Alma Mater Studiorum - Università di Bologna

Department of Statistical Science

Second Cycle Degree in Statistical Sciences

Causal Inference in Multivariate Time Series: a Machine Learning-based prospective

Presented by:

Jacopo Palombarini

0001076600

UNIBO Supervisor:

Prof. Pietro Rigo

ULB Supervisors:

Prof. Gianluca Bontempi

Prof. Gianmarco Paldino

II SESSION

Academic Year 2023/2024

Abstract

An abstract typically under 200 words should summarize the core parts of the thesis. Specifically, the abstract should include the motivation for the study, statement of the research problem, the methodology used and the results and conclusions of the study.

Keywords: keyword1, keyword2, keyword3, keyword4, keyword5

Contents

Introduction

The introduction contains a synthetic description of the thesis, reflecting the outline of the paper. The purpose and the goal of the study should be clearly stated and the motivations behind the study should be given.
- What is the main scope of the thesis
- What does the thesis contain
- Why does the subject interested me.
- What do I expect to achieve with this work.

Causal Inference can be described as the process of evaluate cause-and-effect relationships between variables within a system based on observational data. This analysis is crucial in several fields, improves their explainability and reliability and allows us to determine effective solutions in fields such as healthcare, earth science, politics, business, education and many others ([2]). The information we can obtain thanks to this analysis, for example, the causal factors, are very useful for decision-making of any kind and for predicting the results of potential interventions without actually see them in practice. In fact, while randomized control trials (RCTs) are the gold standard for identifying cause-effect relationships, they are frequently impractical because of high costs and ethical issues. To reach these goals, Causal Inference makes use of Causal discovery algorithms, whose purpose is to uncover the underlying causal structure of a generative processes from a set of observations. These methods rely on specific assumptions and are often represented as a causal graph or a causal adjacency matrix, giving us the possibility to effectively see the relationships between variables and to eventually recognise causes and effects within them ([5]). To illustrate these relationships, causal graphs, often Directed Acyclic Graphs (DAG), use directed arrows, helping us understand the data-generating mechanism more effectively and guiding necessary interventions. Traditional practical applications of predictive systems often overlook causal knowledge, leading to irrational and incomplete decisions when correlations are confused with causation. This can result in significant errors, as highly correlated variables not necessarily influence each other, but could be influenced by hidden factors or latent confounders that have an effect on each of them ([8]).

Some of these Causal Discovery methods are designed for static data (non-temporal), while others focus on time series or temporal data. Given that both data types are prevalent in real-world

scenarios across different problem domains, it is crucial to have methods capable of recovering causal structures from both types. Time series Causal Discovery methods, more specifically for multivariate scenarios, have been less explored so far, and it's in this field that this thesis tries to give its contribute. The main focus is put on the TD2C method, a quite recent (2015-2024) method that relies on distribution asymmetry brought by causal relationships between variables. In Chapter 1 we are going to display the most important theoretical aspects about Causality and Causal Inference, necessary to understand the procedures we will apply in Chapter 2. In this second part, a detailed experimental phase is conducted to explore the potentialities of the TD2C method. We will to investigate its applicability to a wide range of possible real-world scenarios and we will try to validate some of its modified versions. In order to fulfill these tests, we will compare TD2C with the most-known state-of-the-art methods for Causal Discovery on multivariate time series.

More specifications on what we achieved at the end

Chapter 1

Theoretical Background

This first chapter wants to give an overview of all the theoretical concepts behind the methods that are applied in the second part of the thesis. We are going to see what a time series is and what are the main statistical tools to analyze time series from a causal point of view.

1.1 Causality, Causal Inference & Causal Discovery

In this section we are going to define Causality and its derivations in mathematical and statistical fields.

What is causality and why is it interesting to go above the concept of dependence between variables.
What is causal inference and what is it useful for.
What are its main topics and fields of application.
What have been its main contributors during time.
What is causal discovery and what is it useful for.

We can define causality as the consequential relationship between two objects, a cause and an effect. A causal relationship underlines a connection, between two or more events, imposing a directionality, so that one of those events occurs, time-wise, before the others, and it can't be the other way around. J.Pearl - one of the greatest contributors in the causality field - proposed this simple definition: *A causes B if B listens to A* (The Book of Why, Pearl & Mackenzie, 2019).

Causality goes beyond the concept of correlation - *correlation is not causation* - which only requires that two variables are connected and show dependency, but without imposing an order between them, i.e., one variable is not a consequence of the other, but just connected to it.
For the same reasons that discovering, analyzing and predicting the associations between variables is so important, knowing that we can (directly or indirectly) influence some aspects of the world we observe by controlling some of its features, is also quite attractive. Aristotle,

Figure 1.1: Dependency relation (a) and causality relation (b)

one of the most important philosophers of ancient Greece, believed that comprehending the causal structure of a process is a crucial element in understanding this process. ([20]). To cite another important philosopher, David Hume contended that causation relies on experience, which means any perceived cause-and-effect relationship could be erroneous. He maintained that since thoughts are subjective, it is impossible to definitively establish causality. Despite this quite skeptical interpretation of causality, we can still find, in statistical literature, some methods and techniques that aim to find strong enough evidences of consequential connections between events. From 1920 - when Sewall Wright, for the first time, put down mathematically the assumption that X causes Y and not the other way around (1.1) ([16]) - researchers started to develop an interest in what we today call Causal Learning.

Causal Learning encompasses two primary processes, both rely on observational or interventional data. The first, Causal Inference (Ci), aims to quantify the impact of a cause on its effect and relies heavily on a formal representation of the interactions among observed variables, called a causal graph. Despite its simplicity, this graphical representation is very effective for enhancing explainability. When the causal graph is unknown, it is possible to identify cause-effect pairs by integrating available data with prior knowledge. The second process, known as Causal Discovery (CD), aims to derive those causal connections within the system, it's a set of methods meant to recover the structure of the data-generating process from the data generated by that process. Recently, Causal Discovery has seen significant advancements, and this progress has led to the fragmentation of the field into various subfields, each with different assumptions, problems, and solutions, although they share the same ultimate goal ([18]). Causal discovery methods and their application are at the core of this study and some of them are going to be deepened in section 1.4.

1.1.1 Simpson paradox and other examples and images to make causation more clear

Here we display some examples to clarify the essential concept of causality, which may not be easily comprehended by someone who is used to thinking solely in terms of statistical correla-

tion. Let's start with one of the most famous effects of causation: the Simpson's paradox. The paradox refers to the existence of data in which a statistical association that holds for an entire population is reversed in every subpopulation (i.e. considering subpupulations of data marked by a specifica variable's categories). This effect can help us a lot in understanding the logic behind causation because it shows us how a simple association between two variables can be denied and changed by considering a third variable that influences both the other two. Let's take, for example, three variables from a passed investigation about female smokers, their age, their smoker status (positive or negative) and their survival after 20 years from the first data collection (positive or negative). The marginal contingency table is here displayed:

Table 1.1: Marginal table between the variables *Smoker* and *Survival state*. In green, the smallest survival rate between the two.

	Alive	
Smoker	yes	no
yes	76,1%	23,9%
no	68,6%	31,4%

As we notice in Table 1.1, smoking seems to favour the survival of considered subjects. In contrast, if we consider percentages in the partial contingency Table 1.2 we see how this trend is reversed by including the third variable, *Age*, and conditioning on it, i.e. by sorting data using its categories (1 = "Age $\leq$ 24", 2 = "24 $<$ Age $\leq$ 65", 3 = "Age $>$ 65").

Table 1.2: Partial contingency table, both in frequencies and percentages, between the variables *Smoker* and *Survival state*, grouped by variable *Age*. In green, the smallest survival rate per age category.

Percentages

Age	1		2		3	
Smoker / Alive	yes	no	yes	no	yes	no
yes	96.4%	3.6%	80.3%	19.7%	12.2%	87.8%
no	98.6%	1.4%	84.5%	15.4%	13.9%	86.1%

Frequencies

Age	1		2		3	
Smoker / Alive	yes	no	yes	no	yes	no
yes	53	2	384	94	6	43
no	71	1	406	74	25	155

Here, it's clear how smoking has a negative impact on survival rate in every age category. This is

possible because, as we see also in 1.3, the variable *Age* causes both the other two variables, and this changes their relation. This back-and-forth consideration of more variables can continue indefinitely, revealing a persistent challenge. Simple statistics cannot resolve this issue, as no statistical method alone can uncover the true causal relationships within the data. To determine the actual causal connections among the variables, we must first comprehend the underlying story and mechanisms that produced the observed results. Statisticians often interpret data with strong causal assumptions. In such cases, the paradox would disappear because the causal story could align with our example's structure. However, even though the assumption that "smoking does not cause age" might seem obvious, it cannot be tested within the data itself. Moreover, causal information cannot be represented in contingency tables, which are frequently used for statistical inference. ([9])

Other examples

1.1.2 Causality framework exploration

Before proceeding, we would like to clarify, as completely as possible, all the statistical purposes bounded with the concept of causality. These purposes are in constant evolution, but we think it could be useful to portray the current state of this field to better understand the intents and the positioning in the literature of all the following topics.

As previously mentioned, causality learning objectives can be divided into two categories: determining causal effects, referred to as Causal Inference or Parameter Learning, and identifying causal relationships between objects, known as Causal Discovery or Structural Learning. Although these two areas intersect to some extent, their definitions differ slightly. Understanding the causal relationships among variables typically involves determining the correct causal directions, causal lags, and associated causal indices. In contrast, learning about causal effects focuses on infer the impact of altering one variable on the outcome of another, given that some variables are already known to have causal connections.([15])

These two purposes branch out, following various criteria, in different approaches. While studying causality and its application, the amount of methods in the literature could get confusing, that's why we provide, in Figure 1.2, a concise summary of them in a schematic way.

1.2 Fundamental tools

This quite technical section is going to give us some important definitions that revolve around the Causality world. We dedicate this section of the thesis to theoretical concepts that require

