frac{\beta_{YZ|C}}{\beta_{XZ|C}}$. However, in Figure 22c, Z is not an instrument given C because conditioning on C opens the path $Z \to X \to C \leftarrow W \to Y$ (since C is a collider for this path).

Finally, it may be possible to use several variables in order to identify a set of parameters when, individually, none of the variables qualifies as an instrument. In Figure 23, there are no admissible sets in order to identify α_1 and α_2. In fact, C_1 and C_2 are descendants of Y. X_1 is not d-separated from Y in the G_{α_1} (graph with arrow $X_1 \to Y$ removed) given neither Z_1 nor Z_2, since the path $X_1 \leftarrow C_1 \leftarrow Y$ is active. Similarly, X_2 is not d-separated from Y in G_{α_2} (graph with arrow $X_2 \to Y$ removed), given neither Z_1 nor Z_2, since the path $X_2 \leftrightarrow X_1 \to Y$ is active. Thus, there exist no admissible sets for identification due to a violation of Theorem 1. Moreover, neither Z_1 nor Z_2 are instruments since they are not d-separated from Y in G_{α_1} and G_{α_2} respectively (consider the paths $Z_1 \to Z_2 \to X_2 \to Y$ and $Z_2 \leftarrow Z_1 \to X_1 \to Y$, which are both active). Furthermore, Z_1 and Z_2 are not conditional instrumental variables given C_1 and C_2 respectively, since C_1 and C_2 are descendants of Y. However, if we use Wright's equations, we have:

$$\sigma_{Z_1Y} = \sigma_{Z_1Z_2} \cdot \sigma_{Z_2X_2} \cdot \alpha_2 + \sigma_{Z_1X_1} \cdot \alpha_1$$
$$\sigma_{Z_2Y} = \sigma_{Z_2Z_1} \cdot \sigma_{Z_1X_1} \cdot \alpha_1 + \sigma_{Z_2X_2} \cdot \alpha_2$$

which is equivalent to:

$$\sigma_{Z_1Y} = \sigma_{Z_1X_1} \cdot \alpha_1 + \sigma_{Z_1X_2} \cdot \alpha_2$$
$$\sigma_{Z_2Y} = \sigma_{Z_2X_1} \cdot \alpha_1 + \sigma_{Z_2X_2} \cdot \alpha_2$$

Solving these two linearly independent equations for α_1 and α_2 identifies the two parameters. We call a set of variables that enables a solution in this manner an instrumental set.

Definition 10. ***Instrumental set*** *For a path π_h that passes through nodes V_i and V_j, let $\pi_h[V_i \dots V_j]$ denote the sub-path that begins with V_i, ends with V_j, and follows the same sequence of edges and nodes as π_h does from V_i to V_j. Then $\{Z_1, Z_2, \dots, Z_k\}$ is an instrumental set for the coefficients $\alpha_1, \dots, \alpha_k$ associated with edges $X_1 \to Y, \dots, X_k \to Y$ if the following conditions are satisfied:*

1. *Let G be the graph obtained from G by deleting edges $X_1 \to Y, \dots, X_k \to Y$. Then, Z_i is d-separated from Y in G for all $i \notin \{1, 2, \dots, k\}$.*

2. *There exist paths* $\pi_1, \pi_2, \ldots, \pi_k$ *such that* π_i *is a path from* Z_i *to* Y *that includes edge* $X_i \to Y$*, and if paths* π_i *and* π_j *have a common variable* V*, then either:*

   (a) both $\pi_i[Z_i \ldots V]$ *and* $\pi_j[V \ldots Y]$ *point to* V*, or*

   (b) both $\pi_j[Z_j \ldots V]$ *and* $\pi_i[V...Y]$ *point to* V*,*

   for all $i, j \in \{1, 2, \ldots, k\}$ *and* $i \neq j$*.*

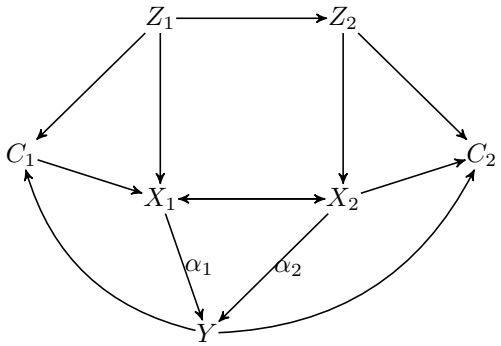
The following theorem explains how instrumental sets can be used to obtain closed form solutions for the relevant coefficients.

Theorem 3. *Let* $\{Z_1, Z_2, \ldots, Z_n\}$ *be an instrumental set for the coefficients* $\alpha_1, \ldots, \alpha_n$ *associated with edges* $X_1 \to Y, \ldots, X_n \to Y$*. Then the linear equations,*

$$\begin{aligned}
\sigma_{Z_1 Y} &= \sigma_{Z_1 X_1}\alpha_1 + \sigma_{Z_1 X_2}\alpha_2 + \ldots + \sigma_{Z_1 X_n}\alpha_n \\
\sigma_{Z_2 Y} &= \sigma_{Z_2 X_1}\alpha_1 + \sigma_{Z_2 X_2}\alpha_2 + \ldots + \sigma_{Z_2 X_n}\alpha_n \\
&\vdots \\
\sigma_{Z_n Y} &= \sigma_{Z_n X_1}\alpha_1 + \sigma_{Z_n X_2}\alpha_2 + \ldots + \sigma_{Z_n X_n}\alpha_n,
\end{aligned}$$

are linearly independent for almost all parameterizations of the model and can be solved to obtain expressions for $\alpha_1, \ldots, \alpha_n$ *in terms of the covariance matrix.*

Figure 23: Using instrumental set to identify causal effect



Returning to Figure 23, we can see graphically that the set $\{Z_1, Z_2\}$ is an instrumental set for the coefficients α_1 and α_2. In fact, both conditions are satisfied:

- Both Z_1 and Z_2 are d-separated from Y when we remove the paths $X_1 \to Y$ and $X_2 \to Y$.
- If we consider the paths $\pi_1 : Z_1 \to Z_2 \to X_2 \leftrightarrow X_1 \to Y$ and $\pi_2 : Z_2 \leftarrow Z_1 \to X_1 \leftrightarrow X_2 \to Y$, one common variable between these paths is X_1, and we have that $\pi_1[Z_1, \ldots, X_1]$ and $\pi_2[X_1, \ldots, Y]$ both point to X_1.

Furthermore, due to Theorem 3, we obtain the following linear independent equations :

$$\begin{aligned}
\sigma_{Z_1 Y} &= \sigma_{Z_1 X_1} \cdot \alpha_1 + \sigma_{Z_1 X_2} \cdot \alpha_2 \\
\sigma_{Z_2 Y} &= \sigma_{Z_2 X_1} \cdot \alpha_1 + \sigma_{Z_2 X_2} \cdot \alpha_2
\end{aligned}$$

Solving the equations identifies α_1 and α_2 giving :

$$\alpha_1 = \frac{\sigma_{Z_2X_2} \cdot \sigma_{Z_1Y} - \sigma_{Z_1X_2} \cdot \sigma_{Z_2Y}}{\sigma_{Z_2X_2} \cdot \sigma_{Z_1X_1} - \sigma_{Z_1X_2} \cdot \sigma_{Z_2X_1}}$$
$$\alpha_2 = \frac{\sigma_{Z_1Y} - \sigma_{Z_1X_1} \cdot \alpha_1}{\sigma_{Z_1X_2}}$$

Till now, we have learned about the process of determining structural coefficients using admissible sets and instrumental variables. To establish the overall identifiability of the model, it becomes imperative to validate the identifiability of each individual causal effect. This task could become time-consuming and resource-intensive when dealing with intricate models. In this regard, we will introduce a criterion that offers a graphical and direct method for assessing whether an entire model is identifiable, thus bypassing the need to verify the identifiability of each parameter.

Theorem 4. ***Model identification***
If a causal model contains a set of variables and associated causal relationships that do not form a bow-arc, then the model is identifiable as a whole. In other words, the absence of bow-arcs within the model indicates its overall identifiability. A bow-arc is a specific configuration within a causal model where a variable, known as the tail, has multiple arrows, or causal pathways, pointing towards it. These arrows, or arcs, collectively resemble the shape of a bow. Such a configuration complicates the identifiability of the model, as the influence on the tail variable is not uniquely attributable to any single causal pathway.

Figure 24: Example of a bow-arc

3.6 Coherent Scenarios and treatment effects

In this subsection, we delve into various metrics that enable us to quantitatively evaluate the impact of interventions, allowing us to measure the causal strength of a variable T, which we refer to as the treatment variable on the target variable Y. To illustrate this concept, consider an example: suppose we aim to assess the effect of a medication T on a patient's heart rate Y, while keeping environmental conditions X constant (such as libido levels and lifestyle). In such scenarios, conducting an intervention is necessary. We begin by setting the value of T to t_0 while conditioning X on x, and we observe the value of Y, and then perturbing the variable T to t_1, we record the new value of Y. This process is repeated multiple times, and we calculate the average of the differences between post-intervention and pre-intervention values of Y. This calculation yields a metric that quantifies the causal effect of T on Y in the form of an expectation, known as the Heterogeneous Treatment Effect. We will provide explanations of several metrics that will assist us in creating coherent scenarios and addressing what-if questions. It is essential to note that when we discuss interventions, we are referring to the $do()$ operator introduced earlier. We follow [41] and [5].

3.6.1 Heterogeneous treatment effect (HTE)

The HTE [41] represents the variation in treatment effects across different levels of covariates. It measures how the treatment effect differs based on the values of a set of covariates X.

$$\text{HTE}(x) = \mathbb{E}\left[Y(t_1) - Y(t_0) \mid X = x\right] \tag{9}$$
$$= \mathbb{E}\left[Y \mid do(T = t_1), X = x\right] - \mathbb{E}\left[Y \mid do(T = t_0), X = x\right] \tag{10}$$

Where:

- $\text{HTE}(x)$ is the heterogeneous treatment effect for a specific value of covariates x.
- $\mathbb{E}$ denotes the expectation.
- $Y(t_1)$ represents the potential outcome when the treatment has the fixed value t_1.
- $Y(t_0)$ represents the potential outcome when the treatment has fixed value t_0.

3.6.2 Heterogeneous marginal effect

If treatments are continuous, then one might also be interested in a local effect around a treatment point. This means estimating a local gradient around a treatment vector conditional on observables :

$$\text{HME}(x) = \mathbb{E}\left[\nabla_t Y(t) \mid X = x\right] \tag{11}$$

3.6.3 Average treatment effect (ATE)

The ATE represents the average difference between potential outcomes for the entire population, regardless of covariate values. It quantifies the overall impact of the treatment.

$$\text{ATE} = \mathbb{E}\left[Y(t_1) - Y(t_0)\right] \tag{12}$$
$$\approx \frac{\partial}{\partial t}\mathbb{E}\left[Y \mid do(T = t)\right] \tag{13}$$

The ATE is an approximation (via Taylor) of the total causal effect. More details about the Average Treatment Effect (ATE) can be found in the following part of the paper.

3.6.4 Conditional average causal effect (CATE)

The CATE focuses on the average treatment effect for a specific value or set of values of the covariates X. It provides a more nuanced understanding of treatment effects.

$$\text{CATE} = \frac{\partial}{\partial t}\mathbb{E}\left[Y \mid do(T = t), X = x\right] \tag{14}$$

4 Case study and results

In this section, we investigate the practical application of our methodology to generate coherent scenarios and extract valuable insights. Our focus is on a case study involving seven different clean-tech indices: hydrogen-sector based leading clean-techs, chemicals, agribusiness, energy, gas, environment, and solar. We will examine their performance within the context of macroeconomic variables.

4.1 Methodology and approach

For the hydrogen-sector based leading clean-techs index, we have predominantly chosen american clean-tech companies with a purity rate exceeding 50%. These companies are known for their commitment to clean energy solutions within the hydrogen sector. Their operations include the development of cutting-edge technologies such as advanced electrolysis systems, hydrogen refueling infrastructure, and energy storage solutions. The selected companies boast a substantial market capitalization, ensuring that their historical valuation trends can be observed over an extended period. Importantly, these clean-techs align with the criteria outlined in the IRA, making them eligible for tax credits under the IRA project.

