Journal of Dynamics and Games
©American Institute of Mathematical Sciences
Volume **8**, Number **3**, July **2021**
doi:10.3934/jdg.2021008
pp. **203–231**

CAUSAL DISCOVERY IN MACHINE LEARNING: THEORIES AND APPLICATIONS

Ana Rita Nogueira*

LIAAD - INESC TEC
Rua Dr. Roberto Frias
Porto, 4200 - 465, Portugal
Faculdade de Ciências da Universidade do Porto
Rua do Campo Alegre 1021/1055
Porto, 4169-007, Portugal

João Gama and Carlos Abreu Ferreira

LIAAD - INESC TEC
Rua Dr. Roberto Frias
Porto, 4200 - 465, Portugal

(Communicated by Carlos Ramos)

Abstract. Determining the cause of a particular event has been a case of study for several researchers over the years. Finding out why an event happens (its cause) means that, for example, if we remove the cause from the equation, we can stop the effect from happening or if we replicate it, we can create the subsequent effect. Causality can be seen as a mean of predicting the future, based on information about past events, and with that, prevent or alter future outcomes. This temporal notion of past and future is often one of the critical points in discovering the causes of a given event. The purpose of this survey is to present a cross-sectional view of causal discovery domain, with an emphasis in the machine learning/data mining area.

1. **Introduction.** The search for causal relationships between events has been a case of study for several researchers through the centuries. From its beginning in philosophy, going through physics and celestial mechanics, mankind has always been interested in understanding and explaining their surroundings. More recently, the definition of causality went from a purely philosophical term to a concept in statistics, machine learning and data mining.

Regarding these two lasts fields (machine learning and data mining), we have the definition of causal discovery, as the study of the possible cause-and-effect relationships in data. With that in mind, we can say that the focal point in the investigation of the causal relationships is in their observation, meaning that, to discover potential causal relationships, it is necessary to observe them first. Ideally,

2020 *Mathematics Subject Classification.* Primary: 58F15, 58F17; Secondary: 53C35.
Key words and phrases. Causality, probabilistic methods, granger causality, graphical models, bayesian networks.
The first author is supported by *Fundação para a Ciência e Tecnologia (FCT)* (Portugal) PhD grant SFRH/BD/146197/2019.
* Corresponding author: Ana Rita Nogueira.

these observations are performed in a controlled environment, and through exhaustive testing, so that we can isolate the desired behaviours (these type of experiments are called randomised controlled experiments or RCE). Unfortunately, this is not always possible, either because it is impossible to follow a particular action during the necessary time for it to happen, or because it is not ethical or even prohibited. In these cases, we have to deal with the available information and draw conclusions from it. In such cases, authors advocate using observational data over RCE data [93], since it is a less expensive method for collecting data.

These causal relationships can be found through several methods, with the most commonly used algorithms being the Bayesian networks [69]. However, there are exceptions: recently, several authors adapted well-known machine learning methods, such as Decision Trees, (among others), into causal discovery methods (some of these methods will be presented in more detail in the next sections).

The application of causality to the data mining and machine learning domain is not as trivial as it may seem since it is necessary to distinguish between cause and correlation. This distinction is so important that there is even a very famous sentence in statistics and that it is assumed to be an absolute truth. This sentence is: **"correlation is not causation"**. Correlation is not the same as causation because, although there might be a causal relationship when there is a strong correlation between events, the fact that two events happen sequentially and always together does not mean that they have a cause-effect relationship. Mere correlation does not give us enough information about the occurrence of the events. There are several reasons why these correlations are similar to causality: omitted data, links that go against established rules are some of them. Nevertheless, the fact that there is a correlation between two events may give clues about the true relationship between these events. The opposite idea (where there is causality exists correlation) is not necessarily correct either since there are cases in which, although there is a clear causal relationship between two events, there is no evidence that there is a correlation. This is the case of the Simpson Paradox [11]. This paradox is a phenomenon in which the relation cause-effect can disappear or be inverted depending on whether the data is studied as a whole or divided (for example, separate the data by gender and study it separately). There are essentially two ways to deal with this paradox: proving that the causal relationship is wrong or by denying the premise that the standard probability calculus governs this relationship.

Related with this topic, we have the Causal Decision Theory [98]. This theory uses the expected utility of every option available in a given decision problem so that it is possible to recommend an option with the maximum utility. This theory also interconnects with game theory. It supports the game theory in the interactive solutions made, by identifying rational options in those problems, thus serving as distinguisher to those methods [92].

The areas of application for causal discovery are immense, from its use in climate research to business and biomedical, among many other areas. For example, in the medical field, this type of causal analysis is quite relevant in the diagnosis of certain diseases. If a patient has a set of symptoms (for example, a sudden increase of creatinine[1]), and we can prove that this specific combination of symptoms is caused by a disease B and only by this disease we can infer that the patient has the disease B (this increase in the creatinine could mean kidney problems).

[1] "a crystalline end product of creatine metabolism, $C_4H_7\ N_3O$, occurring in urine, muscle, and blood."http://www.dictionary.com/browse/creatinine

1.1. **Previous Reviews and Contributions.** Over the years several survey studies in causal discovery have been published. In Table 1, a summary of some of these studies is presented.

TABLE 1. Survey studies overview

Survey Title	Reference	Causal Bayesian Networks				Non-bayesian methods	Causal discovery over Time	Causal discovery in statistics	Tools/Frameworks for causal discovery	Evaluation Metrics	Possible Applications
		Reference	Assumptions	Constraint-Based BN	Score-Bases BN						
Review of Causal Discovery Methods Based on Graphical Models	[30]	✓	✓	✓	✓		✓	✓	✓		
A Review on Algorithms for Constraint-based Causal Discovery	[104]	✓	✓	✓					✓	✓	
A review of causal inference for biomedical informatics	[50]	✓	✓				✓	✓			✓
Causal discovery and inference: concepts and recent methodological advances	[89]	✓	✓				✓				
A Survey of Learning Causality with Data: Problems and Methods	[34]	✓	✓	✓	✓	✓	✓				
Causality and Statistical Learning	[27]	✓						✓			
Machine learning for causal inference in Biostatistics	[79]	✓					✓				
Causal Interpretability for Machine Learning - Problems, Methods and Evaluation	[65]	✓							✓*		

*metrics to measure how explainable an algorithm is

One of the most recent overviews about the topic is the survey presented by Glymour *et al.* [30]. Here, the authors focus on the discussion of the most known causal discovery algorithms (PC, FCI, GES, among others) and their usage in biological data.

Another survey related to the subject is the work of Yun *et al.* [104], in which the authors discuss in detail several constraint-based causal discovery algorithms (both global and local discovery). Besides this, the authors also present several metrics to evaluate graphical causal models (as it is possible to see in Table 1 this survey is the only one that provides evaluation metrics).

Finally, in the work of Moraffah *et al.* [65], a slightly different perspective of causal discovery is presented. In this paper, the authors discuss the relationship between causal discovery and the new topic of explainability and eXplainable Artificial Intelligence (XAI) [22]. Although being a relevant topic, the subject discussed in this paper is outside this survey's spectrum.

This survey's purpose is to present a cross-sectional view of the causal discovery domain, with an emphasis in the machine learning/data mining area. This survey emphasises Bayesian and statistical methods and approaches and new emerging techniques, like causal association rule mining and causal algorithm that deal with time-series data. Furthermore, this survey compiles several new evaluations metrics proposed in the last few years.

1.2. **Organisation.** This survey is organised as follows: Section 2 presents an overview of the evolution of the definition of causality; Section 3 presents Pearl's perspective over causal discovery; Section 4 presents the Bayesian networks, while Section 5 presents the causal Bayesian networks; Section 6 presents other algorithms for causal discovery; Section 7 discusses causal discovery over sequential data; Section 8 discusses other possible applications of causal discovery; Section 9 presents existing frameworks for the application of causal related tasks, Section 10 presents

evaluation metrics used to access the performance of causal discovery algorithms, and finally Section 11 presents possible applications. These applications are:

- Medicine;
- Economics (and stock markets);
- Climatology;
- Causal discovery over Sequential Data.

2. **The evolution of causality.** The idea of causality it is not new, dating back to the Ancient Greece, where Plato defined it as being: *"everything that becomes or changes must be so owing to some cause"* [55]. This definition became the basis on which many other philosophers would support themselves to create their ideas of causality.

Although Plato was the one who first proposed a proper definition of causality, it was Aristotle who studied this idea in more depth, by proposing that it should be distinguished in four different forms: material cause, formal cause, efficient cause and final cause [6], [25].

In the middle ages and with the rising of new ideas, the definition of causality was also altered. For example, Aristotle's ideas of causality were partially rejected, keeping only two of the four forms (efficient and final cause), which were entirely reformulated [39].

Finally, in more recent times, several philosophers proposed a more empirical view of the definition of causality. Despite agreeing with the "empiricalisation" of this definition, there was no consensus on the full definition: Hume [94] defended that the idea of causal necessity was obtained by observing the conjunction of certain events and that in the human mind this was associated with causal necessity between events. Locke [33] and Newton [41] defended that causality does not involve a necessary connection.

In more recent years, the definition of causality shifted from a purely philosophical concept to a more statistical one. Rubin [80] proposed a model for causal inference for randomised and non-randomised studies. In a randomised study, the causal effect of a treatment (cause) in a unit (object of study) is measured thought the difference between the post-exposure response variable if the treatment (t) is applied and the response variable if the control (c) is applied (Eq. (1)).

$$Y_t(u) - Y_c(u) \tag{1}$$

In non-randomised studies, it is not possible to measure the post-exposure-response of both treatment and control. To overcome this, Rubin proposed that the causal effect of a treatment should be measured as the averaged causal effect (T) of the measured controls and treatments in the set of all units:

$$E(Y_t - Y_c) = T \tag{2}$$

At the same time, Clive Granger created (in 1969) a statistical hypothesis test (Granger causality test) [32]. In this test, Granger proposes a probabilistic model for the discovery of causal relationships between time-series. This discovery analyses past events since they influence present and future events.

More recently, Pearl proposed the Bayesian Networks [69]. These networks can represent probability distributions, but, in special cases, the connections between nodes can also represent causal relationships [66]. This theory of causality that can be represented through a probability is defended by Pearl and other authors,

such as Peter Spirtes, Clark Glymour, among others, including Clive Granger [49] (Figure 1).