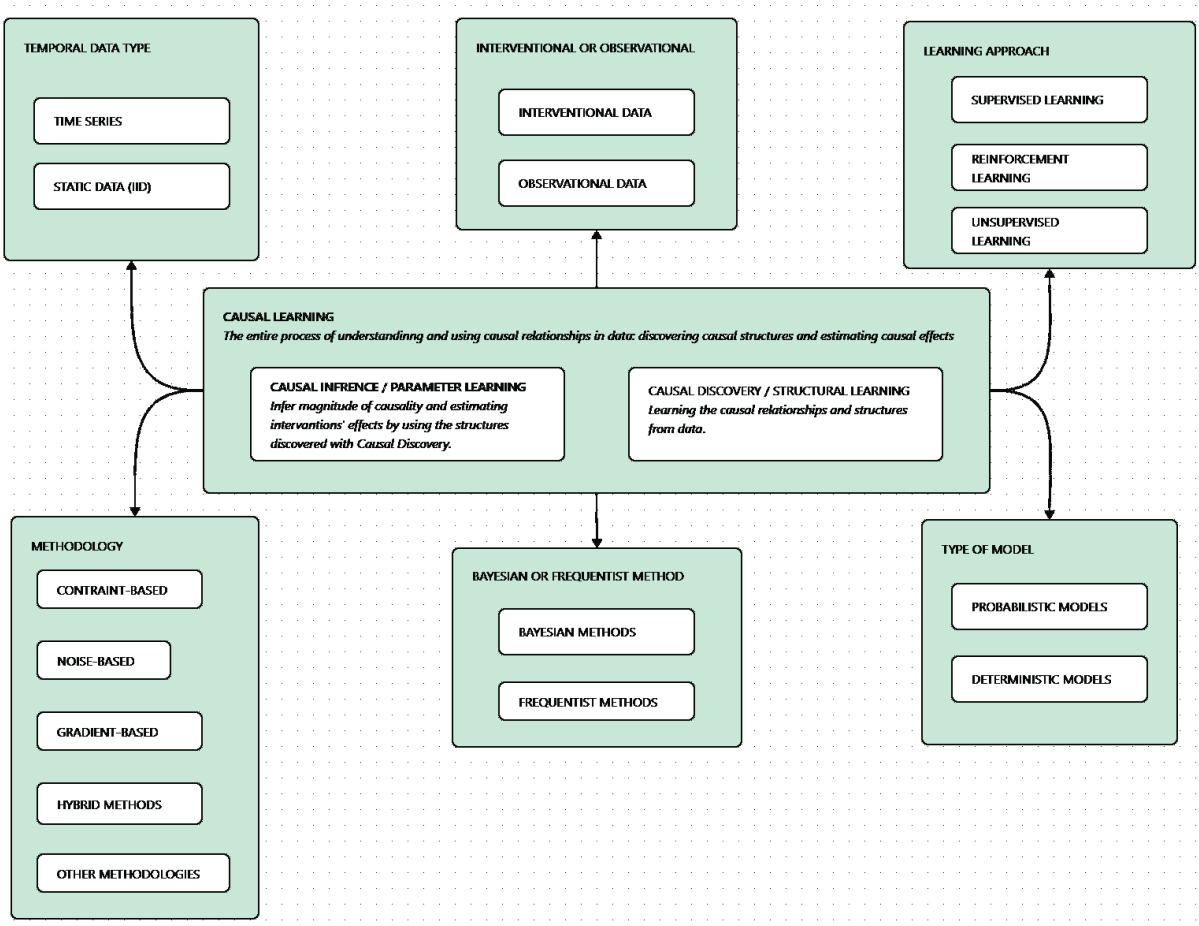

Figure 1.2: Ways of categorizing methods and approaches in Causal Learning context. In blue, we underlined TD2C's characteristic

some specifications and that will be mentioned in the subsequent parts, or that are just useful to have a wider comprehension of the context.

We start with the definition of some fundamental concepts about causal relationships and causal effects.

Definition 1.1 (Conditional Probability)**.** Conditional Probability is the probability of one event, given that another event has occurred. $P(X|Y)$, where "|" stays for "given that".

The complete notation is $P(X = x|Y = y)$: "the probability that the variable X takes the value x, given that the variable Y takes the value y". It can be also adapted to continuous cases, where we refer to probability densities.

Definition 1.2 (Confounding)**.** A confounding variable influences two or more other variables and produces a spurious association between them. From a purely statistical point of view, such associations are indistinguishable from the ones produced by a causal mechanism (1.3).

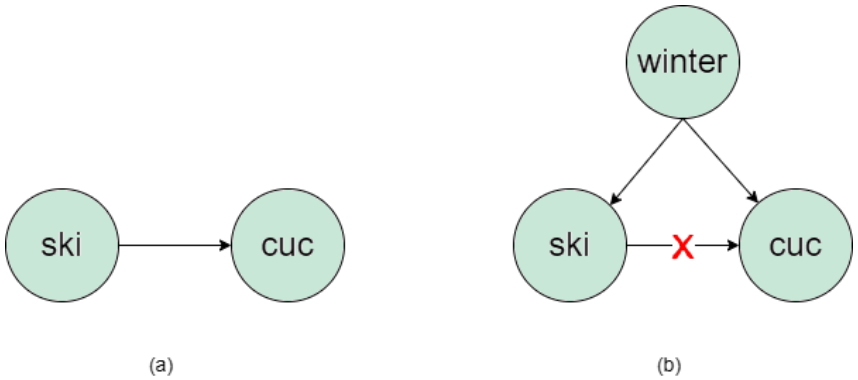

Figure 1.3: Example of confounding: from a first look (a) the increase in people going skiing at a certain time causes the decrease in sales of cucumbers, but, including the *confounder* "winter season" (b), we are able to remove the causality between the previous two variables, since the new one causes them both.

A very important aspect of a properly designed randomized experiment (RCT) is that it allows us to avoid confounding.

Definition 1.3 (Intervention)**.** Intervention is changing one thing in the world and the observing whether and how this change affects another thing in the world.

Interventions are the essence of scientific experiments. To describe interventions mathematically, we use the do-operator 1.5: $P(Y = 1|do(X = 0)$. While conditioning only modifies our view of the data, interventions affects the distribution by actively setting one (or more) variable(s) to a fixed value or distribution. [20]

Definition 1.4 (Counterfactual)**.** Counterfactuals are estimates of how the world would look if we changed the value of one or more variables, holding everything else constant.

Because, we can never observe the same event under two mutually exclusive conditions at the same time, counterfactuals cannot be observed, and so the true causal effect is always unknown. Two different counterfactual causal models can lead to the same interventional distribution.

Interventions and Counterfactuals, togheter with Associations, are the three steps of the so-called *Ladder of Causation* described by Judea Pearl. Each step answers to different causal questions: association is related to observing, using association, we can answer questions about how seeing one thing changes our beliefs about another thing; the action related to step two is doing (1.3); activities associated with step three are imagining and understanding (1.4).

Definition 1.5 (do-operator)**.**

Little digression on do-calcolus (Molak's book)

Then, we need some definitions about DAGs and SCMs.

Definition 1.6 (Graph)**.** A graph $G = (V, E)$ is a mathematical object represented by a tuple of two sets: a finite set of vertices V and a finite set of edges $E \subseteq VV$. If not specified otherwise, this graph is intended as an undirected graph, where the undirected edge (X, Y) is identical to the edge (Y, X) and its graphical representation is $X - Y$.

Definition 1.7 (Directed Graph)**.** A directed graph (DG) is a graph where the edge (X, Y) is distinct from the edge (Y, X).

In particular, a directed edge (X, Y) is graphically represented by an arrow as $X \to Y$, and induces a set of relationships between the vertices of the graph G. Given a vertex X, we denote by $Pa(X)$ its parents, i.e. the set of vertices that have an arrow into X, while we denote by $Ch(X)$ its children, i.e. the set of vertices that have an arrow out of X. Recursively, any parent and parent of a parent (child and child of a child) of X is an ancestor $An(X)$ (descendant $De(X)$) of X. The vertices connected to X are said to be adjacent to X and denoted by $Adj(X)$, while the vertices connected with an undirected edge are the neighbors $Ne(X)$. These two sets of vertices are identical in undirected graphs, but may be different in graphs with other mixed orientations.

Definition 1.8 (Adjacency matrix)**.** Adjacency matrices are square $M \times M$ matrices where M is the number of nodes. Each positive entry in the matrix encodes an edge between a pair of nodes (1.4).

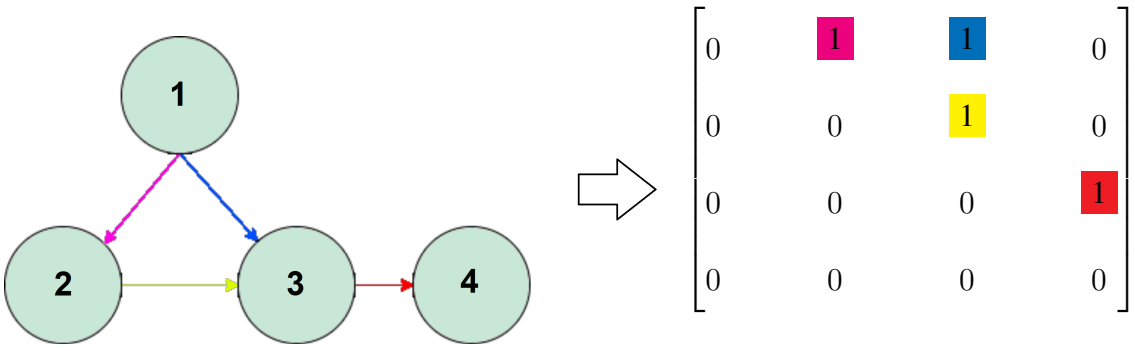

Figure 1.4: A Directed Acyclic Graph (a) and its Adjacency Matrix (b).

Definition 1.9 (Path (Directed path))**.** A path (Directed path) $\pi = (X - ... - Y)$ ($\pi = (X \to ... \to Y)$) is a tuple of non repeating vertices, where each vertex is connected to the next in the sequence with an undirected (directed) edge.

Definition 1.10 (Directd Acyclic Graph)**.** : A directed acyclic graph (DAG) is a directed graph G that has no cycles.

Definition 1.11 (Causal Graph)**.** A causal graph $G = (V, E)$ is a graphical description of a system in terms of cause-effect relationships, i.e. the causal mechanism.

Definition 1.12 (Causal edge assumption)**.** The value assigned to each X is completely determined by the function f given its parents: $X := f(Pa(X))XV$.

Definition 1.13 (Structural causal model)**.** A structural causal model(SCM) is defined by the tuple $M = (V, U, F, P)$, where:

- V is a set of endogenous variables, i.e. observable variables
- U is a set of exogenous variables, i.e. unobservable variables, where $V \cap U = \emptyset$
- F is a set of functions, where each function $f \in F$ is defined as $f_i : (V \cup U)^p \to V$, with p the ariety of f, so that f determines completely the value of V_i
- P is a joint probability distribution over the exogenous variables $P(U) = \prod_i P(U_i)$.

SCMs consist of a set of exogenous (noise variables, root variables) and endogenous (observable variables) variables and a set of functions defining the relationships between these variables. They can be represented as graphs, with nodes representing variables and directed edges representing functions and they can produce interventional and counterfactual distributions. To represent a SCM we can use a set of equations or a graph, as in figure 1.5.

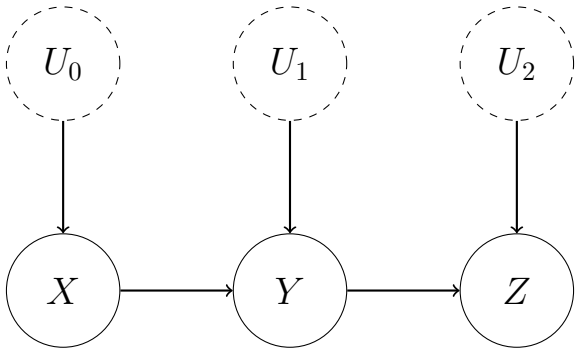

Figure 1.5: An example of Structural Causal Model with exogenous (dashed circles) and endogenous (solid circles) variables

Definition 1.14 (Causal discovery problem)**.** The causal discovery problem consists in recovering the groundtrouth graph G^* (that generated D) from the given dataset D.

Definition 1.15 (Soundness and completeness of a Causal Discovery algorithm)**.** A causal discovery algorithm is sound if it is able to solve the causal discovery problem, and it is complete if it outputs the most informative causal graph G that can be recovered from the input dataset D, without making further assumptions.

Definition 1.16 (Markov property)**.** A graph $G = (V, E)$ is said to satisfy the Markov property if the associated joint probability distribution $P(V)$ can be decomposed recursively as: $P(V) = \prod_{X \in V} P(X|Pa(X))$

The probability factorization expressed in Definition 1.16 relies on the assumption that the relationships encoded by the graph match exactly the underlying conditional probability independencies: $X \perp\!\!\!\perp_P Y|Z \Rightarrow X \perp\!\!\!\perp_G Y|Z$ where Z is a subset of $V/X, Y$. Essentially, it is assumed that the probability independence ($\perp\!\!\!\perp_P$) implies the graphical independence ($\perp\!\!\!\perp_G$).