In the chemicals index, we have focused on american clean-tech companies with a purity rate exceeding 50%. These companies specialize in developing clean and sustainable solutions within the chemical industry. Their innovations include eco-friendly production methods, environmentally conscious chemicals, and reduced environmental impact throughout the entire chemical manufacturing process. The chosen clean-techs have a significant market capitalization, allowing for a comprehensive analysis of their historical valuation trends. Moreover, they meet the criteria of the IRA, making them potential beneficiaries of tax credits provided by the IRA project.

Within the agribusiness index, our selection comprises american clean-tech companies with a purity rate surpassing 50%. These companies contribute to sustainable agriculture and food production through advancements in precision farming, eco-friendly fertilizers, and technologies promoting environmental conservation in agriculture. The chosen clean-techs possess substantial market capitalization, enabling the observation of their historical valuation trends. Additionally, they meet the criteria of the IRA, making them eligible for tax credits within the framework of the IRA project.

The energy index includes american clean-tech companies with a purity rate exceeding 50%. These companies are dedicated to clean energy production and distribution, employing technologies such as renewable energy sources, energy storage solutions, and grid optimization. With a significant market capitalization, the selected clean-techs allow for the analysis of their historical valuation trends. Furthermore, they qualify for tax credits under the IRA, aligning with the criteria for potential benefits from the IRA project.

In the gas index, we have chosen american clean-tech companies with a purity rate above 50%. These companies specialize in clean gas production and utilization, with advancements in extraction, storage, and applications across various industries. The selected clean-techs exhibit a substantial market capitalization, facilitating the examination of their historical valuation trends. Additionally, they meet the criteria of the IRA, making them eligible for tax credits provided by the IRA project.

For the environment index, our selection features american clean-tech companies with a purity rate surpassing 50%. These companies focus on a broad spectrum of environmental clean-tech solutions, including waste management, pollution control, and initiatives for preserving and restoring natural ecosystems. The chosen clean-techs possess a significant market capitalization, enabling the analysis of their historical valuation trends. Moreover,

they meet the criteria of the IRA, making them eligible for tax credits within the framework of the IRA project.

The solar index comprises american clean-tech companies specializing in solar energy solutions, with a purity rate exceeding 50%. These companies are involved in the development of solar panels, solar energy storage systems, and technologies for harnessing solar power across various applications. With a substantial market capitalization, the selected clean-techs facilitate the examination of their historical valuation trends. Additionally, they qualify for tax credits under the IRA, aligning with the criteria for potential benefits from the IRA project.

Our methodology hinges on a systematic approach to understanding the impact of macroeconomic factors on these clean-tech indices. Here's a breakdown of our methodological framework:

1. **Data collection:** We begin by gathering data from various sources, compiling the macroeconomic factors susceptible to driving the clean-tech indices' valuation, and collecting, for each sector, various clean-tech stock performance data from 2007 to 2023. This data is recorded on a weekly basis. The macroeconomic factors include: oil price, US interest rate, EU interest rate, nickel price, carbon price, index of technology company stocks, index of semiconductor and electronic company stocks, inflation, and gas price.

2. **Data discretization:** After meticulously preparing and standardizing the data, the next crucial step involves discretizing values into three distinct levels. Level 0 corresponds to the first quartile, level 1 to the second quartile, and level 2 to the third quartile. This discretization aligns with our approach of working with discrete-time Bayesian networks, where values are confined to a discrete space. The significance of this process lies in its role in facilitating the visualization of intervention effects. For example, if the price of oil shifts from level 1 to level 2, we aim to understand the ensuing repercussions on the value of the Hydrogen Equity index. While continuous-time Bayesian networks may seem ideal, they inherently introduce significant practical complexity. For those interested in a comprehensive introduction to continuous-time Bayesian networks, additional details are available in the annex. However, delving further into continuous-time networks within a continuous state space would deviate substantially from our primary focus and potentially introduce unwarranted intricacy. It's essential to emphasize that our primary focus centers around static scenarios. We don't engage in predicting future outcomes at specific time horizons. Instead, we specialize in generating scenarios tailored to address causal queries, i.e., "what-if" questions. By revisiting historical data, we assess the impact of perturbations on each equity sector-based index's present-day value. This approach holds particular relevance in clustering, where we aim to identify the three (or two) factors with the most substantial influence on our variable of interest. Subsequently, we quantify the sensitivity of our variable to these factors, ultimately enabling the categorization of American clean-tech companies into clusters based on their responsiveness to specific factors.

3. **Causal graph construction:** Following careful data preparation and visualization, we progress to constructing causal graphs using various algorithms such as Hill-climbing, Chow-Liu, Regression-based, NTE, PCMCI, Granger Causality, DirectLinGAM, GES, and an ensemble model combining PCMCI+, Granger, and NTE simultaneously. Each algorithm reveals a unique dependency structure among macroeconomic factors, resulting in the creation of nine distinct causal graphs. Our focus will

be on three types of results from three algorithms: ensemble, regression-based, and DirectLinGAM. Additionally, we'll provide a concise overview of a dynamic causal structure obtained through hill-climbing. Here, "dynamic" refers to a series of causal graphs, each corresponding to a specific time period. For instance, when analyzing the hydrogen equity index, we observe distinct causal relationships in 2020-2021 compared to 2008-2009. This reflects the evolving interdependencies during significant periods such as the financial crisis of 2008 and the COVID-19 pandemic in 2020. Causality algorithms may inherently produce causal relationships that initially appear counterintuitive or illogical due to their reliance on data, which may exhibit suboptimal quality. Consequently, expert oversight is necessary to ensure that the resulting causal models align with domain knowledge. In our specific case, we enforce a critical constraint: each sector-based clean-tech equity index, as a target outcome variable, exclusively serves as a child node within these causal structures, indicating that it has no offspring. Furthermore, specific constraints are imposed to reinforce the integrity of the causal relationships:

- *US_rate* and *EU_rate*, representative of american and european interest rates, are explicitly prevented from being child nodes of *Techno_stocks* and *Semi_conduct*, corresponding to technological and semiconductor stock indices, respectively.
- The variable *Inflation* is only permitted as a child node if its sole parent is *Oil_price*.

4. **Structural equation modeling (SEM):** After establishing the causal graphs, the subsequent critical phase involves estimating direct causal effects through SEM. This pivotal step is executed using the Semopy library, enabling the construction of structural equations following the topological order of the variables and subsequent estimation of the structural coefficients. These estimations are conducted for each distinct causal graph. As we examine the graphs, we identify variables serving as confounders and instrumental variables. These variable types, as discussed in the theoretical section of this paper, play a crucial role in the identifiability criteria. We recall that identification refers to the ability to estimate causal effects (structural coefficients, given that we are working with linear functions) using data.

5. **Scenario generation and analysis:** Finally, we can embark on the critical phase of generating coherent scenarios. These scenarios play a pivotal role in addressing causal queries, shedding light on questions such as : how would the taste of my cake have been affected if I had increased the flour rate by 100g? In a causal model, these scenarios arise from interventions on variables, enabling us to observe the ripple effects of perturbations on the other variables throughout the graph. We recall that an intervention involves making a random variable deterministic by assigning it a specified value. Subsequently, the causal model itself undergoes modification (remember that intervening on a variable in a graph leads to cutting all the arrows pointing towards it), and as each causal model corresponds to a joint probability distribution, we can derive a post-intervention distribution known as the counterfactual (hypothetical) distribution [20], [47]. This distribution, distinct from the pre-intervention distribution, is obtained using the three do-calculus rules introduced in the theoretical section of our paper. We have seen so far that the do() operator serves as the central tool of the do-calculus, akin to a new calculus algebra on the space of random variables, facilitating interventions. Once the post-intervention distribution is obtained, we are able to compute order moments such as post-intervention expectations and standard

deviations. Furthermore, we can perform various operations typical of probability distributions, such as simulation and estimation. As a result, metrics based on these post-intervention order moments can be derived, aiding in the assessment of counterfactual questions. Similar to how expectation provides insight into the average of a random variable, post-intervention expectations offer an interpretation of the average of a target variable after we have perturbed the graph. Our focus lies on three key metrics for answering causal queries: Average treatment effect, Greatest causal influence (GCI), and average causal effect. While our framework encompasses several other metrics, these three are pivotal for our study. Here's a brief overview of each:

1. **Average treatment effect:** ATE, introduced in the previous section, quantifies the average difference in the outcome variable between the post-intervention and pre-intervention states, providing a general measure of the intervention's impact. For example, let's consider the impact on our hydrogen equity index when we modify the price of oil in the causal model. If we intervene by increasing the price of oil from level 1 to level 2, we generate a new post-intervention distribution. From this distribution, we can simulate the counterfactual variable "hydrogen equity index" multiple times and calculate a post-interventional expectation. The ATE is then determined by the difference between this post-intervention expectation and the pre-intervention expectation. To illustrate, if the ATE is -0.26, it implies that, on average, the hydrogen equity index would decrease by 26% after we have perturbed the model by intervening on oil price.

2. **Greatest causal influence:** GCI identifies the variable with the most significant influence on the outcome variable following an intervention, helping pinpoint the primary driver of observed changes. The GCI is not necessarily the variable that has the strongest causal effect on our target since the GCI is obtained by multiple interventions and therefore acts as a "local" influence. On the opposite, the causal effect, as we have seen it in the theoretical section of our paper, is identified with partial regression coefficients which provide an "overall" causal effect over the entire data.

3. **Average causal effect :** ACE represents the average change in the outcome variable resulting from the intervention, offering a comprehensive understanding of the overall causal impact.

While our framework incorporates numerous other metrics, these three metrics form the core of our analytical approach, allowing us to gain meaningful insights into the causal relationships within the clean-tech sector-based equity indices.

Our algorithm, coded in Python, is designed to automate this comprehensive process systematically. The results of our analysis will be presented in the context of clean-tech sector-based indexes. This versatile algorithm is not limited to clean-tech alone and can be applied to a wide range of causal analyses across various domains.

4.2 Causal discovery and structural coefficients

In the following sections, we will present and analyze the various causal graphs obtained for the hydrogen-based clean-techs equity index. Of course, similar causal graphs can be obtained for the other sector-based clean-techs equity indices. Given the comprehensive nature of the algorithms employed, we will not delve into every algorithm used, as such an exposition would be exhaustive and potentially unengaging.

We will focus on the causal graphs derived from the Regression-based approach, the ensemble model integrating NTE, Granger, and PCMCI+, as well as DirectLiNGAM. Moreover, we have devised a method for constructing a dynamic Bayesian network, effectively a

causal graph that evolves discretely over time (with one-year intervals). Consequently, we will unveil two causal graphs, each established during a distinct period. For example, we will explore the graph covering the year 2008-2009 and the graph covering the year 2019-2020. This approach allows us to decipher the intricate interdependencies between factors, taking into account the unique characteristics of each period, such as the global financial crisis of 2008 and the COVID-19 pandemic crisis in 2020. Each edge within the causal graph is associated with a weight, representing structural coefficients or direct causal effects. These coefficients are estimated by the Semopy library using the single-door criterion, defined in Theorem 1. We recall that the direct causal effect of a variable X on a variable Y is given, using the do() operator, by:

$$DCE = \frac{\partial}{\partial x}\mathbb{E}[Y \mid \text{do}(X = x, Z = z)],$$

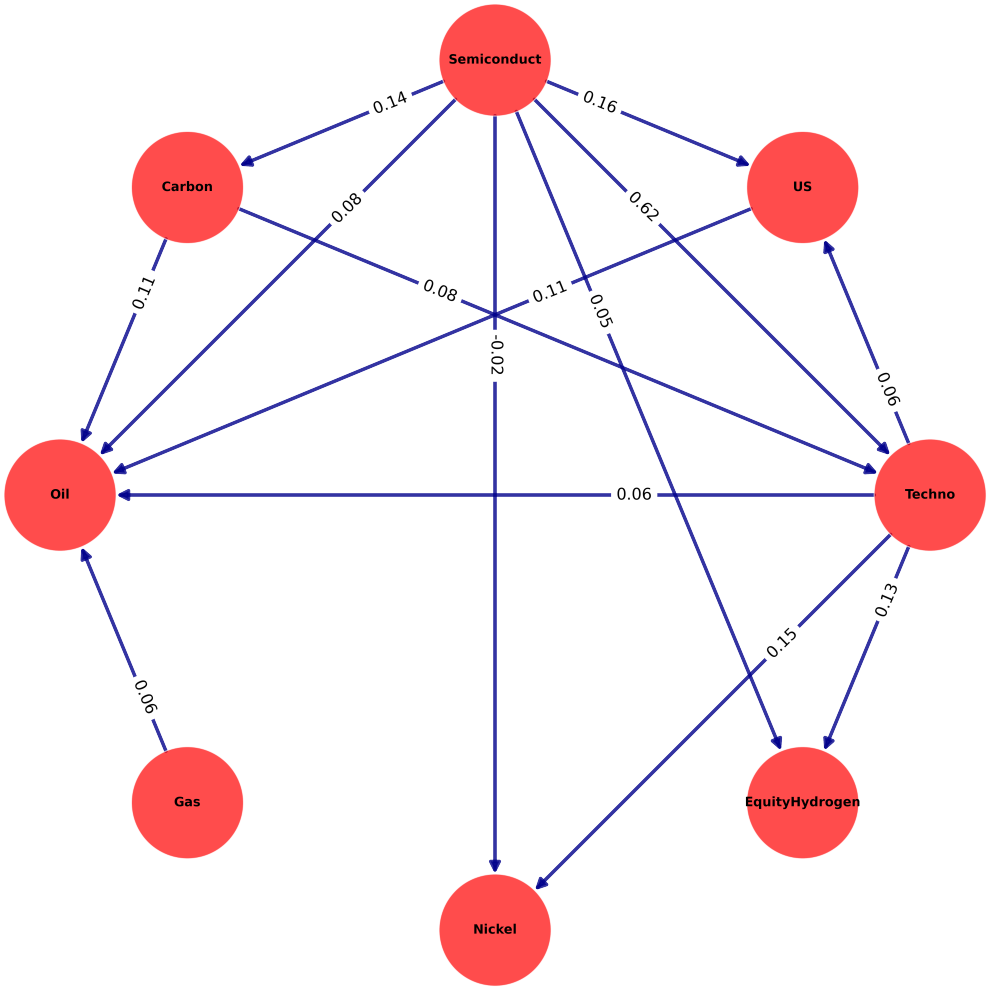
where Z is the set of variables in the causal model containing all the variables other than X and Y. Intuitively, we can justify this relation with a simple logic: For instance, if we want to know how the oil price affects the hydrogen equity index, we fix a value x for the oil price while keeping all other variables unchanged (i.e., we intervene on the oil price using do()). This results in a counterfactual distribution (post-intervention distribution) where we can compute the value of our target variable, the hydrogen equity index. Then, we fix another value x' for the oil price (i.e., we intervene a second time) and compute the new value of our target variable. This process is repeated many times, and the average of the differences between the new value of the hydrogen equity and its initial value quantifies the impact (or causal effect) of the oil price on the hydrogen equity. It is called *direct* since we only intervene on the treatment variable, which is the oil price. Given the causal effects, we can derive the structural equations built with respect to the topological ordering of the variables in the graph.

Figure 25 illustrates the causal graph generated using a regression-based algorithm. This approach involves conducting a regression analysis of each variable against all the others. A causal link is established between two variables if the regression coefficient exceeds a predefined threshold. Notably, we observe that *EquityHydrogen* serves as the target variable, while the key treatment variable in this model is associated with the technology companies index *Techno*. The direct causal effect of this variable on the target variable is quantified at 0.13. In practical terms, this implies that a 1% perturbation in the technology stocks index is expected, on average, to result in a 13% increase in the hydrogen equity index. The variable *Semiconduct*, corresponding to the semiconductor companies index, acts as a confounder: it simultaneously influences both the target variable and the treatment variable. On the other hand, the variable *Carbon*, representing the carbon price, serves as an instrumental variable for the causal effect Techno $\rightarrow$ EquityHydrogen, as per the definition in 8.

The structural equations derived from this causal graph are outlined below:

$$\begin{aligned}
Carbon &\sim Semiconduct \\
EquityHydrogen &\sim Techno + Semiconduct \\
Nickel &\sim Techno + Semiconduct \\
Techno &\sim Carbon + Semiconduct \\
US &\sim Techno + Semiconduct \\
Oil &\sim Gas + Techno + Carbon + Semiconduct + US
\end{aligned}$$

Figure 25: Causal graph obtained by regression



From this causal graph, we can derive Conditional probability tables (CPTs), which are tables providing estimates of the probability of a child variable being in one of three states (Level 0, Level 1, or Level 2) given the states of its parents. For example, Table 1 illustrates the CPT for the hydrogen-based clean-tech equity index. We observe that the probability of the hydrogen equity index being at Level 0, given that the semiconductor stocks index is at Level 2 and the technology stocks index is at Level 1, is 0.27. This table also highlights that probabilities are higher when the hydrogen index is at the same level as its parents, indicating a positive causal influence of the semiconductor index and technology index on the hydrogen index.

Figure 26 showcases the causal graph obtained through the DirectLiNGAM algorithm, representing a linear structural causal model with non-Gaussian additive noise. Once again, the technology stock index is designated as the treatment variable. Remarkably, there are

Table 1: Extract CPT for the hydrogen-based equity index

Semiconduct Techno	**Semiconduct(0.0) Techno(0.0)**	**...**	**Semiconduct(2.0) Techno(1.0)**	**Semiconduct(2.0) Techno(2.0)**
EQHydro(0.0)	0.4521	**...**	0.2747	0.2713
EQHydro(1.0)	0.3129	**...**	0.3067	0.3133
EQHydro(2.0)	0.2349	**...**	0.4186	0.4154

no confounders present in this graph. The isolation of the variable *EU_rate*, which represents the European interest rate, is evident. This isolation is attributed to the construction of the hydrogen equity index, which is based on a portfolio of American hydrogen sector-based clean-tech companies.

The structural equations derived from this graph are provided below:

$$
\begin{aligned}
EquityHydrogen &\sim Techno \\
Semiconduct &\sim US \\
Gas &\sim Oil \\
Oil &\sim US + Techno \\
Techno &\sim Semiconduct + US + Nickel \\
Carbon &\sim Inflation + Semiconduct + Oil \\
Carbon &\sim Semiconduct \\
EquityHydrogen &\sim Techno + Semiconduct
\end{aligned}
$$

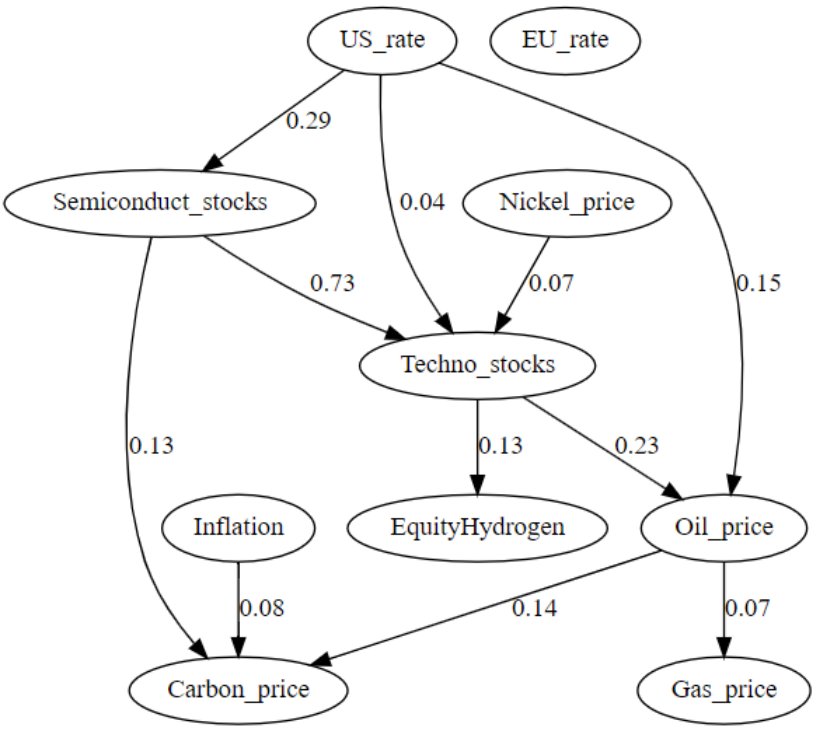
In Figure 27, we've merged three causality discovery methods–PCMCI+, Granger, and NTE—to form a cohesive causal graph. For each algorithm, we built a causal matrix, where rows represent explanatory variables and columns represent dependent variables. The values in the matrix indicate causality coefficients, ranging from 0 to 1. For example, in the Granger algorithm, a coefficient reflects the correlation between the estimated series from the past of the causing variable and the past of the variable of interest, compared to the series estimated solely from the past of the variable of interest. In NTE, coefficients come from standardized mutual information scores based on Shannon entropy. Once we have causality coefficients (and matrices) from each algorithm, we combine them into a single matrix. Then, we keep only the edges corresponding to coefficients surpassing a specified threshold. The graph visually represents the strength of these edges: darker colors indicate higher causality coefficients, signaling a more significant causal effect.

Figure 28 depicts the dynamic Bayesian networks constructed using a one-year time step spanning from 2008 to 2022, employing the Hill-climbing algorithm. Our analysis focuses on two key periods:

1. In the 2008-2009 phase (Figure 28a), corresponding to the financial crisis, we observe a negative effect (-0.02) of the carbon price on the hydrogen-based clean-techs index. Similarly, american interest rates also exhibit a negative effect (-0.04), driven by the fact that during the financial crisis of 2008-2009, higher interest rates negatively impacted borrowing and investment, leading to a downturn in economic activities and, consequently, a decrease in the hydrogen-based clean-techs index.

2. During 2019-2020 (Figure 28b), aligning with the onset of the COVID-19 crisis, semiconductor stocks emerge as the most influential causal factor on the Hydrogen-based

Figure 26: Causal graph obtained by DirectLingam



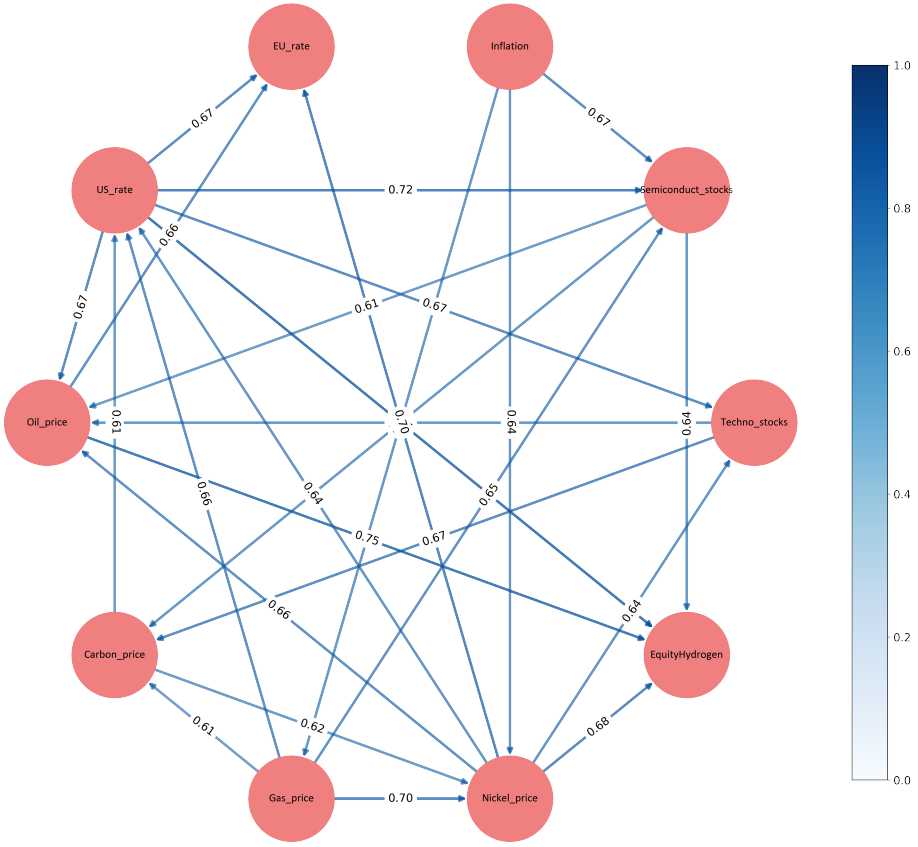
clean-techs index. This phenomenon is attributed to the heightened interdependence between the electronics and clean-tech industries during the climate transition period. Additionally, the COVID-19 crisis propelled hydrogen-related companies, closely linked to the surge in electronic-related companies. Furthermore, we observe the impact of inflation (-0.02) and gas prices (0.09) because, during the 2019-2020 period marked by the onset of the COVID-19 crisis, inflationary pressures influenced the overall economic landscape, affecting production costs and consumer spending. The increase in gas prices is indicative of shifts in energy markets and heightened demand for clean technologies amid the crisis, contributing positively to the valuation of hydrogen-based clean-techs.