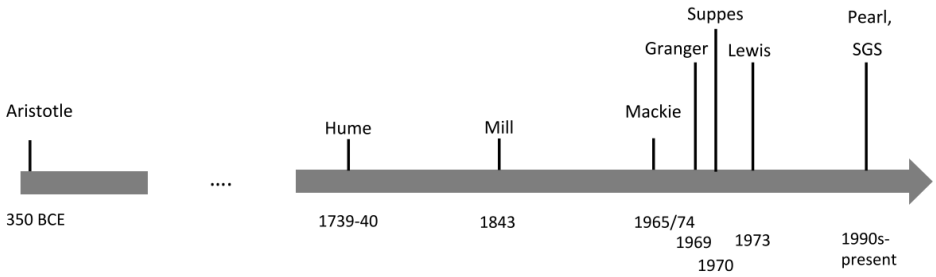

FIGURE 1. Overview of the evolution of the term "causality" and the main contributors [48]

3. **Causal discovery in machine learning (Pearl's perspective).** Causal inference is the process in which we compare the potential outcomes (also called counterfactuals) of an event, in which we have different conditions, but the same variables (for example, a patient with a disease takes a drug or not) [83].

We can look at this idea in two different ways [70]:

- From an observational perspective, in which we formulate the interaction of events A and B as a conditional probability $P(A|B)$, and where $P(A = a|B = b)$ signifies that we want to find the probability of $A = a$ that is conditioned by $B = b$;
- From an interventional perspective, in which we formulate the interaction between the events A and B as $P(A = a|do(B = b))$[2] that represents the probability of A if we set B as b. This means that to find A's probability, we set B with a specific value, but maintain the rest of the system the same.

These two approaches are usually applied in different situations. For example, if the objective is to study how the system interacts naturally and make a diagnosis about it, the observational approach is usually the one applied. If the goal is to actively intervene in the system (for example if we make a patient take a drug what the outcome is), the interventional approach should be applied.

The application of probabilistic theories to causal models is canon in several areas, such as economics, epidemiology, sociology and psychology. In these areas, more critical than finding causal relationships is to discover the probability of the relationship, *i.e.* the degree of belief that an event caused another [29, 82].

4. **The beginning: Bayesian networks.** In order to understand some of the methods presented in this survey, several concepts need to be comprehended. These concepts will be presented in this section.

In 1985, Judea Pearl proposed a new probabilistic graphical model: the Bayesian Networks [69]. These networks are graphical representations of probabilistic dependencies, called conditional dependences. This representation is done through

[2]the notation *do*, was first introduced by Judea Pearl to represent the active intervention of an outside entity in the system [68]

directed acyclic graphs (DAG): graphs that have no directed cycles, meaning that it is not possible to start on a node and return to this same node in a sequence (Figure 2).

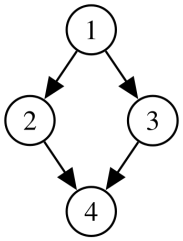

FIGURE 2. Example of a DAG

In these kinds of models, each node represents a variable, and the edges represent dependencies between them (for example, in Figure 2 node 4 depends on nodes 2 and 3). These dependencies can be expressed as joint probability distributions and can be obtained by Eq.(3) [66].

$$P(x_1, ..., x_n) = \prod_i P(x_i|pa_i) \tag{3}$$

In Eq (3), x_i represents the node and pa_i represents its ancestors. We can take Figure 2 as an example: the conditional probability of node 2 and 3 is given by $P(2|1)$ and $P(3|1)$, and node four is given by $P(4|2,3)$ since node 4 has two ancestors.

In order to develop and understand these kinds of models, certain assumptions must be fulfilled (in general). These assumptions are [104]: **Markov condition** and **Faithfulness**.

The first assumption (Markov Condition) states that *"a directed acyclic graph G over V and a probability distribution P(V) satisfy the Markov condition if and only if for every W in V, W is independent of V\(Descendants(W) ∪ Parents(W)) given Parents(W)"* [91], implying that any two unconnected nodes in a graph are independent of each other, given all their intermediate nodes. This assumption is a particular case or another concept: d-separation. This concept defines that three disjoint nodes must be such that one blocks all paths between the other two nodes.

The second assumption (Faithfulness) is related with the Markov Condition: *"If we have a joint probability distribution P of the random variables in some set V and a DAG G =(V,E), (G,P) satisfies the faithfulness condition if, G entails all and only conditional independencies in P."*[66]. This concept defines that if two variables are probabilistically dependent, there must be an edge connecting those variables.

In the combination of these two assumptions, it is possible to relate the conditional dependences given by the probabilities with the edges that represent the causal relationships between the variables. Due to these assumptions, these kinds of models can be interpreted in different ways [18]:

- In a probabilistic way, where the relationships are represented with probabilistic conditional independence. This signifies that, for example $A \perp\!\!\!\perp C|B$[3] can be expressed as $A \leftarrow B \rightarrow C$, $A \leftarrow B \leftarrow C$ and $A \rightarrow B \rightarrow C$ ($X \leftarrow Y$

[3] A is conditional independent from C, given B

meaning **Y depends of X**), since these are all equivalent, in the sense that all these graphs represent the same underlying independencies A and C are conditionally independent given B);

- In a causal way, where the connection between two nodes represents a causal relationship, and for this reason, a change in the direction of the relationship represents a completely different action. For example, $A \perp\!\!\!\perp C|B$, that signifies that B is a common cause of both A and C can only be expressed as $A \leftarrow B \rightarrow C$, since A and C are only causally related due to existence of B, and because of that the DAGs $A \leftarrow B \leftarrow C$ and $A \rightarrow B \rightarrow C$ are not equivalent.

Although there is a clear difference between probabilistic Bayesian networks and causal Bayesian networks, there are cases in which a model may fall in both categories. Moreover, it is also possible to see causal Bayesian networks as a particular case of Bayesian networks.

5. **Bayesian networks based causal discovery algorithms.** A particular case of the Bayesian Networks is the Causal Bayesian Networks. In these Bayesian Networks, the nodes represent the studied variables and the edges the causal relationships between them (instead of mere correlation). In these graphs, the directionality of the edges represents the direction of the causal relationship, *i.e.* in a causal graph a relationship X $\rightarrow$ Y means that X causes Y.

As in the case of probabilistic Bayesian Networks, certain assumptions must be accomplished[4]. Some of these assumptions are common to normal Bayesian networks, but others are specific for causal networks. These assumptions are: **Causal sufficiency**, **Causal Markov condition**, **Faithfulness** and **Independency tests**.

The first assumption (Causal Sufficiency) states that *"for every pair of variables which have their observed values in a given data set, all their common causes also have observations in the data set"*[56], [91]. This condition defines that a pair is causally sufficient if all the common causes of a pair of variables are measured, meaning that there are no hidden causes. Although most of the causal discovery relies on this condition, it cannot be satisfied in all cases. In these situations, some techniques can create causal structures without relying on causal sufficiency, through the assumption of linearity and the existence of latent variables. Nevertheless, how can we assure the causal sufficiency condition? If the data is obtained in a closed system, one can justify that there are no latent variables.

The second assumption (Causal Markov Condition) is a particular case of the Markov Condition: *"a directed acyclic graph G over V and a probability distribution P(V) satisfy the Markov condition if and only if for every W in V, W is independent of V\(Descendants(W) ∪ Parents(W)) given Parents(W)"*[91]. This condition is almost always fulfilled in causal graphs where there are no common hidden causes (if the causal sufficiency assumption is fulfilled).

The last condition (independence tests) is satisfied when the correct independence test is used (*i.e.* applying tests to discrete data in discrete variables and tests to continuous data in continuous variables) and if the test has a reliable result. The most common used independency tests are: G_2 test (used for discrete variables) [2], mutual information (also used for discrete variables)[101] and Fisher's test (used for continuous variables) [71].

[4]as it will be explained later in this document, certain algorithms can circumvent these assumptions

If all these assumptions are fulfilled, we can start applying an algorithm. Typically, a causal discovery algorithm has three steps that are applied [104]: the creation of a skeleton that connects the variables with undirected edges, search for v-structures ($X \rightarrow Y \leftarrow Z$) and orientation of all the possible edges.

In the first step (creation of a skeleton that connects the variables with undirected edges), two different approaches can be applied depending on the situation. The first approach (global approach) builds an undirected graph with all the variables, through the application of independence tests. Typically the algorithm starts with all variables connected through undirected edges. In each iteration, the algorithm removes some of these edges through independence tests (first tests on all variables with one conditional variable, then two conditional variables, *etc.*). For the second approach, called the local approach, the algorithm searches these skeletons locally for one or more variables. Usually, the selected nodes are adjacent nodes or Markov Blanket[5] of the variables. Next, the algorithm aggregates all local skeletons into a global skeleton [63]. The first approach is usually used in relatively small data sets (number of variables) and the second one when we have data sets with a large number of variables.

In the second step (search for v-structures), the objective is to find connections that can be transformed into something similar to Figure 3. To have a v-structure, there has to be a triple of nodes such that there are two connections of type $X \rightarrow Y$ and $Z \rightarrow Y$ and that there is no connection between X and Z. If this proposition holds, we can direct the edges as we see in Figure 3. This process of finding v-structures is done through what is called d-separation.

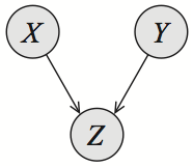

FIGURE 3. Example of a v-structure

Finally, in the third step, we orient the remaining undirected edges. This orientation is done in three different forms [104]:

1. By using a set of established rules (that tell the algorithm how to orient the edges if it comes across specific patterns);
2. By using experimental data to orient the edges (by manipulating the variables and obtaining the statistical association);
3. by using a mixture of both the previous approaches (orient the edges with the first method and then use the second method to orient remaining undirected edges).

It is important to note that there is a particular case of causal algorithms (local discovery causal algorithms). In these algorithms, the first step is altered so that the skeleton is only created around a chosen node, while the second and third step are ignored.

[5]set of variables that protects a given node from the remaining network. This protection makes the knowledge that the node receives restricted to this shield that is constituted by the nodes father, children and parents of the children (the so-called spouse nodes).

This line of work may be different depending on the algorithm. It may vary from algorithm to algorithm, depending on they being constraint-based, score-based or another type, as will be presented in the following sections.

5.1. **Constraint-based algorithms.** One of the most common approaches used to find causal relationships in data are the constraint-based algorithms. In these types of algorithms, the purpose is to find the graphs representing precisely the independence relationships present in the data, through hypothesis tests [61].