Definition 1.17 (Forks, Chains and Colliders)**.** Let $G = (V, E)$ be a DG and π be a path on G. Then, given three vertices X, Y and Z in π, we have the following:

- $X \leftarrow Y \rightarrow Z$ is a fork on π
- $X \rightarrow Y \rightarrow Z$ is a chain on π
- $X \rightarrow Y \leftarrow Z$ is a collider on π

Testing for conditional independence (CI) between the variables is one of the most important techniques to find the causal relationships among the variables (it is the core of 1.4.2). Conditional independence between two variables X and Y results when they are independent of each other given a third variable Z (i.e. $X \perp\!\!\!\perp Y|Z$). In the case of causal discovery, CI testing allows deciding if any two variables are causally connected or disconnected. An important criterion for CI testing is the d-separation criterion which is formally defined in 1.18.
Let's see an example: X is conditionally independent of Z given Y i.e. $X \perp\!\!\!\perp Z|Y$ in Figure 1.6 (a) and in Figure 1.6 (b), X and Z are independent, but are not conditionally independent given Y.

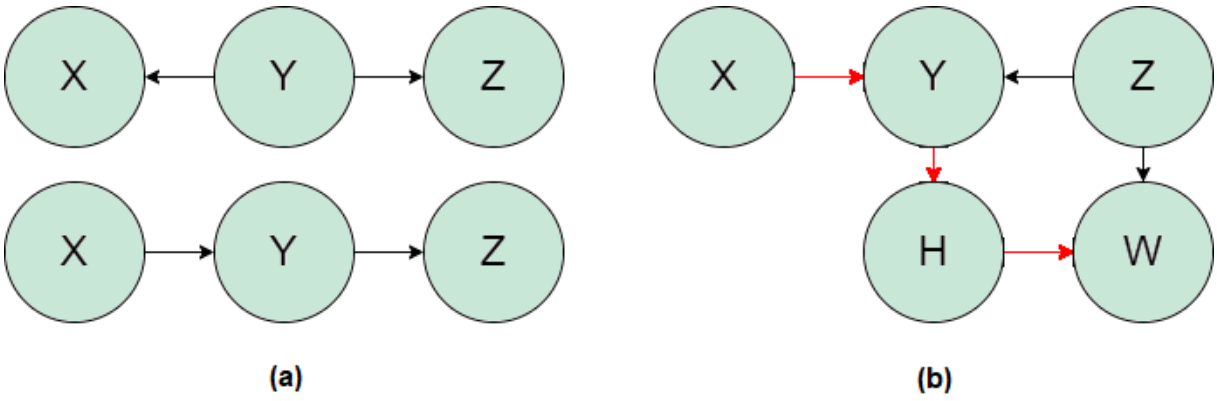

Figure 1.6: Examples of fork and chain structures (a) and of d-separated DAG with a collider between X, Y and Z (b)

Definition 1.18 (d-separation)**.** Let $G = (V, E)$ be a DG, π be a path on G and Z a subset of V. The path π is blocked by Z if and only if π contains:

- a fork $X \leftarrow Y \rightarrow Z$ or a chain $X \rightarrow Y \rightarrow Z$ such that the middle vertex Y is in Z
- a collider $X \rightarrow Y \leftarrow Z$ such that middle vertex Y, or any descendant of it, is not in Z.

The set Z d-separates X from Y if it blocks every path between X and Y (in 1.6 (b) example, we have that X and W are d-separated considering the path that includes Z, but they become connected if we condition on this latter). When we say that a pair of nodes are d-separated, we mean that the variables they represent are definitely independent; when we say that a pair of nodes are d-connected, we mean that they are possibly, or most likely, dependent.

see what to keep here below

In d-separation, the "d" stands for directional. This criterion offers a set of rules to determine if two variables are independent given a specific set of conditioning variables, which can be either a single variable or a collection of variables. Two variables with a directed edge ($\rightarrow$) between them are dependent. The set of testable implications provided by d-separation can be benchmarked with the available data D. If a graph G might have been generated from a dataset D, then d-separation tells us which variables in G must be independent conditional on other variables. If every d-separation condition matches a conditional independence in data, then no further test can refuse the model ([1]). If there is at least one path between two variables that is unblocked, then they are d-connected. If two variables are d-connected, then they are most likely dependent (except intransitive cases) ([1]). The d-separation or conditional independence between the variables in the key structures in 1.17 follow some rules which are discussed below:

- Conditional Independence in Chains: If there is only one unidirectional path between variables X and Z, and Y is any variable or set of variables that intercept that path, then X and Z are conditionally independent given Y , i.e. $X \perp\!\!\!\perp Z|Y$.
- Conditional Independence in Forks: If a variable X is a common cause of variables Y and Z, and there is only one path between Y and Z, then Y and Z are independent conditional on X, i.e. $Y \perp\!\!\!\perp Z|X$.
- Conditional Independence in Colliders: If a variable Z is the collision node between two variables X and Y, and there is only one path between X and Y , then X and Y are unconditionally independent, i.e. $X \perp\!\!\!\perp Y$. But, they become dependent when conditioned on Z or any descendants of Z.

Definition 1.19 (PDAG)**.** The graph G is a partially-directed acyclic graph (PDAG) if it can contain both undirected ($-$) and directed ($\rightarrow$) edges.

Definition 1.20 (Skeleton)**.** Let G be a PDAG. The skeleton of G is the undirected graph resulting from changing any directed edge of G to undirected.

Definition 1.21 (V-structure)**.** Let G be a PDAG. A v-structure in G is a triple $X \to Y \leftarrow Z$ where X and Z are not adjacent. V-structures are also called unshielded colliders.

Definition 1.22 (Observational equivalence)**.** Two DAGs G and H are observationally Markov equivalent if they have the same skeleton and the same V-structures, denoted as $G \equiv H$.

Definition 1.23 (Markov Equivalence Class)**.** Two DAGs G and H belong to the same observational Markov equivalence class (MEC) if they are Markov equivalent. As generalization, the MEC of a graph G, denoted by $[G]$, represents the set of possible DAGs that are observationally equivalent.

Definition 1.24 (Markov Blanket)**.** For any variable X, its Markov blanket (MB) is the set of variables such that X is independent of all other variables given MB. The members in the Markov blanket of any variable will include all of its parents, children, and spouses.

Other useful definitions.

Definition 1.25 (Lagged Causal Effect)**.** Lagged Causal Effects refer to causal relationship between variables at the different time points.

Definition 1.26 (Instantaneous Causal Effect)**.** Instantaneous Causal Effects refer to causal relationship between variables at the same time point.

Definition 1.27 (Expert knowledge)**.** Expert knowledge is a term covering various types of knowledge that can help define or disambiguate causal relations between two or more variables.

Depending on the context, expert knowledge might refer to knowledge from randomized controlled trials, laws of physics, a broad scope of experiences in a given area, and more.

We specify all the assumptions used by the methods in Section 1.4 to restrict their field of application.

Molak e Pearl per approfondimenti e correzioni

Definition 1.28 (Continuous-valued series)**.** All series are assumed to have continuous-valued observations.

However, many interesting data sources - such as social media posts or health states of an individual - are discrete-valued.

Definition 1.29 (Causal Markov assumption)**.** The causal Markov condition states that the node, V i , is independent of all its non-descendants (excluding its parents) given its parents. Therefore, formally, it can be presented as follows:

$$V_i \perp\!\!\!\perp_G V_j | PA(V_i) \ \forall \ j \neq i \in G(V, E) \backslash \{DE(V_i), PA(V_i)\},$$

i.e., the node, V_i , is independent of all other nodes in the graph, G , excluding the descendants and parents of this node, given its parents. If this is not entirely clear to you at this stage, that's perfectly fine.

Definition 1.30 (Linearity)**.** The true data generating process, and correspondingly the causal effects of variables on each other, is assumed to be linear.

The value assigned to each variable X_i is a linear function of the values already assigned to the earlier variables, plus a 'disturbance' (noise) term e_i, and plus an optional constant term c_i, that is

$$X_i = \sum_{k(j)<k(i)} b_{ij} X_j + e_i + c_i$$

In reality, many real-world processes are nonlinear.

Definition 1.31 (Discrete time)**.** The sampling frequency is assumed to be on a discrete, regular grid matching the true causal time lag.

If the data acquisition rate is slower or otherwise irregular, causal effects may not be identifiable. Likewise, the analysis of point processes or other continuous-time processes is precluded.

Definition 1.32 (Known lag)**.** The (linear) dependency on a history of lagged observations is assumed to have a known order. Classically, the order was not estimated and was taken to be uniform across all series.

Definition 1.33 (Stationariety)**.** Some statistical properties of the time series do not change over time.

Definition 1.34 (Complete system)**.** All relevant variables are assumed to be observed and included in the analysis - i.e., there are no unmeasured confounders.

Definition 1.35 (Sufficiency)**.** All relevant variables are measured, and there are no hidden confounders.

Definition 1.36 (Perfectly observed)**.** The variables need to be observed without measurement errors.

Definition 1.37 (Faithfulness)**.** The observed independencies reflect the true causal structure, we write

$$X \perp\!\!\!\perp_P Y|Z \rightarrow X \perp\!\!\!\perp_G Y|Z,$$

i.e., if X and Y are independent in the distribution given Z , they will also be independent in the graph given Z.

It's not very difficult to find situations where the assumption is violated. Intuitively, any situation where one variable influences another through two different paths and these paths cancel each other out completely would lead to the violation of faithfulness.

Definition 1.38 (Non-Gaussianity)**.** The noise terms are non-Gaussian.

While the linear-Gaussian approach usually only leads to a set of possible models, equivalent in their conditional correlation structure, a linear-non-Gaussian setting allows the full causal model to be estimated, with no undetermined parameters.

Definition 1.39 (No Instantaneous effects)**.** Causes do not have immediate effects; there is a time lag between cause and effect.

Definition 1.40 (Time Homogeneity)**.** The causal structure does not change over time.

Definition 1.41 (Additivity)**.** The effect is a sum of the cause and an independent noise term.

Definition 1.42 (Sparsity)**.** The causal graph is sparse, i.e., there are relatively few direct causal connections.

Definition 1.43 (No hidden confounders / Uncorrelated noise variables / Causal sufficience)**.** There are no unobserved variables that affect multiple observed variables; the unobserved noise variables are uncorrelated.

Definition 1.44 (Independence of noise terms)**.** The noise terms in different causal mechanisms are independent,

$$P(e_1, ..., e_n) = \prod_i P_i(e_i)$$

Definition 1.45 (Temporal order / Causal order)**.** The cause precedes the effect in time. No later variable causes any earlier variable.

Definition 1.46 (Acyclicity)**.** The causal graph has no cycles (it is a Directed Acyclic Graph, or DAG).

Definition 1.47 (Model Specification)**.** The form of the causal relationships (e.g., linear, non-linear) is correctly specified.

Definition 1.48 (Kernel assumption)**.** The relationships can be captured in a high-dimensional feature space using kernel functions.

Definition 1.49 (Consistency)**.** The method converges to the correct causal structure as the sample size increases.

Definition 1.50 (Causal minimality)**.** The causal minimality assumption states that DAG G is minimal to distribution, P, if and only if G induces P, but no proper sub-graph of G induces P. In other words, if graph G induces P, removing any edge from G should result in a distribution that is different than P.

Its implications have practical significance for constraint-based causal discovery methods (1.4.2) and their ability to recover correct causal structures.

Definition 1.51. We say that a causal effect (or any other causal quantity) is identifiable when it can be computed unambiguously from a set of (passive) observations summarized by a distribution $P(V)$ and a causal graph G.

In other words, if we have (1) enough information to control for non-causal information flow in the graph and (2) enough data to estimate the effect of interest, the effect is identifiable.

Definition 1.52 (Positivity)**.** Positivity assumptions comprehends two conditions: any estimator needs to have a large enough sample size to return meaningful estimates; the probability of every possible value of treatment in our dataset (possibly conditioned on all important covariates) is greater than 0.