4.3 Counterfactual queries

In this section, our objective is to address causal queries, exploring hypothetical scenarios and uncovering the potential impact of interventions. Some key questions we aim to answer include:

- What would have occurred in each sector-based clean-tech index if the oil price had increased from level 0 to level 2?
- How would each sector-based clean-tech index be affected if the oil price had risen from level 1 to level 2 while the technology stocks index simultaneously decreased from level 2 to level 0?

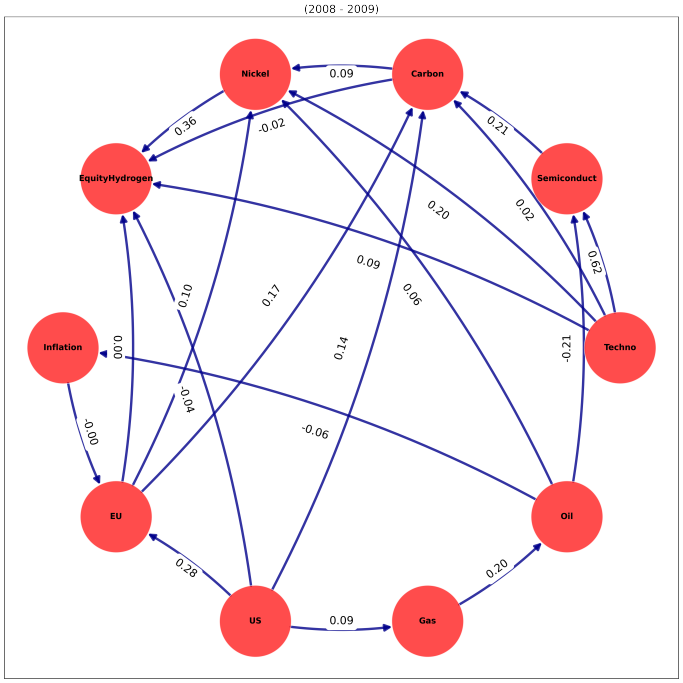
Figure 27: Causal graph obtained by Ensemble



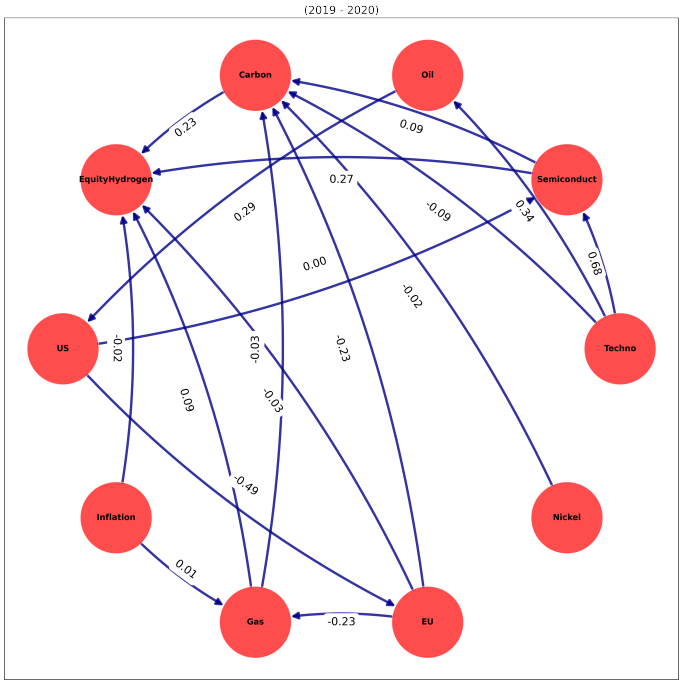
- What constitutes the most significant causal influence on the hydrogen-based index? And on the agribusiness-based index?

The scope of possible scenarios is vast, with each what-if question leading to a unique hypothetical situation. However, we have established theoretical foundations to both generate and evaluate these scenarios. Our approach involves intervention, allowing us to derive a counterfactual probability distribution across the causal graph. From this distribution, we can extract various metrics, expressed as ordered moments (expectation, conditional expectation, etc.) of the counterfactual (post-intervention) variables. We will focus on three metrics : average causal effect, average treatment effect and greatest causal influence. It's

Figure 28: Dynamic Bayesian network obtained by Hill-climbing



(a) Causal graph obtained by Hill-climbing for 2008-2009



(b) Causal graph obtained by hill-climbing for 2018-2019

essential to emphasize that the application of intervention necessitates the use of do-calculus.

This underscores the profound significance of Judea Pearl's contributions, as do-calculus can be regarded as a powerful inference tool for counterfactual models. Furthermore, it's worth highlighting once more that do-calculus unveils a fascinating connection to Bayesian calculus, particularly within the intriguing realm of counterfactual. In this unique conceptual space, do-calculus doesn't just perform mathematical operations; it orchestrates a symphony of probabilities and causal intricacies, akin to the principles underlying Bayesian reasoning. Pearl's work facilitates the transition from the *normal world* to a hypothetical one, where we conduct interventions and assessments before returning to the normal world armed with answers to our queries. This interplay between reality and hypothetical scenarios makes Pearl's do-calculus a truly remarkable analogy to complex numbers: just as we turn to imaginary numbers to solve equations that elude real numbers, we delve into counterfactual models to address questions that transcend the limits of observational data. Do-calculus is not merely a tool; it's a storyteller decoding the intricate language of causation. It navigates the uncertainties of the *what might have been* with the precision of Bayesian reasoning, bringing forth a deeper understanding of the causal tapestry that underlies our observations.

In what follows, all the results are obtained with the causal graph derived by the DirectLiNGAM algorithm. But of course, we can retrieve similar results for the other algorithms.

4.3.1 Average causal effect and greatest causal influence

Table 2 provides an overview of the average causal effects of explanatory variables on each sector-based clean-tech equity index. The largest ACE is highlighted in bold. Positive effects are observed for almost all explanatory variables, with the exception of Inflation. Notably, interest rates exhibit minimal impact on the hydrogen-based index. A detailed analysis of the annual reports for hydrogen-based clean-techs reveals their limited sensitivity to interest rates due to the nature of their business model. Hydrogen-based clean-tech companies often operate within a niche market that is primarily influenced by factors such as technological advancements, government policies, and global demand for clean energy solutions. Unlike industries more directly impacted by interest rate fluctuations, such as finance or real estate, the hydrogen-based clean-tech sector tends to prioritize innovation and sustainable practices. In green, we highlight the greatest causal influence. It's important to note that the GCI may not necessarily align with the variable with the largest ACE, as it specifically quantifies the causal impact on a local aspect. As depicted, both gas price and Oil price consistently emerge as the most common GCIs, except for the Solar-based clean-tech equity index, where the Semiconductor companies stocks index takes precedence. This can be attributed to the intricate relationship between these sectors. Semiconductor companies play a pivotal role in the development and advancement of solar technologies, providing crucial components for solar panels and related applications. Consequently, fluctuations in the Semiconductor companies stocks index can exert a significant cascading effect on the Solar-based clean-tech equity index.

4.3.2 Impact of interventions

We will focus our study on the hydrogen-based clean-tech equity index but again, similar results can be obtained for the other sector-bases groups. Having explored the average causal effects and the most significant causal influences of explanatory variables on each sector-based clean-tech equity group, we will now study the impact of intervention of explanatory variables on the hydrogen equity index. We create two hypothetical scenarios for the Hydrogen-based clean-tech equity index and address counterfactual queries. The ATE will be employed to evaluate the impact of interventions.

Table 2: Average causal effects and greatest causal influence for each sector-based clean-tech equity index (In green : GCI ; In bold : ACE)

	Techno	Semi	EU rate	US rate	Oil	Carbon	Gaz	Nickel	Inflation
Hydro EQ	0.13	0.10	0.00	0.02	**0.25**	0.13	0.14	0.01	0.00
Env. EQ	**0.37**	0.27	0.01	0.08	0.18	0.14	0.13	0.17	-0.05
Energy EQ	0.38	**0.41**	0.01	0.28	0.33	0.09	0.20	0.13	-0.01
Agri EQ	0.18	**0.37**	0.00	0.15	0.16	0.02	0.01	0.09	-0.05
Chemical EQ	0.56	**0.59**	0.03	0.33	0.19	0.03	0.02	0.16	-0.11
Gaz EQ	0.12	**0.38**	0.01	0.21	0.31	0.04	0.11	0.16	-0.02
Solar EQ	0.13	**0.25**	0.00	0.08	0.03	0.14	0.03	0.01	-0.08

Figure 29 presents the hypothetical scenario achieved by altering a single variable: the semiconductor stocks index. This scenario aims to answer the question:

- What would have happened to the Hydrogen Equity index if the Semiconductor stocks index had decreased from level 2 to level 0 between 2007 and 2023?

The computed ATE is -25%. This signifies that, on average, if the semiconductor stocks index decreases from level 2 to level 0, the hydrogen clean-tech equity index would also decrease by 25%.

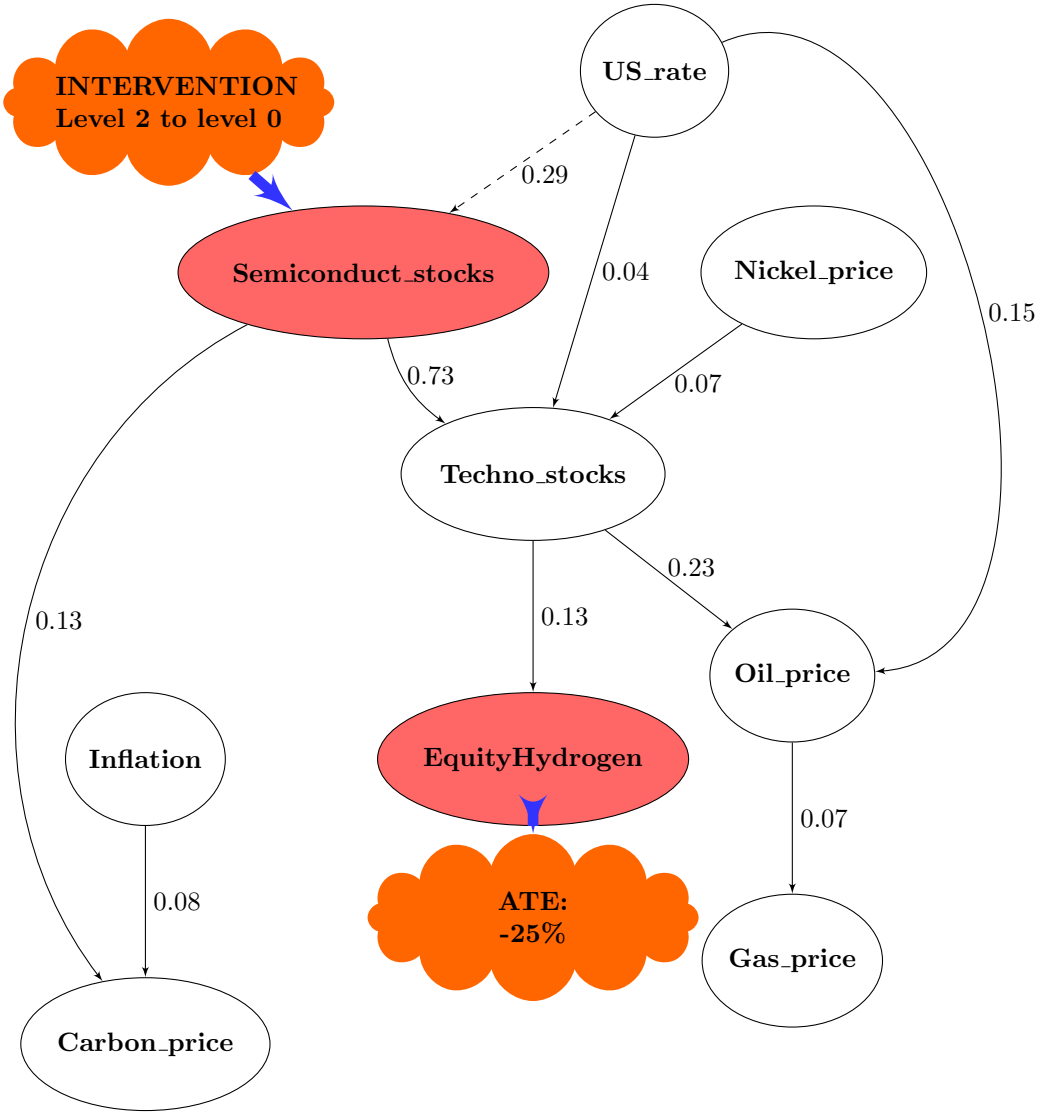
Continuing, Figure 30 depicts a hypothetical scenario achieved by simultaneously altering two variables: the semiconductor stocks index and the nickel price. This scenario aims to address the following question:

- What would have happened to the hydrogen Equity index if the Semiconductor stocks index had increased from level 0 to level 1, and the nickel price had decreased from level 2 to level 0 between 2007 and 2023?