Some of the most common algorithms (and variants) found in the literature will be presented in the following sections.

5.1.1. *Peter-Clark (PC) algorithm.* The PC algorithm a constraint-based algorithm (and perhaps one of the best known) and was proposed by Spirtes *et al.* [91]. This algorithm relies on the *faithfulness* assumption, which means that all the independencies in a DAG need to be under the d-separation criterion (Section 5).

This algorithm, in its initial phase, constructs a skeleton with undirected edges and in next step, it orients these edges [1].

The first phase is, as explained in Section 5, the phase in which the skeleton of the model is created (graph with undirected edges).

In this phase the edges between two nodes X and Y are removed if there is a set of nodes adjacent to X (S) that are contained in the set of all nodes adjacent to X (except Y) and then X and Y are conditionally independent given S ($X \perp\!\!\!\perp Y|S$) [43]. Finally, the edge is removed.

In the second phase, the algorithm orients the edges by first giving direction to all links that appear to be v-structures and then by applying a set of rules. This set of rules together with v-structures are used to direct all possible edges to create what is called Completed Partially Directed Acyclic Graph (CPDAG) [43].

The CPDAG is a type of DAG that is characterised by the existence of directed and undirected edges (this is the main difference to the DAGs) and in which, as in the case of DAGs, it is not possible to have cycles. The existence of undirected edges signifies that equivalent DAGs may exist. These graphs have all the directed edges equal to the original graph, with an average of two equivalent DAGs per each undirected edge (one with to edge $\leftarrow$ and another to edge $\rightarrow$). With these final steps, we create a new graph that is equivalent to the original one, where the faithfulness is respected.

Currently, several algorithms were developed based on the PC and are considered as modifications/increments to it. One of these algorithms is the **PC-stable** [16]. This algorithm is an adaption of the PC algorithm that tries to solve another problem present in the original approach: PC is dependent on the order in which the variables are given to the algorithm. This adaptation modifies the way that the algorithm creates the skeleton: in each level, the nodes that should be removed, are also saved in a queue and are only permanently removed in the next iteration. This change causes the removal of edges to no longer affect which independency tests are performed in the same iteration.

More recently, the **MPC** [96] and **MRPC** [5] algorithms were proposed. The first algorithm (MPC), proposed by Tsagris [96], modifies the original orientation rules by adding a new rule that prevents the creation of cycles (although DAGs do not allow cycles, PC does not explicitly prevent their creation).

The second algorithm (MRPC), proposed by Badsha and Fu [5], was first used in genetic data. This algorithm, identically to PC, has two distinct phases. In the

first phase (skeleton phase), the algorithm searches for independences, starting with a fully connected graph and applies the same independence tests has PC (G^2). It is in the second phase (orientation phase) that this algorithm differs from PC. The MRPC rules are:

1. If the data contains genotype information and if there is an edge that contains a genotype node[6] and a non-genotype node, then the edge should be *non-genotype* $\rightarrow$ *genotype*;
2. Search for v-structures;
3. Directs the remaining undirected edges by searching for groups of three nodes and check if the Principle of Mendelian randomisation [87] is consistent in five basic models. If none of the models match the nodes, one of the edges remains undirected (for example, $A \rightarrow B - C$).

Another extension of PC is the **parallel-PC** [54] which is a parallel adaptation of it. This parallelisation is done in the conditional independence tests that are performed at each level of the algorithm. These tests are grouped and distributed in different cores. At the end of each iteration, the results are combined, producing faster results when compared with the PC algorithm.

Besides the global discovery approaches, there is another type of PC-based algorithms: local discovery. One example of local discovery algorithm is the **PC-simple** [10]. This algorithm is similar to PC since it uses the same methods for detecting associations and searching for the set that conditions a pair of nodes. The difference is that it only analyses the variables that are strongly related to the selected variable (that is, it does not analyse or create the network with the remaining variables). This algorithm was developed to deal with high dimensional data.

The **HITON-PC** [3] is another constraint-based algorithm that is also based in the PC. This algorithm is a mixture between two of the previously presented algorithms since, like PC-stable, it modifies the way the PC creates the skeleton, making it independent of the order. Besides, HITON-PC (like PC-simple) only builds the graph around a variable (target), recreating its Markov Blanket.

5.1.2. *Fast Causal Inference (FCI).* The FCI is another causal discovery algorithm that can be used to find causal relations in observational data. This algorithm differs from the previous ones (despite being also a constraint-based algorithm) because it does not assume causal sufficiency. What this means is that this algorithm assumes that some common causes cannot be measured in real-world data. For this reason, it uses a Maximal Ancestral Graph (MAG) to represent causal relationships (instead of the traditional DAG). In this model, latent variables are represented by bi-directed edges. Thus, we have three types of connections [104]:

- $\leftarrow$ or $\rightarrow$, which represents that a node is a parent of another node;
- $-$, which represents that the nodes are neighbours;
- $\leftrightarrow$, which represents that the nodes are spouses.

A MAG is a particular type of ancestral graph because, for each pair of non-adjacent nodes, there is a set of nodes that separate them. Also, these types of graphs are constructed in such a way that it is not possible to add more edges without having to revise the independences.

[6]Node that represents genetic data

This means that this algorithm can perform well on large data sets, even if there are hidden variables and selection bias[7].

Like PC, FCI consists in two distinct phases: In the first phase the algorithm searches for every conditional independence between every pair of variables to create the skeleton of the underlying MAG (instead of DAG), and in the second phase, it orients all the connections, to identify the direction of the edges($\leftarrow$ / $\rightarrow$,$-$ or $\leftrightarrow$) [14].

At the beginning of this algorithm, it starts with a fully undirected graph, that connects the observed variables. Then, it removes an edge between two variables, if the used oracle[8] states that these variables are d-separated in the subset Z of $O\backslash\{A, B\}$ and saves this subset. In the end of this phase, *Adjacencies(Q,X)* is composed by all the adjacencies of X in the graph Q. After analysing all the edges and adjacencies from the graph, the algorithm starts orientation phase.

Like PC, FCI as had several algorithms that are modifications/increments to it. One example is the **Really Fast Causal Inference (RFCI)**. This algorithm implements the same first phase as FCI, and its innovation is in the second phase: instead of performing conditional independence tests in the subsets of *POSSIBLE-D-STEP*, it performs *"additional tests before orienting v-structures and discriminating paths to ensure soundness"*[17].

5.2. **Score-based algorithms.** Another type of causal discovery algorithms are score-based algorithms. These algorithms differ from the previous ones because they compare the models through some adjustment measures, such as the Bayesian Information Criterion (BIC).

These kinds of algorithms are usually costly in terms of performance since they have to score every model. Because of this, some algorithms apply a greedy approach to restrict the number of scores calculated.

5.2.1. *Greedy Equivalence Search (GES).* The GES [64], [13] was proposed to create models in data sets with a large number of variables since it does not consider all existing patterns [66]. Instead of searching for the optimal DAG, this algorithm chooses a node and analyses all its possible neighbours, giving them scores and choosing as the next node, the one with the highest score, if this score improves the overall score of the DAG [13]. This algorithm generally consists of two phases:

1. Adds dependencies to the model until it reaches a local maximum ;
2. Dependences are removed from the model. This procedure also stops in a local maximum, which is the equivalence DAG.

In the Forward Equivalence Search (FES) phase, the algorithm adds new edges between two nodes X and Y if these nodes are non-adjacent and there is no neighbour of Y that is not adjacent to X. In this phase, the algorithm also directs every edge that is neighbour of Y and not adjacent to X to Y (T $\rightarrow$ Y).

Once the FES algorithm reaches a local optimum, the second phase of the GES algorithm begins, unlike FES, which removes arcs.

---

[7] *"For each set of variables O, and each population Pop such that S=1, there is a causally sufficient set of variables V such that O ∪ S ⊆ V and for all A, B, C ⊆ O, I(A,(C ∪ (S=1),B) in Pop if and only if the causal DAG G(O,S,L) relative to V in Pop entails that I(A,(C ∪ (S= 1),B) in Pop"*[90]

[8] *"procedure that responds with the correct answer to any query about the independence or conditional independence of variables represented by observed nodes in a network"*[31].

In the Backward Equivalence Search (BES) phase, the best link is removed in each iteration with the following criteria: it deletes every edge of type X - Y or X $\rightarrow$ Y if there is a subset of neighbours of Y adjacent to X. Besides, the algorithm transforms all edges H - Y (subset of neighbours of Y adjacent to X) to H $\rightarrow$ Y and all edges X - H for X $\rightarrow$ H.

Like the other presented algorithms, GES has had some variations proposed over the years. One example is the **Greedy Interventional Equivalence Search (GIES)** [37], in which, besides the FES and BES phases, there is a third phase called the Turning Step.

In this turning phase, the algorithm elongates the DAG sequence, gradually moving away from the original graph without losing edges or creating new ones so that the previous graph could be reconstructed by only changing one arrow. This phase was added with the intent to enhance estimation.

Another algorithm proposed as an enhancement of the GES algorithm is the **Fast Greedy Equivalence Search (FGS)** [76]. This algorithm uses parallelisation to optimise the original GES and applies a limited faithfulness assumption. To make the algorithm even faster, it can allow the graph to ignore the Markov factorisation.

6. **Other causal discovery algorithms.** Although the approaches presented above are the most used, other approaches escape these categories of algorithms. The approaches that will be presented in the next sections alter algorithms that traditionally do not represent causality so that it is possible to represent it, also trying to face possible existing flaws.

6.1. **Causal neural networks.** The causal neural networks [100] are an algorithm that adapts a non-causal algorithm to make causal discovery. In this algorithm, an alteration to the *feedforward* neural network is proposed, to be more like a Bayesian Network to represent causal relationships. The Causal Neural Networks are structured to represent the input variables as output and vice-versa. Hence, it is possible to represent causes in the input layer and effects in the output layer (Figure 4).