Definition 1.53 (Exchangeability / Ignorability)**.** The treated subjects, had they been untreated, would have experienced the same average outcome as the untreated did (being actually untreated) and vice versa. Formally:

$$\{Y^0, Y^1\} \perp\!\!\!\perp T|Z.$$

In the preceding formula, Y^0 and Y^1 are counterfactual outcomes under $T = 0$ and $T = 1$ respectively, and Z is a vector of control variables.

Examples from all possible sources.

1.3 Causal Inference for Time Series

Before diving into the state-of-the-art section, it's important to underscore the difference between static and time-varying data (one of the possible cause of distinction in causality framework, as mentioned in 1.1.2), from which derive two different types of approaches, with their different assumptions.

What are time series.
Why is their analysis, inference and prediction useful.
In which field it has been applied and could be applied in the future.
What are the differences with static causal inference and what are the main adding difficulties for it.
Specification of the kind of methods we're gonna use.

Let's start with the definition of Time Series Data.

Definition 1.54. Time Series data (TS) is a collection of observations measured over consistent intervals of time. The observation of a Time Series variable X_j at time t is denoted by X_t^j.

Examples of time series data are retail sales, stock prices, climate data, heart rate of patients, brain activity recordings, temperature readings, daily delays of trains for a specific train station, number of ice-cream sold in a month by a supermarket, etc.

Discriminating between associative dependencies and effective causal relationships in high-dimensional and temporal settings, like time series data, presents significant challenges in causal inference from observational data. The multivariate interactions in these contexts generate substantial correlations among most variables.([11]). Considering the potential confounding effects of additional variables is critical to truly discerning the causal relationships within such systems. For example, we can't assess, in a rigorous way, the causal influence of one variable X_i on another X_j without considering the influence of all other variables in the set $X \backslash Xi, Xj$, where X is the set of all variables considered and the operator $\backslash$ is the subtraction operator. This consideration is crucial because any of these variables, not included in the (X_i, X_j) relationship, could confound it, possibly leading to incorrect conclusions. The unique challenges posed by time series data, due to potential latent confounding factors, are not fully addressed by bivariate models. This has necessitated the development of more sophisticated methods tailored to the complexities of time series analysis. The following section will illustrate some of these methods. ([**last paper on TD2C**])

Add something about it from intros in algos papers

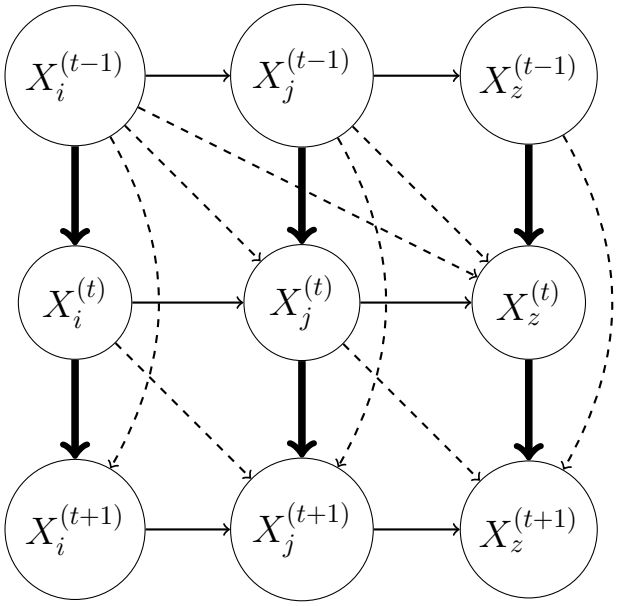

Figure 1.7: An example of three time series variables, X_i, X_j, and X_z, connected through a fork structure. The thick solid lines represent the temporal effects of nodes on their immediate subsequent nodes. The thin solid lines represent the contemporaneous effects between nodes of different variables at the same instant. The dashed lines represent the lagged effects between nodes of different variables (or the same variable) at different (or the same) instants.

1.4 State-of-the-art Methods

In this section we display some of the most valuable strategies for time series causal discovery, their key concepts, their limitations and their applications, in order to create a solid state of the art set of methods we can use to assess the validity of TD2C (2.2).

Introducion of each of them, years of discovery, fields of application, main results achieved, assumption on which they are based and main logic behind them, including some important formulas and/or algos.

1.4.1 Granger-Based methods

[***Maybe to be reduced***]
Let's start our collection of methods with one of the most pivotal concepts in Causal Inference, the Granger Causality.

Based on a statistical version of Hume's regularity theory (Hume, 1738), Granger causality is one of the oldest concepts in causal inference. Hume's theory posits that causal relations can be inferred from the consistent conjunction of causes and effects, with causes preceding their effects. Various authors have investigated probabilistic versions of this theory, which rely on the probability-raising principle (conditioning on a cause increases the probability of the effect occurring). The Granger-based family of methods operates on the premise that if past values of one variable help predict the current value of another, they are causally connected. The simplest implementation, the Pairwise Granger causality test, tests the null hypothesis that X_i does not Granger cause X_j (**cite TD2C paper**). Granger, in 1969, proposed the following definition:

Definition 1.55. A time series X^i Granger-causes X^j if past values of X^i provide unique, statistically significant information about future values of X_j.

We could also express it in this way:
Let H_{t-1} be the history of all relevant information up to time $t-1$ and $P(X_t|H_{t-1})$ be the optimal prediction of X_t given H_{t-1}. Granger defined Y to be causal for X if

$$Var(X_t P(X_t|H_{t-1}) < Var(X_t P(X_t|H_{t-1}/F_t^Y)$$

where, H_{t-1}/F_t^Y indicates excluding the values of F_t^Y (Filtration of Y, i.e., all Y values up to time t-1) from H_{t-1} . That is, the variance of the optimal prediction error of X is reduced by including the history of Y (informally, Y is causal of X if past values of Y improve the prediction of X). This characterization is clearly based on predictability and does not (directly) point to a causal effect of Y on X: Y improving the prediction of X does not mean Y causes

X. Nonetheless, assuming causal effects are ordered in time (i.e., cause before effect), Granger argued that, under some assumptions, if Y can predict X, then there must be a mechanistic (i.e., causal) effect; that is, predictability implies causality. [17]

The concept of Granger Causality has been used in the majority of the exciting causal algorithms today ([12]). Despite its widespread use, ongoing debate has surrounded the validity of the Granger causality framework for inferring causal relationships among time series. Additionally, while the original definition was broad, the constraints of computational tools have limited the application of Granger causality mainly to simple bivariate vector autoregressive processes (VAR).([17]). Let's dig a bit more into this concepts.

Granger causality has traditionally relied on assuming a VAR model and considering tests on its coefficients in the bivariate setting. Denoting the vector of variables at time t by $X_t = (X_t^1, X_t^2, ..., X_t^p)^T$, we consider the linear model

$$A^0 X_t = \sum_k = 1^d A^k X_{tk} + e_t$$

where, $A^0, A^1, ..., A^d$ are $p \times p$ lag matrices (coefficients) and d, the lag, may be finite or infinite. The term e_t, p-dimensional white noise innovation, or error, can have a diagonal or non-diagonal covariance matrix Σ. Granger (1969) highlighted that the model generally lacks identifiability because the matrices A^k are not uniquely defined, except when A^0 is diagonal. This special case, which aligns with the well-known VAR model, was specifically noted by Granger ([3]) — as a "simple causal model," opposed to models with instantaneous causal effects whose A^0 matrix has nonzero off-diagonal values. This latter form is known as a structural vector autoregressive (SVAR) model ([6]) and can be identified under certain parameter restrictions ([10]).

As said, in real-world systems involving multiple time series, examining the relationship between only a pair of series can lead to misleading conclusions due to confounding factors. Specific methods, such as like Network Granger causality, are able address this issue by adjusting for potential confounders and considering multiple series simultaneously. Traditional Granger causal analysis, has also several limitations that restrict its broader application. Specifically, the assumptions required for the (S)VAR model to accurately identify Granger causal relationships include continuous-valued series, linearity, discrete time, known lags, stationarity, perfect observation, and a complete system. Since modern time series data often deviate from these assumptions due to nonlinear dynamics and irregular sampling intervals, recent advancements have made it possible to apply Granger causality to a broader range of scenarios by relaxing these assumptions. [17]

Aside from the most popular GC, there are a few other causality concepts in the field such as

Sims Causality, and Intervention Causality. The first one is often treated as a compliment of Granger Causality, where Granger Causality implies Sims Causality but the inverse is not true. Sims stated that a pairwise Granger Causality for X_t and Y_t can be treated as moving average along several lag terms of the two variables, expressed as

$$Y_t = \alpha_1 Y_{t-1} + \beta_1 X_{t-2} + ... + \alpha_k Y_{t-k} + \beta_k + 1X_{t-2} + C + \epsilon$$

where, k represents the combined noise term from both variables. α and β terms represent the parameter values at each time lag while C is the combined constant term. Sims Causality stated that variable X does not Granger Cause Y if and only if $\beta_1, \beta_2, ..., \beta_k$ is being chosen identically to zero.
Sims' characterizations, which can be shown to be equivalent to Granger's one, can be tested using an F-test comparing two models: the full model, including past values of both x and y, and the reduced model, including only past values of x. We have

$$F = \frac{(RSS_{red} - RSS_{full})/(r-s)}{RSS_{full}/(T-r)}$$

where, RSS_{full} and RSS_{red} are the residual sum of squares for the full and reduced models with r and s parameters, respectively. Using this test, Y is declared Granger causal for X if the observed test statistic F exceeds the (1 α)% quantile of an F-distribution with $r - s$ and $T - r$ degrees of freedom. ([17])

Another causality interpretation, that we call Intervention Causality was proposed by Judea Pearl in 1993 and has been applied to TS data recently. This conception focuses on the idea of counterfactual and calculates the Average Causal Effect (ACE)

$$ACE_S = E(Y_{t^*}) - E(Y_t)$$

where, $E(Y_{t^*})$ represents the resulting outcome for variable Y at time t given the occurrence of intervention S and $E(Y_t)$ represents the expected outcome for variable Y_t without the intervention. While concepts such as Granger Causality and Sims Causality assume an observational framework, Intervention Causality requires counterfactual experiments which are not applicable in many real world applications. Our TD2C algorithm falls into the first group. [15]

1.4.2 The 4, 5, 6, ... Families

In the current literature, there are numerous Causal Discovery methods for both static and time series data. Each method offers unique benefits and faces specific challenges, but the boundaries between them can often be not evident ([19]). Some of these families find application for both

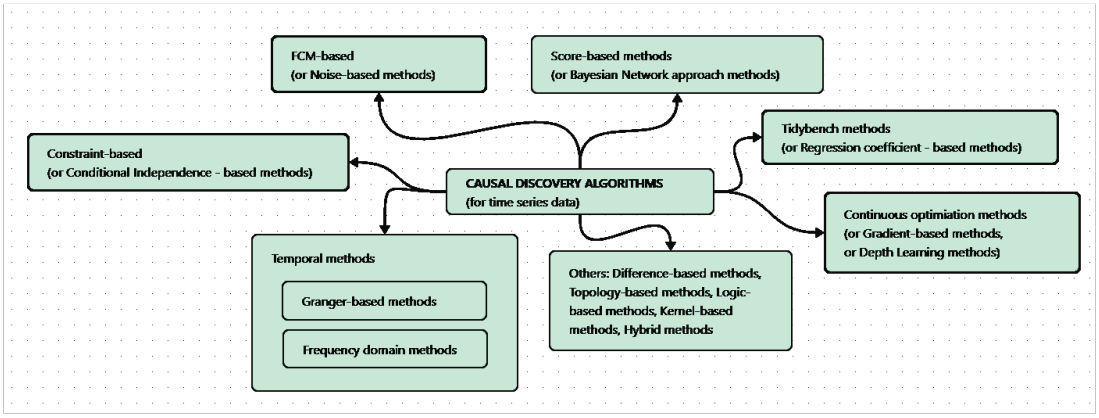

Figure 1.8: Scheme of the most known Causal Discovery methods (in blue, the families we are going to use as benchmarks in the experiments)

types of data, with an obvious more large difficulty for the time-depending ones. In this section we limit our interest in the families whose methods are able to handle successfully this time dependency, and whose application has so far brought at significant results. These methods, as we will see, need to fit in the assumption-wise restricted frame of TD2C method, in order to have a fair and relevant evaluation of this latter.
The families we are going to define are the Constraint-based, the Noise-based, the Gradient-based and the Tidybench Methods. The remaining families include all the hybrid methods and other methods that do not fit our main four categories: Score-based (only valid for IID data, i.e., static data), Kernel-based (from which D2C's logic derives it inspiration), Logic-based, Topology-based, Difference-based, ecc. In Figure 1.8 we summarize schematically what just said.