The calculated ATE is +7%. This implies that, on average, if the semiconductor stocks index increases from level 0 to level 1, and the nickel price decreases from level 2 to level 0, the Hydrogen clean-tech equity index would increase by 7%. This observation suggests that the semiconductor stocks index exerts a greater influence on the target variable than the nickel price. This intuition aligns with the higher average causal effect for the semiconductor index (0.10) compared to the nickel price (0.01).

Finally, to gain deeper insights, let's explore a straightforward scenario: intervening on the oil price by setting it to level 1. We then compare the hydrogen-based clean-tech equity index samples before and after the intervention.
In Figure 31, we illustrate the impact of intervening on oil prices on the hydrogen-based clean-tech equity index samples, both before and after the intervention. We clearly see that the post-intervention curves mitigate oscillations. This effect stems from setting the oil price at a fixed value, which eliminates the inherent randomness associated with its market fluctuations. Consequently, the post-intervention sample of the hydrogen-based clean-tech equity index exhibits reduced oscillations. This is a logical outcome, as the deterministic oil price creates a more controlled and predictable market scenario. With the removal of the stochastic nature of oil prices, the hydrogen equity index responds with greater stability.

Figure 29: Impact of one intervention on the DirectLinGAM causal graph for the target variable hydrogen Equity index

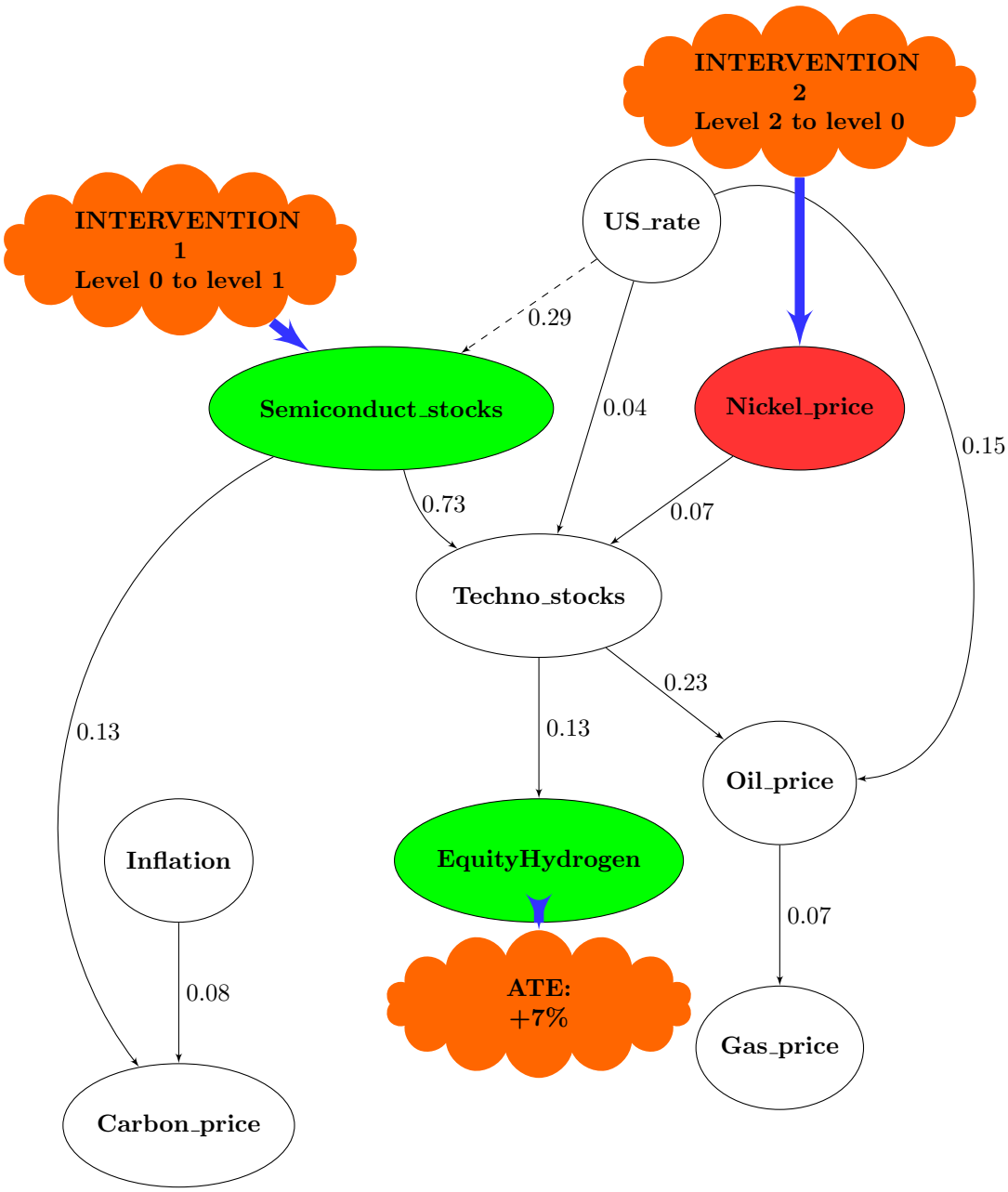


5 Discussion

In summary, our research has utilized a comprehensive methodology to explore the causal intricacies of the clean-tech industry. The process incorporated foundational concepts such as Bayesian networks, causality algorithms, and Judea Pearl's do-calculus, which mathematically encapsulates the fundamental distinction between causation and correlation. Applying these methods resulted in insightful analyses and discoveries.

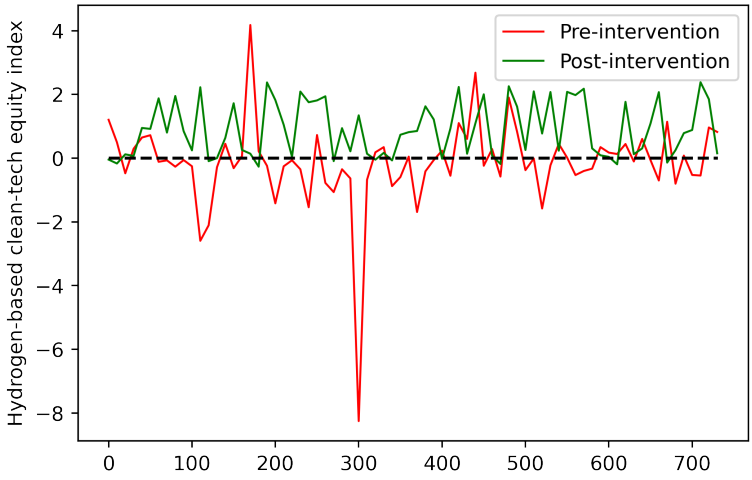
We applied our methodology to scrutinize the causal effects and the impacts of interventions by explanatory variables on identified clean-tech groups categorized by sectors. The

Figure 30: Impact of two simultaneous interventions on the DirectLinGAM causal graph for the target variable hydrogen Equity index



findings revealed that our approach can be extended to portfolios, allowing the assessment of macro variable perturbations on these investment collections. However, it's crucial to acknowledge that the identification of causal structures between variables is not infallible, emphasizing the need for expert oversight to ensure model accuracy. Importantly, our approach does not aim to predict future values but centers around the generation of counterfactual scenarios. This aligns with the concept of regret and underscores the potency of Judea Pearl's

Figure 31: Comparison between the hydrogen-based clean-Tech equity index before and after intervention on oil price



do-calculus in transitioning to a post-intervention world to answer pre-intervention queries. Nevertheless, we must recognize the limitations of our current framework, which primarily deals with static Bayesian networks in discrete spaces. Future extensions may explore continuous-time Bayesian networks, although the theoretical complexity poses challenges.

In the expansive field of causality research, cutting-edge methodologies are emerging, including causal Shapley values [17]. This approach generalizes our work by applying Pearl's do-calculus to derive asymmetric values for general causal graphs that explain the total effect of features on the prediction, offering a more direct and robust way to incorporate causal knowledge. Another approach [54] involves leveraging the power of Partial Dependence Plots (PDP) to extract causal insights from black-box models generated by machine learning algorithms. This method is rooted in the observation that Friedman's partial dependence plot aligns with Pearl's back-door adjustment. It establishes the viability of extracting causal information and demonstrates that Bayesian calculus can be viewed equivalently as a straightforward statistical inference task. Another interesting study would be to follow the approach of [53]. Their methods allow users to identify and specify the subset of parameters associated with causal models and randomize the remaining parameters to generate a range of data generation processes consistent with this method.

In closing, our research contributes valuable insights to the evolving understanding of causation within the clean-tech sector. The combination of advanced methodologies and domain expertise enhances our comprehension of the intricate relationships governing the industry. While the do-calculus proves to be a powerful tool, ongoing research continues to explore novel approaches, underscoring the necessity for a holistic understanding of causation in the complex systems that characterize the clean-tech landscape. In future work, the Pearl approach will be used to construct equity baskets based on causal queries and counterfactuals type intervention on the explanatory variables.

References

[1] Ankan, A., Wortel, I.,Bollen, K. and Textor, J. (2023), Combining Graphical and Algebraic Approaches for Parameter Identification in Latent Variable Structural Equation Models, *arXiv*, 2302.13220.

[2] Badoy, E., Ainley J. and Teja, H. (2023), Macro to Micro: Implications of the European Response to the US Inflation Reduction Act, *Citi Europe Research*.

[3] Barber, D. (2012), *Bayesian Reasoning and Machine Learning*, Cambridge University Press.

[4] Bjornali, E. and Ellingsen, A. (2014), Factors Affecting the Development of Clean-tech Start-ups: A Literature Review, *Energy Procedia*, 58, pp. 43-50.

[5] Blobaum, P. and Shimizu, S. (2017), Estimation of Interventional Effects of Features on Prediction, *2017 IEEE 27th International Workshop on Machine Learning for Signal Processing (MLSP)*.

[6] Blobaum, P., Gotz, P., Budhathoki, K., Mastakouri, A. and Janzing, D. (2022), DoWhy-GCM: An Extension of DoWhy for Causal Inference in Graphical Causal Models, *arXiv*, 2206.06821.

[7] Blondel, G. (2023), Causal Discovery and Prediction: Methods and Algorithms, HAL open science, *arXiv*, 2309.09416.

[8] Bouri, E., Dudda T., Rognone, L. and Walther, T. (2022), Climate Risk and the Dynamic Correlation Between Clean Energy and Technology Stock Markets, *Annals of Operations Research*, Springer.

[9] Bregoli, A., Scutari, M. and Stella, F. (2021), A Constraint-Based Algorithm for the Structural Learning of Continuous-Time Bayesian Networks, *International Journal of Approximate Reasoning*, 138, pp. 105-122.

[10] Bua, G., Kapp, D., Ramella, F. and Rognone, L. (2022), Transition versus Physical Climate Risk Pricing in European Financial Markets: A Text-based Approach, *ECB Working Paper*, 2677.

[11] Byrd, S., Arcaro, D., Percoco, A., Eloizin, P. and Baskar, A. (2023), 2023 Outlook: Strong Multiyear Growth, Despite Near-Term Execution Challenges, *Morgan Stanley Research*.

[12] Chang, A. and Bradsher, K. (2023), Can the World Make an Electric Car Battery Without China, *New York Times*.

[13] Chao, P., Blobaum, P. and Kasiviswanathan, S. (2023), Interventional and Counterfactual Inference with Diffusion Models, *arXiv*, 2302.00860.

[14] Chauhan, R., Riis, C., Adhikari, S., Derrible, S., Zheleva, E., Choudhoury, C. and Pereira, F. (2022), Determining Causality in Travel Mode Choice, *arXiv*, 2208.05624.