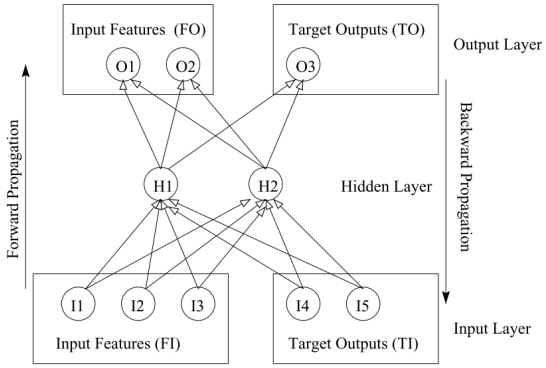

FIGURE 4. Example of a Causal Neural Network with one hidden layer [100]

The algorithm splits the input and output features between input and output layers (FI and FO in Figure 4) with the aid of genetic algorithms, thus creating an

optimal causal structure. To deal with hidden variables, the algorithm applies a second method, called *forward-backward propagation* that mixes the forward propagation theory with the *backpropagation* theory from Rumelhart *et al.* [81]. In this sub-algorithm, the input features are inserted in the FI and the FO's output layers. The unknown features of the input features are inserted in the TI and have as initial value their mean-bias. The output unknown features are inferred by applying a forward propagation (this step is repeated until a termination condition is met), and all the values are updated using backward propagation.

6.2. **Causal association rules discovery.** The Causal Association Rules Discovery algorithms are association rules algorithms in where it is used some test to do causal discovery [58]. This type of approach has the advantage that, in large data sets, we can create causal hypotheses when using association rules, which are no more than candidates for causal relationships. Also, it has the advantage that the associations' left-hand side can be a combination of several variables, which helps in combined cause avoidance (multiple variables influence a single effect).

Within this type of algorithm, causal relations can be obtained in several ways. One is through partial associations (Cochran-Mantel-Haenszel test), which is applied by the CR-PA algorithm [42].
This algorithm starts by searching for frequent itemsets and saves the identified variables (those which to frequent itemsets). After this, the algorithm applies the χ^2 as independence test and it defines that two variables are associated if the value resulting from the test is equal and greater to its critical value, with significance level α. To the variables by χ^2 is then applied the Cochran-Mantel-Haenszel test (Eq.(4)).

$$CMH(A,B) = \frac{(|\sum_{k=1}^{r} \frac{n_{11k}n_{22k} - n_{21k}n_{12k}}{n_{..k}}| - \frac{1}{2})^2}{\sum_{k=1}^{r} \frac{n_{1.k}n_{2.k}n_{.1k}n_{.2k}}{n_{..k}^2(n_{..k}-1)}} \tag{4}$$

In the previous equation, the values n represent the cells of contingency tables identical to Table 2, being that n_{11k} represents the first cell in the first row of table k, n_{12k} the second cell in the first row, n_{21k} the first cell in the second row and n_{22k} the second cell in the second row. $n_{1.k}$, $n_{2.k}$, $n_{.1k}$, $n_{.2k}$ and $n_{..k}$ represent the sum of the cells in the first row, second row, first column, second column and all the cells of a table k (Table 2), respectively. Another way to find causal relationships is through

TABLE 2. Example of a partial contingency table (in where $c_k = \{A = a1, B = b1\}$)

$c_k = \{A, B\}$	$C = c_1$	$C = c_2$	Total
$D = d_1$	n_{11k}	n_{12k}	$n_{1.k}$
$D = d_2$	n_{21k}	n_{22k}	$n_{2.k}$
Total	$n_{.1k}$	$n_{.2k}$	$n_{..k}$

retrospective cohort studies (odds ratio), applied by the CR-CS algorithm [57]. This algorithm selects two types of samples (without knowing its response): exposure samples and control samples. With this data, the algorithm tries to match the two samples, so that the distribution of the control variables of the two groups are as similar as possible. This is done by creating a fair data set (data set containing only matched records), which simulates a cohort study and then matching the records

in such way that the exposure records are only matched with the control records (this matching can be exact, or it may be necessary to apply similarity measures, such as the Euclidean distance or the Jaccard distance).

Like the CR-PA, the CR-CS applies a partial association test to discover causal relationships. Instead of using the Cochran-Mantel-Haenszel test, it uses odd ratio, since it is widely used in retrospective cohort studies [26]. This algorithm defines the association between two variables as being true if the minimum support of the association is greater than the threshold and if the odds ratio (Eq. (5)) of the association is greater than the minimum odds ratio.

$$OR_D(x \to z) = \frac{n_{11} * n_{22}}{n_{12} * n_{21}} \tag{5}$$

6.3. **Causal decision trees.** The Causal Decision Tree [56] ins an algorithm that uses traditional Decision Tree algorithm and alters it. Hence, it is possible to represent causal relationships between the nodes (non-leaf nodes represent causal attributes, leaves represent the outcome's values, and the edges represent the assignment of the value attribute). It does that by applying the Cochran-Mantel-Haenszel test [62] for partial association for binary outcomes [7] with one degree of freedom to evaluate the causal effect of the variables (Eq. (4); this is the same independence test used in CR-PA, in Section 6.2). If the null hypothesis that Q and Y are independent ($CMH(Q,Y) \geq x^2_\alpha$, being α the significance level) is rejected then the partial association of Q and Y is significant (the two variable are causally related).

The authors give a more specific definition: *"In a causal decision tree, a non-leaf node Q represents a context-specific causal relationship between Q and the outcome Y where the context is a series of value assignments of the attributes along the path from the root and to the parent of Q. A leaf node represents a value assignment of Y, which is the most probable value of Y in the context-specific data set where the context is a series of value assignments of the attributes along the path from the root to the leaf"*[56].

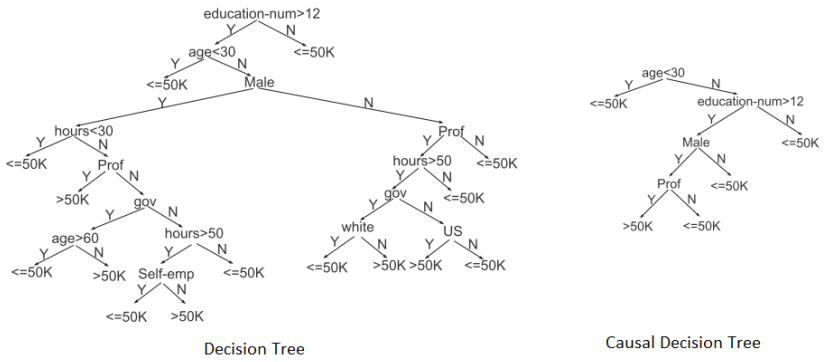

FIGURE 5. Example of a Causal Decision Tree [56] and comparison with a normal Decision Tree

In short, it is possible to represent causality in decision trees, as we can see in Figure 5, which represents the causal relationship between the income of a person, taking into account their age, education, gender and profession. In this figure, it is also possible to see the equivalent Decision Tree. By analysing this tree, we can

conclude that in the traditional Decision Tree, more splits are needed to determine what a person's income is 0.

7. **Causal discovery over time.** Until now, all the algorithms presented do not take into account the temporal component of data. Although this aspect of causal relationships is often overlooked (often because it is simply not available), it impacts how causality is viewed.

When two events are causally related, they are separated by a certain period, with the cause occurring first and the effect second. Nevertheless, this temporal spacing can also lead to false relationships. If two events happen sequentially by mere coincidence, assuming this temporal precedence can lead us to identify these events as causal when they are not.

Although time can be critical to causal discovery, it is not a crucial variable. It can be considered background knowledge that can be used as extra information to direct a causal relationship, thus not being necessary [61], [34].

Nevertheless, time (when available) continues to be very important, since it helps on the identification of correlations with the potential to be causal relationships.

As it was explained previously, Bayesian networks are the representation of conditional probabilities, which do not take time into account. However, this temporal component can be used in these networks as a constraint [46].

Although time is not required for creating causal models, there is one important definition that is used instead: causal ordering (*"fixing a particular time scale and considering only causes happening at time t and effects happening at time* $t + \delta t$*, where* δt *can be made as small as we want"*[36]).

Bayesian networks may not have been designed to handle time. However, some strategies can be applied to analyse temporal data. These strategies are [30]:

1. Handle the data in "windows", *i.e.*, partition the data in parts and take measurements from each window;
2. Estimate the number of lagged effects and deal with the measurements independently, using the lags as data analysis units;
3. Treat each measurement independently of all the other measurements;
4. Finally, there is a method that applies vector auto-regression (also called Granger Causality), that can be applied.

Although causal discovery (using causal Bayesian networks) in time-series is a relatively new area, several algorithms are already available, some of which are going to be presented in the next sections.

7.1. **DOCL.** The Dynamic Online Causal Learning (DOCL) [53] is an online version of PC (presented in Section 5.1.1), meaning that it is adapted to deal with data that has a temporal component. This algorithm, unlike PC, has three phases instead of only two:

1. Online Covariance Matrix Estimator (OCME) module is applied;
2. Causal Model Change Detector (CMCD) module is applied;
3. Causal Model Learner (CML) module is applied;

As it is possible to understand from the phases presented previously, there are three important modules responsible for key tasks. The **Online Covariance Matrix Estimator (OCME)** is the module responsible for estimating the covariance

matrices for incoming data points. To each of these data points, a weight is attributed, such that the weight of the datapoint d_{r+1} is higher than the weight of the datapoint d_r.

The **Causal Model Change Detector (CMCD)** is the module responsible for detecting divergences between the estimated covariance matrix and the incoming data points. The Mahalanobis distance gives this divergence (Eq. (6); C^r represents the covariance matrix, X^r represents the datapoint and $\overrightarrow{\mu}$ (represents the current estimate of the means): if there are consistently larger distances over consecutive data points, this means that the model no longer fits the incoming data. Therefore it is necessary to re-train the model.

$$D_r = (X^r - \overrightarrow{\mu})(C^r)^{-1}(X^r - \overrightarrow{\mu})^T \tag{6}$$

Finally, the module **Causal Model Learner (CML)** is responsible for learning the model with the covariance matrix estimated by the OCME module. To create this model, CML applies the PC. Besides this, the module is also responsible for checking the CMCD module for new covariance matrices.

7.2. **OFCI and FOFCI.** The Online Fast Causal Inference or OFCI [51], as the name says, is a modified version of FCI that considers time. It is important to note that OFCI has a similar structure to DOCL presented previously, the only difference being the causal discovery algorithm applied. This algorithm, unlike FCI, is divided into three phases:

1. In the first phase, the algorithm starts by applying the Online Covariance Estimator (OCME) to each data point sequentially and uses these data points to update the covariance matrix, the sample size and the mean. To each datapoint a sample size of one and a changeable weight are attributed (weight of a data point is always equal or greater than the weight of the previous data point);
2. In the second phase, the algorithm applies the Causal Model Change Detector (CMCD) to detect changes between the current covariance matrix and the input data. This difference (also called the fitness of the model to the data) is given by the Mahalanobis distance [19]: if the distance is consistently larger, over several consecutive data inputs, then the current covariance matrix does not fit the upcoming data points. If this happens, the algorithm starts giving a larger weight to the new data points;
3. Finally, the algorithm applies the Causal Model Learner (CML), that creates the model from the statistics retrieved in point 1. In contrast to the algorithm presented in Section 7.1, here CML uses FCI instead of PC.