Contraint-based Methods

Constraint-based methods in causal discovery rely on graph independencies ([20]). A key goal of these methods is testing for conditional independence (CI), which helps recover the causal skeleton when the observed data's probability distribution is faithful to the underlying causal graph.[14] These approaches aim to determine the presence or absence of edges in the data through the d-separation criterion, i.e. by identifying variables that are d-separated or d-connected. By employing CI tests, constraint-based methods recover the Markov equivalence class of DAGs, assuming faithfulness. Despite their speed, these methods are sensitive to the graph structure and prone to error propagation ([19]). Some of the most used CB methods are in table 1.3.
PCMCI is building upon the foundational principles of the PC algorithm. This latter oper-

Table 1.3: Table of the most known Constraint-based methods in the literature

IID Data	TS Data
PC	tsFCI
FCI	PCMCI
ANYTIME FCI	PCMCI+
RFCI	LPCMCI
PC-STABLE	CD-NOD
PKCL	CDANs

ates operates in three steps: first, identifying the skeleton of the graph by starting with a fully connected undirected graph and then eliminating unconditionally and conditionally independent edges; second, determining V-structures or colliders (e.g., $X \rightarrow Y \leftarrow Z$) using the d-separation set of node pairs; and third, orienting the remaining edges to avoid new V-structures and cycle formation. This process results in a CPDAG, which represents the underlying causal DAG.

PCMCI, proposed by Runge et al. (2019), adapts the PC algorithm for time series data, addressing its limitations. It assumes stationarity, causal sufficiency, and the presence of time-lagged dependencies. The algorithm operates in two main stages. In the first stage, it employs PC1, a variant of the skeleton discovery phase of the PC algorithm, to remove irrelevant variables. In the second stage, the algorithm uses the Momentary Conditional Independence (MCI) test to assess whether two variables are independent given their parent sets:

$$X^i_{t-\tau} \perp\!\!\!\perp X^j_t | P_A(X^j_t), P_A(X^i_{t-\tau})$$

[19].
Even when the assumption of stationarity is violated, typically due to obvious confounders, PCMCI still performs more robustly compared to Lasso regression or the PC algorithm. However, PCMCI is less effective for highly predictable systems where little new information is produced at each time step. One significant issue with PCMCI is autocorrelation, which PCMCI+, a more recent version attempts to address. ([14])

Gradient-based Methods

Gradient-based methods stem from research that treats the graph space search as a continuous optimization problem [20]. Recent advancements in causal discovery have redefined the structure learning problem as a continuous optimization task, employing a least squares objective alongside an algebraic characterization of Directed Acyclic Graphs (DAGs). This approach transforms the combinatorial structure learning problem into a continuous one, solvable through

gradient-based optimization techniques. To further expedite the process, deep learning models capable of capturing intricate nonlinear mappings are often employed. Consequently, these models generally exhibit faster training times, as deep learning is known for its high degree of parallelism on GPUs. They update all edges at each step, considering both the gradient of the score and the acyclicity constraint. [19]

Table 1.4: Table of the most known Gradient-based methods in the literature

IID Data	NOTEARS	DAE	GOLEM	MCSL
	GraN-DAG	DAG-GNN	DAG-NoCurl	ENCO
TS Data	DYNOTEARS	NTS-NOTEARS		

Pamfil et al. ([13]), in 2020, introduced the Dynamic NOTEARS (DYNOTEARS) approach, which is designed for structure learning in dynamic data. DYNOTEARS extends the static NOTEARS method to handle both contemporaneous (intra-slice) and time-lagged (inter-slice) dependencies in dynamic Bayesian networks (DBNs). This methodology involves minimizing a penalized loss function while ensuring that the resulting graph structure is acyclic. The acyclicity constraint is enforced using an algebraic condition that characterizes acyclicity in directed graphs. The method assumes that the structure of the network remains constant over time and is the same across all time series. This fixed-structure assumption simplifies the problem and allows DYNOTEARS to scale effectively to high-dimensional datasets. The approach does not impose implicit constraints on the underlying graph. [19]

The model used in DYNOTEARS is a Structural Vector Autoregressive (SVAR) model, where each time series is represented as $X^T_{m,t} = X^T_{m,t}W + X^T_{m,t-1}A_1 + ... + X^T_{m,t-p}A_p + Z^T_{m,t}$ for $t \in p, ..., T$ and for all $m \in 1, ..., M$.. Here, W captures contemporaneous relationships, while $A_1, ..., A_p$ represent time-lagged dependencies. This can be expressed in matrix form as $X = XW + Y_1A_1 + ... + Y_pA_p + Z$, where X is the matrix of observations, Y represents time-lagged versions of X, and Z denotes the error terms. The optimization problem tackled by DYNOTEARS involves minimizing the penalized loss function $f(W, A) = \frac{1}{2n}||X - XW - YA||^2_F + \lambda_A||A||_1$ subject to the constraint that W is acyclic. The acyclicity constraint is given by $h(W) = tr(e^{WoW}) - d = 0$, where o denotes the Hadamard product. To solve this optimization problem, DYNOTEARS employs the augmented Lagrangian method, transforming the constrained problem into a series of unconstrained subproblems that can be addressed with standard solvers.

Noise-based Methods (Functional Methods)

Drawing inspiration from Independent Component Analysis (ICA), a range of algorithms focuses on unraveling the causal relationships among variables by examining their functional dependencies [20]. These algorithms often operate within the framework of Functional Causal Models (FCMs), which articulate the causal connections between variables through specific functional forms. In FCMs, each variable is modeled as a function of its direct causes, accompanied by an independent noise term. This is captured by the equation

$$X = f(PA_X) + E$$

, where f represents the functional dependency on the parent variables PA_X, and E denotes the stochastic noise component. By incorporating additional constraints on data distributions or function classes, FCM-based approaches can differentiate between various Directed Acyclic Graphs (DAGs) that belong to the same equivalence class. This capacity to distinguish among DAGs is essential for refining causal inferences and understanding the underlying mechanisms driving the observed data. [19]

Table 1.5: Table of the most known Noise-based methods in the literature

IID Data	LiNGAM	ANM	PNL	Direct-LiNGAM
	SAM	CGNN	CAM	CAREFL
TS Data	VarLiNGAM	TiMINo		

To uncover the complete causal structure from continuous-valued data, the Linear Non-Gaussian Acyclic Model (LiNGAM) framework is utilized, capitalizing on the independence of non-Gaussian disturbance terms and the linearity of the data-generating process. This approach assumes a recursive causal model, where observed variables are influenced linearly by earlier variables plus non-Gaussian disturbances, without any unobserved confounders (VarLiNGAM also asssumes acyclicity of contemporaneous causal relations). [4]

The core methodology involves Independent Component Analysis (ICA) to identify the underlying causal structure. In a LiNGAM framework, observed data X can be expressed as $X = BX + E$, where B represents the matrix of connection strengths among variables, and E contains the independent, non-Gaussian disturbances. By applying ICA, the model transforms $X = AE$, where A is derived from $(I - B)^{-1}$. ICA's capability to exploit non-Gaussianity allows for a precise estimation of A, overcoming the limitations inherent in Gaussian settings where different mixing matrices can yield identical covariance structures. The ICA-based approach does face challenges like the indeterminacies of component order and scaling, but these can be managed by normalizing components and adjusting the matrix $W = A^{-1}$, involved in

the 5-steps algorithm. [4]

Building on these principles, the VarLiNGAM algorithm extends the LiNGAM approach by incorporating autoregressive components. This method integrates non-Gaussian instantaneous models with vector autoregressive (VAR) models to estimate both immediate and lagged causal effects. The VarLiNGAM algorithm highlights the importance of accounting for instantaneous influences, as neglecting them can significantly skew the interpretation of time-lagged causal relationships. It demonstrates that, even without prior knowledge of network structure, a non-Gaussian model can be effectively identified, enhancing the accuracy and reliability of causal inference in complex temporal systems. [19]

Tidybench Methods (Regression Coefficient-based Methods)

Evalueate if considering these methods
- SLARAC, QRBS, LASAR, SELVAR

Other eventual Methods relevant for the analysis...

1.5 D2C, caD2C & TD2C

Finally, we reached the section dedicated to the protagonist of this study, the D2C method. Here we explain its logic and its evolution in the last ten years.

Fields of application so far
evolution of the method ($D2C -> caD2C -- > TD2C$)
detailed explanation of methods and processes behind
What is its scope.
Why does it enlarge the causal inference field.
What are the theoretical concepts behind it.
How does its base formulation work.
What are the limits of the base formulation.
How does the last two versions improve the base one.
How does the assumptions changed.

1.5.1 Introduction to D2C

D2C's main idea has been introduced in 2015 (Bontempi & Flausder, [7]): it's a data driven approach that apply supervised machine learning to infer the causality connections within a

multivariate ($n > 2$) set of variables. Its main idea relies on the asymmetry generated by two causally related variables that can be noticed in the distribution of their their Markov Blankets' connected variables. It uses some descriptors able to capture this asymmetry and a classifier which learns from artificial data how to discriminate between causal and non causal relations using such descriptors. From its first version (2015), some modifications of the method have been proposed ([11], ***TD2C paper***) to overcome certain limitations that we will mention later. Anyway, the logic behind the algorithm has remained the same. It is part of the probabilistic methods, (i.e. which infer the probability of existence of a causal link between two variables and then its directionality) relying on observational data, which uses a supervised machine learning approach whose inputs are features able to describe variables' dependency. This ability to reduce uncertainity about the causal relationships using statistical features is common to methods such as ANM (Additive Noise Models), IGCI (Information Geometry Causality Inference), LiNGAM and others. These are able to bypass indistinguishability, which limits the functioning of methods relying on conditional independence (CI) tests, such as Constraint-based approaches. D2C has the merit to expand the field of application of these approaches to multivariate data, a much more complicated operation due to growing parameters and causal interactions.
This method is able to derive the relationships between the n variate distribution $X = (X_1, X_2, ..., X_n)$ and the existence of a directed causal link between two variables X_i and X_j, for $1 \leq i \neq j \leq n$, assuming no confounding, no selection bias and no feedback configurations. For the features, it uses structural quantitative characteristics of the data, with the same asymmetry properties, based on information theory concepts (used to quantify the notion of conditional independence), able to derive causality from dependency. Let's see more in detail how the process works in the next section.

1.5.2 Causality as an asymmetric distribution

Here we report the mathematical tools that allow us to recognise that asymmetry we mentioned above and to create features able to describe it.

First, we need to define dependency

Definition 1.56 (Dependency)**.** A variable X_i is dependent on a variable X_j if

$$p(X_i|X_j = x_j) \neq p(X_i)$$

and we write $X_i \not\perp\!\!\!\perp X_j$.

Dependency is symmetric, so, if $X_i \not\perp\!\!\!\perp X_j \Rightarrow X_j \not\perp\!\!\!\perp X_i$.