[15] Chen, B. and Pearl, J. (2015), Graphical Tools for Linear Structural Equation Modeling, *University of California Los Angeles*.

[16] Gendron, G., Witbrock, M. and Dobbie, G. (2023), A Survey of Methods, Challenges and Perspectives in Causality, *arXiv*, 2302.00293.

[17] Heskes, T., Sijben, E., Bucur, I. and Claassen, T. (2020), Causal Shapley Values: Exploiting Causal Knowledge to Explain Individual Predictions of Complex Models, *NeurIPS Proceedings*.

[18] Imbens, G. (2019), Potential Outcome and Directed Acyclic Graph Approaches to Causality : Relevance for Empirical Practice in Economics, *Journal Of Economic Literature*, 58(4), pp. 1129-79.

[19] Jensen, F. and Nielsen, T. (2007), *Bayesian Networks and Decision Graphs*, Springer.

[20] Karvanen, J., Tikka, S. and Vihola, M. (2023), Simulating Counterfactuals, *arXiv*, 2306.15328.

[21] Koller, D. and Friedman, N. (2009), Probabilistic Graphical Models : Principles and Tecnhiques, MIT Press.

[22] Konstantinos, N., Panayotis, G., Panos, X. and Stravoula, Y. (2023), Carbon Emissions and Sustainability in Covid-19's Waves: Evidence from a Two-state Dynamic Markov-switching Regression (MSR) Model, *Annals of Operations Research*, Springer.

[23] Kosowski, P., Kosowska, K. and Nawalaniec, W. (2022), Application of Bayesian Networks in Modeling of Underground Gas Storage Energy Security, *Energies*, 15(14), pp. 51-85.

[24] Lattimore, F. and Rohde, D. (2019), Causal Inference with Bayes Rule, *arXiv*, 1910.01510.

[25] Lattimore, F. and Rohde, D. (2021), Replacing the Do-calculus with Bayes Rule, *arXiv*, 1906.07125.

[26] Lee, B., Fischer, D., Clarke, Z. Delaney, M., Konig, P., Ritchie, J., Revich, J., Samuelson, A., Mehta, N., Blostein, A. and Singer, B. (2023), Inflation Reduction Act: From Theme to Tailwind - Quantifying the Impact Across Top 20 Ideas, *Goldman Sachs Equity Research*.

[27] Liu, W., Wang, J., Wang, H. and Li, R. (2023), D-Separation for Causal Self-Explanation, *arXiv*, 2309.13391.

[28] Mohammad-Taheri, S., Zucker, J., Tapley Hoyt, C., Sachs, K., Tewari, V., Ness, R. and Vitek, O. (2022), Do-calculus Enables Estimation of Causal Effects in Partially Observed Biomolecular Pathways, *Bioinformatics*.

[29] Morgan, S. and Winship, C. (2007), *Counterfactuals and Causal Inference Methods and Principles for Social Research*, Cambridge University Press.

[30] Nodelman, U., Shelton, C. and Koller, D. (2002), Continuous Time Bayesian Networks, *Proceedings of the Eighteenth International Conference on Uncertainty in Artificial Intelligence*.

[31] Olson E., Kurov, A. and Wolfe, M. (2023), Have the Causal Effects between Equities, Oil Prices, and Monetary Policy Changed Over Time?, *SSRN*, 4573178.

[32] Pearl, J., Glymour, M. and Jewell, N. (2016), *Causal Inference in Statistics: A Primer*, Wiley.

[33] Peters, J., Janzing, D. and Scholkopf, B. (2017), *Elements of Causal Inference Foundations and Learning Algorithm*, The MIT Press Cambridge, Massachusetts.

[34] Pfister, N. and Peters, J. (2022), Identifiability of Sparse Causal Effects Using Instrumental Variables, *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, 180.

[35] Pulakesh, U., Kai, Z., Can, L., Xiaoqian, J. and Yejin, K. (2021), Scalable Causal Structure Learning: Scoping Review of Traditional and Deep Learning Algorithms and New Opportunities in Biomedicine, *MIR Medical Informatics*.

[36] Rebonato, R. (2010), *Coherent Stress Testing, A Bayesian Approach to the Analysis of Financial Stress*, Wiley.

[37] Rohmer, J. (2020), Uncertainties in Conditional Probability Tables of Discrete Bayesian Belief Networks: A Comprehensive Review, *Engineering Applications of Artificial Intelligence*, 88(4).

[38] Runge, J. (2020), Discovering Contemporaneous and Lagged Causal Relations in Autocorrelated Nonlinear Time Series Datasets, *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence*, 124, pp. 1388-1397.

[39] Runge, J., Nowack, P., Kretschmer, M., Flaxman, S. and Sejdinovic, D. (2019), Detecting Causal Associations in Large Nonlinear Time Series Datasets, *Science Advances*, 5(11).

[40] Scholkopf B., and Kugelgen, J. (2022), From Statistical to Causal Learning, *EMS Press*, 7, pp. 5540–5593.

[41] Sharma, A. and Kiciman, E. (2020), DoWhy: An End-to-End Library for Causal Inference, *arXiv*, 2011.04216.

[42] Sharma, A., Syrgkanis, V., Zhang, C. and Kiciman, E. (2021), DoWhy: Addressing Challenges in Expressing and Validating Causal Assumptions, *arXiv*, 2108.13518.

[43] Shimizu, S., Inazumi, T., Sogawa, Y., Hyvarinen, A., Kawahara, Y., Washio, T., Hoyer, P. and Bollen, K. (2011), DirectLiNGAM : A Direct Method for Learning a Linear Non-Gaussian Structural Equation Model, *Journal of Machine Learning Research*, 12, pp. 1225-1248.

[44] Shimizu, S., Hoyer, P., Hyvärinen, A., and Kerminen, A. (2006), A linear non-Gaussian acyclic model for causal discovery, *Journal of Machine Learning Research*, 7, pp. 2003–2030.

[45] Shuai, K., Luo, S., Zhang, Y., Xie, F. and He, Y. (2023), Identifiability of Causal Effects with Non-Gaussianity and Auxiliary Covariates, *arXiv*, 2304.14895.

[47] Smith, B. (2023), Causal Discovery and Counterfactual Recommendations for Personalized Student Learning, *arXiv*, 2309.13066.

[47] Smith, B. (2023), Evaluating Counterfactual Explanations Using Pearl's Counterfactual Method, *arXiv*, 2301.02499.

[48] Spirtes, P., Glymour, C., and Scheines, R. (2001), *Causation, Prediction, and Search*, MIT Press, Cambridge, MA.

[50] Stanley, E., Byrd, S., Agarwal, A., De Maere, J., Sanchez, L. and Duverce, B. (2023), The Price of Purity: Fundamentals, *Morgan Stanley Research*.

[50] Stanley, E., Byrd, S., Agarwal, A., De Maere, J., Sanchez, L. and Duverce, B. (2023), The Price of Purity: Valuation, *Morgan Stanley Research*.

[51] Tucci, R. (2013), Introduction to Judea Pearl's Do-Calculus, *arXiv*, 1305.5506.

[52] Wang, X., Zhao, M., Meng, F., Liu, X., Kong, Z. and Chen, X. (2023), Quantifying Causal Path-Specific Importance in Structural Causal Model, *Computation*, 11(7).

[53] Zamanian, A., Mareis, L. and Ahmidi, N. (2023), Partially Specified Causal Simulations, *arXiv*, 2309.10514.

[54] Zhao, Q. and Hastie, T. (2019), Causal Interpretations of Black-Box Models, *Journal of Business and Economic Statistics*, 39(1), pp. 272-281.

[55] Zheng, Y., Huang, B., Chen, W., Ramsey, J., Gong, M., Cai, R., Shimizu, S., Spirtes, P. and Zhang, K. (2023), Causal-learn: Causal Discovery in Python, *arXiv*, 2307.16405.

[56] Zhipeng, M., Kemmerling, M., Buschmann, D. , Enslin, C., Lutticke, D. and Schmitt, R. (2023), A Data-Driven Two-Phase Multi-Split Causal Ensemble Model for Time Series, *Symmetry*, 15(5), pp. 982.

A Markov random fields

In this appendix, we delve into the concept of Markov random fields, which constitute the second classical type of graphical models after Bayesian networks [3]. A Markov network, also known as a Markov random field, is a graphical model that represents the joint probability distribution of a set of random variables, where the variables are represented as nodes in an undirected graph. Mathematically, a Markov network is defined by an undirected graph $G = (V, E)$, where V represents the set of nodes (random variables) and E represents the set of edges (pairwise dependencies). Let $X = (X_1, X_2, ..., X_n)$ be the set of random variables associated with the nodes of the graph.

In the context of Markov networks, a clique C is a fully connected subgraph within the graphical structure. It consists of a set of nodes in which each pair of nodes is connected by an edge. This is equivalent to the condition that the subgraph of G induced by C is a complete graph. A potential function is associated with each clique in the graph. A potential function, denoted as $\psi_C(X_C)$, represents a non-negative real-valued function defined on the variables X_C corresponding to the nodes in the clique C. It encodes the local interactions or dependencies between the variables in the clique. The joint probability distribution over X in a Markov network is given by:

$$\mathbb{P}(X) = \frac{1}{Z} \prod_{C \in \mathcal{C}} \psi_C(X_C)$$

where:

- X_C denotes the set of variables associated with the clique C in the graph.
- $\psi_C(X_C)$ is a non-negative potential function defined on the clique C.
- $\mathcal{C}$ represents the set of all cliques in the graph.
- Z is a normalization constant called the partition function, defined as

  $$Z = \sum_X \prod_{C \in \mathcal{C}} \psi_C(X_C),$$

  where the summation is taken over all possible assignments of values to the variables X.

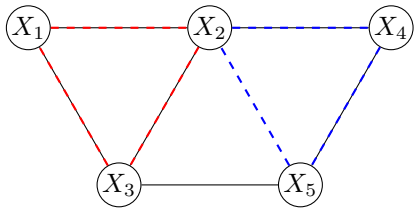
The partition function Z in a Markov network plays a crucial role, similar to its counterpart in statistical physics, as it ensures that the joint probability distribution satisfies the properties of a probability measure. It allows for computations of various quantities of interest and probabilistic inferences about the variables in the network. In the graph below, $X_1 - X_2 - X_3$ form a clique. If X_2 and X_5 were linked, then $X_2 - X_4 - X_5$ would form a clique.

Proposition 4. ***Conditional dependencies in a Markov network***
Following example in Figure 32, we present some rules on conditional dependencies in a Markov network:

- *The joint distribution is given by :*

$$\begin{aligned}\mathbb{P}(X_1, X_2, X_3, X_4, X_5) = \frac{1}{Z} &\psi_{X_1X_2X_3}(X_1, X_2, X_3) \\ &\cdot \psi_{X_2X_4}(X_2, X_4) \\ &\cdot \psi_{X_4X_5}(X_4, X_5) \\ &\cdot \psi_{X_3X_5}(X_3, X_5).\end{aligned}$$

Figure 32: Example of Markov network



- *Marginalising over X_3 makes X_2 and X_5 graphically dependent, i.e. $\mathbb{P}(X_2, X_5) \neq \mathbb{P}(X_2) \cdot \mathbb{P}(X_5)$.*
- *X_2 and X_5 are conditionally independent given X_3, i.e. $\mathbb{P}(X_2, X_5 \mid X_3) = \mathbb{P}(X_2 \mid X_3) \cdot \mathbb{P}(X_5 \mid X_3)$.*
- *A set of variables A is considered independent of a set of variables B, given a set Z, if there exists no path connecting any variable in set A to any variable in set B that goes through the variables in set Z. To put this into practice, we remove all links that directly connect variables in Z with their neighboring variables. If, after removing these connections, there are no remaining paths from any member of set A to any member of set B, then we denote this relationship as $A \perp B \mid Z$.*

Let X be a variable in the Markov network, and let $N(X)$ represent the set of neighboring variables of X. The Markov property with neighbors can be expressed as:

$$X \perp (V - \{X\} - N(X)) \mid N(X),$$

where V represents all the variables in the Markov network. This property implies that a variable is only influenced by its immediate neighbors and is independent of all other variables in the network, given these neighbors.