A similar method to OFCI is FOFCI (Fast Online Fast Causal Inference) [51]. This algorithm was proposed to be a faster version of the first one since it applies RFCI (Section 5.1.2) instead of FCI in CML.

7.3. **Granger causality.** One of the most known algorithms to deal with causal temporal data is the Granger causality, proposed by Clive Granger. Norbert Wiener first proposed this way of inferring causality. He theorised that a variable could be considered the cause of a second variable if it is possible to predict the second using past information from the first one [99]. Granger took this idea and from it created a practical method to find causality in the financial domain, more specifically in the stock returns.

The main idea behind this test is that, if we have two distinct variables A and B in a time-series, we can predict more accurately the first variable value in the future (A_{t+1}) if we use the past values of both A and B, than if we only use data from A (if A can be predicted by both A and B, we can say that B *G-causes* A) [9].

Initially, the Granger test was proposed to calculate one variable's influence on another (bivariate approach) [32]. The bivariate Granger causality is similar to correlation, despite the measures not being symmetric. Mathematically speaking, it can be formalised as a linear regression, in which the previous observations of a cause variable and an effect variable are added, as it is possible to see in the following equations (Eq. (7)) [32]:

$$\begin{aligned} X_t &= \sum_{j=1}^{m} a_j X_{t-j} + \sum_{j=1}^{m} b_j Y_{t-j} + \varepsilon_t, \\ Y_t &= \sum_{j=1}^{m} c_j X_{t-j} + \sum_{j=1}^{m} d_j Y_{t-j} + \eta_t \end{aligned} \tag{7}$$

In these equations m represents the model order or the maximum number of prior observation to be used, a, b, c and d are the contributions of each delayed observation of the predicted values X_t and Y_t and ε and η are the residual errors [32].

Besides the bivariate approach, other approaches involve three or more variables. When we want to evaluate the causal relations of exactly three variables, it is possible to apply the *Conditional Granger Causality* [20]. This approach was designed to analyse situations in which two variables may have a relationship in which a third is intermediate of this relation (*i.e.* X $\rightarrow$ Y $\rightarrow$ Z).

In terms of procedure, we first apply the bivariate Granger test between the cause variable and the assumed conditional variable, and secondly, we apply the conditional Granger Causality equations. These equations are similar to the bivariate approach with the difference that a third variable is added to each Eq. (8):

$$\begin{aligned} X_{\mathrm{t}} &= \sum_{j=1}^{\infty} a_{\mathrm{j}} X_{\mathrm{t\,-\,j}} + \sum_{j=1}^{\infty} b_{\mathrm{j}} y_{\mathrm{t\,-\,j}} + \varepsilon_{\mathrm{t}} + \sum_{j=1}^{\infty} c_{\mathrm{j}} Z_{\mathrm{t\,-\,j}} + \varepsilon_{\mathrm{t}}, \\ Y_{\mathrm{t}} &= \sum_{j=1}^{\infty} d_{\mathrm{j}} X_{\mathrm{t\,-\,j}} + \sum_{j=1}^{\infty} e_{\mathrm{j}} Y_{\mathrm{t\,-\,j}} + \eta_{\mathrm{t}} + \sum_{j=1}^{\infty} f_{\mathrm{j}} Z_{\mathrm{t\,-\,j}} + \varepsilon_{\mathrm{t}}, \\ Z_{\mathrm{t}} &= \sum_{j=1}^{\infty} g_{\mathrm{j}} X_{\mathrm{t\,-\,j}} + \sum_{j=1}^{\infty} h_{\mathrm{j}} Y_{\mathrm{t\,-\,j}} + \eta_{\mathrm{t}} + \sum_{j=1}^{\infty} i_{\mathrm{j}} Z_{\mathrm{t\,-\,j}} + \varepsilon_{\mathrm{t}}, \end{aligned} \tag{8}$$

Finally, we have the multivariate approach in which we extend the bivariate approach to be used with more than three variables. In this multivariate approach, the variables (X, Y, *etc.*) are predicted by extending the conditional Granger Causality model presented previously to n variables (*i.e.* we create n new equations with n variables).

Both the Conditional Granger Causality approach and the multivariate approach have the advantage of not being necessary to carry out the bivariate analysis on each pair of variables[20], [86].

This kind of approach to find causation has its limitations. One of them is the fact that the bivariate Granger causality gives only information about linear features. Besides this, another limitation is that the variables must be stationary.

Finally, this kind of approach is very dependent on the observations that we are dealing with.

In short, we cannot say with certainty that there is a causal relationship between two events, although, in some situations, we are close to it [86].

8. **Other applications.** In addition to the traditional application of the causal discovery algorithms, causality can be applied to other types of problems such as feature selection.

The majority of feature selection algorithms focus on studying how the variables are related through correlation or gain of information, among others, to select them. However, these algorithms do not exploit the possible cause-effect relationships that may exist between these variables, and that could be relevant to the problem.

The application of causality to the selection of features has several advantages. Firstly, it can help in the models' construction since the relationships' mechanics' interpretation is improved. The application of causality may also improve the way we group the variables since the number of variables selected tends to be smaller than in another type of algorithm (we choose only those that are directly related to the target variable) [35].

Typically, causal feature selection algorithms that are used for causal discovery use the Markov Blanket to recreate the network around a target variable, but which can also be used for the selection of variables, since by Markov Blanket's theorem, the variables that influence a target variable are the parents, children and spouses. Therefore these variables could also be those that will be more important for their classification, for example.

One example is the HITON-PC presented in Section 5.1.1. The author proposed this algorithm in [3] as a feature selection algorithm to be used in data sets where there are a high number of variables but a low number of entries. However, although it has been proposed for feature selection, this algorithm can also be used for causal discovery.

Another possible application is in feature engineering. Using causal Bayesian networks is possible to create, for example, features that represent the probability of one variable causing another variable. Despite being an interesting topic and with relevant applications, this has not yet been covered in detail. We have, for example, the work from Nogueira *et al.* [67], who proposed a method that uses the Markov blanket or parents and children of a target variable, obtained by applying a modified version of PC, that used the Cochran-Mantel-Haenszel test as conditional independence test, to create new supposed causal features that represent the relationship between the target and these variables.

9. **Common discovery tools and frameworks.** For the application of causal discovery, several tools can help in its quick and easy implementation.

One of these tools is GeNie (Figure 6) [23], which currently belongs to BayesFusion[9]. In this graphical tool, it is possible to create Bayesian networks manually. However, the system must obey the Bayesian Networks rules (for example the nonexistence of loops). It is also possible to load existing data, treat it (fill in missing data, discretise, among others). There is a great variety of algorithms that can be implemented, among which the PC algorithm stand out. Besides, in this tool, it is still possible to perform tests through cross-validation.

[9] https://www.bayesfusion.com/

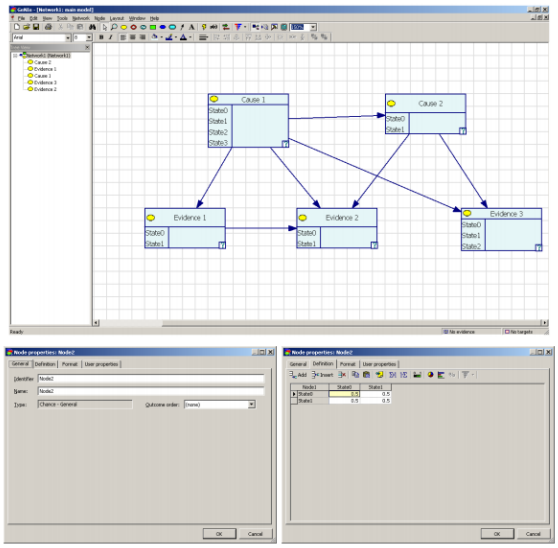

FIGURE 6. GeNie [38]

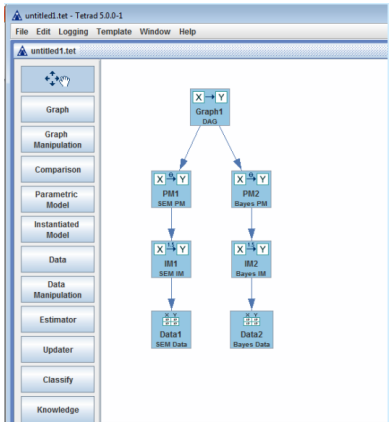

FIGURE 7. Tetrad [21]

Another tool that is used today is the Tetrad[10] (Figure 7) that was created by Scheines *et al.* [84]. This tool is a graphical tool that aids in applying causal algorithms (among other things). It can be used on data simulation, estimation, prediction and searches in both causal and statistical models. This tool is, however, limited only to discrete data. To create a model, it divides into three tasks:

1. The creation of the DAG using the relations of the variables;
2. Specification of the probability distributions and the parameters that are associated with the model;
3. Specification of the values of the parameters.

[10] http://www.phil.cmu.edu/tetrad/index.html

A different kind of tools are the ones that will be presented next. These tools are not graphical, but packages available in R and that help in the application of causal discovery and inference through the provision of implementations of algorithms such as PC, among others.

One of these packages is bnlearn[11] proposed by Scutari [85]. This package helps in the Bayesian networks' learning, by aiding in the parameters' estimation and the models' creation. This package has several types of Bayesian algorithms available:

- Constraint-based algorithms, such as PC algorithm, Grow-Shrink, IAMB, Fast-IAMB, Inter-IAMB, MMPC and SI-HITON-PC;
- Score-based algorithms, such as Hill Climbing and Tabu Search;
- Hybrid algorithms, such as Chow-Liu and ARACNE;
- Bayesian classifiers, such as Naive Bayes and TAN (Tree-Augmented Naive Bayes).