Then, we need to define mutual information, and we do it in form of probabilistic density functions:

$$I(X_1; X_2) = \int\int log\frac{p(X_1, X_2)}{p(X_1)p)(X_2)}p(X_1, X_2)dX_1dX_2 = H(X_1) - H(X_1|X_2) \tag{1.1}$$

where H is the entropy and we use the convention $0\ log\ \frac{0}{0}$. This formula measures the amount of stochastic dependence between X_1 and X_2.

Dependency becomes, in information terms: $I(X_i; X_j) = I(X_j; X_i) > 0$. It follows that, speaking of conditional mutual information, $I(X_i; X_j|X_z) = 0$ if $p(X_i, X_j|X_z) = p(X_i|X_z)p(X_j|X_z)$, i.e. **Conditional mutual information between two variables X_i and X_j, given another variable X_z, is null if X_i and X_j are conditionally independent given X_z, i.e. $X_i \perp\!\!\!\perp X_j|X_z$** (as well as $I(X_1; X_2) = 0$ if $p(X_1, X_2) = p(X_1)p(X_2)$). The link between this notion and our causality problem is the definition of Markov Blanket:

$$I(X_i; (X \backslash (M_i \cup X_i))|M_i) = C_i \cup E_i \cup S_i = 0,$$

where $C_i = X_c|X_c \rightarrow X_i$ is the set of variables that have direct arrows pointing to X_i (causes), $E_i = X_e|X_i \rightarrow X_e$, are the set of variables to which X_i has a direct causal influence, i.e., where X_i is the parent (effects) and $Si = \{X_s|\ \exists\ X_j : Xs \rightarrow Xj \leftarrow Xi$ and $Xs \neq Xi\}$, are the variables that are not direct causes or effects of X_i but are connected to X_i through a common effect (spouses).

Finally, thanks to the *do-operator*, we can define causality, which, instead, is not symmetric in a probabilistic way: We consider, for example, the causal link $X_i \rightarrow X_j$ and we have that

$$p(X_j|do(X_i = x_i)) \neq p(X_j),\ but,\ convercely,$$

$$p(X_i|do(X_j = x_j)) = p(X_i).$$

Here stands the asymmetry in distribution (Figure **??**).

Another way to define causality is graphically, by using a faithful DAG and the rules of d-separation (Figure 1.9).

As last, we need to define the descriptors that will be used to train the classifier. We surely need asymmetric measures (not as correlation or mutual information, which are dependency measures, i.e., symmetric: $\rho(i, j) = \rho(j, i)$, $I(i, j) = I(j, i)$). To find such descriptors we rely on the elements in the MBs of the two variables (M_i, M_j): S, C and E. Thanks to two

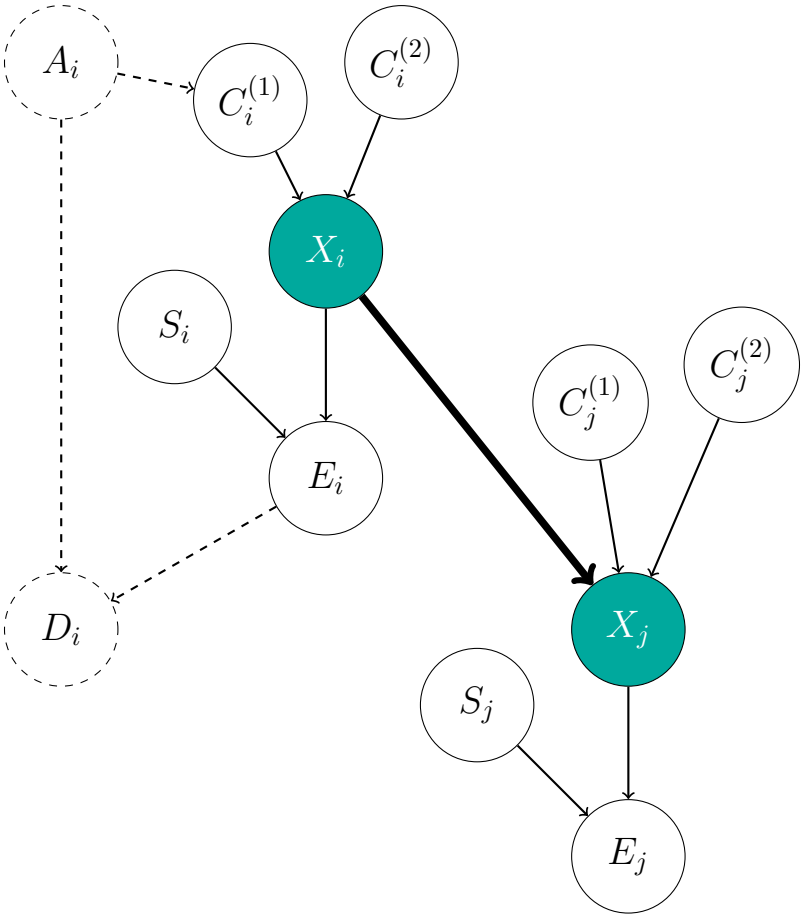

Figure 1.9: Two causally connected variables (X_i and X_j) and their Markov Blankets

assumptions:

1. The only path between the sets $X_i \cup M_i$ and $X_j \cup M_j$ is the edge $X_i \rightarrow X_j$.
2. There are no common ancestors betwween $X_i(X_j)$ and its spouses $S_i(S_j)$,

(DOUBT: how to verify these conditions in real-world data)
we are able to recognise asymmetric conditional independence relationships between M_i and M_j (Table 1.6(a)), which bring to asymmetric mutual information measures (Table 1.7(a)).

The problem is that to know this relationships we need to already know which are the causes, the effects and the spouses in M_i and M_j, which is the same information we are trying to derive.

To solve this problem, we rely on the following: considering the two MB as mixtures of three distributions (for E, S and C) and knowing that there is an asymmetry within there three described by elements in 1.6(a), even without knowing exactly which role each element assumes (E, S or C), we know the two mixtures are different.

Table 1.6: ($\forall k$) All symmetric (b) and asymmetric (a) (un)conditional (in)dependence relationship between M_i and M_j members from 1.9. Relations $S - C$ and $S - S$ are not considered because they include particular relations, not essential for our goal.

(a)

i - j relations	**j - i relations**
$X_i \not\perp\!\!\!\perp C_j^k \mid X_j$	$X_j \perp\!\!\!\perp C_i^k \mid X_i$
$E_i \not\perp\!\!\!\perp C_j^k \mid X_j$	$E_j \perp\!\!\!\perp C_i^k \mid X_i$
$C_i \not\perp\!\!\!\perp C_j^k \mid X_j$	$C_j \perp\!\!\!\perp C_i^k \mid X_i$
$X_i \perp\!\!\!\perp C_j^k$	$X_j \not\perp\!\!\!\perp C_i^k$
$E_i \perp\!\!\!\perp C_j^k$	$E_j \not\perp\!\!\!\perp C_i^k$

(b)

i - j relations	**j - i relations**
$X_i \perp\!\!\!\perp E_j^k \mid X_j$	$X_j \perp\!\!\!\perp E_i^k \mid X_i$
$X_i \perp\!\!\!\perp S_j^k \mid X_j$	$X_j \perp\!\!\!\perp S_i^k \mid X_i$
$E_i \perp\!\!\!\perp E_j^k \mid X_j$	$E_j \perp\!\!\!\perp E_i^k \mid X_i$
$E_i \perp\!\!\!\perp S_j^k \mid X_j$	$E_j \perp\!\!\!\perp S_i^k \mid X_i$
$X_i \not\perp\!\!\!\perp E_j^k$	$X_j \not\perp\!\!\!\perp E_i^k$
$X_i \perp\!\!\!\perp S_j^k$	$X_j \perp\!\!\!\perp S_i^k$
$E_i \not\perp\!\!\!\perp E_j^k$	$E_j \not\perp\!\!\!\perp E_i^k$
$E_i \perp\!\!\!\perp S_j^k$	$E_j \perp\!\!\!\perp S_i^k$
$C_i \perp\!\!\!\perp C_j^k$	$C_j \perp\!\!\!\perp C_i^k$

Table 1.7: ($\forall k$) All symmetric (b) and asymmetric (a) (un)conditional mutual information between M_i and M_j members from 1.9. Relations $S - C$ and $S - S$ are not considered because they include particular relations, not essential for our goal.

(a)

i - j relation	**j - i relation**
$I(X_i; C_j^k \mid X_j) > 0$	$I(X_j; C_i^k \mid X_i) = 0$
$I(E_i; C_j^k \mid X_j) > 0$	$I(E_j; C_i^k \mid X_i) = 0$
$I(X_i; C_j^k \mid X_j) > 0$	$I(C_j; C_i^k \mid X_i) = 0$
$I(C_i; C_j^k) = 0$	$I(X_j \not\perp\!\!\!\perp C_i^k) > 0$
$I(E_i; C_j^k) = 0$	$I(E_j \not\perp\!\!\!\perp C_i^k) > 0$

(b)

i - j relation	**j - i relation**
$I(X_i; E_j^k \mid X_j) = 0$	$I(X_j; E_i^k \mid X_i) = 0$
$I(X_i; S_j^k \mid X_j) = 0$	$I(Z_j; S_i^k \mid X_i) = 0$
$I(E_i; E_j^k \mid X_j) = 0$	$I(E_j; E_i^k \mid X_i) = 0$
$I(E_i; S_j^k \mid X_j) = 0$	$I(E_j; S_i^k \mid X_i) = 0$
$I(X_i; S_j^k) = 0$	$I(X_j \not\perp\!\!\!\perp S_i^k) = 0$
$I(E_i; S_j^k) = 0$	$I(E_j \not\perp\!\!\!\perp S_i^k) = 0$
$I(C_i; C_j^k) = 0$	$I(C_j \not\perp\!\!\!\perp C_i^k) = 0$
$I(E_i; E_j^k) > 0$	$I(E_j \not\perp\!\!\!\perp E_i^k) > 0$
$I(X_i; E_j^k) > 0$	$I(X_j \not\perp\!\!\!\perp E_i^k) > 0$

Observing the elements in 1.7 we notice that, for example,

$$\begin{cases} I(X_i; m_j^{kj} \mid X_j) > I(X_j; m_i^{ki} \mid X_i) = 0 & if \;\; m^{kj} = C_j^{kj} \;\wedge\; m^{ki} = C_i^{ki} \\ I(X_i; m_j^{kj} \mid X_j) = I(X_j; m_i^{ki} \mid X_i) = 0 & else \end{cases}$$

where m^{ki} (m^{kj}) is a member of M_i (M_j) and we are taking the mixtures of the kind $D_1(i,j) = \{I(X_i; m_j^{(kj)}|X_j),\ k_j = 1, ..., K_j\}$ and $D_1(j,i) = \{I(X_j; m_i^{(ki)}|X_i),\ k_i = 1, ..., K_i\}$. So, as said, the populations $D_1(i,j)$ and $D_1(j,i)$ are different, and we can use them (or some of their moments) as our descriptors of causal dependency. In Table 1.8 we put all the asymmetric mixtures D.