Theorem 5 (Hammersley-Clifford). *Let $X = \{X_1, X_2, \ldots, X_n\}$ be a set of random variables, and let $G = (V, E)$ represent the corresponding Markov network, where V denotes the set of nodes and E represents the set of edges between the nodes.*

- *The **local Markov property** states that each variable X_i is conditionally independent of all other variables X_j given its neighbors $X_{ne(X_i)}$ in the Markov network:*

$$X_i \perp X_j \mid X_{ne(X_i)}, \quad \forall i, j \in \{1, 2, \ldots, n\}$$

- *The **global Markov property** states that if a set of variables S separates variables A from variables B in the Markov network G, then variables A are conditionally independent of variables B given S. Specifically, S separates A from B if, for every path between a variable in A and a variable in B in the graph G, there exists at least one variable in S that blocks the path :*

$$A \perp B \mid S, \quad \text{if } S \text{ separates } A \text{ from } B \text{ in } G$$

The Hammersley-Clifford theorem states that a joint probability distribution $\mathbb{P}(X_1, \ldots, X_n)$ factorizes according to the Markov network G if and only if it satisfies both the local Markov property and the global Markov property. Mathematically, this can be expressed as:

$$\mathbb{P}(X_1, X_2, \ldots, X_n) = \frac{1}{Z} \prod_{C \in \mathcal{C}(G)} \psi_C(X_C)$$

where $\mathcal{C}(G)$ represents the set of all maximal cliques in the graph G, X_C denotes the variables in clique C, $\psi_C(X_C)$ represents a potential function associated with clique C, and Z is the partition function.

Remark 3. ***Ising Model***

The Ising model, widely used in statistical physics and related fields, can be seen as a special case of Markov networks. In the Ising model, variables represent spins in a lattice, typically taking values of +1 or -1. The interactions between neighboring spins are captured by assigning weights, denoted as J, to the edges connecting them. To relate the Ising model to Markov networks, we can construct a Markov network that corresponds to the Ising model. Each node in the Markov network represents a spin variable, and the edges between them encode the interactions between neighboring spins. The joint probability distribution of the Ising model can be represented as a product of potentials in the Markov network.

For example, consider a 2D lattice with spin variables $X_{i,j}$, where i and j denote the coordinates of the lattice sites. The joint distribution in the Ising model can be represented in the Markov network as:

$$\mathbb{P}(X_{1,1}, X_{1,2}, X_{2,1}, X_{2,2}) = \frac{1}{Z} \exp(-E(X_{1,1}, X_{1,2}, X_{2,1}, X_{2,2}))$$

where Z is the partition function and $E(X_{1,1}, X_{1,2}, X_{2,1}, X_{2,2})$ represents the energy function of the Ising model. The Markov network representation allows us to apply graphical models techniques, such as determining conditional independencies and performing inference, to analyze the Ising model and understand its properties.

$$\begin{array}{ccc} X_{1,1} & \longleftrightarrow & X_{1,2} \\ \updownarrow & & \updownarrow \\ X_{2,1} & \longleftrightarrow & X_{2,2} \end{array}$$

B Bayesian networks: parameters estimation

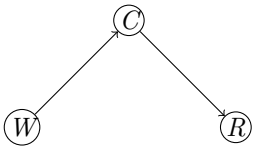
In this appendix, we delve into the process of learning Bayesian Networks, which involves various aspects such as parameter estimation, inference techniques, and structure learning. The content of this appendix is inspired from [3] and [19]. Parameter estimation focuses on determining the conditional probability tables associated with the variables in the network based on observed data. By analyzing the available data, we can estimate the parameters that best fit the given distribution, allowing us to make informed predictions and perform probabilistic reasoning. Smoothing, filtration, and prediction are essential inference techniques in Bayesian Networks. Smoothing refers to estimating the past states of variables given observed evidence up to the present. Filtration involves estimating the current state of variables given the evidence up to a certain point in time. Prediction, on the other hand, is concerned with estimating the future states of variables given the available evidence. Structure learning, another crucial aspect of learning Bayesian Networks, focuses on determining the graphical structure of the network itself. This involves identifying the dependencies and causal relationships among variables based on the observed data. By uncovering the underlying structure, we gain valuable insights into the system and can make accurate predictions and decisions.

In the following sections of this appendix, we will explore these topics in more detail, discussing different methods and algorithms for learning Bayesian Networks and providing practical examples to enhance understanding. Parameter estimation typically involves two main approaches: maximum likelihood estimation (MLE) and Bayesian estimation.

B.1 Maximum Likelihood Estimation (MLE)

MLE seeks to find the parameter values that maximize the likelihood of observing the given data. This approach assumes that the data are independent and identically distributed (i.i.d.), and it aims to find the parameter values that make the observed data most probable. In the case of Bayesian Networks, MLE involves estimating the probabilities in the CPTs that best fit the data. Let's start with a small example :

Example 3. *Consider a Bayesian network that models the relationship between the weather forecast (W), the presence of clouds (C), and the likelihood of rain (R). The network structure is as follows:*



In this network, each variable can take on two states: The weather variable can take the states sunny (S) or cloudy (C).The variables (C) and (R) can take the values 0 or 1, with 1 indicating the presence of clouds (resp. the presence of rain). The variables are related as follows: the weather variable (W) influences the presence of clouds (C), and the presence of clouds affects the likelihood of rain (R). To learn the conditional probabilities for this Bayesian Network, we can collect a dataset of weather observations. Let's assume we have a dataset containing the following information:

W	C	R
S	*0*	*0*
S	*0*	*0*
S	*1*	*0*
C	*1*	*1*
C	*1*	*1*

To estimate the probabilities, we count the occurrences of each combination of states and normalize them to obtain probabilities. For example, to estimate $\mathbb{P}(C = 1 \mid W = S)$*, we count the number of times clouds are present (*$C = 1$*) when the weather is sunny (*$W = S$*), which gives us a count of 1. Dividing this count by the total number of occurrences of* $W = S$ *(which is 3 in this case) gives us the probability* $\mathbb{P}(C = 1 \mid W = S) = \frac{1}{3}$*. Similarly, we can estimate other conditional probabilities, such as* $\mathbb{P}(R = 1 \mid C = 1)$*, by counting the number of times rain occurs (*$R = 1$*) when clouds are present (*$C = 1$*) and dividing it by the total number of occurrences of* $C = 1$*. The resulting conditional probability tables for the Bayesian Network can be filled in using these estimated probabilities, allowing us to perform probabilistic reasoning and make predictions based on observed evidence.*

To estimate the CPTs of a Bayesian network using maximum likelihood estimation, we aim to find the parameter values that maximize the likelihood of observing the given data. Maximizing the log-likelihood is equivalent to minimizing the Kullback-Leibler (KL) divergence with the empirical distribution.

Definition 11. ***Empirical Distribution:*** *The empirical distribution, denoted by* $\hat{P}$*, represents the observed frequencies of the data. For a given dataset* D *with* N *samples, the empirical distribution is defined as follows:*

$$\hat{\mathbb{P}}(X = x) = \frac{\#(X = x)}{N}$$

where $\#(X = x)$ *denotes the number of occurrences of the variable* X *taking the value* x *in the dataset* D*.*

For a given Bayesian network B and dataset D, the maximum likelihood estimate of the CPTs is obtained by maximizing the log-likelihood:

$$\hat{\theta}_{\text{MLE}} = \arg\max_{\theta} \log \mathbb{P}(D \mid \theta)$$

where $\hat{\theta}_{\text{MLE}}$ represents the MLE estimate of the parameters θ.

Definition 12. ***Kullback-Leibler divergence***

The Kullback-Leibler divergence is a measure of the dissimilarity between two probability distributions. Let X *and* Y *be two discrete random variables with supports* $\mathbb{R}_X$ *and* $\mathbb{R}_Y$ *and probability mass functions* $\mathbb{P}_X$ *and* $\mathbb{P}_Y$ *respectively. Then the Kullback-Leibler divergence of the probability mass function* $\mathbb{P}_Y$ *from* $\mathbb{P}_X$ *is given by*

$$D_{KL}(\mathbb{P}_X \parallel \mathbb{P}_Y) = \sum_{x \in \mathbb{R}_X} \mathbb{P}_X(x) \log \frac{\mathbb{P}_X(x)}{\mathbb{P}_Y(x)}.$$

Now, if X *and* Y *be two* ***continuous*** *random variables with probability density functions* $f_X(x)$ *and* $f_Y(y)$*, defined over their respective supports. Then the Kullback-Leibler (KL) divergence of the probability density function* f_Y *from* f_X *is given by*

$$D_{KL}(f_X \parallel f_Y) = \int_{-\infty}^{\infty} f_X(x) \log \frac{f_X(x)}{f_Y(x)}\, dx.$$

Proposition 5. *Maximizing the log-likelihood is equivalent to minimizing the Kullback-Leibler (KL) divergence with the empirical distribution.*

Proof. To prove the proposition, let's start from the expression for the maximum likelihood estimate (MLE) of the parameters, denoted as $\hat{\theta}_{\text{MLE}}$:

$$\hat{\theta}_{\text{MLE}} = \arg\max_{\theta} \log \mathbb{P}(D \mid \theta)$$

Assuming that the data points are i.i.d., let's rewrite the expression for $\hat{\theta}_{\text{MLE}}$ using the empirical distribution defined by the training data, denoted as $\hat{P}$:

$$\hat{\theta}_{\text{MLE}} = \arg\max_{\theta} \frac{1}{m} \sum_{i=1}^{m} \log \mathbb{P}(x^{(i)}|\theta) = \arg\max_{\theta} \mathbb{E}_{x\sim\hat{\mathbb{P}}}[\log \mathbb{P}(x \mid \theta)]$$

Here, we have replaced the sum with an expectation over the empirical distribution $\hat{P}$, which assigns a probability of $\frac{1}{m}$ to each of the m points $x^{(i)}$ in the training data.

Now, let's consider the Kullback-Leibler (KL) divergence between the empirical distribution $\hat{P}$ and the model distribution P:

$$D_{\text{KL}}(\hat{\mathbb{P}}\|\mathbb{P}) = \mathbb{E}_{x\sim\hat{\mathbb{P}}}[\log \hat{\mathbb{P}}(x) - \log \mathbb{P}(x;\theta)]$$

Note that the expectation $\mathbb{E}_{x\sim\hat{P}}[\log \hat{\mathbb{P}}]$ does not depend on θ; it only depends on the data generating process. Therefore, it can be treated as a constant.

Minimizing the KL divergence $D_{\text{KL}}(\hat{\mathbb{P}}\|\mathbb{P})$ is equivalent to minimizing the expression inside the expectation:

$$\arg\min_{\theta} D_{\text{KL}}(\hat{\mathbb{P}}\|\mathbb{P}) = \arg\min_{\theta} \mathbb{E}_{x\sim\hat{\mathbb{P}}}[-\log \mathbb{P}(x;\theta)]$$

Comparing this expression with the earlier expression for $\hat{\theta}_{\text{MLE}}$, we can see that they are equivalent:

$$\hat{\theta}_{\text{MLE}} = \arg\max_{\theta} \mathbb{E}_{x\sim\hat{\mathbb{P}}}[\log \mathbb{P}(x \mid \theta)] = \arg\min_{\theta} \mathbb{E}_{x\sim\hat{\mathbb{P}}}[-\log \mathbb{P}(x;\theta)]$$

Therefore, maximizing the log-likelihood is equivalent to minimizing the Kullback-Leibler divergence with the empirical distribution. □

Now, using the non-negativity property of the Kullback-Leibler divergence, we have:

$$\mathbb{E}_{x\sim\hat{\mathbb{P}}}[\log \hat{\mathbb{P}}(x)] - \mathbb{E}_{x\sim\hat{\mathbb{P}}}[\log \mathbb{P}(x)] \geq 0$$

To minimize the KL divergence, we need to make the second term $\mathbb{E}_{x\sim\hat{\mathbb{P}}}[\log \mathbb{P}(x)]$ as small as possible. In other words, the minimizer of the KL divergence is obtained when $\mathbb{P}$ is equal to the empirical distribution $\hat{\mathbb{P}}$. Hence, we conclude that the MLE estimator of our CPT's are their empirical distributions.

Learning with hidden variables

In many real-world scenarios, the available data may not directly provide information about certain variables of interest. These unobserved variables are referred to as hidden variables. Learning the parameters of a Bayesian network in the presence of hidden variables poses additional challenges. The Expectation Maximization (EM) algorithm is an iterative optimization technique used in statistical inference and machine learning. It is designed to estimate the parameters of probabilistic models when some variables are unobserved or

missing. Let's note by v an observed variable and h a hidden variable. Let N be the number of data-points. Our interest is to set θ by maximising the marginal likelihood $\mathbb{P}(v \mid \theta)$. The model of the data is $\mathbb{P}(v, h \mid \theta)$. The EM algorithm consists of two main steps: the E-step and the M-step.