Finally, we have the pcalg that is another package from R, proposed by Kalisch *et al.* [44], and this package is specific for causal inference and discovery. This package, besides aiding in the application of causal discovery and the creation of the causal structures, also helps estimate causal effects in observational data. In this package, several algorithms are available (some of them were presented in this proposal, in Section 5):

- Algorithms to create the causal structure: PC, FCI, RFCI, GES and GIES;
- algorithms to find the causal effects: Intervention calculus when the DAG is Absent (IDA) [60] and Generalised Backdoor Criterion (GBC) [59].

10. **Evaluation metrics.** To evaluate causal discovery algorithms, more specifically, graphical models, such as PC (Section 5.1.1), FCI (Section 5.1.2) or GES (Section 5.2.1), there are a set of metrics that are used.

These metrics are [104]:

- **Missing edges -** refers to the number of edges that are missing from the model and that are present in the original structure;
- **Extra edges -** refers to the number of edges that are in the model but not in the original structure;
- **Correct undirected edges -** refers to the number of edges that are undirected and are present in both the model and the original structure;
- **Correct directed edges -** refers to the number of edges that are directed and are present in both the model and the original structure;
- **Incorrect directed edges -** refers to the number of directed edges that are different in the model and the original structure;
- **Computational efficiency -** refers to run time and the number of statistical tests;
- **Structural Hamming Distance (SHD)** [97] **-** measures the difference by counting the number of missing edges, extra edges and incorrectly directed edges;
- **Structural Intervention Distance (SID)** [72] **-** measures for each pair of variables A, B if the parents of A in the predicted model are a valid adjustment set in the true graph for the causal effect of A in B. If this is false, one is added to the distance. This adjustment is obtained by computing the marginalised intervention distribution ($p(B|do(A = \hat{a}))$).

[11] http://www.bnlearn.com/

Related with the previous metrics, we have the adjacency and arrowhead precision and recall (Eq. 9) [4, 75]. These metrics are an adaptation of the standard precision and recall evaluations metrics to evaluating graphs [74].

$$\begin{aligned} Adj\ Precision &= \frac{correctly\ predicted\ adjacencies}{predicted\ adjacencies} \\ Adj\ Recall &= \frac{correctly\ predicted\ adjacencies}{true\ adjacencies} \\ Arrhd\ Precision &= \frac{correctly\ predicted\ arrowheads}{predicted\ arrowheads} \\ Arrhd\ Recall &= \frac{correctly\ predicted\ arrowheads}{true\ arrowheads} \end{aligned} \tag{9}$$

The adjacency precision and recall measure the correctness of the model in terms of undirected edges or adjacencies (*i.e.* $A - B$ relationships), while the arrowhead precision and recall measure the correctness of the model considering the direction of the edges (*i.e.* $A \to B$ relationships) [77]:

- **Adjacency precision -** number of undirected edges predicted correctly divided by the number of predicted undirected edges. Reports the proportion of correctly identified undirected edges;
- **Adjacency recall -** number of undirected edges predicted correctly divided by the number of true undirected edges. Reports the proportion of true undirected edges identified;
- **Arrowhead precision -** number of directed edges predicted correctly divided by the number of predicted directed edges. Reports the proportion of correctly identified directed edges;
- **Arrowhead recall -** number of directed edges predicted correctly divided by the number of true directed edges. Reports the proportion of true directed edges identified.

It is important to note that, in order to apply these metrics, it is necessary to have a true representation of the underlying causal relationships present in the data (for example a graph) to be able to compare it with the generated model. When the true graph is not available, one possible solution is to compare the model with a baseline in terms of performance (*accuracy*, *error*, *etc.*).

11. **Possible applications.** As causality exists everywhere (we can see this daily), this can be applied in almost all kinds of problems, be it medicine, climatology, economics, *etc.*.

Beyond this fact, and given that with the technological evolution, data is now obtained in data streams and not in a static form, the application of these algorithms to problems in which the data is obtained in stream form is becoming usual.

In the next sections, some examples of the application of causal discovery algorithms in the most varied areas, both static data and stream data will be presented.

11.1. **Medicine.** The application of causal discovery in the medical field has always been made, even though it was not through the algorithms presented previously. The diagnosis of each disease is no more than discovering the causal relationship between the symptoms and the disease itself. The causal discovery applied through algorithms has been debated over the years [78] given that its application can help in the faster diagnosis of certain diseases.

One possible example is the work of Sokolova *et al.* [88] where an alternative approach is proposed to deal with mixed data (a mix of continuous and discrete data) and missing values, which are very common in medical data. This is done by transforming the continuous and discrete data into a normal distribution (many of the implementations of the algorithms presented in Sections 5.1 and 5.2 can only deal with continuous data) by using a Gaussian copula, consisting in (for each variable X_i):

1. The application of the rescaled empirical distribution (Eq. (10); in this equation $\mathcal{I}$ represents an indicator function; n the total number of observations and j the current observation of X_i) for each variable X_i;
2. The application of the inverse of the cumulative function of the normal distribution (Eq. (11); Φ represents the cumulative distribution function and $\hat{F}(X_i)$ the rescaled empirical distribution).

After having the data fit a normal distribution, the authors apply an Expectation-Maximization algorithm to find the correlation matrix to be used by the Bayesian Constraint-based Causal Discovery (BCCD) algorithm [15].

$$\hat{F}_i(x) = \frac{1}{n+1} \sum_{j+1}^{n} \mathcal{I}(X_{i,j} < x) \tag{10}$$

$$\hat{X}_i = \hat{\Phi}_t^{-1}(\hat{F}(X_i)) \tag{11}$$

This entire procedure is applied to data from patients with attention-deficit/hyperactivity disorder or ADHD.

Another example of the application of causal discovery into the medical field is the work of Kamiński *et al.* [45], who applied this discovery to the area of neurobiology. The authors try to relate the Granger Causality with directed transfer learning, more specifically, how it is possible to interpret directed transfer learning in the Granger causality framework. They accomplish this by comparing the non-normalized DTF with the spectral Granger Causality (this method is related to the autoregressive multivariate or MVAR coefficients, in bivariate cases). In multivariate cases, the authors associated these two topics by applying the Fourier transform of a bivariate model (among other mathematical formulas) to come up with the Granger Causality equation.

Furthermore, another example is the work of Chen *et al.* [12], who applied Multiple Cause Discovery combined with Structure Learning (McDSL) to find the causal risk factors of patients that are in the *Failure* stage. Finally, they select these variables to create a new data set, that is then used to train five classification algorithms (K-nearest neighbour, Decision Tree, Backpropagation Neural Network, Random Forest and an ensemble classifier, that is not specified).

11.2. **Economics.** Another area in which causality can be applied is in economy, as explained by Pol [73]. Although causality has a very intrinsic relationship with the economy (Clive Granger himself was an economist), this relationship is somewhat problematic, since there are four factors that are difficult to transpose. These are:

- There is no objective definition of causality in economics (like it happens in medicine);
- If there is a definition of causality it must be testable;
- The notion of causality should be able to capture the definition of control;

- In the case where an event has multiple causes, the tests are challenging to implement.

Despite this subjectivity underlying the difficulty in objectifying causality in the economy, this fact did not detain several authors from applying it in their work. One example is Giles *et al.* [28] which studied time-series data from the Canadian underground economy and its relationship with the Gross Domestic Product (GDP) of the country, using the Granger causality.

11.2.1. *Stock Market.* The stock market is a case-study case within economics since Clive Granger himself studied the stock market when he created Granger's causality. In this particular case, the interest is almost always in perceiving the relationship between stock prices and other factors that may exist.

The works in this area are about the relation between stock price and exchange rates or price-volume relation. Despite that, there are works in this area that diverge from this, such as the work of Bollen *et al.* [8], who thought outside the box and tried to relate public moods to the value of Dow Jones Industrial Average (DJIA) at closing time. In their research, the authors applied Granger causality and Self-organising Fuzzy Neural Networks to predict that.

Another work in this area belongs to Khorram *et al.* [47], where the authors study the application of Bayesian Networks in stocks prices, more precisely how the stocks relate to each other, using data that is known as being causal (this is another way to obtain causal models). This study is done by analysing the Malaysian stock securities (FBM100) data, and the networks are created using two frameworks: Tetrad IV and Genie 2.0.

Finally, Irfan *et al.* [40] proposed a causal strategic inference framework for networked microfinance economies, specifically in microfinance data from Bangladesh and Bolivia. In their framework, the authors suggest using two-sided networked models, constituted by microfinance institutions and villages (a microfinance institution is comprised of a set of villages). The village's objective is to acquire the maximal amount of loans, while the microfinance institutions purpose is to set their interest rates to obtain market-clearance.

11.3. **Climatology.** Another area where we find cases of application of causal discovery techniques is in climatology, which, contrary to what is said in common sense, is predictable. Climate science is a science that studies climate and tries to explain how changes can occur in order to help society plan their everyday activities, as well as help in choosing how buildings and structures are designed [95], or in short, make the planet a safer and more comfortable place to live.

In this area, it is common to apply causality to try to understand, for example, the influence of carbon dioxide pollution on the planet's temperature. That is why authors such as Kodra *et al.* [52] applied causality in their work. In their paper, the authors discuss the application of the bivariate Granger test, also proposing an adaptation of this test, so that it is better suited to the problem to be solved. This proposed extension is based on a new method (reverse cumulative Granger causality or RCUMGC test) which consists on the application of the Granger test in conjunction with the F test between the variables globally averaged land surface temperature and total radiative forcing with several lag values using additive latest windows (*i.e.* used in the last X years, in this case, 30 years).

Ebert-Uphoff and Deng [24] proposed another similar work. In their paper the objective was to find *"causal relationships between four prominent modes of atmospheric low-frequency variability in boreal winter including the Western Pacific Oscillation (WPO), Eastern Pacific Oscillation (EPO), Pacific–North America (PNA) pattern, and North Atlantic Oscillation (NAO)"*, using Bayesian Networks. In their case study, the authors use two distinct models: the first one called the static model, in which the authors use the data without temporal distinction and the second one, called the temporal model, in which the nodes of the model represent the daily values. In their results, they concluded that the generated models agree with one another, since, the relations that appear in the static model, also appear in the temporal.

11.4. **Causal discovery over Sequential Data.** Besides the previously presented areas, there are many other areas where causality can be applied, especially over sequential data.

One possible example is the work developed by Liu *et al.* [102], in which the authors apply causal discovery over sequential data from traffic. More specifically, they try to find outliers in traffic data and how they relate causally to each other, over time. This is done by creating an outlier tree using STOTree (algorithm also proposed in [102]), that models the different outliers' relationships. In the end, several trees are created (thus creating a forest) and the causal relationships are identified as the most common subtrees.