Asymmetric mixtures
(a) $D_1(i,j) = \{I(X_i; m_j^{(kj)}|X_j),\ k_j = 1, ..., K_j\}$
(b) $D_1(j,i) = \{I(X_j; m_i^{(ki)}|X_i),\ k_i = 1, ..., K_i\}$
(c) $D_2(i,j) = \{I(m_i^{(ki)}; m^{(kj)_j}|X_j),\ k_i = 1, ..., K_i,\ k_j = 1, ..., K_j\}$
(d) $D_2(j,i) = \{I(m_j^{(kj)}; m^{(ki)_i}|X_i),\ k_i = 1, ..., K_i,\ k_j = 1, ..., K_j\}$
(e) $D_3(i,j) = \{I(X_i; m_j^{(kj)}),\ k_j = 1, ..., K_j\}$
(f) $D_3(j,i) = \{I(X_j; m_i^{(ki)}),\ k_i = 1, ..., K_i\}$

Table 1.8: Asymmetric mixtures from which all the causal descriptors will be generated

1.5.3 The algorithm

The algorithm starts from two sets of features: one is used to infer the MBs of the two considered variables (X_i, X_j) through a filter algorithm (could be mIMR), which also creates a previous relevance ranking within M_i and M_j's elements (we saw that C (causes) elements are much more informative on the direction or causal effects with respect to E (effects) and S (spouses) ones); the other is made by a set of quantiles that summarise the asymmetric mixtures distributions found in the previous section (1.8). The combined dataset, intended for the classifier, is made of vectors ($d = (d_1, ..., d_i, ..., d_p)$) composed by: a set of mutual information terms between X_i and X_j (estimated as difference of entropy terms, as in 1.1, through a Lazy Learning algorithm under Gaussian noises assumption), the positions (in the ranking created previously) P_i^{ki} (P_j^{kj}) of members m_i^{ki} (m_i^{ki}) of $M_i \backslash X_j$ ($M_j \backslash X_i$), the quantiles describing populations in Table 1.8 and a binary response variable, *Class*, that indicates the presence (1) or the absence (0) of the causal link. Synthetic dataset of this kind, with complexity $O(Cn + Cn'^2 + K_i K_j N)$ (***specify meanings***) for each test between two variables, are generated and then used by a Random Forest classifier which, as last step, produces a classification on a testset for validation. In Figure 1.10 you can find a graphical representation of the whole process.

1.5.4 Later updates

As said in 1.5.1, D2C was introduced in [7]. Later on, with [11] and **[TD2C paper]**, certain aspects of the original procedure have been changed to solve some of the problematics (e.g.

computational cost).

At firs, the authors introduced a new descriptor, aiming to featurize the context, based on the notion of interaction information:

$$I(m_i^{(1)}; m_i^{(2)}; X_i) = I(m^{(1)}; m_i^{(2)}; X_i) - I(m_i^{(1)}; m_i^{(2)} | X_i),$$

which gives insights on the causal links characterizing the considered data (in particular, positive interactions derive from common effect configurations, while negative ones form common cause configurations), in order to derive an appropriate DAG structure. The underlying context is maybe more clear in Figure ***inserire figura***, where given the latent nature of the nodes F and G, no d-separation (i.e. no independence) occurs between the nodes B,C,D,E and the node A. This means that the features associated to the nodes B,C,D,E are informative about the state of the node A. In other words, by measuring the quantities represented by nodes B,C,D,E we may reduce the uncertainty about the binary state of the node A.***figure3***. As a result, the ranking procedure described in 1.5.3 is simplified (no more bootstrapping is needed, instead, a correlation to the target measure is used), and complexity for each pair X_i - X_j becomes $O(Cn + K^2N)$. This addition wanted to show that context aware learning approaches, as the ones based on Granger causality (methods based on the precedence concept, which is an associative measure not a causal one and so usually utilized to infer strong relevance and not causality), are adequate to infer causal relationships in large variate contexts, as for multivariate TS. This first modification is in [11] and is called caD2C ("context aware").

The last update (**[TD2C paper]**) of the method proposes some variations to better leverage the temporal dependencies of time series data and reduce the dimensionality of the problem. Starting from the same set of assumptions as before: absence of confounding, selection bias, and feedback configurations, faithfulness, stationarity and lagged effects ($X_i^{(t-1)} \rightarrow X_i^{(t)} \; \forall \; (i,t)$), it proposes the followings:

- Thanks to the lagged effects assumption, it is able to reduce the relevant members of a considered MB, i.e. to reduce dimensionality, and so to skip the MB estimation phase of [7]. Looking at the simplified example in Figure 1.7, we see that, studying the causal link $X_i^{(t)} \rightarrow X_j^{(t+1)}$ and only considering effects and causes for them, we can reduce the asymmetric mixtures seen in Table 1.6 to the ones in Table 1.9
- Adopts the knnCMI method to estimate conditional mutual information using a nearest neighbors approach, differently from [11], where a prediction error from a regression model was informing the entropy estimation under the Gaussian assumption.
- precise which descriptors are used
- precise all the data generating process and so how complexity changes

Table 1.9: Shrinked set of descriptors used by TD2C

CMI terms $\forall\ (m_i^{(k_i)}, m_j^{(k_j)})$	**Families of asymmetric mixtures**
$I(X_i; m_j^{(k_j)} \mid X_j)$	$D_{xmx}(i,j) = I(X_i^{(t)}; m_j^{(k_j)} \mid X_j^{(t+1)})$, $m_j^{(k_j)} \in \{X^{(t+2/t)}\}$
$I(X_j; m_i^{(k_i)} \mid X_i)$	$D_{xmx}(j,i) = I(X_j^{(t+1)}; m_i^{(k_i)} \mid X_i^{(t)})$, $m_i^{(k_i)} \in \{X^{(t-1/t+1)}\}$
$I(m_i^{(k_i)}; m_j^{(k_j)} \mid X_i)$	$D_{mmx}(i,j) = I(X_i^{(t-1/t+1)}; m_j^{(k_j)} \mid X_i^{(t)})$, $m_j^{(k_j)} \in \{X^{(t/t+2)}\}$
$I(m_j^{(k_j)}; m_i^{(k_i)} \mid X_j)$	$D_{mmx}(j,i) = I(X_j^{(t/t+2)}; m_i^{(k_i)} \mid X_j^{(t+1)})$, $m_i^{(k_i)} \in \{X^{(t-1/t+1)}\}$

The final algorithm procedure becomes:

1. Estimate the new reduced version of MB_i and MB_j.
2. Derive the set of descriptors and their empirical distributions' quantiles to create the input vectors (d) for the classifier.
3. Label each input vector as causal (1) or non causal (0), to create the target variable (y).
4. Classifier training based on the final dataset.
5. Prediction on an unseen testset.

Maybe insert the formalized algorithm.

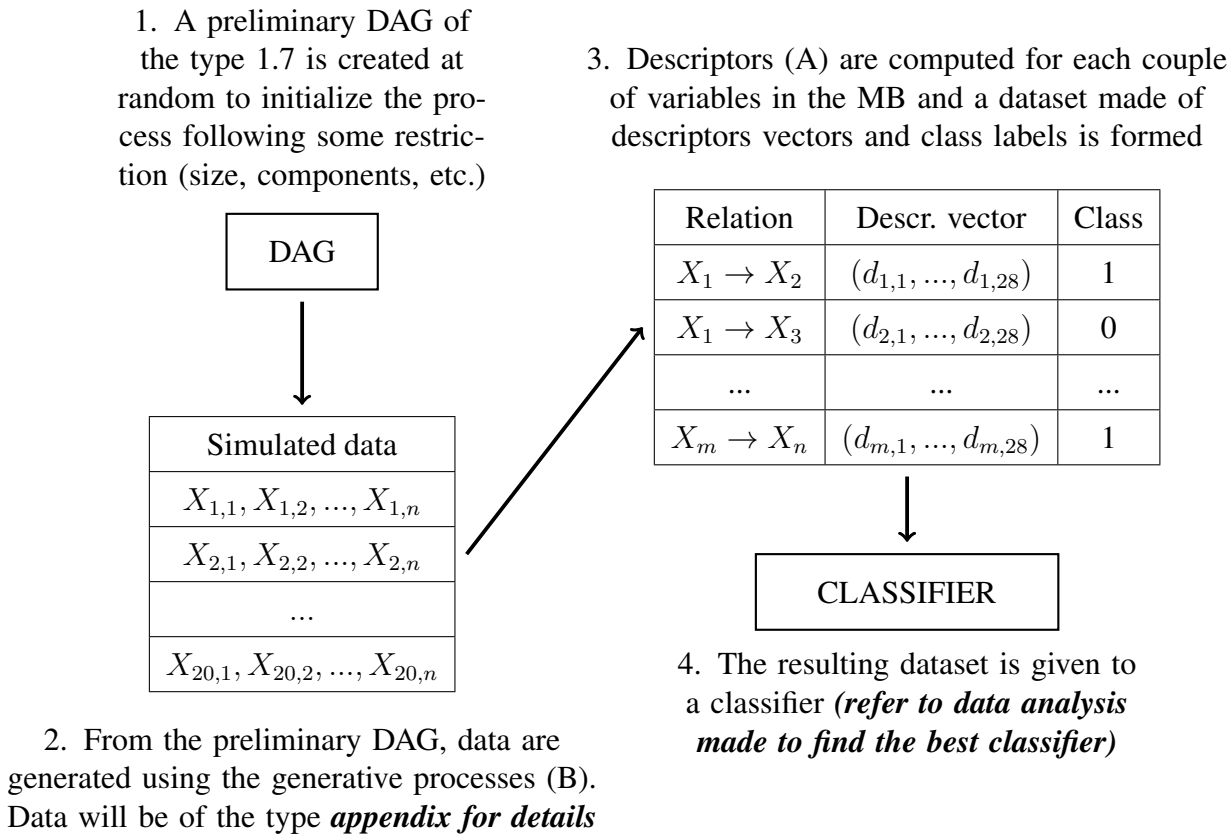

Figure 1.10: D2C process

Chapter 2

Contributions

Introduction to contribution part by describing experimental setting for TD2C (and eventually for past D2C algos), their results, their failure and summarizing limits of TD2C and potential improvements we're going to test for (What are the assumptions and/or the parameters and/or the procedure steps that can be changed/removed/better tested and why?).
Which is the results I wanted to achieve when started the experiments, even if not achieved?
Theoretical explanations of what I've worked on.
Experiments I've done:
- Experimental setting (What are the procedures I followed, What are the settings, in terms of data, software and algorithms i used (python, R, Causeme, ...), What are the evaluation metrics i used and why, What methods i compared to TD2C and/or its modified version, why are they suitable for this scope, Which datasets i used, where did i find them, why are they suitable for my context.)
- Relevant experiments (What are the experiments I conducted, of which kind, only to test already tested scenarios to better confirm them maybe changing the dataset and the procedures and/or testing something new)
- Results (How TD2C or its modified version behave with respect to other sota methods on the base of all metrics, What are the most relevant results to what we're looking for, What proposed modifications gave positive results. What gave negative ones/did't change so much, why, Show some tables/graphics to better explain the results. Other tasks I completed/contribution to the process I gave.).

We reached the second part of the dissertation, where we analyse results from passed studies, we investigate for potential improvements of these latter from a theoretical point of view and we conduct experiments with the scope of verifying such improvements.

2.1 Previous results

Let's take a look at some of the most relevant experiments conducted by similar previous studies, what they put the focus on, their methods and their results.

2.2 Experiments

This section contains a series of tests, conducted with the scope of improving TD2C method from various points of view. We present all the theoretical purposes this study has tried to verify and the way the experiments have been conducted. At last, results form this investigation will be shown and explained.

2.2.1 Problem Solving

This is - from a practical point of view - the most important section of the thesis. Here, we try to give our contribute to the analyzed methods, suggesting solutions to declared problems and limitations and/or simple changes that could improve efficiency, robustness, applicability, reproducibility, precision, validity, ecc. of these methods.

2.2.2 Experimental setting

We specify now the conditions under which the experiments have been conducted. The general setting is similar to the one specified in ***cite TD2C paper***, with some modifications useful to our specific goals. for the experimental phase we relied on Python, Rstudio and Causeme softwares.

Explanation of the phases. could be like this $\rightarrow$ 1st: improvement of TD2C, comparisons with its modified versions, 2nd: validation of improved (or not) version of TD2C, with real valued data, with state of the art methods (or new ones).