E-step: In this step, we compute the posterior distribution over the hidden variables h given the observed variable v and the current parameter estimate $\theta^{(t)}$. This can be expressed as:

$$\mathbb{P}(h^i \mid v^i, \theta^{(t)}) = \frac{\mathbb{P}(v^i, h^i \mid \theta^{(t)})}{\sum_h \mathbb{P}(v^i, h^i \mid \theta^{(t)})}$$

M-step: In this step, we update the parameter estimate by maximizing the expected complete log-likelihood. The expected complete log-likelihood is computed by taking the expectation over the posterior distribution of the hidden variables obtained in the E-step. Mathematically, the parameter update can be written as:

$$\theta^{(t+1)} = \arg\max_{\theta} \sum_{i=1}^{N} \mathbb{E}_{\mathbb{P}(h^i|v^i,\theta^{(t)})}[\log \mathbb{P}(v^i, h^i \mid \theta)]$$

The algorithm starts with an initial parameter estimate $\theta^{(0)}$ and iteratively performs the E-step and M-step until convergence is achieved.

Algorithm 4 Expectation-Maximization Algorithm

Require: Observed variable v, Model parameter θ
Ensure: Maximum likelihood estimates of θ
  Initialize $\theta^{(0)}$
  repeat
    E-step: Compute over all data-points the posterior distribution over hidden variables $p(h^i|v^i, \theta^{(t)})$
    M-step: Update the parameters by maximizing the expected complete log-likelihood $\theta^{(t+1)} = \arg\max_\theta \sum_{i=1}^{N} \mathbb{E}_{\mathbb{P}(h^i|v^i,\theta^{(t)})}[\log \mathbb{P}(v^i, h^i \mid \theta)]$
  until Convergence

Application to Bayesian networks

In a Belief network, we recall that when all the nodes represent visible or observed variables, the conditional probability distribution $\mathbb{P}(X_i \mid Pa(X_i))$, where $Pa(X_i)$ denotes the parents of the node X_i, can be determined by performing a frequency count of each combination of values for the involved variables based on the observed data. When some variables are hidden and others observed, the indicators are replaced by the assumed conditional distribution of the hidden variable found at the E-step. In fact, let's consider the Bayesian network at example 3. We suppose that the states of the variable C are never observed while the states of variable W and R are totally observed. Our goal is to learn the CPTs $\mathbb{P}(R \mid C)$, $\mathbb{P}(C \mid W)$ and $\mathbb{P}(W)$.

- **E-step** The E-step defines a set of distributions on the hidden variable C. We have that $P_t^{i=1}(C|W, R) = P(C|W = S, R = 0)$, $\mathbb{P}_t^{i=2}(C \mid W, R) = \mathbb{P}(C \mid W = S, R = 0)$, $\cdots$, $\mathbb{P}_t^{i=4}(C \mid W, R) = \mathbb{P}(C \mid W = C, R = 1)$, and so on for the 5 training examples.

- **M-step:** Our interest is to maximize the following expression:

$$\sum_{i=1}^{5} \mathbb{E}_{\mathbb{P}_t^i(C|W,R)}[\log \mathbb{P}(C^i, W^i, R^i)] = \sum_{i=1}^{5} \Big\{ \mathbb{E}_{\mathbb{P}_t^i(C|W,R)}[\log \mathbb{P}(R^i \mid C^i)] + \mathbb{E}_{\mathbb{P}_t^i(C|W,R)}[\log \mathbb{P}(C^i \mid W^i)] + \mathbb{E}_{\mathbb{P}_t^i(C|W,R)}[\log \mathbb{P}(W^i)] \Big\}$$

The main goal in the M-step is to maximize this expression with respect to the parameters of the distributions involved (i.e., $\mathbb{P}(R \mid C)$, $\mathbb{P}(C \mid W)$, and $\mathbb{P}(W)$). Suppose we want to estimate $\mathbb{P}(R = 1 \mid C = 0)$. The term that interests us in the above expression is $\mathbb{E}_{\mathbb{P}_t^i(C|W,R)}[\log \mathbb{P}(R^i \mid C^i)]$. We then need to differentiate

$$\log \mathbb{P}(R = 1 \mid C = 0) \sum_i \mathbb{P}_t^i(C = 0 \mid W, R) \mathbb{1}[R^i = 1] + \log \mathbb{P}(R = 0 \mid C = 0) \sum_i \mathbb{P}_t^i(C = 0 \mid W, R) \mathbb{1}[R^i = 0]$$

with respect to $\log \mathbb{P}(R = 1 \mid C = 0)$. Then setting to zero, we have:

$$\frac{1}{\mathbb{P}(R = 1 \mid C = 0)} \sum_i \mathbb{P}_t^i(C = 0 \mid W, R) \mathbb{1}[R^i = 1] - \frac{1}{1 - \mathbb{P}(R = 1 \mid C = 0)} \sum_i \mathbb{P}_t^i(C = 0 \mid W, R) \mathbb{1}[R^i = 0] = 0 \tag{15}$$

and we obtain the following estimate :

$$\hat{\mathbb{P}}(R = 1 \mid C = 0) = \frac{\sum_i \mathbb{P}_t^i(C = 0 \mid W, R) \mathbb{1}[R^i = 1]}{\sum_i \mathbb{P}_t^i(C = 0 \mid W, R) \mathbb{1}[R^i = 0] + \sum_i \mathbb{P}_t^i(C = 0 \mid W, R) \mathbb{1}[R^i = 1]}$$

In conclusion, when we have unobserved variables, we replace the frequency count with the expected count of each combination of values for the involved variables.

Remark 4. ***General case***

Consider a Bayesian Network with the following random variables: The observed variables: $X_1, X_2, \ldots, X_n$. The hidden variables: $H_1, H_2, \ldots, H_m$. The structure of the Bayesian network is represented by a directed acyclic graph where nodes correspond to random variables, and edges indicate probabilistic dependencies between these variables. In this network, hidden variables can influence both observed variables and other hidden variables. To estimate the conditional probabilities, we will use the Expectation-Maximization (EM) algorithm. Here are the steps:

***E-step** : In this step, we calculate the conditional distribution of hidden variables H_i given the observed values $X_1, X_2, \ldots, X_n$ based on the current estimates of the network parameters. For each example in the dataset, we compute the conditional distribution*

$$\mathbb{P}_t^i(H_1, H_2, \ldots, H_m \mid X_1^i, X_2^i, \ldots, X_n^i)$$

.

M-step : In this step, we maximize the following expression to estimate the parameters of the conditional distributions involved in the Bayesian network:

$$\sum_{i=1}^{N} \mathbb{E}_{\mathbb{P}_t^i(H_1,H_2,\ldots,H_m|X_1^i,X_2^i,\ldots,X_n^i)}[\log \mathbb{P}(X_1^i, X_2^i, \ldots, X_n^i, H_1, H_2, \ldots, H_m)]$$

where N is the number of examples in the dataset. In this step, we perform partial maximizations for each conditional distribution to update the parameters.

Let's say we want to estimate the conditional probability $\mathbb{P}(X_j \mid H_k = h_k)$ for a given observed variable X_j and a specific value h_k of the hidden variable H_k. To estimate this probability, we need to calculate the normalized sum of expected indicator values (over all examples in the dataset) :

$$\hat{\mathbb{P}}(X_j = j \mid H_k = h_k) = \frac{\sum_i \mathbb{P}_t^i(H_k = h_k \mid X_1^i, X_2^i, \ldots, X_n^i)\mathbb{I}[X_j^i = j]}{\sum_i \sum_{j'} \mathbb{P}_t^i(H_k = h_k \mid X_1^i, X_2^i, \ldots, X_n^i)\mathbb{I}[X_j^i = j']}$$

where j' is the index running over all the possible values taken by the variable X_j.

B.2 Bayesian Estimation

Bayesian estimation takes a probabilistic approach by incorporating prior beliefs about the parameters and updating them based on the observed data. It provides a framework for combining prior knowledge with the data to obtain posterior probability distributions for the parameters.

Parameter estimation in Bayesian networks involves updating our beliefs about the parameters based on both prior knowledge and observed data. This is achieved through Bayes' theorem, which relates the posterior distribution of parameters $P(\theta|D)$ to the likelihood of the data given the parameters $P(D|\theta)$ and the prior distribution of the parameters $P(\theta)$.

$$\mathbb{P}(\theta \mid D) = \frac{\mathbb{P}(D \mid \theta) \cdot \mathbb{P}(\theta)}{\mathbb{P}(D)}$$

Where:

- $\mathbb{P}(\theta \mid D)$ is the posterior distribution of parameters given data.
- $\mathbb{P}(D \mid \theta)$ is the likelihood of the data given parameters.
- $\mathbb{P}(\theta)$ is the prior distribution of parameters.
- $\mathbb{P}(D)$ is the marginal likelihood of the data.

The joint probability of the observed data $x[1], \ldots, x[M]$ and parameters θ can be expressed using the chain rule and Bayes' rule:

$$\mathbb{P}(x[1], \ldots, x[M], \theta) = \mathbb{P}(x[1], \ldots, x[M]|\theta) \cdot \mathbb{P}(\theta)$$

This expression can be further expanded using the chain rule and the conditional independence properties of Bayesian networks:

$$= \mathbb{P}(\theta) \cdot \prod_{m=1}^{M} \mathbb{P}(x[m] \mid \theta) \cdot \prod_{BN} \mathbb{P}(\theta_i \mid \text{parents}(\theta_i))$$

where parents(θ_i) represents the parents of node θ_i in the Bayesian network.

MLE assumes that θ is an unknown but fixed parameter. It estimates θ^*, the value that maximizes the likelihood function. The prediction is then made based on this estimation:

$$\mathbb{P}(D_{m+1} = H \mid D) = \theta^*$$

In contrast, Bayesian estimation treats θ as a random variable. It assumes a prior probability distribution for θ $\mathbb{P}(\theta)$ and uses the observed data to obtain the posterior probability distribution of θ $\mathbb{P}(\theta \mid D)$.

Chief Editor

Monica DEFEND
Head of Amundi Investment Institute

Editors

Marie BRIÈRE
Head of Investors' Intelligence & Academic Partnership

Thierry RONCALLI
Head of Quant Portfolio Strategy

Amundi ASSET MANAGEMENT | Investment Institute

Working Paper
March 2024

Important Information

This document is solely for informational purposes.

This document does not constitute an offer to sell, a solicitation of an offer to buy, or a recommendation of any security or any other product or service. Any securities, products, or services referenced may not be registered for sale with the relevant authority in your jurisdiction and may not be regulated or supervised by any governmental or similar authority in your jurisdiction.

Any information contained in this document may only be used for your internal use, may not be reproduced or redisseminated in any form and may not be used as a basis for or a component of any financial instruments or products or indices.

Furthermore, nothing in this document is intended to provide tax, legal, or investment advice.

Unless otherwise stated, all information contained in this document is from Amundi Asset Management SAS. Diversification does not guarantee a profit or protect against a loss. This document is provided on an "as is" basis and the user of this information assumes the entire risk of any use made of this information. Historical data and analysis should not be taken as an indication or guarantee of any future performance analysis, forecast or prediction. The views expressed regarding market and economic trends are those of the author and not necessarily Amundi Asset Management SAS and are subject to change at any time based on market and other conditions, and there can be no assurance that countries, markets or sectors will perform as expected. These views should not be relied upon as investment advice, a security recommendation, or as an indication of trading for any Amundi product. Investment involves risks, including market, political, liquidity and currency risks.

Furthermore, in no event shall any person involved in the production of this document have any liability for any direct, indirect, special, incidental, punitive, consequential (including, without limitation, lost profits) or any other damages.

Date of first use: 14 March 2024.

Document issued by Amundi Asset Management, "société par actions simplifiée"- SAS with a capital of €1,143,615,555 - Portfolio manager regulated by the AMF under number GP04000036 – Head office: 91-93 boulevard Pasteur – 75015 Paris – France – 437 574 452 RCS Paris – www.amundi.com

Photo credit: iStock by Getty Images - monsitj

Find out more about
Amundi Investment Institute Publications

research-center.amundi.com