Another work that binds together causality and sequential data is the work of Yamahara, and Shimakawa [103]. In their work, the authors propose a new method to detect the states' transitions in this type of data by representing the transition's causal relationship as rules. This is done by identifying signs that hint that one past event can be the cause of a present event. This cause and signs are represented in the system through rules.

12. **Conclusion and open Issues.** The definition of causality started in Ancient Greece as a philosophical concept. More recently, this concept was redefined as a statistical/ machine learning concept. One of the most known machine learning algorithms to learn causal relationships from data are the Bayesian Networks, more specifically, the Causal Bayesian Networks. Within the Causal Bayesian Networks, there are the constraint-based and the scored-based algorithms. The most commonly used causal Bayesian networks used are PC, FCI and GES (and their respective variants).

Besides this, there are other algorithms that modify well-known machine learning algorithms, by using, for example, independence tests, so that it is possible to infer causal relationships from data. These are the cases of the Causal Decision Tree, CR-PA and CR-CS, and Causal Neural Networks.

The applications of causal discovery in real-world problems are countless: from the stock market to climatology, with an emphasis in medicine, where the usage of causal discovery has several advantages.

Currently, there are several open issues in the domain of causal discovery in machine learning that were not entirely solved:

- Although there are already several approaches that deal, some key issues were not solved entirely yet. For example, from the examples presented in Section 7, the disjoint windows can omit relations depending on the window size, while

the lags approach and the independent measurements fail whenever there is dependence between units;
- As it is now, measurement errors can appear in data, and causal discovery algorithms are particularly susceptible to measurement errors;
- Another open issue is selection bias, since causal discovery algorithms can be used in decision making, and this selection bias can alter the statistical results obtained;
- Yet another still open issue is how to deal with high-dimensional data. Since independence test calculates the (in)dependence between all the variables in a data set, the higher the number of variables, the more computationally costly the algorithm becomes. Although several authors tried to deal with this problem, the proposed solutions exchange the quality of the result by the execution speed, either in terms of creating a local model or by exchanging independence tests;
- Missing data is still an issue (some algorithms have been proposed, but they are only usable with certain assumptions);
- Finally and despite the existence of several metrics to measure a causal algorithm's performance, they all depend on the existence of a true representation, which demonstrates which relationships exist in the data set. Therefore, it is impossible to determine (at least through metrics) if a model is generated from a data set in which its actual representation is not known found causal relationships or only correlation.

In conclusion, the study of causality is not in itself a new subject. However, there is still much to explore. This can be a challenge since determining the existence of causality in the data passes not only by also the study and application of algorithms but by the in-depth study of the problems and their background.

Acknowledgments. This research was carried out in the context of the project FailStopper (DSAIPA/DS/0086/2018) and supported by the *Fundação para a Ciência e Tecnologia (FCT)*, Portugal for the PhD Grant SFRH/BD/146197/2019. This work was also partially supported by the European Commission funded project *Humane AI: Toward AI Systems That Augment and Empower Humans by Understanding Us, our Society and the World Around Us* (grant # 820437). The support is gratefully acknowledged.

REFERENCES

[1] J. Abellán, M. Gómez-Olmedo and S. Moral, Some variations on the PC algorithm, *Proceedings of the Third European Workshop on Probabilistic Graphical Models (PGM' 06)*, 1–8.

[2] A. Agresti and M. Kateri, *Categorical Data Analysis*, in International encyclopedia of statistical science, Springer, 2011,

[3] C. F. Aliferis, I. Tsamardinos and A. Statnikov, HITON: A novel Markov Blanket algorithm for optimal variable selection, *AMIA Annual Symposium Proceedings / AMIA Symposium. AMIA Symposium*, **2003** (2003), 21–25.

[4] B. Andrews, J. Ramsey and G. F. Cooper, Learning high-dimensional directed acyclic graphs with mixed data-types, in *Proceedings of Machine Learning Research* (eds. T. D. Le, J. Li, K. Zhang, E. K. P. Cui and A. Hyvärinen), vol. 104 of Proceedings of Machine Learning Research, PMLR, Anchorage, Alaska, USA, (2019), 4–21.

[5] B. Badsha and A. Q. Fu, Learning causal biological networks with the principle of Mendelian randomization, *Frontiers in Genetics*, **10** (2019).

[6] J. Barnes et al., *Complete Works of Aristotle, Volume 1: The Revised Oxford Translation*, vol. 1, Princeton University Press, 2014.

[7] M. W. Birch, The detection of partial association, I: The 2× 2 case, *Journal of the Royal Statistical Society. Series B (Methodological)*, **26** (1964), 313–324.
[8] J. Bollen, H. Mao and X. Zeng, Twitter mood predicts the stock market, *Journal of Computational Science*, **2** (2011), 1–8.
[9] S. L. Bressler and A. K. Seth, Wiener-Granger Causality: A well established methodology, *NeuroImage*, **58** (2011), 323–329.
[10] P. Bühlmann, M. Kalisch and M. H. Maathuis, Variable selection in high-dimensional linear models: Partially faithful distributions and the pc-simple algorithm, *Biometrika*, **97** (2010), 261–278.
[11] B. W. Carlson, Simpson's paradox — Definition, Example, and Explanation, *Encyclopedia Britannica*, (2019).
[12] W. Chen, Y. Hu, X. Zhang, L. Wu, K. Liu, J. He, Z. Tang, X. Song, L. R. Waitman and M. Liu, Causal risk factor discovery for severe acute kidney injury using electronic health records, *BMC Medical Informatics and Decision Making*, **18** (2018), 13.
[13] D. M. Chickering, Learning equivalence classes of bayesian-network structures, *J. Mach. Learn. Res.*, **2** (2002), 445–498.
[14] T. Claassen and T. Heskes, A structure independent algorithm for causal discovery, *Computational Intelligence*, 27–29.
[15] T. Claassen and T. Heskes, Bayesian probabilities for constraint-based causal discovery, *IJCAI International Joint Conference on Artificial Intelligence*, 2992–2996.
[16] D. Colombo and M. H. Maathuis, Order-independent constraint-based causal structure learning, *J. Mach. Learn. Res.*, **15** (2014), 3741–3782.
[17] D. Colombo, M. H. Maathuis, M. Kalisch and T. S. Richardson, Learning high-dimensional directed acyclic graphs with latent and selection variables, *Annals of Statistics*, **40** (2012), 294–321.
[18] A. P. Dawid, Beware of the dag!, in *NIPS Causality: Objectives and Assessment*, (2008).
[19] R. De Maesschalck, D. Jouan-Rimbaud and D. L. Massart, The mahalanobis distance, *Chemometrics and Intelligent Laboratory Systems*, **50** (2000), 1–18.
[20] M. Ding, Y. Chen and S. L. Bressler, 17 granger causality: basic theory and application to neuroscience, *Handbook of Time Series Analysis: Recent Theoretical Developments and Applications*, **437**.
[21] C. f. C. Discovery, Ccd-2015-1, *Summer Workshop - 2015*.
[22] F. K. Došilović, M. Brčić and N. Hlupić, Explainable artificial intelligence: A survey, in *2018 41st International convention on information and communication technology, electronics and microelectronics (MIPRO)*, IEEE, (2018), 0210–0215.
[23] M. J. Druzdzel, *SMILE: Structural Modeling, Inference, and Learning Engine and GeNie: A Development Environment for Graphical Decision-Theoretic Models*, Technical report, 1999.
[24] I. Ebert-Uphoff and Y. Deng, Causal discovery for climate research using graphical models, *Journal of Climate*, **25** (2012), 5648–5665.
[25] A. Falcon, Aristotle on causality, in *The Stanford Encyclopedia of Philosophy* (ed. E. N. Zalta), spring 2015 edition, Metaphysics Research Lab, Stanford University, (2015).
[26] J. L. Fleiss, B. Levin and M. C. Paik, *Statistical Methods for Rates and Proportions*, John Wiley & Sons, 2003.
[27] A. Gelman, Causality and statistical learning, *American Journal of Sociology*, **117** (2011), 955–966.
[28] D. E. Giles, L. M. Tedds and G. Werkneh, The Canadian underground and measured economies: Granger causality results, *Applied Economics*, **34** (2002), 2347–2352.
[29] D. Gillies, *Causality, Probability, and Medicine*, Routledge, 2018.
[30] C. Glymour, K. Zhang and P. Spirtes, Review of causal discovery methods based on graphical models, *Frontiers in Genetics*, **10** (2019), 524.
[31] C. N. Glymour, *The Mind's Arrows: Bayes Nets and Graphical Causal Models in Psychology*, MIT press, 2001.
[32] C. W. J. Granger, Investigating causal relations by econometric models and cross-spectral methods, *Econometrica*, **37** (1969), 424–438.
[33] H. P. Grice and A. R. White, Symposium: The causal theory of perception, *Proceedings of the Aristotelian Society, Supplementary Volumes*, **35** (1961), 121–168.
[34] R. Guo, L. Cheng, J. Li, P. R. Hahn and H. Liu, A survey of learning causality with data: Problems and methods, *ACM Computing Surveys*, **53** (2020), 37.