Phase 1 settings:
- Data
- Descriptors
- Estimators
- Parameters
- Evaluation metrics

Phase 2 settings:
- Data
- Descriptors

- Estimators
- Parameters
- Evaluation metrics

2.2.3 Results

As final stage, we dispose the most relevant results and their interpretation.

Conclusion

Summarize the most important results of the thesis and try to suggest future research in light of that.

Little recap of what we've seen and done.

What are the most important result obtained.

What could be done in the future and how.

Other noted to be shared.

Appendix A

Descriptors

$$Y_{t+1}^{(j)} = -0.4\left(\frac{3-\bar{Y}_t^{\langle N_j\rangle}}{1+\left(\bar{Y}_t^{\langle N_j\rangle}\right)^2}\right) + 0.6\left(\frac{3-\left(\bar{Y}_{t-1}^{\langle N_j\rangle}-0.5\right)^3}{1+\left(\bar{Y}_{t-1}^{\langle N_j\rangle}-0.5\right)^4}\right) + W_{t+1}^{(j)} \tag{A.1}$$

$$\begin{aligned} Y_{t+1}^{(j)} = &\left(0.4-2\cos\left(40\bar{Y}_{t-2}^{\langle N_j\rangle}\right)\exp\left(-30\left(\bar{Y}_{t-2}^{\langle N_j\rangle}\right)^2\right)\right)\bar{Y}_{t-2}^{\langle N_j\rangle} \\ &+\left(0.5-0.5\exp\left(-50\left(\bar{Y}_{t-1}^{\langle N_j\rangle}\right)^2\right)\right)\bar{Y}_{t-1}^{\langle N_j\rangle} + W_{t+1}^{(j)} \end{aligned} \tag{A.2}$$

$$Y_{t+1}^{(j)} = 1.5\sin\left(\frac{\pi}{2}\bar{Y}_{t-1}^{\langle N_j\rangle}\right) - \sin\left(\frac{\pi}{2}\bar{Y}_{t-2}^{\langle N_j\rangle}\right) + W_{t+1}^{(j)} \tag{A.3}$$

$$\begin{aligned} Y_{t+1}^{(j)} = &2\exp\left(-0.1\left(\bar{Y}_t^{\langle N_j\rangle}\right)^2\right)\bar{Y}_t^{\langle N_j\rangle} \\ &-\exp\left(-0.1\left(\bar{Y}_{t-1}^{\langle N_j\rangle}\right)^2\right)\bar{Y}_{t-1}^{\langle N_j\rangle} + W_{t+1}^{(j)} \end{aligned} \tag{A.4}$$

$$\begin{aligned} Y_{t+1}^{(j)} = &-2\bar{Y}_t^{\langle N_j\rangle} I\left(\bar{Y}_t^{\langle N_j\rangle} < 0\right) \\ &+0.4\bar{Y}_t^{\langle N_j\rangle} I\left(\bar{Y}_t^{\langle N_j\rangle} \geq 0\right) + W_{t+1}^{(j)} \end{aligned} \tag{A.5}$$

$$\begin{aligned} Y_{t+1}^{(j)} = &0.8\log\left(1+3\left(\bar{Y}_t^{\langle N_j\rangle}\right)^2\right) \\ &-0.6\log\left(1+3\left(\bar{Y}_{t-2}^{\langle N_j\rangle}\right)^2\right) + W_{t+1}^{(j)} \end{aligned} \tag{A.6}$$

$$\begin{aligned} Y_{t+1}^{(j)} = &\left(0.4-2\cos\left(40\bar{Y}_{t-2}^{\langle N_j\rangle}\right)\exp\left(-30\left(\bar{Y}_{t-2}^{\langle N_j\rangle}\right)^2\right)\right)\bar{Y}_{t-2}^{\langle N_j\rangle} \\ &+\left(0.5-0.5\exp\left(-50\left(\bar{Y}_{t-1}^{\langle N_j\rangle}\right)^2\right)\right)\bar{Y}_{t-1}^{\langle N_j\rangle} + W_{t+1}^{(j)} \end{aligned} \tag{A.7}$$

$$\begin{aligned} Y_{t+1}^{(j)} = &\left(0.5-1.1\exp\left(-50\left(\bar{Y}_t^{\langle N_j\rangle}\right)^2\right)\right)\bar{Y}_t^{\langle N_j\rangle} \\ &+\left(0.3-0.5\exp\left(-50\left(\bar{Y}_{t-2}^{\langle N_j\rangle}\right)^2\right)\right)\bar{Y}_{t-2}^{\langle N_j\rangle} + W_{t+1}^{(j)} \end{aligned} \tag{A.8}$$

$$\begin{aligned} Y_{t+1}^{(j)} = &0.3\bar{Y}_t^{\langle N_j\rangle} + 0.6\bar{Y}_{t-1}^{\langle N_j\rangle} \\ &+\left(0.1-0.9\bar{Y}_t^{\langle N_j\rangle} + 0.8\bar{Y}_{t-1}^{\langle N_j\rangle}\right)\left(1+\exp\left(-10\bar{Y}_t^{\langle N_j\rangle}\right)\right) + W_{t+1}^{(j)} \end{aligned} \tag{A.9}$$

$$Y_{t+1}^{(j)} = \text{sign}\left(\bar{Y}_t^{\langle N_j\rangle}\right) + W_{t+1}^{(j)} \tag{A.10}$$

$$Y_{t+1}^{(j)} = 0.8\bar{Y}_t^{\langle N_j\rangle} - \frac{0.8\bar{Y}_t^{\langle N_j\rangle}}{1+\exp\left(-10\bar{Y}_t^{\langle N_j\rangle}\right)} + W_{t+1}^{(j)} \tag{A.11}$$

$$\begin{aligned} Y_{t+1}^{(j)} = &0.3\bar{Y}_t^{\langle N_j\rangle} + 0.6\bar{Y}_{t-1}^{\langle N_j\rangle} \\ &+\left(0.1-0.9\bar{Y}_t^{\langle N_j\rangle} + 0.8\bar{Y}_{t-1}^{\langle N_j\rangle}\right)\left(1+\exp\left(-10\bar{Y}_t^{\langle N_j\rangle}\right)\right) + W_{t+1}^{(j)} \end{aligned} \tag{A.12}$$

$$Y_{t+1}^{(j)} = 0.38\bar{Y}_t^{\langle N_j\rangle}\left(1-\bar{Y}_{t-1}^{\langle N_j\rangle}\right) + W_{t+1}^{(j)} \tag{A.13}$$

$$Y_{t+1}^{(j)} = \begin{cases} -0.5\bar{Y}_t^{\langle N_j\rangle} & \text{if } \bar{Y}_t^{\langle N_j\rangle} < 1 \\ 0.4\bar{Y}_t^{\langle N_j\rangle} & \text{otherwise} \end{cases} \tag{A.14}$$

$$Y_{t+1}^{(j)} = \begin{cases} 0.9\bar{Y}_t^{\langle N_j\rangle} + W_{t+1}^{(j)} & \text{if } \bar{Y}_t^{\langle N_j\rangle} < 1 \\ -0.3\bar{Y}_t^{\langle N_j\rangle} + W_{t+1}^{(j)} & \text{otherwise} \end{cases} \tag{A.15}$$

Table A.1: Cross-sectional and temporal series from [3]: Nj denotes the indices of the set of timeseries which are neighbors of the jth component. stands for the average of the value of the neighboring series at time t. Generative processes (27) and (39) have been removed because of their instability with the adopted initial conditions.

Appendix B

Generative Processes

Acknowledgements

Even if it is not required, it provides an opportunity to appreciate the individuals and organizations who supported you and your work. Usually it is located between the sections Discussion and References.

example

The author acknowledges the support of the project WALINNOV 2017 – N 1710030 - CAUSEL funded by the Walloon Region of Belgium and the project "MACHU-PICCHU: Machine Learning for Predictive and Causal modelling of Churn" funded by INNOVIRIS, Brussels (B).

example2

I owe my deepest gratitude and appreciation to my supervisor Dr.John Zelek. Dr. Zelek has always been supportive, approachable, and helpful throughout my master study. His encouragement and understanding helped me go through the difficulties and created the space for me to develop research ideas. I thank my thesis reviewer Dr. Ning Jiang and Dr. PanZhao for agreeing to be on my review committee on a short notice. I would also like to thank postdoctoral fellows Dr. Mohamed Naiel and Dr. Georges Younes for their support and advice. It is a pleasure when to work with them, and I thank them for all their mentorship. I would like to thank the Ontario Centres of Excellence (OCE), Natural Sciences and Engineering Research Council (NSERC) and ATS Automation Tooling Systems Inc., for supporting this research work.

Bibliography

[1] Judea Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan kaufmann, 1988.

[2] Mark Peyrot. "Causal analysis: Theory and application". In: *Journal of Pediatric Psychology* 21.1 (1996), pp. 3–24.

[3] Helmut Lütkepohl. *New introduction to multiple time series analysis*. Springer Science & Business Media, 2005.

[4] Shohei Shimizu et al. "A linear non-Gaussian acyclic model for causal discovery." In: *Journal of Machine Learning Research* 7.10 (2006).

[5] Judea Pearl. *Causality*. Cambridge university press, 2009.

[6] Lutz Kilian. "Structural vector autoregressions". In: *Handbook of research methods and applications in empirical macroeconomics*. Edward Elgar Publishing, 2013, pp. 515–554.

[7] Gianluca Bontempi and Maxime Flauder. "From dependency to causality: a machine learning approach." In: *J. Mach. Learn. Res.* 16.1 (2015), pp. 2437–2457.

[8] Tshilidzi Marwala. *Causality, correlation and artificial intelligence for rational decision making*. World Scientific, 2015.

[9] Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. *Causal inference in statistics: A primer*. John Wiley & Sons, 2016.

[10] Lutz Kilian and Helmut Lütkepohl. *Structural vector autoregressive analysis*. Cambridge University Press, 2017.

[11] Gianluca Bontempi. "Learning causal dependencies in large-variate time series". In: *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE. 2020, pp. 1–7.

[12] Ruocheng Guo et al. "A survey of learning causality with data: Problems and methods". In: *ACM Computing Surveys (CSUR)* 53.4 (2020), pp. 1–37.

[13] Roxana Pamfil et al. "Dynotears: Structure learning from time-series data". In: *International Conference on Artificial Intelligence and Statistics*. Pmlr. 2020, pp. 1595–1605.

[14] Jakob Runge. “Discovering contemporaneous and lagged causal relations in autocorrelated nonlinear time series datasets”. In: *Conference on Uncertainty in Artificial Intelligence*. Pmlr. 2020, pp. 1388–1397.

[15] Bo Yuan Chang. “Multivariate Time Series Data Causal Discovery”. MA thesis. University of Waterloo, 2021.

[16] Judea Pearl. “Causal Inference: history, perspectives, adventures, and unification (an interview with Judea Pearl)”. In: *Observational Studies* 8.2 (2022), pp. 23–36.

[17] Ali Shojaie and Emily B Fox. “Granger causality: A review and recent advances”. In: *Annual Review of Statistics and Its Application* 9.1 (2022), pp. 289–319.

[18] Alessio Zanga, Elif Ozkirimli, and Fabio Stella. “A survey on causal discovery: theory and practice”. In: *International Journal of Approximate Reasoning* 151 (2022), pp. 101–129.

[19] Uzma Hasan, Emam Hossain, and Md Osman Gani. “A survey on causal discovery methods for iid and time series data”. In: *arXiv preprint arXiv:2303.15027* (2023).

[20] Aleksander Molak. *Causal Inference and Discovery in Python: Unlock the secrets of modern causal machine learning with DoWhy, EconML, PyTorch and more*. Packt Publishing Ltd, 2023.

In this section you should provide: authors, year, title, name of the journal or editorial information of the book, issue and pages of the journal.