[35] I. Guyon, A. Elisseeff and C. Aliferis, Causal feature selection, *Training*, **32** (2007), 1–40.
[36] I. Guyon, A. Satnikov and C. Aliferis, Time series analysis with the causality workbench, in *NIPS Mini-Symposium on Causality in Time Series*, (2011), 115–139.
[37] A. Hauser and P. Bühlmann, Characterization and greedy learning of interventional markov equivalence classes of directed acyclic graphs, *Journal of Machine Learning Research*, **13** (2012), 2409–2464.
[38] M. Horný, *Bayesian Networks*, Technical report, 2014.
[39] J. Huyssteen, *Encyclopedia of Science and Religion*, Gale Group, Inc, 2003.
[40] M. T. Irfan and L. E. Ortiz, Causal strategic inference in a game-theoretic model of multiplayer networked microfinance markets, *ACM Trans. Econ. Comput.*, **6** (2018), Art. 6, 58 pp.
[41] A. Janiak, Three concepts of causation in newton, *Studies in History and Philosophy of Science Part A*, **44** (2013), 396 – 407.
[42] Z. Jin, J. Li, L. Liu, T. D. Le, B. Sun and R. Wang, Discovery of causal rules using partial association, *Proceedings - IEEE International Conference on Data Mining, ICDM* , (2012), 309–318.
[43] M. Kalisch and P. Buehlmann, Estimating high-dimensional directed acyclic graphs with the PC-algorithm, *Journal of Machine Learning Research*, **8** (2005), 613–636.
[44] M. Kalisch, M. Mächler and D. Colombo, *Causal Inference with Graphical Models in R Package Pcalg*, Technical Report 11, 2012.
[45] M. Kamiński, M. Ding, W. A. Truccolo and S. L. Bressler, Evaluating causal relations in neural systems: Granger causality, directed transfer function and statistical assessment of significance, *Biological Cybernetics*, **85** (2001), 145–157.
[46] K. Karimi, A brief introduction to temporality and causality, preprint, arXiv:1007.2449.
[47] A. Khorram, C. W. Ping and L. T. Hui, *Causal Knowledge-Driven Approach For Stock Analysis*, Technical report, 2011.
[48] S. Kleinberg, Causal Inference: Prediction, explanation, and intervention Lecture 2: Regularities, counterfactuals and token causality, Cambridge University Press, Cambridge, 2013.
[49] S. Kleinberg, *Why: A Guide to Finding and Using Causes*, O'Reilly, Sebastopol, CA, 2015.
[50] S. Kleinberg and G. Hripcsak, A review of causal inference for biomedical informatics, *Journal of Biomedical Informatics*, **44** (2011), 1102 – 1112.
[51] D. Kocacoban and J. Cussens, Online causal structure learning in the presence of latent variables, in *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, Boca Raton, FL, USA, (2019), 392–395.
[52] E. Kodra, S. Chatterjee and A. R. Ganguly, Exploring Granger causality between global average observed time series of carbon dioxide and temperature, *Theoretical and Applied Climatology*, **104** (2011), 325–335.
[53] E. Kummerfeld, D. Danks and M. Cognition, Online Learning of Time-varying Causal Structures.
[54] T. D. Le, T. Hoang, J. Li, L. Liu, H. Liu and S. Hu, A fast pc algorithm for high dimensional causal discovery with multi-core pcs, *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, **16** (2019), 1483–1495.
[55] H. D. P. Lee et al., *Timaeus and Critias*, Penguin, 1971.
[56] J. Li, S. Ma, T. Le, L. Liu and J. Liu, Causal decision trees, *IEEE Transactions on Knowledge and Data Engineering*, **29** (2017), 257–271.
[57] J. Li, T. D. Le, L. Liu, J. Liu, Z. Jin and B. Sun, Mining causal association rules, in *Proceedings - IEEE 13th International Conference on Data Mining Workshops,* ICDMW, (2013), 114–123.
[58] J. Li, L. Liu and T. D. Le, *Practical Approaches to Causal Relationship Analysis*, 2015.
[59] M. H. Maathuis and D. Colombo, A generalized back-door criterion, *The Annals of Statistics*, **43** (2015), 1060–1088.
[60] M. H. Maathuis, M. Kalisch and P. Bühlmann, Estimating high-dimensional intervention effects from observational data, *Annals of Statistics*, **37** (2009), 3133–3164.
[61] D. Malinsky and D. Danks, Causal discovery algorithms: A practical guide, *Philosophy Compass*, **13** (2017), e12470, 1–11.
[62] N. Mantel and W. Haenszel, Statistical aspects of the analysis of data from retrospective studies of disease, *Journal of the National Cancer Institute*, **22** (1959), 719–748.
[63] D. Margaritis and S. Thrun, Bayesian Network Induction via Local neighborhoods, *Adv. Neural Inf. Process. Syst.*, 505–511.

[64] C. Meek, *Graphical Models: Selecting Causal and Statistical Models*, PhD thesis.
[65] R. Moraffah, M. Karami, R. Guo, A. Raglin and H. Liu, Causal interpretability for machine learning - problems, methods and evaluation, *SIGKDD Explor. Newsl.*, **22** (2020), 18–33.
[66] R. E. Neapolitan et al., *Learning Bayesian Networks*, vol. 38, Pearson Prentice Hall Upper Saddle River, NJ, 2004.
[67] A. R. Nogueira, J. Gama and C. A. Ferreira, Improving prediction with causal probabilistic variables, in *Advances in Intelligent Data Analysis XVIII* (eds. M. R. Berthold, A. Feelders and G. Krempl), Springer International Publishing, Cham, (2020), 379–390.
[68] J. Pearl, On the intepretation of $do(x)$, *Journal of Causal Inference, Causal, Casual, and Curious Section*, **7**.
[69] J. Pearl, Bayesian networks: A model of self-activated memory for evidential reasoning, in *Proceedings of the 7th Conference of the Cognitive Science Society*, (1985), 329–334.
[70] J. Pearl, M. Glymour and N. P. Jewell, *Causal Inference in Statistics - A Primer*, John Wiley & Sons, Ltd., Chichester, 2016.
[71] J. M. Pe na, Learning gaussian graphical models of gene networks with false discovery rate control, in *European Conference on Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics*, Springer, (2008), 165–176.
[72] J. Peters and P. Bühlmann, Structural intervention distance for evaluating causal graphs, *Neural Computation*, **27** (2015), 771–799.
[73] E. Pol, Causality in economics: A menu of approaches, *Journal of Reviews on Global Economics*, **2** (2013), 356–374.
[74] V. K. Raghu, A. Poon and P. V. Benos, Evaluation of causal structure learning methods on mixed data types, *Proceedings of Machine Learning Research*, **92** (2018), 48.
[75] J. Ramsey, Improving accuracy and scalability of the pc algorithm by maximizing p-value, preprint, `arXiv:1610.00378`.
[76] J. Ramsey, M. Glymour, R. Sanchez-Romero and C. Glymour, A million variables and more: The Fast Greedy Equivalence Search algorithm for learning high-dimensional graphical causal models, with an application to functional magnetic resonance images, *Int. J. Data Sci. Anal.*, **3** (2017), 121–129.
[77] J. D. Ramsey, Scaling up greedy causal search for continuous variables, (2015).
[78] D. A. Rizzi and S. A. Pedersen, Causality in medicine: Towards a Theory and Terminology, (1992).
[79] S. Rose and D. Rizopoulos, Machine learning for causal inference in Biostatistics, *Biostatistics*, **21** (2020), 336338.
[80] D. B. Rubin, Estimating causal effects of treatments in randomized and nonrandomized studies, *Journal of Educational Psychology*, **66** (1974), 688–701.
[81] D. E. Rumelhart, G. E. Hinton and R. J. Williams, *Learning Internal Representations by Error Propagation*, Technical report, California Univ San Diego La Jolla Inst for Cognitive Science, 1985.
[82] F. Russo and J. Williamson, Interpreting causality in the health sciences, *International Studies in the Philosophy of Science*, **21** (2007), 157–170.
[83] C. Sammut and G. I. Webb, *Encyclopedia of Machine Learning*, Springer Science & Business Media, 2011.
[84] R. Scheines, P. Spirtes, C. Glymour, C. Meek and T. Richardson, Tetrad 3: Tools for causal modeling–user's manual, *CMU Philosophy*.
[85] M. Scutari, *Learning Bayesian Networks with the Bnlearn R Package*, Technical report, 2009.
[86] A. Seth, Granger causality, 2007.
[87] G. D. Smith and S. Ebrahim, Mendelian randomization: prospects, potentials, and limitations, *International Journal of Epidemiology*, **33** (2004), 30–42.
[88] E. Sokolova, D. von Rhein, J. Naaijen, P. Groot, T. Claassen, J. Buitelaar and T. Heskes, Handling hybrid and missing data in constraint-based causal discovery to study the etiology of ADHD, *International Journal of Data Science and Analytics*, **3** (2017), 105–119.
[89] P. Spirtes and K. Zhang, Causal discovery and inference: Concepts and recent methodological advances, *Applied Informatics*, **3** (2016), 1–28.
[90] P. Spirtes, An anytime algorithm for causal inference, *Proceedings of AISTATS*, 213–231.
[91] P. Spirtes, C. Glymour and R. Scheines, *Causation, Prediction and Search*, Lecture Notes in Statistics, Springer-Verlag, New York, 1993.

[92] R. Stalnaker, *Game Theory and Decision Theory (Causal and Evidential)*, Classic Philosophical Arguments, Cambridge University Press, 2018.
[93] E. W. Steyerberg, *Clinical Prediction Models: A Practical Approach to Development, Validation*, vol. 19, 2009.
[94] B. Stroud, Hume and the idea of causal necessity, *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, **33** (1978), 39–59.
[95] *What is Climatology?*, The National Drought Mitigation Center, Available from: https://drought.unl.edu/Education/DroughtIn-depth/WhatisClimatology.aspx.
[96] M. Tsagris, Bayesian network learning with the pc algorithm: An improved and correct variation, *Applied Artificial Intelligence*, **33** (2019), 101–123.
[97] I. Tsamardinos, L. E. Brown and C. F. Aliferis, The max-min hill-climbing Bayesian network structure learning algorithm, *Machine Learning*, **65** (2006), 31–78.
[98] P. Weirich, Causal decision theory, in *The Stanford Encyclopedia of Philosophy* (ed. E. N. Zalta), winter 2020 edition, Metaphysics Research Lab, Stanford University, (2020).
[99] N. Wiener, The theory of prediction, *Modern Mathematics for Engineers*, **1** (1956), 125–139.
[100] M. A. Wiering, Evolving causal neural networks, in *Benelearn'02: Proceedings of the Twelfth Belgian-Dutch Conference on Machine Learning*, (2002), 103–108.
[101] A. D. Wyner, A definition of conditional mutual information for arbitrary ensembles, *Information and Control*, **38** (1978), 51–59.
[102] H. Yamahara and H. Shimakawa, Monitoring of causal relationships on data stream using time segment characteristic, in *IEEE International Symposium on Communications and Information Technology, ISCIT 2004.*, vol. 2, (2004), 779–782.
[103] H. Yamahara and H. Shimakawa, Monitoring of causal relationships on data stream using time segment characteristic, in *IEEE International Symposium on Communications and Information Technology, 2004. ISCIT 2004.*, vol. 2, 2004, 779–782 vol.2.
[104] K. Yu, J. Li and L. Liu, *A Review on Algorithms for Constraint-based Causal Discovery*, 2016.

Received October 2019; revised March 2020.

E-mail address: ana.r.nogueira@inesctec.pt
E-mail address: jgama@fep.up.pt
E-mail address: cgf@isep.ipp.pt